Differential Geometry and Lie Groups
A Computational Perspective

Jean Gallier and Jocelyn Quaintance
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

August 7, 2019

To my daughter Mia, my wife Anne,

my son Philippe, and my daughter Sylvie.

To my parents Howard and Jane.

Preface

This book is written for a wide audience ranging from upper undergraduate to advanced graduate students in mathematics, physics, and more broadly engineering students, especially in computer science. It covers manifolds, Riemannian geometry, and Lie groups, some central topics of mathematics. However, computer vision, robotics, and machine learning, to list just a few "hot" applied areas, are increasingly consumers of differential geometry tools, so this book is also written for professionals who wish to learn about the concepts and tools from differential geometry used to solve some of their problems.

Although there are many books covering differential geometry and Lie groups, most of them assume that the reader is already quite familar with manifold theory, which is a severe obstacle for a reader who does not possess such a background. In this book, we only assume some modest background in calculus and linear algebra from the reader, and basically develop manifold theory from scratch. Additional review chapters covering some basics of analysis, in particular the notion of derivative of a map between two normed vector spaces, and some basics of topology, are provided for the reader who needs to firm up her/his background in these areas. This book is split into two parts.

1. The basic theory of manifolds and Lie groups.

2. Some of the fundamental topics of Riemannian geometry.

The culmination of the concepts and results presented in this book is the theory of *naturally reductive homogeneous manifolds and symmetric spaces*. It is remarkable that most familiar spaces are naturally reductive manifolds. Remarkably, they all arise from some suitable action of the rotation group $\mathbf{SO}(n)$, a Lie group, which emerges as the master player. The machinery of naturally reductive manifolds, and of symmetric spaces (which are even nicer!), makes it possible to compute explicitly *in terms of matrices* all the notions from differential geometry (Riemannian metrics, geodesics, *etc.*) that are needed to generalize optimization methods to Riemannian manifolds. Such methods are presented in Absil, Mahony and Sepulchre [2], and there is even a software package (MANOPT) that implements some of these procedures.

The interplay between Lie groups, manifolds, and analysis, yields a particularly effective tool. We tried to explain in some detail how these theories all come together to yield such a beautiful and useful tool.

We have also included chapters that present material having significant practical applications. These include

1. Chapter 8, on constructing manifolds from gluing data, which has applications to surface reconstruction from 3D meshes.

2. Chapter 21, on the "Log-Euclidean framework," has applications in medical imaging.

3. Chapter 22, on homogeneous reductive spaces and symmetric spaces, has applications to robotics, machine learning, and computer vision. For example, Stiefel and Grassmannian manifolds come up naturally. Furthermore, in these manifolds, it is possible to compute explicitly geodesics, Riemannian distances, gradients and Hessians. This makes it possible to actually extend optimization methods such as gradient descent and Newton's method to these manifolds. A very good source on these topics is Absil, Mahony and Sepulchre [2].

Let us now give motivations for learning the concepts and tools discussed in this book.

The need to generalize concepts and tools used in "flat spaces" such as the real line, the plane, or more generally $\mathbb{R}^n$, to more general spaces (such as a sphere) arises naturally. Such concepts and tools include

1. Defining functions.

2. Computing derivatives of functions.

3. Finding minima or maxima of functions.

4. More generally, solving optimization problems.

5. Computing the length of curves.

6. Finding shortest paths between two points.

7. Solving differential equations

8. Defining a notion of average or mean.

9. Computing areas and volumes.

10. Integrating functions.

A way to deal with a space M more complicated than $\mathbb{R}^n$ is to cover it with small pieces U_α, such that each piece U_α "looks" like $\mathbb{R}^n$, which means that there is a bijection φ_α from U_α to a subset of $\mathbb{R}^n$. Typically, M is a topological space, so the maps $\varphi_\alpha \colon U_\alpha \to \mathbb{R}^n$, called *charts*, are homeomorphisms of U_α onto some open subset of $\mathbb{R}^n$. From an intuitive point of view, locally, M looks like a piece of $\mathbb{R}^n$.

The mathematical notion formalizing the above idea is the notion of *manifold*. Having a "good" notion of what a space M is, the issue of defining real-valued functions $f\colon M \to \mathbb{R}$, and more generally functions $f\colon M \to N$ between two manifolds M and N, arises. Then it is natural to wonder what is a function with a certain degree of differentiability, and what is the derivative of a function between manifolds.

To answer these questions, one needs to add some structure to the charts $\varphi_\alpha\colon U_\alpha \to \mathbb{R}^n$, namely, whenever two charts $\varphi_\alpha\colon U_\alpha \to \mathbb{R}^n$ and $\varphi_\beta\colon U_\beta \to \mathbb{R}^n$ overlap, which means that $U_\alpha \cap U_\beta \neq \emptyset$, then the map $\varphi_\beta \circ \varphi_\alpha^{-1}$ should behave well; technically, this means that it should be C^k (continuously differentiable up to order k), or smooth.

Another important idea coming from the notion of derivative of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, is the idea of *linear approximation* of a function $f\colon M \to N$ between two manifolds. To accomplish this, we need to define the notion of *tangent space* T_pM to the manifold M at a point $p \in M$. Similarly, we have a tangent space $T_{f(p)}N$ to the manifold N at the point $f(p)$ (the image of p under f), and the *derivative of f at p* is a linear map $df_p\colon T_pM \to T_{f(p)}N$ from the tangent space T_pM (with $p \in M$) to the tangent space $T_{f(p)}N$ (with $f(p) \in N$).

Setting up carefully and rigorously the machinery to define manifolds, maps between them, tangent spaces, and the derivative of a function between manifolds, will occupy the first third of this volume.

If the manifold M is already naturally a subset of $\mathbb{R}^N$ for some N large enough, then matters are simpler, and it is easier to define manifolds, tangent spaces, and derivatives of functions between manifolds. For pedagogical reasons, we begin with this simpler case in Chapters 1–3.

If the manifold M is not embedded in $\mathbb{R}^N$ for some N, which typically occurs when M is obtained as a quotient space, such as real projective space $\mathbb{RP}^n$ (the space of lines through the origin in $\mathbb{R}^{n+1}$) or the Grassmannian $G(k,n)$ (the space of k-dimensional linear subspaces of $\mathbb{R}^n$), then matters are technically more complicated. One needs to introduce charts and atlases, and the definitions of the tangent space and of the derivative of a map between manifolds are more technical. One needs to define tangent vectors in terms of various equivalence relations (on curves, on certain triples, on germs of locally defined functions). We do this very carefully, even in the case of a C^k manifold where $1 \leq k < \infty$ (that is, a manifold which is not necessarily smooth). We give three equivalent definitions of the tangent space T_pM to M at p, and prove their equivalence. The first definition involves equivalence classes of curves through p. The third definition in terms of point derivations applies even to C^k-manifolds, at the price of introducing stationary germs. In the smooth case, this definition is equivalent to the standard definition found in Tu [112] and Warner [114]. Following J.P. Serre, the equivalence of the first and of the third definition is elegantly proved by setting up a bilinear pairing and showing that this pairing is nondegenerate. Chapters 7 and 9 are devoted to the definitions of tangent spaces, tangent bundles, vector fields, and the related concepts such as Lie derivatives and Lie brackets, in the framework of general manifolds.

Chapter 8 presents a more constructive approach for constructing manifolds using what we call gluing data. This chapter has applications to surface reconstruction from 3D meshes.

A very important class of manifolds is the class of groups that are also manifolds and topological groups (which means that multiplication and the inverse operation are smooth). Such groups are called *Lie groups*. The prime example is the group $\mathbf{SO}(3)$ of rotations in $\mathbb{R}^3$, and more generally $\mathbf{SO}(n)$. Remarkably, a large subclass of Lie groups turns out to be the family of *closed* subgroups of $\mathbf{GL}(n, \mathbb{R})$, the group of invertible $n \times n$ real matrices. This is a famous result due to Von Neumann and Cartan, see Theorem 3.8. Such closed subgroups of $\mathbf{GL}(n, \mathbb{R})$ are called *linear Lie groups* or *matrix Lie groups*. If G is a linear Lie group, then its tangent space $T_I G$ at the identity, denoted $\mathfrak{g}$, has some additional structure besides being a vector space. It has a noncommutative and nonassociative skew-symmetric multiplication $[X, Y]$ (with $X, Y \in \mathfrak{g}$) called the *Lie bracket*, which satisfies a strange kind of associativity axiom called the *Jacobi identity*. The vector space $\mathfrak{g}$ with the Lie bracket as multiplication operation has the algebraic structure of what is called a *Lie algebra*. In some sense, $\mathfrak{g}$ is a linearization of G near I, and the Lie bracket is a measure of the noncommutativity of the group operation. Remarkably, there is a way of "recovering" G from its Lie algebra $\mathfrak{g}$ by making use of the (matrix) exponential $\exp\colon \mathfrak{g} \to G$. This map is not injective nor surjective in general. In many cases of interest, such as $\mathbf{SO}(n)$ and $\mathbf{SE}(n)$, it is surjective. Also, "near" I, the exponential is bijective. Since we can move from the tangent space $T_I G = \mathfrak{g}$ at I to to the tangent space $T_g G$ at any other element $g \in G$ by left (or right) multiplication, we obtain a way of parametrizing G using the exponential map.

As a warm-up for the discussion of linear Lie groups and Lie algebras in Chapter 3, we present some properties of the exponential map of matrices in Chapters 1 and 2. In particular, we give a formula for the derivative of exp. A discussion of general Lie groups (not necessarily groups of matrices) is postponed until Chapter 18.

Another important theme of this book is the notion of *group action*. A manifold, such as the sphere S^n (in $\mathbb{R}^{n+1}$), or projective space $\mathbb{RP}^n$, or the Grassmannian $G(k, n)$, may not be a group, but may have a lot of symmetries given by a group G. For example, the sphere S^2 in $\mathbb{R}^3$ has the group of rotations $\mathbf{SO}(3)$ as group of symmetries, in the sense that a rotation in $\mathbf{SO}(3)$ moves any point on the sphere to another point on the sphere, so the sphere is invariant under rotations.

The notion of symmetry of a space under the transformations of a group G is neatly captured by the notion of *action* of a group on a set (or a manifold). A (left) action of a group G on a set X is a binary operation $\cdot\colon G \times X \to X$ satisfying the axioms

$$\begin{aligned} g_1 \cdot (g_2 \cdot x) &= (g_1 g_2) \cdot x && \text{for all } g_1, g_2 \in G \text{ and all } x \in X \\ 1 \cdot x &= x && \text{for all } x \in X. \end{aligned}$$

Here, $g_1 g_2$ denotes the product of the two elements g_1 and g_2 using the group multiplication operation on G, and 1 denotes the identity element of G. Intuitively, we can think of $g \cdot x$, where g is an element of the group G and x is an element of the set X, as the result of moving x using the "transformation" g.

A group action is *transitive* if for any two elements $x, y \in X$, there is some group element $g \in G$ that moves x to y, that is, $y = g \cdot x$. Many actions that arise in practice are transitive. For example, the group $\mathbf{SO}(3)$ acts transitively on S^2, and more generally $\mathbf{SO}(n)$ acts transitively on S^{n-1}. The reason why transitivity is important is that if we consider any fixed element $x \in X$, we can look at the *stabilizer* G_x of x, which is the set of elements of X left fixed by the action of G, namely

$$G_x = \{g \in G \mid g \cdot x = x\}.$$

It can be shown that G_x is a subgroup of G (not necessarily normal), and there is a bijection between the set G/G_x of left cosets of G and X.

This bijection is very crucial, because it allows us to view X as the set of cosets G/G_x, and if the group G is well understood, then this yields a way of inferring information about X using information about G and G_x. So far, X is a just a set, and G is just a group without any additional structure, but if X is also a topological space, and G is a topological group, then we can ask whether the quotient space G/G_x is homeomorphic to X. In general, this is not the case, but if G is a Lie group and if X is a manifold, then G/G_x is a manifold diffeomorphic to X.

The above result is very significant because it allows us to study certain manifolds M that possess a transitive action of a Lie group G in terms of the groups G and G_x. Such spaces are called *homogenous spaces*, and it turns out that many familiar manifolds such as $S^n, \mathbb{RP}^n$, the Grassmannians $G(k, n)$, the space of symmetric positive definite matrices, the Lorentz manifolds, *etc.*, are homogenous manifolds.

We begin our study of group actions and homogenous spaces in Chapter 4. We provide many examples of spaces having a transitive action, and compute explicitly stabilizers for these actions. The study of homogenous spaces is continued in greater depth, also dealing with considerations of Riemannian geometry, in Chapter 22,

As a kind of interlude, in Chapter 5, we spend some time investigating the Lorentz groups $\mathbf{O}(n, 1)$, $\mathbf{SO}(n, 1)$ and $\mathbf{SO}_0(n, 1)$ (and also the groups $\mathbf{O}(1, n)$, $\mathbf{SO}(1, n)$ and $\mathbf{SO}_0(1, n)$). When $n = 3$, these groups arise in the special theory of relativity. It turns out that $\mathbf{O}(3, 1)$ also comes up in computer vision in the study of catadioptric cameras (see Geyer [50], Chapter 5), and this was one of our original motivations for getting interested in homogeneous spaces. In Chapter 6, we also investigate the topological structure of the groups $\mathbf{O}(p, q)$, $\mathbf{SO}(p, q)$, and $\mathbf{SO}_0(p, q)$.

One feature of our exposition worth pointing out is that we give a complete proof of the surjectivity of the exponential map $\exp\colon \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3)$, for the Lorentz group $\mathbf{SO}_0(1, 3)$ (see Section 5.2, Theorem 5.18). Although we searched the literature quite thoroughly, we did not find a proof of this specific fact (the physics books we looked at, even the most reputable ones, seem to take this fact as obvious, and there are also wrong proofs; see the remark following Theorem 5.5).

We are aware of two proofs of the surjectivity of $\exp\colon \mathfrak{so}(1, n) \to \mathbf{SO}_0(1, n)$ in the general case where where n is arbitrary: One due to Nishikawa [90] (1983), and an earlier one due to Marcel Riesz [97] (1957). In both cases, the proof is quite involved (40 pages or so). In the case of $\mathbf{SO}_0(1, 3)$, a much simpler argument can be made using the fact that $\varphi\colon \mathbf{SL}(2, \mathbb{C}) \to \mathbf{SO}_0(1, 3)$ is surjective and that its kernel is $\{I, -I\}$ (see Proposition 5.17). Actually, a proof of this fact is not easy to find in the literature either (and, beware there are wrong proofs, again see the Remark following Theorem 5.5). We have made sure to provide all the steps of the proof of the surjectivity of $\exp\colon \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3)$. For more on this subject, see the discussion in Section 5.2, after Corollary 5.14.

What we have discussed above comprises the basic theory of manifolds, Lie groups, and homogenous spaces. Chapter 10 gathers some technical tools needed later such as partitions of unity and covering spaces. For the sake of the reader who feels rusty on some basics of analysis and topology, we have included two refresher chapters: Chapter 11 on power series and derivative of functions between normed vector spaces, and Chapter 12 on basics of topology. These should be consulted as nedeed, but we strongly advise the reader who has not been exposed to the notion of derivative as a linear map to review Chapter 11.

One of the main gaps in the theory of manifolds that we just sketched is that there is no way to discuss metric notions such as the notion of length of a curve segment, or the notion of angle between two curves. We are in a situation similar to the theory of vector spaces before inner products are introduced. The remedy is to add an inner product to our manifold M, but since the tangent spaces T_pM (with $p \in M$) are unrelated, we actually need to add a family $(\langle -, -\rangle_p)_{p\in M}$ of inner products, one for each tangent space T_pM. We also need to require that these inner products vary smoothly as p moves in M. A family of inner products as above is called a *Riemannian metric*, and a pair $(M, \langle -, -\rangle)$ where M is a smooth manifold and $(\langle -, -\rangle_p)_{p\in M}$ is a Riemannian metric is a *Riemannian manifold*, after B. Riemann who was the first to have this idea. If a manifold is too big, then it may not have a Riemannian metric, but "well-behaved" manifolds, namely second-countable manifolds, always have a Riemannian metric (this is shown using a partition of unity). Riemannian metrics are defined in Chapter 13. Having a Riemannian metric allows us to define the gradient, the Hessian, and the Laplacian, of a function. For functions $f\colon \mathbb{R}^n \to \mathbb{R}$, this is automatic since $\mathbb{R}^n$ is equipped with the Euclidean inner product, but for a manifold M, given a function $f\colon M \to \mathbb{R}$, to convert the linear form df_p into a vector $(\mathrm{grad}\, f)_p \in T_pM$ such that $df_p(u) = \langle (\mathrm{grad}\, f)_p, u\rangle$ for all $u \in T_pM$, an inner product is needed on T_pM, and so a Riemannian metric on M is needed.

The notion of Riemannian metric allows us to discuss metric properties of a manifold, but there is still a serious gap which has to do with the fact that given a manifold M, in general, for any two points $p, q \in M$, there is no "natural" isomorphism between the tangent spaces T_pM and T_qM. Given a curve $c\colon [0, 1] \to M$ on M, as $c(t)$ moves on M, how does the tangent space $T_{c(t)}M$ change as $c(t)$ moves?

If $M = \mathbb{R}^n$, then the spaces $T_{c(t)}\mathbb{R}^n$ are canonically isomorphic to $\mathbb{R}^n$, and any vector $v \in T_{c(0)}\mathbb{R}^n \cong \mathbb{R}^n$ is simply moved along c by *parallel transport*; that is, at $c(t)$, the tangent vector v also belongs to $T_{c(t)}\mathbb{R}^n$. However, if M is curved, for example a sphere, then it is not obvious how to "parallel transport" a tangent vector at $c(0)$ along a curve c. This problem is related to the fact that it is not obvious how to define the derivative $\nabla_X Y$ of a vector field X with respect to another vector field Y. If X and Y are vector fields on a surface S in $\mathbb{R}^3$, then for any point $p \in S$, the derivative $(D_X Y)_p$ given by

$$D_X Y(p) = \lim_{t \to 0} \frac{Y(p + tX(p)) - Y(p)}{t}.$$

(if it exists), is a vector in $\mathbb{R}^3$, but there is *no* reason why it should belong to the tangent space T_pS to S at p.

Gauss solved this problem by introducing the notion of *covariant derivative*, which consists in keeping the projection $(\nabla_Y X)$ of $(D_X Y)_p$ onto the tangent space T_pS, and to discard the normal component.

However, if M is a general manifold not embedded in $\mathbb{R}^N$, then it is not clear how to perform such a projection. Instead, the notion of covariant derivative is defined in terms of a *connection*, which is a bilinear map $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ defined on vector fields and satisfying some properties that make it a generalization of the notion of covariant derivative on a surface. The notion of connection is defined and studied in Chapter 14. Having the notion of connection, we can define the notion of *parallel vector field* along a curve, and of *parallel transport*, which allows us to relate two tangent spaces T_pM and T_qM.

The notion of covariant derivative is also well-defined for vector fields along a curve. This is shown in Section 14.2. Given a vector field X along a curve γ, this covariant derivative is denoted by DX/dt. We then have the crucial notion of a vector field *parallel along a curve* γ, which means that $DX/dt(s) = 0$ for all s (in the domain of γ).

The notion of a connection on a manifold *does not* assume that the manifold is equipped with a Riemannian metric. In Section 14.3, we consider connections having additional properties, such as being compatible with a Riemannian metric or being torsion-free. Then we have a phenomenon called by some people the "miracle" of Riemannian geometry, namely that for every Riemannian manifold, there is a *unique* connection which is torsion-free and compatible with the metric. Furthermore, this connection is determined by an implicit formula known as the *Koszul formula*. Such a connection is called the *Levi-Civita connection*.

If γ is a curve on a smooth Riemannian manifold M, and if $X = \gamma'$ is the vector field of tangent vectors γ' to γ, we can consider the curves γ that satisfy the equation

$$\frac{D\gamma'}{dt} = 0. \tag{$*$}$$

Intuitively, we can view $\frac{D\gamma'}{dt}$ as the tangent component of the acceleration vector γ'' of the curve γ, and such curves have an acceleration normal to the manifold. Curves satisfying equation $(*)$ are called *geodesics*. Geodesics are the Riemannian equivalent of straight lines in $\mathbb{R}^n$. The notion of geodesic is one of the most crucial tools in Riemannian geometry. One of the reasons is that geodesics are locally distance minimizing, and that they provide a way to parametrize a neighborhood U of any point p on a manifold M by a neighborhood of the origin in the tangent space T_pM, using the exponential map (not to be confused with the Lie group exponential) $\exp_p\colon T_pM \to M$. If the exponential map is surjective, then the manifold M is said to be *complete*. A beautiful theorem of Hopf and Rinow states that if a manifold is complete, then any two points can be joined by a minimal geodesic (a geodesic of minimal length). This is an important property because the shortest distance between any two points is achieved by a geodesic. Compact Riemannian manifolds are complete, so many of the familiar compact manifolds (S^n, $\mathbb{RP}^n$, $G(k,n)$) are complete.

Given a curve ω on a Riemannian manifold, the quantity $E(\omega) = \int_0^1 \|\omega'(t)\|^2\,dt$ is called the *energy function*. Geodesics between two points p and q turn out to be critical points of the energy function E on the path space $\Omega(p,q)$ of all piecewise smooth curves from p to q. To define the notion of critical point of the energy function, because the space $\Omega(p,q)$ is not a finite-dimensional manifold, it is necessary to introduce the notion of *variation* of a curve and to prove the *first variation formula*. Here, we make a link with the calculus of variation. Geodesics are studied throroughly in Chapter 15.

Riemannian metrics, connections, and geodesics, are three of the pilars of differential geometry. The fourth pilar is curvature.

For surfaces, the notion of curvature can be defined in terms of the curvatures of curves drawn on the surface. The notion of Gaussian curvature (of course, introduced by Gauss) gives a satisfactory answer. However, for manifolds of dimension greater than 2, it is not obvious what curvature means. Riemann proposed a definition involving the notion of sectional curvature, but his seminal paper (1868) did not contain proofs and did not give a general method to compute such a curvature. It is only fifty years later that the idea emerged that the curvature of a Riemannian manifold should be viewed as a measure $R(X,Y)Z$ of the extent to which the operator $(X,Y) \mapsto \nabla_X\nabla_Y Z$ is symmetric.

The *Riemann curvature operator* R turns out to be C^∞-linear in all of its three arguments, but it is a rather complicated object. Fortunately, there is a simpler object, the *sectional curvature* $K(u,v)$. When ∇ is the Levi-Civita connection, the curvature operator R can be recovered from the sectional curvature K. There is also an important simpler notion of curvature $\mathrm{Ric}(x,y)$, called the *Ricci curvature*, which arises as the trace of the linear map $v \mapsto R(x,v)y$. An even cruder notion of curvature is the *scalar curvature*. These notions of curvature are discussed in Chapter 16.

We pointed out earlier that the energy function $E(\omega) = \int_0^1 \|\omega'(t)\|^2\,dt$ determines the geodesics (between two fixed points p and q) in the sense that its critical points are the geodesics. A deeper understanding of the energy function is achieved by investigating the second derivative of E at critical points. To do this we need the notion of 2-parameter variation and the *second variation formula*. The curvature operator shows up in this formula. Another important technical tool is the notion of Jacobi fields, which are induced by geodesic variations. Jacobi fields can be used to compute the sectional curvature of various manifolds.

Another important theme of differential geometry is the influence of curvature (sectional or Ricci) on the topology of a Riemannian manifold. This is a vast subject and we only discuss three results, one of which being the Hadamard and Cartan theorem about complete manifolds of non-positive curvature.

The goal of Chapter 17 is to understand the behavior of isometries and local isometries, in particular their action on geodesics. We also intoduce Riemannian covering maps and Riemannian submersions. If $\pi\colon M \to B$ is a submersion between two Riemannian manifolds, then for every $b \in B$ and every $p \in \pi^{-1}(b)$, the tangent space T_pM to M at p splits into two orthogonal components, its *vertical component* $\mathcal{V}_p = \mathrm{Ker}\, d\pi_p$, and its *horizontal component* $\mathcal{H}_p$ (the orthogonal complement of $\mathcal{V}_p$). If the map $d\pi_p$ is an isometry between $\mathcal{H}_p$ and T_bB, then most of the differential geometry of B can be studied by lifting B to M, and then projecting down to B again. We also introduce Killing vector fields, which play a technical role in the study of reductive homogeneous spaces.

In Chapter 18, we return to Lie groups. Not every Lie group is a matrix group, so in order to study general Lie groups it is necessary to introduce left-invariant (and right-invariant) vector fields on Lie groups. It turns out that the space of left-invariant vector fields is isomorphic to the tangent space $\mathfrak{g} = T_IG$ to G at the identity, which is a Lie algebra. By considering integral curves of left-invariant vector fields, we define the generalization of the exponential map $\exp\colon \mathfrak{g} \to G$ to an arbitrary Lie group. The notion of immersed Lie subgroup is introduced, and the correspondence between Lie groups and Lie algebra is explored. We also consider the special classes of semidirect products of Lie algebras and Lie groups, the universal covering of a Lie group, and the Lie algebra of Killing vector fields on a Riemannian manifold.

Chapter 19 deals with two topics:

1. A formula for the derivative of the exponential map for a general Lie group (not necessarily a matrix group).

2. A formula for the Taylor expansion of $\mu(X, Y) = \log(\exp(X)\exp(Y))$ near the origin.

The second problem is solved by a formula known as the *Campbell-Baker-Hausdorff formula*. An explicit formula was derived by Dynkin (1947), and we present this formula.

Chapter 20 is devoted to the study of metrics, connections, geodesics, and curvature, on Lie groups. Since a Lie group G is a smooth manifold, we can endow G with a Riemannian metric. Among all the Riemannian metrics on a Lie groups, those for which the left translations (or the right translations) are isometries are of particular interest because they take the group structure of G into account. As a consequence, it is possible to find explicit formulae for the Levi-Civita connection and the various curvatures, especially in the case of metrics which are both left and right-invariant.

In Section 20.2 we give four characterizations of bi-invariant metrics. The first one refines the criterion of the existence of a left-invariant metric and states that every bi-invariant metric on a Lie group G arises from some Ad-invariant inner product on the Lie algebra $\mathfrak{g}$.

In Section 20.3 we show that if G is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of G, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae are be obtained. In particular the geodesics coincide with the one-parameter groups induced by left-invariant vector fields.

Section 20.5 introduces *simple* and *semisimple* Lie algebras. They play a major role in the structure theory of Lie groups

Section 20.6 is devoted to the *Killing form*. It is an important concept, and we establish some of its main properties. Remarkably, the Killing form yields a simple criterion due to Élie Cartan for testing whether a Lie algebra is semisimple.

We conclude this chapter with a section on Cartan connections (Section 20.7). Unfortunately, if a Lie group G does not admit a bi-invariant metric, under the Levi-Civita connection, geodesics are generally not given by the exponential map $\exp\colon \mathfrak{g} \to G$. If we are willing to consider connections not induced by a metric, then it turns out that there is a fairly natural connection for which the geodesics coincide with integral curves of left-invariant vector fields. These connections are called *Cartan connections*. This chapter makes extensive use of results from a beautiful paper of Milnor [84].

In Chapter 21 we present an application of Lie groups and Riemannian geometry. We describe an approach due to Arsigny, Fillard, Pennec and Ayache, to define a Lie group structure and a class of metrics on symmetric, positive-definite matrices (SPD matrices) which yield a new notion of mean on SPD matrices generalizing the standard notion of geometric mean.

SPD matrices are used in diffusion tensor magnetic resonance imaging (for short, DTI), and they are also a basic tool in numerical analysis, for example, in the generation of meshes to solve partial differential equations more efficiently. As a consequence, there is a growing need to interpolate or to perform statistics on SPD matrices, such as computing the mean of a finite number of SPD matrices.

Chapter 22 provides the culmination of the theory presented in the book, the concept of a *homogeneous naturally reductive space*.

The goal is to study the differential geometry of a manifold M presented as the quotient G/H of a Lie group G by a closed subgroup H. We would like to endow G/H with a metric that arises from an inner product on the Lie algebra $\mathfrak{g}$ of G. To do this, we consider G-invariant metrics, which are metrics on G/H such that the left multiplication operations $\tau_g\colon G/H \to G/H$ given by

$$\tau_g(h_2H) = gg_2H$$

are isometries. The existence of G-invariant metrics on G/H depends on properties of a certain representation of H called the isotropy representation (see Proposition 22.21). The isotropy representation is equivalent to another representation $\mathrm{Ad}^{G/H}\colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ of H involving the quotient algebra $\mathfrak{g}/\mathfrak{h}$.

This representation is too complicated to deal with, so we consider the more tractable situation where the Lie algebra $\mathfrak{g}$ of G factors as a direct sum

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m},$$

for some subspace $\mathfrak{m}$ of $\mathfrak{g}$ such that $\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $h \in H$, where $\mathfrak{h}$ is the Lie algebra of H. Then $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, and the representation $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ becomes the representation $\mathrm{Ad} \colon H \to \mathbf{GL}(\mathfrak{m})$, where Ad_h is the restriction of Ad_h to $\mathfrak{m}$ for every $h \in H$. In this situation there is an isomorphism between $T_o(G/H)$ and $\mathfrak{m}$ (where o denotes the point in G/H corresponding to the coset H). It is also the case that if H is "nice" (for example, compact), then $M = G/H$ will carry G-invariant metrics, and that under such metrics, the projection $\pi \colon G \to G/H$ is a Riemannian submersion.

It is remarkable that a simple condition on $\mathfrak{m}$, namely $\mathrm{Ad}(H)$ invariance, yields a one-to-one correspondence between G-invariant metrics on G/H and $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$ (see Proposition 22.22). This is a generalization of the situation of Proposition 20.3 characterizing the existence of bi-invariant metrics on Lie groups. All this is built into the definition of a *reductive homogeneous space* given by Definition 22.8.

It is possible to express the Levi-Civita connection on a reductive homogeneous space in terms of the Lie bracket on $\mathfrak{g}$, but in general this formula is not very useful. A simplification of this formula is obtained if a certain condition holds. The corresponding spaces are said to be *naturally reductive*; see Definition 22.9. A naturally reductive space has the "nice" property that its geodesics at o are given by applying the coset exponential map to $\mathfrak{m}$; see Proposition 22.27. As we will see from the explicit examples provided in Section 22.7, naturally reductive spaces "behave" just as nicely as their Lie group counterpart G, and the coset exponential of $\mathfrak{m}$ will provide *all* the necessary geometric information.

A large supply of naturally reductive homogeneous spaces are the *symmetric spaces*. Such spaces arise from a Lie group G equipped with an involutive automorphism $\sigma \colon G \to G$ (with $\sigma \neq \mathrm{id}$ and $\sigma^2 = \mathrm{id}$). Let G^σ be the set of fixed points of σ, the subgroup of G given by

$$G^\sigma = \{g \in G \mid \sigma(g) = g\},$$

and let G^σ_0 be the identity component of G^σ (the connected component of G^σ containing 1). Consider the $+1$ and -1 eigenspaces of the derivative $d\sigma_1 \colon \mathfrak{g} \to \mathfrak{g}$ of σ, given by

$$\begin{aligned}\mathfrak{k} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}\\ \mathfrak{m} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = -X\}.\end{aligned}$$

Pick a closed subgroup K of G such that $G^\sigma_0 \subseteq K \subseteq G^\sigma$. Then it can be shown that G/K is a reductive homogenous space and that $\mathfrak{g}$ factors as a direct sum $\mathfrak{k} \oplus \mathfrak{m}$, which makes G/K a reductive space. Furthermore, if G is connected and if both G^σ_0 and K are compact, then G/K is naturally reductive.

There is an extensive theory of symmetric spaces and our goal is simply to show that the additional structure afforded by an involutive automorphism of G yields spaces that are naturally reductive. The theory of symmetric spaces was entirely created by one person, Élie Cartan, who accomplished the tour de force of giving a complete classification of these spaces using the classification of semisimple Lie algebras that he had obtained earlier. In Sections 22.8, 22.9, and 22.10, we provide an introduction to symmetric spaces.

In the past five years, we have also come to realize that *Lie groups* and *homogeneous manifolds*, especially naturally reductive ones, are two of the most important topics for their role in applications. It is remarkable that most familiar spaces, spheres, projective spaces, Grassmannian and Stiefel manifolds, symmetric positive definite matrices, are naturally reductive manifolds. Remarkably, they all arise from some suitable action of the rotation group $\mathbf{SO}(n)$, a Lie group, who emerges as the master player. The machinery of naturally reductive manifolds, and of symmetric spaces (which are even nicer!), makes it possible to compute explicitly *in terms of matrices* all the notions from differential geometry (Riemannian metrics, geodesics, *etc.*) that are needed to generalize optimization methods to Riemannian manifolds.

Since we discuss many topics ranging from manifolds to Lie groups, this book is already quite big, so we resolved ourselves, not without regrets, to omit many proofs. The purist may be chagrined, but we feel that it is more important to motivate, demystify, and explain, the reasons for introducing various concepts and to clarify the relationship between these notions rather than spelling out every proof in full detail. Whenever we omit a proof, we provide precise pointers to the literature. In some cases (such as the theorem of Hopf and Rinow), the proof is just too beautiful to be skipped, so we include it.

The motivations for writing these notes arose while the first author was coteaching a seminar on Special Topics in Machine Perception with Kostas Daniilidis in the Spring of 2004. In the Spring of 2005, the first author gave a version of his course *Advanced Geometric Methods in Computer Science* (CIS610), with the main goal of discussing statistics on diffusion tensors and shape statistics in medical imaging. This is when he realized that it was necessary to cover some material on Riemannian geometry but he ran out of time after presenting Lie groups and never got around to doing it! Then, in the Fall of 2006 the first author went on a wonderful and very productive sabbatical year in Nicholas Ayache's group (ACSEPIOS) at INRIA Sophia Antipolis, where he learned about the beautiful and exciting work of Vincent Arsigny, Olivier Clatz, Hervé Delingette, Pierre Fillard, Grégoire Malandin, Xavier Pennec, Maxime Sermesant, and, of course, Nicholas Ayache, on statistics on manifolds and Lie groups applied to medical imaging. This inspired him to write chapters on differential geometry, and after a few additions made during Fall 2007 and Spring 2008, notably on left-invariant metrics on Lie groups, the little set of notes from 2004 had grown into a preliminary version of this manuscript. The first author then joined forces with the second author in 2015, and with her invaluable assistance, produced the present book, as well, as a second volume dealing with more advanced topics.

We must acknowledge our debt to two of our main sources of inspiration: Berger's *Panoramic View of Riemannian Geometry* [14] and Milnor's *Morse Theory* [81]. In our opinion, Milnor's book is still one of the best references on basic differential geometry. His exposition is remarkably clear and insightful, and his treatment of the variational approach to geodesics is unsurpassed. We borrowed heavily from Milnor [81]. Since Milnor's book is typeset in "ancient" typewritten format (1973!), readers might enjoy reading parts of it typeset in LaTeX. We hope that the readers of these notes will be well prepared to read standard differential geometry texts such as do Carmo [39], Gallot, Hulin, Lafontaine [49] and O'Neill [91], but also more advanced sources such as Sakai [100], Petersen [93], Jost [64], Knapp [68], and of course Milnor [81].

The chapters or sections marked with the symbol $\circledast$ contain material that is typically more specialized or more advanced, and they can be omitted upon first (or second) reading.

Acknowledgement: We would like to thank Eugenio Calabi, Ching-Li Chai, Ted Chinburg, Chris Croke, Ron Donagi, Harry Gingold, H.W. Gould, Herman Gluck, David Harbater, Julia Hartmann, Jerry Kazdan, Alexander Kirillov, Florian Pop, Steve Shatz, Jim Stasheff, George Sparling, Doran Zeilberger, and Wolfgand Ziller for their encouragement, advice, inspiration and for what they taught me. We also thank Christine Allen-Blanchette, Arthur Azevedo de Amorim, Kostas Daniilidis, Carlos Esteves, Spyridon Leonardos, Stephen Phillips, João Sedoc, Marcelo Siqueira, and Roberto Tron for reporting typos and for helpful comments.

Contents

Part I

Introduction to Differential Manifolds and Lie Groups

Chapter 1

The Matrix Exponential; Some Matrix Lie Groups

> Le rôle prépondérant de la théorie des groupes en mathématiques a été longtemps insoupçonné; il y a quatre-vingts ans, le nom même de groupe était ignoré. C'est Galois qui, le premier, en a eu une notion claire, mais c'est seulement depuis les travaux de Klein et surtout de Lie que l'on a commencé à voir qu'il n'y a presque aucune théorie mathématique où cette notion ne tienne une place importante.
> **—Henri Poincaré**

The purpose of this chapter and the next two chapters is to give a "gentle" and fairly concrete introduction to manifolds, Lie groups and Lie algebras, our main objects of study.

Most texts on Lie groups and Lie algebras begin with prerequisites in differential geometry that are often formidable to average computer scientists (or average scientists, whatever that means!). We also struggled for a long time, trying to figure out what Lie groups and Lie algebras are all about, but this can be done! A good way to sneak into the wonderful world of Lie groups and Lie algebras is to play with explicit matrix groups such as the group of rotations in $\mathbb{R}^2$ (or $\mathbb{R}^3$) and with the exponential map. After actually computing the exponential $A = e^B$ of a 2×2 skew symmetric matrix B and observing that it is a rotation matrix, and similarly for a 3×3 skew symmetric matrix B, one begins to suspect that there is something deep going on. Similarly, after the discovery that every real invertible $n \times n$ matrix A can be written as $A = RP$, where R is an orthogonal matrix and P is a positive definite symmetric matrix, and that P can be written as $P = e^S$ for some symmetric matrix S, one begins to appreciate the exponential map.

Our goal in this chapter is to give an elementary and concrete introduction to Lie groups and Lie algebras by studying a number of the so-called *classical groups*, such as the general linear group $\mathbf{GL}(n, \mathbb{R})$, the special linear group $\mathbf{SL}(n, \mathbb{R})$, the orthogonal group $\mathbf{O}(n)$, the special orthogonal group $\mathbf{SO}(n)$, and the group of affine rigid motions $\mathbf{SE}(n)$, and their Lie algebras $\mathfrak{gl}(n, \mathbb{R})$ (all matrices), $\mathfrak{sl}(n, \mathbb{R})$ (matrices with null trace), $\mathfrak{o}(n)$, and $\mathfrak{so}(n)$ (skew symmetric matrices). Lie groups are at the same time, groups, topological spaces, and manifolds, so we will also have to introduce the crucial notion of a *manifold*.

The inventors of Lie groups and Lie algebras (starting with Lie!) regarded Lie groups as groups of symmetries of various topological or geometric objects. Lie algebras were viewed as the "infinitesimal transformations" associated with the symmetries in the Lie group. For example, the group $\mathbf{SO}(n)$ of rotations is the group of orientation-preserving isometries of the Euclidean space $\mathbb{E}^n$. The Lie algebra $\mathfrak{so}(n, \mathbb{R})$ consisting of real skew symmetric $n \times n$ matrices is the corresponding set of infinitesimal rotations. The geometric link between a Lie group and its Lie algebra is the fact that the Lie algebra can be viewed as the tangent space to the Lie group at the identity. There is a map from the tangent space to the Lie group, called the *exponential map*. The Lie algebra can be considered as a linearization of the Lie group (near the identity element), and the exponential map provides the "delinearization," i.e., it takes us back to the Lie group. These concepts have a concrete realization in the case of groups of matrices and, for this reason, we begin by studying the behavior of the exponential maps on matrices.

We begin by defining the exponential map on matrices and proving some of its properties. The exponential map allows us to "linearize" certain algebraic properties of matrices. It also plays a crucial role in the theory of linear differential equations with constant coefficients. But most of all, as we mentioned earlier, it is a stepping stone to Lie groups and Lie algebras. On the way to Lie algebras, we derive the classical "Rodrigues-like" formulae for rotations and for rigid motions in $\mathbb{R}^2$ and $\mathbb{R}^3$. We give an elementary proof that the exponential map is surjective for both $\mathbf{SO}(n)$ and $\mathbf{SE}(n)$, not using any topology, just certain normal forms for matrices (see Gallier [48], Chapters 12 and 13).

In Chapter 2, in preparation for defining the Lie bracket on the Lie algebra of a Lie group, we introduce the adjoint representations of the group $\mathbf{GL}(n, \mathbb{R})$ and of the Lie algebra $\mathfrak{gl}(n, \mathbb{R})$. The map $\mathrm{Ad}\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(\mathfrak{gl}(n, \mathbb{R}))$ is defined such that Ad_A is the derivative of the conjugation map $\mathbf{Ad}_A\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(n, \mathbb{R})$ at the identity. The map ad is the derivative of Ad at the identity, and it turns out that $\mathrm{ad}_A(B) = [A, B]$, the Lie bracket of A and B, and in this case, $[A, B] = AB - BA$. We also find a formula for the derivative of the matrix exponential $\mathbf{exp}$.

Chapter 3 gives an introduction to manifolds, Lie groups and Lie algebras. Rather than defining abstract manifolds in terms of charts, atlases, *etc.*, we consider the special case of embedded submanifolds of $\mathbb{R}^N$. This approach has the pedagogical advantage of being more concrete since it uses parametrizations of subsets of $\mathbb{R}^N$, which should be familiar to the reader in the case of curves and surfaces. The general definition of a manifold will be given in Chapter 7.

Also, rather than defining Lie groups in full generality, we define linear Lie groups using the famous result of Cartan (apparently actually due to Von Neumann) that a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$ is a manifold, and thus a Lie group. This way, Lie algebras can be "computed" using tangent vectors to curves of the form $t \mapsto A(t)$, where $A(t)$ is a matrix.

This chapter is inspired from Artin [10], Chevalley [31], Marsden and Ratiu [77], Curtis [34], Howe [62], and Sattinger and Weaver [102].

1.1 The Exponential Map

Given an $n \times n$ (real or complex) matrix $A = (a_{i\,j})$, we would like to define the exponential e^A of A as the sum of the series

$$e^A = I_n + \sum_{p\geq 1} \frac{A^p}{p!} = \sum_{p\geq 0} \frac{A^p}{p!},$$

letting $A^0 = I_n$. The problem is, Why is it well-defined? The following proposition shows that the above series is indeed absolutely convergent. For the definition of absolute convergence see Chapter 2, Section 1.

Proposition 1.1. *Let $A = (a_{i\,j})$ be a (real or complex) $n \times n$ matrix, and let*

$$\mu = \max\{|a_{i\,j}| \mid 1 \leq i, j \leq n\}.$$

If $A^p = (a^{(p)}_{i\,j})$, then

$$\left|a^{(p)}_{i\,j}\right| \leq (n\mu)^p$$

for all i, j, $1 \leq i, j \leq n$. As a consequence, the n^2 series

$$\sum_{p\geq 0} \frac{a^{(p)}_{i\,j}}{p!}$$

converge absolutely, and the matrix

$$e^A = \sum_{p\geq 0} \frac{A^p}{p!}$$

is a well-defined matrix.

Proof. The proof is by induction on p. For $p = 0$, we have $A^0 = I_n$, $(n\mu)^0 = 1$, and the proposition is obvious. Assume that

$$|a^{(p)}_{i\,j}| \leq (n\mu)^p$$

for all i, j, $1 \leq i, j \leq n$. Then we have

$$\left|a^{(p+1)}_{i\,j}\right| = \left|\sum_{k=1}^{n} a^{(p)}_{i\,k} a_{k\,j}\right| \leq \sum_{k=1}^{n} \left|a^{(p)}_{i\,k}\right| \left|a_{k\,j}\right| \leq \mu \sum_{k=1}^{n} \left|a^{(p)}_{i\,k}\right| \leq n\mu(n\mu)^p = (n\mu)^{p+1},$$

for all i, j, $1 \leq i, j \leq n$. For every pair (i, j) such that $1 \leq i, j \leq n$, since

$$\left|a_{i\,j}^{(p)}\right| \leq (n\mu)^p,$$

the series

$$\sum_{p \geq 0} \frac{\left|a_{i\,j}^{(p)}\right|}{p!}$$

is bounded by the convergent series

$$e^{n\mu} = \sum_{p \geq 0} \frac{(n\mu)^p}{p!},$$

and thus it is absolutely convergent. This shows that

$$e^A = \sum_{k \geq 0} \frac{A^k}{k!}$$

is well defined. □

It is instructive to compute explicitly the exponential of some simple matrices. As an example, let us compute the exponential of the real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}.$$

We need to find an inductive formula expressing the powers A^n. Let us observe that

$$\begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix} = \theta \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 & -\theta \\ \theta & 0 \end{pmatrix}^2 = -\theta^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then letting

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

we have

$$\begin{aligned} A^{4n} &= \theta^{4n} I_2, \\ A^{4n+1} &= \theta^{4n+1} J, \\ A^{4n+2} &= -\theta^{4n+2} I_2, \\ A^{4n+3} &= -\theta^{4n+3} J, \end{aligned}$$

and so

$$e^A = I_2 + \frac{\theta}{1!} J - \frac{\theta^2}{2!} I_2 - \frac{\theta^3}{3!} J + \frac{\theta^4}{4!} I_2 + \frac{\theta^5}{5!} J - \frac{\theta^6}{6!} I_2 - \frac{\theta^7}{7!} J + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \cdots\right) I_2 + \left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots\right) J.$$

We recognize the power series for $\cos\theta$ and $\sin\theta$, and thus

$$e^A = \cos\theta I_2 + \sin\theta J,$$

that is

$$e^A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Thus, e^A is a rotation matrix! This is a general fact. If A is a skew symmetric matrix, then e^A is an orthogonal matrix of determinant $+1$, i.e., a rotation matrix. Furthermore, every rotation matrix is of this form; i.e., the exponential map from the set of skew symmetric matrices to the set of rotation matrices is surjective. In order to prove these facts, we need to establish some properties of the exponential map.

But before that, let us work out another example showing that the exponential map is not always surjective. Let us compute the exponential of a real 2×2 matrix with null trace of the form

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}.$$

We need to find an inductive formula expressing the powers A^n. Observe that

$$A^2 = (a^2 + bc)I_2 = -\det(A)I_2.$$

If $a^2 + bc = 0$, we have

$$e^A = I_2 + A.$$

If $a^2 + bc < 0$, let $\omega > 0$ be such that $\omega^2 = -(a^2 + bc)$. Then, $A^2 = -\omega^2 I_2$. We get

$$e^A = I_2 + \frac{A}{1!} - \frac{\omega^2}{2!}I_2 - \frac{\omega^2}{3!}A + \frac{\omega^4}{4!}I_2 + \frac{\omega^4}{5!}A - \frac{\omega^6}{6!}I_2 - \frac{\omega^6}{7!}A + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 - \frac{\omega^2}{2!} + \frac{\omega^4}{4!} - \frac{\omega^6}{6!} + \cdots\right) I_2 + \frac{1}{\omega}\left(\omega - \frac{\omega^3}{3!} + \frac{\omega^5}{5!} - \frac{\omega^7}{7!} + \cdots\right) A.$$

We recognize the power series for $\cos\omega$ and $\sin\omega$, and thus

$$e^A = \cos\omega\, I_2 + \frac{\sin\omega}{\omega}A = \begin{pmatrix} \cos\omega + \frac{\sin\omega}{\omega}a & \frac{\sin\omega}{\omega}b \\ \frac{\sin\omega}{\omega}c & \cos\omega - \frac{\sin\omega}{\omega}a \end{pmatrix}.$$

Note that

$$\begin{aligned}\det(e^A) &= \left(\cos\omega + \frac{\sin\omega}{\omega}a\right)\left(\cos\omega - \frac{\sin\omega}{\omega}a\right) - \frac{\sin^2\omega}{\omega^2}bc\\ &= \cos^2\omega - \frac{\sin^2\omega}{\omega^2}(a^2+bc) = \cos^2\omega + \sin^2\omega = 1.\end{aligned}$$

If $a^2 + bc > 0$, let $\omega > 0$ be such that $\omega^2 = a^2 + bc$. Then $A^2 = \omega^2 I_2$. We get

$$e^A = I_2 + \frac{A}{1!} + \frac{\omega^2}{2!}I_2 + \frac{\omega^2}{3!}A + \frac{\omega^4}{4!}I_2 + \frac{\omega^4}{5!}A + \frac{\omega^6}{6!}I_2 + \frac{\omega^6}{7!}A + \cdots.$$

Rearranging the order of the terms, we have

$$e^A = \left(1 + \frac{\omega^2}{2!} + \frac{\omega^4}{4!} + \frac{\omega^6}{6!} + \cdots\right)I_2 + \frac{1}{\omega}\left(\omega + \frac{\omega^3}{3!} + \frac{\omega^5}{5!} + \frac{\omega^7}{7!} + \cdots\right)A.$$

If we recall that $\cosh\omega = \big(e^\omega + e^{-\omega}\big)/2$ and $\sinh\omega = \big(e^\omega - e^{-\omega}\big)/2$, we recognize the power series for $\cosh\omega$ and $\sinh\omega$, and thus

$$e^A = \cosh\omega\, I_2 + \frac{\sinh\omega}{\omega}A = \begin{pmatrix}\cosh\omega + \frac{\sinh\omega}{\omega}a & \frac{\sinh\omega}{\omega}b\\ \frac{\sinh\omega}{\omega}c & \cosh\omega - \frac{\sinh\omega}{\omega}a\end{pmatrix},$$

and

$$\begin{aligned}\det(e^A) &= \left(\cosh\omega + \frac{\sinh\omega}{\omega}a\right)\left(\cosh\omega - \frac{\sinh\omega}{\omega}a\right) - \frac{\sinh^2\omega}{\omega^2}bc\\ &= \cosh^2\omega - \frac{\sinh^2\omega}{\omega^2}(a^2+bc) = \cosh^2\omega - \sinh^2\omega = 1.\end{aligned}$$

In both cases

$$\det\left(e^A\right) = 1.$$

This shows that the exponential map is a function from the set of 2×2 matrices with null trace to the set of 2×2 matrices with determinant 1. This function is not surjective. Indeed, $\mathrm{tr}(e^A) = 2\cos\omega$ when $a^2 + bc < 0$, $\mathrm{tr}(e^A) = 2\cosh\omega$ when $a^2 + bc > 0$, and $\mathrm{tr}(e^A) = 2$ when $a^2 + bc = 0$. As a consequence, for any matrix A with null trace,

$$\mathrm{tr}\left(e^A\right) \geq -2,$$

and any matrix B with determinant 1 and whose trace is less than -2 is not the exponential e^A of any matrix A with null trace. For example,

$$B = \begin{pmatrix}a & 0\\ 0 & a^{-1}\end{pmatrix},$$

where $a < 0$ and $a \neq -1$, is not the exponential of any matrix A with null trace since

$$\frac{(a+1)^2}{a} = \frac{a^2+2a+1}{a} = \frac{a^2+1}{a} + 2 < 0,$$

which in turn implies $\mathrm{tr}(B) = a + \frac{1}{a} = \frac{a^2+1}{a} < -2$.

A fundamental property of the exponential map is that if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then the eigenvalues of e^A are $e^{\lambda_1}, \ldots, e^{\lambda_n}$. For this we need two propositions.

Proposition 1.2. *Let A and U be (real or complex) matrices, and assume that U is invertible. Then*

$$e^{UAU^{-1}} = Ue^AU^{-1}.$$

Proof. A trivial induction shows that

$$UA^pU^{-1} = (UAU^{-1})^p,$$

and thus

$$\begin{aligned} e^{UAU^{-1}} &= \sum_{p\geq 0} \frac{(UAU^{-1})^p}{p!} = \sum_{p\geq 0} \frac{UA^pU^{-1}}{p!} \\ &= U\left(\sum_{p\geq 0} \frac{A^p}{p!}\right)U^{-1} = Ue^AU^{-1}. \end{aligned}$$

□

Say that a square matrix A is an *upper triangular matrix* if it has the following shape,

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & a_{1\,3} & \ldots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & a_{2\,3} & \ldots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{3\,3} & \ldots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \ldots & 0 & a_{n\,n} \end{pmatrix},$$

i.e., $a_{i\,j} = 0$ whenever $j < i$, $1 \leq i, j \leq n$.

Proposition 1.3. *Given any complex $n \times n$ matrix A, there is an invertible matrix P and an upper triangular matrix T such that*

$$A = PTP^{-1}.$$

Proof. We prove by induction on n that if $f\colon \mathbb{C}^n \to \mathbb{C}^n$ is a linear map, then there is a basis $(u_1, \ldots, u_n)$ with respect to which f is represented by an upper triangular matrix. For $n = 1$ the result is obvious. If $n > 1$, since $\mathbb{C}$ is algebraically closed, f has some eigenvalue $\lambda_1 \in \mathbb{C}$, and let u_1 be an eigenvector for λ_1. We can find $n-1$ vectors $(v_2, \ldots, v_n)$ such that $(u_1, v_2, \ldots, v_n)$ is a basis of $\mathbb{C}^n$, and let W be the subspace of dimension $n-1$ spanned by $(v_2, \ldots, v_n)$. In the basis $(u_1, v_2 \ldots, v_n)$, the matrix of f is of the form

$$\begin{pmatrix} a_{1\,1} & a_{1\,2} & \ldots & a_{1\,n} \\ 0 & a_{2\,2} & \ldots & a_{2\,n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n\,2} & \ldots & a_{n\,n} \end{pmatrix},$$

since its first column contains the coordinates of $\lambda_1 u_1$ over the basis $(u_1, v_2, \ldots, v_n)$. Letting $p\colon \mathbb{C}^n \to W$ be the projection defined such that $p(u_1) = 0$ and $p(v_i) = v_i$ when $2 \leq i \leq n$, the linear map $g\colon W \to W$ defined as the restriction of $p \circ f$ to W is represented by the $(n-1) \times (n-1)$ matrix $(a_{i\,j})_{2 \leq i,j \leq n}$ over the basis $(v_2, \ldots, v_n)$. By the induction hypothesis, there is a basis $(u_2, \ldots, u_n)$ of W such that g is represented by an upper triangular matrix $(b_{i\,j})_{1 \leq i,j \leq n-1}$.

However,

$$\mathbb{C}^n = \mathbb{C}u_1 \oplus W,$$

and thus $(u_1, \ldots, u_n)$ is a basis for $\mathbb{C}^n$. Since p is the projection from $\mathbb{C}^n = \mathbb{C}u_1 \oplus W$ onto W and $g\colon W \to W$ is the restriction of $p \circ f$ to W, we have

$$f(u_1) = \lambda_1 u_1$$

and

$$f(u_{i+1}) = a_{1\,i} u_1 + \sum_{j=1}^{n-1} b_{i\,j} u_{j+1}$$

for some $a_{1\,i} \in \mathbb{C}$, when $1 \leq i \leq n-1$. But then the matrix of f with respect to $(u_1, \ldots, u_n)$ is upper triangular. Thus, there is a change of basis matrix P such that $A = PTP^{-1}$ where T is upper triangular. □

Remark: If E is a Hermitian space, the proof of Proposition 1.3 can be easily adapted to prove that there is an *orthonormal* basis $(u_1, \ldots, u_n)$ with respect to which the matrix of f is upper triangular. In terms of matrices, this means that there is a unitary matrix U and an upper triangular matrix T such that $A = UTU^*$. This is usually known as *Schur's lemma*. Using this result, we can immediately rederive the fact that if A is a Hermitian matrix, i.e. $A = A^*$, then there is a unitary matrix U and a real diagonal matrix D such that $A = UDU^*$.

If $A = PTP^{-1}$ where T is upper triangular, then A and T have the same characteristic polynomial. This is because if A and B are any two matrices such that $A = PBP^{-1}$, then

$$\begin{aligned}
\det(A - \lambda I) &= \det(PBP^{-1} - \lambda P I P^{-1}),\\
&= \det(P(B - \lambda I)P^{-1}),\\
&= \det(P)\det(B - \lambda I)\det(P^{-1}),\\
&= \det(P)\det(B - \lambda I)\det(P)^{-1},\\
&= \det(B - \lambda I).
\end{aligned}$$

Furthermore, it is well known that the determinant of a matrix of the form

$$\begin{pmatrix}
\lambda_1 - \lambda & a_{1\,2} & a_{1\,3} & \dots & a_{1\,n-1} & a_{1\,n}\\
0 & \lambda_2 - \lambda & a_{2\,3} & \dots & a_{2\,n-1} & a_{2\,n}\\
0 & 0 & \lambda_3 - \lambda & \dots & a_{3\,n-1} & a_{3\,n}\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \dots & \lambda_{n-1} - \lambda & a_{n-1\,n}\\
0 & 0 & 0 & \dots & 0 & \lambda_n - \lambda
\end{pmatrix}$$

is $(\lambda_1 - \lambda)\cdots(\lambda_n - \lambda)$, and thus the eigenvalues of $A = PTP^{-1}$ are the diagonal entries of T. We use this property to prove the following proposition.

Proposition 1.4. *Given any complex $n \times n$ matrix A, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A. Furthermore, if u is an eigenvector of A for λ_i, then u is an eigenvector of e^A for e^{λ_i}.*

Proof. By Proposition 1.3 there is an invertible matrix P and an upper triangular matrix T such that

$$A = PTP^{-1}.$$

By Proposition 1.2,

$$e^{PTP^{-1}} = Pe^T P^{-1}.$$

Note that $e^T = \sum_{p\geq 0} \frac{T^p}{p!}$ is upper triangular since T^p is upper triangular for all $p \geq 0$. If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the diagonal entries of T, the properties of matrix multiplication, when combined with an induction on p, imply that the diagonal entries of T^p are $\lambda_1^p, \lambda_2^p, \ldots, \lambda_n^p$. This in turn implies that the diagonal entries of e^T are $\sum_{p\geq 0} \frac{\lambda_i^p}{p!} = e^{\lambda_i}$ for $1 \leq i \leq n$. In the preceding paragraph we showed that A and T have the same eigenvalues, which are the diagonal entries $\lambda_1, \ldots, \lambda_n$ of T. Since $e^A = e^{PTP^{-1}} = Pe^T P^{-1}$, and e^T is upper triangular, we use the same argument to conclude that both e^A and e^T have the same eigenvalues, which are the diagonal entries of e^T, where the diagonal entries of e^T are of the form $e^{\lambda_1}, \ldots, e^{\lambda_n}$. Now, if u is an eigenvector of A for the eigenvalue λ, a simple induction shows that u is an eigenvector of A^n for the eigenvalue λ^n, from which is follows that

$$\begin{aligned}
e^A u &= \left[I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \ldots\right] u = u + Au + \frac{A^2}{2!}u + \frac{A^3}{3!}u + \ldots\\
&= u + \lambda u + \frac{\lambda^2}{2!}u + \frac{\lambda^3}{3!}u + \cdots = \left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] u = e^\lambda u,
\end{aligned}$$

which shows that u is an eigenvector of e^A for e^λ. □

As a consequence, we can show that

$$\det(e^A) = e^{\mathrm{tr}(A)},$$

where $\mathrm{tr}(A)$ is the *trace of* A, i.e., the sum $a_{1\,1} + \cdots + a_{n\,n}$ of its diagonal entries, which is also equal to the sum of the eigenvalues of A. This is because the determinant of a matrix is equal to the product of its eigenvalues, and if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then by Proposition 1.4, $e^{\lambda_1}, \ldots, e^{\lambda_n}$ are the eigenvalues of e^A, and thus

$$\det\left(e^A\right) = e^{\lambda_1} \cdots e^{\lambda_n} = e^{\lambda_1 + \cdots + \lambda_n} = e^{\mathrm{tr}(A)}.$$

This shows that e^A is always an invertible matrix, since e^z is never null for every $z \in \mathbb{C}$. In fact, the inverse of e^A is e^{-A}, but we need to prove another proposition. This is because it is generally not true that

$$e^{A+B} = e^A e^B,$$

unless A and B commute, i.e., $AB = BA$. We need to prove this last fact.

Proposition 1.5. *Given any two complex $n \times n$ matrices A, B, if $AB = BA$, then*

$$e^{A+B} = e^A e^B.$$

Proof. Since $AB = BA$, we can expand $(A+B)^p$ using the binomial formula:

$$(A+B)^p = \sum_{k=0}^{p} \binom{p}{k} A^k B^{p-k},$$

and thus

$$\frac{1}{p!}(A+B)^p = \sum_{k=0}^{p} \frac{A^k B^{p-k}}{k!(p-k)!}.$$

Note that for any integer $N \geq 0$, we can write

$$\begin{aligned}
\sum_{p=0}^{2N} \frac{1}{p!}(A+B)^p &= \sum_{p=0}^{2N} \sum_{k=0}^{p} \frac{A^k B^{p-k}}{k!(p-k)!} \\
&= \left(\sum_{p=0}^{N} \frac{A^p}{p!}\right)\left(\sum_{p=0}^{N} \frac{B^p}{p!}\right) + \sum_{\substack{\max(k,l) > N \\ k+l \leq 2N}} \frac{A^k}{k!} \frac{B^l}{l!},
\end{aligned}$$

where there are $N(N+1)$ pairs (k,l) in the second term. Letting

$$\|A\| = \max\{|a_{i\,j}| \mid 1 \leq i, j \leq n\}, \quad \|B\| = \max\{|b_{i\,j}| \mid 1 \leq i, j \leq n\},$$

and $\mu = \max(\|A\|, \|B\|)$, note that for every entry $c_{i\,j}$ in $\left(A^k/k!\right)\left(B^l/l!\right)$, the first inequality of Proposition 1.1, along with the fact that $N < \max(k, l)$ and $k + l \le 2N$, implies that

$$|c_{i\,j}| \le n\frac{(n\mu)^k}{k!}\frac{(n\mu)^l}{l!} \le \frac{n(n\mu)^{k+l}}{k!l!} \le \frac{n^{k+l}(n\mu)^{k+l}}{k!l!} \le \frac{(n^2\mu)^{k+l}}{k!l!} \le \frac{(n^2\mu)^{2N}}{N!}.$$

As a consequence, the absolute value of every entry in

$$\sum_{\substack{\max(k,l) > N \\ k+l \le 2N}} \frac{A^k}{k!}\frac{B^l}{l!}$$

is bounded by

$$N(N+1)\frac{(n^2\mu)^{2N}}{N!},$$

which goes to 0 as $N \mapsto \infty$. To see why this is the case, note that

$$\begin{aligned} \lim_{N\to\infty} N(N+1)\frac{(n^2\mu)^{2N}}{N!} &= \lim_{N\to\infty} \frac{N(N+1)}{N(N-1)}\frac{(n^2\mu)^{2N}}{(N-2)!} = \lim_{N\to\infty} \frac{(n^4\mu^2)^{N-2+2}}{(N-2)!} \\ &= (n^4\mu^2)^2 \lim_{N\to\infty} \frac{(n^4\mu^2)^{N-2}}{(N-2)!} = 0, \end{aligned}$$

where the last equality follows from the well known identity $\lim_{N\to\infty} \frac{x^N}{N!} = 0$. From this it immediately follows that

$$e^{A+B} = e^A e^B.$$

□

Now, using Proposition 1.5, since A and $-A$ commute, we have

$$e^A e^{-A} = e^{A+-A} = e^{0_n} = I_n,$$

which shows that the inverse of e^A is e^{-A}.

We will now use the properties of the exponential that we have just established to show how various matrices can be represented as exponentials of other matrices.

1.2 The Lie Groups $\mathbf{GL}(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{R})$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, the Lie Algebras $\mathfrak{gl}(n, \mathbb{R})$, $\mathfrak{sl}(n, \mathbb{R})$, $\mathfrak{o}(n)$, $\mathfrak{so}(n)$, and the Exponential Map

First, we recall some basic facts and definitions. The set of real invertible $n \times n$ matrices forms a group under multiplication, denoted by $\mathbf{GL}(n, \mathbb{R})$. The subset of $\mathbf{GL}(n, \mathbb{R})$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{GL}(n, \mathbb{R})$, denoted by $\mathbf{SL}(n, \mathbb{R})$.

It is also easy to check that the set of real $n \times n$ orthogonal matrices forms a group under multiplication, denoted by $\mathbf{O}(n)$. The subset of $\mathbf{O}(n)$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{O}(n)$, denoted by $\mathbf{SO}(n)$. We will also call matrices in $\mathbf{SO}(n)$ *rotation matrices*. Staying with easy things, we can check that the set of real $n \times n$ matrices with null trace forms a vector space under addition, and similarly for the set of skew symmetric matrices.

Definition 1.1. The group $\mathbf{GL}(n, \mathbb{R})$ is called the *general linear group*, and its subgroup $\mathbf{SL}(n, \mathbb{R})$ is called the *special linear group*. The group $\mathbf{O}(n)$ of orthogonal matrices is called the *orthogonal group*, and its subgroup $\mathbf{SO}(n)$ is called the *special orthogonal group* (or *group of rotations*). The vector space of real $n \times n$ matrices with null trace is denoted by $\mathfrak{sl}(n, \mathbb{R})$, and the vector space of real $n \times n$ skew symmetric matrices is denoted by $\mathfrak{so}(n)$.

Remark: The notation $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ is rather strange and deserves some explanation. The groups $\mathbf{GL}(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{R})$, $\mathbf{O}(n)$, and $\mathbf{SO}(n)$ are more than just groups. They are also topological groups, which means that they are topological spaces (viewed as subspaces of $\mathbb{R}^{n^2}$) and that the multiplication and the inverse operations are continuous (in fact, smooth). Furthermore, they are smooth real manifolds.[1] Such objects are called *Lie groups*. The real vector spaces $\mathfrak{sl}(n)$ and $\mathfrak{so}(n)$ are what is called *Lie algebras*. However, we have not defined the algebra structure on $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{so}(n)$ yet. The algebra structure is given by what is called the *Lie bracket*, which is defined as

$$[A,\, B] = AB - BA.$$

Lie algebras are associated with Lie groups. What is going on is that the Lie algebra of a Lie group is its tangent space at the identity, i.e., the space of all tangent vectors at the identity (in this case, I_n). In some sense, the Lie algebra achieves a "linearization" of the Lie group. The exponential map is a map from the Lie algebra to the Lie group, for example,

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

and

$$\exp\colon \mathfrak{sl}(n, \mathbb{R}) \to \mathbf{SL}(n, \mathbb{R}).$$

The exponential map often allows a parametrization of the Lie group elements by simpler objects, the Lie algebra elements.

One might ask, What happened to the Lie algebras $\mathfrak{gl}(n, \mathbb{R})$ and $\mathfrak{o}(n)$ associated with the Lie groups $\mathbf{GL}(n, \mathbb{R})$ and $\mathbf{O}(n)$? We will see later that $\mathfrak{gl}(n, \mathbb{R})$ is the set of *all* real $n \times n$ matrices, and that $\mathfrak{o}(n) = \mathfrak{so}(n)$.

[1] We refrain from defining manifolds right now, not to interrupt the flow of intuitive ideas.

The properties of the exponential map play an important role in studying a Lie group. For example, it is clear that the map

$$\exp\colon \mathfrak{gl}(n,\mathbb{R}) \to \mathbf{GL}(n,\mathbb{R})$$

is well-defined, but since $\det(e^A) = e^{\mathrm{tr}(A)}$, every matrix of the form e^A has a positive determinant and exp is not surjective. Similarly, the fact $\det(e^A) = e^{\mathrm{tr}(A)}$ implies that the map

$$\exp\colon \mathfrak{sl}(n,\mathbb{R}) \to \mathbf{SL}(n,\mathbb{R})$$

is well-defined. However, we showed in Section 1.1 that it is not surjective either. As we will see in the next theorem, the map

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is well-defined and surjective. The map

$$\exp\colon \mathfrak{o}(n) \to \mathbf{O}(n)$$

is well-defined, but it is not surjective, since there are matrices in $\mathbf{O}(n)$ with determinant -1.

Remark: The situation for matrices over the field $\mathbb{C}$ of complex numbers is quite different, as we will see later.

We now show the fundamental relationship between $\mathbf{SO}(n)$ and $\mathfrak{so}(n)$.

Theorem 1.6. *The exponential map*

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is well-defined and surjective.

Proof. First we need to prove that if A is a skew symmetric matrix, then e^A is a rotation matrix. For this we quickly check that

$$\left(e^A\right)^\top = e^{A^\top}.$$

This is consequence of the definition $e^A = \sum_{p\geq 0} \frac{A^p}{p!}$ as a absolutely convergent series, the observation that $(A^p)^\top = (A^\top)^p$, and the linearity of the transpose map, i.e $(A+B)^\top = A^\top + B^\top$. Then since $A^\top = -A$, we get

$$\left(e^A\right)^\top = e^{A^\top} = e^{-A},$$

and so

$$\left(e^A\right)^\top e^A = e^{-A}e^A = e^{-A+A} = e^{0_n} = I_n,$$

and similarly,

$$e^A\left(e^A\right)^\top = I_n,$$

showing that e^A is orthogonal. Also,

$$\det\left(e^A\right) = e^{\mathrm{tr}(A)},$$

and since A is real skew symmetric, its diagonal entries are 0, i.e., $\mathrm{tr}(A) = 0$, and so $\det(e^A) = +1$.

For the surjectivity, we use Theorem 12.5, from Chapter 12 of Gallier [48]. Theorem 12.5 says that for every orthogonal matrix R there is an orthogonal matrix P such that $R = PE\,P^\top$, where E is a block diagonal matrix of the form

$$E = \begin{pmatrix} E_1 & & \dots & \\ & E_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & E_p \end{pmatrix},$$

such that each block E_i is either 1, -1, or a two-dimensional matrix of the form

$$E_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix},$$

with $0 < \theta_i < \pi$. Furthermore, if R is a rotation matrix, then we may assume that $0 < \theta_i \leq \pi$ and that the scalar entries are $+1$. Then we can form the block diagonal matrix

$$D = \begin{pmatrix} D_1 & & \dots & \\ & D_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & D_p \end{pmatrix}$$

such that each block D_i is either 0 when E_i consists of $+1$, or the two-dimensional matrix

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

when

$$E_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix},$$

and we let $A = PD\,P^\top$. It is clear that A is skew symmetric since $A^\top = \left(PDP^\top\right)^\top = PD^\top P^\top = -PDP^\top$. By Proposition 1.2,

$$e^A = e^{PDP^{-1}} = Pe^DP^{-1},$$

and since D is a block diagonal matrix, we can compute e^D by computing the exponentials of its blocks. If $D_i = 0$, we get $E_i = e^0 = +1$, and if

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix},$$

we showed earlier that

$$e^{D_i} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix},$$

exactly the block E_i. Thus, $E = e^D$, and as a consequence,

$$e^A = e^{PDP^{-1}} = Pe^D P^{-1} = PEP^{-1} = PE\,P^\top = R.$$

This shows the surjectivity of the exponential. □

When $n = 3$ (and A is skew symmetric), it is possible to work out an explicit formula for e^A. For any 3×3 real skew symmetric matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

letting $\theta = \sqrt{a^2 + b^2 + c^2}$ and

$$B = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix},$$

we have the following result known as *Rodrigues's formula* (1840).

Proposition 1.7. *The exponential map* $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ *is given by*

$$e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1-\cos\theta)}{\theta^2} B,$$

or, equivalently, by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1-\cos\theta)}{\theta^2} A^2$$

if $\theta \neq 0$*, with* $e^{0_3} = I_3$.

Proof sketch. First observe that

$$A^2 = -\theta^2 I_3 + B,$$

since

$$\begin{aligned}
A^2 &= \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} = \begin{pmatrix} -c^2-b^2 & ba & ca \\ ab & -c^2-a^2 & cb \\ ac & cb & -b^2-a^2 \end{pmatrix} \\
&= \begin{pmatrix} -a^2-b^2-c^2 & 0 & 0 \\ 0 & -a^2-b^2-c^2 & 0 \\ 0 & 0 & -a^2-b^2-c^2 \end{pmatrix} + \begin{pmatrix} a^2 & ba & ca \\ ab & b^2 & cb \\ ac & cb & c^2 \end{pmatrix} \\
&= -\theta^2 I_3 + B,
\end{aligned}$$

and that

$$AB = BA = 0.$$

From the above, deduce that

$$A^3 = -\theta^2 A,$$

and for any $k \geq 0$,

$$\begin{aligned}
A^{4k+1} &= \theta^{4k} A, \\
A^{4k+2} &= \theta^{4k} A^2, \\
A^{4k+3} &= -\theta^{4k+2} A, \\
A^{4k+4} &= -\theta^{4k+2} A^2.
\end{aligned}$$

Then prove the desired result by writing the power series for e^A and regrouping terms so that the power series for $\cos\theta$ and $\sin\theta$ show up. In particular

$$\begin{aligned}
e^A &= I_3 + \sum_{p\geq 1} \frac{A^p}{p!} = I_3 + \sum_{p\geq 0} \frac{A^{2p+1}}{(2p+1)!} + \sum_{p\geq 1} \frac{A^{2p}}{(2p)!} \\
&= I_3 + \sum_{p\geq 0} \frac{(-1)^p \theta^{2p}}{(2p+1)!} A + \sum_{p\geq 1} \frac{(-1)^{p-1}\theta^{2(p-1)}}{(2p)!} A^2 \\
&= I_3 + \frac{A}{\theta} \sum_{p\geq 0} \frac{(-1)^p \theta^{2p+1}}{(2p+1)!} - \frac{A^2}{\theta^2} \sum_{p\geq 1} \frac{(-1)^p \theta^{2p}}{(2p)!} \\
&= I_3 + \frac{\sin\theta}{\theta} A - \frac{A^2}{\theta^2} \sum_{p\geq 0} \frac{(-1)^p \theta^{2p}}{(2p)!} + \frac{A^2}{\theta^2} \\
&= I_3 + \frac{\sin\theta}{\theta} A + \frac{(1-\cos\theta)}{\theta^2} A^2.
\end{aligned}$$

□

The above formulae are the well-known formulae expressing a rotation of axis specified by the vector (a, b, c) and angle θ. Since the exponential is surjective, it is possible to write down an explicit formula for its inverse (but it is a multivalued function!). This has applications in kinematics, robotics, and motion interpolation.

1.3 Symmetric Matrices, Symmetric Positive Definite Matrices, and the Exponential Map

Recall that a real symmetric matrix is called *positive* (or *positive semidefinite*) if its eigenvalues are all positive or null, and *positive definite* if its eigenvalues are all strictly positive. We denote the vector space of real symmetric $n \times n$ matrices by $\mathbf{S}(n)$, the set of symmetric positive matrices by $\mathbf{SP}(n)$, and the set of symmetric positive definite matrices by $\mathbf{SPD}(n)$.

The next proposition shows that every symmetric positive definite matrix A is of the form e^B for some unique symmetric matrix B. The set of symmetric matrices is a vector space, but it is not a Lie algebra because the Lie bracket $[A, B]$ is not symmetric unless A and B commute, and the set of symmetric (positive) definite matrices is not a multiplicative group, so this result is of a different flavor as Theorem 1.6.

Proposition 1.8. *For every symmetric matrix B, the matrix e^B is symmetric positive definite. For every symmetric positive definite matrix A, there is a unique symmetric matrix B such that $A = e^B$.*

Proof. We showed earlier that

$$\left(e^B\right)^\top = e^{B^\top}.$$

If B is a symmetric matrix, then since $B^\top = B$, we get

$$\left(e^B\right)^\top = e^{B^\top} = e^B,$$

and e^B is also symmetric. Since the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the symmetric matrix B are real and the eigenvalues of e^B are $e^{\lambda_1}, \ldots, e^{\lambda_n}$, and since $e^\lambda > 0$ if $\lambda \in \mathbb{R}$, e^B is positive definite.

To show the surjectivity of the exponential map, note that if A is symmetric positive definite, then by Theorem 12.3 from Chapter 12 of Gallier [48], there is an orthogonal matrix P such that $A = PD\,P^\top$, where D is a diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & & \ldots & \\ & \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \lambda_n \end{pmatrix},$$

where $\lambda_i > 0$, since A is positive definite. Letting

$$L = \begin{pmatrix} \log \lambda_1 & & \ldots & \\ & \log \lambda_2 & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & \log \lambda_n \end{pmatrix},$$

by using the power series representation of e^L, it is obvious that $e^L = D$, with $\log \lambda_i \in \mathbb{R}$, since $\lambda_i > 0$.

Let

$$B = PL\, P^\top.$$

By Proposition 1.2, we have

$$e^B = e^{PL\, P^\top} = e^{PLP^{-1}} = Pe^L P^{-1} = Pe^L\, P^\top = PD\, P^\top = A.$$

Finally, we prove that if B_1 and B_2 are symmetric and $A = e^{B_1} = e^{B_2}$, then $B_1 = B_2$. We use an argument due to Chevalley [31] (see Chapter I, Proposition 5, pages 13-14). Since B_1 is symmetric, there is an orthonormal basis $(u_1, \ldots, u_n)$ of eigenvectors of B_1. Let $\mu_1, \ldots, \mu_n$ be the corresponding eigenvalues. Similarly, there is an orthonormal basis $(v_1, \ldots, v_n)$ of eigenvectors of B_2. We are going to prove that B_1 and B_2 agree on the basis $(v_1, \ldots, v_n)$, thus proving that $B_1 = B_2$.

Let μ be some eigenvalue of B_2, and let $v = v_i$ be some eigenvector of B_2 associated with μ. We can write

$$v = \alpha_1 u_1 + \cdots + \alpha_n u_n.$$

Since v is an eigenvector of B_2 for μ and $A = e^{B_2}$, by Proposition 1.4

$$A(v) = e^\mu v = e^\mu \alpha_1 u_1 + \cdots + e^\mu \alpha_n u_n.$$

On the other hand,

$$A(v) = A(\alpha_1 u_1 + \cdots + \alpha_n u_n) = \alpha_1 A(u_1) + \cdots + \alpha_n A(u_n),$$

and since $A = e^{B_1}$ and $B_1(u_i) = \mu_i u_i$, by Proposition 1.4 we get

$$A(v) = e^{\mu_1} \alpha_1 u_1 + \cdots + e^{\mu_n} \alpha_n u_n.$$

Therefore, $\alpha_i = 0$ if $\mu_i \neq \mu$. Letting

$$I = \{i \mid \mu_i = \mu,\ i \in \{1, \ldots, n\}\},$$

we have

$$v = \sum_{i \in I} \alpha_i u_i.$$

Now,

$$\begin{aligned} B_1(v) &= B_1\biggl(\sum_{i \in I} \alpha_i u_i\biggr) = \sum_{i \in I} \alpha_i B_1(u_i) = \sum_{i \in I} \alpha_i \mu_i u_i \\ &= \sum_{i \in I} \alpha_i \mu u_i = \mu\biggl(\sum_{i \in I} \alpha_i u_i\biggr) = \mu v, \end{aligned}$$

since $\mu_i = \mu$ when $i \in I$. Since v is an eigenvector of B_2 for μ,

$$B_2(v) = \mu v,$$

which shows that

$$B_1(v) = B_2(v).$$

Since the above holds for every eigenvector v_i, we have $B_1 = B_2$. □

Proposition 1.8 can be reformulated as stating that the map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a bijection. It can be shown that it is a homeomorphism. In the case of invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space $\mathbf{GL}(n, \mathbb{R})$ of real $n \times n$ invertible matrices (also a group) and $\mathbf{O}(n) \times \mathbf{SPD}(n)$.

As a corollary of the polar form theorem (Theorem 13.1 in Chapter 13 of Gallier [48]) and Proposition 1.8, we have the following result: For every invertible matrix A there is a unique orthogonal matrix R and a unique symmetric matrix S such that

$$A = R\, e^S.$$

Thus, we have a bijection between $\mathbf{GL}(n, \mathbb{R})$ and $\mathbf{O}(n) \times \mathbf{S}(n)$. But $\mathbf{S}(n)$ itself is isomorphic to $\mathbb{R}^{n(n+1)/2}$. Thus, there is a bijection between $\mathbf{GL}(n, \mathbb{R})$ and $\mathbf{O}(n) \times \mathbb{R}^{n(n+1)/2}$. It can also be shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this homeomorphism essentially reduces the study of the topology of $\mathbf{GL}(n, \mathbb{R})$ to the study of the topology of $\mathbf{O}(n)$. This is nice, since it can be shown that $\mathbf{O}(n)$ is compact.

In $A = R\, e^S$, if $\det(A) > 0$, then R must be a rotation matrix (i.e., $\det(R) = +1$), since $\det\left(e^S\right) > 0$. In particular, if $A \in \mathbf{SL}(n, \mathbb{R})$, since $\det(A) = \det(R) = +1$, the symmetric matrix S must have a null trace, i.e., $S \in \mathbf{S}(n) \cap \mathfrak{sl}(n, \mathbb{R})$. Thus, we have a bijection between $\mathbf{SL}(n, \mathbb{R})$ and $\mathbf{SO}(n) \times (\mathbf{S}(n) \cap \mathfrak{sl}(n, \mathbb{R}))$.

We can also show that the exponential map is a surjective map from the skew Hermitian matrices to the unitary matrices (use Theorem 12.7 from Chapter 12 in Gallier [48]).

1.4 The Lie Groups $\mathbf{GL}(n, \mathbb{C})$, $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{U}(n)$, $\mathbf{SU}(n)$, the Lie Algebras $\mathfrak{gl}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, $\mathfrak{su}(n)$, and the Exponential Map

The set of complex invertible $n \times n$ matrices forms a group under multiplication, denoted by $\mathbf{GL}(n, \mathbb{C})$. The subset of $\mathbf{GL}(n, \mathbb{C})$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{GL}(n, \mathbb{C})$, denoted by $\mathbf{SL}(n, \mathbb{C})$. It is also easy to check that the set of complex $n \times n$ unitary matrices forms a group under multiplication, denoted by $\mathbf{U}(n)$. The subset of $\mathbf{U}(n)$ consisting of those matrices having determinant $+1$ is a subgroup of $\mathbf{U}(n)$, denoted by $\mathbf{SU}(n)$. We can also check that the set of complex $n \times n$ matrices with null trace forms a real vector space under addition, and similarly for the set of skew Hermitian matrices and the set of skew Hermitian matrices with null trace.

Definition 1.2. The group $\mathbf{GL}(n, \mathbb{C})$ is called the *general linear group*, and its subgroup $\mathbf{SL}(n, \mathbb{C})$ is called the *special linear group*. The group $\mathbf{U}(n)$ of unitary matrices is called the *unitary group*, and its subgroup $\mathbf{SU}(n)$ is called the *special unitary group*. The real vector space of complex $n \times n$ matrices with null trace is denoted by $\mathfrak{sl}(n, \mathbb{C})$, the real vector space of skew Hermitian matrices is denoted by $\mathfrak{u}(n)$, and the real vector space $\mathfrak{u}(n) \cap \mathfrak{sl}(n, \mathbb{C})$ is denoted by $\mathfrak{su}(n)$.

Remarks:

(1) As in the real case, the groups $\mathbf{GL}(n, \mathbb{C})$, $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$ are also topological groups (viewed as subspaces of $\mathbb{R}^{2n^2}$), and in fact, smooth real manifolds. Such objects are called *(real) Lie groups*. The real vector spaces $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ are *Lie algebras* associated with $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$. The algebra structure is given by the *Lie bracket*, which is defined as

$$[A,\, B] = AB - BA.$$

(2) It is also possible to define complex Lie groups, which means that they are topological groups and smooth *complex* manifolds. It turns out that $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{SL}(n, \mathbb{C})$ are complex manifolds, but not $\mathbf{U}(n)$ and $\mathbf{SU}(n)$.

One should be very careful to observe that even though the Lie algebras $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$ consist of matrices with complex coefficients, we view them as *real* vector spaces. The Lie algebra $\mathfrak{sl}(n, \mathbb{C})$ is also a complex vector space, but $\mathfrak{u}(n)$ and $\mathfrak{su}(n)$ are not! Indeed, if A is a skew Hermitian matrix, iA is *not* skew Hermitian, but Hermitian!

Again the Lie algebra achieves a "linearization" of the Lie group. In the complex case, the Lie algebras $\mathfrak{gl}(n, \mathbb{C})$ is the set of *all* complex $n \times n$ matrices, but $\mathfrak{u}(n) \neq \mathfrak{su}(n)$, because a skew Hermitian matrix does not necessarily have a null trace.

The properties of the exponential map also play an important role in studying complex Lie groups. For example, it is clear that the map

$$\exp\colon \mathfrak{gl}(n, \mathbb{C}) \to \mathbf{GL}(n, \mathbb{C})$$

is well-defined, but this time, it is surjective! One way to prove this is to use the Jordan normal form. Similarly, since

$$\det\left(e^{A}\right) = e^{\mathrm{tr}(A)},$$

the map

$$\exp\colon \mathfrak{sl}(n, \mathbb{C}) \to \mathbf{SL}(n, \mathbb{C})$$

is well-defined, but it is not surjective! As we will see in the next theorem, the maps

$$\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n)$$

and

$$\exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$$

are well-defined and surjective.

Theorem 1.9. *The exponential maps*

$$\exp\colon \mathfrak{u}(n) \to \mathbf{U}(n) \quad \textit{and} \quad \exp\colon \mathfrak{su}(n) \to \mathbf{SU}(n)$$

are well-defined and surjective.

Proof. First we need to prove that if A is a skew Hermitian matrix, then e^A is a unitary matrix. Recall that $A^* = \overline{A}^\top$. Then since $(e^A)^\top = e^{A^\top}$, we readily deduce that

$$\left(e^A\right)^* = e^{A^*}.$$

Then since $A^* = -A$, we get

$$\left(e^A\right)^* = e^{A^*} = e^{-A},$$

and so

$$\left(e^A\right)^* e^A = e^{-A} e^A = e^{-A+A} = e^{0_n} = I_n,$$

and similarly, $e^A \left(e^A\right)^* = I_n$, showing that e^A is unitary. Since

$$\det\left(e^A\right) = e^{\mathrm{tr}(A)},$$

if A is skew Hermitian and has null trace, then $\det(e^A) = +1$.

For the surjectivity we will use Theorem 12.7 in Chapter 12 of Gallier [48]. First assume that A is a unitary matrix. By Theorem 12.7, there is a unitary matrix U and a diagonal matrix D such that $A = UDU^*$. Furthermore, since A is unitary, the entries $\lambda_1, \ldots, \lambda_n$ in D (the eigenvalues of A) have absolute value $+1$. Thus, the entries in D are of the form $\cos\theta + i\sin\theta = e^{i\theta}$. Thus, we can assume that D is a diagonal matrix of the form

$$D = \begin{pmatrix} e^{i\theta_1} & & \ldots & \\ & e^{i\theta_2} & \ldots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \ldots & e^{i\theta_p} \end{pmatrix}.$$

If we let E be the diagonal matrix

$$E = \begin{pmatrix} i\theta_1 & & \dots & \\ & i\theta_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & i\theta_p \end{pmatrix}$$

it is obvious that E is skew Hermitian and that

$$e^E = D.$$

Then letting $B = UEU^*$, we have

$$e^B = A,$$

and it is immediately verified that B is skew Hermitian, since E is.

If A is a unitary matrix with determinant $+1$, since the eigenvalues of A are $e^{i\theta_1}, \dots, e^{i\theta_p}$ and the determinant of A is the product

$$e^{i\theta_1} \cdots e^{i\theta_p} = e^{i(\theta_1 + \cdots + \theta_p)}$$

of these eigenvalues, we must have

$$\theta_1 + \cdots + \theta_p = 0,$$

and so, E is skew Hermitian and has zero trace. As above, letting

$$B = UEU^*,$$

we have

$$e^B = A,$$

where B is skew Hermitian and has null trace. □

We now extend the result of Section 1.3 to Hermitian matrices.

1.5 Hermitian Matrices, Hermitian Positive Definite Matrices, and the Exponential Map

Recall that a Hermitian matrix is called *positive* (or *positive semidefinite*) if its eigenvalues are all positive or null, and *positive definite* if its eigenvalues are all strictly positive. We denote the real vector space of Hermitian $n \times n$ matrices by $\mathbf{H}(n)$, the set of Hermitian positive matrices by $\mathbf{HP}(n)$, and the set of Hermitian positive definite matrices by $\mathbf{HPD}(n)$.

The next proposition shows that every Hermitian positive definite matrix A is of the form e^B for some unique Hermitian matrix B. As in the real case, the set of Hermitian matrices is a real vector space, but it is not a Lie algebra because the Lie bracket $[A, B]$ is not Hermitian unless A and B commute, and the set of Hermitian (positive) definite matrices is not a multiplicative group.

Proposition 1.10. *For every Hermitian matrix B, the matrix e^B is Hermitian positive definite. For every Hermitian positive definite matrix A, there is a unique Hermitian matrix B such that $A = e^B$.*

Proof. It is basically the same as the proof of Theorem 1.8, except that a Hermitian matrix can be written as $A = UDU^*$, where D is a real diagonal matrix and U is unitary instead of orthogonal. □

Proposition 1.10 can be reformulated as stating that the map $\exp\colon \mathbf{H}(n) \to \mathbf{HPD}(n)$ is a bijection. In fact, it can be shown that it is a homeomorphism. In the case of complex invertible matrices, the polar form theorem can be reformulated as stating that there is a bijection between the topological space $\mathbf{GL}(n, \mathbb{C})$ of complex $n \times n$ invertible matrices (also a group) and $\mathbf{U}(n) \times \mathbf{HPD}(n)$. As a corollary of the polar form theorem and Proposition 1.10, we have the following result: For every complex invertible matrix A, there is a unique unitary matrix U and a unique Hermitian matrix S such that

$$A = U\, e^S.$$

Thus, we have a bijection between $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbf{H}(n)$. But $\mathbf{H}(n)$ itself is isomorphic to $\mathbb{R}^{n^2}$, and so there is a bijection between $\mathbf{GL}(n, \mathbb{C})$ and $\mathbf{U}(n) \times \mathbb{R}^{n^2}$. It can also be shown that this bijection is a homeomorphism. This is an interesting fact. Indeed, this homeomorphism essentially reduces the study of the topology of $\mathbf{GL}(n, \mathbb{C})$ to the study of the topology of $\mathbf{U}(n)$. This is nice, since it can be shown that $\mathbf{U}(n)$ is compact (as a real manifold).

In the polar decomposition $A = Ue^S$, we have $|\det(U)| = 1$, since U is unitary, and $\mathrm{tr}(S)$ is real, since S is Hermitian (since it is the sum of the eigenvalues of S, which are real), so that $\det\left(e^S\right) > 0$. Thus, if $\det(A) = 1$, we must have $\det\left(e^S\right) = 1$, which implies that $S \in \mathbf{H}(n) \cap \mathfrak{sl}(n, \mathbb{C})$. Thus, we have a bijection between $\mathbf{SL}(n, \mathbb{C})$ and $\mathbf{SU}(n) \times (\mathbf{H}(n) \cap \mathfrak{sl}(n, \mathbb{C}))$.

In the next section we study the group $\mathbf{SE}(n)$ of affine maps induced by orthogonal transformations, also called rigid motions, and its Lie algebra. We will show that the exponential map is surjective. The groups $\mathbf{SE}(2)$ and $\mathbf{SE}(3)$ play play a fundamental role in robotics, dynamics, and motion planning.

1.6 The Lie Group $\mathbf{SE}(n)$ and the Lie Algebra $\mathfrak{se}(n)$

First, we review the usual way of representing affine maps of $\mathbb{R}^n$ in terms of $(n+1) \times (n+1)$ matrices.

Definition 1.3. The set of affine maps ρ of $\mathbb{R}^n$, defined such that

$$\rho(X) = RX + U,$$

where R is a rotation matrix ($R \in \mathbf{SO}(n)$) and U is some vector in $\mathbb{R}^n$, is a group under composition called the group of *direct affine isometries, or rigid motions*, denoted by $\mathbf{SE}(n)$.

Every rigid motion can be represented by the $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

in the sense that

$$\begin{pmatrix} \rho(X) \\ 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

iff

$$\rho(X) = RX + U.$$

Definition 1.4. The vector space of real $(n+1) \times (n+1)$ matrices of the form

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix},$$

where Ω is an $n \times n$ skew symmetric matrix and U is a vector in $\mathbb{R}^n$, is denoted by $\mathfrak{se}(n)$.

Remark: The group $\mathbf{SE}(n)$ is a Lie group, and its Lie algebra turns out to be $\mathfrak{se}(n)$.

We will show that the exponential map $\exp\colon \mathfrak{se}(n) \to \mathbf{SE}(n)$ is surjective. First we prove the following key proposition.

Proposition 1.11. *Given any $(n+1) \times (n+1)$ matrix of the form*

$$A = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}$$

where Ω is any $n \times n$ matrix and $U \in \mathbb{R}^n$,

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix},$$

where $\Omega^0 = I_n$. As a consequence,

$$e^A = \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix},$$

where

$$V = I_n + \sum_{k \geq 1} \frac{\Omega^k}{(k+1)!} = \sum_{k \geq 1} \frac{\Omega^{k-1}}{k!}.$$

Proof. A trivial induction on k shows that

$$A^k = \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}.$$

Then we have

$$\begin{aligned} e^A &= \sum_{k\geq 0} \frac{A^k}{k!}, \\ &= I_{n+1} + \sum_{k\geq 1} \frac{1}{k!} \begin{pmatrix} \Omega^k & \Omega^{k-1}U \\ 0 & 0 \end{pmatrix}, \\ &= \begin{pmatrix} I_n + \sum_{k\geq 1} \frac{\Omega^k}{k!} & \sum_{k\geq 1} \frac{\Omega^{k-1}}{k!} U \\ 0 & 1 \end{pmatrix}, \\ &= \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

□

We can now prove our main theorem. We will need to prove that V is invertible when Ω is a skew symmetric matrix. It would be tempting to write V as

$$V = \Omega^{-1}(e^\Omega - I).$$

Unfortunately, for odd n, a skew symmetric matrix of order n is not invertible! Thus, we have to find another way of proving that V is invertible. However, observe that we have the following useful fact:

$$V = I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!} = \int_0^1 e^{\Omega t} dt,$$

since $e^{\Omega t}$ is absolutely convergent and term by term integration yields

$$\begin{aligned} \int_0^1 e^{\Omega t} dt &= \int_0^1 \sum_{k\geq 0} \frac{(\Omega t)^k}{k!}\, dt = \sum_{k\geq 0} \frac{1}{k!} \int_0^1 (\Omega t)^k\, dt \\ &= \sum_{k\geq 0} \frac{\Omega^k}{k!} \int_0^1 t^k\, dt = \sum_{k\geq 0} \frac{\Omega^k}{k!} \left[\frac{t^{k+1}}{k+1}\right]_0^1 \\ &= \sum_{k\geq 1} \frac{\Omega^{k-1}}{k!} = I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!}. \end{aligned}$$

This is what we will use in Theorem 1.12 to prove surjectivity.

Theorem 1.12. *The exponential map*

$$\exp\colon \mathfrak{se}(n) \to \mathbf{SE}(n)$$

is well-defined and surjective.

Proof. Since Ω is skew symmetric, e^{Ω} is a rotation matrix, and by Theorem 1.6, the exponential map

$$\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is surjective. Thus it remains to prove that for every rotation matrix R, there is some skew symmetric matrix Ω such that $R = e^{\Omega}$ and

$$V = I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!}$$

is invertible. This is because Proposition 1.11 will then imply

$$e^{\begin{pmatrix} \Omega & V^{-1}U \\ 0 & 0 \end{pmatrix}} = \begin{pmatrix} e^{\Omega} & VV^{-1}U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}.$$

Theorem 12.5 from Chapter 12 of Gallier [48] says that for every orthogonal matrix R there is an orthogonal matrix P such that $R = PE\,P^{\top}$, where E is a block diagonal matrix of the form

$$E = \begin{pmatrix} E_1 & & \dots & \\ & E_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & E_p \end{pmatrix},$$

such that each block E_i is either 1, -1, or a two-dimensional matrix of the form

$$E_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}.$$

Furthermore, if R is a rotation matrix, then we may assume that $0 < \theta_i \leq \pi$ and that the scalar entries are $+1$. Then we can form the block diagonal matrix

$$D = \begin{pmatrix} D_1 & & \dots & \\ & D_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & D_p \end{pmatrix}$$

such that each block D_i is either 0 when E_i consists of $+1$, or the two-dimensional matrix

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

when

$$E_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix},$$

with $0 < \theta_i \leq \pi$. If we let $\Omega = PD\,P^\top$, then

$$e^\Omega = R,$$

as in the proof of Theorem 1.6. To compute V, since $\Omega = PD\,P^\top = PDP^{-1}$, observe that

$$\begin{aligned} V &= I_n + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!} \\ &= I_n + \sum_{k\geq 1} \frac{PD^kP^{-1}}{(k+1)!} \\ &= P\left(I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!}\right) P^{-1} \\ &= PWP^{-1}, \end{aligned}$$

where

$$W = I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!}.$$

We can compute

$$W = I_n + \sum_{k\geq 1} \frac{D^k}{(k+1)!} = \int_0^1 e^{Dt}dt,$$

by computing

$$W = \begin{pmatrix} W_1 & & \dots & \\ & W_2 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ & & \dots & W_p \end{pmatrix}$$

by blocks. Since

$$e^{D_it} = \begin{pmatrix} \cos(\theta_i t) & -\sin(\theta_i t) \\ \sin(\theta_i t) & \cos(\theta_i t) \end{pmatrix}$$

when D_i is a 2×2 skew symmetric matrix

$$D_i = \begin{pmatrix} 0 & -\theta_i \\ \theta_i & 0 \end{pmatrix}$$

and $W_i = \int_0^1 e^{D_it}dt$, we get

$$W_i = \begin{pmatrix} \int_0^1 \cos(\theta_i t)dt & \int_0^1 -\sin(\theta_i t)dt \\ \int_0^1 \sin(\theta_i t)dt & \int_0^1 \cos(\theta_i t)dt \end{pmatrix} = \frac{1}{\theta_i}\begin{pmatrix} \sin(\theta_i t)\mid_0^1 & \cos(\theta_i t)\mid_0^1 \\ -\cos(\theta_i t)\mid_0^1 & \sin(\theta_i t)\mid_0^1 \end{pmatrix},$$

that is,

$$W_i = \frac{1}{\theta_i}\begin{pmatrix} \sin\theta_i & -(1-\cos\theta_i) \\ 1-\cos\theta_i & \sin\theta_i \end{pmatrix},$$

and $W_i = 1$ when $D_i = 0$. Now, in the first case, the determinant is

$$\frac{1}{\theta_i^2}\left((\sin\theta_i)^2 + (1-\cos\theta_i)^2\right) = \frac{2}{\theta_i^2}(1-\cos\theta_i),$$

which is nonzero, since $0 < \theta_i \leq \pi$. Thus, each W_i is invertible, and so is W, and thus, $V = PWP^{-1}$ is invertible. □

In the case $n = 3$, given a skew symmetric matrix

$$\Omega = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

letting $\theta = \sqrt{a^2+b^2+c^2}$, it it easy to prove that if $\theta = 0$, then

$$e^A = \begin{pmatrix} I_3 & U \\ 0 & 1 \end{pmatrix},$$

and that if $\theta \neq 0$ (using the fact that $\Omega^3 = -\theta^2\Omega$), then by adjusting the calculation found at the end of Section 1.2

$$e^\Omega = I_3 + \frac{\sin\theta}{\theta}\Omega + \frac{(1-\cos\theta)}{\theta^2}\Omega^2 \qquad (*_1)$$

and

$$V = I_3 + \frac{(1-\cos\theta)}{\theta^2}\Omega + \frac{(\theta-\sin\theta)}{\theta^3}\Omega^2. \qquad (*_2)$$

1.7 Problems

Problem 1.1. (a) Find two symmetric matrices, A and B, such that AB is not symmetric.

(b) Find two matrices A and B such that

$$e^A e^B \neq e^{A+B}.$$

Hint. Try

$$A = \pi\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \pi\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},$$

and use the Rodrigues formula.

(c) Find some square matrices A, B such that $AB \neq BA$, yet

$$e^A e^B = e^{A+B}.$$

Hint. Look for 2×2 matrices with zero trace.

Problem 1.2. Given any matrix

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{C}),$$

if $\omega^2 = a^2 + bc$ and ω is any of the two complex roots of $a^2 + bc$, prove that if $\omega \neq 0$, then

$$e^B = \cosh \omega \, I + \frac{\sinh \omega}{\omega} B,$$

and $e^B = I + B$, if $a^2 + bc = 0$. Observe that $\mathrm{tr}(e^B) = 2 \cosh \omega$.

Prove that the exponential map, $\exp\colon \mathfrak{sl}(2, \mathbb{C}) \to \mathbf{SL}(2, \mathbb{C})$, is *not* surjective. For instance, prove that

$$\begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$$

is not the exponential of any matrix in $\mathfrak{sl}(2, \mathbb{C})$.

Problem 1.3. (Advanced) (a) Recall that a matrix N is *nilpotent* iff there is some $m \geq 0$ so that $N^m = 0$. Let A be any $n \times n$ matrix of the form $A = I - N$, where N is nilpotent. Why is A invertible? Prove that there is some B so that $e^B = I - N$ as follows: Recall that for any $y \in \mathbb{R}$ so that $|y - 1|$ is small enough, we have

$$\log(y) = -(1 - y) - \frac{(1 - y)^2}{2} - \cdots - \frac{(1 - y)^k}{k} - \cdots .$$

As N is nilpotent, we have $N^m = 0$, where m is the smallest integer with this propery. Then, the expression

$$B = \log(I - N) = -N - \frac{N^2}{2} - \cdots - \frac{N^{m-1}}{m - 1}$$

is well defined. Use a formal power series argument to show that

$$e^B = A.$$

We denote B by $\log(A)$.

(b) Let $A \in \mathbf{GL}(n, \mathbb{C})$. Prove that there is some matrix, B, so that $e^B = A$. Thus, the exponential map, $\exp\colon \mathfrak{gl}(n, \mathbb{C}) \to \mathbf{GL}(n, \mathbb{C})$, is surjective.

Hint. First, use the fact that A has a Jordan form, PJP^{-1}. Then, show that finding a log of A reduces to finding a log of every Jordan block of J. As every Jordan block, J, has a fixed nonzero constant, λ, on the diagonal, with 1's immediately above each diagonal entry and zero's everywhere else, we can write J as $(\lambda I)(I - N)$, where N is nilpotent. Find B_1 and B_2 so that $\lambda I = e^{B_1}$, $I - N = e^{B_2}$, and $B_1 B_2 = B_2 B_1$. Conclude that $J = e^{B_1 + B_2}$.

Problem 1.4. (a) Let $\mathfrak{so}(3)$ be the space of 3×3 skew symmetric matrices

$$\mathfrak{so}(3) = \left\{ \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \;\middle|\; a, b, c \in \mathbb{R} \right\}.$$

For any matrix

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix} \in \mathfrak{so}(3),$$

if we let $\theta = \sqrt{a^2 + b^2 + c^2}$ and

$$B = \begin{pmatrix} a^2 & ab & ac \\ ab & b^2 & bc \\ ac & bc & c^2 \end{pmatrix},$$

prove that

$$\begin{aligned} A^2 &= -\theta^2 I + B, \\ AB &= BA = 0. \end{aligned}$$

From the above, deduce that

$$A^3 = -\theta^2 A.$$

(b) Prove that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is given by

$$\exp A = e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

or, equivalently, by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2, \quad \text{if } \theta \neq 0,$$

with $\exp(0_3) = I_3$.

(c) Prove that e^A is an orthogonal matrix of determinant $+1$, i.e., a rotation matrix.

(d) Prove that the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective. For this, proceed as follows: Pick any rotation matrix $R \in \mathbf{SO}(3)$;

(1) The case $R = I$ is trivial.

(2) If $R \neq I$ and $\mathrm{tr}(R) \neq -1$, then

$$\exp^{-1}(R) = \left\{ \frac{\theta}{2\sin\theta}(R - R^T) \;\middle|\; 1 + 2\cos\theta = \mathrm{tr}(R) \right\}.$$

(Recall that $\mathrm{tr}(R) = r_{1\,1} + r_{2\,2} + r_{3\,3}$, the *trace* of the matrix R).

Show that there is a unique skew-symmetric B with corresponding θ satisfying $0 < \theta < \pi$ such that $e^B = R$.

(3) If $R \neq I$ and $\mathrm{tr}(R) = -1$, then prove that the eigenvalues of R are $1, -1, -1$, that $R = R^\top$, and that $R^2 = I$. Prove that the matrix

$$S = \frac{1}{2}(R - I)$$

is a symmetric matrix whose eigenvalues are $-1, -1, 0$. Thus, S can be diagonalized with respect to an orthogonal matrix Q as

$$S = Q \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} Q^\top.$$

Prove that there exists a skew symmetric matrix

$$U = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

so that

$$U^2 = S = \frac{1}{2}(R - I).$$

Observe that

$$U^2 = \begin{pmatrix} -(c^2 + d^2) & bc & bd \\ bc & -(b^2 + d^2) & cd \\ bd & cd & -(b^2 + c^2) \end{pmatrix},$$

and use this to conclude that if $U^2 = S$, then $b^2 + c^2 + d^2 = 1$. Then, show that

$$\exp^{-1}(R) = \left\{ (2k+1)\pi \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},\ k \in \mathbb{Z} \right\},$$

where (b, c, d) is any unit vector such that for the corresponding skew symmetric matrix U, we have $U^2 = S$.

(e) To find a skew symmetric matrix U so that $U^2 = S = \frac{1}{2}(R - I)$ as in (d), we can solve the system

$$\begin{pmatrix} b^2 - 1 & bc & bd \\ bc & c^2 - 1 & cd \\ bd & cd & d^2 - 1 \end{pmatrix} = S.$$

We immediately get b^2, c^2, d^2, and then, since one of b, c, d is nonzero, say b, if we choose the positive square root of b^2, we can determine c and d from bc and bd.

Implement a computer program to solve the above system.

(f) The previous questions show that we can compute a log of a rotation matrix, although when $\theta \approx 0$, we have to be careful in computing $\frac{\sin\theta}{\theta}$; in this case, we may want to use

$$\frac{\sin\theta}{\theta} = 1 - \frac{\theta^2}{3!} + \frac{\theta^4}{5!} + \cdots .$$

Given two rotations, $R_1, R_2 \in \mathbf{SO}(3)$, there are three natural interpolation formulae:

$$e^{(1-t)\log R_1 + t\log R_2}; \quad R_1 e^{t\log(R_1^\top R_2)}; \quad e^{t\log(R_2 R_1^\top)} R_1,$$

with $0 \leq t \leq 1$.

Write a computer program to investigate the difference between these interpolation formulae.

The position of a rigid body spinning around its center of gravity is determined by a rotation matrix, $R \in \mathbf{SO}(3)$. If R_1 denotes the initial position and R_2 the final position of this rigid body, by computing interpolants of R_1 and R_2, we get a motion of the rigid body and we can create an animation of this motion by displaying several interpolants. The rigid body can be a "funny" object, for example a banana, a bottle, etc.

Problem 1.5. Consider the affine maps $\rho\colon \mathbb{R}^2 \to \mathbb{R}^2$ defined such that

$$\rho\begin{pmatrix} x \\ y \end{pmatrix} = \alpha \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix},$$

where $\theta, w_1, w_2, \alpha \in \mathbb{R}$, with $\alpha > 0$. These maps are called (direct) *affine similitudes* (for short, *similitudes*). The number $\alpha > 0$ is the *scale factor* of the similitude. These affine maps are the composition of a rotation of angle θ, a rescaling by $\alpha > 0$, and a translation.

(a) Prove that these maps form a group that we denote by $\mathbf{SIM}(2)$.

Given any map ρ as above, if we let

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad X = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \text{and} \quad W = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix},$$

then ρ can be represented by the 3×3 matrix

$$A = \begin{pmatrix} \alpha R & W \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \alpha\cos\theta & -\alpha\sin\theta & w_1 \\ \alpha\sin\theta & \alpha\cos\theta & w_2 \\ 0 & 0 & 1 \end{pmatrix}$$

in the sense that

$$\begin{pmatrix} \rho(X) \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha R & W \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

iff

$$\rho(X) = \alpha RX + W.$$

(b) Consider the set of matrices of the form

$$\begin{pmatrix} \lambda & -\theta & u \\ \theta & \lambda & v \\ 0 & 0 & 0 \end{pmatrix}$$

where $\theta, \lambda, u, v \in \mathbb{R}$. Verify that this set of matrices is a vector space isomorphic to $(\mathbb{R}^4, +)$. This vector space is denoted by $\mathfrak{sim}(2)$.

(c) Given a matrix

$$\Omega = \begin{pmatrix} \lambda & -\theta \\ \theta & \lambda \end{pmatrix},$$

prove that

$$e^{\Omega} = e^{\lambda} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Hint. Write

$$\Omega = \lambda I + \theta J,$$

with

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Observe that $J^2 = -I$, and prove by induction on k that

$$\Omega^k = \frac{1}{2}\left((\lambda + i\theta)^k + (\lambda - i\theta)^k\right) I + \frac{1}{2i}\left((\lambda + i\theta)^k - (\lambda - i\theta)^k\right) J.$$

(d) As in (c), write

$$\Omega = \begin{pmatrix} \lambda & -\theta \\ \theta & \lambda \end{pmatrix},$$

let

$$U = \begin{pmatrix} u \\ v \end{pmatrix},$$

and let

$$B = \begin{pmatrix} \Omega & U \\ 0 & 0 \end{pmatrix}.$$

Prove that

$$B^n = \begin{pmatrix} \Omega^n & \Omega^{n-1}U \\ 0 & 0 \end{pmatrix}$$

where $\Omega^0 = I_2$.

Prove that

$$e^B = \begin{pmatrix} e^{\Omega} & VU \\ 0 & 1 \end{pmatrix},$$

where

$$V = I_2 + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!}.$$

(e) Use the formula

$$V = I_2 + \sum_{k\geq 1} \frac{\Omega^k}{(k+1)!} = \int_0^1 e^{\Omega t} dt$$

to prove that if $\lambda = \theta = 0$, then

$$V = I_2,$$

else

$$V = \frac{1}{\lambda^2 + \theta^2} \begin{pmatrix} \lambda(e^\lambda \cos\theta - 1) + e^\lambda \theta \sin\theta & -\theta(1 - e^\lambda \cos\theta) - e^\lambda \lambda \sin\theta \\ \theta(1 - e^\lambda \cos\theta) + e^\lambda \lambda \sin\theta & \lambda(e^\lambda \cos\theta - 1) + e^\lambda \theta \sin\theta \end{pmatrix}.$$

Conclude that if $\lambda = \theta = 0$, then

$$e^B = \begin{pmatrix} I & U \\ 0 & 1 \end{pmatrix},$$

else

$$e^B = \begin{pmatrix} e^\Omega & VU \\ 0 & 1 \end{pmatrix},$$

with

$$e^\Omega = e^\lambda \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

and

$$V = \frac{1}{\lambda^2 + \theta^2} \begin{pmatrix} \lambda(e^\lambda \cos\theta - 1) + e^\lambda \theta \sin\theta & -\theta(1 - e^\lambda \cos\theta) - e^\lambda \lambda \sin\theta \\ \theta(1 - e^\lambda \cos\theta) + e^\lambda \lambda \sin\theta & \lambda(e^\lambda \cos\theta - 1) + e^\lambda \theta \sin\theta \end{pmatrix},$$

and that $e^B \in \mathbf{SIM}(2)$, with scale factor e^λ.

(f) Prove that the exponential map $\exp\colon \mathfrak{sim}(2) \to \mathbf{SIM}(2)$ is surjective.

(g) Similitudes can be used to describe certain deformations (or flows) of a deformable body $\mathcal{B}_t$ in the plane. Given some initial shape $\mathcal{B}$ in the plane (for example, a circle), a deformation of $\mathcal{B}$ is given by a piecewise differentiable curve

$$\mathcal{D}\colon [0, T] \to \mathbf{SIM}(2),$$

where each $\mathcal{D}(t)$ is a similitude (for some $T > 0$). The deformed body $\mathcal{B}_t$ at time t is given by

$$\mathcal{B}_t = \mathcal{D}(t)(\mathcal{B}).$$

The surjectivity of the exponential map $\exp\colon \mathfrak{sim}(2) \to \mathbf{SIM}(2)$ implies that there is a map $\log\colon \mathbf{SIM}(2) \to \mathfrak{sim}(2)$, although it is multivalued. The exponential map and the log “function” allows us to work in the simpler (noncurved) Euclidean space $\mathfrak{sim}(2)$.

For instance, given two similitudes $A_1, A_2 \in \mathbf{SIM}(2)$ specifying the shape of $\mathcal{B}$ at two different times, we can compute $\log(A_1)$ and $\log(A_2)$, which are just elements of the Euclidean space $\mathfrak{sim}(2)$, form the linear interpolant $(1-t)\log(A_1) + t\log(A_2)$, and then apply the exponential map to get an interpolating deformation

$$t \mapsto e^{(1-t)\log(A_1)+t\log(A_2)}, \quad t \in [0,1].$$

Also, given a sequence of "snapshots" of the deformable body $\mathcal{B}$, say $A_0, A_1, \ldots, A_m$, where each is A_i is a similitude, we can try to find an interpolating deformation (a curve in $\mathbf{SIM}(2)$) by finding a simpler curve $t \mapsto C(t)$ in $\mathfrak{sim}(2)$ (say, a B-spline) interpolating $\log A_1, \log A_1, \ldots, \log A_m$. Then, the curve $t \mapsto e^{C(t)}$ yields a deformation in $\mathbf{SIM}(2)$ interpolating $A_0, A_1, \ldots, A_m$.

(1) Write a program interpolating between two deformations.

(2) If you know about cubic spline interpolation, write a program to interpolate a sequence of deformations given by similitudes $A_0, A_1, \ldots, A_m$ by a C^2-curve.

Problem 1.6. Derive Equations $(*_1)$ and $(*_2)$ of Section 1.6.

Chapter 2

Adjoint Representations and the Derivative of exp

In this chapter, in preparation for defining the Lie bracket on the Lie algebra of a Lie group, we introduce the adjoint representations of the group $\mathbf{GL}(n, \mathbb{R})$ and of the Lie algebra $\mathfrak{gl}(n, \mathbb{R})$. The map $\mathrm{Ad}\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(\mathfrak{gl}(n, \mathbb{R}))$ is defined such that Ad_A is the derivative of the conjugation map $\mathbf{Ad}_A\colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(n, \mathbb{R})$ at the identity. The map ad is the derivative of Ad at the identity, and it turns out that $\mathrm{ad}_A(B) = [A, B]$, the Lie bracket of A and B, and in this case, $[A, B] = AB - BA$. We also find a formula for the derivative of the matrix exponential **exp**. This formula has an interesting application to the problem of finding a natural sets of real matrices over which the exponential is injective, which is used in numerical linear algebra.

2.1 The Adjoint Representations Ad and ad

Given any two vector spaces E and F, recall that the vector space of all linear maps from E to F is denoted by $\mathrm{Hom}(E, F)$. The set of all invertible linear maps from E to itself is a group (under composition) denoted $\mathbf{GL}(E)$. When $E = \mathbb{R}^n$, we often denote $\mathbf{GL}(\mathbb{R}^n)$ by $\mathbf{GL}(n, \mathbb{R})$ (and if $E = \mathbb{C}^n$, we often denote $\mathbf{GL}(\mathbb{C}^n)$ by $\mathbf{GL}(n, \mathbb{C})$). The vector space $\mathrm{M}_n(\mathbb{R})$ of all $n \times n$ matrices is also denoted by $\mathfrak{gl}(n, \mathbb{R})$ (and $\mathrm{M}_n(\mathbb{C})$ by $\mathfrak{gl}(n, \mathbb{C})$). Then $\mathbf{GL}(\mathfrak{gl}(n, \mathbb{R}))$ is the group of all invertible linear maps from $\mathfrak{gl}(n, \mathbb{R}) = \mathrm{M}_n(\mathbb{R})$ to itself.

For any matrix $A \in \mathrm{M}_n(\mathbb{R})$ (or $A \in \mathrm{M}_n(\mathbb{C})$), define the maps $L_A\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ and $R_A\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ by

$$L_A(B) = AB, \quad R_A(B) = BA, \quad \text{for all } B \in \mathrm{M}_n(\mathbb{R}).$$

Observe that $L_A \circ R_B = R_B \circ L_A$ for all $A, B \in \mathrm{M}_n(\mathbb{R})$.

For any matrix $A \in \mathbf{GL}(n, \mathbb{R})$, let

$$\mathbf{Ad}_A\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R}) \quad \text{(conjugation by } A)$$

be given by

$$\mathbf{Ad}_A(B) = ABA^{-1} \quad \text{for all } B \in \mathrm{M}_n(\mathbb{R}).$$

Observe that $\mathbf{Ad}_A = L_A \circ R_{A^{-1}}$ and that $\mathbf{Ad}_A$ is an invertible linear map with inverse $\mathbf{Ad}_{A^{-1}}$. The restriction of $\mathbf{Ad}_A$ to invertible matrices $B \in \mathbf{GL}(n, \mathbb{R})$ yields the map

$$\mathbf{Ad}_A \colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(n, \mathbb{R})$$

also given by

$$\mathbf{Ad}_A(B) = ABA^{-1} \quad \text{for all } B \in \mathbf{GL}(n, \mathbb{R}).$$

This time, observe that $\mathbf{Ad}_A$ is a group homomorphism of $\mathbf{GL}(n, \mathbb{R})$ (with respect to multiplication), since

$$\mathbf{Ad}_A(BC) = ABCA^{-1} = ABA^{-1}ACA^{-1} = \mathbf{Ad}_A(B)\mathbf{Ad}_A(C), \quad \text{for all } B, C \in \mathbf{GL}(n, \mathbb{R}).$$

In fact, $\mathbf{Ad}_A$ is a group isomorphism (since its inverse is $\mathbf{Ad}_{A^{-1}}$).

Beware that $\mathbf{Ad}_A$ is **not** a linear map on $\mathbf{GL}(n, \mathbb{R})$ because $\mathbf{GL}(n, \mathbb{R})$ is not a vector space! Indeed, $\mathbf{GL}(n, \mathbb{R})$ is not closed under addition.

Nevertheless, we can define the derivative of $\mathbf{Ad}_A \colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ with $A \in \mathbf{GL}(n, \mathbb{R})$ and $B, X \in \mathrm{M}_n(\mathbb{R})$ by

$$\mathbf{Ad}_A(B + X) - \mathbf{Ad}_A(B) = A(B + X)A^{-1} - ABA^{-1} = AXA^{-1},$$

which shows that $d(\mathbf{Ad}_A)_B$ exists and is given by

$$d(\mathbf{Ad}_A)_B(X) = AXA^{-1}, \quad \text{for all } X \in \mathrm{M}_n(\mathbb{R}).$$

In particular, for $B = I$, we see that the derivative $d(\mathbf{Ad}_A)_I$ of $\mathbf{Ad}_A$ at I is a linear map of $\mathfrak{gl}(n, \mathbb{R}) = \mathrm{M}_n(\mathbb{R})$ denoted by $\mathrm{Ad}(A)$ or Ad_A (or $\mathrm{Ad}\, A$), and given by

$$\mathrm{Ad}_A(X) = AXA^{-1} \quad \text{for all } X \in \mathfrak{gl}(n, \mathbb{R}).$$

The inverse of Ad_A is $\mathrm{Ad}_{A^{-1}}$, so $\mathrm{Ad}_A \in \mathbf{GL}(\mathfrak{gl}(n, \mathbb{R}))$. Note that

$$\mathrm{Ad}_{AB} = \mathrm{Ad}_A \circ \mathrm{Ad}_B,$$

so the map $A \mapsto \mathrm{Ad}_A$ is a group homomorphism of $\mathbf{GL}(\mathfrak{gl}(n, \mathbb{R}))$ denoted

$$\mathrm{Ad} \colon \mathbf{GL}(n, \mathbb{R}) \to \mathbf{GL}(\mathfrak{gl}(n, \mathbb{R})).$$

The homomorphism Ad is called the *adjoint representation* of $\mathbf{GL}(n, \mathbb{R})$.

We also would like to compute the derivative $d(\mathrm{Ad})_I$ of Ad at I. If it exists, it is a linear map

$$d(\mathrm{Ad})_I \colon \mathfrak{gl}(n, \mathbb{R}) \to \mathrm{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R})).$$

For all $X, Y \in \mathrm{M}_n(\mathbb{R})$, with $\|X\|$ small enough we have $I + X \in \mathbf{GL}(n, \mathbb{R})$, and

$$\begin{aligned}\mathrm{Ad}_{I+X}(Y) - \mathrm{Ad}_I(Y) - (XY - YX) &= (I+X)Y(I+X)^{-1} - Y - XY + YX \\ &= [(I+X)Y - Y(I+X) - XY(I+X) \\ &\quad + YX(I+X)](I+X)^{-1} \\ &= [Y + XY - Y - YX - XY - XYX \\ &\quad + YX + YX^2](I+X)^{-1} \\ &= (YX^2 - XYX)(I+X)^{-1}.\end{aligned}$$

If we let

$$\epsilon(X, Y) = \frac{(YX^2 - XYX)(I+X)^{-1}}{\|X\|},$$

since $\| \; \|$ is a matrix norm, we get

$$\begin{aligned}\|\epsilon(X,Y)\| &= \frac{\|YX^2 - XYX\| \, \|(I+X)^{-1}\|}{\|X\|} \leq \frac{(\|YX^2\| + \|XYX\|) \, \|(I+X)^{-1}\|}{\|X\|} \\ &\leq \frac{\big(\|Y\| \, \|X\|^2 + \|X\| \, \|Y\| \, \|X\|\big) \, \|(I+X)^{-1}\|}{\|X\|} = \frac{2 \, \|Y\| \, \|X\|^2 \, \|(I+X)^{-1}\|}{\|X\|} \\ &= 2 \, \|X\| \, \|Y\| \, \big\|(I+X)^{-1}\big\| .\end{aligned}$$

Therefore, we proved that for $\|X\|$ small enough

$$\mathrm{Ad}_{I+X}(Y) - \mathrm{Ad}_I(Y) = (XY - YX) + \epsilon(X, Y) \, \|X\| ,$$

with $\|\epsilon(X, Y)\| \leq 2 \, \|X\| \, \|Y\| \, \|(I + X)^{-1}\|$, and $\epsilon(X, Y)$ linear in Y.

Let $\mathrm{ad}_X \colon \mathfrak{gl}(n, \mathbb{R}) \to \mathfrak{gl}(n, \mathbb{R})$ be the linear map given by

$$\mathrm{ad}_X(Y) = XY - YX = [X, Y],$$

and ad be the linear map

$$\mathrm{ad} \colon \mathfrak{gl}(n, \mathbb{R}) \to \mathrm{Hom}(\mathfrak{gl}(n, \mathbb{R}), \mathfrak{gl}(n, \mathbb{R}))$$

given by

$$\mathrm{ad}(X) = \mathrm{ad}_X.$$

We also define $\epsilon_X \colon \mathfrak{gl}(n, \mathbb{R}) \to \mathfrak{gl}(n, \mathbb{R})$ as the linear map given by

$$\epsilon_X(Y) = \epsilon(X, Y).$$

If $\|\epsilon_X\|$ is the operator norm of ϵ_X, we have

$$\|\epsilon_X\| = \max_{\|Y\|=1} \|\epsilon(X, Y)\| \leq 2 \, \|X\| \, \big\|(I + X)^{-1}\big\| .$$

Then the equation

$$\mathrm{Ad}_{I+X}(Y) - \mathrm{Ad}_I(Y) = (XY - YX) + \epsilon(X,Y)\left\|X\right\|,$$

which holds for all Y, yields

$$\mathrm{Ad}_{I+X} - \mathrm{Ad}_I = \mathrm{ad}_X + \epsilon_X\left\|X\right\|,$$

and because $\left\|\epsilon_X\right\| \leq 2\left\|X\right\|\left\|(I+X)^{-1}\right\|$, we have $\lim_{X\mapsto 0}\epsilon_X = 0$, which shows that $d(\mathrm{Ad})_I(X) = \mathrm{ad}_X$; that is,

$$d(\mathrm{Ad})_I = \mathrm{ad}.$$

The notation $\mathrm{ad}(X)$ (or $\mathrm{ad}\,X$) is also used instead ad_X. The map ad is a linear map

$$\mathrm{ad}\colon \mathfrak{gl}(n,\mathbb{R}) \to \mathrm{Hom}(\mathfrak{gl}(n,\mathbb{R}), \mathfrak{gl}(n,\mathbb{R}))$$

called the *adjoint representation* of $\mathfrak{gl}(n,\mathbb{R})$. The Lie algebra $\mathrm{Hom}(\mathfrak{gl}(n,\mathbb{R}), \mathfrak{gl}(n,\mathbb{R}))$ of the group $\mathbf{GL}(\mathfrak{gl}(n,\mathbb{R}))$ is also denoted by $\mathfrak{gl}(\mathfrak{gl}(n,\mathbb{R}))$.

Since

$$\begin{aligned}\mathrm{ad}([X,Y])(Z) &= \mathrm{ad}(XY-YX)(Z) = (XY-YX)Z - Z(XY-YX)\\ &= XYZ - YXZ - ZXY + ZYX\\ &= XYZ - XZY - YZX + ZYX - (YXZ - YZX - XZY + ZXY)\\ &= X(YZ-ZY) - (YZ-ZY)X - (Y(XZ-ZX) - (XZ-ZX)Y)\\ &= \mathrm{ad}(X)(YZ-ZY) - \mathrm{ad}(Y)(XZ-ZX)\\ &= \mathrm{ad}(X)\mathrm{ad}(Y)(Z) - \mathrm{ad}(Y)\mathrm{ad}(X)(Z)\end{aligned}$$

whenever $X,Y,Z \in \mathfrak{gl}(n,\mathbb{R})$, we find that

$$\mathrm{ad}([X,Y]) = \mathrm{ad}(X)\mathrm{ad}(Y) - \mathrm{ad}(Y)\mathrm{ad}(X) = [\mathrm{ad}(X), \mathrm{ad}(Y)].$$

This means that ad is a Lie algebra homomorphism. It can be checked that this property is equivalent to the following identity known as the *Jacobi identity*:

$$[X,\,[Y,\,Z]] + [Z,\,[X,\,Y]] + [Y,\,[Z,\,X]] = 0,$$

for all $X,Y,Z \in \mathfrak{gl}(n,\mathbb{R})$. Note that

$$\mathrm{ad}_X = L_X - R_X.$$

Next we prove a formula relating Ad and ad through the exponential. For this, we view ad_X and Ad_A as $n^2 \times n^2$ matrices, for example, over the basis (E_{ij}) of $n \times n$ matrices whose entries are all 0 except for the entry of index (i,j) which is equal to 1.

Proposition 2.1. *For any $X \in \mathrm{M}_n(\mathbb{R}) = \mathfrak{gl}(n, \mathbb{R})$, we have*

$$\mathrm{Ad}_{e^X} = e^{\mathrm{ad}_X} = \sum_{k=0}^{\infty} \frac{(\mathrm{ad}_X)^k}{k!};$$

that is,

$$e^X Y e^{-X} = e^{\mathrm{ad}_X} Y = Y + [X, Y] + \frac{1}{2!}[X, [X, Y]] + \frac{1}{3!}[X, [X, [X, Y]]] + \cdots$$

for all $X, Y \in \mathrm{M}_n(\mathbb{R})$

Proof. Let

$$A(t) = \mathrm{Ad}_{e^{tX}},$$

pick any $Y \in \mathrm{M}_n(\mathbb{R})$, and compute the derivative of $A(t)Y$. By the product rule we have

$$\begin{aligned}(A(t)Y)'(t) &= (e^{tX} Y e^{-tX})'(t) \\ &= X e^{tX} Y e^{-tX} + e^{tX} Y e^{-tX}(-X) \\ &= X e^{tX} Y e^{-tX} - e^{tX} Y e^{-tX} X \\ &= \mathrm{ad}_X(\mathrm{Ad}_{e^{tX}} Y) = \mathrm{ad}_X(A(t)Y).\end{aligned}$$

We also have $A(0)Y = \mathrm{Ad}_I Y = Y$. Therefore, the curve $t \mapsto A(t)Y$ is an integral curve for the vector field X_{ad_X} with initial condition Y, and by Proposition 11.25 (with n replaced by n^2), this unique integral curve is given by

$$\gamma(t) = e^{t\mathrm{ad}_X} Y,$$

which proves our assertion. □

2.2 The Derivative of exp

It is also possible to find a formula for the derivative $d\exp_A$ of the exponential map at A, but this is a bit tricky. It can be shown that

$$d(\exp)_A = e^A \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} (\mathrm{ad}_A)^k = e^{L_A} \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (L_A - R_A)^j,$$

so

$$d(\exp)_A(B) = e^A \left(B - \frac{1}{2!}[A, B] + \frac{1}{3!}[A, [A, B]] - \frac{1}{4!}[A, [A, [A, B]]] + \cdots \right).$$

It is customary to write

$$\frac{\mathrm{id} - e^{-\mathrm{ad}_A}}{\mathrm{ad}_A}$$

for the power series

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} (\mathrm{ad}_A)^k,$$

and the formula for the derivative of exp is usually stated as

$$d(\exp)_A = e^A \left(\frac{\mathrm{id} - e^{-\mathrm{ad}_A}}{\mathrm{ad}_A} \right).$$

Most proofs I am aware of use some tricks involving ODE's, but there is a simple and direct way to prove the formula based on the fact that $\mathrm{ad}_A = L_A - R_A$ and that L_A and R_A commute. First, one can show that

$$d(\exp)_A = \sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!}.$$

Thus, we need to prove that

$$e^{L_A} \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (L_A - R_A)^j = \sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!}.$$

To simplify notation, write a for L_A and b for L_B. We wish to prove that

$$e^a \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!} (a-b)^j = \sum_{h,k \geq 0} \frac{a^h b^k}{(h+k+1)!}, \tag{$*$}$$

assuming that $ab = ba$. This can be done by finding the coefficient of the monomial $a^h b^k$ on the left hand side. We find that this coefficient is

$$\frac{1}{(h+k+1)!} \sum_{i=0}^{h} (-1)^{h-i} \binom{h+k+1}{i} \binom{h+k-i}{k}.$$

Therefore, to prove $(*)$, we need to prove that

$$\sum_{i=0}^{h} (-1)^{h-i} \binom{h+k+1}{i} \binom{h+k-i}{k} = 1.$$

The above identity can be shown in various ways. A brute force method is to use induction. One can also use "negation of the upper index" and a Vandermonde convolution to obtain a two line proof. The details are left as an exercise.

The formula for the exponential tells us when the derivative $d(\exp)_A$ is invertible. Indeed, if the eigenvalues of the matrix X are $\lambda_1, \ldots, \lambda_n$, then the eigenvalues of the matrix

$$\frac{\mathrm{id} - e^{-X}}{X} = \sum_{k=0}^{\infty} \frac{(-1)^k}{(k+1)!} X^k$$

are

$$\frac{1-e^{-\lambda_j}}{\lambda_j} \quad \text{if } \lambda_j \neq 0, \text{ and } 1 \text{ if } \lambda_j = 0.$$

To see why this is the case, assume $\lambda \neq 0$ is an eigenvalue of X with eigenvector u, i.e. $Xu = \lambda u$. Then $(-X)^k u = -\lambda^k u$ for any nonnegative integer k and

$$\begin{aligned}
\frac{\mathrm{id} - e^{-X}}{X} u &= \sum_{k=0}^{\infty} \frac{(-X)^k}{(k+1)!} u = \left[1 + \frac{-X}{2!} + \frac{X^2}{3!} + \frac{-X^3}{4!} + \frac{X^4}{5!} + \dots\right] u \\
&= \left[1 - \frac{1}{2!}\lambda + \frac{1}{3!}\lambda^2 - \frac{1}{4!}\lambda^3 + \frac{1}{5!}\lambda^4 + \dots\right] u \\
&= \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{(k+1)!} u = \frac{1}{\lambda} \sum_{k=0}^{\infty} \frac{(-\lambda)^{k+1}}{(k+1)!} u \\
&= \frac{1-e^{-\lambda}}{\lambda} u.
\end{aligned}$$

It follows that the matrix $\frac{\mathrm{id}-e^{-X}}{X}$ is invertible iff no λ_j is of the form $k2\pi i$ for some $k \in \mathbb{Z} - \{0\}$, so $d(\exp)_A$ is invertible iff no eigenvalue of ad_A is of the form $k2\pi i$ for some $k \in \mathbb{Z} - \{0\}$; this result is also found in Duistermaat and Kolk [43] (Chapter I, Section 5, Corollary 1.5.4) and Varadarajan [113] (Chapter 2, Section 14, Theorem 2.14.3). However, it can also be shown that if the eigenvalues of A are $\lambda_1, \dots, \lambda_n$, then the eigenvalues of ad_A are the $\lambda_i - \lambda_j$, with $1 \leq i, j \leq n$. In conclusion, $d(\exp)_A$ is invertible iff for all i, j we have

$$\lambda_i - \lambda_j \neq k2\pi i, \quad k \in \mathbb{Z} - \{0\}. \tag{$*$}$$

This suggests defining the following subset $\mathcal{E}(n)$ of $\mathrm{M}_n(\mathbb{R})$. The set $\mathcal{E}(n)$ consists of all matrices $A \in \mathrm{M}_n(\mathbb{R})$ whose eigenvalues $\lambda + i\mu$ of A ($\lambda, \mu \in \mathbb{R}$) lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. It is clear that the matrices in $\mathcal{E}(n)$ satisfy Condition ($*$), so $d(\exp)_A$ is invertible for all $A \in \mathcal{E}(n)$. By the inverse function theorem, the exponential map is a local diffeomorphism between $\mathcal{E}(n)$ and $\exp(\mathcal{E}(n))$. Remarkably, more is true: the exponential map is diffeomorphism between $\mathcal{E}(n)$ and $\exp(\mathcal{E}(n))$ (in particular, it is a bijection). This takes quite a bit of work to be proved. For example, see Mnemné and Testard [86], Chapter 3, Theorem 3.8.4 (see also Bourbaki [19], Chapter III, Section 6.9, Proposition 17, Theorem 6, and Varadarajan [113], Chapter 2, Section 14, Lemma 2.14.4). We have the following result.

Theorem 2.2. *The restriction of the exponential map to $\mathcal{E}(n)$ is a diffeomorphism of $\mathcal{E}(n)$ onto its image $\exp(\mathcal{E}(n))$. Furthermore, $\exp(\mathcal{E}(n))$ consists of all invertible matrices that have no real negative eigenvalues; it is an open subset of $\mathbf{GL}(n,\mathbb{R})$; it contains the open ball $B(I,1) = \{A \in \mathbf{GL}(n,\mathbb{R}) \mid \|A - I\| < 1\}$, for every matrix norm $\| \ \|$ on $n \times n$ matrices.*

Theorem 2.2 has some practical applications because there are algorithms for finding a real log of a matrix with no real negative eigenvalues; for more on applications of Theorem 2.2 to medical imaging, see Section 9.4.

2.3 Problems

Problem 2.1. Let $\mathrm{M}_n(\mathbb{C})$ denote the vector space of $n \times n$ matrices with complex coefficients (and $\mathrm{M}_n(\mathbb{R})$ denote the vector space of $n \times n$ matrices with real coefficients). For any matrix $A \in \mathrm{M}_n(\mathbb{C})$, let R_A and L_A be the maps from $\mathrm{M}_n(\mathbb{C})$ to itself defined so that

$$L_A(B) = AB, \quad R_A(B) = BA, \quad \text{for all } B \in \mathrm{M}_n(\mathbb{C}).$$

Check that L_A and R_A are linear, and that L_A and R_B commute for all A, B.

Let $\mathrm{ad_A} \colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the linear map given by

$$\mathrm{ad}_A(B) = L_A(B) - R_A(B) = AB - BA = [A, B], \quad \text{for all } B \in \mathrm{M}_n(\mathbb{C}).$$

Note that $[A, B]$ is the Lie bracket.

(1) Prove that if A is invertible, then L_A and R_A are invertible; in fact, $(L_A)^{-1} = L_{A^{-1}}$ and $(R_A)^{-1} = R_{A^{-1}}$. Prove that if $A = PBP^{-1}$ for some invertible matrix P, then

$$L_A = L_P \circ L_B \circ L_P^{-1}, \quad R_A = R_P^{-1} \circ R_B \circ R_P.$$

(2) Recall that the n^2 matrices E_{ij} defined such that all entries in E_{ij} are zero except the (i, j)th entry, which is equal to 1, form a basis of the vector space $\mathrm{M}_n(\mathbb{C})$. Consider the partial ordering of the E_{ij} defined such that for $i = 1, \ldots, n$, if $n \geq j > k \geq 1$, then then E_{ij} precedes E_{ik}, and for $j = 1, \ldots, n$, if $1 \leq i < h \leq n$, then E_{ij} precedes E_{hj}.

Draw the Hasse diagam of the partial order defined above when $n = 3$.

There are total orderings extending this partial ordering. How would you find them algorithmically? Check that the following is such a total order:

$$(1, 3), \ (1, 2), \ (1, 1), \ (2, 3), \ (2, 2), \ (2, 1), \ (3, 3), \ (3, 2), \ (3, 1).$$

(3) Let the total order of the basis (E_{ij}) extending the partial ordering defined in (2) be given by

$$(i, j) < (h, k) \quad \text{iff} \quad \begin{cases} i = h \text{ and } j > k \\ \text{or } i < h. \end{cases}$$

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of A (not necessarily distinct). Using Schur's theorem, A is similar to an upper triangular matrix B, that is, $A = PBP^{-1}$ with B upper triangular, and we may assume that the diagonal entries of B in descending order are $\lambda_1, \ldots, \lambda_n$. If the E_{ij} are listed according to the above total order, prove that R_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_n, \ldots, \lambda_1, \ldots, \lambda_n, \ldots, \lambda_1}_{n^2}),$$

and that L_B is an upper triangular matrix whose diagonal entries are

$$(\underbrace{\lambda_1, \ldots, \lambda_1}_{n} \ldots, \underbrace{\lambda_n, \ldots, \lambda_n}_{n}).$$

Hint. Figure out what are $R_B(E_{ij}) = E_{ij}B$ and $L_B(E_{ij}) = BE_{ij}$.

Use the fact that

$$L_A = L_P \circ L_B \circ L_P^{-1}, \quad R_A = R_P^{-1} \circ R_B \circ R_P,$$

to express $\mathrm{ad}_A = L_A - R_A$ in terms of $L_B - R_B$, and conclude that the eigenvalues of ad_A are $\lambda_i - \lambda_j$, for $i = 1, \ldots, n$, and for $j = n, \ldots, 1$.

(4) (**Extra Credit**) Let R be the $n \times n$ permutation matrix given by

$$R = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 & 0 \\ 1 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

Observe that $R^{-1} = R$. I checked for $n = 3$ that in the basis (E_{ij}) ordered as above, the matrix of L_A is given by $A \otimes I_3$, and the matrix of R_A is given by $I_3 \otimes RA^\top R$. Here, $\otimes$ the *Kronecker product* (also called *tensor product*) of matrices. It is natural to conjecture that for any $n \geq 1$, the matrix of L_A is given by $A \otimes I_n$, and the matrix of R_A is given by $I_n \otimes RA^\top R$. Prove this conjecture.

Problem 2.2.

(i) First show that

$$d(\exp)_A = \sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!}.$$

(ii) Next show that

$$\sum_{h,k \geq 0} \frac{L_A^h R_A^k}{(h+k+1)!} = e^{L_A} \sum_{j=0}^{\infty} \frac{(-1)^j}{(j+1)!}(L_A - R_A)^j.$$

Chapter 3

Introduction to Manifolds and Lie Groups

In this chapter we define precisely manifolds, Lie groups, and Lie algebras. One of the reasons that Lie groups are nice is that they have a differential structure, which means that the notion of tangent space makes sense at any point of the group. Furthermore, the tangent space at the identity happens to have some algebraic structure, that of a Lie algebra. Roughly speaking, the tangent space at the identity provides a "linearization" of the Lie group, and it turns out that many properties of a Lie group are reflected in its Lie algebra, and that the loss of information is not too severe. The challenge that we are facing is that unless our readers are already familiar with manifolds, the amount of basic differential geometry required to define Lie groups and Lie algebras in full generality is overwhelming.

Fortunately, most of the Lie groups that we will consider are subspaces of $\mathbb{R}^N$ for some sufficiently large N. In fact, most of them are isomorphic to subgroups of $\mathbf{GL}(N, \mathbb{R})$ for some suitable N, even $\mathbf{SE}(n)$, which is isomorphic to a subgroup of $\mathbf{SL}(n+1)$. Such groups are called *linear Lie groups* (or *matrix groups*). Since these groups are subspaces of $\mathbb{R}^N$, in a first stage, we do not need the definition of an abstract manifold. We just have to define embedded submanifolds (also called submanifolds) of $\mathbb{R}^N$ (in the case of $\mathbf{GL}(n, \mathbb{R})$, $N = n^2$). This is the path that we will follow. The general definition of manifold will be given in Chapter 7.

3.1 Introduction to Embedded Manifolds

In this section we provide the definition of an embedded submanifold. For simplicity, we restrict our attention to smooth manifolds. For detailed presentations, see DoCarmo [38, 39], Milnor [83], Marsden and Ratiu [77], Berger and Gostiaux [15], or Warner [114]. For the sake of brevity, we use the terminology *manifold* (but other authors would say *embedded submanifolds*, or something like that).

The intuition behind the notion of a smooth manifold in $\mathbb{R}^N$ is that a subspace M is a manifold of dimension m if every point $p \in M$ is contained in some open subset U of M (in the subspace topology) that can be parametrized by some function $\varphi \colon \Omega \to U$ from some open subset Ω in $\mathbb{R}^m$ containing the origin, and that φ has some nice properties that allow the definition of smooth functions on M and of the tangent space at p. For this, φ has to be at least a homeomorphism, but more is needed: φ must be smooth, and the derivative $\varphi'(0_m)$ at the origin must be injective (letting $0_m = \underbrace{(0, \ldots, 0)}_{m}$).

Definition 3.1. Given any integers N, m, with $N \geq m \geq 1$, an *m-dimensional smooth manifold in $\mathbb{R}^N$, for short a manifold*, is a nonempty subset M of $\mathbb{R}^N$ such that for every point $p \in M$ there are two open subsets $\Omega \subseteq \mathbb{R}^m$ and $U \subseteq M$, with $p \in U$, and a smooth function $\varphi \colon \Omega \to \mathbb{R}^N$ such that φ is a homeomorphism between Ω and $U = \varphi(\Omega)$, and $\varphi'(t_0)$ is injective, where $t_0 = \varphi^{-1}(p)$; see Figure 3.1. The function $\varphi \colon \Omega \to U$ is called a *(local) parametrization of M at p*. If $0_m \in \Omega$ and $\varphi(0_m) = p$, we say that $\varphi \colon \Omega \to U$ is *centered at p*.

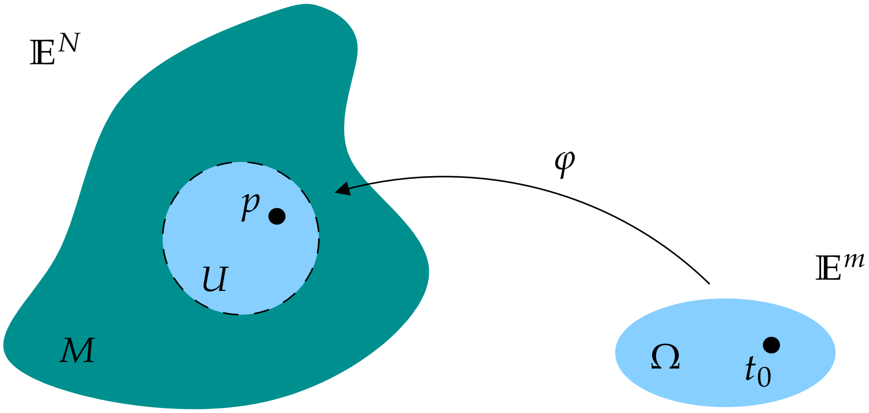


Figure 3.1: A manifold in $\mathbb{R}^N$.

Saying that $\varphi'(t_0)$ is injective is equivalent to saying that φ is an immersion at t_0.

Recall that $M \subseteq \mathbb{R}^N$ is a topological space under the subspace topology, and U is some open subset of M in the subspace topology, which means that $U = M \cap W$ for some open subset W of $\mathbb{R}^N$. Since $\varphi \colon \Omega \to U$ is a homeomorphism, it has an inverse $\varphi^{-1} \colon U \to \Omega$ that is also a homeomorphism, called a *(local) chart*. Since $\Omega \subseteq \mathbb{R}^m$, for every point $p \in M$ and every parametrization $\varphi \colon \Omega \to U$ of M at p, we have $\varphi^{-1}(p) = (z_1, \ldots, z_m)$ for some $z_i \in \mathbb{R}$, and we call $z_1, \ldots, z_m$ the *local coordinates of p (w.r.t. φ^{-1})*. We often refer to a manifold M without explicitly specifying its dimension (the integer m).

Intuitively, a chart provides a "flattened" local map of a region on a manifold. For instance, in the case of surfaces (2-dimensional manifolds), a chart is analogous to a planar map of a region on the surface. For a concrete example, consider a map giving a planar representation of a country, a region on the earth, a curved surface.

Remark: We could allow $m = 0$ in Definition 3.1. If so, a manifold of dimension 0 is just a set of isolated points, and thus it has the discrete topology. In fact, it can be shown that a discrete subset of $\mathbb{R}^N$ is countable. Such manifolds are not very exciting, but they do correspond to discrete subgroups.

Example 3.1. The unit sphere S^2 in $\mathbb{R}^3$ defined such that

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$$

is a smooth 2-manifold because it can be parametrized using the following two maps φ_1 and φ_2:

$$\varphi_1 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right)$$

and

$$\varphi_2 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{1 - u^2 - v^2}{u^2 + v^2 + 1}\right).$$

The map φ_1 corresponds to the inverse of the stereographic projection from the north pole $N = (0, 0, 1)$ onto the plane $z = 0$, and the map φ_2 corresponds to the inverse of the stereographic projection from the south pole $S = (0, 0, -1)$ onto the plane $z = 0$, as illustrated in Figure 3.2.

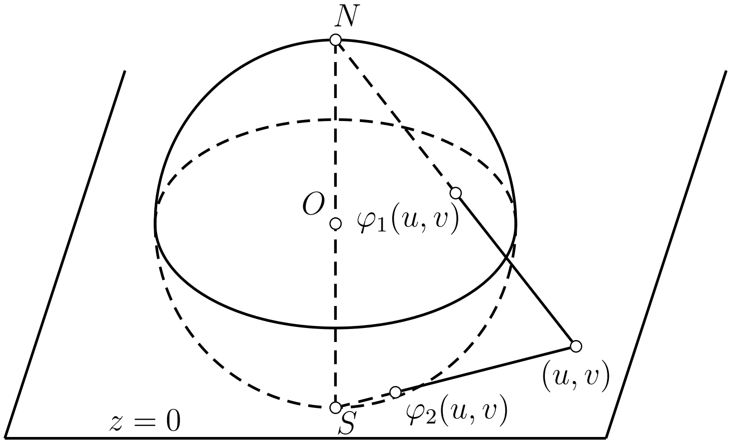


Figure 3.2: Inverse stereographic projections.

We demonstrate the algebraic constructions of φ_1 and φ_1^{-1}, leaving the constructions of φ_2 and φ_2^{-1} to the reader. Take S^2 and a point $Z = (x_1, x_2, x_3) \in S^2 - \{(0, 0, 1)\}$ and form

l, the line connecting $(0,0,1)$ and Z. Line l intersects the xy-plane at point $(u,v,0)$ and has equation $p+(1-t)\overrightarrow{v}$ where $p=(0,0,1)$ and $\overrightarrow{v}=(u,v,0)-(0,0,1)=(u,v,-1)$. See Figure 3.3.

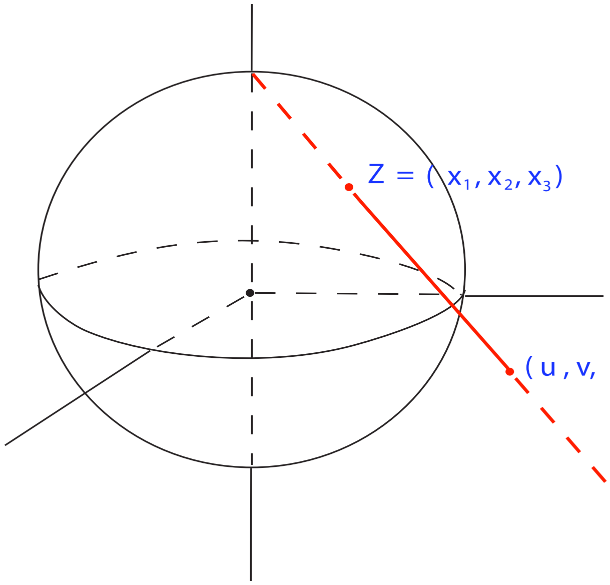


Figure 3.3: Line l is in red.

In other words, the line segment on Line l between $(u,v,0)$ and $(0,0,1)$ is parametrized by $((1-t)u,(1-t)v,t)$ for $0\le t\le 1$. The intersection of this line segment and S^2 is characterized by the equation

$$(1-t)^2u^2+(1-t)^2v^2+t^2=1, \qquad 0<t<1.$$

Take this equation, subtract t^2, and divide by $1-t$ to obtain

$$(1-t)(u^2+v^2)=1+t.$$

Solving this latter equation for t yields

$$t=\frac{u^2+v^2-1}{u^2+v^2+1} \quad \text{and} \quad 1-t=\frac{2}{u^2+v^2+1}.$$

By construction we know the intersection of the line segment with S^2 is $Z=(x_1,x_2,x_3)$. Hence, we conclude that

$$x_1=(1-t)u=\frac{2u}{u^2+v^2+1}, \quad x_2=(1-t)v=\frac{2v}{u^2+v^2+1}, \quad x_3=t=\frac{u^2+v^2-1}{u^2+v^2+1}.$$

To calculate φ_1^{-1}, we parameterize l by $((1-t)x_1, (1-t)x_2, (1-t)(x_3-1)+1)$. The intersection of Line l with the xy-plane is characterized by $((1-t)x_1, (1-t)x_2, (1-t)(x_3-1)+1) = (u, v, 0)$ and gives

$$(1-t)(x_3-1)+1 = 0.$$

Solving this equation for t implies that

$$t = -\frac{x_3}{1-x_3} \quad \text{and} \quad 1-t = \frac{1}{1-x_3}.$$

Hence $\varphi_1^{-1}(x_1, x_2, x_3) = (u, v)$, where

$$u = (1-t)x_1 = \frac{x_1}{1-x_3}, \qquad v = (1-t)x_2 = \frac{x_2}{1-x_3}.$$

We leave as an exercise to check that the map φ_1 parametrizes $S^2 - \{N\}$ and that the map φ_2 parametrizes $S^2 - \{S\}$ (and that they are smooth, homeomorphisms, etc.). Using φ_1, the open lower hemisphere is parametrized by the open disk of center O and radius 1 contained in the plane $z = 0$.

The chart φ_1^{-1} assigns local coordinates to the points in the open lower hemisphere. If we draw a grid of coordinate lines parallel to the x and y axes inside the open unit disk and map these lines onto the lower hemisphere using φ_1, we get curved lines on the lower hemisphere. These "coordinate lines" on the lower hemisphere provide local coordinates for every point on the lower hemisphere. For this reason, older books often talk about *curvilinear coordinate systems* to mean the coordinate lines on a surface induced by a chart. See Figure 3.4.

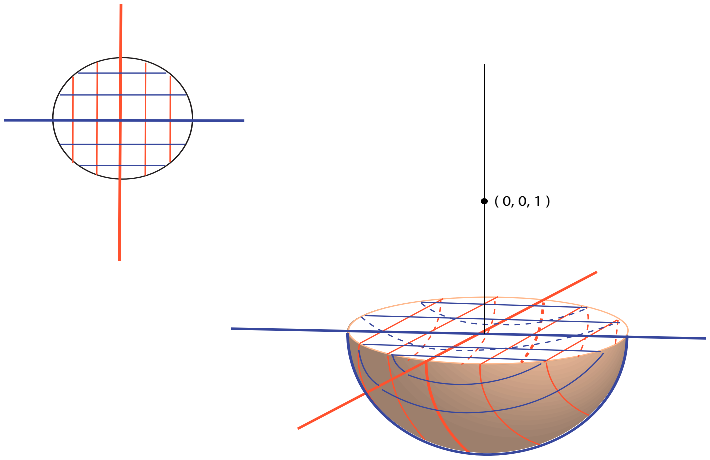


Figure 3.4: The curvilinear coordinates on the lower hemisphere of S^2 induced by φ_1.

We urge our readers to define a manifold structure on a torus. This can be done using four charts.

Every open subset of $\mathbb{R}^N$ is a manifold in a trivial way. Indeed, we can use the inclusion map as a parametrization. In particular, $\mathbf{GL}(n,\mathbb{R})$ is an open subset of $\mathbb{R}^{n^2}$, since its complement is closed (the set of invertible matrices is the inverse image of the determinant function, which is continuous). Thus, $\mathbf{GL}(n,\mathbb{R})$ is a manifold. We can view $\mathbf{GL}(n,\mathbb{C})$ as a subset of $\mathbb{R}^{(2n)^2}$ using the embedding defined as follows: For every complex $n\times n$ matrix A, construct the real $2n\times 2n$ matrix such that every entry $a+ib$ in A is replaced by the 2×2 block

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

where $a, b \in \mathbb{R}$. It is immediately verified that this map is in fact a group isomorphism. Thus we can view $\mathbf{GL}(n,\mathbb{C})$ as a subgroup of $\mathbf{GL}(2n,\mathbb{R})$, and as a manifold in $\mathbb{R}^{(2n)^2}$.

A 1-manifold is called a *(smooth) curve*, and a 2-manifold is called a *(smooth) surface* (although some authors require that they also be connected).

The following two lemmas provide the link with the definition of an abstract manifold. The first lemma is shown using Proposition 3.4 and is Condition (2) of Theorem 3.6; see below.

Lemma 3.1. *Given an m-dimensional manifold M in $\mathbb{R}^N$, for every $p\in M$ there are two open sets $O, W\subseteq \mathbb{R}^N$ with $0_N\in O$ and $p\in M\cap W$, and a smooth diffeomorphism $\varphi\colon O\to W$, such that $\varphi(0_N)=p$ and*

$$\varphi(O\cap(\mathbb{R}^m\times\{0_{N-m}\})) = M\cap W.$$

There is an open subset Ω of $\mathbb{R}^m$ such that

$$O\cap(\mathbb{R}^m\times\{0_{N-m}\}) = \Omega\times\{0_{N-m}\},$$

and the map $\psi\colon \Omega\to\mathbb{R}^N$ given by

$$\psi(x) = \varphi(x, 0_{N-m})$$

is an immersion and a homeomorphism onto $U = W\cap M$; so ψ is a parametrization of M at p. We can think of φ as a promoted version of ψ which is actually a diffeomorphism between open subsets of $\mathbb{R}^N$; see Figure 3.5.

The next lemma is easily shown from Lemma 3.1 (see Berger and Gostiaux [15], Theorem 2.1.9 or DoCarmo [39], Chapter 0, Section 4). It is a key technical result used to show that interesting properties of maps between manifolds do not depend on parametrizations.

Lemma 3.2. *Given an m-dimensional manifold M in $\mathbb{R}^N$, for every $p\in M$ and any two parametrizations $\varphi_1\colon \Omega_1\to U_1$ and $\varphi_2\colon\Omega_2\to U_2$ of M at p, if $U_1\cap U_2\neq\emptyset$, the map $\varphi_2^{-1}\circ\varphi_1\colon \varphi_1^{-1}(U_1\cap U_2)\to\varphi_2^{-1}(U_1\cap U_2)$ is a smooth diffeomorphism.*

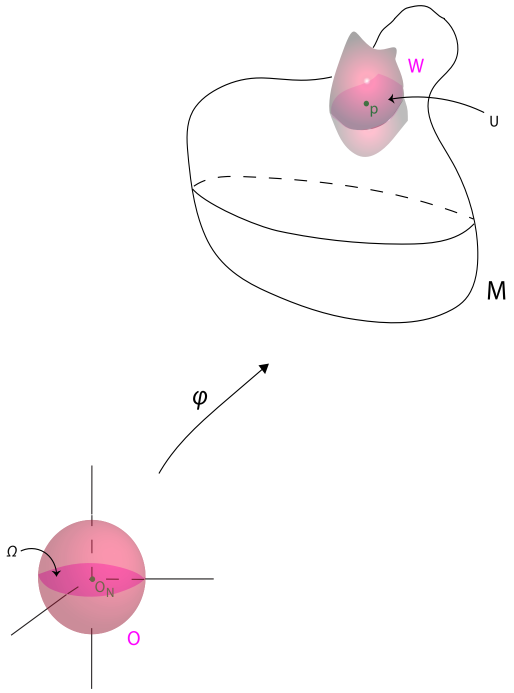


Figure 3.5: An illustration of Lemma 3.1, where M is a surface embedded in $\mathbb{R}^3$, namely $m = 2$ and $N = 3$.

The maps $\varphi_2^{-1} \circ \varphi_1 \colon \varphi_1^{-1}(U_1 \cap U_2) \to \varphi_2^{-1}(U_1 \cap U_2)$ are called *transition maps*. Lemma 3.2 is illustrated in Figure 3.6.

Using Definition 3.1, it may be quite hard to prove that a space is a manifold. Therefore, it is handy to have alternate characterizations such as those given in the next Proposition, which is Condition (3) of Theorem 3.6. An illustration of Proposition 3.3 is given by Figure 3.7.

Proposition 3.3. *A subset $M \subseteq \mathbb{R}^{m+k}$ is an m-dimensional manifold iff either*

(1) For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^{m+k}$ with $p \in W$, and a (smooth) submersion $f \colon W \to \mathbb{R}^k$, so that $W \cap M = f^{-1}(0)$,
or

(2) For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^{m+k}$ with $p \in W$, and a (smooth) map $f \colon W \to \mathbb{R}^k$, so that $f'(p)$ is surjective and $W \cap M = f^{-1}(0)$.

Observe that Condition (2), although apparently weaker than Condition (1), is in fact equivalent to it, but more convenient in practice. This is because to say that $f'(p)$ is surjective means that the Jacobian matrix of $f'(p)$ has rank k, which means that some determinant is

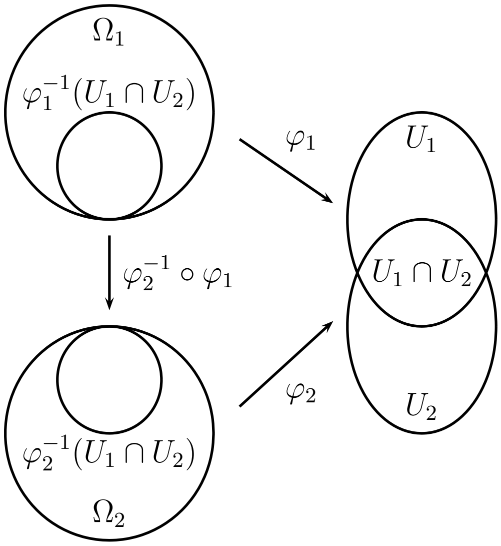

Ω_1
$\varphi_1^{-1}(U_1 \cap U_2)$
φ_1
U_1
$\varphi_2^{-1} \circ \varphi_1$
$U_1 \cap U_2$
φ_2
U_2
$\varphi_2^{-1}(U_1 \cap U_2)$
Ω_2

Figure 3.6: Parametrizations and transition functions.

nonzero, and because the determinant function is continuous this must hold in some open subset $W_1 \subseteq W$ containing p. Consequently, the restriction f_1 of f to W_1 is indeed a submersion, and $f_1^{-1}(0) = W_1 \cap f^{-1}(0) = W_1 \cap W \cap M = W_1 \cap M$.

A proof of Proposition 3.3 can be found in Lafontaine [72] or Berger and Gostiaux [15]. Lemma 3.1 and Proposition 3.3 are actually *equivalent* to Definition 3.1. This equivalence is also proved in Lafontaine [72] and Berger and Gostiaux [15].

Theorem 3.6, which combines Propositions 3.1 and 3.3, provides four equivalent characterizations of when a subspace of $\mathbb{R}^N$ is a manifold of dimension m. Its proof, which is somewhat illuminating, is based on two technical lemmas that are proved using the inverse function theorem (for example, see Guillemin and Pollack [55], Chapter 1, Sections 3 and 4).

Lemma 3.4. *Let $U \subseteq \mathbb{R}^m$ be an open subset of $\mathbb{R}^m$ and pick some $a \in U$. If $f \colon U \to \mathbb{R}^n$ is a smooth immersion at a, i.e., df_a is injective (so, $m \leq n$), then there is an open set $V \subseteq \mathbb{R}^n$ with $f(a) \in V$, an open subset $U' \subseteq U$ with $a \in U'$ and $f(U') \subseteq V$, an open subset $O \subseteq \mathbb{R}^{n-m}$, and a diffeomorphism $\theta \colon V \to U' \times O$, so that*

$$\theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

for all $(x_1, \ldots, x_m) \in U'$, as illustrated in the diagram below

$$\begin{tikzcd} U' \subseteq U \arrow[r, "f"] \arrow[dr, "in_1"'] & f(U') \subseteq V \arrow[d, "\theta"] \\ & U' \times O \end{tikzcd}$$

where $in_1(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0)$; see Figure 3.8.

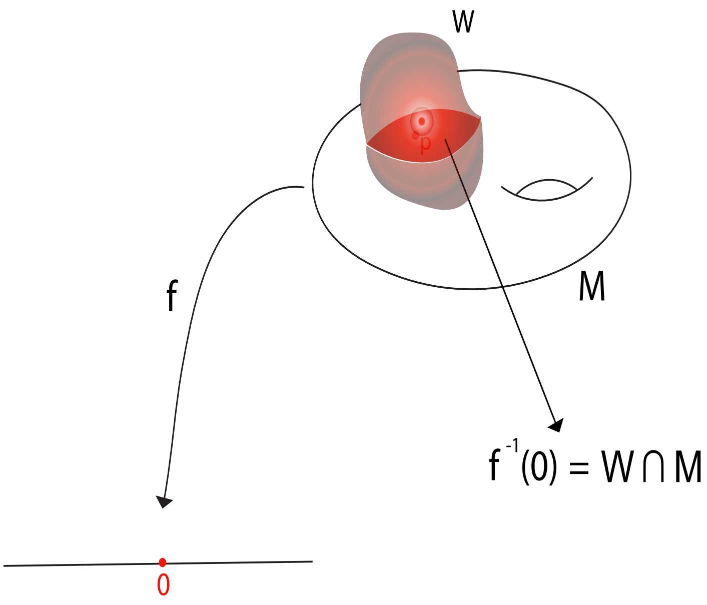


Figure 3.7: An illustration of Proposition 3.3, where M is the torus, $m = 2$, and $k = 1$. Note that $f^{-1}(0)$ is the pink patch of the torus, i.e. the zero level set of the open ball W.

Proof. Since f is an immersion, its Jacobian matrix $J(f)$ (an $n \times m$ matrix) has rank m, and by permuting coordinates if needed, we may assume that the first m rows of $J(f)$ are linearly independent and we let

$$A = \left(\frac{\partial f_i}{\partial x_j}(a) \right)$$

be this invertible $m \times m$ matrix. Define the map $g \colon U \times \mathbb{R}^{n-m} \to \mathbb{R}^n$ by

$$g(x, y) = (f_1(x), \ldots, f_m(x), y_1 + f_{m+1}(x), \ldots, y_{n-m} + f_n(x)),$$

for all $x \in U$ and all $y \in \mathbb{R}^{n-m}$. The Jacobian matrix of g at $(a, 0)$ is of the form

$$J = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix},$$

so $\det(J) = \det(A)\det(I) = \det(A) \neq 0$, since A is invertible. By the inverse function theorem, there are some open subsets $W \subseteq U \times \mathbb{R}^{n-m}$ with $(a, 0) \in W$ and $V \subseteq \mathbb{R}^n$ such that the restriction of g to W is a diffeomorphism between W and V. Since $W \subseteq U \times \mathbb{R}^{n-m}$ is an open set, we can find some open subsets $U' \subseteq U$ and $O \subseteq \mathbb{R}^{n-m}$ so that $U' \times O \subseteq W$, $a \in U'$, and we can replace W by $U' \times O$ and restrict further g to this open set so that we obtain a diffeomorphism from $U' \times O$ to (a smaller) V. If $\theta \colon V \to U' \times O$ is the inverse of this diffeomorphism, then $f(U') \subseteq V$ and since $g(x, 0) = f(x)$,

$$\theta(g(x, 0)) = \theta(f(x_1, \ldots, x_m)) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

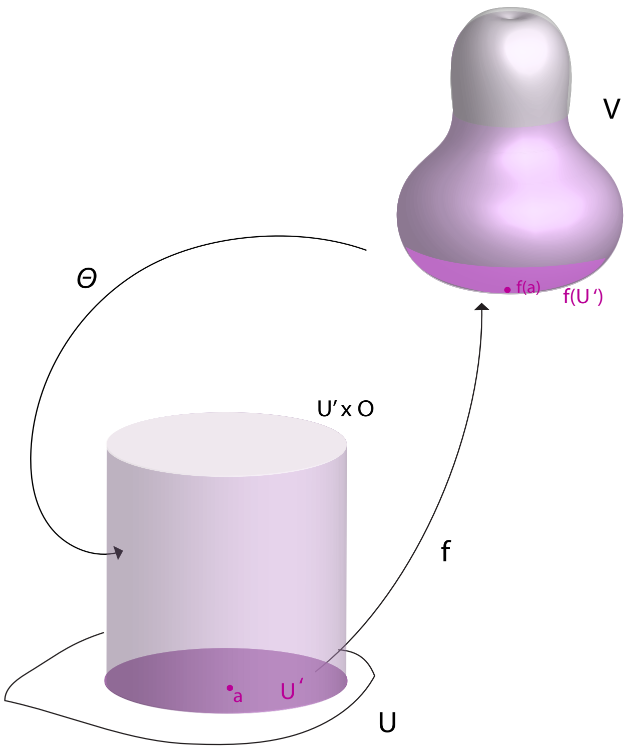


Figure 3.8: An illustration of Lemma 3.4, where $m = 2$ and $n = 3$. Note that U' is the base of the solid cylinder and θ is the diffeomorphism between the solid cylinder and the solid gourd shaped V. The composition $\theta \circ f$ injects U' into $U' \times O$.

for all $x = (x_1, \ldots, x_m) \in U'$. $\square$

Lemma 3.5. *Let $W \subseteq \mathbb{R}^m$ be an open subset of $\mathbb{R}^m$ and pick some $a \in W$. If $f \colon W \to \mathbb{R}^n$ is a smooth submersion at a, i.e., df_a is surjective (so, $m \geq n$), then there is an open set $V \subseteq W \subseteq \mathbb{R}^m$ with $a \in V$, and a diffeomorphism $\psi \colon O \to V$ with domain $O \subseteq \mathbb{R}^m$, so that*

$$f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n),$$

for all $(x_1, \ldots, x_m) \in O$, as illustrated in the diagram below

$$\begin{tikzcd}
O \subseteq \mathbb{R}^m \arrow[r, "\psi"] \arrow[d, "\pi"'] & V \subseteq W \subseteq \mathbb{R}^m \arrow[dl, "f"] \\
\mathbb{R}^n, &
\end{tikzcd}$$

where $\pi(x_1, \ldots, x_m) = (x_1, \ldots, x_n)$; see Figure 3.9.

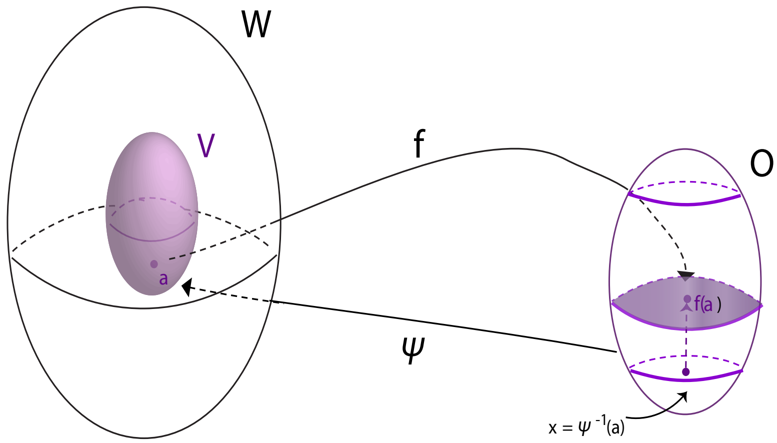


Figure 3.9: An illustration of Lemma 3.5, where $m = 3$ and $n = 2$. Note that ψ is the diffeomorphism between the 0 and the solid purple ball V. The composition $f \circ \psi$ projects O onto its equatorial pink disk.

Proof. Since f is a submersion, its Jacobian matrix $J(f)$ (an $n \times m$ matrix) has rank n, and by permuting coordinates if needed, we may assume that the first n columns of $J(f)$ are linearly independent and we let

$$A = \left(\frac{\partial f_i}{\partial x_j}(a) \right)$$

be this invertible $n \times n$ matrix. Define the map $g \colon W \to \mathbb{R}^m$ by

$$g(x) = (f(x), x_{n+1}, \ldots, x_m),$$

for all $x \in W$. The Jacobian matrix of g at a is of the form

$$J = \begin{pmatrix} A & B \\ 0 & I \end{pmatrix},$$

so $\det(J) = \det(A)\det(I) = \det(A) \neq 0$, since A is invertible. By the inverse function theorem, there are some open subsets $V \subseteq W$ with $a \in V$ and $O \subseteq \mathbb{R}^m$ such that the restriction of g to V is a diffeomorphism between V and O. Let $\psi \colon O \to V$ be the inverse of this diffeomorphism. Because $g \circ \psi = \mathrm{id}$, we have

$$(x_1, \ldots, x_m) = g(\psi(x)) = (f(\psi(x)), \psi_{n+1}(x), \ldots, \psi_m(x)),$$

that is,

$$f(\psi(x_1, \ldots, x_m)) = (x_1, \ldots, x_n)$$

for all $(x_1, \ldots, x_m) \in O$, as desired. $\square$

Using Lemmas 3.4 and 3.5, we can prove the following theorem which confirms that all our characterizations of a manifold are equivalent.

Theorem 3.6. *A nonempty subset $M \subseteq \mathbb{R}^N$ is an m-manifold (with $1 \leq m \leq N$) iff any of the following conditions hold:*

(1) For every $p \in M$, there are two open subsets $\Omega \subseteq \mathbb{R}^m$ and $U \subseteq M$ with $p \in U$, and a smooth function $\varphi\colon \Omega \to \mathbb{R}^N$ such that φ is a homeomorphism between Ω and $U = \varphi(\Omega)$, and $\varphi'(0)$ is injective, where $p = \varphi(0)$.

(2) For every $p \in M$, there are two open sets $O, W \subseteq \mathbb{R}^N$ with $0_N \in O$ and $p \in M \cap W$, and a smooth diffeomorphism $\varphi\colon O \to W$, such that $\varphi(0_N) = p$ and

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$

(3) For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^N$ with $p \in W$, and a smooth submersion $f\colon W \to \mathbb{R}^{N-m}$, so that $W \cap M = f^{-1}(0)$.

(4) For every $p \in M$, there is some open subset $W \subseteq \mathbb{R}^N$ with $p \in W$, and $N - m$ smooth functions $f_i\colon W \to \mathbb{R}$, so that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent, and

$$W \cap M = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0).$$

See Figure 3.10.

Proof. If (1) holds, then by Lemma 3.4, replacing Ω by a smaller open subset $\Omega' \subseteq \Omega$ if necessary, there is some open subset $V \subseteq \mathbb{R}^N$ with $p \in V$ and $\varphi(\Omega') \subseteq V$, an open subset $O' \subseteq \mathbb{R}^{N-m}$, and some diffeomorphism $\theta\colon V \to \Omega' \times O'$, so that

$$(\theta \circ \varphi)(x_1, \ldots, x_m) = (x_1, \ldots, x_m, 0, \ldots, 0),$$

for all $(x_1, \ldots, x_m) \in \Omega'$. Observe that the above condition implies that

$$(\theta \circ \varphi)(\Omega') = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).$$

Since φ is a homeomorphism between Ω and its image in M and since $\Omega' \subseteq \Omega$ is an open subset, $\varphi(\Omega') = M \cap W'$ for some open subset $W' \subseteq \mathbb{R}^N$, so if we let $W = V \cap W'$, because $\varphi(\Omega') \subseteq V$, it follows that $\varphi(\Omega') = M \cap W$ and

$$\theta(W \cap M) = \theta(\varphi(\Omega')) = \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}).$$

However, θ is injective and $\theta(W \cap M) \subseteq \theta(W)$, so

$$\begin{aligned} \theta(W \cap M) &= \theta(W) \cap \theta(V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) \\ &= \theta(W \cap V) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}) \\ &= \theta(W) \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\}). \end{aligned}$$

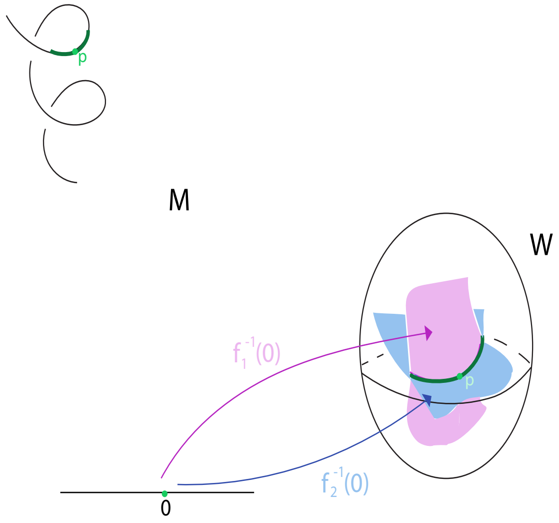


Figure 3.10: An illustration of Condition (4) in Theorem 3.6, where $N = 3$ and $m = 1$. The manifold M is the helix in $\mathbb{R}^3$. The dark green portion of M is magnified in order to show that it is the intersection of the pink surface, $f_1^{-1}(0)$, and the blue surface, $f_2^{-1}(0)$.

If we let $O = \theta(W)$, we get

$$\theta^{-1}(O \cap (\mathbb{R}^m \times \{(0, \ldots, 0)\})) = M \cap W,$$

which is (2).

If (2) holds, we can write $\varphi^{-1} = (f_1, \ldots, f_N)$ and because $\varphi^{-1} \colon W \to O$ is a diffeomorphism, $df_1(q), \ldots, df_N(q)$ are linearly independent for all $q \in W$, so the map

$$f = (f_{m+1}, \ldots, f_N)$$

is a submersion $f \colon W \to \mathbb{R}^{N-m}$, and we have $f(x) = 0$ iff $f_{m+1}(x) = \cdots = f_N(x) = 0$ iff

$$\varphi^{-1}(x) = (f_1(x), \ldots, f_m(x), 0, \ldots, 0)$$

iff $\varphi^{-1}(x) \in O \cap (\mathbb{R}^m \times \{0_{N-m}\})$ iff $x \in \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = M \cap W$, because

$$\varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})) = M \cap W.$$

Thus, $M \cap W = f^{-1}(0)$, which is (3).

The proof that (3) implies (2) uses Lemma 3.5 instead of Lemma 3.4. If $f \colon W \to \mathbb{R}^{N-m}$ is the submersion such that $M \cap W = f^{-1}(0)$ given by (3), then by Lemma 3.5, there are open subsets $V \subseteq W$, $O \subseteq \mathbb{R}^N$ and a diffeomorphism $\psi \colon O \to V$, so that

$$f(\psi(x_1, \ldots, x_N)) = (x_1, \ldots, x_{N-m})$$

for all $(x_1, \ldots, x_N) \in O$. If σ is the permutation of variables given by

$$\sigma(x_1, \ldots, x_m, x_{m+1}, \ldots, x_N) = (x_{m+1}, \ldots, x_N, x_1, \ldots, x_m),$$

then $\varphi = \psi \circ \sigma$ is a diffeomorphism such that

$$f(\varphi(x_1, \ldots, x_N)) = (x_{m+1}, \ldots, x_N)$$

for all $(x_1, \ldots, x_N) \in O$. If we denote the restriction of f to V by g, it is clear that

$$M \cap V = g^{-1}(0),$$

and because $g(\varphi(x_1, \ldots, x_N)) = 0$ iff $(x_{m+1}, \ldots, x_N) = 0_{N-m}$ and φ is a bijection,

$$\begin{aligned} M \cap V &= \{(y_1, \ldots, y_N) \in V \mid g(y_1, \ldots, y_N) = 0\} \\ &= \{\varphi(x_1, \ldots, x_N) \mid (\exists (x_1, \ldots, x_N) \in O)(g(\varphi(x_1, \ldots, x_N)) = 0)\} \\ &= \varphi(O \cap (\mathbb{R}^m \times \{0_{N-m}\})), \end{aligned}$$

which is (2).

If (2) holds, then $\varphi\colon O \to W$ is a diffeomorphism,

$$O \cap (\mathbb{R}^m \times \{0_{N-m}\}) = \Omega \times \{0_{N-m}\}$$

for some open subset $\Omega \subseteq \mathbb{R}^m$, and the map $\psi\colon \Omega \to \mathbb{R}^N$ given by

$$\psi(x) = \varphi(x, 0_{N-m})$$

is an immersion on Ω and a homeomorphism onto $W \cap M$, which implies (1).

If (3) holds, then if we write $f = (f_1, \ldots, f_{N-m})$, with $f_i\colon W \to \mathbb{R}$, then the fact that $df(p)$ is a submersion is equivalent to the fact that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent and

$$M \cap W = f^{-1}(0) = f_1^{-1}(0) \cap \cdots \cap f_{N-m}^{-1}(0).$$

Finally, if (4) holds, then if we define $f\colon W \to \mathbb{R}^{N-m}$ by

$$f = (f_1, \ldots, f_{N-m}),$$

because $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent we get a smooth map which is a submersion at p such that

$$M \cap W = f^{-1}(0).$$

Now, f is a submersion at p iff $df(p)$ is surjective, which means that a certain determinant is nonzero, and since the determinant function is continuous, this determinant is nonzero on some open subset $W' \subseteq W$ containing p, so if we restrict f to W', we get a submersion on W' such that $M \cap W' = f^{-1}(0)$. □

Condition (4) says that locally (that is, in a small open set of M containing $p \in M$), M is "cut out" by $N - m$ smooth functions $f_i \colon W \to \mathbb{R}$, in the sense that the portion of the manifold $M \cap W$ is the intersection of the $N - m$ hypersurfaces $f_i^{-1}(0)$ (the zero-level sets of the f_i), and that this intersection is "clean," which means that the linear forms $df_1(p), \ldots, df_{N-m}(p)$ are linearly independent.

As an illustration of Theorem 3.6, we can show again that the sphere

$$S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\|_2^2 - 1 = 0\}$$

is an n-dimensional manifold in $\mathbb{R}^{n+1}$. Indeed, the map $f \colon \mathbb{R}^{n+1} \to \mathbb{R}$ given by $f(x) = \|x\|_2^2 - 1$ is a submersion (for $x \neq 0$), since

$$df(x)(y) = 2\sum_{k=1}^{n+1} x_k y_k.$$

We can also show that the rotation group $\mathbf{SO}(n)$ is an $\frac{n(n-1)}{2}$-dimensional manifold in $\mathbb{R}^{n^2}$.

Indeed, $\mathbf{GL}^+(n)$ is an open subset of $\mathbb{R}^{n^2}$ of dimension n^2 (recall, $\mathbf{GL}^+(n) = \{A \in \mathbf{GL}(n) \mid \det(A) > 0\}$), and if f is defined by

$$f(A) = A^\top A - I,$$

where $A \in \mathbf{GL}^+(n)$, then $f(A)$ is symmetric, so $f(A) \in \mathbf{S}(n) = \mathbb{R}^{\frac{n(n+1)}{2}}$. We proved in Section 11.2 that

$$df(A)(H) = A^\top H + H^\top A.$$

But then, $df(A)$ is surjective for all $A \in \mathbf{SO}(n)$, because if S is any symmetric matrix, we see that

$$df(A)(AS/2) = A^\top \frac{AS}{2} + \left(\frac{AS}{2}\right)^\top A = A^\top A \frac{S}{2} + \frac{S^\top}{2} A^\top A = \frac{S}{2} + \frac{S^\top}{2} = S.$$

As $\mathbf{SO}(n) = f^{-1}(0)$, we conclude that $\mathbf{SO}(n)$ is indeed a manifold.

A similar argument proves that $\mathbf{O}(n)$ is an $\frac{n(n-1)}{2}$-dimensional manifold.

Using the map $f \colon \mathbf{GL}(n) \to \mathbb{R}$ given by $A \mapsto \det(A)$, we can prove that $\mathbf{SL}(n)$ is a manifold of dimension $n^2 - 1$.

Remark: We have $df(A)(B) = \det(A)\mathrm{tr}(A^{-1}B)$ for every $A \in \mathbf{GL}(n)$, where $f(A) = \det(A)$.

A class of manifolds generalizing the spheres and the orthogonal groups are the *Stiefel manifolds*. For any $n \geq 1$ and any k with $1 \leq k \leq n$, let $S(k, n)$ be the set of all *orthonormal k-frames*; that is, of k-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$. Obviously

$S(1,n) = S^{n-1}$, and $S(n,n) = \mathbf{O}(n)$. Every orthonormal k-frame $(u_1, \ldots, u_k)$ can be represented by an $n \times k$ matrix Y over the canonical basis of $\mathbb{R}^n$, and such a matrix Y satisfies the equation

$$Y^\top Y = I.$$

Thus, $S(k,n)$ can be viewed as a subspace of $\mathrm{M}_{n,k}(\mathbb{R})$, where $\mathrm{M}_{n,k}(\mathbb{R})$ denotes the vector space of all $n \times k$ matrices with real entries. We claim that $S(k,n)$ is a manifold. Let $W = \{A \in \mathrm{M}_{n,k}(\mathbb{R}) \mid \det(A^\top A) > 0\}$, an open subset of $\mathrm{M}_{n,k}(\mathbb{R})$ such that $S(k,n) \subseteq W$ (since if $A \in S(k,n)$, then $A^\top A = I$, so $\det(A^\top A) = 1$). Generalizing the situation involving $\mathbf{SO}(n)$, define the function $f\colon W \to \mathbf{S}(k)$ by

$$f(A) = A^\top A - I.$$

Basically the same computation as in the case of $\mathbf{SO}(n)$ yields

$$df(A)(H) = A^\top H + H^\top A.$$

The proof that $df(A)$ is surjective for all $A \in S(k,n)$ is the same as before, because only the equation $A^\top A = I$ is needed. Indeed, given any symmetric matrix $S \in \mathbf{S}(k) \cong \mathbb{R}^{\frac{k(k+1)}{2}}$, we have from our previous calculation that

$$df(A)\left(\frac{AS}{2}\right) = S.$$

As $S(k,n) = f^{-1}(0)$, we conclude that $S(k,n)$ is a smooth manifold of dimension

$$nk - \frac{k(k+1)}{2} = k(n-k) + \frac{k(k-1)}{2}.$$

The third characterization of Theorem 3.6 suggests the following definition.

Definition 3.2. Let $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$ be a smooth function. A point $p \in \mathbb{R}^{m+k}$ is called a *critical point (of f)* iff df_p is *not* surjective, and a point $q \in \mathbb{R}^k$ is called a *critical value (of f)* iff $q = f(p)$ for some critical point $p \in \mathbb{R}^{m+k}$. A point $p \in \mathbb{R}^{m+k}$ is a *regular point (of f)* iff p is not critical, i.e., df_p is surjective, and a point $q \in \mathbb{R}^k$ is a *regular value (of f)* iff it is not a critical value. In particular, any $q \in \mathbb{R}^k - f(\mathbb{R}^{m+k})$ is a regular value, and $q \in f(\mathbb{R}^{m+k})$ is a regular value iff *every* $p \in f^{-1}(q)$ is a regular point (in contrast, q is a critical value iff *some* $p \in f^{-1}(q)$ is critical).

Part (3) of Theorem 3.6 implies the following useful proposition:

Proposition 3.7. *Given any smooth function $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$, for every regular value $q \in f(\mathbb{R}^{m+k})$, the preimage $Z = f^{-1}(q)$ is a manifold of dimension m.*

Definition 3.2 and Proposition 3.7 can be generalized to manifolds. Regular and critical values of smooth maps play an important role in differential topology. Firstly, given a smooth map $f\colon \mathbb{R}^{m+k} \to \mathbb{R}^k$, almost every point of $\mathbb{R}^k$ is a regular value of f. To make this statement precise, one needs the notion of a *set of measure zero*. Then *Sard's theorem* says that the set of critical values of a smooth map has measure zero. Secondly, if we consider smooth functions $f\colon \mathbb{R}^{m+1} \to \mathbb{R}$, a point $p \in \mathbb{R}^{m+1}$ is critical iff $df_p = 0$. Then we can use second order derivatives to further classify critical points. The *Hessian matrix* of f (at p) is the matrix of second-order partials

$$H_f(p) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(p)\right),$$

and a critical point p is a *nondegenerate critical point* if $H_f(p)$ is a nonsingular matrix. The remarkable fact is that, at a nondegenerate critical point p, the local behavior of f is completely determined, in the sense that after a suitable change of coordinates (given by a smooth diffeomorphism)

$$f(x) = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{m+1}^2$$

near p, where λ, called the *index of f at p*, is an integer which depends only on p (in fact, λ is the number of negative eigenvalues of $H_f(p)$). This result is known as *Morse lemma* (after Marston Morse, 1892-1977).

Smooth functions whose critical points are all nondegenerate are called *Morse functions*. It turns out that every smooth function $f\colon \mathbb{R}^{m+1} \to \mathbb{R}$ gives rise to a large supply of Morse functions by adding a linear function to it. More precisely, the set of $a \in \mathbb{R}^{m+1}$ for which the function f_a given by

$$f_a(x) = f(x) + a_1x_1 + \cdots + a_{m+1}x_{m+1}$$

is not a Morse function has measure zero.

Morse functions can be used to study topological properties of manifolds. In a sense to be made precise and under certain technical conditions, a Morse function can be used to reconstruct a manifold by attaching cells, up to homotopy equivalence. However, these results are way beyond the scope of this book. A fairly elementary exposition of nondegenerate critical points and Morse functions can be found in Guillemin and Pollack [55] (Chapter 1, Section 7). Sard's theorem is proved in Appendix 1 of Guillemin and Pollack [55] and also in Chapter 2 of Milnor [83]. Morse theory (starting with Morse lemma) and much more, is discussed in Milnor [81], widely recognized as a mathematical masterpiece. An excellent and more leisurely introduction to Morse theory is given in Matsumoto [80], where a proof of Morse lemma is also given.

Let us now introduce the definitions of a smooth curve in a manifold and the tangent vector at a point of a curve.

Definition 3.3. Let M be an m-dimensional manifold in $\mathbb{R}^N$. A *smooth curve γ in M* is any function $\gamma\colon I \to M$ where I is an open interval in $\mathbb{R}$ and such that for every $t \in I$, letting $p = \gamma(t)$, there is some parametrization $\varphi\colon \Omega \to U$ of M at p and some open interval $(t - \epsilon,\, t + \epsilon) \subseteq I$ such that the curve $\varphi^{-1} \circ \gamma\colon (t - \epsilon,\, t + \epsilon) \to \mathbb{R}^m$ is smooth.

The notion of a smooth curve is illustrated in Figure 3.11.

Using Lemma 3.2, it is easily shown that Definition 3.3 does not depend on the choice of the parametrization $\varphi\colon \Omega \to U$ at p.

Lemma 3.2 also implies that γ viewed as a curve $\gamma\colon I \to \mathbb{R}^N$ is smooth. Then the *tangent vector to the curve $\gamma\colon I \to \mathbb{R}^N$ at t*, denoted by $\gamma'(t)$, is the value of the derivative of γ at t (a vector in $\mathbb{R}^N$) computed as usual:

$$\gamma'(t) = \lim_{h \mapsto 0} \frac{\gamma(t+h) - \gamma(t)}{h}.$$

Given any point $p \in M$, we will show that the set of tangent vectors to all smooth curves in M through p is a vector space isomorphic to the vector space $\mathbb{R}^m$. The tangent vector at p to a curve γ on a manifold M is illustrated in Figure 3.12.

Given a smooth curve $\gamma\colon I \to M$, for any $t \in I$, letting $p = \gamma(t)$, since M is a manifold, there is a parametrization $\varphi\colon \Omega \to U$ such that $\varphi(0_m) = p \in U$ and some open interval $J \subseteq I$ with $t \in J$ and such that the function

$$\varphi^{-1} \circ \gamma\colon J \to \mathbb{R}^m$$

is a smooth curve, since γ is a smooth curve. Letting $\alpha = \varphi^{-1} \circ \gamma$, the derivative $\alpha'(t)$ is well-defined, and it is a vector in $\mathbb{R}^m$. But $\varphi \circ \alpha\colon J \to M$ is also a smooth curve, which agrees with γ on J, and by the chain rule,

$$\gamma'(t) = \varphi'(0_m)(\alpha'(t)),$$

since $\alpha(t) = 0_m$ (because $\varphi(0_m) = p$ and $\gamma(t) = p$). See Figure 3.11. Observe that $\gamma'(t)$ is a vector in $\mathbb{R}^N$. Now for every vector $v \in \mathbb{R}^m$, the curve $\alpha\colon J \to \mathbb{R}^m$ defined such that

$$\alpha(u) = (u - t)v$$

for all $u \in J$ is clearly smooth, and $\alpha'(t) = v$. This shows that the set of tangent vectors at t to all smooth curves (in $\mathbb{R}^m$) passing through 0_m is the entire vector space $\mathbb{R}^m$. Since every smooth curve $\gamma\colon I \to M$ agrees with a curve of the form $\varphi \circ \alpha\colon J \to M$ for some smooth curve $\alpha\colon J \to \mathbb{R}^m$ (with $J \subseteq I$) as explained above, and since it is assumed that $\varphi'(0_m)$ is injective, $\varphi'(0_m)$ maps the vector space $\mathbb{R}^m$ injectively to the set of tangent vectors to γ at p, as claimed. All this is summarized in the following definition.

Definition 3.4. Let M be an m-dimensional manifold in $\mathbb{R}^N$. For every point $p \in M$, the *tangent space T_pM at p* is the set of all vectors in $\mathbb{R}^N$ of the form $\gamma'(0)$, where $\gamma\colon I \to M$ is any smooth curve in M such that $p = \gamma(0)$. The set T_pM is a vector space isomorphic to $\mathbb{R}^m$. Every vector $v \in T_pM$ is called a *tangent vector to M at p*.

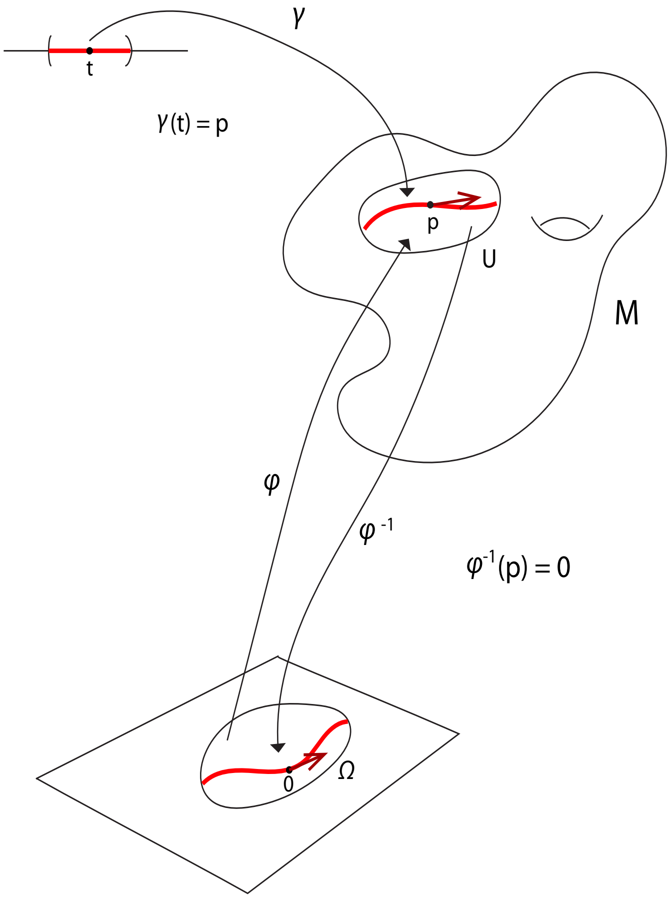


Figure 3.11: A smooth curve in a manifold M.

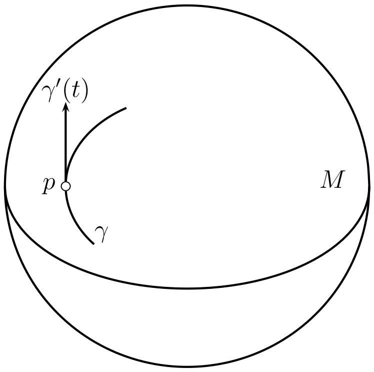


Figure 3.12: Tangent vector to a curve on a manifold.

Remark: The definition of a tangent vector at p involves smooth curves, where a smooth curve is defined in Definition 3.3. Actually, because of Lemma 3.1, it is only necessary to use curves that are C^1 viewed as curves in $\mathbb{R}^N$. The potential problem is that if φ is a parametrization at p and γ is a C^1 curve, it is not obvious that $\varphi^{-1} \circ \gamma$ is C^1 in $\mathbb{R}^m$. However, Lemma 3.1 allows us to promote φ to a diffeomorphism between open subsets of $\mathbb{R}^N$, and since both γ and (this new) φ^{-1} are C^1, so is $\varphi^{-1} \circ \gamma$. However, in the more general case of an abstract manifold M not assumed to be contained in some $\mathbb{R}^N$, smooth curves have to be defined as in Definition 3.3.

3.2 Linear Lie Groups

We can now define Lie groups (postponing defining smooth maps). In general, the difficult part in proving that a subgroup of $\mathbf{GL}(n, \mathbb{R})$ is a Lie group is to prove that it is a manifold. Fortunately, there is a characterization of the linear groups that obviates much of the work. This characterization rests on two theorems. First, a Lie subgroup H of a Lie group G (where H is an embedded submanifold of G) is closed in G (see Warner [114], Chapter 3, Theorem 3.21, page 97). Second, a theorem of Von Neumann and Cartan asserts that a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$ is an embedded submanifold, and thus, a Lie group (see Warner [114], Chapter 3, Theorem 3.42, page 110). Thus, a linear Lie group G is a **closed subgroup** of $\mathbf{GL}(n, \mathbb{R})$. Recall that this means that for every sequence $(A_n)_{n \geq 1}$ of matrices $A_n \in G$, if this sequence converges to a limit $A \in \mathbf{GL}(n, \mathbb{R})$, then actually $A \in G$.

Since our Lie groups are subgroups (or isomorphic to subgroups) of $\mathbf{GL}(n, \mathbb{R})$ for some suitable n, it is easy to define the Lie algebra of a Lie group using curves. This approach to define the Lie algebra of a matrix group is followed by a number of authors, such as Curtis [34]. However, Curtis is rather cavalier, since he does not explain why the required curves actually exist, and thus, according to his definition, Lie algebras could be the trivial vector space reduced to the zero element.

A small annoying technical problem will arise in our approach, the problem with discrete subgroups. If A is a subset of $\mathbb{R}^N$, recall that A inherits a topology from $\mathbb{R}^N$ called the *subspace topology*, defined such that a subset V of A is open if

$$V = A \cap U$$

for some open subset U of $\mathbb{R}^N$. A point $a \in A$ is said to be *isolated* if there is some open subset U of $\mathbb{R}^N$ such that

$$\{a\} = A \cap U,$$

in other words, if $\{a\}$ is an open set in A.

The group $\mathbf{GL}(n, \mathbb{R})$ of real invertible $n \times n$ matrices can be viewed as a subset of $\mathbb{R}^{n^2}$, and as such, it is a topological space under the subspace topology (in fact, a dense open subset of $\mathbb{R}^{n^2}$). One can easily check that multiplication and the inverse operation are continuous, and in fact smooth (i.e., C^∞-continuously differentiable). This makes $\mathbf{GL}(n, \mathbb{R})$ a *topological* *group*. Any subgroup G of $\mathbf{GL}(n, \mathbb{R})$ is also a topological space under the subspace topology. A subgroup G is called a *discrete subgroup* if it has some isolated point. This turns out to be equivalent to the fact that every point of G is isolated, and thus, G has the discrete topology (every subset of G is open). Because $\mathbf{GL}(n, \mathbb{R})$ is a topological group, every discrete subgroup of $\mathbf{GL}(n, \mathbb{R})$ is closed (which means that its complement is open); see Proposition 4.5. Moreover, since $\mathbf{GL}(n, \mathbb{R})$ is the union of countably many compact subsets, discrete subgroups of $\mathbf{GL}(n, \mathbb{R})$ must be countable. Thus, discrete subgroups of $\mathbf{GL}(n, \mathbb{R})$ are Lie groups (and countable)! But these are not very interesting Lie groups, and so we will consider only closed subgroups of $\mathbf{GL}(n, \mathbb{R})$ that are not discrete.

Definition 3.5. A *Lie group* is a nonempty subset G of $\mathbb{R}^N$ ($N \geq 1$) satisfying the following conditions:

(a) G is a group.

(b) G is a manifold in $\mathbb{R}^N$.

(c) The group operation $\cdot\colon G \times G \to G$ and the inverse map ${}^{-1}\colon G \to G$ are smooth.

(Smooth maps are defined in Definition 3.8). It is immediately verified that $\mathbf{GL}(n, \mathbb{R})$ is a Lie group. Since all the Lie groups that we are considering are subgroups of $\mathbf{GL}(n, \mathbb{R})$, the following definition is in order.

Definition 3.6. A *linear Lie group* is a subgroup G of $\mathbf{GL}(n, \mathbb{R})$ (for some $n \geq 1$) which is a smooth manifold in $\mathbb{R}^{n^2}$.

Let $\mathrm{M}_n(\mathbb{R})$ denote the set of all real $n \times n$ matrices (invertible or not). If we recall that the exponential map

$$\exp\colon A \mapsto e^A$$

is well defined on $\mathrm{M}_n(\mathbb{R})$, we have the following crucial theorem due to Von Neumann and Cartan.

Theorem 3.8. *(Von Neumann and Cartan, 1927) A closed subgroup G of $\mathbf{GL}(n, \mathbb{R})$ is a linear Lie group. Furthermore, the set $\mathfrak{g}$ defined such that*

$$\mathfrak{g} = \{X \in \mathrm{M}_n(\mathbb{R}) \mid e^{tX} \in G \textit{ for all } t \in \mathbb{R}\}$$

is a nontrivial vector space equal to the tangent space $T_I G$ at the identity I, and $\mathfrak{g}$ is closed under the Lie bracket $[-,-]$ defined such that $[A, B] = AB - BA$ for all $A, B \in \mathrm{M}_n(\mathbb{R})$.

Theorem 3.8 applies even when G is a discrete subgroup, but in this case, $\mathfrak{g}$ is trivial (i.e., $\mathfrak{g} = \{0\}$). For example, the set of nonnull reals $\mathbb{R}^* = \mathbb{R} - \{0\} = \mathbf{GL}(1, \mathbb{R})$ is a Lie group under multiplication, and the subgroup

$$H = \{2^n \mid n \in \mathbb{Z}\}$$

is a discrete subgroup of $\mathbb{R}^*$. Thus, H is a Lie group. On the other hand, the set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ of nonnull rational numbers is a multiplicative subgroup of $\mathbb{R}^*$, but it is not closed, since $\mathbb{Q}$ is dense in $\mathbb{R}$. Hence $\mathbb{Q}^*$ is not a Lie subgroup of $\mathbf{GL}(1, \mathbb{R})$.

The first step in proving Theorem 3.8 is to show that if G is a closed and nondiscrete subgroup of $\mathbf{GL}(n, \mathbb{R})$ and if we define $\mathfrak{g}$ just as $T_I G$ (even though we don't know yet that G is a manifold), then $\mathfrak{g}$ is a vector space satisfying the properties of Theorem 3.8. We follow the treatment in Kosmann [70], which we find one of the simplest and clearest.

Proposition 3.9. *Given any closed subgroup G in $\mathbf{GL}(n, \mathbb{R})$, the set*

$$\mathfrak{g} = \{X \in \mathrm{M}_n(\mathbb{R}) \mid X = \gamma'(0), \gamma\colon J \to G \text{ is a } C^1 \text{ curve in } \mathrm{M}_n(\mathbb{R}) \text{ such that } \gamma(0) = I\}$$

satisfies the following properties:

(1) $\mathfrak{g}$ is a vector subspace of $\mathrm{M}_n(\mathbb{R})$.

(2) For every $X \in \mathrm{M}_n(\mathbb{R})$, we have $X \in \mathfrak{g}$ iff $e^{tX} \in G$ for all $t \in \mathbb{R}$.

(3) For every $X \in \mathfrak{g}$ and for every $g \in G$, we have $gXg^{-1} \in \mathfrak{g}$.

(4) $\mathfrak{g}$ is closed under the Lie bracket.

Proof. If γ is a C^1 curve in G such that $\gamma(0) = I$ and $\gamma'(0) = X$, then for any $\lambda \in \mathbb{R}$, the curve $\alpha(t) = \gamma(\lambda t)$ passes through I and $\alpha'(0) = \lambda X$. If γ_1 and γ_2 are two C^1 curves in G such that $\gamma_1(0) = \gamma_2(0) = I$, $\gamma_1'(0) = X$, and $\gamma_2'(0) = Y$, then the curve $\alpha(t) = \gamma_1(t)\gamma_2(t)$ passes through I and the product rule implies

$$\alpha'(0) = (\gamma_1(t)\gamma_2(t))'(0) = X + Y.$$

Therefore, $\mathfrak{g}$ is a vector space.

(2) If $e^{tX} \in G$ for all $t \in \mathbb{R}$, then $\gamma\colon t \mapsto e^{tX}$ is a smooth curve through I in G such that $\gamma'(0) = X$, so $X \in \mathfrak{g}$.

Conversely, if $X = \gamma'(0)$ for some C^1 curve in G such that $\gamma(0) = I$, using the Taylor expansion of γ near 0, for every $t \in \mathbb{R}$ and for any positive integer k large enough t/k is small enough so that $\gamma(t/k) \in G$ and we have

$$\gamma\left(\frac{t}{k}\right) = I + \frac{t}{k}X + \epsilon_1(k) = \exp\left(\frac{t}{k}X + \epsilon_2(k)\right),$$

where $\epsilon_1(k)$ is $O(1/k^2)$, i.e. $|\epsilon_1(k)| \leq \frac{C}{k^2}$ for some nonnegative C, and $\epsilon_2(k)$ is also $O(1/k^2)$. Raising to the kth power, we deduce that

$$\gamma\left(\frac{t}{k}\right)^k = \exp\left(tX + \epsilon_3(k)\right),$$

where $\epsilon_3(k)$ is $O(1/k)$, and by the continuity of the exponential, we get

$$\lim_{k \mapsto \infty} \gamma \left(\frac{t}{k} \right)^k = \exp(tX).$$

For all k large enough, since G is a closed subgroup, $(\gamma(t/k))^k \in G$ and

$$\lim_{k \mapsto \infty} \gamma \left(\frac{t}{k} \right)^k \in G,$$

and thus $e^{tX} \in G$.

(3) We know by Proposition 1.2 that

$$e^{tgXg^{-1}} = ge^{tX}g^{-1},$$

and by (2), if $X \in \mathfrak{g}$, then $e^{tX} \in G$ for all t, and since $g \in G$, we have $e^{tgXg^{-1}} = ge^{tX}g^{-1} \in G$. Since $(ge^{tX}g^{-1})'(t) = gXe^{tX}g^{-1}$, the definition of $\mathfrak{g}$ implies that

$$(e^{tgXg^{-1}})'(0) = (ge^{tX}g^{-1})'(0) = gXg^{-1} \in \mathfrak{g}.$$

(4) if $X, Y \in \mathfrak{g}$, then by (2), for all $t \in \mathbb{R}$ we have $e^{tX} \in G$, and by (3), $e^{tX}Ye^{-tX} \in \mathfrak{g}$. By the product rule we obtain

$$(e^{tX}Ye^{-tX})'(t) = Xe^{tX}Ye^{-tX} - e^{tX}YXe^{-tX},$$

which in turn implies

$$(e^{tX}Ye^{-tX})'(0) = XY - YX$$

and proves that $\mathfrak{g}$ is a Lie algebra. □

The second step in the proof of Theorem 3.8 is to prove that when G is not a discrete subgroup, there is an open subset $\Omega \subseteq \mathrm{M}_n(\mathbb{R})$ such that $0 \in \Omega$, an open subset $W \subseteq \mathbf{GL}(n, \mathbb{R})$ such that $I \in W$, and a diffeomorphism $\Phi \colon \Omega \to W$ such that

$$\Phi(\Omega \cap \mathfrak{g}) = W \cap G.$$

If G is closed and not discrete, we must have $m \geq 1$, and $\mathfrak{g}$ has dimension m.

We begin by observing that the exponential map is a diffeomorphism between some open subset of 0 and some open subset of I. This is because $d(\exp)_0 = \mathrm{id}$, which is easy to see since

$$e^X - I = X + \|X\| \, \epsilon(X)$$

with

$$\epsilon(X) = \frac{1}{\|X\|} \sum_{k=0}^{\infty} \frac{X^{k+2}}{(k+2)!},$$

and so $\lim_{X \mapsto 0} \epsilon(X) = 0$. By the inverse function theorem, exp is a diffeomorphism between some open subset U_0 of $\mathrm{M}_n(\mathbb{R})$ containing 0 and some open subset V_0 of $\mathbf{GL}(n, \mathbb{R})$ containing I.

Proposition 3.10. *Let G be a subgroup of $\mathbf{GL}(n, \mathbb{R})$, and assume that G is closed and not discrete. Then $\dim(\mathfrak{g}) \geq 1$, and the exponential map is a diffeomorphism of a neighborhood of 0 in $\mathfrak{g}$ onto a neighborhood of I in G. Furthermore, there is an open subset $\Omega \subseteq \mathrm{M}_n(\mathbb{R})$ with $0 \in \Omega$, an open subset $W \subseteq \mathbf{GL}(n, \mathbb{R})$ with $I \in W$, and a diffeomorphism $\Phi \colon \Omega \to W$ such that*

$$\Phi(\Omega \cap \mathfrak{g}) = W \cap G.$$

Proof. We follow the proof in Kosmann [70] (Chapter 4, Section 5). A similar proof is given in Helgason [58] (Chapter 2, §2), Mneimné and Testard [86] (Chapter 3, Section 3.4), and in Duistermaat and Kolk [43] (Chapter 1, Section 10). As explained above, by the inverse function theorem, exp is a diffeomorphism between some open subset U_0 of $\mathrm{M}_n(\mathbb{R})$ containing 0 and some open subset V_0 of $\mathbf{GL}(n, \mathbb{R})$ containing I. Let $\mathfrak{p}$ be any subspace of $\mathrm{M}_n(\mathbb{R})$ such that $\mathfrak{g}$ and $\mathfrak{p}$ form a direct sum

$$\mathrm{M}_n(\mathbb{R}) = \mathfrak{g} \oplus \mathfrak{p},$$

and let $\Phi \colon \mathfrak{g} \oplus \mathfrak{p} \to G$ be the map defined by

$$\Phi(X + Y) = e^X e^Y.$$

We claim that $d\Phi_0 = \mathrm{id}$. One way to prove this is to observe that for $\|X\|$ and $\|Y\|$ small,

$$e^X = I + X + \|X\| \, \epsilon_1(X) \quad e^Y = I + Y + \|Y\| \, \epsilon_2(Y),$$

with $\lim_{X \mapsto 0} \epsilon_1(X) = 0$ and $\lim_{Y \mapsto 0} \epsilon_2(Y) = 0$, so we get

$$\begin{aligned} e^X e^Y &= I + X + Y + XY \\ &\quad + \|X\| \, \epsilon_1(X)(I + Y) + \|Y\| \, \epsilon_2(Y)(I + X) + \|X\| \, \|Y\| \, \epsilon_1(X)\epsilon_2(X) \\ &= I + X + Y + \left(\sqrt{\|X\|^2 + \|Y\|^2}\right)\epsilon(X, Y), \end{aligned}$$

with

$$\begin{aligned} \epsilon(X, Y) = &\frac{\|X\|}{\sqrt{\|X\|^2 + \|Y\|^2}} \epsilon_1(X)(I + Y) + \frac{\|Y\|}{\sqrt{\|X\|^2 + \|Y\|^2}} \epsilon_2(Y)(I + X) \\ &+ \frac{XY + \|X\| \, \|Y\| \, \epsilon_1(X)\epsilon_2(X)}{\sqrt{\|X\|^2 + \|Y\|^2}}. \end{aligned}$$

Since $\lim_{X \mapsto 0} \epsilon_1(X) = 0$ and $\lim_{Y \mapsto 0} \epsilon_2(Y) = 0$, the first two terms go to 0 when X and Y go to 0, and since

$$\begin{aligned} \|XY + \|X\| \, \|Y\| \, \epsilon_1(X)\epsilon_2(X)\| &\leq \|X\| \, \|Y\| \, (1 + \|\epsilon_1(X)\epsilon_2(X)\|) \\ &\leq \frac{1}{2}(\|X\|^2 + \|Y\|^2)(1 + \|\epsilon_1(X)\epsilon_2(X)\|), \end{aligned}$$

we have

$$\left\| \frac{XY + \|X\| \, \|Y\| \, \epsilon_1(X)\epsilon_2(X)}{\sqrt{\|X\|^2 + \|Y\|^2}} \right\| \leq \frac{1}{2}\Big(\sqrt{\|X\|^2 + \|Y\|^2}\Big)(1 + \|\epsilon_1(X)\epsilon_2(X)\|),$$

so the third term also goes to 0 when X and Y to 0. Therefore, $\lim_{X\mapsto 0, Y\mapsto 0} \epsilon(X, Y) = 0$, and $d\Phi_0(X + Y) = X + Y$, as claimed.

By the inverse function theorem, there exists an open subset of $\mathrm{M}_n(\mathbb{R})$ containing 0 of the form $U' + U''$ with $U' \subseteq \mathfrak{g}$ and $U'' \subseteq \mathfrak{p}$ and some open subset W' of $\mathbf{GL}(n, \mathbb{R})$ such that Φ is a diffeomorphism of $U' + U''$ onto W'. By considering $U_0 \cap (U' + U'')$, we may assume that $U_0 = U' + U''$, and write $W' = \Phi(U_0)$; the maps exp and Φ are diffeomorphisms on U_0.

Since $U' \subseteq \mathfrak{g}$, we have $\exp(U') \subseteq W' \cap G$, but we would like equality to hold.

Suppose we can show that there is some open subset $U_0'' \subseteq U'' \subseteq \mathfrak{p}$ such that for all $X \in U_0''$, if $e^X \in G$, then $X = 0$. If so, consider the restriction of Φ to $U' \oplus U_0''$, and let $W = \Phi(U' \oplus U_0'')$; clearly, $\exp(U') \subseteq W \cap G$. Then, since Φ maps $U' + U_0''$ onto W, for any $g \in W \cap G$, we have $g = e^{X'}e^{X''}$ for some $X' \in U' \subseteq \mathfrak{g}$ and some $X'' \in U_0'' \subseteq \mathfrak{p}$. Then, $e^{X'} \in G$ since $X' \in \mathfrak{g}$, so $e^{X''} = e^{-X'}g \in G$. However, as $X'' \in U_0''$, we must have $X'' = 0$, and thus $W \cap G = \exp(U')$. This proves that exp is a diffeomorphism of $U' \subseteq \mathfrak{g}$ onto $W \cap G$, which is the first statement of Proposition 3.10.

For the second part of Proposition 3.10, if we let $\Omega = U' + U_0''$ and $W = \exp(\Omega)$, then Ω is an open subset of $\mathrm{M}_n(\mathbb{R})$ containing 0, W is an open subset of $\mathbf{GL}(n, \mathbb{R})$ containing I, $U' = \Omega \cap \mathfrak{g}$, and Φ is a diffeomorphism of Ω onto W such that $\Phi(\Omega \cap \mathfrak{g}) = W \cap G$, as desired.

We still need to prove the following claim:

Claim. There exists an open subset $U_0'' \subseteq U'' \subseteq \mathfrak{p}$ such that for all $X \in U_0''$, if $e^X \in G$, then $X = 0$

The proof of the claim relies on the fact that G is closed.

Proof of the Claim. We proceed by contradiction. If the claim is false, then in every open subset of $\mathfrak{p}$ containing 0, there is some $X \neq 0$ such that $e^X \in G$. In particular, for every positive integer n, there is some $X_n \in B(0, 1/n) \cap \mathfrak{p}$ such that $X_n \neq 0$ and $e^{X_n} \in G$ (where $B(0, 1/n)$ denotes the open ball of center 0 and radius $1/n$). We obtain a sequence (X_n) in $\mathfrak{p}$ whose limit is 0, and thus the sequence (e^{X_n}) converges to I in G. Define the sequence (Z_n) by

$$Z_n = \frac{X_n}{\|X_n\|},$$

so that $\|Z_n\| = 1$. Since the unit sphere is compact, there is some subsequence of (Z_n) that converges to a limit Z in $\mathfrak{p}$ of unit norm (since $\mathfrak{p}$ is closed); from now on, consider this converging subsequence of (Z_n) and the corresponding subsequence of X_n (which still converges to 0, with $X_n \neq 0$ for all n).

Lemma 3.11. *Let G be a closed subgroup of* $\mathbf{GL}(n, \mathbb{R})$ *and let* $\mathfrak{m}$ *be any subspace of* $\mathrm{M}_n(\mathbb{R})$*. For any sequence (X_n) of nonzero matrices in $\mathfrak{m}$, if $e^{X_n} \in G$ for all n, if (X_n) converges to 0, and if the sequence (Z_n) given by*

$$Z_n = \frac{X_n}{\|X_n\|}$$

converges to a limit Z (necessarily in $\mathfrak{m}$ and with $\|Z\| = 1$), then $Z \in \mathfrak{g}$.

Proof. We would like to prove that $e^{tZ} \in G$ for all $t \in \mathbb{R}$, because then, by Proposition 3.9(2), $Z \in \mathfrak{g}$. For any $t \in \mathbb{R}$, write

$$\frac{t}{\|X_n\|} = p_n(t) + u_n(t), \quad \text{with } p_n(t) \in \mathbb{Z} \text{ and } u_n(t) \in [0, 1).$$

Then we have

$$e^{tZ_n} = e^{\left(\frac{t}{\|X_n\|} X_n\right)} = (e^{X_n})^{p_n(t)} e^{u_n(t) X_n}.$$

Since $u_n(t) \in [0, 1)$ and since the sequence (X_n) converges to 0, the sequence $(u_n(t)X_n)$ also converges to 0, so the sequence $e^{u_n(t)X_n}$ converges to I. Furthermore, since $p_n(t)$ is an integer, $e^{X_n} \in G$, and G is a group, we have $(e^{X_n})^{p_n(t)} \in G$. Since G is closed, the limit of the sequence $e^{tZ_n} = (e^{X_n})^{p_n(t)} e^{u_n(t)X_n}$ belongs to G, and since $\lim_{n \mapsto \infty} Z_n = Z$, by the continuity of the exponential, we conclude that $e^{tZ} \in G$. Since this holds for all $t \in \mathbb{R}$, we have $Z \in \mathfrak{g}$. □

Applying Lemma 3.11 to $\mathfrak{m} = \mathfrak{p}$, we deduce that $Z \in \mathfrak{g} \cap \mathfrak{p} = (0)$, so $Z = 0$, contradicting the fact that $\|Z\| = 1$. Therefore, the claim holds. □

It remains to prove that $\mathfrak{g}$ is nontrivial. This is where the assumption that G is not discrete is needed. Indeed, if G is not discrete, we can find a sequence (g_n) of elements of G such that $g_n \neq I$ and the sequence converges to I. Since the exponential is a diffeomorphism between a neighborhood of 0 and a neighborhood of I, we may assume by dropping some initial segment of the sequence that $g_n = e^{X_n}$ for some nonzero matrices X_n, and that the sequence (X_n) converges to 0. For n large enough, the sequence

$$Z_n = \frac{X_n}{\|X_n\|}$$

makes sense and belongs to the unit sphere. By compactness of the unit sphere, (Z_n) has some subsequence that converges to some matrix Z with $\|Z\| = 1$. The corresponding subsequence of X_n still consists of nonzero matrices and converges to 0. We can apply Lemma 3.11 to $\mathfrak{m} = \mathrm{M}_n(\mathbb{R})$ and to the converging subsequences of (X_n) and (Z_n) to conclude that $Z \in \mathfrak{g}$, with $Z \neq 0$. This proves that $\dim(\mathfrak{g}) \geq 1$, and completes the proof of Proposition 3.10. □

Remark: The first part of Proposition 3.10 shows that exp is a diffeomorphism of an open subset $U' \subseteq \mathfrak{g}$ containing 0 onto $W \cap G$, which is Condition (1) of Theorem 3.6; that is, the restriction of exp to U' is a parametrization of G.

Theorem 3.8 now follows immediately from Propositions 3.9 and 3.10.

Proof of Theorem 3.8. Proposition 3.9 shows that $\mathfrak{g} = T_I G$ and that it is a Lie algebra. Proposition 3.10 shows that Condition (2) of Theorem 3.6 holds; that is, there is an open subset $\Omega \subseteq \mathrm{M}_n(\mathbb{R})$ with $0 \in \Omega$, an open subset $W \subseteq \mathbf{GL}(n, \mathbb{R})$ with $I \in W$, and a diffeomorphism $\Phi \colon \Omega \to W$ such that

$$\Phi(\Omega \cap \mathfrak{g}) = W \cap G.$$

To prove that this condition holds for every $g \in G$ besides I is easy. Indeed, $L_g \colon G \to G$ is a diffeomorphism, so $L_g \circ \Phi \colon \Omega \to L_g(W)$ is a diffeomorphism such that

$$(L_g \circ \Phi)(\Omega \cap \mathfrak{g}) = L_g(W) \cap G,$$

which shows that Condition (2) of Theorem 3.6 also holds for any $g \in G$, and thus G is a manifold. □

It should be noted that the assumption that G is closed is crucial, as shown by the following example from Tapp [111].

Pick any irrational multiple λ of 2π, and define

$$G = \left\{ g_t = \begin{pmatrix} e^{ti} & 0 \\ 0 & e^{\lambda t i} \end{pmatrix} \;\middle|\; t \in \mathbb{R} \right\}.$$

It is clear that G is a subgroup of $\mathbf{GL}(2, \mathbb{C})$. We leave it as an exercise to prove that the map $\varphi \colon t \mapsto g_t$ is a continuous isomorphism of $(\mathbb{R}, +)$ onto G, but that φ^{-1} is not continuous. Geometrically, φ is a curve embedded in $\mathbb{R}^4$ (by viewing $\mathbb{C}^2$ as $\mathbb{R}^4$). It is easy to check that $\mathfrak{g}$ (as defined in Proposition 3.9) is the one dimensional vector space spanned by

$$W = \begin{pmatrix} i & 0 \\ 0 & \lambda i \end{pmatrix},$$

and that $e^{tW} = g_t$ for all $t \in \mathbb{R}$. For every $r > 0$ ($r \in \mathbb{R}$), we leave it as an exercise to prove that

$$\exp(\{tW \mid t \in (-r, r)\}) = \{g_t \mid t \in (-r, r)\}$$

is **not** a neighborhood of I in G. The problem is that there are elements of G of the form $g_{2\pi n}$ for some large n that are arbitrarily close to I, so they are exponential images of very short vectors in $\mathrm{M}_2(\mathbb{C})$, but they are exponential images only of very long vectors in $\mathfrak{g}$. The reader should prove that the closure of the group G is the group

$$\overline{G} = \left\{ \begin{pmatrix} e^{ti} & 0 \\ 0 & e^{si} \end{pmatrix} \;\middle|\; t, s \in \mathbb{R} \right\},$$

and that G is dense in $\overline{G}$. Geometrically, G is a curve in $\mathbb{R}^4$ and $\overline{G}$ is the product of two circles, that is, a torus (in $\mathbb{R}^4$). Due to the the irrationality of λ, the curve G winds around the torus and forms a dense subset.

With the help of Theorem 3.8 it is now very easy to prove that $\mathbf{SL}(n)$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, $\mathbf{SL}(n,\mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$ are Lie groups and to figure out what are their Lie algebras. (Of course, $\mathbf{GL}(n,\mathbb{R})$ is a Lie group, as we already know.) It suffices to show that these subgroups of $\mathbf{GL}(n,\mathbb{R})$ ($\mathbf{GL}(2n,\mathbb{R})$ in the case of $\mathbf{SL}(n,\mathbb{C})$, $\mathbf{U}(n)$, and $\mathbf{SU}(n)$) are closed, which is easy to show since these groups are zero sets of simple continuous functions. For example, $\mathbf{SL}(n)$ is the zero set of the function $A \mapsto \det(A) - 1$, $\mathbf{O}(n)$ is the zero set of the function $R \mapsto R^\top R - I$, $\mathbf{SO}(n) = \mathbf{SL}(n) \cap \mathbf{O}(n)$, *etc.*

For example, if $G = \mathbf{GL}(n,\mathbb{R})$, as e^{tA} is invertible for *every* matrix $A \in \mathrm{M}_n(\mathbb{R})$, we deduce that the Lie algebra $\mathfrak{gl}(n,\mathbb{R})$ of $\mathbf{GL}(n,\mathbb{R})$ is equal to $\mathrm{M}_n(\mathbb{R})$. We also claim that the Lie algebra $\mathfrak{sl}(n,\mathbb{R})$ of $\mathbf{SL}(n,\mathbb{R})$ is the set of all matrices with zero trace. Indeed, $\mathfrak{sl}(n,\mathbb{R})$ is the subalgebra of $\mathfrak{gl}(n,\mathbb{R})$ consisting of all matrices $X \in \mathfrak{gl}(n,\mathbb{R})$ such that

$$\det(e^{tX}) = 1$$

for all $t \in \mathbb{R}$, and because $\det(e^{tX}) = e^{\mathrm{tr}(tX)}$, for $t = 1$, we get $\mathrm{tr}(X) = 0$, as claimed.

We can also prove that $\mathbf{SE}(n)$ is a Lie group as follows. Recall that we can view every element of $\mathbf{SE}(n)$ as a real $(n+1) \times (n+1)$ matrix

$$\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$$

where $R \in \mathbf{SO}(n)$ and $U \in \mathbb{R}^n$. In fact, such matrices belong to $\mathbf{SL}(n+1)$. This embedding of $\mathbf{SE}(n)$ into $\mathbf{SL}(n+1)$ is a group homomorphism, since the group operation on $\mathbf{SE}(n)$ corresponds to multiplication in $\mathbf{SL}(n+1)$:

$$\begin{pmatrix} RS & RV + U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S & V \\ 0 & 1 \end{pmatrix}.$$

Note that the inverse of $\begin{pmatrix} R & U \\ 0 & 1 \end{pmatrix}$ is given by

$$\begin{pmatrix} R^{-1} & -R^{-1}U \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} R^\top & -R^\top U \\ 0 & 1 \end{pmatrix}.$$

It is easy to show that $\mathbf{SE}(n)$ is a closed subgroup of $\mathbf{GL}(n+1,\mathbb{R})$ (because $\mathbf{SO}(n)$ and $\mathbb{R}^n$ are closed). Also note that the embedding shows that, as a manifold, $\mathbf{SE}(n)$ is diffeomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ (given a manifold M_1 of dimension m_1 and a manifold M_2 of dimension m_2, the product $M_1 \times M_2$ can be given the structure of a manifold of dimension $m_1 + m_2$ in a natural way). Thus, $\mathbf{SE}(n)$ is a Lie group with underlying manifold $\mathbf{SO}(n) \times \mathbb{R}^n$, and in fact, a closed subgroup of $\mathbf{SL}(n+1)$.

Even though $\mathbf{SE}(n)$ is diffeomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ as a manifold, it is *not* isomorphic to $\mathbf{SO}(n) \times \mathbb{R}^n$ as a group, because the group multiplication on $\mathbf{SE}(n)$ is not the multiplication on $\mathbf{SO}(n) \times \mathbb{R}^n$. Instead, $\mathbf{SE}(n)$ is a *semidirect product* of $\mathbf{SO}(n)$ by $\mathbb{R}^n$; see Section 18.5 or Gallier [48] (Chapter 2, Problem 2.19).

An application of Theorem 3.8 shows that the Lie algebra of $\mathbf{SE}(n)$, $\mathfrak{se}(n)$, is as described in Section 1.6; is easily determined as the subalgebra of $\mathfrak{sl}(n+1)$ consisting of all matrices of the form

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix}$$

where $B \in \mathfrak{so}(n)$ and $U \in \mathbb{R}^n$. Thus, $\mathfrak{se}(n)$ has dimension $n(n+1)/2$. The Lie bracket is given by

$$\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix}\begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} C & V \\ 0 & 0 \end{pmatrix}\begin{pmatrix} B & U \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} BC - CB & BV - CU \\ 0 & 0 \end{pmatrix}.$$

Returning to Theorem 3.8, the vector space $\mathfrak{g}$ is called the *Lie algebra* of the Lie group G. Lie algebras are defined as follows.

Definition 3.7. A *(real) Lie algebra* $\mathcal{A}$ is a real vector space together with a bilinear map $[\cdot, \cdot] \colon \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ called the *Lie bracket* on $\mathcal{A}$ such that the following two identities hold for all $a, b, c \in \mathcal{A}$:

$$[a,\, a] = 0,$$

and the so-called *Jacobi identity*

$$[a,\, [b,\, c]] + [c,\, [a,\, b]] + [b,\, [c,\, a]] = 0.$$

By using the Jacobi identity, it is readily verified that $[b, a] = -[a, b]$.

In view of Theorem 3.8, the vector space $\mathfrak{g} = T_I G$ associated with a Lie group G is indeed a Lie algebra. Furthermore, the exponential map $\exp \colon \mathfrak{g} \to G$ is well-defined. In general, exp is neither injective nor surjective, as we observed earlier. Theorem 3.8 also provides a kind of recipe for "computing" the Lie algebra $\mathfrak{g} = T_I G$ of a Lie group G. Indeed, $\mathfrak{g}$ is the tangent space to G at I, and thus we can use curves to compute tangent vectors. Actually, for every $X \in T_I G$, the map

$$\gamma_X \colon t \mapsto e^{tX}$$

is a smooth curve in G, and it is easily shown that $\gamma_X'(0) = X$. Thus, we can use these curves. As an illustration, we show that the Lie algebras of $\mathbf{SL}(n)$ and $\mathbf{SO}(n)$ are the matrices with null trace and the skew symmetric matrices.

Let $t \mapsto R(t)$ be a smooth curve in $\mathbf{SL}(n)$ such that $R(0) = I$. We have $\det(R(t)) = 1$ for all $t \in (-\epsilon, \epsilon)$. Using the chain rule, we can compute the derivative of the function

$$t \mapsto \det(R(t))$$

at $t = 0$, and since $\det(R(t)) = 1$ we get

$$\det{}_I'(R'(0)) = 0.$$

We leave it as an exercise for the reader to prove that

$$\det{}'_I(X) = \mathrm{tr}(X),$$

and thus $\mathrm{tr}(R'(0)) = 0$, which says that the tangent vector $X = R'(0)$ has null trace. Clearly, $\mathfrak{sl}(n, \mathbb{R})$ has dimension $n^2 - 1$.

Let $t \mapsto R(t)$ be a smooth curve in $\mathbf{SO}(n)$ such that $R(0) = I$. Since each $R(t)$ is orthogonal, we have

$$R(t)\, R(t)^\top = I$$

for all $t \in (-\epsilon, \epsilon)$. By using the product rule and taking the derivative at $t = 0$, we get

$$R'(0)\, R(0)^\top + R(0)\, R'(0)^\top = 0,$$

but since $R(0) = I = R(0)^\top$, we get

$$R'(0) + R'(0)^\top = 0,$$

which says that the tangent vector $X = R'(0)$ is skew symmetric. Since the diagonal elements of a skew symmetric matrix are null, the trace is automatically null, and the condition $\det(R) = 1$ yields nothing new. This shows that $\mathfrak{o}(n) = \mathfrak{so}(n)$. It is easily shown that $\mathfrak{so}(n)$ has dimension $n(n-1)/2$.

By appropriately adjusting the above methods, we readily calculate $\mathfrak{gl}(n, \mathbb{C})$, $\mathfrak{sl}(n, \mathbb{C})$, $\mathfrak{u}(n)$, and $\mathfrak{su}(n)$, confirming the claims of Section 1.4. It is easy to show that $\mathfrak{gl}(n, \mathbb{C})$ has dimension $2n^2$, $\mathfrak{sl}(n, \mathbb{C})$ has dimension $2(n^2 - 1)$, $\mathfrak{u}(n)$ has dimension n^2, and $\mathfrak{su}(n)$ has dimension $n^2 - 1$.

As a concrete example, the Lie algebra $\mathfrak{so}(3)$ of $\mathbf{SO}(3)$ is the real vector space consisting of all 3×3 real skew symmetric matrices. Every such matrix is of the form

$$\begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}$$

where $b, c, d \in \mathbb{R}$. The Lie bracket $[A, B]$ in $\mathfrak{so}(3)$ is also given by the usual commutator, $[A, B] = AB - BA$.

Let × represent the cross product of two vectors in $\mathbb{R}^3$ where for $u = (u_1, u_2, u_3)$ and $v = (v_1, v_2, v_3)$, we have

$$u \times v = -v \times u = (u_2 v_3 - u_3 v_2, -u_1 v_3 + u_3 v_1, u_1 v_2 - u_2 v_1).$$

It is easily checked that the vector space $\mathbb{R}^3$ is a Lie algebra if we define the Lie bracket on $\mathbb{R}^3$ as the usual cross product $u \times v$ of vectors. We can define an isomorphism of Lie algebras $\psi \colon (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ by the formula

$$\psi(b, c, d) = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix}.$$

A basic algebraic computation verifies that

$$\psi(u \times v) = [\psi(u),\, \psi(v)].$$

It is also verified that for any two vectors $u = (b, c, d)$ and $v = (b', c', d')$ in $\mathbb{R}^3$

$$\psi(u)(v) = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix} \begin{pmatrix} b' \\ c' \\ d' \end{pmatrix} = \begin{pmatrix} -dc' + cd' \\ db' - bd' \\ -cb' + bc' \end{pmatrix} = u \times v.$$

In robotics and in computer vision, $\psi(u)$ is often denoted by $u_\times$.

The exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is given by Rodrigues's formula (see Proposition 1.7):

$$e^A = \cos\theta\, I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} B,$$

or equivalently by

$$e^A = I_3 + \frac{\sin\theta}{\theta} A + \frac{(1 - \cos\theta)}{\theta^2} A^2$$

if $\theta \neq 0$, where

$$A = \begin{pmatrix} 0 & -d & c \\ d & 0 & -b \\ -c & b & 0 \end{pmatrix},$$

$\theta = \sqrt{b^2 + c^2 + d^2}$, $B = A^2 + \theta^2 I_3$, and with $e^0 = I_3$.

For another concrete example, the Lie algebra $\mathfrak{su}(2)$ of $\mathbf{SU}(2)$ (or S^3) is the real vector space consisting of all 2×2 (complex) skew Hermitian matrices of null trace. Every such matrix is of the form

$$i(d\sigma_1 + c\sigma_2 + b\sigma_3) = \begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix},$$

where $b, c, d \in \mathbb{R}$, and $\sigma_1, \sigma_2, \sigma_3$ are the Pauli spin matrices

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and thus the matrices $i\sigma_1, i\sigma_2, i\sigma_3$ form a basis of the Lie algebra $\mathfrak{su}(2)$. The Lie bracket $[A, B]$ in $\mathfrak{su}(2)$ is given by the usual commutator, $[A, B] = AB - BA$.

Let × represent the cross product of two vectors in $\mathbb{R}^3$. Then we can define an isomorphism of Lie algebras $\varphi\colon (\mathbb{R}^3, \times) \to \mathfrak{su}(2)$ by the formula

$$\varphi(b, c, d) = \frac{i}{2}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{2}\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}.$$

A tedious but basic algebraic computation verifies that

$$\varphi(u \times v) = [\varphi(u),\, \varphi(v)].$$

Returning to $\mathfrak{su}(2)$, letting $\theta = \sqrt{b^2 + c^2 + d^2}$, we can write

$$d\sigma_1 + c\sigma_2 + b\sigma_3 = \begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix} = \theta A,$$

where

$$A = \frac{1}{\theta}(d\sigma_1 + c\sigma_2 + b\sigma_3) = \frac{1}{\theta}\begin{pmatrix} b & -ic + d \\ ic + d & -b \end{pmatrix},$$

so that $A^2 = I$, and it can be shown that the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is given by

$$\exp(i\theta A) = \cos\theta\, I + i \sin\theta\, A.$$

In view of the isomorphism $\varphi\colon (\mathbb{R}^3, \times) \to \mathfrak{su}(2)$, where

$$\varphi(b, c, d) = \frac{1}{2}\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = i\frac{\theta}{2}A,$$

the exponential map can be viewed as a map $\exp\colon (\mathbb{R}^3, \times) \to \mathbf{SU}(2)$ given by the formula

$$\exp(\theta v) = \left[\cos\frac{\theta}{2},\, \sin\frac{\theta}{2}\, v\right],$$

for every vector θv, where v is a unit vector in $\mathbb{R}^3$ and $\theta \in \mathbb{R}$. Recall that $[a, (b, c, d)]$ is another way of denoting the quaternion $a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$; see Section 5.3 and Problem 18.16 for the definition of the quaternions and their properties. In this form, $\exp(\theta v)$ is a unit quaternion corresponding to a rotation of axis v and angle θ.

3.3 Homomorphisms of Linear Lie groups and Lie Algebras

In this section we will discuss the relationship between homomorphisms of Lie groups and homomorphisms of Lie algebras. But in order to do so, we first need to explain what is meant by a smooth map between manifolds.

Definition 3.8. Let M_1 (m_1-dimensional) and M_2 (m_2-dimensional) be manifolds in $\mathbb{R}^N$. A function $f\colon M_1 \to M_2$ is *smooth* if for every $p \in M_1$ there are parameterizations $\varphi\colon \Omega_1 \to U_1$ of M_1 at p and $\psi\colon \Omega_2 \to U_2$ of M_2 at $f(p)$ such that $f(U_1) \subseteq U_2$ and

$$\psi^{-1} \circ f \circ \varphi\colon \Omega_1 \to \mathbb{R}^{m_2}$$

is smooth; see Figure 3.13.

Using Lemma 3.2, it is easily shown that Definition 3.8 does not depend on the choice of the parametrizations $\varphi\colon \Omega_1 \to U_1$ and $\psi\colon \Omega_2 \to U_2$. A smooth map f between manifolds is a *smooth diffeomorphism* if f is bijective and both f and f^{-1} are smooth maps.

We now define the derivative of a smooth map between manifolds.

Definition 3.9. Let M_1 (m_1-dimensional) and M_2 (m_2-dimensional) be manifolds in $\mathbb{R}^N$. For any smooth function $f\colon M_1 \to M_2$ and any $p \in M_1$, the function $f'_p\colon T_pM_1 \to T_{f(p)}M_2$, called the *tangent map of f at p, or derivative of f at p, or differential of f at p*, is defined as follows: For every $v \in T_pM_1$ and every smooth curve $\gamma\colon I \to M_1$ such that $\gamma(0) = p$ and $\gamma'(0) = v$,

$$f'_p(v) = (f \circ \gamma)'(0).$$

See Figure 3.14.

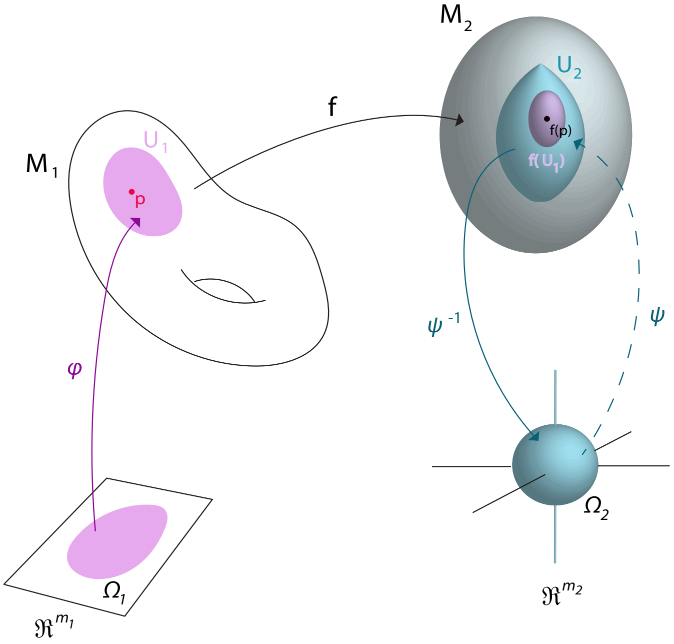


Figure 3.13: An illustration of a smooth map from the torus, M_1, to the solid ellipsoid M_2. The pink patch on M_1 is mapped into interior pink ellipsoid of M_2.

The map f'_p is also denoted by df_p or T_pf. Doing a few calculations involving the facts that

$$f \circ \gamma = (f \circ \varphi) \circ (\varphi^{-1} \circ \gamma) \quad \text{and} \quad \gamma = \varphi \circ (\varphi^{-1} \circ \gamma)$$

and using Lemma 3.2, it is not hard to show that $f'_p(v)$ does not depend on the choice of the curve γ. It is easily shown that f'_p is a linear map.

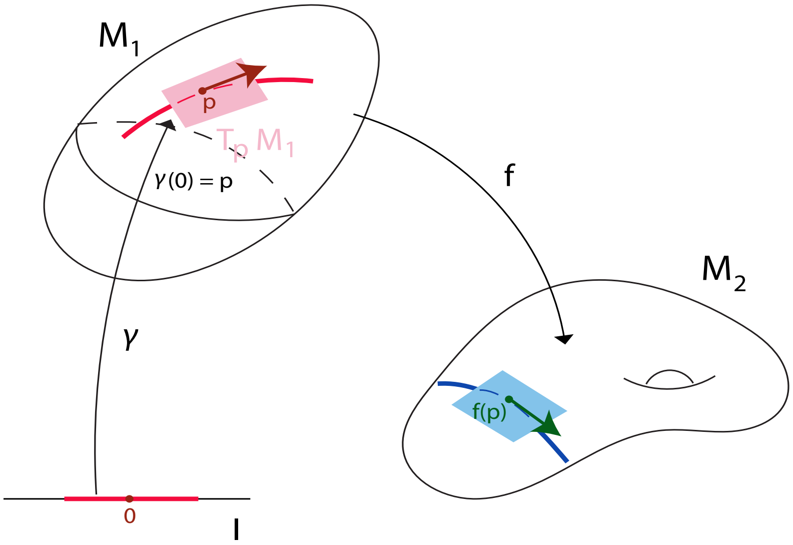


Figure 3.14: An illustration of the tangent map from T_pM_1 to $T_{f(p)}M_2$.

Given a linear Lie group G, since L_a and R_a are diffeomorphisms for every $a \in G$, the maps $d(L_a)_I \colon \mathfrak{g} \to T_aG$ and $d(R_a)_I \colon \mathfrak{g} \to T_aG$ are linear isomorphisms between the Lie algebra $\mathfrak{g}$ and the tangent space T_aG to G at a. Since G is a linear group, both L_a and R_a are linear, we have $(dL_a)_b = L_a$ and $(dR_a)_b = R_a$ for all $b \in G$, and so

$$T_aG = a\mathfrak{g} = \{aX \mid X \in \mathfrak{g}\} = \{Xa \mid X \in \mathfrak{g}\} = \mathfrak{g}a.$$

Finally we define homomorphisms of Lie groups and Lie algebras and see how they are related.

Definition 3.10. Given two Lie groups G_1 and G_2, a *homomorphism (or map) of Lie groups* is a function $f \colon G_1 \to G_2$ that is a homomorphism of groups and a smooth map (between the manifolds G_1 and G_2). Given two Lie algebras $\mathcal{A}_1$ and $\mathcal{A}_2$, a *homomorphism (or map) of Lie algebras* is a function $f \colon \mathcal{A}_1 \to \mathcal{A}_2$ that is a linear map between the vector spaces $\mathcal{A}_1$ and $\mathcal{A}_2$ and that preserves Lie brackets, i.e.,

$$f([A, B]) = [f(A), f(B)]$$

for all $A, B \in \mathcal{A}_1$.

An *isomorphism of Lie groups* is a bijective function f such that both f and f^{-1} are homomorphisms of Lie groups, and an *isomorphism of Lie algebras* is a bijective function f such that both f and f^{-1} are maps of Lie algebras. If $f \colon G_1 \to G_2$ is a homomorphism of linear Lie groups, then $f'_I\colon \mathfrak{g}_1 \to \mathfrak{g}_2$ is a homomorphism of Lie algebras, but in order to prove this, we need the adjoint representation Ad, so we postpone the proof.

The notion of a one-parameter group plays a crucial role in Lie group theory.

Definition 3.11. A smooth homomorphism $h\colon (\mathbb{R}, +) \to G$ from the additive group $\mathbb{R}$ to a Lie group G is called a *one-parameter group* in G.

All one-parameter groups of a linear Lie group can be determined explicitly.

Proposition 3.12. *Let G be any linear Lie group.*

1. *For every $X \in \mathfrak{g}$, the map $h(t) = e^{tX}$ is a one-parameter group in G.*

2. *Every one-parameter group $h\colon \mathbb{R} \to G$ is of the form $h(t) = e^{tZ}$, with $Z = h'(0)$.*

In summary, for every $Z \in \mathfrak{g}$, there is a unique one-parameter group h such that $h'(0) = Z$ given by $h(t) = e^{Zt}$.

Proof. The proof of (1) is easy and left as an exercise. To prove (2), since h is a homomorphism, for all $s, t \in \mathbb{R}$, we have

$$h(s + t) = h(s)h(t).$$

Taking the derivative with respect to s for $s = 0$ and holding t constant, the product rule implies that

$$h'(t) = h'(0)h(t).$$

If we write $Z = h'(0)$ we we have

$$h'(t) = Zh(t) = X_Z(h(t)) \quad \text{for all } t \in \mathbb{R}.$$

This means that $h(t)$ is an integral curve for all t passing through I for the linear vector field X_Z, and by Proposition 11.25, it must be equal to e^{tZ}. □

The exponential map is natural in the following sense:

Proposition 3.13. *Given any two linear Lie groups G and H, for every Lie group homomorphism $f\colon G \to H$, the following diagram commutes:*

$$\begin{tikzcd}
G \arrow[r, "f"] & H \\
\mathfrak{g} \arrow[u, "\exp"] \arrow[r, "df_I"'] & \mathfrak{h} \arrow[u, "\exp"']
\end{tikzcd}$$

Proof. Observe that for every $v \in \mathfrak{g}$, the map $h\colon t \mapsto f(e^{tv})$ is a homomorphism from $(\mathbb{R}, +)$ to G such that $h'(0) = df_I(v)$. On the other hand, by Proposition 3.12 the map $t \mapsto e^{tdf_I(v)}$ is the unique one-parameter group whose tangent vector at 0 is $df_I(v)$, so $f(e^v) = e^{df_I(v)}$. □

Alert readers must have noticed that in Theorem 3.8 we only defined the Lie algebra of a linear group. In the more general case, we can still define the Lie algebra $\mathfrak{g}$ of a Lie group G as the tangent space $T_I G$ at the identity I. The tangent space $\mathfrak{g} = T_I G$ is a vector space, but we need to define the Lie bracket. This can be done in several ways. We explain briefly how this can be done in terms of so-called adjoint representations. This has the advantage of not requiring the definition of left-invariant vector fields, but it is still a little bizarre!

Given a Lie group G, for every $a \in G$ we define *left translation* as the map $L_a \colon G \to G$ such that $L_a(b) = ab$ for all $b \in G$, and *right translation* as the map $R_a \colon G \to G$ such that $R_a(b) = ba$ for all $b \in G$. The maps L_a and R_a are diffeomorphisms, and their derivatives play an important role.

The inner automorphisms $\mathbf{Ad}_a \colon G \to G$ defined by $\mathbf{Ad}_a = R_{a^{-1}} \circ L_a$ $(= R_{a^{-1}} L_a)$ also play an important role. Note that

$$\mathbf{Ad}_a(b) = aba^{-1}.$$

The derivative

$$(\mathbf{Ad}_a)'_I \colon T_I G \to T_I G$$

of $\mathbf{Ad}_a$ at I is an isomorphism of Lie algebras, and since $T_I G = \mathfrak{g}$, if we denote $(\mathbf{Ad}_a)'_I$ by Ad_a, we get a map

$$\mathrm{Ad}_a \colon \mathfrak{g} \to \mathfrak{g}.$$

The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad} \colon G \to \mathbf{GL}(\mathfrak{g}),$$

called the *adjoint representation of* G (where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$).

In the case of a linear group, we have

$$\mathrm{Ad}(a)(X) = \mathrm{Ad}_a(X) = aXa^{-1}$$

for all $a \in G$ and all $X \in \mathfrak{g}$. Indeed, for any $X \in \mathfrak{g}$, the curve $\gamma(t) = e^{tX}$ is a curve in G such that $\gamma(0) = I$ and $\gamma'(0) = X$. Then by the definition of the tangent map, we have

$$\begin{aligned} d(\mathbf{Ad}_a)_I(X) &= (\mathbf{Ad}_a(\gamma(t)))'(0) \\ &= (ae^{tX}a^{-1})'(0) \\ &= aXa^{-1}. \end{aligned}$$

We are now almost ready to prove that if $f \colon G_1 \to G_2$ is a homomorphism of linear Lie groups, then $f'_I \colon \mathfrak{g}_1 \to \mathfrak{g}_2$ is a homomorphism of Lie algebras. What we need is to express the Lie bracket $[A, B]$ in terms of the derivative of an expression involving the adjoint representation Ad. For any $A, B \in \mathfrak{g}$, we have

$$\big(\mathrm{Ad}_{e^{tA}}(B)\big)'(0) = (e^{tA}Be^{-tA})'(0) = AB - BA = [A, B].$$

Proposition 3.14. *If $f\colon G_1 \to G_2$ is a homomorphism of linear Lie groups, then the linear map $df_I\colon \mathfrak{g}_1 \to \mathfrak{g}_2$ satisfies the equation*

$$df_I(\mathrm{Ad}_a(X)) = \mathrm{Ad}_{f(a)}(df_I(X)), \quad \textit{for all } a \in G \textit{ and all } X \in \mathfrak{g}_1,$$

that is, the following diagram commutes

$$\begin{tikzcd}
\mathfrak{g}_1 \arrow[r, "df_I"] \arrow[d, "\mathrm{Ad}_a"'] & \mathfrak{g}_2 \arrow[d, "\mathrm{Ad}_{f(a)}"] \\
\mathfrak{g}_1 \arrow[r, "df_I"'] & \mathfrak{g}_2
\end{tikzcd}$$

Furthermore, df_I is a homomorphism of Lie algebras.

Proof. Since f is a group homomorphism, for all $X \in \mathfrak{g}_1$, we have

$$f(ae^{tX}a^{-1}) = f(a)f(e^{tX})f(a^{-1}) = f(a)f(e^{tX})f(a)^{-1}.$$

The curve α given by $\alpha(t) = ae^{tX}a^{-1}$ passes through I and $\alpha'(0) = aXa^{-1} = \mathrm{Ad}_a(X)$, so we have

$$\begin{aligned}
df_I(\mathrm{Ad}_a(X)) &= (f(\alpha(t)))'(0) \\
&= (f(ae^{tX}a^{-1}))'(0) \\
&= (f(a)f(e^{tX})f(a)^{-1})'(0) \\
&= \mathrm{Ad}_{f(a)}(df_I(X)),
\end{aligned}$$

as claimed. Now pick any $X, Y \in \mathfrak{g}_1$. The plan is to use the identity we just proved with $a = e^{tX}$ and $X = Y$, namely

$$df_I(\mathrm{Ad}_{e^{tX}}(Y)) = \mathrm{Ad}_{f(e^{tX})}(df_I(Y)), \tag{$*$}$$

and to take the derivative of both sides for $t = 0$. We make use of the fact that since $df_I\colon \mathfrak{g} \to \mathfrak{g}$ is linear, for any $Z \in \mathfrak{g}_1$, we have

$$d(df_I)_Z = df_I.$$

Then, if we write $\beta(t) = \mathrm{Ad}_{e^{tX}}Y$, we have $df_I(\mathrm{Ad}_{e^{tX}}Y) = df_I(\beta(t))$, and as df_I is linear, the derivative of the left hand side of $(*)$ is

$$(df_I(\beta(t)))'(0) = d(df_I)_{\beta(0)}(\beta'(0)) = df_I(\beta'(0)).$$

On the other hand, by the fact proven just before stating Proposition 3.14,

$$\beta'(0) = (\mathrm{Ad}_{e^{tX}}Y)'(0) = [X, Y],$$

so the the derivative of the left hand side of $(*)$ is equal to $df_I(\beta'(0)) = df_I([X, Y])$. When we take the derivative of the right hand side, since f is a group homomorphism, we get

$$\begin{aligned}(\mathrm{Ad}_{f(e^{tX})}(df_I(Y)))'(0) &= (f(e^{tX})df_I(Y)(f(e^{tX}))^{-1})'(0) \\ &= (f(e^{tX})df_I(Y)f(e^{-tX}))'(0) = [df_I(X), df_I(Y)],\end{aligned}$$

and we conclude that

$$df_I([X, Y]) = [df_I(X), df_I(Y)];$$

that is, f_I is a Lie algebra homomorphism. □

If some additional assumptions are made about G_1 and G_2 (for example, connected, simply connected), it can be shown that f is pretty much determined by f'_I.

The derivative

$$\mathrm{Ad}'_I \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ at I is map of Lie algebras, and if we denote Ad'_I by ad, it is a map

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

called the *adjoint representation of* $\mathfrak{g}$. (Recall that Theorem 3.8 immediately implies that the Lie algebra $\mathfrak{gl}(\mathfrak{g})$ of $\mathbf{GL}(\mathfrak{g})$ is the vector space $\mathrm{Hom}(\mathfrak{g}, \mathfrak{g})$ of all linear maps on $\mathfrak{g}$).

In the case of linear Lie groups, if we apply Proposition 3.13 to $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$, we obtain the equation

$$\mathrm{Ad}_{e^A} = e^{\mathrm{ad}_A} \quad \text{for all } A \in \mathfrak{g},$$

or equivalently

$$\begin{tikzcd} G \arrow[r, "\mathrm{Ad}"] & \mathbf{GL}(\mathfrak{g}) \\ \mathfrak{g} \arrow[u, "\exp"] \arrow[r, "\mathrm{ad}"'] & \mathfrak{gl}(\mathfrak{g}) \arrow[u, "\exp"'] \end{tikzcd},$$

which is a generalization of the identity of Proposition 2.1.

In the case of a linear group we have

$$\mathrm{ad}(A)(B) = [A,\, B]$$

for all $A, B \in \mathfrak{g}$. This can be shown as follows.

Proof. For any $A, B \in \mathfrak{g}$, the curve $\gamma(t) = e^{tA}$ is a curve in G passing through I and such that $\gamma'(0) = A$, so we have

$$\begin{aligned}\mathrm{ad}_A(B) &= ((\mathrm{Ad}_{e^{tA}})'(0))(B) \\ &= \big((\mathrm{Ad}_{e^{tA}})(B)\big)'(0) \\ &= (e^{tA}Be^{-tA})'(0) \\ &= AB - BA,\end{aligned}$$

which proves our result. □

Remark: The equation

$$((\mathrm{Ad}_{e^{tA}})'(0))(B) = \big((\mathrm{Ad}_{e^{tA}})(B)\big)'(0)$$

requires some justification. Define $\mathrm{eval}_B\colon \mathrm{Hom}(\mathfrak{g},\mathfrak{g}) \to \mathfrak{g}$ by $\mathrm{eval}_B(f) = f(B)$ for any $f \in \mathrm{Hom}(\mathfrak{g},\mathfrak{g})$. Note that eval_B is a linear map, and hence $d(\mathrm{eval}_B)_f = \mathrm{eval}_B$ for all $f \in \mathrm{Hom}(\mathfrak{g},\mathfrak{g})$. By definition $\mathrm{Ad}_{e^{tA}}(B) = \mathrm{eval}_B(\mathrm{Ad}_{e^{tA}})$, and an application of the chain rule implies that

$$\begin{aligned}\big((\mathrm{Ad}_{e^{tA}})(B)\big)'(0) &= \big(\mathrm{eval}_B(\mathrm{Ad}_{e^{tA}})\big)'(0) = d(\mathrm{eval}_B)_{\mathrm{Ad}_{e0}} \circ \big(\mathrm{Ad}_{e^{tA}}\big)'(0)\\ &= \mathrm{eval}_B(\mathrm{Ad}_{e^{tA}}\big)'(0) = ((\mathrm{Ad}_{e^{tA}})'(0))(B).\end{aligned}$$

Another proof of the fact that $\mathrm{ad}_A(B) = [A, B]$ can be given using Propositions 2.1 and 3.13. To avoid confusion, let us temporarily write $ad_A(B) = [A, B]$ to distinguish it from $\mathrm{ad}_A(B) = (d(\mathrm{Ad})_I(A))(B)$. Both ad and ad are linear. For any fixed $t \in \mathbb{R}$, by Proposition 2.1 we have

$$\mathrm{Ad}_{e^{tA}} = e^{ad_{tA}} = e^{tad_A},$$

and by Proposition 3.13 applied to Ad, we have

$$\mathrm{Ad}_{e^{tA}} = e^{\mathrm{ad}_{tA}} = e^{t\mathrm{ad}_A}.$$

It follows that

$$e^{tad_A} = e^{t\mathrm{ad}_A} \quad \text{for all } t \in \mathbb{R},$$

and by taking the derivative at $t = 0$, we get $ad_A = \mathrm{ad}_A$.

One can also check that the Jacobi identity on $\mathfrak{g}$ is equivalent to the fact that ad preserves Lie brackets, i.e., ad is a map of Lie algebras:

$$\mathrm{ad}([A,\, B]) = [\mathrm{ad}(A),\, \mathrm{ad}(B)]$$

for all $A, B \in \mathfrak{g}$ (where on the right, the Lie bracket is the commutator of linear maps on $\mathfrak{g}$). Thus we recover the Lie bracket from ad.

This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group). We define the Lie bracket on $\mathfrak{g}$ as

$$[A,\, B] = \mathrm{ad}(A)(B).$$

To be complete, we have to define the exponential map $\exp\colon \mathfrak{g} \to G$ for a general Lie group. For this we need to introduce some left-invariant vector fields induced by the derivatives of the left translations, and integral curves associated with such vector fields. We will do this in Chapter 18 but for this we will need a deeper study of manifolds (see Chapters 7 and 9).

We conclude this section by computing explicitly the adjoint representations ad of $\mathfrak{so}(3)$ and Ad of $\mathbf{SO}(3)$. Recall that for every $X \in \mathfrak{so}(3)$, ad_X is a linear map $\mathrm{ad}_X\colon \mathfrak{so}(3) \to \mathfrak{so}(3)$. Also, for every $R \in \mathbf{SO}(3)$, the map $\mathrm{Ad}_R\colon \mathfrak{so}(3) \to \mathfrak{so}(3)$ is an invertible linear map of $\mathfrak{so}(3)$. As we saw at the end of Section 3.2, $\mathfrak{so}(3)$ is isomorphic to $(\mathbb{R}^3, \times)$, where $\times$ is the cross-product on $\mathbb{R}^3$, *via* the isomorphism $\psi\colon (\mathbb{R}^3, \times) \to \mathfrak{so}(3)$ given by the formula

$$\psi(a,b,c) = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}.$$

In robotics and in computer vision, $\psi(u)$ is often denoted by $u_\times$. Recall that

$$\psi(u)v = u_\times v = u \times v \quad \text{for all } u, v \in \mathbb{R}^3.$$

The image of the canonical basis (e_1, e_2, e_3) of $\mathbb{R}^3$ is the following basis of $\mathfrak{so}(3)$:

$$\left(E_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix},\ E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix},\ E_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right).$$

Observe that

$$[E_1, E_2] = E_3, \quad [E_2, E_3] = E_1, \quad [E_3, E_1] = E_2.$$

Using the isomorphism ψ, we obtain an isomorphism Ψ between $\mathrm{Hom}(\mathfrak{so}(3), \mathfrak{so}(3))$ and $\mathrm{M}_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R})$ such that every linear map $f\colon \mathfrak{so}(3) \to \mathfrak{so}(3)$ corresponds to the matrix of the linear map

$$\Psi(f) = \psi^{-1} \circ f \circ \psi$$

in the basis (e_1, e_2, e_3). By restricting Ψ to $\mathbf{GL}(\mathfrak{so}(3))$, we obtain an isomorphism between $\mathbf{GL}(\mathfrak{so}(3))$ and $\mathbf{GL}(3, \mathbb{R})$. It turns out that if we use the basis (E_1, E_2, E_3) in $\mathfrak{so}(3)$, for every $X \in \mathfrak{so}(3)$, the matrix representing $\mathrm{ad}_X \in \mathrm{Hom}(\mathfrak{so}(3), \mathfrak{so}(3))$ is X itself, and for every $R \in \mathbf{SO}(3)$, the matrix representing $\mathrm{Ad}_R \in \mathbf{GL}(\mathfrak{so}(3))$ is R itself.

Proposition 3.15. *For all $X \in \mathfrak{so}(3)$ and all $R \in \mathbf{SO}(3)$, we have*

$$\Psi(\mathrm{ad}_X) = X, \qquad \Psi(\mathrm{Ad}_R) = R,$$

which means that $\Psi \circ \mathrm{ad}$ is the inclusion map from $\mathfrak{so}(3)$ to $\mathrm{M}_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R})$, and that $\Psi \circ \mathrm{Ad}$ is the inclusion map from $\mathbf{SO}(3)$ to $\mathbf{GL}(3, \mathbb{R})$. Equivalently, for all $u \in \mathbb{R}^3$, we have

$$\mathrm{ad}_X(\psi(u)) = \psi(Xu), \quad \mathrm{Ad}_R(\psi(u)) = \psi(Ru).$$

These equations can also be written as

$$[X, u_\times] = (Xu)_\times, \qquad R u_\times R^{-1} = (Ru)_\times.$$

Proof. Since ad is linear, it suffices to prove the equation for the basis (E_1, E_2, E_3). For E_1, since $\psi(e_i) = E_i$, we have

$$\mathrm{ad}_{E_1}(\psi(e_i)) = [E_1, \psi(e_i)] = \begin{cases} 0 & \text{if } i = 1 \\ E_3 & \text{if } i = 2 \\ -E_2 & \text{if } i = 3. \end{cases}$$

Since

$$E_1 e_1 = 0, \quad E_1 e_2 = e_3, \quad E_1 e_3 = -e_2, \quad \psi(0) = 0, \quad \psi(e_3) = E_3, \quad \psi(e_2) = E_2,$$

we proved that

$$\mathrm{ad}_{E_1}(\psi(e_i)) = \psi(E_1 e_i), \quad i = 1, 2, 3.$$

Similarly, the reader should check that

$$\mathrm{ad}_{E_j}(\psi(e_i)) = \psi(E_j e_i), \quad j = 2, 3, \quad i = 1, 2, 3,$$

and so,

$$\mathrm{ad}_X(\psi(u)) = \psi(Xu) \quad \text{for all } X \in \mathfrak{so}(3) \text{ and all } u \in \mathbb{R}^3,$$

or equivalently

$$\psi^{-1}(\mathrm{ad}_X(\psi(u))) = X(u) \quad \text{for all } X \in \mathfrak{so}(3) \text{ and all } u \in \mathbb{R}^3;$$

that is, $\Psi \circ \mathrm{ad}$ is the inclusion map from $\mathfrak{so}(3)$ to $\mathrm{M}_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R})$.

Since every one-parameter group in $\mathbf{SO}(3)$ is of the form $t \mapsto e^{tX}$ for some $X \in \mathfrak{so}(3)$ and since $\Psi \circ \mathrm{ad}$ is the inclusion map from $\mathfrak{so}(3)$ to $\mathrm{M}_3(\mathbb{R}) = \mathfrak{gl}(3, \mathbb{R})$, the map $\Psi \circ \mathrm{Ad}$ maps every one-parameter group in $\mathbf{SO}(3)$ to itself in $\mathbf{GL}(3, \mathbb{R})$. Since the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective, every $R \in \mathbf{SO}(3)$ is of the form $R = e^X$ for some $X \in \mathfrak{so}(3)$, so R is contained in some one-parameter group, and thus R is mapped to itself by $\Psi \circ \mathrm{Ad}$. □

Readers who wish to learn more about Lie groups and Lie algebras should consult (more or less listed in order of difficulty) Tapp [111], Rossmann [98], Kosmann [70], Curtis [34], Sattinger and Weaver [102], Hall [56], and Marsden and Ratiu [77]. The excellent lecture notes by Carter, Segal, and Macdonald [29] constitute a very efficient (although somewhat terse) introduction to Lie algebras and Lie groups. Classics such as Weyl [118] and Chevalley [31] are definitely worth consulting, although the presentation and the terminology may seem a bit old fashioned. For more advanced texts, one may consult Abraham and Marsden [1], Warner [114], Sternberg [110], Bröcker and tom Dieck [24], and Knapp [68]. For those who read French, Mneimné and Testard [86] is very clear and quite thorough, and uses very little differential geometry, although it is more advanced than Curtis. Chapter 1, by Bryant, in Freed and Uhlenbeck [25] is also worth reading, but the pace is fast.

3.4 Problems

Problem 3.1. Recall that

$$S^2 = \left\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\right\}.$$

Let $N = (0, 0, 1)$ and $S = (0, 0, -1)$. Define two maps $\varphi_1 \colon \mathbb{R}^2 \to S^2 - \{N\}$ and $\varphi_2 \colon \mathbb{R}^2 \to S^2 - \{S\}$ as follows:

$$\varphi_1 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right)$$

$$\varphi_2 \colon (u, v) \mapsto \left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{1 - u^2 - v^2}{u^2 + v^2 + 1}\right).$$

(i) Prove that both of these maps are smooth homeomorphisms.

(ii) Prove that φ_1 and φ_2 are immersions for each point in their respective domains.

Problem 3.2. Show that the torus $T^2 = S_1 \times S_1$ is an embedded manifold in $\mathbb{R}^3$.

Problem 3.3. Prove Lemma 3.1.

Problem 3.4. (a) Consider the map $f \colon \mathbf{GL}(n, \mathbb{R}) \to \mathbb{R}$, given by

$$f(A) = \det(A).$$

Prove that $df(I)(B) = \mathrm{tr}(B)$, the trace of B, for any matrix B (here, I is the identity matrix). Then, prove that

$$df(A)(B) = \det(A)\mathrm{tr}(A^{-1}B),$$

where $A \in \mathbf{GL}(n, \mathbb{R})$.

(b) Use the map $A \mapsto \det(A) - 1$ to prove that $\mathbf{SL}(n, \mathbb{R})$ is a manifold of dimension $n^2 - 1$.

(c) Let J be the $(n+1) \times (n+1)$ diagonal matrix

$$J = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix}.$$

We denote by $\mathbf{SO}(n, 1)$ the group of real $(n+1) \times (n+1)$ matrices

$$\mathbf{SO}(n, 1) = \{A \in \mathbf{GL}(n+1, \mathbb{R}) \mid A^\top J A = J \quad \text{and} \quad \det(A) = 1\}.$$

Check that $\mathbf{SO}(n, 1)$ is indeed a group with the inverse of A given by $A^{-1} = JA^\top J$ (this is the *special Lorentz group*.) Consider the function $f \colon \mathbf{GL}^+(n+1) \to \mathbf{S}(n+1)$, given by

$$f(A) = A^\top J A - J,$$

where $\mathbf{S}(n+1)$ denotes the space of $(n+1)\times(n+1)$ symmetric matrices. Prove that

$$df(A)(H) = A^\top JH + H^\top JA$$

for any matrix, H. Prove that $df(A)$ is surjective for all $A \in \mathbf{SO}(n,1)$ and that $\mathbf{SO}(n,1)$ is a manifold of dimension $\frac{n(n+1)}{2}$.

Problem 3.5. Prove Proposition 3.7.

Problem 3.6. Recall that a matrix $B \in \mathrm{M}_n(\mathbb{R})$ is skew-symmetric if

$$B^\top = -B.$$

Check that the set $\mathfrak{so}(n)$ of skew-symmetric matrices is a vector space of dimension $n(n-1)/2$, and thus is isomorphic to $\mathbb{R}^{n(n-1)/2}$.

(a) Given a rotation matrix

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

where $0 < \theta < \pi$, prove that there is a skew symmetric matrix B such that

$$R = (I-B)(I+B)^{-1}.$$

(b) Prove that the eigenvalues of a skew-symmetric matrix are either 0 or pure imaginary (that is, of the form $i\mu$ for $\mu \in \mathbb{R}$.).

Let $C\colon \mathfrak{so}(n) \to \mathrm{M}_n(\mathbb{R})$ be the function given by

$$C(B) = (I-B)(I+B)^{-1}.$$

Prove that if B is skew-symmetric, then $I-B$ and $I+B$ are invertible, and so C is well-defined. Prove that

$$(I+B)(I-B) = (I-B)(I+B),$$

and that

$$(I+B)(I-B)^{-1} = (I-B)^{-1}(I+B).$$

Prove that

$$(C(B))^\top C(B) = I$$

and that

$$\det C(B) = +1,$$

so that $C(B)$ is a rotation matrix. Furthermore, show that $C(B)$ does not admit -1 as an eigenvalue.

(c) Let $\mathbf{SO}(n)$ be the group of $n \times n$ rotation matrices. Prove that the map

$$C\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$$

is bijective onto the subset of rotation matrices that do not admit -1 as an eigenvalue. Show that the inverse of this map is given by

$$B = (I + R)^{-1}(I - R) = (I - R)(I + R)^{-1},$$

where $R \in \mathbf{SO}(n)$ does not admit -1 as an eigenvalue. Check that C is a homeomorphism between $\mathfrak{so}(n)$ and $C(\mathfrak{so}(n))$.

(d) Use Problem 11.9 to prove that

$$dC(B)(A) = -[I + (I - B)(I + B)^{-1}]A(I + B)^{-1} = -2(I + B)^{-1}A(I + B)^{-1}.$$

Prove that $dC(B)$ is injective, for every skew-symmetric matrix B. Prove that C a parametrization of $\mathbf{SO}(n)$.

Problem 3.7. Recall from Problem 3.6, the Cayley parametrization of rotation matrices in $\mathbf{SO}(n)$ given by

$$C(B) = (I - B)(I + B)^{-1},$$

where B is any $n \times n$ skew symmetric matrix.

(a) Now, consider $n = 3$, i.e., $\mathbf{SO}(3)$. Let E_1, E_2 and E_3 be the rotations about the x-axis, y-axis, and z-axis, respectively, by the angle π, i.e.,

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad E_2 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad E_3 = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Prove that the four maps

$$\begin{aligned} B &\mapsto C(B) \\ B &\mapsto E_1C(B) \\ B &\mapsto E_2C(B) \\ B &\mapsto E_3C(B) \end{aligned}$$

where B is skew symmetric, are parametrizations of $\mathbf{SO}(3)$ and that the union of the images of C, E_1C, E_2C and E_3C covers $\mathbf{SO}(3)$, so that $\mathbf{SO}(3)$ is a manifold.

(b) Let A be *any* matrix (not necessarily invertible). Prove that there is some diagonal matrix, E, with entries $+1$ or -1, so that $EA + I$ is invertible.

(c) Prove that every rotation matrix, $A \in \mathbf{SO}(n)$, is of the form

$$A = E(I - B)(I + B)^{-1},$$

for some skew symmetric matrix, B, and some diagonal matrix, E, with entries $+1$ and -1, and where the number of -1 is even. Moreover, prove that every orthogonal matrix $A \in \mathbf{O}(n)$ is of the form

$$A = E(I - B)(I + B)^{-1},$$

for some skew symmetric matrix, B, and some diagonal matrix, E, with entries $+1$ and -1. The above provide parametrizations for $\mathbf{SO}(n)$ (resp. $\mathbf{O}(n)$) that show that $\mathbf{SO}(n)$ and $\mathbf{O}(n)$ are manifolds. However, observe that the number of these charts grows exponentially with n.

Problem 3.8. Consider the parametric surface given by

$$\begin{aligned} x(u,v) &= \frac{8uv}{(u^2+v^2+1)^2}, \\ y(u,v) &= \frac{4v(u^2+v^2-1)}{(u^2+v^2+1)^2}, \\ z(u,v) &= \frac{4(u^2-v^2)}{(u^2+v^2+1)^2}. \end{aligned}$$

The trace of this surface is called a *crosscap*. In order to plot this surface, make the change of variables

$$\begin{aligned} u &= \rho\cos\theta \\ v &= \rho\sin\theta. \end{aligned}$$

Prove that we obtain the parametric definition

$$\begin{aligned} x &= \frac{4\rho^2}{(\rho^2+1)^2}\,\sin 2\theta, \\ y &= \frac{4\rho(\rho^2-1)}{(\rho^2+1)^2}\,\sin\theta, \\ z &= \frac{4\rho^2}{(\rho^2+1)^2}\,\cos 2\theta. \end{aligned}$$

Show that the entire trace of the surface is obtained for $\rho \in [0, 1]$ and $\theta \in [-\pi, \pi]$.

Hint. What happens if you change ρ to $1/\rho$?

Plot the trace of the surface using the above parametrization. Show that there is a line of self-intersection along the portion of the z-axis corresponding to $0 \leq z \leq 1$. What can you say about the point corresponding to $\rho = 1$ and $\theta = 0$?

Plot the portion of the surface for $\rho \in [0, 1]$ and $\theta \in [0, \pi]$.

(b) Express the trigonometric functions in terms of $u = \tan(\theta/2)$, and letting $v = \rho$, show that we get

$$\begin{aligned} x &= \frac{16uv^2(1-u^2)}{(u^2+1)^2(v^2+1)^2}, \\ y &= \frac{8uv(u^2+1)(v^2-1)}{(u^2+1)^2(v^2+1)^2}, \\ z &= \frac{4v^2(u^4-6u^2+1)}{(u^2+1)^2(v^2+1)^2}. \end{aligned}$$

Problem 3.9. Consider the parametric surface given by

$$\begin{aligned} x(u,v) &= \frac{4v(u^2+v^2-1)}{(u^2+v^2+1)^2}, \\ y(u,v) &= \frac{4u(u^2+v^2-1)}{(u^2+v^2+1)^2}, \\ z(u,v) &= \frac{4(u^2-v^2)}{(u^2+v^2+1)^2}. \end{aligned}$$

The trace of this surface is called the *Steiner Roman surface*. In order to plot this surface, make the change of variables

$$\begin{aligned} u &= \rho\cos\theta \\ v &= \rho\sin\theta. \end{aligned}$$

Prove that we obtain the parametric definition

$$\begin{aligned} x &= \frac{4\rho(\rho^2-1)}{(\rho^2+1)^2}\sin\theta, \\ y &= \frac{4\rho(\rho^2-1)}{(\rho^2+1)^2}\cos\theta, \\ z &= \frac{4\rho^2}{(\rho^2+1)^2}\cos 2\theta. \end{aligned}$$

Show that the entire trace of the surface is obtained for $\rho \in [0,1]$ and $\theta \in [-\pi,\pi]$. Plot the trace of the surface using the above parametrization.

Plot the portion of the surface for $\rho \in [0,1]$ and $\theta \in [0,\pi]$.

Prove that this surface has five singular points.

(b) Express the trigonometric functions in terms of $u = \tan(\theta/2)$, and letting $v = \rho$, show that we get

$$x = \frac{8uv(u^2+1)(v^2-1)}{(u^2+1)^2(v^2+1)^2},$$
$$y = \frac{4v(1-u^4)(v^2-1)}{(u^2+1)^2(v^2+1)^2},$$
$$z = \frac{4v^2(u^4-6u^2+1)}{(u^2+1)^2(v^2+1)^2}.$$

Problem 3.10. Consider the map $\mathcal{H}\colon \mathbb{R}^3 \to \mathbb{R}^4$ defined such that

$$(x, y, z) \mapsto (xy, yz, xz, x^2 - y^2).$$

Prove that when it is restricted to the sphere S^2 (in $\mathbb{R}^3$), we have $\mathcal{H}(x, y, z) = \mathcal{H}(x', y', z')$ iff $(x', y', z') = (x, y, z)$ or $(x', y', z') = (-x, -y, -z)$. In other words, the inverse image of every point in $\mathcal{H}(S^2)$ consists of two antipodal points.

(a) Prove that the map $\mathcal{H}$ induces an injective map from the projective plane onto $\mathcal{H}(S^2)$, and that it is a homeomorphism.

(b) The map $\mathcal{H}$ allows us to realize concretely the projective plane in $\mathbb{R}^4$ as an embedded manifold. Consider the three maps from $\mathbb{R}^2$ to $\mathbb{R}^4$ given by

$$\begin{aligned}
\psi_1(u,v) &= \left(\frac{uv}{u^2+v^2+1}, \frac{v}{u^2+v^2+1}, \frac{u}{u^2+v^2+1}, \frac{u^2-v^2}{u^2+v^2+1}\right),\\
\psi_2(u,v) &= \left(\frac{u}{u^2+v^2+1}, \frac{v}{u^2+v^2+1}, \frac{uv}{u^2+v^2+1}, \frac{u^2-1}{u^2+v^2+1}\right),\\
\psi_3(u,v) &= \left(\frac{u}{u^2+v^2+1}, \frac{uv}{u^2+v^2+1}, \frac{v}{u^2+v^2+1}, \frac{1-u^2}{u^2+v^2+1}\right).
\end{aligned}$$

Observe that ψ_1 is the composition $\mathcal{H} \circ \alpha_1$, where $\alpha_1 \colon \mathbb{R}^2 \longrightarrow S^2$ is given by

$$(u, v) \mapsto \left(\frac{u}{\sqrt{u^2+v^2+1}}, \frac{v}{\sqrt{u^2+v^2+1}}, \frac{1}{\sqrt{u^2+v^2+1}}\right),$$

that ψ_2 is the composition $\mathcal{H} \circ \alpha_2$, where $\alpha_2 \colon \mathbb{R}^2 \longrightarrow S^2$ is given by

$$(u, v) \mapsto \left(\frac{u}{\sqrt{u^2+v^2+1}}, \frac{1}{\sqrt{u^2+v^2+1}}, \frac{v}{\sqrt{u^2+v^2+1}}\right).$$

and ψ_3 is the composition $\mathcal{H} \circ \alpha_3$, where $\alpha_3 \colon \mathbb{R}^2 \longrightarrow S^2$ is given by

$$(u, v) \mapsto \left(\frac{1}{\sqrt{u^2+v^2+1}}, \frac{u}{\sqrt{u^2+v^2+1}}, \frac{v}{\sqrt{u^2+v^2+1}}\right),$$

Prove that each ψ_i is injective, continuous and nonsingular (i.e., the Jacobian has rank 2).

(c) Prove that if $\psi_1(u,v) = (x,y,z,t)$, then

$$y^2+z^2 \leq \frac{1}{4} \quad \text{and} \quad y^2+z^2 = \frac{1}{4} \quad \text{iff} \quad u^2+v^2 = 1.$$

Prove that if $\psi_1(u,v) = (x,y,z,t)$, then u and v satisfy the equations

$$\begin{aligned}(y^2+z^2)u^2 - zu + z^2 &= 0\\ (y^2+z^2)v^2 - yv + y^2 &= 0.\end{aligned}$$

Prove that if $y^2+z^2 \neq 0$, then

$$u = \frac{z(1-\sqrt{1-4(y^2+z^2)})}{2(y^2+z^2)} \quad \text{if} \quad u^2+v^2 \leq 1,$$

else

$$u = \frac{z(1+\sqrt{1-4(y^2+z^2)})}{2(y^2+z^2)} \quad \text{if} \quad u^2+v^2 \geq 1,$$

and there are similar formulae for v. Prove that the expression giving u in terms of y and z is continuous everywhere in $\{(y,z) \mid y^2+z^2 \leq \frac{1}{4}\}$ and similarly for the expression giving v in terms of y and z. Conclude that $\psi_1 \colon \mathbb{R}^2 \to \psi_1(\mathbb{R}^2)$ is a homeomorphism onto its image. Therefore, $U_1 = \psi_1(\mathbb{R}^2)$ is an open subset of $\mathcal{H}(S^2)$.

Prove that if $\psi_2(u,v) = (x,y,z,t)$, then

$$x^2+y^2 \leq \frac{1}{4} \quad \text{and} \quad x^2+y^2 = \frac{1}{4} \quad \text{iff} \quad u^2+v^2 = 1.$$

Prove that if $\psi_2(u,v) = (x,y,z,t)$, then u and v satisfy the equations

$$\begin{aligned}(x^2+y^2)u^2 - xu + x^2 &= 0\\ (x^2+y^2)v^2 - yv + y^2 &= 0.\end{aligned}$$

Conclude that $\psi_2 \colon \mathbb{R}^2 \to \psi_2(\mathbb{R}^2)$ is a homeomorphism onto its image and that the set $U_2 = \psi_2(\mathbb{R}^2)$ is an open subset of $\mathcal{H}(S^2)$.

Prove that if $\psi_3(u,v) = (x,y,z,t)$, then

$$x^2+z^2 \leq \frac{1}{4} \quad \text{and} \quad x^2+z^2 = \frac{1}{4} \quad \text{iff} \quad u^2+v^2 = 1.$$

Prove that if $\psi_3(u,v) = (x,y,z,t)$, then u and v satisfy the equations

$$\begin{aligned}(x^2+z^2)u^2 - xu + x^2 &= 0\\ (x^2+z^2)v^2 - zv + z^2 &= 0.\end{aligned}$$

Conclude that $\psi_3\colon \mathbb{R}^2 \to \psi_3(\mathbb{R}^2)$ is a homeomorphism onto its image and that the set $U_3 = \psi_3(\mathbb{R}^2)$ is an open subset of $\mathcal{H}(S^2)$.

Prove that the union of the U_i's covers $\mathcal{H}(S^2)$. Conclude that ψ_1, ψ_2, ψ_3 are parametrizations of $\mathbb{RP}^2$ as a smooth manifold in $\mathbb{R}^4$.

(d) Plot the surfaces obtained by dropping the fourth coordinate and the third coordinates, respectively (with $u, v \in [-1, 1]$).

(e) Prove that if $(x, y, z, t) \in \mathcal{H}(S^2)$, then

$$\begin{aligned} x^2y^2 + x^2z^2 + y^2z^2 &= xyz \\ x(z^2 - y^2) &= yzt. \end{aligned}$$

Prove that the zero locus of these equations strictly contains $\mathcal{H}(S^2)$. This is a "famous mistake" of Hilbert and Cohn-Vossen in *Geometry and the Immagination*!

Finding a set of equations defining exactly $\mathcal{H}(S^2)$ appears to be an open problem.

Problem 3.11. Pick any irrational multiple λ of 2π, and define

$$G = \left\{ g_t = \begin{pmatrix} e^{ti} & 0 \\ 0 & e^{\lambda ti} \end{pmatrix} \;\middle|\; t \in \mathbb{R} \right\}.$$

(1) Check that G is a subgroup of $\mathbf{GL}(2, \mathbb{C})$.

(2) Prove that the map $\varphi\colon t \mapsto g_t$ is a continuous isomorphism of $(\mathbb{R}, +)$ onto G, but that φ^{-1} is not continuous.

Check that $\mathfrak{g}$ (as defined in Proposition 3.9) is the one dimensional vector space spanned by

$$W = \begin{pmatrix} i & 0 \\ 0 & \lambda i \end{pmatrix},$$

and that $e^{tW} = g_t$ for all $t \in \mathbb{R}$.

(3) For every $r > 0$ ($r \in \mathbb{R}$), prove that

$$\exp(\{tW \mid t \in (-r, r)\}) = \{g_t \mid t \in (-r, r)\}$$

is **not** a neighborhood of I in G.

The problem is that there are elements of G of the form $g_{2\pi n}$ for some large n that are arbitrarily close to I, so they are exponential images of very short vectors in $\mathrm{M}_2(\mathbb{C})$, but they are exponential images only of very long vectors in $\mathfrak{g}$.

(4) Prove that the closure of the group G is the group

$$\overline{G} = \left\{ \begin{pmatrix} e^{ti} & 0 \\ 0 & e^{si} \end{pmatrix} \;\middle|\; t, s \in \mathbb{R} \right\},$$

and that G is dense in $\overline{G}$.

Chapter 4

Groups and Group Actions

This chapter provides the foundations for deriving a class of manifolds known as homogeneous spaces. It begins with a short review of group theory, introduces the concept of a group acting on a set, and defines the Grassmanians and Stiefel manifolds as homogenous manifolds arising from group actions of Lie groups. The last section provides an overview of topological groups, of which Lie groups are a special example, and contains more advanced material that may be skipped upon first reading.

4.1 Basic Concepts of Groups

We begin with a brief review of the group theory necessary for understanding the concept of a group acting on a set. Readers familiar with this material may proceed to the next section.

Definition 4.1. A *group* is a set G equipped with a binary operation $\cdot\colon G \times G \to G$ that associates an element $a \cdot b \in G$ to every pair of elements $a, b \in G$, and having the following properties: $\cdot$ is associative, has an identity element $e \in G$, and every element in G is invertible (w.r.t. $\cdot$). More explicitly, this means that the following equations hold for all $a, b, c \in G$:

(G1) $a \cdot (b \cdot c) = (a \cdot b) \cdot c.$ (associativity)

(G2) $a \cdot e = e \cdot a = a.$ (identity)

(G3) For every $a \in G$, there is some $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e.$ (inverse)

A group G is *abelian* (or *commutative*) if

$$a \cdot b = b \cdot a \quad \text{for all } a, b \in G.$$

A set M together with an operation $\cdot\colon M \times M \to M$ and an element e satisfying only conditions (G1) and (G2) is called a *monoid*. For example, the set $\mathbb{N} = \{0, 1, \ldots, n, \ldots\}$ of natural numbers is a (commutative) monoid under addition. However, it is not a group.

Some examples of groups are given below.

Example 4.1.

1. The set $\mathbb{Z} = \{\ldots, -n, \ldots, -1, 0, 1, \ldots, n, \ldots\}$ of integers is an abelian group under addition, with identity element 0. However, $\mathbb{Z}^* = \mathbb{Z} - \{0\}$ is not a group under multiplication, but rather a commutative monoid.

2. The set $\mathbb{Q}$ of rational numbers (fractions p/q with $p, q \in \mathbb{Z}$ and $q \neq 0$) is an abelian group under addition, with identity element 0. The set $\mathbb{Q}^* = \mathbb{Q} - \{0\}$ is also an abelian group under multiplication, with identity element 1.

3. Similarly, the sets $\mathbb{R}$ of real numbers and $\mathbb{C}$ of complex numbers are abelian groups under addition (with identity element 0), and $\mathbb{R}^* = \mathbb{R} - \{0\}$ and $\mathbb{C}^* = \mathbb{C} - \{0\}$ are abelian groups under multiplication (with identity element 1).

4. The sets $\mathbb{R}^n$ and $\mathbb{C}^n$ of n-tuples of real or complex numbers are groups under componentwise addition:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n),$$

   with identity element $(0, \ldots, 0)$. All these groups are abelian.

5. Given any nonempty set S, the set of bijections $f\colon S \to S$, also called *permutations of S*, is a group under function composition (i.e., the multiplication of f and g is the composition $g \circ f$), with identity element the identity function id_S. This group is not abelian as soon as S has more than two elements.

6. The set of $n \times n$ matrices with real (or complex) coefficients is an abelian group under addition of matrices, with identity element the null matrix. It is denoted by $\mathrm{M}_n(\mathbb{R})$ (or $\mathrm{M}_n(\mathbb{C})$).

7. The set $\mathbb{R}[X]$ of all polynomials in one variable with real coefficients is an abelian group under addition of polynomials.

8. The set of $n \times n$ invertible matrices with real (or complex) coefficients is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *general linear group* and is usually denoted by $\mathbf{GL}(n, \mathbb{R})$ (or $\mathbf{GL}(n, \mathbb{C})$).

9. The set of $n \times n$ invertible matrices with real (or complex) coefficients and determinant $+1$ is a group under matrix multiplication, with identity element the identity matrix I_n. This group is called the *special linear group* and is usually denoted by $\mathbf{SL}(n, \mathbb{R})$ (or $\mathbf{SL}(n, \mathbb{C})$).

10. The set of $n \times n$ invertible matrices with real coefficients such that $RR^\top = I_n$ and of determinant $+1$ is a group called the *orthogonal group* and is usually denoted by $\mathbf{SO}(n)$ (where $R^\top$ is the *transpose* of the matrix R, i.e., the rows of $R^\top$ are the columns of R). It corresponds to the rotations in $\mathbb{R}^n$.

11. Given an open interval (a,b), the set $\mathcal{C}((a,b))$ of continuous functions $f\colon (a,b)\to\mathbb{R}$ is an abelian group under the operation $f+g$ defined such that
$$(f+g)(x) = f(x) + g(x)$$
for all $x\in(a,b)$.

It is customary to denote the operation of an abelian group G by $+$, in which case the inverse a^{-1} of an element $a\in G$ is denoted by $-a$.

The identity element of a group is *unique*. In fact, we can prove a more general fact:

Fact 1. If a binary operation $\cdot\colon M\times M\to M$ is associative and if $e'\in M$ is a left identity and $e''\in M$ is a right identity, which means that
$$e'\cdot a = a \quad \text{for all} \quad a\in M \tag{G2l}$$
and
$$a\cdot e'' = a \quad \text{for all} \quad a\in M, \tag{G2r}$$
then $e' = e''$.

Proof. If we let $a = e''$ in equation (G2l), we get
$$e'\cdot e'' = e'',$$
and if we let $a = e'$ in equation (G2r), we get
$$e'\cdot e'' = e',$$
and thus
$$e' = e'\cdot e'' = e'',$$
as claimed. □

Fact 1 implies that the identity element of a monoid is unique, and since every group is a monoid, the identity element of a group is unique. Furthermore, every element in a group has a *unique inverse*. This is a consequence of a slightly more general fact:

Fact 2. In a monoid M with identity element e, if some element $a\in M$ has some left inverse $a'\in M$ and some right inverse $a''\in M$, which means that
$$a'\cdot a = e \tag{G3l}$$
and
$$a\cdot a'' = e, \tag{G3r}$$
then $a' = a''$.

Proof. Using (G3l) and the fact that e is an identity element, we have

$$(a' \cdot a) \cdot a'' = e \cdot a'' = a''.$$

Similarly, Using (G3r) and the fact that e is an identity element, we have

$$a' \cdot (a \cdot a'') = a' \cdot e = a'.$$

However, since M is monoid, the operation $\cdot$ is associative, so

$$a' = a' \cdot (a \cdot a'') = (a' \cdot a) \cdot a'' = a'',$$

as claimed. □

Remark: Axioms (G2) and (G3) can be weakened a bit by requiring only (G2r) (the existence of a right identity) and (G3r) (the existence of a right inverse for every element) (or (G2l) and (G3l)). It is a good exercise to prove that the group axioms (G2) and (G3) follow from (G2r) and (G3r).

Given a group G, for any two subsets $R, S \subseteq G$, we let

$$RS = \{r \cdot s \mid r \in R,\ s \in S\}.$$

In particular, for any $g \in G$, if $R = \{g\}$, we write

$$gS = \{g \cdot s \mid s \in S\},$$

and similarly, if $S = \{g\}$, we write

$$Rg = \{r \cdot g \mid r \in R\}.$$

From now on, we will drop the multiplication sign and write $g_1 g_2$ for $g_1 \cdot g_2$.

Definition 4.2. Given a group G, a subset H of G is a *subgroup of* G iff

(1) The identity element e of G also belongs to H ($e \in H$);

(2) For all $h_1, h_2 \in H$, we have $h_1 h_2 \in H$;

(3) For all $h \in H$, we have $h^{-1} \in H$.

It is easily checked that a subset $H \subseteq G$ is a subgroup of G iff H is nonempty and whenever $h_1, h_2 \in H$, then $h_1 h_2^{-1} \in H$.

Definition 4.3. If H is a subgroup of G and $g \in G$ is any element, the sets of the form gH are called *left cosets of H in G* and the sets of the form Hg are called *right cosets of H in G.*

The left cosets (resp. right cosets) of H induce an equivalence relation $\sim$ defined as follows: For all $g_1, g_2 \in G$,

$$g_1 \sim g_2 \quad \text{iff} \quad g_1H = g_2H$$

(resp. $g_1 \sim g_2$ iff $Hg_1 = Hg_2$).

Obviously, $\sim$ is an equivalence relation. It is easy to see that $g_1H = g_2H$ iff $g_2^{-1}g_1 \in H$, so the equivalence class of an element $g \in G$ is the coset gH (resp. Hg). The set of left cosets of H in G (which, in general, is **not** a group) is denoted G/H. The "points" of G/H are obtained by "collapsing" all the elements in a coset into a single element. This is the same intuition used for constructing the quotient space topology. The set of right cosets is denoted by $H\backslash G$.

It is tempting to define a multiplication operation on left cosets (or right cosets) by setting

$$(g_1H)(g_2H) = (g_1g_2)H,$$

but this operation is not well defined in general, unless the subgroup H possesses a special property. This property is typical of the kernels of group homomorphisms, so we are led to

Definition 4.4. Given any two groups G and G', a function $\varphi\colon G \to G'$ is a *homomorphism* iff

$$\varphi(g_1g_2) = \varphi(g_1)\varphi(g_2), \quad \text{for all } g_1, g_2 \in G.$$

Taking $g_1 = g_2 = e$ (in G), we see that

$$\varphi(e) = e',$$

and taking $g_1 = g$ and $g_2 = g^{-1}$, we see that

$$\varphi(g^{-1}) = \varphi(g)^{-1}.$$

If $\varphi\colon G \to G'$ and $\psi\colon G' \to G''$ are group homomorphisms, then $\psi \circ \varphi\colon G \to G''$ is also a homomorphism. If $\varphi\colon G \to G'$ is a homomorphism of groups, and $H \subseteq G$, $H' \subseteq G'$ are two subgroups, then it is easily checked that

$$\mathrm{Im}\, H = \varphi(H) = \{\varphi(g) \mid g \in H\}$$

is a subgroup of G' called the *image of H by φ*, and

$$\varphi^{-1}(H') = \{g \in G \mid \varphi(g) \in H'\}$$

is a subgroup of G. In particular, when $H' = \{e'\}$, we obtain the *kernel* $\mathrm{Ker}\ \varphi$ of φ. Thus,

$$\mathrm{Ker}\ \varphi = \{g \in G \mid \varphi(g) = e'\}.$$

It is immediately verified that $\varphi\colon G \to G'$ is injective iff $\mathrm{Ker}\ \varphi = \{e\}$. (We also write $\mathrm{Ker}\ \varphi = (0)$.) We say that φ is an *isomorphism* if there is a homomorphism $\psi\colon G' \to G$, so that

$$\psi \circ \varphi = \mathrm{id}_G \quad \text{and} \quad \varphi \circ \psi = \mathrm{id}_{G'}.$$

In this case, ψ is unique and it is denoted φ^{-1}. When φ is an isomorphism, we say the the groups G and G' are *isomorphic* and we write $G \cong G'$ (or $G \approx G'$). When $G' = G$, a group isomorphism is called an *automorphism*.

We claim that $H = \mathrm{Ker}\ \varphi$ satisfies the following property:

$$gH = Hg, \quad \text{for all } g \in G. \tag{$*$}$$

Note that $(*)$ is equivalent to

$$gHg^{-1} = H, \quad \text{for all } g \in G,$$

and the above is equivalent to

$$gHg^{-1} \subseteq H, \quad \text{for all } g \in G. \tag{$**$}$$

This is because $gHg^{-1} \subseteq H$ implies $H \subseteq g^{-1}Hg$, and this for all $g \in G$. But

$$\varphi(ghg^{-1}) = \varphi(g)\varphi(h)\varphi(g^{-1}) = \varphi(g)e'\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e',$$

for all $h \in H = \mathrm{Ker}\ \varphi$ and all $g \in G$. Thus, by definition of $H = \mathrm{Ker}\ \varphi$, we have $gHg^{-1} \subseteq H$.

Definition 4.5. For any group G, a subgroup N of G is a *normal subgroup* of G iff

$$gNg^{-1} = N, \quad \text{for all } g \in G.$$

This is denoted by $N \lhd G$.

If N is a normal subgroup of G, the equivalence relation induced by left cosets is the same as the equivalence induced by right cosets. Furthermore, this equivalence relation $\sim$ is a *congruence*, which means that: For all $g_1, g_2, g_1', g_2' \in G$,

(1) If $g_1N = g_1'N$ and $g_2N = g_2'N$, then $g_1g_2N = g_1'g_2'N$, and

(2) If $g_1N = g_2N$, then $g_1^{-1}N = g_2^{-1}N$.

As a consequence, we can define a group structure on the set $G/\sim$ of equivalence classes modulo $\sim$, by setting

$$(g_1N)(g_2N) = (g_1g_2)N.$$

This group is denoted G/N. The equivalence class gN of an element $g \in G$ is also denoted $\overline{g}$. The map $\pi\colon G \to G/N$, given by

$$\pi(g) = \overline{g} = gN$$

is clearly a group homomorphism called the *canonical projection*.

Given a homomorphism of groups $\varphi\colon G \to G'$, we easily check that the groups $G/\mathrm{Ker}\ \varphi$ and $\mathrm{Im}\ \varphi = \varphi(G)$ are isomorphic.

4.2 Group Actions: Part I, Definition and Examples

If X is a set (usually some kind of geometric space, for example, the sphere in $\mathbb{R}^3$, the upper half-plane, etc.), the "symmetries" of X are often captured by the action of a group G on X. In fact, if G is a Lie group and the action satisfies some simple properties, the set X can be given a manifold structure which makes it a projection (quotient) of G, a so-called "homogeneous space."

Definition 4.6. Given a set X and a group G, a *left action of* G *on* X (for short, an *action of* G *on* X) is a function $\varphi\colon G \times X \to X$, such that:

(1) For all $g, h \in G$ and all $x \in X$,

$$\varphi(g, \varphi(h, x)) = \varphi(gh, x),$$

(2) For all $x \in X$,

$$\varphi(1, x) = x,$$

where $1 \in G$ is the identity element of G.

To alleviate the notation, we usually write $g \cdot x$ or even gx for $\varphi(g, x)$, in which case the above axioms read:

(1) For all $g, h \in G$ and all $x \in X$,

$$g \cdot (h \cdot x) = gh \cdot x,$$

(2) For all $x \in X$,

$$1 \cdot x = x.$$

The set X is called a *(left) G-set*. The action φ is *faithful* or *effective* iff for every g, if $g \cdot x = x$ for all $x \in X$, then $g = 1$. Faithful means that if the action of some element g behaves like the identity, then g must be the identity element. The action φ is *transitive* iff for any two elements $x, y \in X$, there is some $g \in G$ so that $g \cdot x = y$.

Given an action $\varphi\colon G\times X \to X$, for every $g \in G$, we have a function $\varphi_g\colon X \to X$ defined by

$$\varphi_g(x) = g\cdot x, \quad \text{for all } x \in X.$$

Observe that φ_g has $\varphi_{g^{-1}}$ as inverse, since

$$\varphi_{g^{-1}}(\varphi_g(x)) = \varphi_{g^{-1}}(g\cdot x) = g^{-1}\cdot(g\cdot x) = (g^{-1}g)\cdot x = 1\cdot x = x,$$

and similarly, $\varphi_g\circ\varphi_{g^{-1}} = \mathrm{id}$. Therefore, φ_g is a bijection of X; that is, φ_g is a permutation of X. Moreover, we check immediately that

$$\varphi_g\circ\varphi_h = \varphi_{gh},$$

so the map $g \mapsto \varphi_g$ is a group homomorphism from G to $\mathfrak{S}_X$, the group of permutations of X. With a slight abuse of notation, this group homomorphism $G \longrightarrow \mathfrak{S}_X$ is also denoted φ.

Conversely, it is easy to see that any group homomorphism $\varphi\colon G \to \mathfrak{S}_X$ yields a group action $\cdot\colon G\times X \longrightarrow X$, by setting

$$g\cdot x = \varphi(g)(x).$$

Observe that an action φ is faithful iff the group homomorphism $\varphi\colon G \to \mathfrak{S}_X$ is injective, i.e. iff φ has a trivial kernel. Also, we have $g\cdot x = y$ iff $g^{-1}\cdot y = x$, since $(gh)\cdot x = g\cdot(h\cdot x)$ and $1\cdot x = x$, for all $g, h \in G$ and all $x \in X$.

Definition 4.7. Given two G-sets X and Y, a function $f\colon X \to Y$ is said to be *equivariant*, or a *G-map*, iff for all $x \in X$ and all $g \in G$, we have

$$f(g\cdot x) = g\cdot f(x).$$

Equivalently, if the G-actions are denoted by $\varphi\colon G\times X \to X$ and $\psi\colon G\times Y \to Y$, we have the following commutative diagram for all $g \in G$:

$$\begin{tikzcd}
X \arrow[r, "\varphi_g"] \arrow[d, "f"'] & X \arrow[d, "f"] \\
Y \arrow[r, "\psi_g"'] & Y.
\end{tikzcd}$$

Remark: We can also define a *right action* $\cdot\colon X\times G \to X$ of a group G on a set X as a map satisfying the conditions

(1) For all $g, h \in G$ and all $x \in X$,

$$(x\cdot g)\cdot h = x\cdot gh,$$

(2) For all $x \in X$,

$$x\cdot 1 = x.$$

Every notion defined for left actions is also defined for right actions in the obvious way.

However, one change is necessary. For every $g \in G$, the map $\varphi_g \colon X \to X$ must be defined as

$$\varphi_g(x) = x \cdot g^{-1},$$

in order for the map $g \mapsto \varphi_g$ from G to $\mathfrak{S}_X$ to be a homomorphism ($\varphi_g \circ \varphi_h = \varphi_{gh}$). Conversely, given a homomorphism $\varphi \colon G \to \mathfrak{S}_X$, we get a right action $\cdot \colon X \times G \longrightarrow X$ by setting

$$x \cdot g = \varphi(g^{-1})(x).$$

Here are some examples of (left) group actions.

Example 4.2. The unit sphere S^2 (more generally, S^{n-1}).

Recall that for any $n \geq 1$, the *(real) unit sphere* S^{n-1} is the set of points in $\mathbb{R}^n$ given by

$$S^{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 = 1\}.$$

In particular, S^2 is the usual sphere in $\mathbb{R}^3$. Since the group $\mathbf{SO}(3) = \mathbf{SO}(3, \mathbb{R})$ consists of (orientation preserving) linear isometries, i.e., *linear* maps that are distance preserving (and of determinant $+1$), and every linear map leaves the origin fixed, we see that any rotation maps S^2 into itself.

Beware that this would be false if we considered the group of *affine* isometries $\mathbf{SE}(3)$ of $\mathbb{E}^3$. For example, a screw motion does *not* map S^2 into itself, even though it is distance preserving, because the origin is translated.

Thus, for $X = S^2$ and $G = \mathbf{SO}(3)$, we have an action $\cdot \colon \mathbf{SO}(3) \times S^2 \to S^2$, given by the matrix multiplication

$$R \cdot x = Rx.$$

The verification that the above is indeed an action is trivial. This action is transitive. This is because, for any two points x, y on the sphere S^2, there is a rotation whose axis is perpendicular to the plane containing x, y and the center O of the sphere (this plane is not unique when x and y are antipodal, i.e., on a diameter) mapping x to y. See Figure 4.1.

Similarly, for any $n \geq 1$, let $X = S^{n-1}$ and $G = \mathbf{SO}(n)$ and define the action $\cdot \colon \mathbf{SO}(n) \times S^{n-1} \to S^{n-1}$ as $R \cdot x = Rx$. It is easy to show that this action is transitive.

Analogously, we can define the *(complex) unit sphere* Σ^{n-1}, as the set of points in $\mathbb{C}^n$ given by

$$\Sigma^{n-1} = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1\overline{z}_1 + \cdots + z_n\overline{z}_n = 1\}.$$

If we write $z_j = x_j + iy_j$, with $x_j, y_j \in \mathbb{R}$, then

$$\Sigma^{n-1} = \{(x_1, \ldots, x_n, y_1, \ldots, y_n) \in \mathbb{R}^{2n} \mid x_1^2 + \cdots + x_n^2 + y_1^2 + \cdots + y_n^2 = 1\}.$$

Therefore, we can view the complex sphere Σ^{n-1} (in $\mathbb{C}^n$) as the real sphere S^{2n-1} (in $\mathbb{R}^{2n}$). By analogy with the real case, we can define for $X = \Sigma^{n-1}$ and $G = \mathbf{SU}(n)$ an action $\cdot \colon \mathbf{SU}(n) \times \Sigma^{n-1} \to \Sigma^{n-1}$ of the group $\mathbf{SU}(n)$ of *linear* maps of $\mathbb{C}^n$ preserving the Hermitian inner product (and the origin, as all linear maps do), and this action is transitive.

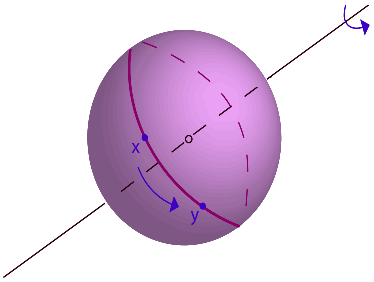


Figure 4.1: The rotation which maps x to y.

One should not confuse the unit sphere Σ^{n-1} with the hypersurface $S^{n-1}_{\mathbb{C}}$, given by

$$S^{n-1}_{\mathbb{C}} = \{(z_1, \ldots, z_n) \in \mathbb{C}^n \mid z_1^2 + \cdots + z_n^2 = 1\}.$$

For instance, one should check that a line L through the origin intersects Σ^{n-1} in a circle, whereas it intersects $S^{n-1}_{\mathbb{C}}$ in exactly two points! Recall for a fixed $u = (x_1, \ldots x_n, y_1, \ldots y_n) \in \mathbb{C}^n$, that $L = \{\gamma u \mid \gamma \in \mathbb{C}\}$. Since $\gamma = \rho(\cos\theta + i\sin\theta)$, we deduce that L is actually the two dimensional subspace through the origin spanned by the orthogonal vectors $(x_1, \ldots x_n, y_1, \ldots y_n)$ and $(-y_1, \cdots - y_n, x_1, \ldots x_n)$.

Example 4.3. The upper half-plane.

The *upper half-plane* H is the open subset of $\mathbb{R}^2$ consisting of all points $(x, y) \in \mathbb{R}^2$, with $y > 0$. It is convenient to identify H with the set of complex numbers $z \in \mathbb{C}$ such that $\Im z > 0$. Then we can let $X = H$ and $G = \mathbf{SL}(2, \mathbb{R})$ and define an action $\cdot\colon \mathbf{SL}(2, \mathbb{R}) \times H \to H$ of the group $\mathbf{SL}(2, \mathbb{R})$ on H, as follows: For any $z \in H$, for any $A \in \mathbf{SL}(2, \mathbb{R})$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $ad - bc = 1$.

It is easily verified that $A \cdot z$ is indeed always well defined and in H when $z \in H$ (check this). To see why this action is transitive, let z and w be two arbitrary points of H where $z = x + iy$ and $w = u + iv$ with $x, u \in \mathbb{R}$ and $y, v \in \mathbb{R}^+$ (i.e. y and v are positive real numbers). Define $A = \begin{pmatrix} \sqrt{\frac{v}{y}} & \frac{uy - vx}{\sqrt{yv}} \\ 0 & \sqrt{\frac{y}{v}} \end{pmatrix}$. Note that $A \in \mathbf{SL}(2, \mathbb{R})$. A routine calculation shows that $A \cdot z = w$.

Before introducing Example 4.4, we need to define the groups of Möbius transformations and the Riemann sphere. Maps of the form

$$z \mapsto \frac{az+b}{cz+d},$$

where $z \in \mathbb{C}$ and $ad - bc = 1$, are called *Möbius transformations*. Here, $a, b, c, d \in \mathbb{R}$, but in general, we allow $a, b, c, d \in \mathbb{C}$. Actually, these transformations are not necessarily defined everywhere on $\mathbb{C}$, for example, for $z = -d/c$ if $c \neq 0$. To fix this problem, we add a "point at infinity" ∞ to $\mathbb{C}$, and define Möbius transformations as functions $\mathbb{C} \cup \{\infty\} \longrightarrow \mathbb{C} \cup \{\infty\}$. If $c = 0$, the Möbius transformation sends ∞ to itself, otherwise, $-d/c \mapsto \infty$ and $\infty \mapsto a/c$.

The space $\mathbb{C} \cup \{\infty\}$ can be viewed as the plane $\mathbb{R}^2$ extended with a point at infinity. Using a stereographic projection from the sphere S^2 to the plane (say from the north pole to the equatorial plane), we see that there is a bijection between the sphere S^2 and $\mathbb{C} \cup \{\infty\}$. More precisely, the *stereographic projection* σ_N of the sphere S^2 from the north pole $N = (0, 0, 1)$ to the plane $z = 0$ (extended with the point at infinity ∞) is given by

$$(x, y, z) \in S^2 - \{(0, 0, 1)\} \mapsto \left(\frac{x}{1-z}, \frac{y}{1-z}\right) = \frac{x+iy}{1-z} \in \mathbb{C}, \quad \text{with} \quad (0, 0, 1) \mapsto \infty.$$

The inverse stereographic projection σ_N^{-1} is given by

$$(x, y) \mapsto \left(\frac{2x}{x^2+y^2+1}, \frac{2y}{x^2+y^2+1}, \frac{x^2+y^2-1}{x^2+y^2+1}\right), \quad \text{with} \quad \infty \mapsto (0, 0, 1).$$

Intuitively, the inverse stereographic projection "wraps" the equatorial plane around the sphere. See Figure 3.3.

The space $\mathbb{C} \cup \{\infty\}$ is known as the *Riemann sphere*. We will see shortly that $\mathbb{C} \cup \{\infty\} \cong S^2$ is also the complex projective line $\mathbb{CP}^1$. In summary, Möbius transformations are bijections of the Riemann sphere. It is easy to check that these transformations form a group under composition for all $a, b, c, d \in \mathbb{C}$, with $ad - bc = 1$. This is the *Möbius group*, denoted $\mathbf{M\ddot{o}b}^+$. The Möbius transformations corresponding to the case $a, b, c, d \in \mathbb{R}$, with $ad - bc = 1$ form a subgroup of $\mathbf{M\ddot{o}b}^+$ denoted $\mathbf{M\ddot{o}b}^+_{\mathbb{R}}$.

The map from $\mathbf{SL}(2, \mathbb{C})$ to $\mathbf{M\ddot{o}b}^+$ that sends $A \in \mathbf{SL}(2, \mathbb{C})$ to the corresponding Möbius transformation is a surjective group homomorphism, and one checks easily that its kernel is $\{-I, I\}$ (where I is the 2×2 identity matrix). Therefore, the Möbius group $\mathbf{M\ddot{o}b}^+$ is isomorphic to the quotient group $\mathbf{SL}(2, \mathbb{C})/\{-I, I\}$, denoted $\mathbf{PSL}(2, \mathbb{C})$. This latter group turns out to be the group of projective transformations of the projective space $\mathbb{CP}^1$. The same reasoning shows that the subgroup $\mathbf{M\ddot{o}b}^+_{\mathbb{R}}$ is isomorphic to $\mathbf{SL}(2, \mathbb{R})/\{-I, I\}$, denoted $\mathbf{PSL}(2, \mathbb{R})$.

Example 4.4. The Riemann sphere $\mathbb{C} \cup \{\infty\}$.

Let $X = \mathbb{C} \cup \{\infty\}$ and $G = \mathbf{SL}(2, \mathbb{C})$. The group $\mathbf{SL}(2, \mathbb{C})$ acts on $\mathbb{C} \cup \{\infty\} \cong S^2$ the same way that $\mathbf{SL}(2, \mathbb{R})$ acts on H, namely: For any $A \in \mathbf{SL}(2, \mathbb{C})$, for any $z \in \mathbb{C} \cup \{\infty\}$,

$$A \cdot z = \frac{az + b}{cz + d},$$

where

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad ad - bc = 1.$$

This action is transitive, an exercise we leave for the reader.

Example 4.5. The unit disk.

One may recall from complex analysis that the scaled (complex) Möbius transformation

$$z \mapsto \frac{z - i}{z + i}$$

is a biholomorphic or analytic isomorphism between the upper half plane H and the open unit disk

$$D = \{z \in \mathbb{C} \mid |z| < 1\}.$$

As a consequence, it is possible to define a transitive action of $\mathbf{SL}(2, \mathbb{R})$ on D. This can be done in a more direct fashion, using a group isomorphic to $\mathbf{SL}(2, \mathbb{R})$, namely, $\mathbf{SU}(1, 1)$ (a group of complex matrices), but we don't want to do this right now.

Example 4.6. The unit Riemann sphere revisited.

Another interesting action is the action of $\mathbf{SU}(2)$ on the extended plane $\mathbb{C} \cup \{\infty\}$. Recall that the group $\mathbf{SU}(2)$ consists of all complex matrices of the form

$$A = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \qquad \alpha, \beta \in \mathbb{C}, \qquad \alpha\overline{\alpha} + \beta\overline{\beta} = 1,$$

Let $X = \mathbb{C} \cup \{\infty\}$ and $G = \mathbf{SU}(2)$. The action $\cdot \colon \mathbf{SU}(2) \times (\mathbb{C} \cup \{\infty\}) \to \mathbb{C} \cup \{\infty\}$ is given by

$$A \cdot w = \frac{\alpha w + \beta}{-\overline{\beta} w + \overline{\alpha}}, \quad w \in \mathbb{C} \cup \{\infty\}.$$

This action is transitive, but the proof of this fact relies on the surjectivity of the group homomorphism

$$\rho \colon \mathbf{SU}(2) \to \mathbf{SO}(3)$$

defined below, and the stereographic projection σ_N from S^2 onto $\mathbb{C} \cup \{\infty\}$. In particular, take $z, w \in \mathbb{C} \cup \{\infty\}$, use the inverse stereographic projection to obtain two points on S^2, namely $\sigma_N^{-1}(z)$ and $\sigma_N^{-1}(w)$. Then apply the appropriate rotation $R \in \mathbf{SO}(3)$ to map $\sigma_N^{-1}(z)$ onto $\sigma_N^{-1}(w)$. Such a rotation exists by the argument presented in Example 4.2. Since

$\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ is surjective (see below), we know there must exist $A \in \mathbf{SU}(2)$ such that $\rho(A) = R$ and $A \cdot z = w$.

Using the stereographic projection σ_N from S^2 onto $\mathbb{C} \cup \{\infty\}$ and its inverse σ_N^{-1}, we can define an action of $\mathbf{SU}(2)$ on S^2 by

$$A \cdot (x, y, z) = \sigma_N^{-1}(A \cdot \sigma_N(x, y, z)), \quad (x, y, z) \in S^2.$$

Although this is not immediately obvious, it turns out that $\mathbf{SU}(2)$ acts on S^2 by maps that are restrictions of linear maps to S^2, and since these linear maps preserve S^2, they are orthogonal transformations. Thus, we obtain a continuous (in fact, smooth) group homomorphism

$$\rho\colon \mathbf{SU}(2) \to \mathbf{O}(3).$$

Since $\mathbf{SU}(2)$ is connected and ρ is continuous, the image of $\mathbf{SU}(2)$ is contained in the connected component of I in $\mathbf{O}(3)$, namely $\mathbf{SO}(3)$, so ρ is a homomorphism

$$\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3).$$

We will see that this homomorphism is surjective and that its kernel is $\{I, -I\}$. The upshot is that we have an isomorphism

$$\mathbf{SO}(3) \cong \mathbf{SU}(2)/\{I, -I\}.$$

The homomorphism ρ is a way of describing how a unit quaternion (any element of $\mathbf{SU}(2)$) induces a rotation, *via* the stereographic projection and its inverse. If we write $\alpha = a + ib$ and $\beta = c + id$, a rather tedious computation yields

$$\rho(A) = \begin{pmatrix} a^2 - b^2 - c^2 + d^2 & -2ab - 2cd & -2ac + 2bd \\ 2ab - 2cd & a^2 - b^2 + c^2 - d^2 & -2ad - 2bc \\ 2ac + 2bd & 2ad - 2bc & a^2 + b^2 - c^2 - d^2 \end{pmatrix}.$$

One can check that $\rho(A)$ is indeed a rotation matrix which represents the rotation whose axis is the line determined by the vector $(d, -c, b)$ and whose angle $\theta \in [-\pi, \pi]$ is determined by

$$\cos\frac{\theta}{2} = |a|.$$

We can also compute the derivative $d\rho_I\colon \mathfrak{su}(2) \to \mathfrak{so}(3)$ of ρ at I as follows. Recall that $\mathfrak{su}(2)$ consists of all complex matrices of the form

$$\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix}, \quad b, c, d \in \mathbb{R},$$

so pick the following basis for $\mathfrak{su}(2)$,

$$X_1 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and define the curves in $\mathbf{SU}(2)$ through I given by

$$c_1(t) = \begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix}, \quad c_2(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}, \quad c_3(t) = \begin{pmatrix} \cos t & i\sin t \\ i\sin t & \cos t \end{pmatrix}.$$

It is easy to check that $c_i'(0) = X_i$ for $i = 1, 2, 3$, and that

$$d\rho_I(X_1) = 2\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_2) = 2\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_3) = 2\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Thus we have

$$d\rho_I(X_1) = 2E_3, \quad d\rho_I(X_2) = -2E_2, \quad d\rho_I(X_3) = 2E_1,$$

where (E_1, E_2, E_3) is the basis of $\mathfrak{so}(3)$ given in Section 3.1, which means that $d\rho_I$ is an isomorphism between the Lie algebras $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$.

Recall from Proposition 3.13 that we have the commutative diagram

$$\begin{tikzcd}
\mathbf{SU}(2) \arrow[r, "\rho"] & \mathbf{SO}(3) \\
\mathfrak{su}(2) \arrow[u, "\exp"] \arrow[r, "d\rho_I"'] & \mathfrak{so}(3) . \arrow[u, "\exp"']
\end{tikzcd}$$

Since $d\rho_I$ is surjective and the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective, we conclude that ρ is surjective. (We also know from Section 3.1 that $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is surjective.) Observe that $\rho(-A) = \rho(A)$, and it is easy to check that $\operatorname{Ker}\rho = \{I, -I\}$.

Example 4.7. The set of $n \times n$ symmetric, positive, definite matrices, $\mathbf{SPD}(n)$.

Let $X = \mathbf{SPD}(n)$ and $G = \mathbf{GL}(n)$. The group $\mathbf{GL}(n) = \mathbf{GL}(n, \mathbb{R})$ acts on $\mathbf{SPD}(n)$ as follows: for all $A \in \mathbf{GL}(n)$ and all $S \in \mathbf{SPD}(n)$,

$$A \cdot S = ASA^\top.$$

It is easily checked that $ASA^\top$ is in $\mathbf{SPD}(n)$ if S is in $\mathbf{SPD}(n)$. First observe that $ASA^\top$ is symmetric since

$$(ASA^\top)^\top = AS^\top A^\top = ASA^\top.$$

Next recall the following characterization of positive definite matrix, namely

$$y^\top S y > 0, \qquad \text{whenever } y \neq 0.$$

We want to show $x^\top(A^\top SA)x > 0$ for all $x \neq 0$. Since A is invertible, we have $x = A^{-1}y$ for some nonzero y, and hence

$$\begin{aligned} x^\top(A^\top SA)x &= y^\top(A^{-1})^\top A^\top SAA^{-1}y \\ &= y^\top Sy > 0. \end{aligned}$$

Hence $A^\top SA$ is positive definite. This action is transitive because every SPD matrix S can be written as $S = AA^\top$, for some invertible matrix A (prove this as an exercise). Given any two SPD matrices $S_1 = A_1A_1^\top$ and $S_2 = A_2A_2^\top$ with A_1 and A_2 invertible, if $A = A_2A_1^{-1}$, we have

$$\begin{aligned} A \cdot S_1 &= A_2A_1^{-1}S_1(A_2A_1^{-1})^\top = A_2A_1^{-1}S_1(A_1^\top)^{-1}A_2^\top \\ &= A_2A_1^{-1}A_1A_1^\top(A_1^\top)^{-1}A_2^\top = A_2A_2^\top = S_2. \end{aligned}$$

Example 4.8. The projective spaces $\mathbb{RP}^n$ and $\mathbb{CP}^n$.

The *(real) projective space* $\mathbb{RP}^n$ is the set of all lines through the origin in $\mathbb{R}^{n+1}$; that is, the set of one-dimensional subspaces of $\mathbb{R}^{n+1}$ (where $n \geq 0$). Since a one-dimensional subspace $L \subseteq \mathbb{R}^{n+1}$ is spanned by any nonzero vector $u \in L$, we can view $\mathbb{RP}^n$ as the set of equivalence classes of nonzero vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

In terms of this definition, there is a projection $pr\colon (\mathbb{R}^{n+1} - \{0\}) \to \mathbb{RP}^n$, given by $pr(u) = [u]_\sim$, the equivalence class of u modulo $\sim$. Write $[u]$ for the line defined by the nonzero vector u. Since every line L in $\mathbb{R}^{n+1}$ intersects the sphere S^n in two antipodal points, we can view $\mathbb{RP}^n$ as the quotient of the sphere S^n by identification of antipodal points. See Figures 4.2 and 4.3.

Let $X = \mathbb{RP}^n$ and $G = \mathbf{SO}(n+1)$. We define an action of $\mathbf{SO}(n+1)$ on $\mathbb{RP}^n$ as follows: For any line $L = [u]$, for any $R \in \mathbf{SO}(n+1)$,

$$R \cdot L = [Ru].$$

Since R is linear, the line $[Ru]$ is well defined; that is, does not depend on the choice of $u \in L$. The reader can show that this action is transitive.

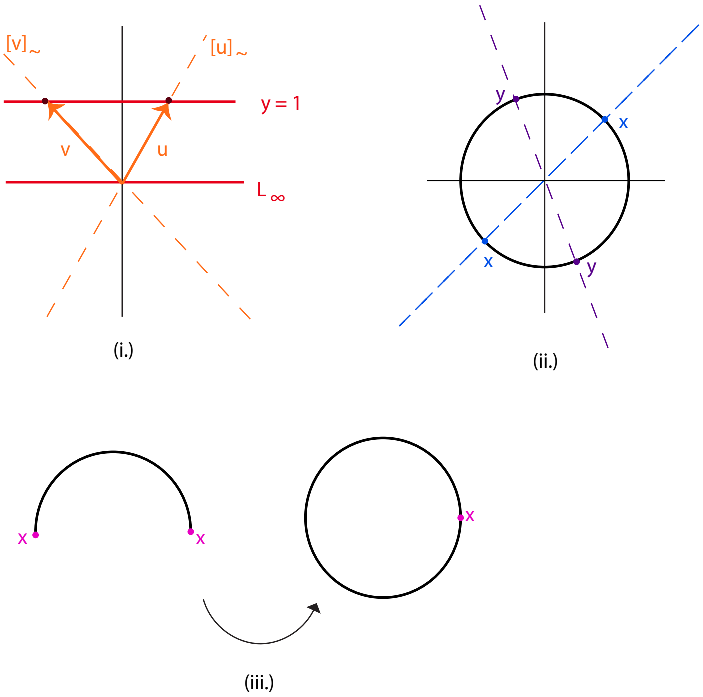
The *(complex) projective space* $\mathbb{CP}^n$ is defined analogously as the set of all lines through the origin in $\mathbb{C}^{n+1}$; that is, the set of one-dimensional subspaces of $\mathbb{C}^{n+1}$ (where $n \geq 0$). This time, we can view $\mathbb{CP}^n$ as the set of equivalence classes of vectors in $\mathbb{C}^{n+1} - \{0\}$ modulo the equivalence relation

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \neq 0 \in \mathbb{C}.$$

We have the projection $pr\colon \mathbb{C}^{n+1} - \{0\} \to \mathbb{CP}^n$, given by $pr(u) = [u]_\sim$, the equivalence class of u modulo $\sim$. Again, write $[u]$ for the line defined by the nonzero vector u. Let $X = \mathbb{CP}^n$ and $G = \mathbf{SU}(n+1)$. We define an action of $\mathbf{SU}(n+1)$ on $\mathbb{CP}^n$ as follows: For any line $L = [u]$, for any $R \in \mathbf{SU}(n+1)$,

$$R \cdot L = [Ru].$$

Again, this action is well defined and it is transitive. (Check this.)



(i.)

(ii.)

(iii.)

Figure 4.2: Three constructions for $\mathbb{RP}^1 \cong S^1$. Illustration (i.) applies the equivalence relation. Since any line through the origin, excluding the x-axis, intersects the line $y = 1$, its equivalence class is represented by its point of intersection on $y = 1$. Hence, $\mathbb{RP}^n$ is the disjoint union of the line $y = 1$ and the point of infinity given by the x-axis. Illustration (ii.) represents $\mathbb{RP}^1$ as the quotient of the circle S^1 by identification of antipodal points. Illustration (iii.) is a variation which glues the equatorial points of the upper semicircle.

Before progressing to our final example of group actions, we take a moment to construct $\mathbb{CP}^n$ as a quotient space of S^{2n+1}. Recall that $\Sigma^n \subseteq \mathbb{C}^{n+1}$, the unit sphere in $\mathbb{C}^{n+1}$, is defined by

$$\Sigma^n = \{(z_1, \ldots, z_{n+1}) \in \mathbb{C}^{n+1} \mid z_1\overline{z}_1 + \cdots + z_{n+1}\overline{z}_{n+1} = 1\}.$$

For any line $L = [u]$, where $u \in \mathbb{C}^{n+1}$ is a nonzero vector, writing $u = (u_1, \ldots, u_{n+1})$, a point $z \in \mathbb{C}^{n+1}$ belongs to L iff $z = \lambda(u_1, \ldots, u_{n+1})$, for some $\lambda \in \mathbb{C}$. Therefore, the intersection $L \cap \Sigma^n$ of the line L and the sphere Σ^n is given by

$$L \cap \Sigma^n = \{\lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \mid \lambda \in \mathbb{C},\ \lambda\overline{\lambda}(u_1\overline{u}_1 + \cdots + u_{n+1}\overline{u}_{n+1}) = 1\},$$

i.e.,

$$L \cap \Sigma^n = \left\{ \lambda(u_1, \ldots, u_{n+1}) \in \mathbb{C}^{n+1} \;\middle|\; \lambda \in \mathbb{C},\ |\lambda| = \frac{1}{\sqrt{|u_1|^2 + \cdots + |u_{n+1}|^2}} \right\}.$$

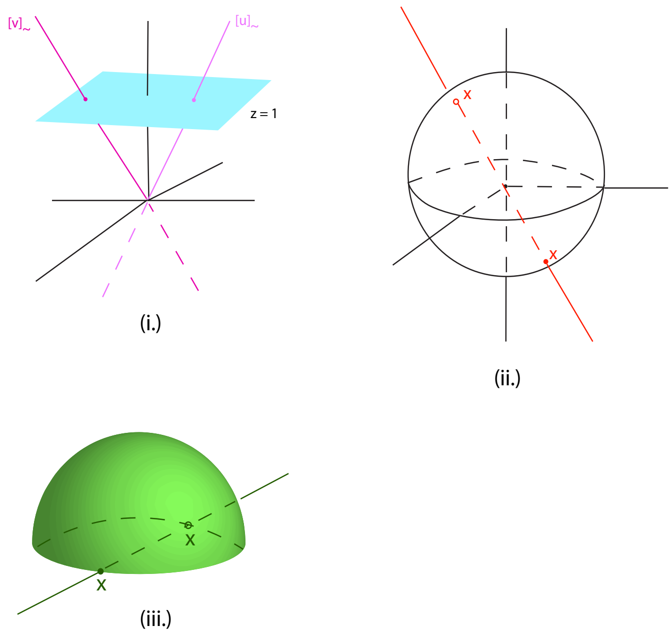


Figure 4.3: Three constructions for $\mathbb{RP}^2$. Illustration (i.) applies the equivalence relation. Since any line through the origin which is not contained in the xy-plane intersects the plane $z = 1$, its equivalence class is represented by its point of intersection on $z = 1$. Hence, $\mathbb{RP}^2$ is the disjoint union of the plane $z = 1$ and the copy of $\mathbb{RP}^1$ provided by the xy-plane. Illustration (ii.) represents $\mathbb{RP}^2$ as the quotient of the sphere S^2 by identification of antipodal points. Illustration (iii.) is a variation which glues the antipodal points on boundary of the unit disk, which is represented as as the upper hemisphere.

Thus, we see that there is a bijection between $L \cap \Sigma^n$ and the circle S^1; that is, geometrically $L \cap \Sigma^n$ is a circle. Moreover, since any line L through the origin is determined by just one other point, we see that for any two lines L_1 and L_2 through the origin,

$$L_1 \neq L_2 \quad \text{iff} \quad (L_1 \cap \Sigma^n) \cap (L_2 \cap \Sigma^n) = \emptyset.$$

However, Σ^n is the sphere S^{2n+1} in $\mathbb{R}^{2n+2}$. It follows that $\mathbb{CP}^n$ is the quotient of S^{2n+1} by the equivalence relation $\sim$ defined such that

$$y \sim z \quad \text{iff} \quad y, z \in L \cap \Sigma^n, \quad \text{for some line, } L\text{, through the origin.}$$

Therefore, we can write

$$S^{2n+1}/S^1 \cong \mathbb{CP}^n.$$

The case $n = 1$ is particularly interesting, as it turns out that

$$S^3/S^1 \cong S^2.$$

This is the famous *Hopf fibration*. To show this, proceed as follows: As

$$S^3 \cong \Sigma^1 = \{(z, z') \in \mathbb{C}^2 \mid |z|^2 + |z'|^2 = 1\},$$

define a map, HF: $S^3 \to S^2$, by

$$\mathrm{HF}((z, z')) = (2z\overline{z'}, |z|^2 - |z'|^2).$$

We leave as a homework exercise to prove that this map has range S^2 and that

$$\mathrm{HF}((z_1, z_1')) = \mathrm{HF}((z_2, z_2')) \quad \text{iff} \quad (z_1, z_1') = \lambda(z_2, z_2'), \quad \text{for some } \lambda \text{ with } |\lambda| = 1.$$

In other words, for any point, $p \in S^2$, the inverse image $\mathrm{HF}^{-1}(p)$ (also called *fibre* over p) is a circle on S^3. Consequently, S^3 can be viewed as the union of a family of disjoint circles. This is the *Hopf fibration*. It is possible to visualize the Hopf fibration using the stereographic projection from S^3 onto $\mathbb{R}^3$. This is a beautiful and puzzling picture. For example, see Berger [13]. Therefore, HF induces a bijection from $\mathbb{CP}^1$ to S^2, and it is a homeomorphism.

Example 4.9. Affine spaces.

Let X be a set and E a real vector space. A transitive and faithful action $\cdot\colon E \times X \to X$ of the additive group of E on X makes X into an *affine space*. The intuition is that the members of E are translations.

Those familiar with affine spaces as in Gallier [48] (Chapter 2) or Berger [13] will point out that if X is an affine space, then not only is the action of E on X transitive, but more is true: For any two points $a, b \in E$, there is a *unique* vector $u \in E$, such that $u \cdot a = b$. By the way, the action of E on X is usually considered to be a right action and is written additively, so $u \cdot a$ is written $a + u$ (the result of translating a by u). Thus, it would seem that we have to require more of our action. However, this is not necessary because E (under addition) is *abelian*. More precisely, we have the proposition

Proposition 4.1. *If G is an abelian group acting on a set X and the action $\cdot\colon G \times X \to X$ is transitive and faithful, then for any two elements $x, y \in X$, there is a unique $g \in G$ so that $g \cdot x = y$ (the action is simply transitive).*

Proof. Since our action is transitive, there is at least some $g \in G$ so that $g \cdot x = y$. Assume that we have $g_1, g_2 \in G$ with

$$g_1 \cdot x = g_2 \cdot x = y.$$

We shall prove that

$$g_1 \cdot z = g_2 \cdot z, \quad \text{for all } z \in X.$$

This implies that

$$g_1 g_2^{-1} \cdot z = z, \quad \text{for all } z \in X.$$

As our action is faithful, $g_1 g_2^{-1} = 1$, and we must have $g_1 = g_2$, which proves our proposition.

Pick any $z \in X$. As our action is transitive, there is some $h \in G$ so that $z = h \cdot x$. Then, we have

$$\begin{aligned} g_1 \cdot z &= g_1 \cdot (h \cdot x) \\ &= (g_1 h) \cdot x \\ &= (h g_1) \cdot x \qquad \text{(since } G \text{ is abelian)} \\ &= h \cdot (g_1 \cdot x) \\ &= h \cdot (g_2 \cdot x) \qquad \text{(since} \quad g_1 \cdot x = g_2 \cdot x) \\ &= (h g_2) \cdot x \\ &= (g_2 h) \cdot x \qquad \text{(since } G \text{ is abelian)} \\ &= g_2 \cdot (h \cdot x) \\ &= g_2 \cdot z. \end{aligned}$$

Therefore, $g_1 \cdot z = g_2 \cdot z$ for all $z \in X$, as claimed. □

4.3 Group Actions: Part II, Stabilizers and Homogeneous Spaces

Now that we have an understanding of how a group G acts on a set X, we may use this action to form new topological spaces, namely homogeneous spaces. In the construction of homogeneous spaces, the subset of group elements that leaves some given element $x \in X$ fixed plays an important role.

Definition 4.8. Given an action $\cdot \colon G \times X \to X$ of a group G on a set X, for any $x \in X$, the group G_x (also denoted $\mathrm{Stab}_G(x)$), called the *stabilizer* of x or *isotropy group at* x, is given by

$$G_x = \{g \in G \mid g \cdot x = x\}.$$

We have to verify that G_x is indeed a subgroup of G, but this is easy. Indeed, if $g \cdot x = x$ and $h \cdot x = x$, then we also have $h^{-1} \cdot x = x$ and so, we get $gh^{-1} \cdot x = x$, proving that G_x is a subgroup of G. In general, G_x is **not** a normal subgroup.

Observe that

$$G_{g \cdot x} = g G_x g^{-1},$$

for all $g \in G$ and all $x \in X$. Indeed,

$$\begin{aligned} G_{g \cdot x} &= \{h \in G \mid h \cdot (g \cdot x) = g \cdot x\} \\ &= \{h \in G \mid hg \cdot x = g \cdot x\} \\ &= \{h \in G \mid g^{-1} hg \cdot x = x\}, \end{aligned}$$

which shows $g^{-1}G_{g\cdot x}g \subseteq G_x$, or equivalently that $G_{g\cdot x} \subseteq gG_xg^{-1}$. It remains to show that $gG_xg^{-1} \subseteq G_{g\cdot x}$. Take an element of gG_xg^{-1}, which has the form ghg^{-1} with $h \cdot x = x$. Since $h \cdot x = x$, we have $(ghg^{-1}) \cdot gx = gx$, which shows that $ghg^{-1} \in G_{g\cdot x}$.

Because $G_{g\cdot x} = gG_xg^{-1}$, the stabilizers of x and $g \cdot x$ are conjugate of each other.

When the action of G on X is transitive, for any fixed $x \in G$, the set X is a quotient (as a set, not as group) of G by G_x. Indeed, we can define the map, $\pi_x \colon G \to X$, by

$$\pi_x(g) = g \cdot x, \quad \text{for all } g \in G.$$

Observe that

$$\pi_x(gG_x) = (gG_x) \cdot x = g \cdot (G_x \cdot x) = g \cdot x = \pi_x(g).$$

This shows that $\pi_x \colon G \to X$ induces a quotient map $\overline{\pi}_x \colon G/G_x \to X$, from the set G/G_x of (left) cosets of G_x to X, defined by

$$\overline{\pi}_x(gG_x) = g \cdot x.$$

Since

$$\pi_x(g) = \pi_x(h) \quad \text{iff} \quad g \cdot x = h \cdot x \quad \text{iff} \quad g^{-1}h \cdot x = x \quad \text{iff} \quad g^{-1}h \in G_x \quad \text{iff} \quad gG_x = hG_x,$$

we deduce that $\overline{\pi}_x \colon G/G_x \to X$ is injective. However, since our action is transitive, for every $y \in X$, there is some $g \in G$ so that $g \cdot x = y$, and so $\overline{\pi}_x(gG_x) = g \cdot x = y$; that is, the map $\overline{\pi}_x$ is also surjective. Therefore, the map $\overline{\pi}_x \colon G/G_x \to X$ is a bijection (of sets, not groups). The map $\pi_x \colon G \to X$ is also surjective. Let us record this important fact as

Proposition 4.2. *If $\cdot \colon G \times X \to X$ is a transitive action of a group G on a set X, for every fixed $x \in X$, the surjection $\pi_x \colon G \to X$ given by*

$$\pi_x(g) = g \cdot x$$

induces a bijection

$$\overline{\pi}_x \colon G/G_x \to X,$$

where G_x is the stabilizer of x. See Figure 4.4.

The map $\pi_x \colon G \to X$ (corresponding to a fixed $x \in X$) is sometimes called a *projection* of G onto X. Proposition 4.2 shows that for every $y \in X$, the subset $\pi_x^{-1}(y)$ of G (called the *fibre above* y) is equal to some coset gG_x of G, and thus is in bijection with the group G_x itself. We can think of G as a moving family of fibres G_x parametrized by X. This point of view of viewing a space as a moving family of simpler spaces is typical in (algebraic) geometry, and underlies the notion of (principal) fibre bundle.

Note that if the action $\cdot \colon G \times X \to X$ is transitive, then the stabilizers G_x and G_y of any two elements $x, y \in X$ are isomorphic, as they are conjugates. Thus, in this case, it is enough to compute one of these stabilizers for a "convenient" x.

As the situation of Proposition 4.2 is of particular interest, we make the following definition:

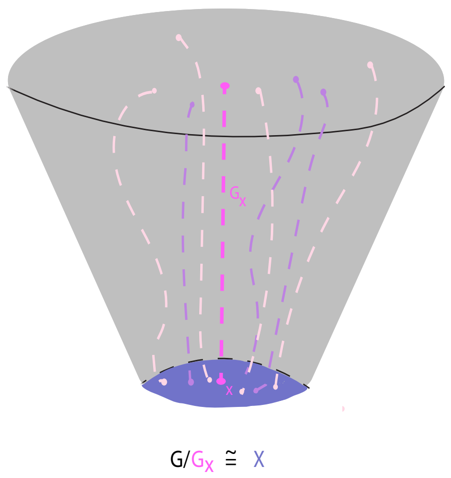


Figure 4.4: A schematic representation of $G/G_x \cong X$, where G is the gray solid, X is its purple circular base, and G_x is the pink vertical strand. The dotted strands are the fibres gG_x.

Definition 4.9. A set X is said to be a *homogeneous space* if there is a transitive action $\cdot\colon G \times X \to X$ of some group G on X.

We see that all the spaces of Examples 4.2–4.9, are homogeneous spaces. Another example that will play an important role when we deal with Lie groups is the situation where we have a group G, a subgroup H of G (not necessarily normal), and where $X = G/H$, the set of left cosets of G modulo H. The group G acts on G/H by left multiplication:

$$a \cdot (gH) = (ag)H,$$

where $a, g \in G$. This action is clearly transitive and one checks that the stabilizer of gH is gHg^{-1}. If G is a topological group and H is a closed subgroup of G (see later for an explanation), it turns out that G/H is Hausdorff. If G is a Lie group, we obtain a manifold.

Even if G and X are topological spaces and the action $\cdot\colon G \times X \to X$ is continuous, in general, the space G/G_x under the quotient topology is **not** homeomorphic to X.

We will give later sufficient conditions that insure that X is indeed a topological space or even a manifold. In particular, X will be a manifold when G is a Lie group.

In general, an action $\cdot\colon G \times X \to X$ is not transitive on X, but for every $x \in X$, it is transitive on the set

$$O(x) = G \cdot x = \{g \cdot x \mid g \in G\}.$$

Such a set is called the *orbit* of x. The orbits are the equivalence classes of the following equivalence relation:

Definition 4.10. Given an action $\cdot\colon G \times X \to X$ of some group G on X, the equivalence relation $\sim$ on X is defined so that, for all $x, y \in X$,

$$x \sim y \quad \text{iff} \quad y = g \cdot x, \quad \text{for some } g \in G.$$

For every $x \in X$, the equivalence class of x is the *orbit of* x, denoted $O(x)$ or $G \cdot x$, with

$$G \cdot x = O(x) = \{g \cdot x \mid g \in G\}.$$

The set of orbits is denoted X/G.

We warn the reader that some authors use the notation $G\backslash X$ for the the set of orbits $G \cdot x$, because these orbits can be considered as right orbits, by analogy with right cosets Hg of a subgroup H of G.

The orbit space X/G is obtained from X by an identification (or merging) process: For every orbit, all points in that orbit are merged into a single point. This akin to the process of forming the identification topology. For example, if $X = S^2$ and G is the group consisting of the restrictions of the two linear maps I and $-I$ of $\mathbb{R}^3$ to S^2 (where $(-I)(x) = -x$ for all $x \in \mathbb{R}^3$), then

$$X/G = S^2/\{I, -I\} \cong \mathbb{RP}^2.$$

See Figure 4.3. More generally, if S^n is the n-sphere in $\mathbb{R}^{n+1}$, then we have a bijection between the orbit space $S^n/\{I, -I\}$ and $\mathbb{RP}^n$:

$$S^n/\{I, -I\} \cong \mathbb{RP}^n.$$

Many manifolds can be obtained in this fashion, including the torus, the Klein bottle, the Möbius band, *etc.*

Since the action of G is transitive on $O(x)$, by Proposition 4.2, we see that for every $x \in X$, we have a bijection

$$O(x) \cong G/G_x.$$

As a corollary, if both X and G are finite, for any set $A \subseteq X$ of representatives from every orbit, we have the *orbit formula*:

$$|X| = \sum_{a \in A} [G\colon G_a] = \sum_{a \in A} |G|/|G_a|.$$

Even if a group action $\cdot\colon G \times X \to X$ is not transitive, when X is a manifold, we can consider the set of orbits X/G, and if the action of G on X satisfies certain conditions, X/G is actually a manifold. Manifolds arising in this fashion are often called *orbifolds*. In summary, we see that manifolds arise in at least two ways from a group action:

(1) As homogeneous spaces G/G_x, if the action is transitive.

(2) As orbifolds X/G (under certain conditions on the action).

Of course, in both cases, the action must satisfy some additional properties.

For the rest of this section, we reconsider Examples 4.2–4.9 in the context of homogeneous space by determining some stabilizers for those actions.

(a) Consider the action $\cdot \colon \mathbf{SO}(n) \times S^{n-1} \to S^{n-1}$ of $\mathbf{SO}(n)$ on the sphere S^{n-1} $(n \geq 1)$ defined in Example 4.2. Since this action is transitive, we can determine the stabilizer of any convenient element of S^{n-1}, say $e_1 = (1, 0, \ldots, 0)$. In order for any $R \in \mathbf{SO}(n)$ to leave e_1 fixed, the first column of R must be e_1, so R is an orthogonal matrix of the form

$$R = \begin{pmatrix} 1 & U \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

where U is a $1 \times (n-1)$ row vector. As the rows of R must be unit vectors, we see that $U = 0$ and $S \in \mathbf{SO}(n-1)$. Therefore, the stabilizer of e_1 is isomorphic to $\mathbf{SO}(n-1)$, and we deduce the bijection

$$\mathbf{SO}(n)/\mathbf{SO}(n-1) \cong S^{n-1}.$$

Strictly speaking, $\mathbf{SO}(n-1)$ is not a subgroup of $\mathbf{SO}(n)$, and in all rigor, we should consider the subgroup $\widetilde{\mathbf{SO}}(n-1)$ of $\mathbf{SO}(n)$ consisting of all matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \det(S) = 1,$$

and write

$$\mathbf{SO}(n)/\widetilde{\mathbf{SO}}(n-1) \cong S^{n-1}.$$

However, it is common practice to identify $\mathbf{SO}(n-1)$ with $\widetilde{\mathbf{SO}}(n-1)$.

When $n = 2$, as $\mathbf{SO}(1) = \{1\}$, we find that $\mathbf{SO}(2) \cong S^1$, a circle, a fact that we already knew. When $n = 3$, we find that $\mathbf{SO}(3)/\mathbf{SO}(2) \cong S^2$. This says that $\mathbf{SO}(3)$ is somehow the result of glueing circles to the surface of a sphere (in $\mathbb{R}^3$), in such a way that these circles do not intersect. This is hard to visualize!

A similar argument for the complex unit sphere Σ^{n-1} shows that

$$\mathbf{SU}(n)/\mathbf{SU}(n-1) \cong \Sigma^{n-1} \cong S^{2n-1}.$$

Again, we identify $\mathbf{SU}(n-1)$ with a subgroup of $\mathbf{SU}(n)$, as in the real case. In particular, when $n = 2$, as $\mathbf{SU}(1) = \{1\}$, we find that

$$\mathbf{SU}(2) \cong S^3;$$

that is, the group $\mathbf{SU}(2)$ is topologically the sphere S^3! Actually, this is not surprising if we remember that $\mathbf{SU}(2)$ is in fact the group of unit quaternions.

(b) We saw in Example 4.3 that the action $\cdot\colon \mathbf{SL}(2,\mathbb{R}) \times H \to H$ of the group $\mathbf{SL}(2,\mathbb{R})$ on the upper half plane is transitive. Let us find out what the stabilizer of $z = i$ is. We should have

$$\frac{ai+b}{ci+d} = i,$$

that is, $ai + b = -c + di$, i.e.,

$$(d-a)i = b + c.$$

Since a, b, c, d are real, we must have $d = a$ and $b = -c$. Moreover, $ad - bc = 1$, so we get $a^2 + b^2 = 1$. We conclude that a matrix in $\mathbf{SL}(2,\mathbb{R})$ fixes i iff it is of the form

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad \text{with} \quad a^2 + b^2 = 1.$$

Clearly, these are the rotation matrices in $\mathbf{SO}(2)$, and so the stabilizer of i is $\mathbf{SO}(2)$. We conclude that

$$\mathbf{SL}(2,\mathbb{R})/\mathbf{SO}(2) \cong H.$$

This time we can view $\mathbf{SL}(2,\mathbb{R})$ as the result of glueing circles to the upper half plane. This is not so easy to visualize. There is a better way to visualize the topology of $\mathbf{SL}(2,\mathbb{R})$ by making it act on the open disk D. We will return to this action in a little while.

(c) Now consider the action of $\mathbf{SL}(2,\mathbb{C})$ on $\mathbb{C} \cup \{\infty\} \cong S^2$ given in Example 4.4. As it is transitive, let us find the stabilizer of $z = 0$. We must have

$$\frac{b}{d} = 0,$$

and as $ad - bc = 1$, we must have $b = 0$ and $ad = 1$. Thus the stabilizer of 0 is the subgroup $\mathbf{SL}(2,\mathbb{C})_0$ of $\mathbf{SL}(2,\mathbb{C})$ consisting of all matrices of the form

$$\begin{pmatrix} a & 0 \\ c & a^{-1} \end{pmatrix}, \quad \text{where} \quad a \in \mathbb{C} - \{0\} \quad \text{and} \quad c \in \mathbb{C}.$$

We get

$$\mathbf{SL}(2,\mathbb{C})/\mathbf{SL}(2,\mathbb{C})_0 \cong \mathbb{C} \cup \{\infty\} \cong S^2,$$

but this is not very illuminating.

(d) In Example 4.7 we considered the action $\cdot\colon \mathbf{GL}(n) \times \mathbf{SPD}(n) \to \mathbf{SPD}(n)$ of $\mathbf{GL}(n)$ on $\mathbf{SPD}(n)$, the set of symmetric positive definite matrices. As this action is transitive, let us find the stabilizer of I. For any $A \in \mathbf{GL}(n)$, the matrix A stabilizes I iff

$$AIA^\top = AA^\top = I.$$

Therefore the stabilizer of I is $\mathbf{O}(n)$, and we find that

$$\mathbf{GL}(n)/\mathbf{O}(n) = \mathbf{SPD}(n).$$

Observe that if $\mathbf{GL}^+(n)$ denotes the subgroup of $\mathbf{GL}(n)$ consisting of all matrices with a strictly positive determinant, then we have an action $\cdot\colon \mathbf{GL}^+(n) \times \mathbf{SPD}(n) \to \mathbf{SPD}(n)$ of $\mathbf{GL}^+(n)$ on $\mathbf{SPD}(n)$. This action is transitive and we find that the stabilizer of I is $\mathbf{SO}(n)$; consequently, we get

$$\mathbf{GL}^+(n)/\mathbf{SO}(n) = \mathbf{SPD}(n).$$

(e) In Example 4.8 we considered the action $\cdot\colon \mathbf{SO}(n+1) \times \mathbb{RP}^n \to \mathbb{RP}^n$ of $\mathbf{SO}(n+1)$ on the (real) projective space $\mathbb{RP}^n$. As this action is transitive, let us find the stabilizer of the line $L = [e_1]$, where $e_1 = (1, 0, \ldots, 0)$. For any $R \in \mathbf{SO}(n+1)$, the line L is fixed iff either $R(e_1) = e_1$ or $R(e_1) = -e_1$, since e_1 and $-e_1$ define the same line. As R is orthogonal with $\det(R) = 1$, this means that R is of the form

$$R = \begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha.$$

But, S must be orthogonal, so we conclude $S \in \mathbf{O}(n)$. Therefore, the stabilizer of $L = [e_1]$ is isomorphic to the group $\mathbf{O}(n)$, and we find that

$$\mathbf{SO}(n+1)/\mathbf{O}(n) \cong \mathbb{RP}^n.$$

Strictly speaking, $\mathbf{O}(n)$ is not a subgroup of $\mathbf{SO}(n+1)$, so the above equation does not make sense. We should write

$$\mathbf{SO}(n+1)/\widetilde{\mathbf{O}}(n) \cong \mathbb{RP}^n,$$

where $\widetilde{\mathbf{O}}(n)$ is the subgroup of $\mathbf{SO}(n+1)$ consisting of all matrices of the form

$$\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad S \in \mathbf{O}(n),\ \alpha = \pm 1 \quad \text{and} \quad \det(S) = \alpha.$$

This group is also denoted $S(\mathbf{O}(1) \times \mathbf{O}(n))$. However, the common practice is to write $\mathbf{O}(n)$ instead of $S(\mathbf{O}(1) \times \mathbf{O}(n))$.

We should mention that $\mathbb{RP}^3$ and $\mathbf{SO}(3)$ are homeomorphic spaces. This is shown using the quaternions; for example, see Gallier [48], Chapter 8.

A similar argument applies to the action $\cdot\colon \mathbf{SU}(n+1) \times \mathbb{CP}^n \to \mathbb{CP}^n$ of $\mathbf{SU}(n+1)$ on the (complex) projective space $\mathbb{CP}^n$. We find that

$$\mathbf{SU}(n+1)/\mathbf{U}(n) \cong \mathbb{CP}^n.$$

Again, the above is a bit sloppy as $\mathbf{U}(n)$ is not a subgroup of $\mathbf{SU}(n+1)$. To be rigorous, we should use the subgroup $\widetilde{\mathbf{U}}(n)$ consisting of all matrices of the form

$$\begin{pmatrix} \alpha & 0 \\ 0 & S \end{pmatrix}, \quad \text{with} \quad S \in \mathbf{U}(n),\ |\alpha| = 1 \quad \text{and} \quad \det(S) = \overline{\alpha}.$$

This group is also denoted $S(\mathbf{U}(1) \times \mathbf{U}(n))$. The common practice is to write $\mathbf{U}(n)$ instead of $S(\mathbf{U}(1) \times \mathbf{U}(n))$. In particular, when $n = 1$, we find that

$$\mathbf{SU}(2)/\mathbf{U}(1) \cong \mathbb{CP}^1.$$

But, we know that $\mathbf{SU}(2) \cong S^3$, and clearly $\mathbf{U}(1) \cong S^1$. So, again, we find that $S^3/S^1 \cong \mathbb{CP}^1$ (we know more, namely, $S^3/S^1 \cong S^2 \cong \mathbb{CP}^1$.)

Observe that $\mathbb{CP}^n$ can also be viewed as the orbit space of the action $\cdot\colon S^1 \times S^{2n+1} \to S^{2n+1}$ given by

$$\lambda \cdot (z_1, \ldots, z_{n+1}) = (\lambda z_1, \ldots, \lambda z_{n+1}),$$

where $S^1 = \mathbf{U}(1)$ (the group of complex numbers of modulus 1) and S^{2n+1} is identified with Σ^n.

We now return to Case (b) to give a better picture of $\mathbf{SL}(2, \mathbb{R})$. Instead of having $\mathbf{SL}(2, \mathbb{R})$ act on the upper half plane, we define an action of $\mathbf{SL}(2, \mathbb{R})$ on the open unit disk D as we did in Example 4.5. Technically, it is easier to consider the group $\mathbf{SU}(1, 1)$, which is isomorphic to $\mathbf{SL}(2, \mathbb{R})$, and to make $\mathbf{SU}(1, 1)$ act on D. The group $\mathbf{SU}(1, 1)$ is the group of 2×2 complex matrices of the form

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}, \quad \text{with} \quad a\overline{a} - b\overline{b} = 1.$$

The reader should check that if we let

$$g = \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix},$$

then the map from $\mathbf{SL}(2, \mathbb{R})$ to $\mathbf{SU}(1, 1)$ given by

$$A \mapsto gAg^{-1}$$

is an isomorphism. Observe that the scaled Möbius transformation associated with g is

$$z \mapsto \frac{z - i}{z + i},$$

which is the holomorphic isomorphism mapping H to D mentionned earlier! We can define a bijection between $\mathbf{SU}(1, 1)$ and $S^1 \times D$ given by

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \mapsto (a/|a|, b/a).$$

We conclude that $\mathbf{SL}(2, \mathbb{R}) \cong \mathbf{SU}(1, 1)$ is topologically an open solid torus (i.e., with the surface of the torus removed). It is possible to further classify the elements of $\mathbf{SL}(2, \mathbb{R})$ into three categories and to have geometric interpretations of these as certain regions of the torus.

For details, the reader should consult Carter, Segal and Macdonald [29] or Duistermatt and Kolk [43] (Chapter 1, Section 1.2).

The group $\mathbf{SU}(1,1)$ acts on D by interpreting any matrix in $\mathbf{SU}(1,1)$ as a Möbius tranformation; that is,

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \mapsto \left(z \mapsto \frac{az+b}{\overline{b}z+\overline{a}} \right).$$

The reader should check that these transformations preserve D.

Both the upper half-plane and the open disk are models of Lobachevsky's non-Euclidean geometry (where the parallel postulate fails). They are also models of hyperbolic spaces (Riemannian manifolds with constant negative curvature, see Gallot, Hulin and Lafontaine [49], Chapter III). According to Dubrovin, Fomenko, and Novikov [41] (Chapter 2, Section 13.2), the open disk model is due to Poincaré and the upper half-plane model to Klein, although Poincaré was the first to realize that the upper half-plane is a hyperbolic space.

4.4 The Grassmann and Stiefel Manifolds

In this section we introduce two very important homogeneous manifolds, the Grassmann manifolds and the Stiefel manifolds. The Grassmann manifolds are generalizations of projective spaces (real and complex), while the Stiefel manifold are generalizations of $\mathbf{O}(n)$. Both of these manifolds are examples of reductive homogeneous spaces; see Chapter 22. We begin by defining the Grassmann manifolds $G(k,n)$.

First consider the real case.

Definition 4.11. Given any $n \geq 1$, for any k with $0 \leq k \leq n$, the set $G(k,n)$ of all linear k-dimensional subspaces of $\mathbb{R}^n$ (also called *k-planes*) is called a *Grassmannian* (or *Grassmann manifold*).

Any k-dimensional subspace U of $\mathbb{R}^n$ is spanned by k linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$; write $U = \mathrm{span}(u_1, \ldots, u_k)$. We can define an action $\cdot\colon \mathbf{O}(n) \times G(k,n) \to G(k,n)$ as follows: For any $R \in \mathbf{O}(n)$, for any $U = \mathrm{span}(u_1, \ldots, u_k)$, let

$$R \cdot U = \mathrm{span}(Ru_1, \ldots, Ru_k).$$

We have to check that the above is well defined. If $U = \mathrm{span}(v_1, \ldots, v_k)$ for any other k linearly independent vectors $v_1, \ldots, v_k$, we have

$$v_i = \sum_{j=1}^{k} a_{ij} u_j, \quad 1 \leq i \leq k,$$

for some $a_{ij} \in \mathbb{R}$, and so

$$Rv_i = \sum_{j=1}^{k} a_{ij} Ru_j, \quad 1 \leq i \leq k,$$

which shows that

$$\mathrm{span}(Ru_1, \ldots, Ru_k) = \mathrm{span}(Rv_1, \ldots, Rv_k);$$

that is, the above action is well defined.

We claim this action is transitive. This is because if U and V are any two k-planes, we may assume that $U = \mathrm{span}(u_1, \ldots, u_k)$ and $V = \mathrm{span}(v_1, \ldots, v_k)$, where the u_i's form an orthonormal family and similarly for the v_i's. Then we can extend these families to orthonormal bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ on $\mathbb{R}^n$, and w.r.t. the orthonormal basis $(u_1, \ldots, u_n)$, the matrix of the linear map sending u_i to v_i is orthogonal. Hence $G(k, n)$ is a homogeneous space.

In order to represent $G(k, n)$ as a quotient space, Proposition 4.2 implies it is enough to find the stabilizer of any k-plane. Pick $U = \mathrm{span}(e_1, \ldots, e_k)$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$ (i.e., $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, with the 1 in the ith position). Any $R \in \mathbf{O}(n)$ stabilizes U iff R maps $e_1, \ldots, e_k$ to k linearly independent vectors in the subspace $U = \mathrm{span}(e_1, \ldots, e_k)$, i.e., R is of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

where S is $k \times k$ and T is $(n-k) \times (n-k)$. Moreover, as R is orthogonal, S and T must be orthogonal, that is $S \in \mathbf{O}(k)$ and $T \in \mathbf{O}(n-k)$. We deduce that the stabilizer of U is isomorphic to $\mathbf{O}(k) \times \mathbf{O}(n-k)$ and we find that

$$\mathbf{O}(n)/(\mathbf{O}(k) \times \mathbf{O}(n-k)) \cong G(k, n).$$

It turns out that this makes $G(k, n)$ into a smooth manifold of dimension

$$\frac{n(n-1)}{2} - \frac{k(k-1)}{2} - \frac{(n-k)(n-k-1)}{2} = k(n-k)$$

called a *Grassmannian*.

The restriction of the action of $\mathbf{O}(n)$ on $G(k, n)$ to $\mathbf{SO}(n)$ yields an action $\cdot\colon \mathbf{SO}(n) \times G(k, n) \to G(k, n)$ of $\mathbf{SO}(n)$ on $G(k, n)$. Then it is easy to see that this action is transitive and that the stabilizer of the subspace U is isomorphic to the subgroup $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$ of $\mathbf{SO}(n)$ consisting of the rotations of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in \mathbf{O}(k)$, $T \in \mathbf{O}(n-k)$ and $\det(S)\det(T) = 1$. Thus, we also have

$$\mathbf{SO}(n)/S(\mathbf{O}(k) \times \mathbf{O}(n-k)) \cong G(k, n).$$

If we recall the projection map of Example 4.8 in Section 4.2, namely $pr\colon \mathbb{R}^{n+1} - \{0\} \to \mathbb{RP}^n$, by definition, a *k-plane* in $\mathbb{RP}^n$ is the image under pr of any $(k+1)$-plane in $\mathbb{R}^{n+1}$.

So, for example, a line in $\mathbb{RP}^n$ is the image of a 2-plane in $\mathbb{R}^{n+1}$, and a hyperplane in $\mathbb{RP}^n$ is the image of a hyperplane in $\mathbb{R}^{n+1}$. The advantage of this point of view is that the k-planes in $\mathbb{RP}^n$ are arbitrary; that is, they do not have to go through "the origin" (which does not make sense, anyway!). Then we see that we can interpret the Grassmannian, $G(k+1, n+1)$, as a space of "parameters" for the k-planes in $\mathbb{RP}^n$. For example, $G(2, n+1)$ parametrizes the lines in $\mathbb{RP}^n$. In this viewpoint, $G(k+1, n+1)$ is usually denoted $\mathbb{G}(k, n)$.

It can be proved (using some exterior algebra) that $G(k, n)$ can be embedded in $\mathbb{RP}^{\binom{n}{k}-1}$. Much more is true. For example, $G(k, n)$ is a projective variety, which means that it can be defined as a subset of $\mathbb{RP}^{\binom{n}{k}-1}$ equal to the zero locus of a set of homogeneous equations. There is even a set of quadratic equations known as the *Plücker equations* defining $G(k, n)$. In particular, when $n = 4$ and $k = 2$, we have $G(2, 4) \subseteq \mathbb{RP}^5$, and $G(2, 4)$ is defined by a single equation of degree 2. The Grassmannian $G(2, 4) = \mathbb{G}(1, 3)$ is known as the *Klein quadric*. This hypersurface in $\mathbb{RP}^5$ parametrizes the lines in $\mathbb{RP}^3$.

Complex Grassmannians are defined in a similar way, by replacing $\mathbb{R}$ by $\mathbb{C}$ and $\mathbf{O}(n)$ by $\mathbf{U}(n)$ throughout. The complex Grassmannian $G_{\mathbb{C}}(k, n)$ is a complex manifold as well as a real manifold, and we have

$$\mathbf{U}(n)/(\mathbf{U}(k) \times \mathbf{U}(n-k)) \cong G_{\mathbb{C}}(k, n).$$

As in the case of the real Grassmannians, the action of $\mathbf{U}(n)$ on $G_{\mathbb{C}}(k, n)$ yields an action of $\mathbf{SU}(n)$ on $G_{\mathbb{C}}(k, n)$, and we get

$$\mathbf{SU}(n)/S(\mathbf{U}(k) \times \mathbf{U}(n-k)) \cong G_{\mathbb{C}}(k, n),$$

where $S(\mathbf{U}(k) \times \mathbf{U}(n-k))$ is the subgroup of $\mathbf{SU}(n)$ consisting of all matrices $R \in \mathbf{SU}(n)$ of the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

with $S \in \mathbf{U}(k)$, $T \in \mathbf{U}(n-k)$ and $\det(S)\det(T) = 1$.

Closely related to Grassmannians are the *Stiefel manifolds* $S(k, n)$. Again we begin with the real case.

Definition 4.12. For any $n \geq 1$ and any k with $1 \leq k \leq n$, the set $S(k, n)$ of all orthonormal k-frames, that is, of k-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$, is called a *Stiefel manifold*.

Obviously, $S(1, n) = S^{n-1}$ and $S(n, n) = \mathbf{O}(n)$, so assume $k \leq n-1$. There is a natural action $\cdot \colon \mathbf{SO}(n) \times S(k, n) \to S(k, n)$ of $\mathbf{SO}(n)$ on $S(k, n)$ given by

$$R \cdot (u_1, \ldots, u_k) = (Ru_1, \ldots, Ru_k).$$

This action is transitive, because if $(u_1, \ldots, u_k)$ and $(v_1, \ldots, v_k)$ are any two orthonormal k-frames, then they can be extended to orthonormal bases (for example, by Gram-Schmidt)

$(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_n)$ with the same orientation (since we can pick u_n and v_n so that our bases have the same orientation), and there is a unique orthogonal transformation $R \in \mathbf{SO}(n)$ such that $Ru_i = v_i$ for $i = 1, \ldots, n$.

In order to apply Proposition 4.2, we need to find the stabilizer of the orthonormal k-frame $(e_1, \ldots, e_k)$ consisting of the first canonical basis vectors of $\mathbb{R}^n$. A matrix $R \in \mathbf{SO}(n)$ stabilizes $(e_1, \ldots, e_k)$ iff it is of the form

$$R = \begin{pmatrix} I_k & 0 \\ 0 & S \end{pmatrix}$$

where $S \in \mathbf{SO}(n-k)$. Therefore, for $1 \leq k \leq n-1$, we have

$$\mathbf{SO}(n)/\mathbf{SO}(n-k) \cong S(k, n).$$

This makes $S(k, n)$ a smooth manifold of dimension

$$\frac{n(n-1)}{2} - \frac{(n-k)(n-k-1)}{2} = nk - \frac{k(k+1)}{2} = k(n-k) + \frac{k(k-1)}{2}.$$

Remark: It should be noted that we can define another type of Stiefel manifolds, denoted by $V(k, n)$, using linearly independent k-tuples $(u_1, \ldots, u_k)$ that do not necessarily form an orthonormal system. In this case, there is an action $\cdot \colon \mathbf{GL}(n, \mathbb{R}) \times V(k, n) \to V(k, n)$, and the stabilizer H of the first k canonical basis vectors $(e_1, \ldots, e_k)$ is a closed subgroup of $\mathbf{GL}(n, \mathbb{R})$, but it doesn't have a simple description (see Warner [114], Chapter 3). We get an isomorphism

$$V(k, n) \cong \mathrm{GL}(n, \mathbb{R})/H.$$

The version of the Stiefel manifold $S(k, n)$ using orthonormal frames is sometimes denoted by $V^0(k, n)$ (Milnor and Stasheff [85] use the notation $V_k^0(\mathbb{R}^n)$). Beware that the notation is not standardized. Certain authors use $V(k, n)$ for what we denote by $S(k, n)$!

Complex Stiefel manifolds are defined in a similar way by replacing $\mathbb{R}$ by $\mathbb{C}$ and $\mathbf{SO}(n)$ by $\mathbf{SU}(n)$. For $1 \leq k \leq n-1$, the complex Stiefel manifold $S_{\mathbb{C}}(k, n)$ is isomorphic to the quotient

$$\mathbf{SU}(n)/\mathbf{SU}(n-k) \cong S_{\mathbb{C}}(k, n).$$

If $k = 1$, we have $S_{\mathbb{C}}(1, n) = S^{2n-1}$, and if $k = n$, we have $S_{\mathbb{C}}(n, n) = \mathbf{U}(n)$.

The Grassmannians can also be viewed as quotient spaces of the Stiefel manifolds. Every orthonomal k-frame $(u_1, \ldots, u_k)$ can be represented by an $n \times k$ matrix Y over the canonical basis of $\mathbb{R}^n$, and such a matrix Y satisfies the equation

$$Y^\top Y = I.$$

We have a right action $\cdot \colon S(k, n) \times \mathbf{O}(k) \to S(k, n)$ given by

$$Y \cdot R = YR,$$

for any $R \in \mathbf{O}(k)$. This action is well defined since

$$(YR)^\top YR = R^\top Y^\top YR = I.$$

However, this action is not transitive (unless $k = 1$), but the orbit space $S(k,n)/\mathbf{O}(k)$ is isomorphic to the Grassmannian $G(k,n)$, so we can write

$$G(k,n) \cong S(k,n)/\mathbf{O}(k).$$

Similarly, the complex Grassmannian is isomorphic to the orbit space $S_{\mathbb{C}}(k,n)/\mathbf{U}(k)$:

$$G_{\mathbb{C}}(k,n) \cong S_{\mathbb{C}}(k,n)/\mathbf{U}(k).$$

4.5 Topological Groups ⊛

Since Lie groups are topological groups (and manifolds), it is useful to gather a few basic facts about topological groups.

Definition 4.13. A set G is a *topological group* iff

(a) G is a Hausdorff topological space;

(b) G is a group (with identity 1);

(c) Multiplication $\cdot\colon G \times G \to G$, and the inverse operation $G \longrightarrow G\colon g \mapsto g^{-1}$, are continuous, where $G \times G$ has the product topology.

It is easy to see that the two requirements of Condition (c) are equivalent to

(c′) The map $G \times G \longrightarrow G\colon (g,h) \mapsto gh^{-1}$ is continuous.

Proposition 4.3. *If G is a topological group and H is any subgroup of G, then the closure $\overline{H}$ of H is a subgroup of G.*

Proof. We use the fact that if $f\colon X \to Y$ is a continuous map between two topological spaces X and Y, then $f(\overline{A}) \subseteq \overline{f(A)}$ for any subset A of X. For any $a \in \overline{A}$, we need to show that for any open subset $W \subseteq Y$ containing $f(a)$, we have $W \cap f(A) \neq \emptyset$. Since f is continuous, $V = f^{-1}(W)$ is an open subset containing a, and since $a \in \overline{A}$, we have $f^{-1}(W) \cap A \neq \emptyset$, so there is some $x \in f^{-1}(W) \cap A$, which implies that $f(x) \in W \cap f(A)$, so $W \cap f(A) \neq \emptyset$, as desired. The map $f\colon G \times G \to G$ given by $f(x,y) = xy^{-1}$ is continuous, and since H is a subgroup of G, $f(H \times H) \subseteq H$. By the above property, if $a \in \overline{H}$ and if $b \in \overline{H}$, that is, $(a,b) \in \overline{H \times H}$, then $f(a,b) = ab^{-1} \in \overline{H}$, which shows that $\overline{H}$ is a subgroup of G. □

Given a topological group G, for every $a \in G$ we define the *left translation* L_a as the map $L_a \colon G \to G$ such that $L_a(b) = ab$, for all $b \in G$, and the *right translation* R_a as the map $R_a \colon G \to G$ such that $R_a(b) = ba$, for all $b \in G$. Observe that $L_{a^{-1}}$ is the inverse of L_a and similarly, $R_{a^{-1}}$ is the inverse of R_a. As multiplication is continuous, we see that L_a and R_a are continuous. Moreover, since they have a continuous inverse, they are homeomorphisms. As a consequence, if U is an open subset of G, then so is $gU = L_g(U)$ (resp. $Ug = R_gU$), for all $g \in G$. Therefore, the topology of a topological group is *determined* by the knowledge of the open subsets containing the identity 1.

Given any subset $S \subseteq G$, let $S^{-1} = \{s^{-1} \mid s \in S\}$; let $S^0 = \{1\}$, and $S^{n+1} = S^nS$, for all $n \geq 0$. Property (c) of Definition 4.13 has the following useful consequences, which shows there exists an open set containing 1 which has a special symmetrical structure.

Proposition 4.4. *If G is a topological group and U is any open subset containing* 1*, then there is some open subset $V \subseteq U$, with $1 \in V$, so that $V = V^{-1}$ and $V^2 \subseteq U$. Furthermore, $\overline{V} \subseteq U$.*

Proof. Since multiplication $G \times G \longrightarrow G$ is continuous and $G \times G$ is given the product topology, there are open subsets U_1 and U_2, with $1 \in U_1$ and $1 \in U_2$, so that $U_1U_2 \subseteq U$. Let $W = U_1 \cap U_2$ and $V = W \cap W^{-1}$. Then V is an open set containing 1, and clearly $V = V^{-1}$ and $V^2 \subseteq U_1U_2 \subseteq U$. If $g \in \overline{V}$, then gV is an open set containing g (since $1 \in V$) and thus, $gV \cap V \neq \emptyset$. This means that there are some $h_1, h_2 \in V$ so that $gh_1 = h_2$, but then, $g = h_2h_1^{-1} \in VV^{-1} = VV \subseteq U$. □

Definition 4.14. A subset U containing 1 and such that $U = U^{-1}$ is called *symmetric*.

Proposition 4.4 is used in the proofs of many the propositions and theorems on the structure of topological groups. For example, it is key in verifying the following proposition regarding discrete topological subgroups.

Definition 4.15. A subgroup H of a topological group G is *discrete* iff the induced topology on H is discrete; that is, for every $h \in H$, there is some open subset U of G so that $U \cap H = \{h\}$.

Proposition 4.5. *If G is a topological group and H is a discrete subgroup of G, then H is closed.*

Proof. As H is discrete, there is an open subset U of G so that $U \cap H = \{1\}$, and by Proposition 4.4, we may assume that $U = U^{-1}$. Our goal is to show $H = \overline{H}$. Clearly $H \subseteq \overline{H}$. Thus it remains to show $\overline{H} \subseteq H$. If $g \in \overline{H}$, as gU is an open set containing g, we have $gU \cap H \neq \emptyset$. Consequently, there is some $y \in gU \cap H = gU^{-1} \cap H$, so $g \in yU$ with $y \in H$. We claim that $yU \cap H = \{y\}$. Note that $x \in yU \cap H$ means $x = yu_1$ with $yu_1 \in H$ and $u_1 \in U$. Since H is a subgroup of G and $y \in H$, $y^{-1}yu_1 = u_1 \in H$. Thus $u_1 \in U \cap H$, which implies $u_1 = 1$ and $x = yu_1 = y$, and we have

$$g \in yU \cap \overline{H} \subseteq \overline{yU \cap H} = \overline{\{y\}} = \{y\}.$$

since G is Hausdorff. Therefore, $g = y \in H$. □

Using Proposition 4.4, we can give a very convenient characterization of the Hausdorff separation property in a topological group.

Proposition 4.6. *If G is a topological group, then the following properties are equivalent:*

(1) G is Hausdorff;

(2) The set $\{1\}$ is closed;

(3) The set $\{g\}$ is closed, for every $g \in G$.

Proof. The implication (1) $\longrightarrow$ (2) is true in any Hausdorff topological space. We just have to prove that $G - \{1\}$ is open, which goes as follows: For any $g \neq 1$, since G is Hausdorff, there exists disjoint open subsets U_g and V_g, with $g \in U_g$ and $1 \in V_g$. Thus, $\bigcup U_g = G - \{1\}$, showing that $G - \{1\}$ is open. Since L_g is a homeomorphism, (2) and (3) are equivalent. Let us prove that (3) $\longrightarrow$ (1). Let $g_1, g_2 \in G$ with $g_1 \neq g_2$. Then, $g_1^{-1}g_2 \neq 1$ and if U and V are disjoint open subsets such that $1 \in U$ and $g_1^{-1}g_2 \in V$, then $g_1 \in g_1U$ and $g_2 \in g_1V$, where g_1U and g_1V are still open and disjoint. Thus, it is enough to separate 1 and $g \neq 1$. Pick any $g \neq 1$. If every open subset containing 1 also contained g, then 1 would be in the closure of $\{g\}$, which is absurd since $\{g\}$ is closed and $g \neq 1$. Therefore, there is some open subset U such that $1 \in U$ and $g \notin U$. By Proposition 4.4, we can find an open subset V containing 1, so that $VV \subseteq U$ and $V = V^{-1}$. We claim that V and gV are disjoint open sets with $1 \in V$ and $g \in gV$.

Since $1 \in V$, it is clear that $g \in gV$. If we had $V \cap gV \neq \emptyset$, then by the last sentence in the proof of Proposition 4.4 we would have $g \in VV^{-1} = VV \subseteq U$, a contradiction. □

If H is a subgroup of G (not necessarily normal), we can form the set of left cosets G/H, and we have the projection $p\colon G \to G/H$, where $p(g) = gH = \overline{g}$. If G is a topological group, then G/H can be given the *quotient topology*, where a subset $U \subseteq G/H$ is open iff $p^{-1}(U)$ is open in G. With this topology, p is continuous. The trouble is that G/H is not necessarily Hausdorff. However, we can neatly characterize when this happens.

Proposition 4.7. *If G is a topological group and H is a subgroup of G, then the following properties hold:*

(1) The map $p\colon G \to G/H$ is an open map, which means that $p(V)$ is open in G/H whenever V is open in G.

(2) The space G/H is Hausdorff iff H is closed in G.

(3) If H is open, then H is closed and G/H has the discrete topology (every subset is open).

(4) The subgroup H is open iff $1 \in \overset{\circ}{H}$ (i.e., there is some open subset U so that $1 \in U \subseteq H$).

Proof. (1) Observe that if V is open in G, then $VH = \bigcup_{h\in H} Vh$ is open, since each Vh is open (as right translation is a homeomorphism). However, it is clear that

$$p^{-1}(p(V)) = VH,$$

i.e., $p^{-1}(p(V))$ is open which, by definition of the quotient topology, means that $p(V)$ is open.

(2) If G/H is Hausdorff, then by Proposition 4.6, every point of G/H is closed, i.e., each coset gH is closed, so H is closed. Conversely, assume H is closed. Let $\overline{x}$ and $\overline{y}$ be two distinct point in G/H and let $x, y \in G$ be some elements with $p(x) = \overline{x}$ and $p(y) = \overline{y}$. As $\overline{x} \neq \overline{y}$, the elements x and y are not in the same coset, so $x \notin yH$. As H is closed, so is yH, and since $x \notin yH$, there is some open containing x which is disjoint from yH, and we may assume (by translation) that it is of the form Ux, where U is an open containing 1. By Proposition 4.4, there is some open V containing 1 so that $VV \subseteq U$ and $V = V^{-1}$. Thus, we have

$$V^2x \cap yH = \emptyset$$

and in fact,

$$V^2xH \cap yH = \emptyset,$$

since H is a group; if $z \in V^2xH \cap yH$, then $z = v_1v_2xh_1 = yh_2$ for some $v_1, v_2 \in V$, and some $h_1, h_2 \in H$, but then $v_1v_2x = yh_2h_1^{-1}$ so that $V^2x \cap yH \neq \emptyset$, a contradiction. Since $V = V^{-1}$, we get

$$VxH \cap VyH = \emptyset,$$

and then, since V is open, both VxH and VyH are disjoint, open, so $p(VxH)$ and $p(VyH)$ are open sets (by (1)) containing $\overline{x}$ and $\overline{y}$ respectively and $p(VxH)$ and $p(VyH)$ are disjoint (because $p^{-1}(p(VxH)) = VxHH = VxH$, $p^{-1}(p(VyH)) = VyHH = VyH$, and $VxH \cap VyH = \emptyset$). See Figure 4.5.

(3) If H is open, then every coset gH is open, so every point of G/H is open and G/H is discrete. Also, $\bigcup_{g\notin H} gH$ is open, i.e., H is closed.

(4) Say U is an open subset such that $1 \in U \subseteq H$. Then for every $h \in H$, the set hU is an open subset of H with $h \in hU$, which shows that H is open. The converse is trivial. □

We next provide a criterion relating the connectivity of G with that of G/H.

Proposition 4.8. *Let G be a topological group and H be any subgroup of G. If H and G/H are connected, then G is connected.*

Proof. It is a standard fact of topology that a space G is connected iff every continuous function f from G to the discrete space $\{0, 1\}$ is constant; see Proposition 12.15. Pick any continuous function f from G to $\{0, 1\}$. As H is connected and left translations are homeomorphisms, all cosets gH are connected. Thus, f is constant on every coset gH. It follows that the function $f\colon G \to \{0, 1\}$ induces a continuous function $\overline{f}\colon G/H \to \{0, 1\}$

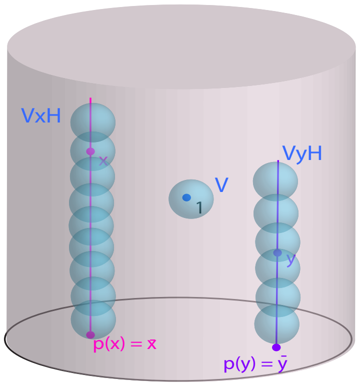


Figure 4.5: A schematic illustration of $VxH \cap VyH = \emptyset$, where G is the pink cylinder, H is the vertical edge, and G/H is the circular base. Note xH and yH are vertical fibres.

such that $f = \overline{f} \circ p$ (where $p\colon G \to G/H$; the continuity of $\overline{f}$ follows immediately from the definition of the quotient topology on G/H). As G/H is connected, $\overline{f}$ is constant, and so $f = \overline{f} \circ p$ is constant. □

The next three propositions describe how to generate a topological group from its symmetric neighborhoods of 1.

Proposition 4.9. *If G is a connected topological group, then G is generated by any symmetric neighborhood V of 1. In fact,*

$$G = \bigcup_{n \geq 1} V^n.$$

Proof. Since $V = V^{-1}$, it is immediately checked that $H = \bigcup_{n \geq 1} V^n$ is the group generated by V. As V is a neighborhood of 1, there is some open subset $U \subseteq V$, with $1 \in U$, and so $1 \in \overset{\circ}{H}$. From Proposition 4.7 (3), the subgroup H is open and closed, and since G is connected, $H = G$. □

Proposition 4.10. *Let G be a topological group and let V be any connected symmetric open subset containing 1. Then, if G_0 is the connected component of the identity, we have*

$$G_0 = \bigcup_{n \geq 1} V^n,$$

and G_0 is a normal subgroup of G. Moreover, the group G/G_0 is discrete.

Proof. First, as V is open, every V^n is open, so the group $\bigcup_{n \geq 1} V^n$ is open, and thus closed, by Proposition 4.7 (3). For every $n \geq 1$, we have the continuous map

$$\underbrace{V \times \cdots \times V}_{n} \longrightarrow V^n : (g_1, \ldots, g_n) \mapsto g_1 \cdots g_n.$$

As V is connected, $V \times \cdots \times V$ is connected, and so V^n is connected; see Theorem 12.18 and Proposition 12.11. Since $1 \in V^n$ for all $n \geq 1$ and every V^n is connected, we use Lemma 12.12 to conclude that $\bigcup_{n\geq 1} V^n$ is connected. Now, $\bigcup_{n\geq 1} V^n$ is connected, open and closed, so it is the connected component of 1. Finally, for every $g \in G$, the group gG_0g^{-1} is connected and contains 1, so it is contained in G_0, which proves that G_0 is normal. Since G_0 is open, Proposition 4.7 (3) implies that the group G/G_0 is discrete. □

Recall that a topological space X is *locally compact* iff for every point $p \in X$, there is a compact neighborhood C of p; that is, there is a compact C and an open U, with $p \in U \subseteq C$. For example, manifolds are locally compact.

Proposition 4.11. *Let G be a topological group and assume that G is connected and locally compact. Then, G is countable at infinity, which means that G is the union of a countable family of compact subsets. In fact, if V is any symmetric compact neighborhood of* 1*, then*

$$G = \bigcup_{n\geq 1} V^n.$$

Proof. Since G is locally compact, there is some compact neighborhood K of 1. Then, $V = K \cap K^{-1}$ is also compact and a symmetric neighborhood of 1. By Proposition 4.9, we have

$$G = \bigcup_{n\geq 1} V^n.$$

An argument similar to the one used in the proof of Proposition 4.10 to show that V^n is connected if V is connected proves that each V^n compact if V is compact. □

We end this section by combining the various properties of a topological group G to characterize when G/G_x is homeomorphic to X. In order to do so, we need two definitions.

Definition 4.16. Let G be a topological group and let X be a topological space. An action $\varphi\colon G \times X \to X$ is *continuous* (and G *acts continuously on* X) if the map φ is continuous.

If an action $\varphi\colon G\times X \to X$ is continuous, then each map $\varphi_g\colon X \to X$ is a homeomorphism of X (recall that $\varphi_g(x) = g \cdot x$, for all $x \in X$).

Under some mild assumptions on G and X, the quotient space G/G_x is homeomorphic to X. For example, this happens if X is a Baire space.

Definition 4.17. A *Baire space* X is a topological space with the property that if $\{F\}_{i\geq 1}$ is any countable family of closed sets F_i such that each F_i has empty interior, then $\bigcup_{i\geq 1} F_i$ also has empty interior. By complementation, this is equivalent to the fact that for every countable family of open sets U_i such that each U_i is dense in X (i.e., $\overline{U}_i = X$), then $\bigcap_{i\geq 1} U_i$ is also dense in X.

Remark: A subset $A \subseteq X$ is *rare* if its closure $\overline{A}$ has empty interior. A subset $Y \subseteq X$ is *meager* if it is a countable union of rare sets. Then, it is immediately verified that a space X is a Baire space iff every nonempty open subset of X is not meager.

The following theorem shows that there are plenty of Baire spaces:

Theorem 4.12. *(Baire) (1) Every locally compact topological space is a Baire space.*

(2) Every complete metric space is a Baire space.

A proof of Theorem 4.12 can be found in Bourbaki [21], Chapter IX, Section 5, Theorem 1.

We can now greatly improve Proposition 4.2 when G and X are topological spaces having some "nice" properties.

Theorem 4.13. *Let G be a topological group which is locally compact and countable at infinity, X a Hausdorff topological space which is a Baire space, and assume that G acts transitively and continuously on X. Then, for any $x \in X$, the map $\varphi\colon G/G_x \to X$ is a homeomorphism.*

Proof. We follow the proof given in Bourbaki [21], Chapter IX, Section 5, Proposition 6 (Essentially the same proof can be found in Mneimné and Testard [86], Chapter 2). First, observe that if a topological group acts continuously and transitively on a Hausdorff topological space, then for every $x \in X$, the stabilizer G_x is a closed subgroup of G. This is because, as the action is continuous, the projection $\pi_x\colon G \longrightarrow X\colon g \mapsto g{\cdot}x$ is continuous, and $G_x = \pi^{-1}(\{x\})$, with $\{x\}$ closed. Therefore, by Proposition 4.7, the quotient space G/G_x is Hausdorff. As the map $\pi_x\colon G \longrightarrow X$ is continuous, the induced map $\varphi_x\colon G/G_x \to X$ is continuous, and by Proposition 4.2, it is a bijection. Therefore, to prove that φ_x is a homeomorphism, it is enough to prove that φ_x is an open map. For this, it suffices to show that π_x is an open map. Given any open U in G, we will prove that for any $g \in U$, the element $\pi_x(g) = g \cdot x$ is contained in the interior of $U \cdot x$. However, observe that this is equivalent to proving that x belongs to the interior of $(g^{-1} \cdot U) \cdot x$. Therefore, we are reduced to the following case: if U is any open subset of G containing 1, then x belongs to the interior of $U \cdot x$.

Since G is locally compact, using Proposition 4.4, we can find a compact neighborhood of the form $W = \overline{V}$, such that $1 \in W$, $W = W^{-1}$ and $W^2 \subseteq U$, where V is open with $1 \in V \subseteq U$. As G is countable at infinity, $G = \bigcup_{i\geq 1} K_i$, where each K_i is compact. Since V is open, all the cosets gV are open, and as each K_i is covered by the gV's, by compactness of K_i, finitely many cosets gV cover each K_i, and so

$$G = \bigcup_{i\geq 1} g_i V = \bigcup_{i\geq 1} g_i W,$$

for countably many $g_i \in G$, where each $g_i W$ is compact. As our action is transitive, we deduce that

$$X = \bigcup_{i\geq 1} g_i W \cdot x,$$

where each $g_iW \cdot x$ is compact, since our action is continuous and the g_iW are compact. As X is Hausdorff, each $g_iW \cdot x$ is closed, and as X is a Baire space expressed as a union of closed sets, one of the $g_iW \cdot x$ must have nonempty interior; that is, there is some $w \in W$, with $g_iw \cdot x$ in the interior of $g_iW \cdot x$, for some i. But then, as the map $y \mapsto g \cdot y$ is a homeomorphism for any given $g \in G$ (where $y \in X$), we see that x is in the interior of

$$w^{-1}g_i^{-1} \cdot (g_iW \cdot x) = w^{-1}W \cdot x \subseteq W^{-1}W \cdot x = W^2 \cdot x \subseteq U \cdot x,$$

as desired. □

By Theorem 4.12, we get the following important corollary:

Theorem 4.14. *Let G be a topological group which is locally compact and countable at infinity, X a Hausdorff locally compact topological space, and assume that G acts transitively and continuously on X. Then, for any $x \in X$, the map $\varphi_x \colon G/G_x \to X$ is a homeomorphism.*

Readers who wish to learn more about topological groups may consult Sagle and Walde [99] and Chevalley [31] for an introductory account, and Bourbaki [20], Weil [116] and Pontryagin [94, 95], for a more comprehensive account (especially the last two references).

4.6 Problems

Problem 4.1. Recall that the group $\mathbf{SU}(2)$ consists of all complex matrices of the form

$$A = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \qquad \alpha, \beta \in \mathbb{C}, \qquad \alpha\overline{\alpha} + \beta\overline{\beta} = 1,$$

and the action $\cdot \colon \mathbf{SU}(2) \times (\mathbb{C} \cup \{\infty\}) \to \mathbb{C} \cup \{\infty\}$ is given by

$$A \cdot w = \frac{\alpha w + \beta}{-\overline{\beta} w + \overline{\alpha}}, \quad w \in \mathbb{C} \cup \{\infty\}.$$

This is a transitive action. Using the stereographic projection σ_N from S^2 onto $\mathbb{C} \cup \{\infty\}$ and its inverse σ_N^{-1}, we can define an action of $\mathbf{SU}(2)$ on S^2 by

$$A \cdot (x, y, z) = \sigma_N^{-1}(A \cdot \sigma_N(x, y, z)), \quad (x, y, z) \in S^2,$$

and we denote by $\rho(A)$ the corresponding map from S^2 to S^2.

(1) If we write $\alpha = a + ib$ and $\beta = c + id$, prove that $\rho(A)$ is given by the matrix

$$\rho(A) = \begin{pmatrix} a^2 - b^2 - c^2 + d^2 & -2ab - 2cd & -2ac + 2bd \\ 2ab - 2cd & a^2 - b^2 + c^2 - d^2 & -2ad - 2bc \\ 2ac + 2bd & 2ad - 2bc & a^2 + b^2 - c^2 - d^2 \end{pmatrix}.$$

Prove that $\rho(A)$ is indeed a rotation matrix which represents the rotation whose axis is the line determined by the vector $(d, -c, b)$ and whose angle $\theta \in [-\pi, \pi]$ is determined by

$$\cos\frac{\theta}{2} = |a|.$$

Hint. Recall that the axis of a rotation matrix $R \in \mathbf{SO}(3)$ is specified by any eigenvector of 1 for R, and that the angle of rotation θ satisfies the equation

$$\mathrm{tr}(R) = 2\cos\theta + 1.$$

(2) We can compute the derivative $d\rho_I \colon \mathfrak{su}(2) \to \mathfrak{so}(3)$ of ρ at I as follows. Recall that $\mathfrak{su}(2)$ consists of all complex matrices of the form

$$\begin{pmatrix} ib & c+id \\ -c+id & -ib \end{pmatrix}, \quad b, c, d \in \mathbb{R},$$

so pick the following basis for $\mathfrak{su}(2)$,

$$X_1 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and define the curves in $\mathbf{SU}(2)$ through I given by

$$c_1(t) = \begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix}, \quad c_2(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}, \quad c_3(t) = \begin{pmatrix} \cos t & i\sin t \\ i\sin t & \cos t \end{pmatrix}.$$

Prove that $c_i'(0) = X_i$ for $i = 1, 2, 3$, and that

$$d\rho_I(X_1) = 2\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_2) = 2\begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad d\rho_I(X_3) = 2\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Thus, we have

$$d\rho_I(X_1) = 2E_3, \quad d\rho_I(X_2) = -2E_2, \quad d\rho_I(X_3) = 2E_1,$$

where (E_1, E_2, E_3) is the basis of $\mathfrak{so}(3)$ given in Section 2.5. Conclude that $d\rho_I$ is an isomorphism between the Lie algebras $\mathfrak{su}(2)$ and $\mathfrak{so}(3)$.

(3) Recall from Proposition 3.13 that we have the commutative diagram

$$\begin{tikzcd} \mathbf{SU}(2) \arrow[r, "\rho"] & \mathbf{SO}(3) \\ \mathfrak{su}(2) \arrow[u, "\exp"] \arrow[r, "d\rho_I"'] & \mathfrak{so}(3) \arrow[u, "\exp"'] \end{tikzcd}.$$

Since $d\rho_I$ is surjective and the exponential map $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective, conclude that ρ is surjective. Prove that $\mathrm{Ker}\,\rho = \{I, -I\}$.

Problem 4.2. Consider the action of the group $\mathbf{SL}(2,\mathbb{R})$ on the upper half-plane, $H = \{z = x+iy \in \mathbb{C} \mid y > 0\}$, given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot z = \frac{az+b}{cz+d}.$$

(a) Check that for any $g \in \mathbf{SL}(2,\mathbb{R})$,

$$\Im(g\cdot z) = \frac{\Im(z)}{|cz+d|^2},$$

and conclude that if $z \in H$, then $g \cdot z \in H$, so that the action of $\mathbf{SL}(2,\mathbb{R})$ on H is indeed well-defined (Recall, $\Re(z) = x$ and $\Im(z) = y$, where $z = x+iy$.)

(b) Check that if $c \neq 0$, then

$$\frac{az+b}{cz+d} = \frac{-1}{c^2z+cd} + \frac{a}{c}.$$

Prove that the group of Möbius transformations induced by $\mathbf{SL}(2,\mathbb{R})$ is generated by Möbius transformations of the form

1. $z \mapsto z+b$,
2. $z \mapsto kz$,
3. $z \mapsto -1/z$,

where $b \in \mathbb{R}$ and $k \in \mathbb{R}$, with $k > 0$. Deduce from the above that the action of $\mathbf{SL}(2,\mathbb{R})$ on H is transitive and that transformations of type (1) and (2) suffice for transitivity.

(c) Now, consider the action of the discrete group $\mathbf{SL}(2,\mathbb{Z})$ on H, where $\mathbf{SL}(2,\mathbb{Z})$ consists of all matrices

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad ad-bc = 1, \quad a,b,c,d \in \mathbb{Z}.$$

Why is this action not transitive? Consider the two transformations

$$S\colon z \mapsto -1/z$$

associated with $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and

$$T\colon z \mapsto z+1$$

associated with $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

Define the subset, D, of H, as the set of points, $z = x+iy$, such that $-1/2 \leq x \leq -1/2$ and $x^2+y^2 \geq 1$. Observe that D contains the three special points, i, $\rho = e^{2\pi i/3}$ and $-\overline{\rho} = e^{\pi i/3}$.

Draw a picture of this set, known as a *fundamental domain* of the action of $G = \mathbf{SL}(2, \mathbb{Z})$ on H.

Remark: Gauss proved that the group $G = \mathbf{SL}(2, \mathbb{Z})$ is generated by S and T.

Problem 4.3. Let J be the 2×2 matrix

$$J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

and let $\mathbf{SU}(1, 1)$ be the set of 2×2 complex matrices

$$\mathbf{SU}(1, 1) = \{A \mid A^* J A = J, \quad \det(A) = 1\},$$

where A^* is the conjugate transpose of A.

(a) Prove that $\mathbf{SU}(1, 1)$ is the group of matrices of the form

$$A = \begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix}, \quad \text{with} \quad a\overline{a} - b\overline{b} = 1.$$

If

$$g = \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}$$

prove that the map from $\mathbf{SL}(2, \mathbb{R})$ to $\mathbf{SU}(1, 1)$ given by

$$A \mapsto gAg^{-1}$$

is a group isomorphism.

(b) Prove that the Möbius transformation associated with g,

$$z \mapsto \frac{z - i}{z + i}$$

is a bijection between the upper half-plane, H, and the unit open disk, $D = \{z \in \mathbb{C} \mid |z| < 1\}$. Prove that the map from $\mathbf{SU}(1, 1)$ to $S^1 \times D$ given by

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \mapsto (a/|a|, b/a)$$

is a continuous bijection (in fact, a homeomorphism). Conclude that $\mathbf{SU}(1, 1)$ is topologically an open solid torus.

(c) Check that $\mathbf{SU}(1, 1)$ acts transitively on D by

$$\begin{pmatrix} a & b \\ \overline{b} & \overline{a} \end{pmatrix} \cdot z = \frac{az + b}{\overline{b}z + \overline{a}}.$$

Find the stabilizer of $z = 0$ and conclude that

$$\mathbf{SU}(1, 1)/\mathbf{SO}(2) \cong D.$$

Chapter 5

The Lorentz Groups ⊛

In this chapter we study a class of linear Lie groups known as the Lorentz groups. As we will see, the Lorentz groups provide interesting examples of homogeneous spaces. Moreover, the Lorentz group $\mathbf{SO}(3,1)$ shows up in an interesting way in computer vision.

5.1 The Lorentz Groups $\mathbf{O}(n,1)$, $\mathbf{SO}(n,1)$ and $\mathbf{SO}_0(n,1)$

Denote the $p \times p$-identity matrix by I_p, for $p, q, \geq 1$, and define

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.$$

If $n = p + q$, the matrix $I_{p,q}$ is associated with the nondegenerate symmetric bilinear form

$$\varphi_{p,q}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{n} x_j y_j$$

with associated quadratic form

$$\Phi_{p,q}((x_1, \ldots, x_n)) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n} x_j^2.$$

In particular, when $p = 1$ and $q = 3$, we have the *Lorentz metric*

$$x_1^2 - x_2^2 - x_3^2 - x_4^2.$$

In physics, x_1 is interpreted as time and written t, and x_2, x_3, x_4 as coordinates in $\mathbb{R}^3$ and written x, y, z. Thus, the Lorentz metric is usually written a

$$t^2 - x^2 - y^2 - z^2,$$

although it also appears as

$$x^2 + y^2 + z^2 - t^2,$$

which is equivalent but slightly less convenient for certain purposes, as we will see later. The space $\mathbb{R}^4$ with the Lorentz metric is called *Minkowski space*. It plays an important role in Einstein's theory of special relativity.

Definition 5.1. For any $p, q \geq 1$, the group $\mathbf{O}(p, q)$ is the set of all $n \times n$-matrices

$$\mathbf{O}(p, q) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}\}.$$

This is the group of all invertible linear maps of $\mathbb{R}^n$ that preserve the quadratic form $\Phi_{p,q}$, i.e., the group of isometries of $\Phi_{p,q}$.

Let us check that $\mathbf{O}(p, q)$ is indeed a group.

Proposition 5.1. *For any $p, q \geq 1$, the set $\mathbf{O}(p, q)$ is a group, with the inverse A^{-1} of any element $A \in \mathbf{O}(p, q)$ given by $A^{-1} = I_{p,q} A^\top I_{p,q}$. If $A \in \mathbf{O}(p, q)$, then $A^\top \in \mathbf{O}(p, q)$.*

Proof. If $A, B \in \mathbf{O}(p, q)$, then $A^\top I_{p,q} A = I_{p,q}$ and $B^\top I_{p,q} B = I_{p,q}$, so we get

$$(AB^\top) I_{p,q} AB = B^\top A^\top I_{p,q} AB = B^\top I_{p,q} B = I_{p,q},$$

which shows that $AB \in \mathbf{O}(p, q)$. Since $I_{p,q}^2 = I$ we have, $I \in \mathbf{O}(p, q)$. Since $I_{p,q}^2 = I$, the condition $A^\top I_{p,q} A = I_{p,q}$ is equivalent to $I_{p,q} A^\top I_{p,q} A = I$, which means that

$$A^{-1} = I_{p,q} A^\top I_{p,q}.$$

Consequently $I = AA^{-1} = A I_{p,q} A^\top I_{p,q}$, so

$$A I_{p,q} A^\top = I_{p,q} \tag{$*$}$$

also holds, which shows that $\mathbf{O}(p, q)$ is closed under transposition (i.e., if $A \in \mathbf{O}(p, q)$, then $A^\top \in \mathbf{O}(p, q)$). Using the fact that $I_{p,q}^2 = I$ and $I_{p,q}^\top = I_{p,q}$, we have

$$(A^{-1})^\top I_{p,q} A^{-1} = (I_{p,q} A^\top I_{p,q})^\top I_{p,q} A^{-1} = I_{p,q} A I_{p,q} I_{p,q} A^{-1} = I_{p,q} A A^{-1} = I_{p,q}.$$

Therefore, $A^{-1} \in \mathbf{O}(p, q)$, so $\mathbf{O}(p, q)$ is indeed a subgroup of $\mathbf{GL}(n, \mathbb{R})$ with inverse given by $A^{-1} = I_{p,q} A^\top I_{p,q}$. □

Definition 5.2. For any $p, q \geq 1$, the subgroup $\mathbf{SO}(p, q)$ of $\mathbf{O}(p, q)$ consisting of the isometries of $(\mathbb{R}^n, \Phi_{p,q})$ with determinant $+1$ is given by

$$\mathbf{SO}(p, q) = \{A \in \mathbf{O}(p, q) \mid \det(A) = 1\}.$$

It is clear that $\mathbf{SO}(p,q)$ is indeed a subgroup of of $\mathbf{O}(p,q)$ also closed under transposition.

The condition $A^\top I_{p,q} A = I_{p,q}$ has an interpretation in terms of the inner product $\varphi_{p,q}$ and the columns (and rows) of A. Indeed, if we denote the jth column of A by A_j, then

$$A^\top I_{p,q} A = (\varphi_{p,q}(A_i, A_j)),$$

so $A \in \mathbf{O}(p,q)$ iff the columns of A form an "orthonormal basis" w.r.t. $\varphi_{p,q}$, i.e.,

$$\varphi_{p,q}(A_i, A_j) = \begin{cases} \delta_{ij} & \text{if } 1 \leq i,j \leq p; \\ -\delta_{ij} & \text{if } p+1 \leq i,j \leq p+q. \end{cases}$$

The difference with the usual orthogonal matrices is that $\varphi_{p,q}(A_i, A_i) = -1$, if $p+1 \leq i \leq p+q$. As $\mathbf{O}(p,q)$ is closed under transposition, the rows of A also form an orthonormal basis w.r.t. $\varphi_{p,q}$.

It turns out that $\mathbf{SO}(p,q)$ has two connected components, and the component containing the identity is a subgroup of $\mathbf{SO}(p,q)$ denoted $\mathbf{SO}_0(p,q)$. The group $\mathbf{SO}_0(p,q)$ is actually homeomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q) \times \mathbb{R}^{pq}$. This is not immediately obvious. A way to prove this fact is to work out the polar decomposition for matrices in $\mathbf{O}(p,q)$. This is nicely done in Dragon [40] (see Section 6.2). A close examination of the factorization obtained in Section 6.1 also shows that there is bijection between $\mathbf{O}(p,q)$ and $\mathbf{O}(p) \times \mathbf{O}(q) \times \mathbb{R}^{pq}$. Another way to prove these results (in a stronger form, namely that there is a homeomorphism) is to use results on pseudo-algebraic subgroups of $\mathbf{GL}(n,\mathbb{C})$; see Sections 6.2 and 6.3. It can also be shown that there are isomorphisms $\psi\colon \mathbf{O}(p,q) \to \mathbf{O}(q,p)$, $\psi\colon \mathbf{SO}(p,q) \to \mathbf{SO}(q,p)$, and $\psi\colon \mathbf{SO}_0(p,q) \to \mathbf{SO}_0(q,p)$; see Proposition 6.8.

We will now determine the polar decomposition and the SVD decomposition of matrices in the Lorentz groups $\mathbf{O}(n,1)$ and $\mathbf{SO}(n,1)$. Write $J = I_{n,1}$, and given any $A \in \mathbf{O}(n,1)$, write

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix},$$

where B is an $n \times n$ matrix, u, v are (column) vectors in $\mathbb{R}^n$ and $c \in \mathbb{R}$. We begin with the polar decomposition of matrices in the Lorentz groups $\mathbf{O}(n,1)$.

Proposition 5.2. *Every matrix $A \in \mathbf{O}(n,1)$ has a polar decomposition of the form*

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$.

Proof. Write A in block form as above. As the condition for A to be in $\mathbf{O}(n,1)$ is $A^\top J A = J$, we get

$$\begin{pmatrix} B^\top & v \\ u^\top & c \end{pmatrix} \begin{pmatrix} B & u \\ -v^\top & -c \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

i.e.,

$$\begin{aligned} B^\top B &= I_n + vv^\top \\ u^\top u &= c^2 - 1 \\ B^\top u &= cv. \end{aligned}$$

If we remember that we also have $AJA^\top = J$, then

$$\begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} \begin{pmatrix} B^\top & v \\ -u^\top & -c \end{pmatrix} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

and

$$\begin{aligned} BB^\top &= I_n + uu^\top \\ v^\top v &= c^2 - 1 \\ Bv &= cu. \end{aligned}$$

From $u^\top u = \|u\|^2 = c^2 - 1$, we deduce that $|c| \geq 1$. From $B^\top B = I_n + vv^\top$, we deduce that $B^\top B$ is clearly symmetric; we also deduce that $B^\top B$ positive definite since

$$x^\top (I_n + vv^\top)x = \|x\|^2 + x^\top vv^\top x = \|x\|^2 + \left\|v^\top x\right\|^2,$$

and $\|x\|^2 + \left\|v^\top x\right\|^2$ whenever $x \neq 0$. Now, geometrically, it is well known that $vv^\top / v^\top v$ is the orthogonal projection onto the line determined by v. Consequently, the kernel of $vv^\top$ is the orthogonal complement of v, and $vv^\top$ has the eigenvalue 0 with multiplicity $n - 1$ and the eigenvalue $c^2 - 1 = \|v\|^2 = v^\top v$ with multiplicity 1. The eigenvectors associated with 0 are orthogonal to v, and the eigenvectors associated with $c^2 - 1$ are proportional with v since $\left(vv^\top / \|v\|^2\right) v = (c^2 - 1)v$. It follows that $I_n + vv^\top$ has the eigenvalue 1 with multiplicity $n - 1$ and the eigenvalue c^2 with multiplicity 1, the eigenvectors being as before. Now, B has polar form $B = QS_1$, where Q is orthogonal and S_1 is symmetric positive definite and $S_1^2 = B^\top B = I_n + vv^\top$. Therefore, if $c > 0$, then $S_1 = \sqrt{I_n + vv^\top}$ is a symmetric positive definite matrix with eigenvalue 1 with multiplicity $n - 1$ and eigenvalue c with multiplicity 1, the eigenvectors being as before. If $c < 0$, then change c to $-c$.

Case 1: $c > 0$. Then v is an eigenvector of S_1 for c and we must also have $Bv = cu$, which implies

$$Bv = QS_1 v = Q(cv) = cQv = cu,$$

so

$$Qv = u.$$

It follows that

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} = \begin{pmatrix} QS_1 & Qv \\ v^\top & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$.

Case 2: $c < 0$. Then v is an eigenvector of S_1 for $-c$ and we must also have $Bv = cu$, which implies

$$Bv = QS_1v = Q(-cv) = cQ(-v) = cu,$$

so

$$Q(-v) = u.$$

It follows that

$$A = \begin{pmatrix} B & u \\ v^\top & c \end{pmatrix} = \begin{pmatrix} QS_1 & Q(-v) \\ v^\top & c \end{pmatrix} = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & -v \\ -v^\top & -c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = -\sqrt{\|v\|^2 + 1}$.

We conclude that any $A \in \mathbf{O}(n,1)$ has a factorization of the form

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \quad \text{or} \quad A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$. Note that the matrix $\begin{pmatrix} Q & 0 \\ 0 & \pm 1 \end{pmatrix}$ is orthogonal and $\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$ is symmetric. Proposition 5.3 will show that $\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$ is positive definite. Hence the above factorizations are polar decompositions. □

In order to show that $S = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$ is positive definite, we show that the eigenvalues are strictly positive. Such a matrix is called a *Lorentz boost*. Observe that if $v = 0$, then $c = 1$ and $S = I_{n+1}$.

Proposition 5.3. *Assume $v \neq 0$. The eigenvalues of the symmetric positive definite matrix*

$$S = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $c = \sqrt{\|v\|^2 + 1}$, are 1 with multiplicity $n-1$, and e^{α} and $e^{-\alpha}$ each with multiplicity 1 (for some $\alpha \geq 0$). An orthonormal basis of eigenvectors of S consists of vectors of the form

$$\begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} u_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ -\frac{1}{\sqrt{2}} \end{pmatrix},$$

where the $u_i \in \mathbb{R}^n$ are all orthogonal to v and pairwise orthogonal.

Proof. Let us solve the linear system

$$\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} v \\ d \end{pmatrix} = \lambda \begin{pmatrix} v \\ d \end{pmatrix}.$$

We get

$$\begin{aligned} \left(\sqrt{I_n + vv^\top}\right) v + dv &= \lambda v \\ v^\top v + cd &= \lambda d. \end{aligned}$$

Since the proof of Proposition 5.2 implies that $c = \sqrt{\|v\|^2 + 1}$ and $\left(\sqrt{I_n + vv^\top}\right) v = cv$, the previous two equations are equivalent to

$$\begin{aligned} (c+d)v &= \lambda v \\ c^2 - 1 + cd &= \lambda d. \end{aligned}$$

Because $v \neq 0$, we get $\lambda = c + d$. Substituting in the second equation, we get

$$c^2 - 1 + cd = (c+d)d,$$

that is,

$$d^2 = c^2 - 1.$$

In other words $d = \pm\sqrt{c^2 - 1}$, which in turn implies $\lambda = c + d = c \pm \sqrt{c^2 - 1}$. Thus, either $\lambda_1 = c + \sqrt{c^2 - 1}$ and $d = \sqrt{c^2 - 1}$, or $\lambda_2 = c - \sqrt{c^2 - 1}$ and $d = -\sqrt{c^2 - 1}$. Since $c \geq 1$ and $\lambda_1 \lambda_2 = 1$, set $\alpha = \log(c + \sqrt{c^2 - 1}) \geq 0$, so that $-\alpha = \log(c - \sqrt{c^2 - 1})$, and then $\lambda_1 = e^\alpha$ and $\lambda_2 = e^{-\alpha}$. On the other hand, if u is orthogonal to v, observe that

$$\begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix} \begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix},$$

since the kernel of $vv^\top$ is the orthogonal complement of v. The rest is clear. □

Corollary 5.4. *The singular values of any matrix $A \in \mathbf{O}(n,1)$ are 1 with multiplicity $n-1$, e^α, and $e^{-\alpha}$, for some $\alpha \geq 0$.*

Note that the case $\alpha = 0$ is possible, in which case A is an orthogonal matrix of the form

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix},$$

with $Q \in \mathbf{O}(n)$. The two singular values e^α and $e^{-\alpha}$ tell us how much A deviates from being orthogonal.

By using Proposition 5.2 we see that $\mathbf{O}(n,1)$ has four components corresponding to the cases:

(1) $Q \in \mathbf{O}(n)$; $\det(Q) < 0$; $+1$ as the lower right entry of the orthogonal matrix;

(2) $Q \in \mathbf{SO}(n)$; -1 as the lower right entry of the orthogonal matrix;

(3) $Q \in \mathbf{O}(n)$; $\det(Q) < 0$; -1 as the lower right entry of the orthogonal matrix;

(4) $Q \in \mathbf{SO}(n)$; $+1$ as the lower right entry of the orthogonal matrix.

Observe that $\det(A) = -1$ in Cases (1) and (2) and that $\det(A) = +1$ in Cases (3) and (4). Thus, Cases (3) and (4) correspond to the group $\mathbf{SO}(n,1)$, in which case the polar decomposition is of the form

$$A = \begin{pmatrix} Q & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

where $Q \in \mathbf{O}(n)$, with $\det(Q) = -1$ and $c = \sqrt{\|v\|^2 + 1}$, or

$$A = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$$

where $Q \in \mathbf{SO}(n)$ and $c = \sqrt{\|v\|^2 + 1}$.

The components in Cases (1), (2) and (3) are not groups. We will show later that all four components are connected and that Case (4) corresponds to a group (Proposition 5.7). This group is the connected component of the identity and it is denoted $\mathbf{SO}_0(n,1)$ (see Corollary 5.11). For the time being, note that $A \in \mathbf{SO}_0(n,1)$ iff $A \in \mathbf{SO}(n,1)$ and $a_{n+1\,n+1} = c > 0$ (here, $A = (a_{i\,j})$.) In fact, we proved above that if $a_{n+1\,n+1} > 0$, then $a_{n+1\,n+1} \geq 1$.

Remark: If we let

$$\Lambda_P = \begin{pmatrix} I_{n-1,1} & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Lambda_T = I_{n,1}, \quad \text{where} \quad I_{n,1} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

then we have the disjoint union

$$\mathbf{O}(n,1) = \mathbf{SO}_0(n,1) \cup \Lambda_P \mathbf{SO}_0(n,1) \cup \Lambda_T \mathbf{SO}_0(n,1) \cup \Lambda_P \Lambda_T \mathbf{SO}_0(n,1).$$

We can now determine a convenient form for the SVD of matrices in $\mathbf{O}(n,1)$.

Theorem 5.5. *Every matrix $A \in \mathbf{O}(n,1)$ can be written as*

$$A = \begin{pmatrix} P & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix}$$

with $\epsilon = \pm 1$, $P \in \mathbf{O}(n)$ and $Q \in \mathbf{SO}(n)$. When $A \in \mathbf{SO}(n,1)$, we have $\det(P)\epsilon = +1$, and when $A \in \mathbf{SO}_0(n,1)$, we have $\epsilon = +1$ and $P \in \mathbf{SO}(n)$; that is,

$$A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix}$$

with $P \in \mathbf{SO}(n)$ and $Q \in \mathbf{SO}(n)$.

Proof. By Proposition 5.2, any matrix $A \in \mathbf{O}(n,1)$ can be written as

$$A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$$

where $\epsilon = \pm 1$, $R \in \mathbf{O}(n)$ and $c = \sqrt{\|v\|^2 + 1}$. The case where $c = 1$ is trivial, so assume $c > 1$, which means that α from Proposition 5.3 is such that $\alpha > 0$. The key fact is that the eigenvalues of the matrix

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix}$$

are e^α and $e^{-\alpha}$. To verify this fact, observe that

$$\det \begin{pmatrix} \cosh\alpha - \lambda & \sinh\alpha \\ \sinh\alpha & \cosh\alpha - \lambda \end{pmatrix} = (\cosh\alpha - \lambda)^2 - \sinh^2\alpha = \lambda^2 - 2\lambda\cosh\alpha + 1 = 0,$$

which in turn implies

$$\lambda = \cosh\alpha \pm \sinh\alpha,$$

and the conclusion follows from the definitions of $\cosh\alpha = \frac{e^\alpha + e^{-\alpha}}{2}$ and $\sinh\alpha = \frac{e^\alpha - e^{-\alpha}}{2}$.

Also observe that the definitions of $\cosh\alpha$ and $\sinh\alpha$ imply that

$$\begin{pmatrix} e^\alpha & 0 \\ 0 & e^{-\alpha} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

which is equivalent to the observation that $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ is the eigenvector associated with e^α, while $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$ is the eigenvector associated with $e^{-\alpha}$.

From these two facts we see that the diagonal matrix

$$D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & e^{\alpha} & 0 \\ 0 & \cdots & 0 & 0 & e^{-\alpha} \end{pmatrix}$$

of eigenvalues of $S = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix}$ is given by

$$D = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

By Proposition 5.3, an orthonormal basis of eigenvectors of S consists of vectors of the form

$$\begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \ldots, \begin{pmatrix} u_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{v}{\sqrt{2}\|v\|} \\ -\frac{1}{\sqrt{2}} \end{pmatrix},$$

where the $u_i \in \mathbb{R}^n$ are all orthogonal to v and pairwise orthogonal. Now, if we multiply the matrices

$$\begin{pmatrix} u_1 & \cdots & u_{n-1} & \frac{v}{\sqrt{2}\|v\|} & \frac{v}{\sqrt{2}\|v\|} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix},$$

we get an orthogonal matrix of the form

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}$$

where the columns of Q are the vectors

$$u_1, \cdots, u_{n-1}, \frac{v}{\|v\|}.$$

By flipping u_1 to $-u_1$ if necessary, we can make sure that this matrix has determinant $+1$. Consequently,

$$S = \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix},$$

so

$$A = \begin{pmatrix} R & 0 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & 1 \end{pmatrix},$$

and if we let $P = RQ$, we get the desired decomposition. $\square$

Remark: We warn our readers about Chapter 6 of Baker's book [12]. Indeed, this chapter is seriously flawed. The main two Theorems (Theorem 6.9 and Theorem 6.10) are false, and as consequence, the proof of Theorem 6.11 is wrong too. Theorem 6.11 states that the exponential map $\exp\colon \mathfrak{so}(n,1) \to \mathbf{SO}_0(n,1)$ is surjective, which is correct, but known proofs are nontrivial and quite lengthy (see Section 5.2). The proof of Theorem 6.12 is also false, although the theorem itself is correct (this is our Theorem 5.18, see Section 5.2). The main problem with Theorem 6.9 (in Baker) is that the existence of the normal form for matrices in $\mathbf{SO}_0(n,1)$ claimed by this theorem is unfortunately false on several accounts. Firstly, it would imply that every matrix in $\mathbf{SO}_0(n,1)$ can be diagonalized, but this is false for $n \geq 2$. Secondly, even if a matrix $A \in \mathbf{SO}_0(n,1)$ is diagonalizable as $A = PDP^{-1}$, Theorem 6.9 (and Theorem 6.10) miss some possible eigenvalues and the matrix P is not necessarily in $\mathbf{SO}_0(n,1)$ (as the case $n = 1$ already shows). For a thorough analysis of the eigenvalues of Lorentz isometries (and much more), one should consult Riesz [97] (Chapter III).

Clearly, a result similar to Theorem 5.5 also holds for the matrices in the groups $\mathbf{O}(1,n)$, $\mathbf{SO}(1,n)$ and $\mathbf{SO}_0(1,n)$. For example, every matrix $A \in \mathbf{SO}_0(1,n)$ can be written as

$$A = \begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & \cdots & 0 \\ \sinh\alpha & \cosh\alpha & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q^\top \end{pmatrix},$$

where $P, Q \in \mathbf{SO}(n)$.

In the case $n = 3$, we obtain the *proper orthochronous Lorentz group* $\mathbf{SO}_0(1,3)$, also denoted $\mathbf{Lor}(1,3)$. By the way, $\mathbf{O}(1,3)$ is called the *(full) Lorentz group* and $\mathbf{SO}(1,3)$ is the *special Lorentz group*.

Theorem 5.5 (really, the version for $\mathbf{SO}_0(1,n)$) shows that the Lorentz group $\mathbf{SO}_0(1,3)$ is generated by the matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3)$$

and the matrices of the form

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & 0 \\ \sinh\alpha & \cosh\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This fact will be useful when we prove that the homomorphism $\varphi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ is surjective.

Remark: Unfortunately, unlike orthogonal matrices which can always be diagonalized over $\mathbb{C}$, **not** every matrix in $\mathbf{SO}(1,n)$ can be diagonalized for $n \geq 2$. This has to do with the fact that the Lie algebra $\mathfrak{so}(1,n)$ has non-zero idempotents (see Section 5.2).

It turns out that the group $\mathbf{SO}_0(1,3)$ admits another interesting characterization involving the hypersurface

$$\mathcal{H} = \{(t,x,y,z) \in \mathbb{R}^4 \mid t^2 - x^2 - y^2 - z^2 = 1\}.$$

This surface has two sheets, and it is not hard to show that $\mathbf{SO}_0(1,3)$ is the subgroup of $\mathbf{SO}(1,3)$ that preserves these two sheets (does not swap them). Actually, we will prove this fact for any n. In preparation for this, we need some definitions and a few propositions.

Let us switch back to $\mathbf{SO}(n,1)$. First, as a matter of notation, we write every $u \in \mathbb{R}^{n+1}$ as $u = (\mathbf{u}, t)$, where $\mathbf{u} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, so that the Lorentz inner product can be expressed as

$$\langle u, v\rangle = \langle (\mathbf{u}, t), (\mathbf{v}, s)\rangle = \mathbf{u}\cdot\mathbf{v} - ts,$$

where $\mathbf{u}\cdot\mathbf{v}$ is the standard Euclidean inner product (the Euclidean norm of x is denoted $\|x\|$). Then we can classify the vectors in $\mathbb{R}^{n+1}$ as follows:

Definition 5.3. A nonzero vector $u = (\mathbf{u}, t) \in \mathbb{R}^{n+1}$ is called

(a) *spacelike* iff $\langle u, u\rangle > 0$, i.e., iff $\|\mathbf{u}\|^2 > t^2$;

(b) *timelike* iff $\langle u, u\rangle < 0$, i.e., iff $\|\mathbf{u}\|^2 < t^2$;

(c) *lightlike* or *isotropic* iff $\langle u, u\rangle = 0$, i.e., iff $\|\mathbf{u}\|^2 = t^2$.

A spacelike (resp. timelike, resp. lightlike) vector is said to be *positive* iff $t > 0$ and *negative* iff $t < 0$. The set of all isotropic vectors

$$\mathcal{H}_n(0) = \{u = (\mathbf{u}, t) \in \mathbb{R}^{n+1} \mid \|\mathbf{u}\|^2 = t^2\}$$

is called the *light cone*. For every $r > 0$, let

$$\mathcal{H}_n(r) = \{u = (\mathbf{u}, t) \in \mathbb{R}^{n+1} \mid \|\mathbf{u}\|^2 - t^2 = -r\},$$

a hyperboloid of two sheets.

It is easy to check that $\mathcal{H}_n(r)$ has two connected components as follows: First, since $r > 0$ and

$$\|\mathbf{u}\|^2 + r = t^2,$$

we have $|t| \geq \sqrt{r}$. For any $x = (x_1, \ldots, x_n, t) \in \mathcal{H}_n(r)$ with $t \geq \sqrt{r}$, we have the continuous path from $(0, \ldots, 0, \sqrt{r})$ to x given by

$$\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, \sqrt{r + \lambda^2(t^2 - r)}),$$

where $\lambda \in [0, 1]$, proving that the component of $(0, \ldots, 0, \sqrt{r})$ is connected. Similarly, when $t \leq -\sqrt{r}$, we have the continuous path from $(0, \ldots, 0, -\sqrt{r})$ to x given by

$$\lambda \mapsto (\lambda x_1, \ldots, \lambda x_n, -\sqrt{r + \lambda^2(t^2 - r)}),$$

where $\lambda \in [0, 1]$, proving that the component of $(0, \ldots, 0, -\sqrt{r})$ is connected. We denote the sheet containing $(0, \ldots, 0, \sqrt{r})$ by $\mathcal{H}_n^+(r)$ and sheet containing $(0, \ldots, 0, -\sqrt{r})$ by $\mathcal{H}_n^-(r)$

Since every Lorentz isometry $A \in \mathbf{SO}(n, 1)$ preserves the Lorentz inner product, we conclude that A globally preserves every hyperboloid $\mathcal{H}_n(r)$, for $r > 0$. We claim that every $A \in \mathbf{SO}_0(n, 1)$ preserves both $\mathcal{H}_n^+(r)$ and $\mathcal{H}_n^-(r)$. This follows immediately from

Proposition 5.6. *If $a_{n+1\,n+1} > 0$, then every isometry $A \in \mathbf{O}(n, 1)$ preserves all positive (resp. negative) timelike vectors and all positive (resp. negative) lightlike vectors. Moreover, if $A \in \mathbf{O}(n, 1)$ preserves all positive timelike vectors, then $a_{n+1\,n+1} > 0$.*

Proof. Let $u = (\mathbf{u}, t)$ be a nonzero timelike or lightlike vector. This means that

$$\|\mathbf{u}\|^2 \leq t^2 \quad \text{and} \quad t \neq 0.$$

Since $A \in \mathbf{O}(n, 1)$, the matrix A preserves the inner product; if $\langle u, u \rangle = \|\mathbf{u}\|^2 - t^2 < 0$, we get $\langle Au, Au \rangle < 0$, which shows that Au is also timelike. Similarly, if $\langle u, u \rangle = 0$, then $\langle Au, Au \rangle = 0$. Define $A_{n+1} = (\mathbf{A}_{n+1}, a_{n+1\,n+1})$ is the $(n+1)$th row of the matrix A. As $A \in \mathbf{O}(n, 1)$, we know that

$$\langle A_{n+1}, A_{n+1} \rangle = -1,$$

that is,

$$\|\mathbf{A}_{n+1}\|^2 - a_{n+1\,n+1}^2 = -1,$$

or equivalently

$$\|\mathbf{A}_{n+1}\|^2 = a_{n+1\,n+1}^2 - 1.$$

The $(n+1)$th component of the vector Au is

$$\mathbf{u} \cdot \mathbf{A}_{n+1} + a_{n+1\,n+1} t.$$

By Cauchy-Schwarz,

$$(\mathbf{u} \cdot \mathbf{A}_{n+1})^2 \leq \|\mathbf{u}\|^2 \|\mathbf{A}_{n+1}\|^2,$$

so we get,

$$\begin{aligned}(\mathbf{u}\cdot\mathbf{A}_{n+1})^2 &\le \|\mathbf{u}\|^2\,\|\mathbf{A}_{n+1}\|^2 = \|\mathbf{u}\|^2\,(a_{n+1\,n+1}^2-1)\\ &\le t^2(a_{n+1\,n+1}^2-1) = t^2a_{n+1\,n+1}^2 - t^2\\ &< t^2a_{n+1\,n+1}^2,\end{aligned}$$

since $t\neq 0$. These calculations imply that

$$(\mathbf{u}\cdot\mathbf{A}_{n+1})^2 - t^2a_{n+1\,n+1}^2 = (\mathbf{u}\cdot\mathbf{A}_{n+1} - ta_{n+1\,n+1})(\mathbf{u}\cdot\mathbf{A}_{n+1} + ta_{n+1\,n+1}) < 0,$$

and that

$$|\mathbf{u}\cdot\mathbf{A}_{n+1}| < |t|a_{n+1\,n+1}.$$

Note that either $(\mathbf{u}\cdot\mathbf{A}_{n+1} - ta_{n+1\,n+1}) < 0$ or $(\mathbf{u}\cdot\mathbf{A}_{n+1} + ta_{n+1\,n+1}) < 0$, but not both. If $t<0$, since $|\mathbf{u}\cdot\mathbf{A}_{n+1}| < |t|a_{n+1\,n+1}$ and $a_{n+1\,n+1}>0$, then $(\mathbf{u}\cdot\mathbf{A}_{n+1} - ta_{n+1\,n+1}) > 0$ and $(\mathbf{u}\cdot\mathbf{A}_{n+1} + ta_{n+1\,n+1}) < 0$. On the other hand, if $t>0$, the fact that $|\mathbf{u}\cdot\mathbf{A}_{n+1}| < |t|a_{n+1\,n+1}$ and $a_{n+1\,n+1}>0$ implies $(\mathbf{u}\cdot\mathbf{A}_{n+1} - ta_{n+1\,n+1}) < 0$ and $(\mathbf{u}\cdot\mathbf{A}_{n+1} + ta_{n+1\,n+1}) > 0$. From this it follows that $\mathbf{u}\cdot\mathbf{A}_{n+1} + a_{n+1,\,n+1}t$ has the same sign as t, since $a_{n+1\,n+1}>0$. Consequently, if $a_{n+1\,n+1}>0$, we see that A maps positive timelike (resp. lightlike) vectors to positive timelike (resp. lightlike) vectors and similarly with negative timelight (resp. lightlike) vectors.

Conversely, as $e_{n+1} = (0,\ldots,0,1)$ is timelike and positive, if A preserves all positive timelike vectors, then Ae_{n+1} is timelike positive, which implies $a_{n+1\,n+1}>0$. □

Let $\mathbf{O}^+(n,1)$ denote the subset of $\mathbf{O}(n,1)$ consisting of all matrices $A=(a_{i\,j})$ such that $a_{n+1\,n+1}>0$. Using Proposition 5.6, we can now show that $\mathbf{O}^+(n,1)$ is a subgroup of $\mathbf{O}(n,1)$ and that $\mathbf{SO}_0(n,1)$ is a subgroup of $\mathbf{SO}(n,1)$. Recall that

$$\mathbf{SO}_0(n,1) = \{A\in\mathbf{SO}(n,1) \mid a_{n+1\,n+1}>0\}.$$

Note that $\mathbf{SO}_0(n,1) = \mathbf{O}^+(n,1)\cap\mathbf{SO}(n,1)$.

Proposition 5.7. *The set $\mathbf{O}^+(n,1)$ is a subgroup of $\mathbf{O}(n,1)$ and the set $\mathbf{SO}_0(n,1)$ is a subgroup of $\mathbf{SO}(n,1)$.*

Proof. Let $A\in\mathbf{O}^+(n,1)\subseteq\mathbf{O}(n,1)$, so that $a_{n+1\,n+1}>0$. The inverse of A in $\mathbf{O}(n,1)$ is $JA^\top J$, where

$$J = \begin{pmatrix} I_n & 0\\ 0 & -1\end{pmatrix},$$

which implies that $a_{n+1\,n+1}^{-1} = a_{n+1\,n+1} > 0$, and so $A^{-1}\in\mathbf{O}^+(n,1)$. If $A,B\in\mathbf{O}^+(n,1)$, then by Proposition 5.6, both A and B preserve all positive timelike vectors, so AB preserves all positive timelike vectors. By Proposition 5.6 again, $AB\in\mathbf{O}^+(n,1)$. Therefore, $\mathbf{O}^+(n,1)$ is a group. But then, $\mathbf{SO}_0(n,1) = \mathbf{O}^+(n,1)\cap\mathbf{SO}(n,1)$ is also a group. □

Since any matrix $A \in \mathbf{SO}_0(n,1)$ preserves the Lorentz inner product and all positive timelike vectors and since $\mathcal{H}_n^+(1)$ consists of timelike vectors, we see that every $A \in \mathbf{SO}_0(n,1)$ maps $\mathcal{H}_n^+(1)$ into itself. Similarly, every $A \in \mathbf{SO}_0(n,1)$ maps $\mathcal{H}_n^-(1)$ into itself. Thus, we can define an action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ by

$$A \cdot u = Au$$

and similarly, we have an action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}_n^-(1) \longrightarrow \mathcal{H}_n^-(1)$.

Proposition 5.8. *The group $\mathbf{SO}_0(n,1)$ is the subgroup of $\mathbf{SO}(n,1)$ that preserves $\mathcal{H}_n^+(1)$ (and $\mathcal{H}_n^-(1)$); that is,*

$$\mathbf{SO}_0(n,1) = \{A \in \mathbf{SO}(n,1) \mid A(\mathcal{H}_n^+(1)) = \mathcal{H}_n^+(1) \quad \text{and} \quad A(\mathcal{H}_n^-(1)) = \mathcal{H}_n^-(1)\}.$$

Proof. We already observed that $A(\mathcal{H}_n^+(1)) = \mathcal{H}_n^+(1)$ if $A \in \mathbf{SO}_0(n,1)$ (and similarly, $A(\mathcal{H}_n^-(1)) = \mathcal{H}_n^-(1)$). Conversely, for any $A \in \mathbf{SO}(n,1)$ such that $A(\mathcal{H}_n^+(1)) = \mathcal{H}_n^+(1)$, as $e_{n+1} = (0,\ldots,0,1) \in \mathcal{H}_n^+(1)$, the vector Ae_{n+1} must be positive timelike, but this says that $a_{n+1\,n+1} > 0$, i.e., $A \in \mathbf{SO}_0(n,1)$. □

Next we wish to prove that the action $\mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ is transitive. For this, we need the next two propositions.

Proposition 5.9. *Let $u = (\mathbf{u}, t)$ and $v = (\mathbf{v}, s)$ be nonzero vectors in $\mathbb{R}^{n+1}$ with $\langle u, v\rangle = 0$. If u is timelike, then v is spacelike (i.e., $\langle v, v\rangle > 0$).*

Proof. Since u is timelike, we have $\|\mathbf{u}\|^2 < t^2$, so $t \neq 0$. The condition $\langle u, v\rangle = 0$ is equivalent to $\mathbf{u}\cdot\mathbf{v} - ts = 0$. If $\mathbf{u} = 0$, then $ts = 0$, and since $t \neq 0$, then $s = 0$. Then $\langle v, v\rangle = \|\mathbf{v}\|^2 - s^2 = \|\mathbf{v}\|^2 > 0$ since v is a nonzero vector in $\mathbb{R}^{n+1}$. We now assume $\mathbf{u} \neq 0$. In this case $\mathbf{u}\cdot\mathbf{v} - ts = 0$, and we get

$$\langle v, v\rangle = \|\mathbf{v}\|^2 - s^2 = \|\mathbf{v}\|^2 - \frac{(\mathbf{u}\cdot\mathbf{v})^2}{t^2}.$$

But when $u \neq 0$ Cauchy-Schwarz implies that $(\mathbf{u}\cdot\mathbf{v})^2/\|\mathbf{u}\|^2 \leq \|\mathbf{v}\|^2$, so we get

$$\langle v, v\rangle = \|\mathbf{v}\|^2 - \frac{(\mathbf{u}\cdot\mathbf{v})^2}{t^2} > \|\mathbf{v}\|^2 - \frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{u}\|^2} \geq 0,$$

as $\|\mathbf{u}\|^2 < t^2$. □

Lemma 5.9 also holds if $u = (\mathbf{u}, t)$ is a nonzero isotropic vector and $v = (\mathbf{v}, s)$ is a nonzero vector that is not collinear with u: If $\langle u, v\rangle = 0$, then v is spacelike (i.e., $\langle v, v\rangle > 0$). The proof is left as an exercise to the reader.

Proposition 5.10. *The action $\mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ is transitive.*

Proof. Let $e_{n+1} = (0,\ldots,0,1) \in \mathcal{H}_n^+(1)$. It is enough to prove that for every $u = (\mathbf{u}, t) \in \mathcal{H}_n^+(1)$, there is some $A \in \mathbf{SO}_0(n,1)$ such that $Ae_{n+1} = u$. By hypothesis,

$$\langle u, u\rangle = \|\mathbf{u}\|^2 - t^2 = -1.$$

We show that we can construct an orthonormal basis, $e_1, \ldots, e_n, u$, with respect to the Lorentz inner product. Consider the hyperplane

$$H = \{v \in \mathbb{R}^{n+1} \mid \langle u, v\rangle = 0\}.$$

Since u is timelike, by Proposition 5.9, every nonzero vector $v \in H$ is spacelike, that is $\langle v, v\rangle > 0$. Let $v_1, \ldots, v_n$ be a basis of H. Since all (nonzero) vectors in H are spacelike, we can apply the Gram-Schmidt orthonormalization procedure and we get a basis $e_1, \ldots, e_n$ of H, such that

$$\langle e_i, e_j\rangle = \delta_{i\,j}, \quad 1 \leq i, j \leq n.$$

By construction, we also have

$$\langle e_i, u\rangle = 0, \quad 1 \leq i \leq n, \quad \text{and} \quad \langle u, u\rangle = -1.$$

Therefore, $e_1, \ldots, e_n, u$ are the column vectors of a Lorentz matrix A such that $Ae_{n+1} = u$, proving our assertion. ☐

Let us find the stabilizer of $e_{n+1} = (0,\ldots,0,1)$. We must have $Ae_{n+1} = e_{n+1}$, and the polar form implies that

$$A = \begin{pmatrix} P & 0 \\ 0 & 1 \end{pmatrix}, \quad \text{with} \quad P \in \mathbf{SO}(n).$$

Therefore, the stabilizer of e_{n+1} is isomorphic to $\mathbf{SO}(n)$, and we conclude that $\mathcal{H}_n^+(1)$, as a homogeneous space, is

$$\mathcal{H}_n^+(1) \cong \mathbf{SO}_0(n,1)/\mathbf{SO}(n).$$

We will return to this homogeneous space in Chapter 22, and see that it is actually a symmetric space.

We end this section by showing that the Lorentz group $\mathbf{SO}_0(n,1)$ is connected. Firstly, it is easy to check that $\mathbf{SO}_0(n,1)$ and $\mathcal{H}_n^+(1)$ satisfy the assumptions of Theorem 4.14 because they are both manifolds, although this notion has not been discussed yet (but will be in Chapter 7). Since the action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ of $\mathbf{SO}_0(n,1)$ on $\mathcal{H}_n^+(1)$ is transitive, Theorem 4.14 implies that as topological spaces,

$$\mathbf{SO}_0(n,1)/\mathbf{SO}(n) \cong \mathcal{H}_n^+(1).$$

We already showed that $\mathcal{H}_n^+(1)$ is connected, so by Proposition 4.8, the connectivity of $\mathbf{SO}_0(n,1)$ follows from the connectivity of $\mathbf{SO}(n)$ for $n \geq 1$. The connectivity of $\mathbf{SO}(n)$ is a consequence of the surjectivity of the exponential map (for instance, see Gallier [48], Chapter 14) but we can also give a quick proof using Proposition 4.8. Indeed, $\mathbf{SO}(n+1)$ and S^n are both manifolds and we saw in Section 4.2 that

$$\mathbf{SO}(n+1)/\mathbf{SO}(n) \cong S^n.$$

Now, S^n is connected for $n \geq 1$ and $\mathbf{SO}(1) \cong S^1$ is connected. We finish the proof by induction on n.

Corollary 5.11. *The Lorentz group $\mathbf{SO}_0(n,1)$ is connected; it is the component of the identity in $\mathbf{O}(n,1)$.*

5.2 The Lie Algebra of the Lorentz Group $\mathbf{SO}_0(n,1)$

In this section we take a closer look at the Lorentz group $\mathbf{SO}_0(n,1)$, and in particular, at the relationship between $\mathbf{SO}_0(n,1)$ and its Lie algebra $\mathfrak{so}(n,1)$. The Lie algebra of $\mathbf{SO}_0(n,1)$ is easily determined by computing the tangent vectors to curves $t \mapsto A(t)$ on $\mathbf{SO}_0(n,1)$ through the identity I. Since $A(t)$ satisfies

$$A^\top JA = J, \qquad J = I_{n,1} = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

differentiating and using the fact that $A(0) = I$, we get

$$A'^\top J + JA' = 0.$$

Therefore,

$$\mathfrak{so}(n,1) = \{A \in \mathrm{M}_{n+1}(\mathbb{R}) \mid A^\top J + JA = 0\}.$$

Since $J = J^\top$, this means that JA is skew-symmetric, and so

$$\mathfrak{so}(n,1) = \left\{ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} \in \mathrm{M}_{n+1}(\mathbb{R}) \mid u \in \mathbb{R}^n, \quad B^\top = -B \right\}.$$

Since $J^2 = I$, the condition $A^\top J + JA = 0$ is equivalent to

$$A^\top = -JAJ.$$

Observe that every matrix $A \in \mathfrak{so}(n,1)$ can be written uniquely as

$$\begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix},$$

where the first matrix is skew-symmetric, the second one is symmetric, and both belong to $\mathfrak{so}(n,1)$. Thus, it is natural to define

$$\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \mid B \in \mathrm{M}_n(\mathbb{R}),\ B^\top = -B \right\},$$

and

$$\mathfrak{p} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\}.$$

It is immediately verified that both $\mathfrak{k}$ and $\mathfrak{p}$ are subspaces of $\mathfrak{so}(n,1)$ (as vector spaces) and that $\mathfrak{k}$ is a Lie subalgebra isomorphic to $\mathfrak{so}(n)$, but $\mathfrak{p}$ is *not* a Lie subalgebra of $\mathfrak{so}(n,1)$ because it is not closed under the Lie bracket. Still, we have

$$[\mathfrak{k},\mathfrak{k}] \subseteq \mathfrak{k}, \quad [\mathfrak{k},\mathfrak{p}] \subseteq \mathfrak{p}, \quad [\mathfrak{p},\mathfrak{p}] \subseteq \mathfrak{k}.$$

Clearly, we have the direct sum decomposition

$$\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{p},$$

known as *Cartan decomposition*.

There is also an automorphism of $\mathfrak{so}(n,1)$ known as the *Cartan involution*, namely

$$\theta(A) = -A^\top = JAJ,$$

and we see that

$$\mathfrak{k} = \{A \in \mathfrak{so}(n,1) \mid \theta(A) = A\} \quad \text{and} \quad \mathfrak{p} = \{A \in \mathfrak{so}(n,1) \mid \theta(A) = -A\}.$$

The involution θ defined on $\mathfrak{so}(n,1)$ is the derivative at I of the involutive isomorphism σ of the group $\mathbf{SO}_0(n,1)$ also defined by

$$\sigma(A) = JAJ, \quad A \in \mathbf{SO}_0(n,1).$$

To justify this claim, let $\gamma(t)$ be a curve in $\mathbf{SO}_0(n,1)$ through I. Define $h(t) = \sigma \circ \gamma(t) = J\gamma(t)J$. The product rule implies $h'(0) = J\gamma'(0)J$. On the other hand, the chain rule implies $h'(0) = D\sigma_I \circ \gamma'(0)$. Combining the two equivalent forms of $h'(0)$ implies $D\sigma_I(X) = JXJ$, whenever $X \in \mathbf{SO}_0(n,1)$.

Since the inverse of an element $A \in \mathbf{SO}_0(n,1)$ is given by $A^{-1} = JA^\top J$, we see that σ is also given by

$$\sigma(A) = (A^{-1})^\top.$$

Unfortunately, there does not appear to be any simple way of obtaining a formula for $\exp(A)$, where $A \in \mathfrak{so}(n,1)$ (except for small n–there is such a formula for $n = 3$ due to Chris Geyer). However, it is possible to obtain an explicit formula for the matrices in $\mathfrak{p}$. This is because for such matrices A, if we let $\omega = \|u\| = \sqrt{u^\top u}$, we have

$$A^3 = \omega^2 A.$$

Thus we get

Proposition 5.12. *For every matrix $A \in \mathfrak{p}$ of the form*

$$A = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix},$$

we have

$$e^A = \begin{pmatrix} I_n + \frac{(\cosh \omega - 1)}{\omega^2} uu^\top & \frac{\sinh \omega}{\omega} u \\ \frac{\sinh \omega}{\omega} u^\top & \cosh \omega \end{pmatrix} = \begin{pmatrix} \sqrt{I_n + \frac{\sinh^2 \omega}{\omega^2} uu^\top} & \frac{\sinh \omega}{\omega} u \\ \frac{\sinh \omega}{\omega} u^\top & \cosh \omega \end{pmatrix}.$$

Proof. Using the fact that $A^3 = \omega^2 A$, we easily prove (by adjusting the calculations of Section 1.1) that

$$e^A = I + \frac{\sinh \omega}{\omega} A + \frac{\cosh \omega - 1}{\omega^2} A^2,$$

which is the first equation of the proposition, since

$$A^2 = \begin{pmatrix} uu^\top & 0 \\ 0 & u^\top u \end{pmatrix} = \begin{pmatrix} uu^\top & 0 \\ 0 & \omega^2 \end{pmatrix}.$$

We leave as an exercise the fact that

$$\left(I_n + \frac{(\cosh \omega - 1)}{\omega^2} uu^\top \right)^2 = I_n + \frac{\sinh^2 \omega}{\omega^2} uu^\top.$$

□

It clear from the above formula that each e^B with $B \in \mathfrak{p}$ is a Lorentz boost. Conversely, every Lorentz boost is the exponential of some $B \in \mathfrak{p}$, as shown below.

Proposition 5.13. *Every Lorentz boost*

$$A = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

with $c = \sqrt{\|v\|^2 + 1}$, is of the form $A = e^B$ for some $B \in \mathfrak{p}$; that is, for some $B \in \mathfrak{so}(n, 1)$ of the form

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}.$$

Proof. Given

$$A = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

we need to find some

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}$$

such that $A = e^B$. This is done by solving the equation

$$\begin{pmatrix} \sqrt{I_n + \frac{\sinh^2 \omega}{\omega^2} uu^\top} & \frac{\sinh \omega}{\omega} u \\ \frac{\sinh \omega}{\omega} u^\top & \cosh \omega \end{pmatrix} = \begin{pmatrix} \sqrt{I_n + vv^\top} & v \\ v^\top & c \end{pmatrix},$$

with $\omega = \|u\|$ and $c = \sqrt{\|v\|^2 + 1}$. When $v = 0$, we have $A = I$, and the matrix $B = 0$ corresponding to $u = 0$ works. So assume $v \neq 0$. In this case, $c > 1$. We have to solve the equation $\cosh \omega = c$, that is,

$$e^{2\omega} - 2ce^{\omega} + 1 = 0.$$

The roots of the corresponding algebraic equation $X^2 - 2cX + 1 = 0$ are

$$X = c \pm \sqrt{c^2 - 1}.$$

As $c > 1$, both roots are strictly positive, so we can solve for ω, say $\omega = \log(c + \sqrt{c^2 - 1}) \neq 0$. Then, $\sinh \omega \neq 0$, so we can solve the equation

$$\frac{\sinh \omega}{\omega} u = v$$

for u, which yields a $B \in \mathfrak{so}(n, 1)$ of the right form with $A = e^B$. □

Combining Proposition 5.2 and Proposition 5.13, we have the corollary:

Corollary 5.14. *Every matrix $A \in \mathbf{O}(n, 1)$ can be written as*

$$A = \begin{pmatrix} Q & 0 \\ 0 & \epsilon \end{pmatrix} e^{\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}},$$

where $Q \in \mathbf{O}(n)$, $\epsilon = \pm 1$, and $u \in \mathbb{R}^n$.

Remarks:

(1) It is easy to show that the eigenvalues of matrices

$$B = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}$$

are 0, with multiplicity $n - 1$, $\|u\|$, and $-\|u\|$. In particular, the eigenvalue relation

$$\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \lambda \begin{pmatrix} c \\ d \end{pmatrix}, \qquad c \in \mathbb{R}^n,\ d, \lambda \in \mathbb{R}$$

implies

$$du = \lambda c, \qquad u^\top c = \lambda d.$$

If $\lambda \neq 0$, $c = \frac{du}{\lambda}$, which in turn implies $u^\top ud = \lambda^2 d$, i.e. $\lambda^2 = u^\top u = \|u\|^2$. If $\lambda = 0$, $u^\top c = 0$, which implies that c is in the $n - 1$-dimensional hyperplane perpendicular to u. Eigenvectors are then easily determined.

(2) The matrices $B \in \mathfrak{so}(n,1)$ of the form

$$B = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \alpha \\ 0 & \cdots & \alpha & 0 \end{pmatrix}$$

are easily seen to form an abelian Lie subalgebra $\mathfrak{a}$ of $\mathfrak{so}(n,1)$ (which means that for all $B, C \in \mathfrak{a}$, $[B,C] = 0$, i.e., $BC = CB$). Proposition 5.12 implies that any $B \in \mathfrak{a}$ as above, we get

$$e^B = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & \cosh\alpha & \sinh\alpha \\ 0 & \cdots & 0 & \sinh\alpha & \cosh\alpha \end{pmatrix}$$

The matrices of the form e^B with $B \in \mathfrak{a}$ form an abelian subgroup A of $\mathbf{SO}_0(n,1)$ isomorphic to $\mathbf{SO}_0(1,1)$. As we already know, the matrices $B \in \mathfrak{so}(n,1)$ of the form

$$\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix},$$

where B is skew-symmetric, form a Lie subalgebra $\mathfrak{k}$ of $\mathfrak{so}(n,1)$. Clearly, $\mathfrak{k}$ is isomorphic to $\mathfrak{so}(n)$, and using the exponential, we get a subgroup K of $\mathbf{SO}_0(n,1)$ isomorphic to $\mathbf{SO}(n)$. It is also clear that $\mathfrak{k} \cap \mathfrak{a} = (0)$, but $\mathfrak{k} \oplus \mathfrak{a}$ is *not* equal to $\mathfrak{so}(n,1)$. What is the missing piece?

Consider the matrices $N \in \mathfrak{so}(n,1)$ of the form

$$N = \begin{pmatrix} 0 & -u & u \\ u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix},$$

where $u \in \mathbb{R}^{n-1}$. The reader should check that these matrices form an abelian Lie subalgebra $\mathfrak{n}$ of $\mathfrak{so}(n,1)$. Furthermore, since

$$\begin{aligned} \mathfrak{so}(n,1) &= \begin{pmatrix} B_1 & u_1 & u \\ -u_1^\top & 0 & \alpha \\ u^\top & \alpha & 0 \end{pmatrix} \\ &= \begin{pmatrix} B_1 & u_1 + u & 0 \\ -u_1^\top - u^\top & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha \\ 0 & \alpha & 0 \end{pmatrix} + \begin{pmatrix} 0 & -u & u \\ u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix}, \end{aligned}$$

where $B_1 \in \mathfrak{so}(n-1)$, $u, u_1 \in \mathbb{R}^{n-1}$, and $\alpha \in \mathbb{R}$, we conclude that

$$\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{a} \oplus \mathfrak{n}.$$

This is the *Iwasawa decomposition* of the Lie algebra $\mathfrak{so}(n,1)$. Furthermore, the reader should check that every $N \in \mathfrak{n}$ is nilpotent; in fact, $N^3 = 0$. (It turns out that $\mathfrak{n}$ is a nilpotent Lie algebra, see Knapp [68]).

The connected Lie subgroup of $\mathbf{SO}_0(n,1)$ associated with $\mathfrak{n}$ is denoted N and it can be shown that we have the *Iwasawa decomposition* of the Lie group $\mathbf{SO}_0(n,1)$:

$$\mathbf{SO}_0(n,1) = KAN.$$

It is easy to check that $[\mathfrak{a}, \mathfrak{n}] \subseteq \mathfrak{n}$, so $\mathfrak{a} \oplus \mathfrak{n}$ is a Lie subalgebra of $\mathfrak{so}(n,1)$ and $\mathfrak{n}$ is an ideal of $\mathfrak{a} \oplus \mathfrak{n}$. This implies that N is normal in the group corresponding to $\mathfrak{a} \oplus \mathfrak{n}$, so AN is a subgroup (in fact, solvable) of $\mathbf{SO}_0(n,1)$. For more on the Iwasawa decomposition, see Knapp [68].

Observe that the image $\overline{\mathfrak{n}}$ of $\mathfrak{n}$ under the Cartan involution θ is the Lie subalgebra

$$\overline{\mathfrak{n}} = \left\{ \begin{pmatrix} 0 & u & u \\ -u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix} \mid u \in \mathbb{R}^{n-1} \right\}.$$

By using the Iwasawa decomposition, we can show that the centralizer of $\mathfrak{a}$, namely $\{m \in \mathfrak{so}(n,1) \mid ma = am \text{ whenever } a \in \mathfrak{a}\}$, is the Lie subalgebra

$$\mathfrak{m} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \in \mathrm{M}_{n+1}(\mathbb{R}) \mid B \in \mathfrak{so}(n-1) \right\}.$$

Hence

$$\mathfrak{so}(n,1) = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n} \oplus \overline{\mathfrak{n}},$$

since

$$\begin{aligned}
\mathfrak{so}(n,1) &= \begin{pmatrix} B_1 & u_1 & u \\ -u_1^\top & 0 & \alpha \\ u^\top & \alpha & 0 \end{pmatrix} \\
&= \begin{pmatrix} B_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \alpha \\ 0 & \alpha & 0 \end{pmatrix} + \begin{pmatrix} 0 & (u_1 - u)/2 & (u - u_1)/2 \\ (u^\top - u_1^\top)/2 & 0 & 0 \\ (u^\top - u_1^\top)/2 & 0 & 0 \end{pmatrix} \\
&+ \begin{pmatrix} 0 & (u + u_1)/2 & (u + u_1)/2 \\ (-u^\top - u_1^\top)/2 & 0 & 0 \\ (u^\top + u_1^\top)/2 & 0 & 0 \end{pmatrix}
\end{aligned}$$

where $B_1 \in \mathfrak{so}(n-1)$, $u, u_1 \in \mathbb{R}^{n-1}$, and $\alpha \in \mathbb{R}$. We also have

$$[\mathfrak{m}, \mathfrak{n}] \subseteq \mathfrak{n},$$

so $\mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n}$ is a subalgebra of $\mathfrak{so}(n, 1)$.

The group M associated with $\mathfrak{m}$ is isomorphic to $\mathbf{SO}(n-1)$, and it can be shown that $B = MAN$ is a subgroup of $\mathbf{SO}_0(n, 1)$. In fact,

$$\mathbf{SO}_0(n, 1)/(MAN) = KAN/MAN = K/M = \mathbf{SO}(n)/\mathbf{SO}(n-1) = S^{n-1}.$$

It is customary to denote the subalgebra $\mathfrak{m} \oplus \mathfrak{a}$ by $\mathfrak{g}_0$, the algebra $\mathfrak{n}$ by $\mathfrak{g}_1$, and $\overline{\mathfrak{n}}$ by $\mathfrak{g}_{-1}$, so that $\mathfrak{so}(n, 1) = \mathfrak{m} \oplus \mathfrak{a} \oplus \mathfrak{n} \oplus \overline{\mathfrak{n}}$ is also written

$$\mathfrak{so}(n, 1) = \mathfrak{g}_0 \oplus \mathfrak{g}_{-1} \oplus \mathfrak{g}_1.$$

By the way, if $N \in \mathfrak{n}$, then

$$e^N = I + N + \frac{1}{2}N^2,$$

and since $N + \frac{1}{2}N^2$ is also nilpotent, e^N can't be diagonalized when $N \neq 0$. This provides a simple example of matrices in $\mathbf{SO}_0(n, 1)$ that can't be diagonalized.

Observe that Corollary 5.14 proves that every matrix $A \in \mathbf{SO}_0(n, 1)$ can be written as

$$A = Pe^S, \quad \text{with } P \in K \cong \mathbf{SO}(n) \text{ and } S \in \mathfrak{p},$$

i.e.,

$$\mathbf{SO}_0(n, 1) = K \exp(\mathfrak{p}),$$

a version of the polar decomposition for $\mathbf{SO}_0(n, 1)$.

5.3 The Surjectivity of $\exp\colon \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3)$

It is known that the exponential map $\exp\colon \mathfrak{so}(n) \to \mathbf{SO}(n)$ is surjective. So when $A \in \mathbf{SO}_0(n, 1)$, since then $Q \in \mathbf{SO}(n)$ and $\epsilon = +1$, the matrix

$$\begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix}$$

is the exponential of some skew symmetric matrix

$$C = \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \in \mathfrak{so}(n, 1),$$

and we can write $A = e^C e^Z$, with $C \in \mathfrak{k}$ and $Z \in \mathfrak{p}$. Unfortunately, C and Z generally don't commute, so it is generally not true that $A = e^{C+Z}$. Thus, we don't get an "easy" proof of the surjectivity of the exponential, $\exp\colon \mathfrak{so}(n,1) \to \mathbf{SO}_0(n,1)$.

This is not too surprising because to the best of our knowledge, proving surjectivity for all n is not a simple matter. One proof is due to Nishikawa [90] (1983). Nishikawa's paper is rather short, but this is misleading. Indeed, Nishikawa relies on a classic paper by Djokovic [37], which itself relies heavily on another fundamental paper by Burgoyne and Cushman [26], published in 1977. Burgoyne and Cushman determine the conjugacy classes for some linear Lie groups and their Lie algebras, where the linear groups arise from an inner product space (real or complex). This inner product is nondegenerate, symmetric, or Hermitian or skew-symmetric or skew-Hermitian. Altogether, one has to read over 40 pages to fully understand the proof of surjectivity.

In his introduction, Nishikawa states that he is not aware of any other proof of the surjectivity of the exponential for $\mathbf{SO}_0(n,1)$. However, such a proof was also given by Marcel Riesz as early as 1957, in some lectures notes that he gave while visiting the University of Maryland in 1957-1958. These notes were probably not easily available until 1993, when they were published in book form, with commentaries, by Bolinder and Lounesto [97].

Interestingly, these two proofs use very different methods. The Nishikawa–Djokovic–Burgoyne and Cushman proof makes heavy use of methods in Lie groups and Lie algebra, although not far beyond linear algebra. Riesz's proof begins with a deep study of the structure of the minimal polynomial of a Lorentz isometry (Chapter III). This is a beautiful argument that takes about 10 pages. The story is not over, as it takes most of Chapter IV (some 40 pages) to prove the surjectivity of the exponential (actually, Riesz proves other things along the way). In any case, the reader can see that both proofs are quite involved.

It is worth noting that Milnor (1969) also uses techniques very similar to those used by Riesz (in dealing with minimal polynomials of isometries) in his paper on isometries of inner product spaces [82].

What we will do to close this section is to give a relatively simple proof that the exponential map $\exp\colon \mathfrak{so}(1,3) \to \mathbf{SO}_0(1,3)$ is surjective. The reader may wonder why we are considering the groups $\mathbf{SO}_0(1,3)$ instead of the group $\mathbf{SO}_0(3,1)$. This is simply a matter of technical convenience, for instance, in the proof of Proposition 5.17.

In the case of $\mathbf{SO}_0(1,3)$, we can use the fact that $\mathbf{SL}(2,\mathbb{C})$ is a two-sheeted covering space of $\mathbf{SO}_0(1,3)$, which means that there is a homomorphism $\phi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ which is surjective and that $\mathrm{Ker}\,\phi = \{-I, I\}$. Then the small miracle is that, although the exponential $\exp\colon \mathfrak{sl}(2,\mathbb{C}) \to \mathbf{SL}(2,\mathbb{C})$ is *not* surjective, for every $A \in \mathbf{SL}(2,\mathbb{C})$, *either A or $-A$ is in the image of the exponential!*

Proposition 5.15. *Given any matrix*

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{C}),$$

let ω be any of the two complex roots of $a^2 + bc$. If $\omega \neq 0$, then

$$e^B = \cosh \omega \, I + \frac{\sinh \omega}{\omega} \, B,$$

and $e^B = I + B$ if $a^2 + bc = 0$. Furthermore, every matrix $A \in \mathbf{SL}(2, \mathbb{C})$ is in the image of the exponential map, unless $A = -I + N$, where N is a nonzero nilpotent (i.e., $N^2 = 0$ with $N \neq 0$). Consequently, for any $A \in \mathbf{SL}(2, \mathbb{C})$, either A or $-A$ is of the form e^B, for some $B \in \mathfrak{sl}(2, \mathbb{C})$.

Proof. Observe that

$$B^2 = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \begin{pmatrix} a & b \\ c & -a \end{pmatrix} = (a^2 + bc) I.$$

Then, it is straightforward to prove that

$$e^B = \cosh \omega \, I + \frac{\sinh \omega}{\omega} \, B,$$

where ω is a square root of $a^2 + bc$ if $\omega \neq 0$, otherwise, $e^B = I + B$.

Let

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \qquad \alpha\delta - \gamma\beta = 1$$

be any matrix in $\mathbf{SL}(2, \mathbb{C})$. We would like to find a matrix $B \in \mathfrak{sl}(2, \mathbb{C})$ so that $A = e^B$. In view of the above, we need to solve the system

$$\begin{aligned} \cosh \omega + \frac{\sinh \omega}{\omega} a &= \alpha \\ \cosh \omega - \frac{\sinh \omega}{\omega} a &= \delta \\ \frac{\sinh \omega}{\omega} b &= \beta \\ \frac{\sinh \omega}{\omega} c &= \gamma \end{aligned}$$

for a, b, c, and ω. From the first two equations we get

$$\begin{aligned} \cosh \omega &= \frac{\alpha + \delta}{2} \\ \frac{\sinh \omega}{\omega} a &= \frac{\alpha - \delta}{2}. \end{aligned}$$

Thus, we see that we need to know whether complex cosh is surjective and when complex sinh is zero. We claim:

(1) cosh is surjective.

(2) $\sinh z = 0$ iff $z = n\pi i$, where $n \in \mathbb{Z}$.

Given any $c \in \mathbb{C}$, we have $\cosh \omega = c$ iff

$$e^{2\omega} - 2e^{\omega}c + 1 = 0.$$

The corresponding algebraic equation

$$Z^2 - 2cZ + 1 = 0$$

has discriminant $4(c^2 - 1)$ and it has two complex roots

$$Z = c \pm \sqrt{c^2 - 1}$$

where $\sqrt{c^2-1}$ is some square root of $c^2 - 1$. Observe that these roots are *never zero*. Therefore, we can find a complex log of $c + \sqrt{c^2-1}$, say ω, so that $e^{\omega} = c + \sqrt{c^2-1}$ is a solution of $e^{2\omega} - 2e^{\omega}c + 1 = 0$. This proves the surjectivity of cosh.

We have $\sinh \omega = 0$ iff $e^{2\omega} = 1$; this holds iff $2\omega = n2\pi i$, i.e., $\omega = n\pi i$.

Observe that

$$\frac{\sinh n\pi i}{n\pi i} = 0 \quad \text{if } n \neq 0, \text{ but} \quad \frac{\sinh n\pi i}{n\pi i} = 1 \quad \text{when } n = 0.$$

We know that

$$\cosh \omega = \frac{\alpha + \delta}{2}$$

can always be solved.

Case 1. If $\omega \neq n\pi i$, with $n \neq 0$, then

$$\frac{\sinh \omega}{\omega} \neq 0$$

and the other equations can also be solved (this includes the case $\omega = 0$). We still have to check that

$$a^2 + bc = \omega^2.$$

This is because, using the fact that $\cosh \omega = \frac{\alpha+\delta}{2}$, $\alpha\delta - \beta\gamma = 1$, and $\cosh^2 \omega - \sinh^2 \omega = 1$, we have

$$\begin{aligned}
a^2 + bc &= \frac{(\alpha - \delta)^2\omega^2}{4\sinh^2 \omega} + \frac{\beta\gamma\omega^2}{\sinh^2 \omega} \\
&= \frac{\omega^2(\alpha^2 + \delta^2 - 2\alpha\delta + 4\beta\gamma)}{4\sinh^2 \omega} \\
&= \frac{\omega^2(\alpha^2 + \delta^2 + 2\alpha\delta - 4(\alpha\delta - \beta\gamma))}{4\sinh^2 \omega} \\
&= \frac{\omega^2((\alpha + \delta)^2 - 4(\alpha\delta - \beta\gamma))}{4\sinh^2 \omega} \\
&= \frac{4\omega^2(\cosh^2 \omega - 1)}{4\sinh^2 \omega} \\
&= \omega^2.
\end{aligned}$$

Therefore, in this case, the exponential is surjective. It remains to examine the other case.

Case 2. Assume $\omega = n\pi i$, with $n \neq 0$. If n is even, then $e^{\omega} = 1$, which implies

$$\alpha + \delta = 2.$$

However, $\alpha\delta - \beta\gamma = 1$ (since $A \in \mathbf{SL}(2, \mathbb{C})$), so from the facts that $\det(A)$ is the product of the eigenvalues and $\mathrm{tr}(A)$ is the sum of the eigenvalues, we deduce that A has the double eigenvalue 1. Thus, $N = A - I$ is nilpotent (i.e., $N^2 = 0$) and has zero trace; but then, $N \in \mathfrak{sl}(2, \mathbb{C})$ and

$$e^N = I + N = I + A - I = A.$$

If n is odd, then $e^{\omega} = -1$, which implies

$$\alpha + \delta = -2.$$

In this case, A has the double eigenvalue -1 and $A + I = N$ is nilpotent. So $A = -I + N$, where N is nilpotent. If $N \neq 0$, then A cannot be diagonalized. We claim that there is no $B \in \mathfrak{sl}(2, \mathbb{C})$ so that $e^B = A$.

Indeed, any matrix $B \in \mathfrak{sl}(2, \mathbb{C})$ has zero trace, which means that if λ_1 and λ_2 are the eigenvalues of B, then $\lambda_1 = -\lambda_2$. If $\lambda_1 \neq 0$, then $\lambda_1 \neq \lambda_2$ so B can be diagonalized, but then Proposition 1.4 implies that e^B can also be diagonalized, contradicting the fact that A can't be diagonalized. If $\lambda_1 = \lambda_2 = 0$, then e^B has the double eigenvalue $+1$, but by Proposition 1.4, A has eigenvalues -1. Therefore, the only matrices $A \in \mathbf{SL}(2, \mathbb{C})$ that are not in the image of the exponential are those of the form $A = -I + N$, where N is a nonzero nilpotent. However, note that $-A = I - N$ *is* in the image of the exponential. □

Remark: If we restrict our attention to $\mathbf{SL}(2, \mathbb{R})$, then we have the following proposition that can be used to prove that the exponential map $\exp\colon \mathfrak{so}(1, 2) \to \mathbf{SO}_0(1, 2)$ is surjective:

Proposition 5.16. *Given any matrix*

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{R}),$$

if $a^2 + bc > 0$, then let $\omega = \sqrt{a^2 + bc} > 0$, and if $a^2 + bc < 0$, then let $\omega = \sqrt{-(a^2 + bc)} > 0$ (i.e., $\omega^2 = -(a^2 + bc)$). In the first case ($a^2 + bc > 0$), we have

$$e^B = \cosh \omega \, I + \frac{\sinh \omega}{\omega} B,$$

and in the second case ($a^2 + bc < 0$), we have

$$e^B = \cos \omega \, I + \frac{\sin \omega}{\omega} B.$$

If $a^2 + bc = 0$, then $e^B = I + B$. Furthermore, every matrix $A \in \mathbf{SL}(2, \mathbb{R})$ whose trace satisfies $\mathrm{tr}(A) \geq -2$ is in the image of the exponential map, unless $A = -I + N$ with $N \neq 0$ nilpotent. Consequently, for any $A \in \mathbf{SL}(2, \mathbb{R})$, either A or $-A$ is of the form e^B, for some $B \in \mathfrak{sl}(2, \mathbb{R})$.

Proof. For any matrix

$$B = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2,\mathbb{R}),$$

some simple calculations show that if $a^2 + bc > 0$, then

$$e^B = \cosh\omega\, I + \frac{\sinh\,\omega}{\omega}\, B$$

with $\omega = \sqrt{a^2 + bc} > 0$, and if $a^2 + bc < 0$, then

$$e^B = \cos\omega\, I + \frac{\sin\,\omega}{\omega}\, B$$

with $\omega = \sqrt{-(a^2 + bc)} > 0$ (and $e^B = I + B$ when $a^2 + bc = 0$). Let

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \qquad \alpha\delta - \beta\gamma = 1$$

be any matrix in $\mathbf{SL}(2,\mathbb{R})$.

First, assume that $\mathrm{tr}(A) = \alpha + \delta > 2$. We would like to find a matrix $B \in \mathfrak{sl}(2,\mathbb{R})$ so that $A = e^B$. In view of the above, we need to solve the system

$$\begin{aligned} \cosh\omega + \frac{\sinh\omega}{\omega} a &= \alpha \\ \cosh\omega - \frac{\sinh\omega}{\omega} a &= \delta \\ \frac{\sinh\omega}{\omega} b &= \beta \\ \frac{\sinh\omega}{\omega} c &= \gamma \end{aligned}$$

for a, b, c, and ω. From the first two equations we get

$$\begin{aligned} \cosh\omega &= \frac{\alpha+\delta}{2} \\ \frac{\sinh\omega}{\omega} a &= \frac{\alpha-\delta}{2}. \end{aligned}$$

As in the proof of Proposition 5.15, $\cosh\omega = c$ iff e^ω is a root of the quadratic equation

$$Z^2 - 2cZ + 1 = 0.$$

This equation has real roots iff $c^2 \geq 1$, and since $c = \frac{\alpha+\delta}{2}$ and $\alpha + \delta > 2$, our equation has real roots. Furthermore, the root $c + \sqrt{c^2 - 1}$ is greater than 1, so $\log c$ is a positive real number. Then, as in the proof of Proposition 5.15, we find solutions of our system above. Moreover, these solutions are real and satisfy $a^2 + bc = \omega^2$.

Let us now consider the case where $-2 \leq \alpha+\delta \leq 2$. This time we try to solve the system

$$\begin{aligned} \cos\omega + \frac{\sin\omega}{\omega}a &= \alpha \\ \cos\omega - \frac{\sin\omega}{\omega}a &= \delta \\ \frac{\sin\omega}{\omega}b &= \beta \\ \frac{\sin\omega}{\omega}c &= \gamma. \end{aligned}$$

We get

$$\begin{aligned} \cos\omega &= \frac{\alpha+\delta}{2} \\ \frac{\sin\omega}{\omega}a &= \frac{\alpha-\delta}{2}. \end{aligned}$$

Because $-2 \leq \alpha+\delta \leq 2$, the first equation has (real) solutions, and we may assume that $0 \leq \omega \leq \pi$.

If $\omega = 0$ is a solution, then $\alpha+\beta = 2$ and we already know via the arguments of Proposition 5.15 that $N = A - I$ is nilpotent and that $e^N = I + N = A$. If $\omega = \pi$, then $\alpha+\beta = -2$ and we know that $N = A+I$ is nilpotent. If $N = 0$, then $A = -I$, and otherwise we already know that $A = -I + N$ is not in the image of the exponential.

If $0 < \omega < \pi$, then $\sin\omega \neq 0$ and the other equations have a solution. We still need to check that

$$a^2 + bc = -\omega^2.$$

Because $\cos\omega = \frac{\alpha+\delta}{2}$, $\alpha\delta - \beta\gamma = 1$ and $\cos^2\omega + \sin^2\omega = 1$, we have

$$\begin{aligned} a^2 + bc &= \frac{(\alpha-\delta)^2\omega^2}{4\sin^2\omega} + \frac{\beta\gamma\omega^2}{\sin^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2-2\alpha\delta+4\beta\gamma)}{4\sinh^2\omega} \\ &= \frac{\omega^2(\alpha^2+\delta^2+2\alpha\delta-4(\alpha\delta-\beta\gamma))}{4\sin^2\omega} \\ &= \frac{\omega^2((\alpha+\delta)^2-4(\alpha\delta-\beta\gamma))}{4\sin^2\omega} \\ &= \frac{4\omega^2(\cos^2\omega-1)}{4\sin^2\omega} \\ &= -\omega^2. \end{aligned}$$

This proves that every matrix $A \in \mathbf{SL}(2,\mathbb{R})$ whose trace satisfies $\mathrm{tr}(A) \geq -2$ is in the image of the exponential map, unless $A = -I + N$ with $N \neq 0$ nilpotent. □

We now return to the relationship between $\mathbf{SL}(2,\mathbb{C})$ and $\mathbf{SO}_0(1,3)$. In order to define a homomorphism $\phi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$, we begin by defining a linear bijection h between $\mathbb{R}^4$ and $\mathbf{H}(2)$, the set of complex 2×2 Hermitian matrices, by

$$(t,x,y,z) \mapsto \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix}.$$

Those familiar with quantum physics will recognize a linear combination of the Pauli matrices! The inverse map is easily defined For instance, given a Hermitian matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad a,d \in \mathbb{R}, \quad c = \overline{b} \in \mathbb{C}$$

by setting

$$\begin{pmatrix} a & b \\ \overline{b} & d \end{pmatrix} = \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix},$$

we find that

$$t = \frac{a+d}{2},\ x = \frac{a-d}{2},\ y = \frac{b+\overline{b}}{2},\ z = \frac{b-\overline{b}}{2i}.$$

For any $A \in \mathbf{SL}(2,\mathbb{C})$, we define a map $l_A\colon \mathbf{H}(2) \to \mathbf{H}(2)$, *via*

$$S \mapsto ASA^*.$$

(Here, $A^* = \overline{A}^\top$.) Using the linear bijection $h\colon \mathbb{R}^4 \to \mathbf{H}(2)$ and its inverse, we obtain a map $\mathrm{lor}_A\colon \mathbb{R}^4 \to \mathbb{R}^4$, where

$$\mathrm{lor}_A = h^{-1} \circ l_A \circ h.$$

As ASA^* is Hermitian, we see that l_A is well defined. It is obviously linear and since $\det(A) = 1$ (recall, $A \in \mathbf{SL}(2,\mathbb{C})$) and

$$\det \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix} = t^2 - x^2 - y^2 - z^2,$$

we see that lor_A preserves the Lorentz metric! Furthermore, it is not hard to prove that $\mathbf{SL}(2,\mathbb{C})$ is connected (use the polar form or analyze the eigenvalues of a matrix in $\mathbf{SL}(2,\mathbb{C})$, for example, as in Duistermatt and Kolk [43] (Chapter 1, Section 1.2)) and that the map $\phi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{GL}(4,\mathbb{R})$ with

$$\phi\colon A \mapsto \mathrm{lor}_A$$

is a continuous group homomorphism. Thus the range of ϕ is a connected subgroup of $\mathbf{SO}_0(1,3)$. This shows that $\phi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ is indeed a homomorphism. It remains to prove that it is surjective and that its kernel is $\{I, -I\}$.

Proposition 5.17. *The homomorphism $\phi\colon \mathbf{SL}(2,\mathbb{C}) \to \mathbf{SO}_0(1,3)$ is surjective and its kernel is $\{I, -I\}$.*

Proof. Recall that from Theorem 5.5, the Lorentz group $\mathbf{SO}_0(1,3)$ is generated by the matrices of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3)$$

and the matrices of the form

$$\begin{pmatrix} \cosh\alpha & \sinh\alpha & 0 & 0 \\ \sinh\alpha & \cosh\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Thus, to prove the surjectivity of ϕ, it is enough to check that the above matrices are in the range of ϕ. For matrices of the second kind

$$A = \begin{pmatrix} e^{\frac{1}{2}\alpha} & 0 \\ 0 & e^{-\frac{1}{2}\alpha} \end{pmatrix}$$

does the job. Let e_1, e_2, e_3, and e_4 be the standard basis for $\mathbb{R}^4$. Then

$$\begin{aligned} \mathrm{lor}_A(e_1) &= h^{-1} \circ l_A \circ h(e_1) = h^{-1} \circ l_A \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ &= h^{-1} \left(\begin{pmatrix} e^{\alpha} & 0 \\ 0 & e^{-\alpha} \end{pmatrix} \right) = \left(\frac{e^{\alpha} + e^{-\alpha}}{2}, \frac{e^{\alpha} - e^{-\alpha}}{2}, 0, 0 \right) \\ &= (\cosh\alpha, \sinh\alpha, 0, 0). \end{aligned}$$

Similar calculations show that

$$\begin{aligned} &\mathrm{lor}_A(e_2) = (\sinh\alpha, \cosh\alpha, 0, 0) \\ &\mathrm{lor}_A(e_3) = (0,0,1,0) \qquad \mathrm{lor}_A(e_4) = (0,0,0,1). \end{aligned}$$

For matrices of the first kind, we recall that the group of unit quaternions $q = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ can be viewed as $\mathbf{SU}(2)$, *via* the correspondence

$$a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mapsto \begin{pmatrix} a + ib & c + id \\ -c + id & a - ib \end{pmatrix},$$

where $a, b, c, d \in \mathbb{R}$ and $a^2 + b^2 + c^2 + d^2 = 1$. Moreover, the algebra of quaternions $\mathbb{H}$ is the real algebra of matrices as above, without the restriction $a^2 + b^2 + c^2 + d^2 = 1$, and $\mathbb{R}^3$ is embedded in $\mathbb{H}$ as the *pure quaternions*, i.e., those for which $a = 0$. Observe that when $a = 0$,

$$\begin{pmatrix} ib & c + id \\ -c + id & -ib \end{pmatrix} = i \begin{pmatrix} b & d - ic \\ d + ic & -b \end{pmatrix} = ih(0, b, d, c).$$

Therefore, we have a bijection between the pure quaternions and the subspace of the Hermitian matrices

$$\begin{pmatrix} b & d-ic \\ d+ic & -b \end{pmatrix}$$

for which $a = 0$, the inverse being division by i, i.e., multiplication by $-i$. Also, when q is a unit quaternion, let $\overline{q} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$, and observe that $\overline{q} = q^{-1}$. Using the embedding $\mathbb{R}^3 \hookrightarrow \mathbb{H}$, for every unit quaternion $q \in \mathbf{SU}(2)$, define the map $\rho_q\colon \mathbb{R}^3 \to \mathbb{R}^3$ by

$$\rho_q(X) = qX\overline{q} = qXq^{-1},$$

for all $X \in \mathbb{R}^3 \hookrightarrow \mathbb{H}$. It is well known that ρ_q is a rotation (i.e., $\rho_q \in \mathbf{SO}(3)$), and moreover the map $q \mapsto \rho_q$ is a surjective homomorphism $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$, and $\mathrm{Ker}\, \phi = \{I, -I\}$ (For example, see Gallier [48], Chapter 8).

Now consider a matrix A of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} \quad \text{with } P \in \mathbf{SO}(3).$$

We claim that we can find a matrix $B \in \mathbf{SL}(2,\mathbb{C})$, such that $\phi(B) = \mathrm{lor}_B = A$. We claim that we can pick $B \in \mathbf{SU}(2) \subseteq \mathbf{SL}(2,\mathbb{C})$. Indeed, if $B \in \mathbf{SU}(2)$, then $B^* = B^{-1}$, so

$$B \begin{pmatrix} t+x & y-iz \\ y+iz & t-x \end{pmatrix} B^* = t \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - iB \begin{pmatrix} ix & z+iy \\ -z+iy & -ix \end{pmatrix} B^{-1}.$$

The above shows that lor_B leaves the coordinate t invariant. The term

$$B \begin{pmatrix} ix & z+iy \\ -z+iy & -ix \end{pmatrix} B^{-1}$$

is a pure quaternion corresponding to the application of the rotation ρ_B induced by the unit quaternion B to the pure quaternion associated with (x,y,z) and multiplication by $-i$ is just the corresponding Hermitian matrix, as explained above. But, we know that for any $P \in \mathbf{SO}(3)$, there is a unit quaternion B so that $\rho_B = P$, so we can find our $B \in \mathbf{SU}(2)$ so that

$$\mathrm{lor}_B = \begin{pmatrix} 1 & 0 \\ 0 & P \end{pmatrix} = A.$$

Finally, assume that $\phi(A) = \mathrm{lor}_A = I$. This means that

$$ASA^* = S,$$

for all Hermitian matrices S defined above. In particular, for $S = I$, we get $AA^* = I$, i.e., $A \in \mathbf{SU}(2)$. Thus

$$AS = SA$$

for all Hermitian matrices S defined above, so in particular, this holds for diagonal matrices of the form

$$\begin{pmatrix} t+x & 0 \\ 0 & t-x \end{pmatrix},$$

with $t + x \neq t - x$. We deduce that A is a diagonal matrix, and since it is unitary, we must have $A = \pm I$. Therefore, $\mathrm{Ker}\, \phi = \{I, -I\}$. □

Remark: The group $\mathbf{SL}(2, \mathbb{C})$ is isomorphic to the group $\mathbf{Spin}(1, 3)$, which is a (simply-connected) double-cover of $\mathbf{SO}_0(1, 3)$. This is a standard result of Clifford algebra theory; see Bröcker and tom Dieck [24] or Fulton and Harris [46]. What we just did is to provide a direct proof of this fact.

We just proved that there is an isomorphism

$$\mathbf{SL}(2, \mathbb{C})/\{I, -I\} \cong \mathbf{SO}_0(1, 3).$$

However, the reader may recall that $\mathbf{SL}(2, \mathbb{C})/\{I, -I\} = \mathbf{PSL}(2, \mathbb{C}) \cong \mathbf{M\ddot{o}b}^+$. Therefore, the Lorentz group is isomorphic to the Möbius group.

We now have all the tools to prove that the exponential map $\exp\colon \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3)$ is surjective.

Theorem 5.18. *The exponential map* $\exp\colon \mathfrak{so}(1, 3) \to \mathbf{SO}_0(1, 3)$ *is surjective.*

Proof. First recall from Proposition 3.13 that the following diagram commutes:

$$\begin{tikzcd} \mathbf{SL}(2, \mathbb{C}) \arrow[r, "\phi"] & \mathbf{SO}_0(1, 3) . \\ \mathfrak{sl}(2, \mathbb{C}) \arrow[u, "\exp"] \arrow[r, "d\phi_1"'] & \mathfrak{so}(1, 3) \arrow[u, "\exp"'] \end{tikzcd}$$

Pick any $A \in \mathbf{SO}_0(1, 3)$. By Proposition 5.17, the homomorphism ϕ is surjective and as $\mathrm{Ker}\, \phi = \{I, -I\}$, there exists some $B \in \mathbf{SL}(2, \mathbb{C})$ so that

$$\phi(B) = \phi(-B) = A.$$

Now by Proposition 5.15, for any $B \in \mathbf{SL}(2, \mathbb{C})$, either B or $-B$ is of the form e^C, for some $C \in \mathfrak{sl}(2, \mathbb{C})$. By the commutativity of the diagram, if we let $D = d\phi_1(C) \in \mathfrak{so}(1, 3)$, we get

$$A = \phi(\pm e^C) = e^{d\phi_1(C)} = e^D,$$

with $D \in \mathfrak{so}(1, 3)$, as required. □

Remark: We can restrict the bijection $h\colon \mathbb{R}^4 \to \mathbf{H}(2)$ defined earlier to a bijection between $\mathbb{R}^3$ and the space of real symmetric matrices of the form

$$\begin{pmatrix} t+x & y \\ y & t-x \end{pmatrix}.$$

Then, if we also restrict ourselves to $\mathbf{SL}(2,\mathbb{R})$, for any $A \in \mathbf{SL}(2,\mathbb{R})$ and any symmetric matrix S as above, we get a map

$$S \mapsto ASA^{\top}.$$

The reader should check that these transformations correspond to isometries in $\mathbf{SO}_0(1,2)$ and we get a homomorphism $\phi\colon \mathbf{SL}(2,\mathbb{R}) \to \mathbf{SO}_0(1,2)$. Just as $\mathbf{SL}(2,\mathbb{C})$ is connected, the group $\mathbf{SL}(2,\mathbb{R})$ is also connected (but not simply connected, unlike $\mathbf{SL}(2,\mathbb{C})$). Then we have a version of Proposition 5.17 for $\mathbf{SL}(2,\mathbb{R})$ and $\mathbf{SO}_0(1,2)$:

Proposition 5.19. *The homomorphism $\phi\colon \mathbf{SL}(2,\mathbb{R}) \to \mathbf{SO}_0(1,2)$ is surjective and its kernel is $\{I, -I\}$.*

Using Proposition 5.19, Proposition 5.16, and the commutative diagram

$$\begin{tikzcd}
\mathbf{SL}(2,\mathbb{R}) \arrow[r, "\phi"] & \mathbf{SO}_0(1,2) \\
\mathfrak{sl}(2,\mathbb{R}) \arrow[u, "\exp"] \arrow[r, "d\phi_1"'] & \mathfrak{so}(1,2) \arrow[u, "\exp"']
\end{tikzcd},$$

we get a version of Theorem 5.18 for $\mathbf{SO}_0(1,2)$:

Theorem 5.20. *The exponential map $\exp\colon \mathfrak{so}(1,2) \to \mathbf{SO}_0(1,2)$ is surjective.*

Also observe that $\mathbf{SO}_0(1,1)$ consists of the matrices of the form

$$A = \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix},$$

and a direct computation shows that

$$e^{\begin{pmatrix} 0 & \alpha \\ \alpha & 0 \end{pmatrix}} = \begin{pmatrix} \cosh\alpha & \sinh\alpha \\ \sinh\alpha & \cosh\alpha \end{pmatrix}.$$

Thus, we see that the map $\exp\colon \mathfrak{so}(1,1) \to \mathbf{SO}_0(1,1)$ is also surjective. Therefore, we have proved that $\exp\colon \mathfrak{so}(1,n) \to \mathbf{SO}_0(1,n)$ is surjective for $n = 1, 2, 3$. This actually holds for all $n \geq 1$, but the proof is much more involved, as we already discussed earlier.

5.4 Problems

Problem 5.1. Define $\mathfrak{k}$ and $\mathfrak{p}$ by

$$\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \mid B \in \mathrm{M}_n(\mathbb{R}),\ B^\top = -B \right\},$$

and

$$\mathfrak{p} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\}.$$

(1) Check that both $\mathfrak{k}$ and $\mathfrak{p}$ are subspaces of $\mathfrak{so}(n, 1)$ (as vector spaces) and that $\mathfrak{k}$ is a Lie subalgebra isomorphic to $\mathfrak{so}(n)$.

(2) Show that $\mathfrak{p}$ is not a Lie subalgebra of $\mathfrak{so}(n, 1)$ because it is not closed under the Lie bracket. Still, check that

$$[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}, \quad [\mathfrak{k}, \mathfrak{p}] \subseteq \mathfrak{p}, \quad [\mathfrak{p}, \mathfrak{p}] \subseteq \mathfrak{k}.$$

Problem 5.2. Consider the subset $\mathfrak{n}$ of $\mathfrak{so}(n, 1)$ consisting of the matrices of the form

$$N = \begin{pmatrix} 0 & -u & u \\ u^\top & 0 & 0 \\ u^\top & 0 & 0 \end{pmatrix},$$

where $u \in \mathbb{R}^{n-1}$.

(1) Check that $\mathfrak{n}$ is an abelian Lie subalgebra of $\mathfrak{so}(n, 1)$.

(2) Prove that every $N \in \mathfrak{n}$ is nilpotent; in fact, $N^3 = 0$.

(3) Prove that $[\mathfrak{a}, \mathfrak{n}] \subseteq \mathfrak{n}$, and that $\mathfrak{a} \oplus \mathfrak{n}$ is a Lie subalgebra of $\mathfrak{so}(n, 1)$ and $\mathfrak{n}$ is an ideal of $\mathfrak{a} \oplus \mathfrak{n}$.

Problem 5.3. The map

$$(x, y, t) \mapsto \begin{pmatrix} t + x & y \\ y & t - x \end{pmatrix}$$

is a bijection between $\mathbb{R}^3$ and the space of real symmetric matrices of the above form. For any $A \in \mathbf{SL}(2, \mathbb{R})$ and any symmetric matrix S as above, we get a map

$$S \mapsto ASA^\top.$$

(1) Check that these transformations correspond to isometries in $\mathbf{SO}_0(1, 2)$, and that we get a homomorphism $\phi\colon \mathbf{SL}(2, \mathbb{R}) \to \mathbf{SO}_0(1, 2)$.

(2) Prove Proposition 5.19, namely that the homomorphism $\phi\colon \mathbf{SL}(2, \mathbb{R}) \to \mathbf{SO}_0(1, 2)$ is surjective and its kernel is $\{I, -I\}$.

Chapter 6

The Structure of $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$

In this chapter, we take a closer look at the stucture of the groups $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$ (also $\mathbf{SO}_0(p,q)$). We begin with the polar form of matrices in $\mathbf{O}(p,q)$, and then we describe the topological structure of the groups $\mathbf{O}(p,q)$, $\mathbf{SO}(p,q)$, and $\mathbf{SO}_0(p,q)$. For this, we briefly investigate a class of groups called pseudo-algebraic groups.

6.1 Polar Forms for Matrices in $\mathbf{O}(p,q)$

Recall from Section 5.1 that the group $\mathbf{O}(p,q)$ is the set of all $n \times n$-matrices

$$\mathbf{O}(p,q) = \{A \in \mathbf{GL}(n,\mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q}\}.$$

We deduce immediately that $|\det(A)| = 1$, and we also know that $A I_{p,q} A^\top = I_{p,q}$ holds. Unfortunately, when $p \neq 0, 1$ and $q \neq 0, 1$, it does not seem possible to obtain a formula as nice as that given in Proposition 5.2. Nevertheless, we can obtain a formula for a polar form factorization of matrices in $\mathbf{O}(p,q)$.

Recall (for example, see Gallier [48], Chapter 12) that if S is a symmetric positive definite matrix, then there is a unique symmetric positive definite matrix, T, so that

$$S = T^2.$$

We denote T by $S^{\frac{1}{2}}$ or $\sqrt{S}$. By $S^{-\frac{1}{2}}$, we mean the inverse of $S^{\frac{1}{2}}$. In order to obtain the polar form of a matrix in $\mathbf{O}(p,q)$, we begin with the following proposition:

Proposition 6.1. *Every matrix $X \in \mathbf{O}(p,q)$ can be written as*

$$X = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix},$$

where $\alpha = (I_p - Z^\top Z)^{-1}$ and $\delta = (I_q - Z Z^\top)^{-1}$, for some orthogonal matrices $U \in \mathbf{O}(p)$, $V \in \mathbf{O}(q)$ and for some $q \times p$ matrix, Z, such that $I_p - Z^\top Z$ and $I_q - Z Z^\top$ are symmetric positive definite matrices. Moreover, U, V, Z are uniquely determined by X.

Proof. If we write

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

with A a $p \times p$ matrix, D a $q \times q$ matrix, B a $p \times q$ matrix and C a $q \times p$ matrix, then the equations $X^\top I_{p,q} X = I_{p,q}$ and $X I_{p,q} X^\top = I_{p,q}$ yield the (not independent) conditions

$$\begin{aligned} A^\top A &= I_p + C^\top C \\ D^\top D &= I_q + B^\top B \\ A^\top B &= C^\top D \\ AA^\top &= I_p + BB^\top \\ DD^\top &= I_q + CC^\top \\ AC^\top &= BD^\top. \end{aligned}$$

Since $C^\top C$ is symmetric and since

$$x^\top C^\top C x = \|Cx\|^2 \geq 0,$$

we see that $C^\top C$ is a positive semi-definite matrix with nonnegative eigenvalues. We then deduce, (via the argument used in Proposition 5.2), that $A^\top A$ is symmetric positive definite and similarly for $D^\top D$. If we assume that the above decomposition of X holds, we deduce that

$$\begin{aligned} A &= U\alpha^{\frac{1}{2}} = U(I_p - Z^\top Z)^{-\frac{1}{2}} \\ B &= U\alpha^{\frac{1}{2}} Z^\top = U(I_p - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ C &= V\delta^{\frac{1}{2}} Z = V(I_q - ZZ^\top)^{-\frac{1}{2}} Z \\ D &= V\delta^{\frac{1}{2}} = V(I_q - ZZ^\top)^{-\frac{1}{2}}, \end{aligned}$$

which implies

$$Z = D^{-1}C \quad \text{and} \quad Z^\top = A^{-1}B.$$

We must check that

$$(D^{-1}C)^\top = A^{-1}B$$

i.e.,

$$C^\top (D^\top)^{-1} = A^{-1}B,$$

namely,

$$AC^\top = BD^\top,$$

which is indeed the last of our identities. Thus, we must have $Z = D^{-1}C = (A^{-1}B)^\top$. The above expressions for A and D also imply that

$$A^\top A = (I_p - Z^\top Z)^{-1} \quad \text{and} \quad D^\top D = (I_q - ZZ^\top)^{-1},$$

so we must check that the choice $Z = D^{-1}C = (A^{-1}B)^\top$ yields the above equations.

Since $Z^\top = A^{-1}B$, we have

$$\begin{aligned} Z^\top Z &= A^{-1}BB^\top (A^\top)^{-1} \\ &= A^{-1}(AA^\top - I_p)(A^\top)^{-1}, \qquad \text{since } AA^\top = I_p + BB^\top \\ &= I_p - A^{-1}(A^\top)^{-1} \\ &= I_p - (A^\top A)^{-1}. \end{aligned}$$

Therefore,

$$(A^\top A)^{-1} = I_p - Z^\top Z,$$

i.e.,

$$A^\top A = (I_p - Z^\top Z)^{-1},$$

as desired. We also have, this time, with $Z = D^{-1}C$,

$$\begin{aligned} ZZ^\top &= D^{-1}CC^\top (D^\top)^{-1} \\ &= D^{-1}(DD^\top - I_q)(D^\top)^{-1}, \qquad \text{since } DD^\top = I_p + CC^\top \\ &= I_q - D^{-1}(D^\top)^{-1} \\ &= I_q - (D^\top D)^{-1}. \end{aligned}$$

Therefore,

$$(D^\top D)^{-1} = I_q - ZZ^\top,$$

i.e.,

$$D^\top D = (I_q - ZZ^\top)^{-1},$$

as desired. Now since $A^\top A$ and $D^\top D$ are positive definite, the polar form implies that

$$A = U(A^\top A)^{\frac{1}{2}} = U(I_p - Z^\top Z)^{-\frac{1}{2}}$$

and

$$D = V(D^\top D)^{\frac{1}{2}} = V(I_q - ZZ^\top)^{-\frac{1}{2}},$$

for some unique matrices, $U \in \mathbf{O}(p)$ and $V \in \mathbf{O}(q)$. Since $Z = D^{-1}C$ and $Z^\top = A^{-1}B$, we get $C = DZ$ and $B = AZ^\top$, but this is

$$\begin{aligned} B &= U(I_p - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ C &= V(I_q - ZZ^\top)^{-\frac{1}{2}} Z, \end{aligned}$$

as required. Therefore, the unique choice of $Z = D^{-1}C = (A^{-1}B)^\top$, U and V does yield the formula of the proposition. □

We next show that the matrix

$$\begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} (I_p - Z^\top Z)^{-\frac{1}{2}} & (I_p - Z^\top Z)^{-\frac{1}{2}} Z^\top \\ (I_q - ZZ^\top)^{-\frac{1}{2}} Z & (I_q - ZZ^\top)^{-\frac{1}{2}} \end{pmatrix}$$

is symmetric. To prove this we use power series.

Proposition 6.2. *For any $q \times p$ matrix Z such that $I_p - Z^\top Z$ and $I_q - ZZ^\top$ are symmetric positive definite, the matrix*

$$S = \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix}$$

is symmetric, where $\alpha = (I_p - Z^\top Z)^{-1}$ and $\delta = (I_q - ZZ^\top)^{-1}$.

Proof. The matrix S is symmetric iff $Z\alpha^{\frac{1}{2}} = \delta^{\frac{1}{2}} Z$, that is iff $Z(I_p - Z^\top Z)^{-\frac{1}{2}} = (I_q - ZZ^\top)^{-\frac{1}{2}} Z$ iff

$$(I_q - ZZ^\top)^{\frac{1}{2}} Z = Z(I_p - Z^\top Z)^{\frac{1}{2}}.$$

If $Z = 0$, the equation holds trivially. If $Z \neq 0$, we know from linear algebra that $ZZ^\top$ and $Z^\top Z$ are symmetric positive semidefinite, and they have the same positive eigenvalues. Thus, $I_p - Z^\top Z$ is positive definite iff $I_q - ZZ^\top$ is positive definite, and if so, we must have $\rho(ZZ^\top) = \rho(Z^\top Z) < 1$ (where $\rho(ZZ^\top)$ denotes the largest modulus of the eigenvalues of $ZZ^\top$; in this case, since the eigenvalues of $ZZ^\top$ are nonnegative, this is the largest eigenvalue of $ZZ^\top$). If we use the spectral norm $\| \ \|$ (the operator norm induced by the 2-norm), we have

$$\left\| ZZ^\top \right\| = \sqrt{\rho((ZZ^\top)^\top ZZ^\top)} = \rho(ZZ^\top) < 1,$$

and similarly

$$\left\| Z^\top Z \right\| = \rho(Z^\top Z) < 1.$$

Therefore, the following series converge absolutely:

$$(I_p - Z^\top Z)^{\frac{1}{2}} = 1 + \frac{1}{2} Z^\top Z - \frac{1}{8}(Z^\top Z)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2} - 1\right)\cdots\left(\frac{1}{2} - k + 1\right)}{k!}(Z^\top Z)^k + \cdots$$

and

$$(I_q - ZZ^\top)^{\frac{1}{2}} = 1 + \frac{1}{2} ZZ^\top - \frac{1}{8}(ZZ^\top)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2} - 1\right)\cdots\left(\frac{1}{2} - k + 1\right)}{k!}(ZZ^\top)^k + \cdots .$$

We get

$$Z(I_p - Z^\top Z)^{\frac{1}{2}} = Z + \frac{1}{2} ZZ^\top Z - \frac{1}{8} Z(Z^\top Z)^2 + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2} - 1\right)\cdots\left(\frac{1}{2} - k + 1\right)}{k!} Z(Z^\top Z)^k + \cdots$$

and

$$(I_q - ZZ^\top)^{\frac{1}{2}} Z = Z + \frac{1}{2} ZZ^\top Z - \frac{1}{8}(ZZ^\top)^2 Z + \cdots + \frac{\frac{1}{2}\left(\frac{1}{2} - 1\right)\cdots\left(\frac{1}{2} - k + 1\right)}{k!}(ZZ^\top)^k Z + \cdots .$$

However

$$Z(Z^\top Z)^k = Z \underbrace{Z^\top Z \cdots Z^\top Z}_{k} = \underbrace{ZZ^\top \cdots ZZ^\top}_{k} Z = (ZZ^\top)^k Z,$$

which proves that $(I_q - ZZ^\top)^{\frac{1}{2}} Z = Z(I_p - Z^\top Z)^{\frac{1}{2}}$, as required. □

Another proof of Proposition 6.2 can be given using the SVD of Z. Indeed, we can write

$$Z = PDQ^\top$$

where P is a $q \times q$ orthogonal matrix, Q is a $p \times p$ orthogonal matrix, and D is a $q \times p$ matrix whose diagonal entries are (strictly) positive and all other entries zero. Then,

$$I_p - Z^\top Z = I_p - QD^\top P^\top PDQ^\top = Q(I_p - D^\top D)Q^\top,$$

a symmetric positive definite matrix by assumption. Furthermore, $(I_p - Z^\top Z)^{\frac{1}{2}} = Q(I_q - DD^\top)^{\frac{1}{2}} Q^\top$ since

$$Q(I_q - DD^\top)^{\frac{1}{2}} Q^\top Q(I_q - DD^\top)^{\frac{1}{2}} Q^\top = Q(I_p - D^\top D)Q^\top.$$

We also have

$$I_q - ZZ^\top = I_q - PDQ^\top QD^\top P^\top = P(I_q - DD^\top)P^\top,$$

another symmetric positive definite matrix by assumption, which has unique square root $(I_q - ZZ^\top)^{\frac{1}{2}} = P(I_q - DD^\top)^{\frac{1}{2}} P^\top$. Then,

$$Z(I_p - Z^\top Z)^{-\frac{1}{2}} = PDQ^\top Q(I_p - D^\top D)^{-\frac{1}{2}} Q^\top = PD(I_p - D^\top D)^{-\frac{1}{2}} Q^\top$$

and

$$(I_q - ZZ^\top)^{-\frac{1}{2}} Z = P(I_q - DD^\top)^{-\frac{1}{2}} P^\top PDQ^\top = P(I_q - DD^\top)^{-\frac{1}{2}} DQ^\top,$$

so it suffices to prove that

$$D(I_p - D^\top D)^{-\frac{1}{2}} = (I_q - DD^\top)^{-\frac{1}{2}} D.$$

However, D is essentially a diagonal matrix and the above is easily verified, as the reader should check.

Remark: The polar form of matrices in $\mathbf{O}(p, q)$ can be obtained *via* the exponential map and the Lie algebra, $\mathfrak{o}(p, q)$, of $\mathbf{O}(p, q)$, see Section 6.3. Indeed, every matrix $X \in \mathbf{O}(p, q)$ has a polar form of the form

$$X = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix},$$

with $P \in \mathbf{O}(p), Q \in \mathbf{O}(q)$, and with $\begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix}$ symmetric positive definite. This implies that

$$x^\top S_1 x = \begin{pmatrix} x^\top & 0 \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} > 0$$

for all $x \in \mathbb{R}^p$, $x \neq 0$, and that

$$y^\top S_3 y = \begin{pmatrix} 0 & y^\top \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix} \begin{pmatrix} 0 \\ y \end{pmatrix} > 0$$

for all $y \in \mathbb{R}^q$, $y \neq 0$. Therefore, S_1 and S_3 are symmetric positive definite. But then if we write

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

from

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} S_1 & S_2 \\ S_2^\top & S_3 \end{pmatrix},$$

we get $A = PS_1$ and $D = QS_3$, which are polar decompositions of A and D respectively. On the other hand, our factorization

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} \begin{pmatrix} \alpha^{\frac{1}{2}} & \alpha^{\frac{1}{2}} Z^\top \\ \delta^{\frac{1}{2}} Z & \delta^{\frac{1}{2}} \end{pmatrix}$$

yields $A = U\alpha^{\frac{1}{2}}$ and $D = V\delta^{\frac{1}{2}}$, with $U \in \mathbf{O}(p), V \in \mathbf{O}(q)$, and $\alpha^{\frac{1}{2}}, \delta^{\frac{1}{2}}$ symmetric positive definite. By uniqueness of the polar form, $P = U, Q = V$ ($S_1 = \alpha^{\frac{1}{2}}$ and $S_3 = \delta^{\frac{1}{2}}$), which shows that our factorization is the polar decomposition of X after all! This can also be proved more directly using the fact that $I - Z^\top Z$ (and $I - ZZ^\top$) being positive definite implies that the spectral norms $\|Z\|$ and $\|Z^\top\|$ of Z and $Z^\top$ are both strictly less than one.

We also have the following amusing property of the determinants of A and D:

Proposition 6.3. *For any matrix $X \in \mathbf{O}(p,q)$, if we write*

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

then

$$\det(X) = \det(A)\det(D)^{-1} \quad \textit{and} \quad |\det(A)| = |\det(D)| \geq 1.$$

Proof. Using the identities $A^\top B = C^\top D$ and $D^\top D = I_q + B^\top B$ proven in Proposition 6.1, observe that

$$\begin{pmatrix} A^\top & 0 \\ B^\top & -D^\top \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A^\top A & A^\top B \\ B^\top A - D^\top C & B^\top B - D^\top D \end{pmatrix} = \begin{pmatrix} A^\top A & A^\top B \\ 0 & -I_q \end{pmatrix}.$$

If we compute determinants, we get

$$\det(A)(-1)^q \det(D)\det(X) = \det(A)^2(-1)^q.$$

It follows that

$$\det(X) = \det(A)\det(D)^{-1}.$$

From $A^\top A = I_p + C^\top C$ and $D^\top D = I_q + B^\top B$, we conclude, via an eigenvalue argument, that $|\det(A)| \geq 1$ and $|\det(D)| \geq 1$. Since $|\det(X)| = 1$, we have $|\det(A)| = |\det(D)| \geq 1$. □

Remark: It is easy to see that the equations relating A, B, C, D established in the proof of Proposition 6.1 imply that

$$\det(A) = \pm 1 \quad \text{iff} \quad C = 0 \quad \text{iff} \quad B = 0 \quad \text{iff} \quad \det(D) = \pm 1.$$

We end this section by exhibiting a bijection between $\mathbf{O}(p,q)$ and $\mathbf{O}(p) \times \mathbf{O}(q) \times \mathbb{R}^{pq}$, and in essence justifying the statement that $\mathbf{SO_0(p,q)}$ is homeomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q) \times \mathbb{R}^{pq}$. The construction of the bijection begins with the following claim: for every $q \times p$ matrix Y, there is a unique $q \times p$ matrix Z such that $I_q - ZZ^\top$ is positive definite symmetric matrix and

$$(I_q - ZZ^\top)^{-\frac{1}{2}} Z = Y, \tag{$*$}$$

given by

$$Z = (I_q + YY^\top)^{-\frac{1}{2}} Y.$$

To verify the claim, we start with a given Y and define $Z = (I_q + YY^\top)^{-\frac{1}{2}} Y$, and show that Z satisfies $(*)$. Indeed, $I_q + YY^\top$ is symmetric positive definite, and we have

$$\begin{aligned} ZZ^\top &= (I_q + YY^\top)^{-\frac{1}{2}} YY^\top (I_q + YY^\top)^{-\frac{1}{2}} \\ &= (I_q + YY^\top)^{-\frac{1}{2}} (I_q + YY^\top - I_q)(I_q + YY^\top)^{-\frac{1}{2}} \\ &= I_q - (I_q + YY^\top)^{-1}, \end{aligned}$$

so

$$I_q - ZZ^\top = (I_q + YY^\top)^{-1},$$

from which we deduce that $I_q - ZZ^\top$ is positive definite (since it is the inverse of a positive definite matrix, and hence must have positive eigenvalues). Note that $I_q - ZZ^\top$ is also symmetric since it is the inverse of a symmetric matrix. It follows that

$$(I_q - ZZ^\top)^{-\frac{1}{2}} Z = (I_q + YY^\top)^{\frac{1}{2}} (I_q + YY^\top)^{-\frac{1}{2}} Y = Y,$$

which shows that $Z = (I_q + YY^\top)^{-\frac{1}{2}} Y$ is a solution of $(*)$.

We now verify the uniqueness of the solution. Assume that Z is a solution of $(*)$. Then we have

$$\begin{aligned} YY^\top &= (I_q - ZZ^\top)^{-\frac{1}{2}} ZZ^\top (I_q - ZZ^\top)^{-\frac{1}{2}} \\ &= (I_q - ZZ^\top)^{-\frac{1}{2}} (I_q - (I_q - ZZ^\top))(I_q - ZZ^\top)^{-\frac{1}{2}} \\ &= (I_q - ZZ^\top)^{-1} - I_q, \end{aligned}$$

so $(I_q - ZZ^\top)^{-1} = I_q + YY^\top$, which implies that

$$Z = (I_q - ZZ^\top)^{\frac{1}{2}} Y = (I_q + YY^\top)^{-\frac{1}{2}} Y.$$

Therefore, the map $Y \mapsto (I_q + YY^\top)^{-\frac{1}{2}} Y$ is a bijection between R^{qp} and the set of $q \times p$ matrices Z such that $I_q - ZZ^\top$ is symmetric positive definite, whose inverse is the map

$$Z \mapsto (I_q - ZZ^\top)^{-\frac{1}{2}} Z = \delta^{\frac{1}{2}} Z.$$

As a corollary, there is a bijection between $\mathbf{O}(p,q)$ and $\mathbf{O}(p) \times \mathbf{O}(q) \times \mathbb{R}^{pq}$.

6.2 Pseudo-Algebraic Groups

The topological structure of certain linear Lie groups determined by equations among the real and the imaginary parts of their entries can be determined by refining the polar form of matrices. Such groups are called pseudo-algebraic groups. For example, the groups $\mathbf{SO}(p,q)$ and $\mathbf{SU}(p,q)$ are pseudo-algebraic, where $\mathbf{U}(p,q)$ is the set of all $n \times n$-matrices

$$\mathbf{U}(p,q) = \{A \in \mathbf{GL}(n,\mathbb{C}) \mid A^* I_{p,q} A = I_{p,q}\},$$

and $\mathbf{SU}(p,q)$ is the subgroup

$$\mathbf{SU}(p,q) = \{A \in \mathbf{U}(p,q) \mid \det(A) = 1\}.$$

Consider the group $\mathbf{GL}(n,\mathbb{C})$ of invertible $n \times n$ matrices with complex coefficients. If $A = (a_{kl})$ is such a matrix, denote by x_{kl} the real part (resp. y_{kl}, the imaginary part) of a_{kl} (so, $a_{kl} = x_{kl} + iy_{kl}$).

Definition 6.1. A subgroup G of $\mathbf{GL}(n,\mathbb{C})$ is *pseudo-algebraic* iff there is a finite set of polynomials in $2n^2$ variables with real coefficients $\{P_j(X_1,\ldots,X_{n^2},Y_1,\ldots,Y_{n^2})\}_{j=1}^t$, so that

$$A = (x_{kl} + iy_{kl}) \in G \quad \text{iff} \quad P_j(x_{11},\ldots,x_{nn},y_{11},\ldots,y_{nn}) = 0, \quad \text{for } j = 1,\ldots,t.$$

Since a pseudo-algebraic subgroup is the zero locus of a set of polynomials, it is a closed subgroup, and thus a Lie group.

Recall that if A is a complex $n \times n$-matrix, its *adjoint* A^* is defined by $A^* = (\overline{A})^\top$. Also, $\mathbf{U}(n)$ denotes the group of unitary matrices, i.e., those matrices $A \in \mathbf{GL}(n,\mathbb{C})$ so that $AA^* = A^*A = I$, and $\mathbf{H}(n)$ denotes the vector space of Hermitian matrices i.e., those matrices A so that $A^* = A$.

The following proposition is needed.

Proposition 6.4. *Let $P(x_1,\ldots,x_n)$ be a polynomial with real coefficients. For any $(a_1,\ldots,$ $a_n) \in \mathbb{R}^n$, assume that $P(e^{ka_1},\ldots,e^{ka_n}) = 0$ for all $k \in \mathbb{N}$. Then,*

$$P(e^{ta_1},\ldots,e^{ta_n}) = 0 \quad \textit{for all } t \in \mathbb{R}.$$

Proof. Any monomial $\alpha x_1^{i_1} \cdots x_n^{i_n}$ in P when evaluated at $(e^{ta_1},\ldots,e^{ta_n})$ becomes $\alpha e^{t\sum a_j i_j}$. Collecting terms with the same exponential part, we may assume that we have an expression of the form

$$P(e^{ta_1},\ldots,e^{ta_n}) = \sum_{k=1}^{N} \alpha_k e^{tb_k} = \alpha_N e^{tb_N} + \sum_{k=1}^{N-1} \alpha_k e^{tb_k}$$

which vanishes for all $t \in \mathbb{N}$. We may also assume that $\alpha_k \neq 0$ for all k and that the b_k are sorted so that $b_1 < b_2 < \cdots < b_N$. Assume by contradiction that $N > 0$. If we multiply the above expression by e^{-tb_N}, by relabeling the coefficients b_k in the exponentials, we may assume that $b_1 < b_2 < \cdots < b_{N-1} < 0 = b_N$. Now, if we let t go to $+\infty$, the terms $\alpha_k e^{tb_k}$ go to 0 for $k = 1,\ldots,N-1$, and we get $\alpha_N = 0$, a contradiction. $\square$

We now have the following theorem which is essentially a refined version of the polar decomposition of matrices:

Theorem 6.5. *Let G be a pseudo-algebraic subgroup of $\mathbf{GL}(n, \mathbb{C})$ stable under adjunction (i.e., we have $A^* \in G$ whenever $A \in G$). There is some integer $d \in \mathbb{N}$ so that G is homeomorphic to $(G \cap \mathbf{U}(n)) \times \mathbb{R}^d$. Moreover, if $\mathfrak{g}$ is the Lie algebra of G, the map*

$$(\mathbf{U}(n) \cap G) \times (\mathbf{H}(n) \cap \mathfrak{g}) \longrightarrow G \quad \textit{given by} \quad (U, H) \mapsto Ue^H,$$

is a homeomorphism onto G.

Proof. We follow the proof in Mneimné and Testard [86] (Chapter 3); a similar proof is given in Knapp [68] (Chapter 1). First we observe that for every invertible matrix P, the group G is pseudo-algebraic iff PGP^{-1} is pseudo-algebraic, since the map $X \mapsto PXP^{-1}$ is linear.

By the polar decomposition, every matrix $A \in G$ can be written uniquely as $A = US$, where $U \in \mathbf{U}(n)$ and $S \in \mathbf{HPD}(n)$. Furthermore, by Proposition 1.10, the matrix S can be written (uniquely) as $S = e^H$, for some unique Hermitian matrix $H \in \mathbf{H}(n)$, so we have $A = Ue^H$. We need to prove that $H \in \mathfrak{g}$ and that $U \in G$. Since G is closed under adjunction, $A^* \in G$, that is $e^H U^* \in G$, so $e^H U^* U e^H = e^{2H} \in G$. If we can prove that $e^{tH} \in G$ for all $t \in \mathbb{R}$, then $H \in \mathfrak{g}$ and $e^H \in G$, so $U \in e^{-H}A \in G$.

Since $2H$ is Hermitian, it has real eigenvalues $\lambda_1, \ldots, \lambda_n$ and it can be diagonalized as $2H = V\Lambda V^{-1}$, where V is unitary and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. By a previous observation, the group VGV^{-1} is also pseudo-algebraic, so we may assume that $2H$ is a diagonal matrix with real entries, and to say that $e^{2H} \in G$ means that $e^{\lambda_1}, \ldots, e^{\lambda_n}$ satisfy a set of algebraic equations. Since G is a group, for every $k \in \mathbb{Z}$, we have $e^{k2H} \in G$, so $e^{k\lambda_1}, \ldots, e^{k\lambda_n}$ satisfy the same set of algebraic equations. By Proposition 6.4, $e^{t\lambda_1}, \ldots, e^{t\lambda_n}$ satisfy the same set of algebraic equations for all $t \in \mathbb{R}$, which means that $e^{tH} \in G$ for all $t \in \mathbb{R}$. It follows that $H \in \mathfrak{g}$, $e^H \in G$, and thus $U \in e^{-H}A \in G$.

For invertible matrices, the polar decomposition is unique, so we found a unique $U \in \mathbf{U}(n) \cap G$ and a unique matrix $H \in \mathbf{H}(n) \cap \mathfrak{g}$ so that

$$A = Ue^H.$$

The fact that the map $(U, H) \mapsto Ue^H$ is a homeomorphism takes a little bit of work. This follows from the fact that polar decomposition and the bijection between $\mathbf{H}(n)$ and $\mathbf{HPD}(n)$ are homeomorphisms (see Section 1.5); these facts are proved in Mneimné and Testard [86]; see Theorem 1.6.3 for the first homeomorphism and Theorem 3.3.4 for the second homeomorphism. Since $\mathbf{H}(n) \cap \mathfrak{g}$ is a real vector space, it is isomorphic to $\mathbb{R}^d$ for some $d \in \mathbb{N}$, and so G is homeomorphic to $(G \cap \mathbf{U}(n)) \times \mathbb{R}^d$. □

Observe that if G is also compact then $d = 0$, and $G \subseteq \mathbf{U}(n)$.

Remark: A subgroup G of $\mathbf{GL}(n,\mathbb{R})$ is called *algebraic* if there is a finite set of polynomials in n^2 variables with real coefficients $\{P_j(X_1,\ldots,X_{n^2})\}_{j=1}^t$, so that

$$A=(x_{kl})\in G \quad \text{iff} \quad P_j(x_{11},\ldots,x_{nn})=0, \quad \text{for } j=1,\ldots,t.$$

Then it can be shown that every compact subgroup of $\mathbf{GL}(n,\mathbb{R})$ is algebraic. The proof is quite involved and uses the existence of the Haar measure on a compact Lie group; see Mneimné and Testard [86] (Theorem 3.7).

6.3 More on the Topology of $\mathbf{O}(p,q)$ and $\mathbf{SO}(p,q)$

It turns out that the topology of the group $\mathbf{O}(p,q)$ is completely determined by the topology of $\mathbf{O}(p)$ and $\mathbf{O}(q)$. This result can be obtained as a simple consequence of some standard Lie group theory. The key notion is that of a pseudo-algebraic group defined in Section 6.2.

We can apply Theorem 6.5 to determine the structure of the space $\mathbf{O}(p,q)$. We know that $\mathbf{O}(p,q)$ consists of the matrices A in $\mathbf{GL}(p+q,\mathbb{R})$ such that

$$A^\top I_{p,q}A=I_{p,q},$$

and so $\mathbf{O}(p,q)$ is clearly pseudo-algebraic. Using the above equation, and the curve technique demonstrated at the beginning of Section 5.2, it is easy to determine the Lie algebra $\mathfrak{o}(p,q)$ of $\mathbf{O}(p,q)$. We find that $\mathfrak{o}(p,q)$ is given by

$$\mathfrak{o}(p,q)=\left\{\begin{pmatrix} X_1 & X_2\\ X_2^\top & X_3\end{pmatrix}\;\middle|\; X_1^\top=-X_1,\; X_3^\top=-X_3,\; X_2 \text{ arbitrary}\right\}$$

where X_1 is a $p\times p$ matrix, X_3 is a $q\times q$ matrix, and X_2 is a $p\times q$ matrix.

Consequently, it immediately follows that

$$\mathfrak{o}(p,q)\cap\mathbf{H}(p+q)=\left\{\begin{pmatrix} 0 & X_2\\ X_2^\top & 0\end{pmatrix}\;\middle|\; X_2 \text{ arbitrary}\right\},$$

a vector space of dimension pq.

Some simple calculations also show that

$$\mathbf{O}(p,q)\cap\mathbf{U}(p+q)=\left\{\begin{pmatrix} X_1 & 0\\ 0 & X_2\end{pmatrix}\;\middle|\; X_1\in\mathbf{O}(p),\; X_2\in\mathbf{O}(q)\right\}\cong\mathbf{O}(p)\times\mathbf{O}(q).$$

Therefore, we obtain the structure of $\mathbf{O}(p,q)$:

Proposition 6.6. *The topological space $\mathbf{O}(p,q)$ is homeomorphic to $\mathbf{O}(p)\times\mathbf{O}(q)\times\mathbb{R}^{pq}$.*

Since $\mathbf{O}(p)$ has two connected components when $p \geq 1$, we deduce (via the decomposition of Proposition 6.1) that $\mathbf{O}(p,q)$ has four connected components when $p, q \geq 1$. It is also obvious that

$$\mathbf{SO}(p,q) \cap \mathbf{U}(p+q) = \left\{ \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \;\middle|\; X_1 \in \mathbf{O}(p),\;\; X_2 \in \mathbf{O}(q),\;\; \det(X_1)\det(X_2) = 1 \right\}.$$

This is a subgroup of $\mathbf{O}(p) \times \mathbf{O}(q)$ that we denote $S(\mathbf{O}(p) \times \mathbf{O}(q))$. Furthermore, it can be shown that $\mathfrak{so}(p,q) = \mathfrak{o}(p,q)$. Thus, we also have

Proposition 6.7. *The topological space $\mathbf{SO}(p,q)$ is homeomorphic to $S(\mathbf{O}(p) \times \mathbf{O}(q)) \times \mathbb{R}^{pq}$.*

Observe that the dimension of all these spaces depends only on $p+q$. It is $p(p-1)/2 + q(q-1)/2 + pq = (p+q)(p+q-1)/2$, where we used the fact that $\mathbf{O}(n)$ is a smooth manifold of dimension $n(n-1)/2$. Also, $\mathbf{SO}(p,q)$ has two connected components when $p, q \geq 1$. The connected component of I_{p+q} is the group $\mathbf{SO}_0(p,q)$. This latter space is homeomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q) \times \mathbb{R}^{pq}$. If we write

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix},$$

then it is shown in O'Neill [91] (Chapter 9, Lemma 6) that the connected component $\mathbf{SO}_0(p,q)$ of $\mathbf{SO}(p,q)$ containing I is given by

$$\mathbf{SO}_0(p,q) = \{A \in \mathbf{GL}(n,\mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q},\; \det(P) > 0,\; \det(S) > 0\}.$$

For both $\mathbf{SO}(p,q)$ and $\mathbf{SO}_0(p,q)$, the inverse is given by

$$A^{-1} = I_{p,q} A^\top I_{p,q}.$$

We can show that $\mathbf{SO}(p,q)$ and $\mathbf{SO}(q,p)$ are isomorphic (similarly, $\mathbf{O}(p,q)$ and $\mathbf{O}(q,p)$ are isomorphic, and $\mathbf{SO}_0(p,q)$ and $\mathbf{SO}_0(q,p)$ are isomorphic) as follows. Let $J_{p,q}$ be the permutation matrix

$$J_{p,q} = \begin{pmatrix} 0 & I_q \\ I_p & 0 \end{pmatrix}.$$

Observe that $J_{p,q} J_{q,p} = I_{p+q}$ and that $J_{p,q}^\top = J_{q,p}$.

Proposition 6.8. *If ψ is the map given by $\psi(A) = J_{p,q} A J_{q,p}$, then $\psi\colon \mathbf{O}(p,q) \to \mathbf{O}(q,p)$, $\psi\colon \mathbf{SO}(p,q) \to \mathbf{SO}(q,p)$, and $\psi\colon \mathbf{SO}_0(p,q) \to \mathbf{SO}_0(q,p)$, are isomorphisms.*

Proof sketch. Since $J_{p,q} J_{q,p} = I_{p+q}$, we have

$$\psi(A)\psi(B) = J_{p,q} A J_{q,p} J_{p,q} B J_{q,p} = J_{p,q} A B J_{q,p}.$$

Observe that

$$J_{q,p} I_{q,p} J_{q,p}^\top = J_{q,p} I_{q,p} J_{p,q} = -I_{p,q}.$$

Using the above equation, if $A \in \mathbf{O}(p,q)$, that is, $A^\top I_{p,q} A = I_{p,q}$, then we have

$$\begin{aligned}
(\psi(A))^\top I_{q,p} \psi(A) &= (J_{p,q} A J_{q,p})^\top I_{q,p} J_{p,q} A J_{q,p} \\
&= J_{p,q} A^\top J_{q,p} I_{q,p} J_{p,q} A J_{q,p} \\
&= -J_{p,q} A^\top I_{p,q} A J_{q,p} \\
&= -J_{p,q} I_{p,q} J_{q,p} \\
&= - - I_{q,p} = I_{q,p}.
\end{aligned}$$

Therefore $\psi(A) \in \mathbf{O}(q,p)$, and $\psi\colon \mathbf{O}(p,q) \to \mathbf{O}(q,p)$ is a homomorphism.

Since $J_{p,q}^\top = J_{q,p}$, and since $\det(J_{p,q}) = \pm 1$ because $J_{p,q}$ is a permutation matrix, we have $J_{p,q} A J_{q,p} = J_{p,q} A J_{p,q}^\top$, so

$$\det(\psi(A)) = \det(J_{p,q} A J_{p,q}^\top) = \det(J_{p,q}) \det(A) \det(J_{p,q}^\top) = \det(A).$$

Therefore, if $A \in \mathbf{SO}(p,q)$, then $\det(A) = 1$, so $\det(\psi(A)) = \det(A) = 1$, so $\psi(A) \in \mathbf{SO}(q,p)$, and $\psi\colon \mathbf{SO}(p,q) \to \mathbf{SO}(q,p)$ is a homomorphism.

If we write

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix},$$

then we have

$$\psi(A) = J_{p,q} A J_{q,p} = \begin{pmatrix} 0 & I_q \\ I_p & 0 \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} \begin{pmatrix} 0 & I_p \\ I_q & 0 \end{pmatrix} = \begin{pmatrix} S & R \\ Q & P \end{pmatrix}.$$

If $A \in \mathbf{SO}_0(p,q)$ then $\det(P) > 0$ and $\det(S) > 0$, so $\psi(A) \in \mathbf{SO}_0(q,p)$, and $\psi\colon \mathbf{SO}_0(p,q) \to \mathbf{SO}_0(q,p)$ is a homomorphism. It is easy to verify that the inverse of ψ is given by $\psi^{-1}(B) = J_{q,p} B J_{p,q}$, so the above maps are indeed isomorphisms. □

Theorem 6.5 gives the polar form of a matrix $A \in \mathbf{O}(p,q)$. We have

$$A = Ue^S, \quad \text{with} \quad U \in \mathbf{O}(p) \times \mathbf{O}(q) \quad \text{and} \quad S \in \mathfrak{so}(p,q) \cap \mathbf{S}(p+q),$$

where U is of the form

$$U = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix}, \quad \text{with} \quad P \in \mathbf{O}(p) \quad \text{and} \quad Q \in \mathbf{O}(q),$$

and $\mathfrak{so}(p,q) \cap \mathbf{S}(p+q)$ consists of all $(p+q) \times (p+q)$ symmetric matrices of the form

$$S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},$$

with X an arbitrary $p \times q$ matrix.

It turns out that it is not very hard to compute explicitly the exponential e^S of such matrices (see Mneimné and Testard [86]). Recall that the functions cosh and sinh also make sense for matrices (since the exponential makes sense) and are given by

$$\cosh(A) = \frac{e^A + e^{-A}}{2} = I + \frac{A^2}{2!} + \cdots + \frac{A^{2k}}{(2k)!} + \cdots$$

and

$$\sinh(A) = \frac{e^A - e^{-A}}{2} = A + \frac{A^3}{3!} + \cdots + \frac{A^{2k+1}}{(2k+1)!} + \cdots .$$

We also set

$$\frac{\sinh(A)}{A} = I + \frac{A^2}{3!} + \cdots + \frac{A^{2k}}{(2k+1)!} + \cdots ,$$

which is defined for all matrices A (even when A is singular). Then we have

Proposition 6.9. *For any matrix S of the form*

$$S = \begin{pmatrix} 0 & X \\ X^\top & 0 \end{pmatrix},$$

we have

$$e^S = \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh((X^\top X)^{\frac{1}{2}})X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix}.$$

Proof. By induction, it is easy to see that

$$S^{2k} = \begin{pmatrix} (XX^\top)^k & 0 \\ 0 & (X^\top X)^k \end{pmatrix}$$

and

$$S^{2k+1} = \begin{pmatrix} 0 & (XX^\top)^k X \\ (X^\top X)^k X^\top & 0 \end{pmatrix}.$$

The rest is left as an exercise. □

Remark: Although at first glance, e^S does not look symmetric, it is!

As a consequence of Proposition 6.9, every matrix $A \in \mathbf{O}(p,q)$ has the polar form

$$A = \begin{pmatrix} P & 0 \\ 0 & Q \end{pmatrix} \begin{pmatrix} \cosh((XX^\top)^{\frac{1}{2}}) & \frac{\sinh((XX^\top)^{\frac{1}{2}})X}{(XX^\top)^{\frac{1}{2}}} \\ \frac{\sinh((X^\top X)^{\frac{1}{2}})X^\top}{(X^\top X)^{\frac{1}{2}}} & \cosh((X^\top X)^{\frac{1}{2}}) \end{pmatrix},$$

with $P \in \mathbf{O}(p)$, $Q \in \mathbf{O}(q)$, and X an arbitrary $p \times q$ matrix.

6.4 Problems

Problem 6.1. Finish the proof of Proposition 6.8.

Problem 6.2. Provide the details of the proof of Proposition 6.9.

Chapter 7

Manifolds, Tangent Spaces, Cotangent Spaces, Submanifolds

In Chapter 3 we defined the notion of a manifold embedded in some ambient space $\mathbb{R}^N$. In order to maximize the range of applications of the theory of manifolds, it is necessary to generalize the concept of a manifold to spaces that are not a priori embedded in some $\mathbb{R}^N$. The basic idea is still that, whatever a manifold is, it is a topological space that can be covered by a collection of open subsets U_α, where each U_α is isomorphic to some "standard model," e.g., some open subset of Euclidean space $\mathbb{R}^n$. Of course, manifolds would be very dull without functions defined on them and between them. This is a general fact learned from experience: *Geometry arises not just from spaces but from spaces and interesting classes of functions between them*. In particular, we still would like to "do calculus" on our manifold and have good notions of curves, tangent vectors, differential forms, etc.

The small drawback with the more general approach is that the definition of a tangent vector is more abstract. We can still define the notion of a curve on a manifold, but such a curve does not live in any given $\mathbb{R}^n$, so it it not possible to define tangent vectors in a simple-minded way using derivatives. Instead, we have to resort to the notion of chart. This is not such a strange idea. For example, a geography atlas gives a set of maps of various portions of the earth and this provides a very good description of what the earth is, without actually imagining the earth embedded in 3-space.

The material of this chapter borrows from many sources, including Warner [114], Berger and Gostiaux [15], O'Neill [91], Do Carmo [39, 38], Gallot, Hulin and Lafontaine [49], Lang [75], Schwartz [104], Hirsch [61], Sharpe [107], Guillemin and Pollack [55], Lafontaine [72], Dubrovin, Fomenko and Novikov [42] and Boothby [16]. A nice (not very technical) exposition is given in Morita [87] (Chapter 1). The recent book by Tu [112] is also highly recommended for its clarity. Among the many texts on manifolds and differential geometry, the book by Choquet-Bruhat, DeWitt-Morette and Dillard-Bleick [32] stands apart because it is one of the clearest and most comprehensive. (Many proofs are omitted, but this can be an advantage!) Being written for (theoretical) physicists, it contains more examples and applications than most other sources.

7.1 Charts and Manifolds

Given $\mathbb{R}^n$, recall that the projection functions $pr_i \colon \mathbb{R}^n \to \mathbb{R}$ are defined by

$$pr_i(x_1, \ldots, x_n) = x_i, \quad 1 \leq i \leq n.$$

For technical reasons (in particular, to ensure the existence of partitions of unity, a crucial tool in manifold theory; see Section 10.1) and to avoid "esoteric" manifolds that do not arise in practice, from now on, **all topological spaces under consideration will be assumed to be Hausdorff and second-countable** (which means that the topology has a countable basis).

The first step in generalizing the notion of a manifold is to define charts, a way to say that locally a manifold "looks like" an open subset of $\mathbb{R}^n$.

Definition 7.1. Given a topological space M, a *chart* (or *local coordinate map*) is a pair (U, φ), where U is an open subset of M and $\varphi \colon U \to \Omega$ is a homeomorphism onto an open subset $\Omega = \varphi(U)$ of $\mathbb{R}^{n_\varphi}$ (for some $n_\varphi \geq 1$). For any $p \in M$, a chart (U, φ) is a *chart at* p iff $p \in U$. If (U, φ) is a chart, then the functions $x_i = pr_i \circ \varphi$ are called *local coordinates* and for every $p \in U$, the tuple $(x_1(p), \ldots, x_n(p))$ is the set of *coordinates of* p w.r.t. the chart. The inverse (Ω, φ^{-1}) of a chart is called a *local parametrization*. Given any two charts (U_i, φ_i) and (U_j, φ_j), if $U_i \cap U_j \neq \emptyset$, we have the *transition maps* $\varphi_i^j \colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ and $\varphi_j^i \colon \varphi_j(U_i \cap U_j) \to \varphi_i(U_i \cap U_j)$, defined by

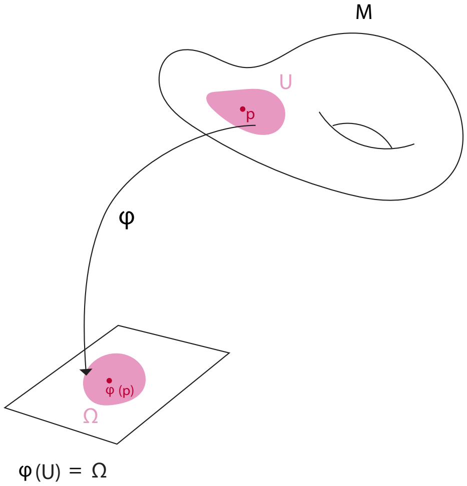


Figure 7.1: A chart (U, φ) on M.

$$\varphi_i^j = \varphi_j \circ \varphi_i^{-1} \quad \text{and} \quad \varphi_j^i = \varphi_i \circ \varphi_j^{-1}.$$

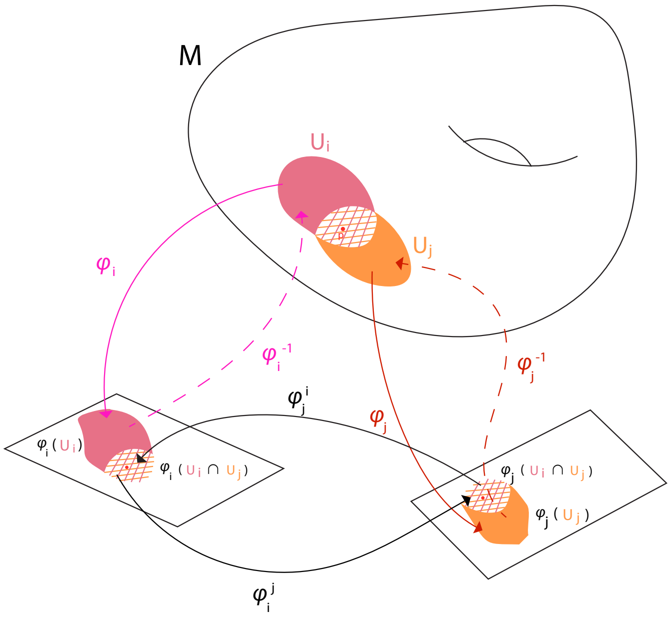


Figure 7.2: The transition maps φ_i^j and φ_j^i.

Clearly, $\varphi_j^i = (\varphi_i^j)^{-1}$. Observe that the transition maps φ_i^j (resp. φ_j^i) are maps between *open subsets of* $\mathbb{R}^n$. This is good news! Indeed, the whole arsenal of calculus is available for functions on $\mathbb{R}^n$, and we will be able to promote many of these results to manifolds by imposing suitable conditions on transition functions.

As in Section 3.1, whatever our generalized notion of a manifold is, we would like to define the notion of tangent space at a point of manifold, the notion of smooth function between manifolds, and the notion of derivative of a function (at a point) between manifolds. Unfortunately, even though our parametrizations $\varphi^{-1}\colon \Omega \to U$ are homeomorphisms, since U is a subset of a space M *which is not assumed to be contained in* $\mathbb{R}^N$ *(for any* N*)*, the derivative $d\varphi_{t_0}^{-1}$ does not make sense, unlike in the situation of Definition 3.1. Therefore, some extra conditions on the charts must be imposed in order to recapture the fact that for manifolds embedded in $\mathbb{R}^N$, the parametrizations are immersions. An invaluable hint is provided by Lemma 3.2: we require the transition maps $\varphi_i^j\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j)$ to be sufficiently differentiable. This makes perfect sense since the φ_i^j are functions between open subsets of $\mathbb{R}^n$. It also turns out that these conditions on transition maps guarantee that notions, such as tangent vectors, whose definition seems to depend on the choice of a chart, are in fact independent of the choice of charts. The above motivations suggest the following requirements on charts.

Definition 7.2. Given a topological space M, given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a C^k *n-atlas* (or *n-atlas of class C^k*) $\mathcal{A}$ is a family of charts $\{(U_i, \varphi_i)\}$, such that

(1) $\varphi_i(U_i) \subseteq \mathbb{R}^n$ for all i;

(2) The U_i cover M, i.e.,
$$M = \bigcup_i U_i;$$

(3) Whenever $U_i \cap U_j \neq \emptyset$, the transition map φ_i^j (and φ_j^i) is a C^k-diffeomorphism. When $k = \infty$, the φ_i^j are smooth diffeomorphisms.

We must ensure that we have enough charts in order to carry out our program of generalizing calculus on $\mathbb{R}^n$ to manifolds. For this, we must be able to add new charts whenever necessary, provided that they are consistent with the previous charts in an existing atlas.

Definition 7.3. Given a C^k n-atlas $\mathcal{A}$ on a topological space M, for any other chart (U, φ), we say that (U, φ) is *compatible* with the atlas $\mathcal{A}$ iff every map $\varphi_i \circ \varphi^{-1}$ and $\varphi \circ \varphi_i^{-1}$ is C^k (whenever $U \cap U_i \neq \emptyset$). Two atlases $\mathcal{A}$ and $\mathcal{A}'$ on M are *compatible* iff every chart of one is compatible with the other atlas. This is equivalent to saying that the union of the two atlases is still an atlas.

It is immediately verified that compatibility induces an equivalence relation on C^k n-atlases on M. In fact, given an atlas $\mathcal{A}$ for M, it is easy to see that the collection $\widetilde{\mathcal{A}}$ of all charts compatible with $\mathcal{A}$ is a maximal atlas in the equivalence class of atlases compatible with $\mathcal{A}$. Finally we have our generalized notion of a manifold.

Definition 7.4. Given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a *C^k-manifold of dimension n* consists of a topological space M together with an equivalence class $\overline{\mathcal{A}}$ of C^k n-atlases on M. Any atlas $\mathcal{A}$ in the equivalence class $\overline{\mathcal{A}}$ is called a *differentiable structure of class C^k (and dimension n) on M*. We say that M is *modeled on* $\mathbb{R}^n$. When $k = \infty$, we say that M is a *smooth manifold*.

Remark: It might have been better to use the terminology *abstract manifold* rather than manifold to emphasize the fact that the space M is not a priori a subspace of $\mathbb{R}^N$, for some suitable N.

We can allow $k = 0$ in the above definitions. In this case, Condition (3) in Definition 7.2 is void, since a C^0-diffeomorphism is just a homeomorphism, but φ_i^j is always a homeomorphism.

Definition 7.5. If $k = 0$ in Definition 7.4, then M is called a *topological manifold of dimension n*.

We do not require a manifold to be connected but we require *all* the components to have the same dimension n.

On every connected component of M, it can be shown that the dimension n_φ of the range of every chart is the same. This is quite easy to show if $k \geq 1$ (use Proposition 11.20) but for $k = 0$ this requires a deep theorem of Brouwer. (Brouwer's *Invariance of Domain Theorem* states that if $U \subseteq \mathbb{R}^n$ is an open set and if $f\colon U \to \mathbb{R}^n$ is a continuous and injective map, then $f(U)$ is open in $\mathbb{R}^n$. Using Brouwer's theorem, we can show the following fact: If $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^n$ are two open subsets and if $f\colon U \to V$ is a homeomorphism between U and V, then $m = n$. If $m > n$, then consider the injection, $i\colon \mathbb{R}^n \to \mathbb{R}^m$, where $i(x) = (x, 0_{m-n})$. Clearly, i is injective and continuous, so $i \circ f\colon U \to i(V)$ is injective and continuous and Brouwer's Theorem implies that $i(V)$ is open in $\mathbb{R}^m$, which is a contradiction, as $i(V) = V \times \{0_{m-n}\}$ is not open in $\mathbb{R}^m$. If $m < n$, consider the homeomorphism $f^{-1}\colon V \to U$.)

What happens if $n = 0$? In this case, every one-point subset of M is open, so every subset of M is open; that is, M is any (countable if we assume M to be second-countable) set with the discrete topology!

Observe that since $\mathbb{R}^n$ is locally compact and locally connected, so is every manifold (check this!).

In order to get a better grasp of the notion of manifold it is useful to consider examples of non-manifolds. First, consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^2 - x^3,$$

namely, the set of points

$$M_1 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^2 - x^3\}.$$

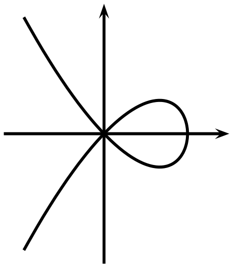


Figure 7.3: A nodal cubic; not a manifold.

This curve, shown in Figure 7.3, is called a *nodal cubic* and is also defined as the parametric curve

$$\begin{aligned} x &= 1 - t^2 \\ y &= t(1 - t^2). \end{aligned}$$

We claim that M_1 is not even a topological manifold. The problem is that the nodal cubic has a self-intersection at the origin. If M_1 was a topological manifold, then there would be a connected open subset $U \subseteq M_1$ containing the origin $O = (0, 0)$, namely the intersection of a small enough open disc centered at O with M_1, and a local chart $\varphi\colon U \to \Omega$, where Ω is some connected open subset of $\mathbb{R}$ (that is, an open interval), since φ is a homeomorphism. However, $U - \{O\}$ consists of four disconnected components, and $\Omega - \varphi(O)$ of two disconnected components, contradicting the fact that φ is a homeomorphism.

Let us now consider the curve in $\mathbb{R}^2$ given by the zero locus of the equation

$$y^2 = x^3,$$

namely, the set of points

$$M_2 = \{(x, y) \in \mathbb{R}^2 \mid y^2 = x^3\}.$$

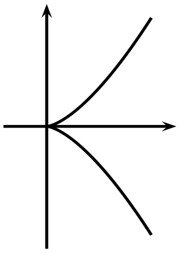


Figure 7.4: A cuspidal cubic.

This curve showed in Figure 7.4 and called a *cuspidal cubic* is also defined as the parametric curve

$$\begin{aligned} x &= t^2 \\ y &= t^3. \end{aligned}$$

Consider the map, $\varphi\colon M_2 \to \mathbb{R}$, given by

$$\varphi(x, y) = y^{1/3}.$$

Since $x = y^{2/3}$ on M_2, we see that φ^{-1} is given by

$$\varphi^{-1}(t) = (t^2, t^3)$$

and clearly φ is a homeomorphism, so M_2 is a topological manifold. However, with the atlas consisting of the single chart $\{\varphi\colon M_2 \to \mathbb{R}\}$, the space M_2 is also a smooth manifold! Indeed, as there is a single chart, Condition (3) of Definition 7.2 holds vacuously.

This fact is somewhat unexpected because the cuspidal cubic is not smooth at the origin, since the tangent vector of the parametric curve $c\colon t \mapsto (t^2, t^3)$ at the origin is the zero vector (the velocity vector at t is $c'(t) = (2t, 3t^2)$). However, this apparent paradox has to do with the fact that, as a parametric curve, M_2 is not immersed in $\mathbb{R}^2$ since c' is not injective (see Definition 7.27 (a)), whereas as an abstract manifold, with this single chart, M_2 is diffeomorphic to $\mathbb{R}$.

We also have the chart $\psi\colon M_2 \to \mathbb{R}$, given by

$$\psi(x, y) = y,$$

with ψ^{-1} given by

$$\psi^{-1}(u) = (u^{2/3}, u).$$

With the atlas consisting of the single chart $\{\psi\colon M_2 \to \mathbb{R}\}$, the space M_2 is also a smooth manifold. Observe that

$$\varphi \circ \psi^{-1}(u) = u^{1/3},$$

a map that is *not* differentiable at $u = 0$. Therefore, the atlas $\{\varphi\colon M_2 \to \mathbb{R}, \psi\colon M_2 \to \mathbb{R}\}$ is not C^1, and thus with respect to that atlas, M_2 is *not* a C^1-manifold. This example also shows that the atlases $\{\varphi\colon M_2 \to \mathbb{R}\}$ and $\{\psi\colon M_2 \to \mathbb{R}\}$ are inequivalent.

The example of the cuspidal cubic reveals one of the subtleties of the definition of a C^k (or C^∞) manifold: whether a topological space is a C^k-manifold or a smooth manifold depends on the choice of atlas. As a consequence, if a space M happens to be a topological manifold because it has an atlas consisting of a single chart, or more generally if it has an atlas whose transition functions "avoid" singularities, then it is automatically a smooth manifold. In particular, if $f\colon U \to \mathbb{R}^m$ is any *continuous* function from some open subset U of $\mathbb{R}^n$ to $\mathbb{R}^m$, then the graph $\Gamma(f) \subseteq \mathbb{R}^{n+m}$ of f given by

$$\Gamma(f) = \{(x, f(x)) \in \mathbb{R}^{n+m} \mid x \in U\}$$

is a smooth manifold of dimension n with respect to the atlas consisting of the single chart $\varphi\colon \Gamma(f) \to U$, given by

$$\varphi(x, f(x)) = x,$$

with its inverse $\varphi^{-1}\colon U \to \Gamma(f)$ given by

$$\varphi^{-1}(x) = (x, f(x)).$$

The notion of a submanifold using the concept of "adapted chart" (see Definition 7.26 in Section 7.6) gives a more satisfactory treatment of C^k (or smooth) submanifolds of $\mathbb{R}^n$.

It should also be noted that determining the number of inequivalent differentiable structures on a topological space is a very difficult problem, even for $\mathbb{R}^n$. In the case of $\mathbb{R}^n$, it turns out that any two smooth differentiable structures are diffeomorphic, except for $n = 4$. For $n = 4$, it took some very hard and deep work to show that there are uncountably many distinct diffeomorphism classes of smooth differentiable structures. The case of the spheres S^n is even more mysterious. It is known that there is a single diffeomorphism class for $n = 1, 2, 3$, but for $n = 4$ the answer is unknown! For $n = 15$, there are $16,256$ distinct classes; for more about these issues, see Conlon [33] (Chapter 3). It is also known that every topological manifold admits a smooth structure for $n = 1, 2, 3$. However, for $n = 4$, there exist nonsmoothable manifolds; see Conlon [33] (Chapter 3).

In some cases, M does not come with a topology in an obvious (or natural) way and a slight variation of Definition 7.2 is more convenient in such a situation:

Definition 7.6. Given a set M, given some integer $n \geq 1$ and given some k such that k is either an integer $k \geq 1$ or $k = \infty$, a C^k *n-atlas* (or *n-atlas of class* C^k) $\mathcal{A}$ is a family of charts $\{(U_i, \varphi_i)\}$, such that

(1) Each U_i is a subset of M and $\varphi_i \colon U_i \to \varphi_i(U_i)$ is a bijection onto an open subset $\varphi_i(U_i) \subseteq \mathbb{R}^n$, for all i;

(2) The U_i cover M; that is,
$$M = \bigcup_i U_i;$$

(3) Whenever $U_i \cap U_j \neq \emptyset$, the sets $\varphi_i(U_i \cap U_j)$ and $\varphi_j(U_i \cap U_j)$ are open in $\mathbb{R}^n$ and the transition maps φ_i^j and φ_j^i are C^k-diffeomorphisms.

Then the notion of a chart being compatible with an atlas and of two atlases being compatible is just as before, and we get a new definition of a manifold analogous to Definition 7.4. But this time we give M the topology in which the open sets are arbitrary unions of domains of charts U_i, more precisely, the U_i's of the maximal atlas defining the differentiable structure on M.

It is not difficult to verify that the axioms of a topology are verified, and M is indeed a topological space with this topology. It can also be shown that when M is equipped with the above topology, then the maps $\varphi_i \colon U_i \to \varphi_i(U_i)$ are homeomorphisms, so M is a manifold according to Definition 7.4. We also require that under this topology, M is Hausdorff and second-countable. A sufficient condition (in fact, also necessary!) for being second-countable is that some atlas be countable. A sufficient condition of M to be Hausdorff is that for all $p, q \in M$ with $p \neq q$, either $p, q \in U_i$ for some U_i, or $p \in U_i$ and $q \in U_j$ for some disjoint U_i, U_j. Thus, we are back to the original notion of a manifold where it is assumed that M is already a topological space.

One can also define the topology on M in terms of any of the atlases $\mathcal{A}$ defining M (not only the maximal one) by requiring $U \subseteq M$ to be open iff $\varphi_i(U \cap U_i)$ is open in $\mathbb{R}^n$, for every chart (U_i, φ_i) in the atlas $\mathcal{A}$. Then one can prove that we obtain the same topology as the topology induced by the maximal atlas. For details, see Berger and Gostiaux [15], Chapter 2.

If the underlying topological space of a manifold is compact, then M has some finite atlas. Also, if $\mathcal{A}$ is some atlas for M and (U, φ) is a chart in $\mathcal{A}$, for any (nonempty) open subset $V \subseteq U$, we get a chart $(V, \varphi \restriction V)$, and it is obvious that this chart is compatible with $\mathcal{A}$. Thus, $(V, \varphi \restriction V)$ is also a chart for M. This observation shows that if U is any open subset of a C^k-manifold M, then U is also a C^k-manifold whose charts are the restrictions of charts on M to U.

We are now fully prepared to present a variety of examples.

Example 7.1. The sphere S^n.

Using the stereographic projections (from the north pole and the south pole), we can define two charts on S^n and show that S^n is a smooth manifold. Let $\sigma_N\colon S^n - \{N\} \to \mathbb{R}^n$ and $\sigma_S\colon S^n - \{S\} \to \mathbb{R}^n$, where $N = (0, \cdots, 0, 1) \in \mathbb{R}^{n+1}$ (the north pole) and $S = (0, \cdots, 0, -1) \in \mathbb{R}^{n+1}$ (the south pole) be the maps called respectively *stereographic projection from the north pole* and *stereographic projection from the south pole*, given by

$$\sigma_N(x_1, \ldots, x_{n+1}) = \frac{1}{1 - x_{n+1}}\,(x_1, \ldots, x_n) \quad \text{and} \quad \sigma_S(x_1, \ldots, x_{n+1}) = \frac{1}{1 + x_{n+1}}\,(x_1, \ldots, x_n).$$

The inverse stereographic projections are given by

$$\sigma_N^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right) + 1}\left(2x_1, \ldots, 2x_n, \left(\sum_{i=1}^n x_i^2\right) - 1\right)$$

and

$$\sigma_S^{-1}(x_1, \ldots, x_n) = \frac{1}{\left(\sum_{i=1}^n x_i^2\right) + 1}\left(2x_1, \ldots, 2x_n, -\left(\sum_{i=1}^n x_i^2\right) + 1\right).$$

See Example 3.1 for the case of $n = 2$. Thus, if we let $U_N = S^n - \{N\}$ and $U_S = S^n - \{S\}$, we see that U_N and U_S are two open subsets covering S^n, both homeomorphic to $\mathbb{R}^n$. Furthermore, it is easily checked that on the overlap $U_N \cap U_S = S^n - \{N, S\}$, the transition maps

$$\mathcal{I} = \sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1}$$

defined on $\varphi_N(U_N \cap U_S) = \varphi_S(U_N \cap U_S) = \mathbb{R}^n - \{0\}$, are given by

$$(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2}\,(x_1, \ldots, x_n);$$

that is, the inversion $\mathcal{I}$ of center $O = (0, \ldots, 0)$ and power 1. Clearly, this map is smooth on $\mathbb{R}^n - \{O\}$, so we conclude that (U_N, σ_N) and (U_S, σ_S) form a smooth atlas for S^n.

Example 7.2. Smooth manifolds in $\mathbb{R}^N$.

Any m-dimensional embeddded manifold M in $\mathbb{R}^N$ is a smooth manifold, because by Lemma 3.2, the inverse maps $\varphi^{-1}\colon U \to \Omega$ of the parametrizations $\varphi\colon \Omega \to U$ are charts that yield smooth transition functions. In particular, by Theorem 3.8, any linear Lie group is a smooth manifold.

Example 7.3. The projective space $\mathbb{RP}^n$. To define an atlas on $\mathbb{RP}^n$, it is convenient to view $\mathbb{RP}^n$ as the set of equivalence classes of vectors in $\mathbb{R}^{n+1} - \{0\}$ modulo the equivalence relation

$$u \sim v \quad \text{iff} \quad v = \lambda u, \quad \text{for some} \quad \lambda \neq 0 \in \mathbb{R}.$$

Given any $p = [x_1, \ldots, x_{n+1}] \in \mathbb{RP}^n$, we call $(x_1, \ldots, x_{n+1})$ the *homogeneous coordinates* of p. It is customary to write $(x_1\colon \cdots \colon x_{n+1})$ instead of $[x_1, \ldots, x_{n+1}]$. (Actually, in most books, the indexing starts with 0, i.e., homogeneous coordinates for $\mathbb{RP}^n$ are written as $(x_0\colon \cdots \colon x_n)$.) Now, $\mathbb{RP}^n$ can also be viewed as the quotient of the sphere S^n under the equivalence relation where any two antipodal points x and $-x$ are identified. It is not hard to show that the projection $\pi\colon S^n \to \mathbb{RP}^n$ is both open and closed. Since S^n is compact and second-countable, we can apply Propositions 12.31 and 12.33 to prove that under the quotient topology, $\mathbb{RP}^n$ is Hausdorff, second-countable, and compact.

We define charts in the following way. For any i, with $1 \leq i \leq n+1$, let

$$U_i = \{(x_1\colon \cdots \colon x_{n+1}) \in \mathbb{RP}^n \mid x_i \neq 0\}.$$

Observe that U_i is well defined, because if $(y_1\colon \cdots \colon y_{n+1}) = (x_1\colon \cdots \colon x_{n+1})$, then there is some $\lambda \neq 0$ so that $y_j = \lambda x_j$, for $j = 1, \ldots, n+1$. We can define a homeomorphism φ_i of U_i onto $\mathbb{R}^n$ as follows:

$$\varphi_i(x_1\colon \cdots \colon x_{n+1}) = \left(\frac{x_1}{x_i}, \ldots, \frac{x_{i-1}}{x_i}, \frac{x_{i+1}}{x_i}, \ldots, \frac{x_{n+1}}{x_i}\right),$$

where the ith component is omitted. Again, it is clear that this map is well defined since it only involves ratios. We can also define the maps ψ_i from $\mathbb{R}^n$ to $U_i \subseteq \mathbb{RP}^n$, given by

$$\psi_i(x_1, \ldots, x_n) = (x_1\colon \cdots \colon x_{i-1}\colon 1\colon x_i\colon \cdots \colon x_n),$$

where the 1 goes in the ith slot, for $i = 1, \ldots, n+1$.

One easily checks that φ_i and ψ_i are mutual inverses, so the φ_i are homeomorphisms. On the overlap $U_i \cap U_j$, (where $i \neq j$), as $x_j \neq 0$, we have

$$(\varphi_j \circ \varphi_i^{-1})(x_1, \ldots, x_n) = \left(\frac{x_1}{x_{j-1}}, \ldots, \frac{x_{i-1}}{x_{j-1}}, \frac{1}{x_{j-1}}, \frac{x_i}{x_{j-1}}, \ldots, \frac{x_j}{x_{j-1}}, \frac{x_{j+1}}{x_{j-1}}, \ldots, \frac{x_n}{x_{j-1}}\right).$$

(We assumed that $i < j$; the case $j < i$ is similar.) This is clearly a smooth function from $\varphi_i(U_i \cap U_j)$ to $\varphi_j(U_i \cap U_j)$. As the U_i cover $\mathbb{RP}^n$, we conclude that the (U_i, φ_i) are $n+1$ charts making a smooth atlas for $\mathbb{RP}^n$. Intuitively, the space $\mathbb{RP}^n$ is obtained by gluing the open subsets U_i on their overlaps. Even for $n = 3$, this is not easy to visualize!

Example 7.4. The Grassmannian $G(k, n)$. Recall that $G(k, n)$ is the set of all k-dimensional linear subspaces of $\mathbb{R}^n$, also called k-planes. Every k-plane W is the linear span of k linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$; furthermore, $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ both span W iff there is an invertible $k \times k$-matrix $\Lambda = (\lambda_{ij})$ such that

$$v_j = \sum_{i=1}^{k} \lambda_{ij} u_i, \quad 1 \leq j \leq k.$$

Obviously there is a bijection between the collection of k linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$ and the collection of $n \times k$ matrices of rank k. Furthermore, two $n \times k$ matrices A and B of rank k represent the same k-plane iff

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda. \tag{$*$}$$

The Grassmannian $G(k, n)$ can be viewed of the set of equivalence classes of $n \times k$ matrices of rank k under the equivalence relation given by $(*)$. (Note the analogy with projective spaces where two vectors u, v represent the same point iff $v = \lambda u$ for some invertible $\lambda \in \mathbb{R}$.)

The set of $n \times k$ matrices of rank k is a subset of $\mathbb{R}^{n \times k}$, in fact an open subset.

One can show that the equivalence relation on $n \times k$ matrices of rank k given by

$$B = A\Lambda, \quad \text{for some invertible } k \times k \text{ matrix, } \Lambda,$$

is open, and that the graph of this equivalence relation is closed. For some help proving these facts, see Problem 7.2 in Tu [112]. By Proposition 12.32, the Grassmannian $G(k, n)$ is Hausdorff and second-countable.

We can define the domain of charts (according to Definition 7.2) on $G(k, n)$ as follows: For every subset $S = \{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$, let U_S be the subset of equivalence classes of $n \times k$ matrices A of rank k whose rows of index in $S = \{i_1, \ldots, i_k\}$ form an invertible $k \times k$ matrix denoted A_S. Note U_S is open in the quotient topology of $G(k, n)$ since the existence of an invertible $k \times k$ matrix is equivalent to the open condition of $\det A_S \neq 0$. Observe that the $k \times k$ matrix consisting of the rows of the matrix AA_S^{-1} whose index belong to S is the identity matrix I_k. Therefore, we can define a map $\varphi_S \colon U_S \to \mathbb{R}^{(n-k) \times k}$ where $\varphi_S(A)$ is equal to the $(n-k) \times k$ matrix obtained by deleting the rows of index in S from AA_S^{-1}.

We need to check that this map is well defined, i.e., that it does not depend on the matrix A representing W. Let us do this in the case where $S = \{1, \ldots, k\}$, which is notationally simpler. The general case can be reduced to this one using a suitable permutation.

If $B = A\Lambda$, with Λ invertible, if we write

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where A_1 and B_1 are $k \times k$ matrices and A_2 and B_2 are $(n-k) \times k$ matrices, as $B = A\Lambda$, we get $B_1 = A_1\Lambda$ and $B_2 = A_2\Lambda$, from which we deduce that

$$\begin{pmatrix} B_1 \\ B_2 \end{pmatrix} B_1^{-1} = \begin{pmatrix} I_k \\ B_2 B_1^{-1} \end{pmatrix} = \begin{pmatrix} I_k \\ A_2\Lambda\Lambda^{-1}A_1^{-1} \end{pmatrix} = \begin{pmatrix} I_k \\ A_2 A_1^{-1} \end{pmatrix} = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} A_1^{-1}.$$

Therefore, our map is indeed well-defined.

Here is an example for $n = 6$ and $k = 3$. Let A be the matrix

$$A = \begin{pmatrix} 2 & 3 & 5 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & -1 & 2 \\ 1 & 0 & 0 \\ 2 & -1 & 2 \end{pmatrix}$$

and let

$$S = \{2, 3, 5\}.$$

Then we have

$$A_{\{2,3,5\}} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

and we find that

$$A_{\{2,3,5\}}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix},$$

and

$$AA_{\{2,3,5\}}^{-1} = \begin{pmatrix} 5 & -2 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -3 & 2 \\ 0 & 0 & 1 \\ 2 & -3 & 3 \end{pmatrix}.$$

Therefore,

$$\varphi_{\{2,3,5\}}(A) = \begin{pmatrix} 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix}.$$

We can define its inverse ψ_S as follows: let π_S be the permutation of $\{1, \ldots, n\}$ sending $\{1, \ldots, k\}$ to S defined such that if $S = \{i_1 < \cdots < i_k\}$, then $\pi_S(j) = i_j$ for $j = 1, \ldots, k$, and if $\{h_1 < \cdots < h_{n-k}\} = \{1, \ldots, n\} - S$, then $\pi_S(k + j) = h_j$ for $j = 1, \ldots, n-k$ (this is a k-shuffle). If P_S is the permutation matrix associated with π_S, for any $(n-k)\times k$ matrix M, let

$$\psi_S(M) = P_S \begin{pmatrix} I_k \\ M \end{pmatrix},$$

actually the equivalence class of $P_S\binom{I_k}{M}$ in U_S. The effect of ψ_S is to "insert into M" the rows of the identity matrix I_k as the rows of index from S. Using our previous example where $n = 6, k = 3$ and $S = \{2, 3, 5\}$, we have

$$M = \begin{pmatrix} 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix},$$

the permutation π_S is given by

$$\pi_S = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 1 & 4 & 6 \end{pmatrix},$$

whose permutation matrix is

$$P_{\{2,3,5\}} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\psi_{\{2,3,5\}}(M) = P_{\{2,3,5\}} \begin{pmatrix} I_3 \\ M \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix} = \begin{pmatrix} 5 & -2 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -3 & 2 \\ 0 & 0 & 1 \\ 2 & -3 & 3 \end{pmatrix}.$$

Since the permutation π_S is a k-shuffle that sends $\{1, \ldots, k\}$ to S, we see that $\varphi_S(A)$ is also obtained by first forming $P_S^{-1}A$, which brings the rows of index in S to the first k rows, then forming $P_S^{-1}A(P_S^{-1}A)^{-1}_{\{1,\ldots,k\}}$, and finally deleting the first k rows. If we write A and P_S^{-1} in block form as

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad P_S^{-1} = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix},$$

with A_1 a $k\times k$ matrix, A_2 a $(n-k)\times k$ matrix, P_1 a $k\times k$ matrix, P_4 a $(n-k)\times(n-k)$ matrix, P_2 a $k\times(n-k)$ matrix, and P_3 a $(n-k)\times k$ matrix, then

$$P_S^{-1}A = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = \begin{pmatrix} P_1A_1 + P_2A_2 \\ P_3A_1 + P_4A_2 \end{pmatrix},$$

so

$$\begin{aligned} P_S^{-1}A(P_S^{-1}A)_{\{1,\ldots,k\}}^{-1} &= \begin{pmatrix} P_1A_1 + P_2A_2 \\ P_3A_1 + P_4A_2 \end{pmatrix} (P_1A_1 + P_2A_2)^{-1} \\ &= \begin{pmatrix} I_k \\ (P_3A_1 + P_4A_2)(P_1A_1 + P_2A_2)^{-1} \end{pmatrix}, \end{aligned}$$

and thus,

$$\varphi_S(A) = (P_3A_1 + P_4A_2)(P_1A_1 + P_2A_2)^{-1}.$$

With the above example,

$$P_{\{2,3,5\}}^{-1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and then

$$P_{\{2,3,5\}}^{-1}A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 2 & 3 & 5 \\ 1 & -1 & 2 \\ 2 & -1 & 2 \end{pmatrix},$$

$$(P_{\{2,3,5\}}^{-1}A)_{\{1,2,3\}}^{-1} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix},$$

and

$$P_{\{2,3,5\}}^{-1}A(P_{\{2,3,5\}}^{-1}A)_{\{1,2,3\}}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix},$$

which does yield

$$\varphi_{\{2,3,5\}}(A) = \begin{pmatrix} 5 & -2 & -1 \\ 2 & -3 & 2 \\ 2 & -3 & 3 \end{pmatrix}.$$

At this stage, we have charts that are bijections from subsets U_S of $G(k,n)$ to open subsets, namely, $\mathbb{R}^{(n-k)\times k}$. The reader can check that the transition map $\varphi_T \circ \varphi_S^{-1}$ from $\varphi_S(U_S \cap U_T)$ to $\varphi_T(U_S \cap U_T)$ is given by

$$M \mapsto (P_3 + P_4M)(P_1 + P_2M)^{-1},$$

where

$$\begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix} = P_T^{-1} P_S$$

is the matrix of the permutation $\pi_T^{-1} \circ \pi_S$ and M is an $(n-k) \times k$ matrix. This map is smooth, as the inversion of a matrix uses the cofactor matrix which relies on the smoothness of the determinant, and so the charts (U_S, φ_S) form a smooth atlas for $G(k,n)$. Finally, one can check that the conditions of Definition 7.2 are satisfied, so the atlas just defined makes $G(k,n)$ into a topological space and a smooth manifold.

The Grassmannian $G(k,n)$ is actually compact. To see this, observe that if W is any k-plane, then using the Gram-Schmidt orthonormalization procedure, every basis $B = (b_1, \ldots, b_k)$ for W yields an orthonormal basis $U = (u_1, \ldots, u_k)$, and there is an invertible $k \times k$ matrix Λ such that

$$U = B\Lambda,$$

where the the columns of B are the b_j's and the columns of U are the u_j's. Thus we may assume that the representatives of W are matrices U which have orthonormal columns and are characterized by the equation

$$U^\top U = I_k.$$

The space of such matrices is closed and clearly bounded in $\mathbb{R}^{n \times k}$, and thus compact. In fact, the space of $n \times k$ matrices U satisfying $U^\top U = I$ is the Stiefel manifold $S(k,n)$. Observe that if U and V are two $n \times k$ matrices such that $U^\top U = I$ and $V^\top V = I$ and if $V = U\Lambda$ for some invertible $k \times k$ matrix Λ, then $\Lambda \in \mathbf{O}(k)$. Then $G(k,n)$ is the orbit space obtained by making $\mathbf{O}(k)$ act on $S(k,n)$ on the right, i.e. $S(k,n)/\mathbf{O}(k) \cong G(k,n)$, and since $S(k,n)$ is compact, we conclude that $G(k,n)$ is also compact as it is the continuous image of a projection map.

Remark: The reader should have no difficulty proving that the collection of k-planes represented by matrices in U_S is precisely the set of k-planes W supplementary to the $(n-k)$-plane spanned by the canonical basis vectors $e_{j_{k+1}}, \ldots, e_{j_n}$ (i.e., $\mathrm{span}(W \cup \{e_{j_{k+1}}, \ldots, e_{j_n}\}) = \mathbb{R}^n$, where $S = \{i_1, \ldots, i_k\}$ and $\{j_{k+1}, \ldots, j_n\} = \{1, \ldots, n\} - S$).

Example 7.5. Product Manifolds.

Let M_1 and M_2 be two C^k-manifolds of dimension n_1 and n_2, respectively. The topological space $M_1 \times M_2$ with the product topology (the open sets of $M_1 \times M_2$ are arbitrary unions of sets of the form $U \times V$, where U is open in M_1 and V is open in M_2) can be given the structure of a C^k-manifold of dimension $n_1 + n_2$ by defining charts as follows: For any two charts (U_i, φ_i) on M_1 and (V_j, ψ_j) on M_2, we declare that $(U_i \times V_j, \varphi_i \times \psi_j)$ is a chart on $M_1 \times M_2$, where $\varphi_i \times \psi_j \colon U_i \times V_j \to \mathbb{R}^{n_1+n_2}$ is defined so that

$$\varphi_i \times \psi_j(p,q) = (\varphi_i(p), \psi_j(q)), \quad \text{for all } (p,q) \in U_i \times V_j.$$

See Figure 7.5.

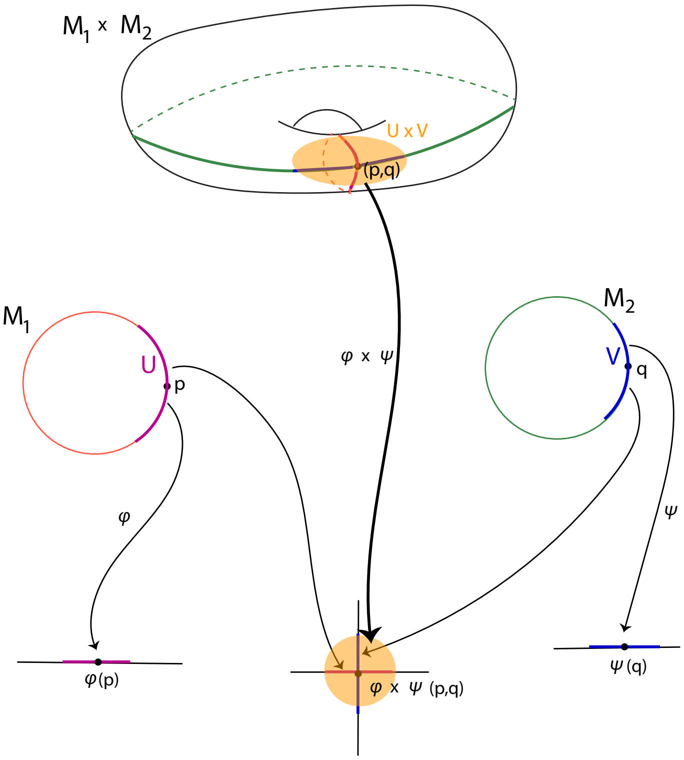


Figure 7.5: A chart for the torus as the product manifold $S^1 \times S^1$.

We define C^k-maps between manifolds as follows:

Definition 7.7. Given any two C^k-manifolds M and N of dimension m and n respectively, a C^k*-map* is a continuous function $h\colon M \to N$ satisfying the following property: For every $p \in M$, there is some chart (U, φ) at p and some chart (V, ψ) at $q = h(p)$, with $h(U) \subseteq V$ and

$$\psi \circ h \circ \varphi^{-1} \colon \varphi(U) \longrightarrow \psi(V)$$

a C^k-function. See Figure 7.6.

It is easily shown that Definition 7.7 does not depend on the choice of charts.

The requirement in Definition 7.7 that $h\colon M \to N$ should be continuous is actually redundant. Indeed, since φ and ψ are homeomorphisms, φ and ψ^{-1} are continuous, and since $\varphi(U)$ is an open subset of $\mathbb{R}^m$ and $\psi(V)$ is an open subset of $\mathbb{R}^n$, the function $\psi \circ h \circ \varphi^{-1} \colon \varphi(U) \to \psi(V)$ being a C^k-function is continuous, so the restriction of h to U being equal to the composition of the three continuous maps

$$\psi^{-1} \circ (\psi \circ h \circ \varphi^{-1}) \circ \varphi$$

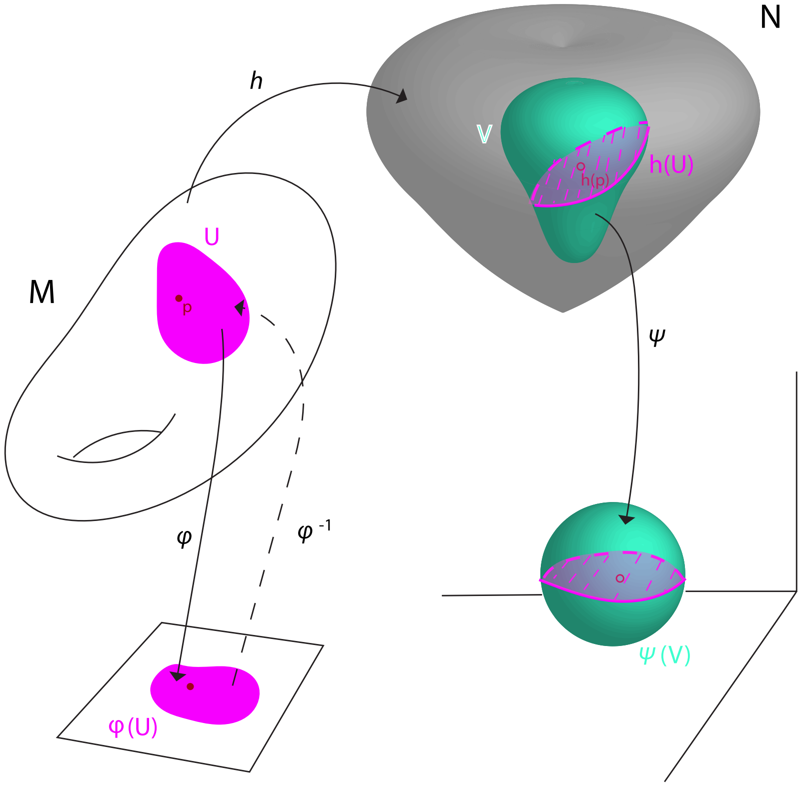


Figure 7.6: The C^k map from M to N, where M is a 2-dimensional manifold and N is a 3-dimensional manifold.

is also continuous on U. Since this holds on some open subset containing p, for every $p \in M$, the function h is continuous on M.

Other definitions of a C^k-map appear in the literature, some requiring continuity. The following proposition from Berger and Gostiaux [15] (Theorem 2.3.3) helps clarifying how these definitions relate.

Proposition 7.1. *Let $h\colon M \to N$ be a function between two manifolds M and N. The following equivalences hold.*

(1) The map h is continuous, and for every $p \in M$, for every chart (U, φ) at p and every chart (V, ψ) at $h(p)$, the function $\psi \circ h \circ \varphi^{-1}$ from $\varphi(U \cap h^{-1}(V))$ to $\psi(V)$ is a C^k-function.

(2) The map h is continuous, and for every $p \in M$, for every chart (U, φ) at p and every chart (V, ψ) at $h(p)$, if $h(U) \subseteq V$, then the function $\psi \circ h \circ \varphi^{-1}$ from $\varphi(U)$ to $\psi(V)$ is a C^k-function.

(3) For every $p \in M$, there is some chart (U, φ) at p and some chart (V, ψ) at $q = h(p)$ with $h(U) \subseteq V$, such that the function $\psi \circ h \circ \varphi^{-1}$ from $\varphi(U)$ to $\psi(V)$ is a C^k-function.

Observe that Condition (3) states exactly the conditions of Definition 7.7, with the continuity requirement omitted. Condition (1) is used by many texts. The continuity of h is required to ensure that $h^{-1}(V)$ is an open set. The implication (2) ⇒ (3) also requires the continuity of h.

Even though the continuity requirement in Definition 7.7 is redundant, it seems to us that it does not hurt to emphasize that C^k-maps are continuous.

In the special case where $N = \mathbb{R}$, we obtain the notion of a *C^k-function on M*. One checks immediately that a function $f \colon M \to \mathbb{R}$ is a C^k-map iff the following condition holds.

Definition 7.8. A function $f \colon M \to \mathbb{R}$ is a *C^k-map* iff for every $p \in M$, there is some chart (U, φ) at p so that

$$f \circ \varphi^{-1} \colon \varphi(U) \longrightarrow \mathbb{R}$$

is a C^k-function. See Figure 7.7.

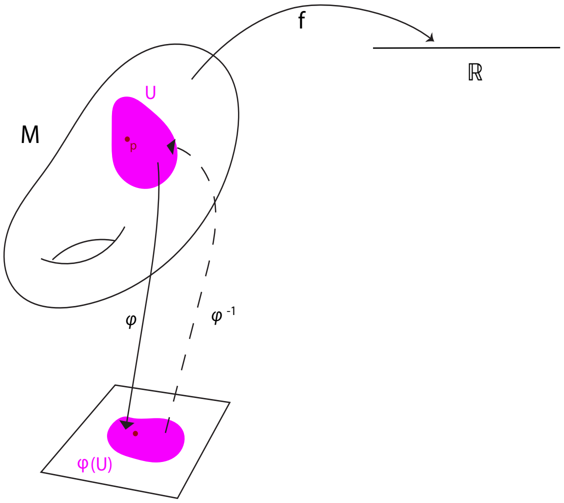


Figure 7.7: A schematic illustration of a C^k-function on the torus M.

If U is an open subset of M, the set of C^k-functions on U is denoted by $\mathcal{C}^k(U)$. In particular, $\mathcal{C}^k(M)$ denotes the set of C^k-functions on the manifold, M. Observe that $\mathcal{C}^k(U)$ is a commutative ring.

On the other hand, if M is an open interval of $\mathbb{R}$, say $M = (a, b)$, then $\gamma \colon (a, b) \to N$ is called a *C^k-curve* in N. One checks immediately that a function $\gamma \colon (a, b) \to N$ is a C^k-map iff the following condition holds.

Definition 7.9. A function $\gamma\colon (a,b) \to N$ is a C^k-map iff for every $q \in N$, there is some chart (V, ψ) at q and some open subinterval (c,d) of (a,b), so that $\gamma((c,d)) \subseteq V$ and

$$\psi \circ \gamma\colon (c,d) \longrightarrow \psi(V)$$

is a C^k-function. See Figure 7.8.

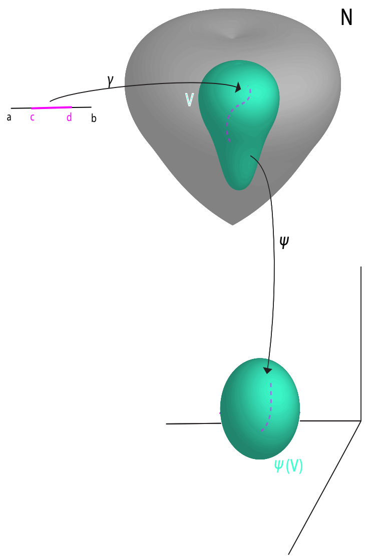


Figure 7.8: A schematic illustration of a C^k-curve in the solid spheroid N.

It is clear that the composition of C^k-maps is a C^k-map.

Definition 7.10. A C^k-map $h\colon M \to N$ between two manifolds is a C^k-*diffeomorphism* iff h has an inverse $h^{-1}\colon N \to M$ (i.e., $h^{-1} \circ h = \mathrm{id}_M$ and $h \circ h^{-1} = \mathrm{id}_N$), and both h and h^{-1} are C^k-maps (in particular, h and h^{-1} are homeomorphisms).

Next we define tangent vectors.

7.2 Tangent Vectors, Tangent Spaces

Let M be a C^k manifold of dimension n, with $k \geq 1$. The purpose of the next three sections is to define the tangent space $T_p(M)$ at a point p of a manifold M. We provide three definitions of the notion of a tangent vector to a manifold and prove their equivalence.

The first definition uses equivalence classes of curves on a manifold and is the most intuitive.

The second definition makes heavy use of the charts and of the transition functions. It is also quite intuitive and it is easy to see that that it is equivalent to the first definition. The second definition is the most convenient one to define the manifold structure of the tangent bundle $T(M)$ (see Section 9.1).

The third definition (given in the next section) is based on the view that a tangent vector v, at p, induces a differential operator on real-valued functions f, defined locally near p; namely, the map $f \mapsto v(f)$ is a linear form satisfying an additional property akin to the rule for taking the derivative of a product (the Leibniz property). Such linear forms are called *point-derivations*. This third definition is more intrinsic than the first two but more abstract. However, for any point p on the manifold M and for any chart whose domain contains p, there is a convenient basis of the tangent space $T_p(M)$. The third definition is also the most convenient one to define vector fields. A few technical complications arise when M is not a smooth manifold (when $k \neq \infty$), but these are easily overcome using "stationary germs."

As pointed out by Serre in [105] (Chapter III, Section 8), the relationship between the first definition and the third definition of the tangent space at p is best described by a nondegenerate pairing which shows that $T_p(M)$ is the dual of the space of point derivations at p that vanish on stationary germs. This pairing is presented in Section 7.4.

The most intuitive method to define tangent vectors is to use curves. Let $p \in M$ be any point on M and let $\gamma\colon (-\epsilon, \epsilon) \to M$ be a C^1-curve passing through p, that is, with $\gamma(0) = p$. Unfortunately, if M is not embedded in any $\mathbb{R}^N$, the derivative $\gamma'(0)$ does not make sense. However, for any chart, (U, φ), at p, the map $\varphi\circ\gamma$ is a C^1-curve in $\mathbb{R}^n$ and the tangent vector $v = (\varphi \circ \gamma)'(0)$ is well defined. The trouble is that different curves may yield the same v!

To remedy this problem, we define an equivalence relation on curves through p as follows:

Definition 7.11. Given a C^k manifold, M, of dimension n, for any $p \in M$, two C^1-curves, $\gamma_1\colon (-\epsilon_1, \epsilon_1) \to M$ and $\gamma_2\colon (-\epsilon_2, \epsilon_2) \to M$, through p (i.e., $\gamma_1(0) = \gamma_2(0) = p$) are *equivalent* iff there is some chart, (U, φ), at p so that

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0).$$

See Figure 7.9.

The problem is that this definition seems to depend on the choice of the chart. Fortunately, this is not the case. For if (V, ψ) is another chart at p, as p belongs both to U and V, we have $U \cap V \neq 0$, so the transition function $\eta = \psi \circ \varphi^{-1}$ is C^k and, by the chain rule, we have

$$\begin{aligned}
(\psi \circ \gamma_1)'(0) &= (\eta \circ \varphi \circ \gamma_1)'(0)\\
&= \eta'(\varphi(p))((\varphi \circ \gamma_1)'(0))\\
&= \eta'(\varphi(p))((\varphi \circ \gamma_2)'(0))\\
&= (\eta \circ \varphi \circ \gamma_2)'(0)\\
&= (\psi \circ \gamma_2)'(0).
\end{aligned}$$

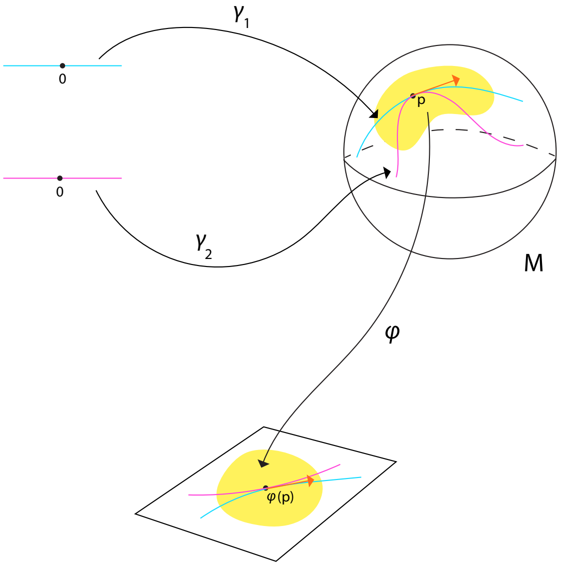


Figure 7.9: Equivalent curves γ_1, in blue, and γ_2, in pink.

This leads us to the first definition of a tangent vector.

Definition 7.12. (Tangent Vectors, Version 1) Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, a *tangent vector to* M *at* p is any equivalence class $u = [\gamma]$ of C^1-curves γ through p on M, modulo the equivalence relation defined in Definition 7.11. The set of all tangent vectors at p is denoted by $T_p(M)$ (or T_pM).

In order to make T_pM into a vector space, given a chart (U, φ) with $p \in U$, we observe that the map $\overline{\varphi}_U \colon T_pM \to \mathbb{R}^n$ given by

$$\overline{\varphi}_U([\gamma]) = (\varphi \circ \gamma)'(0)$$

is a bijection, where $[\gamma]$ is the equivalence class of a curve γ in M through p (with $\gamma(0) = p$). The map $\overline{\varphi}_U$ is injective by definition of the equivalence relation on curves; it is surjective, because for every vector $v \in \mathbb{R}^n$, if γ_v is the curve given by $\gamma_v(t) = \varphi^{-1}(\varphi(p) + tv)$, then $(\varphi \circ \gamma_v)'(0) = v$, and so $\overline{\varphi}_U([\gamma_v]) = v$.

Observe that for any chart (U, φ) at p, the equivalence class $[\gamma]$ of all curves through p such that $(\varphi \circ \gamma)'(0) = v$ for some given vector $v \in \mathbb{R}^n$ is determined by the special curve γ_v defined above.

The vector space structure on T_pM is defined as follows. For any chart (U, φ) at p, given any two equivalences classes $[\gamma_1]$ and $[\gamma_2]$ in T_pM, for any real λ, we set

$$\begin{aligned} [\gamma_1] + [\gamma_2] &= \overline{\varphi}_U^{-1}(\overline{\varphi}_U([\gamma_1]) + \overline{\varphi}_U([\gamma_2])) \\ \lambda[\gamma_1] &= \overline{\varphi}_U^{-1}(\lambda\overline{\varphi}_U([\gamma_1])). \end{aligned}$$

If (V, ψ) is any other chart at p, since by the chain rule

$$(\psi \circ \gamma)'(0) = (\psi \circ \varphi^{-1})'_{\varphi(p)} \circ (\varphi \circ \gamma)'(0),$$

it follows that

$$\overline{\psi}_V = (\psi \circ \varphi^{-1})'_{\varphi(p)} \circ \overline{\varphi}_U.$$

Since $(\psi \circ \varphi^{-1})'_{\varphi(p)}$ is a linear isomorphism, we see that the vector space structure defined above does not depend on the choice of chart at p. Therefore, with this vector space structure on T_pM, the map $\overline{\varphi}_U \colon T_pM \to \mathbb{R}^n$ is a linear isomorphism. This shows that T_pM is a vector space of dimension $n =$ dimension of M.

In particular, if M is an n-dimensional smooth embedded manifold in $\mathbb{R}^N$ and if γ is a curve in M through p, then $\gamma'(0) = u$ is well defined as a vector in $\mathbb{R}^N$, and the equivalence class of all curves γ through p such that $(\varphi \circ \gamma)'(0)$ is the same vector in some chart $\varphi \colon U \to \Omega$ can be identified with u. Thus, the tangent space T_pM to M at p is isomorphic to

$$\{\gamma'(0) \mid \gamma \colon (-\epsilon, \epsilon) \to M \text{ is a } C^1\text{-curve with } \gamma(0) = p\}.$$

In the special case of a linear Lie group G, Proposition 3.10 shows that the exponential map $\exp \colon \mathfrak{g} \to G$ is a diffeomorphism from some open subset of $\mathfrak{g}$ containing 0 to some open subset of G containing I. For every $g \in G$, since $L_g \colon G \to G$ is a diffeomorphism, the map $L_g \circ \exp \colon \mathfrak{g} \to G$ is a diffeomorphism from some open subset of $\mathfrak{g}$ containing 0 to some open subset of G containing g. Furthermore,

$$dL_g(\mathfrak{g}) = L_g(\mathfrak{g}) = g\mathfrak{g} = \{gX \mid X \in \mathfrak{g}\}.$$

Thus, we obtain smooth parametrizations of G whose inverses are charts on G, and since by definition of $\mathfrak{g}$, for every $X \in \mathfrak{g}$, the curve $\gamma(t) = ge^{tX}$ is a curve through g in G such that $\gamma'(0) = gX$, we see that the tangent space T_gG to G at g is isomorphic to $g\mathfrak{g}$.

One should observe that unless $M = \mathbb{R}^n$, in which case, for any $p, q \in \mathbb{R}^n$, the tangent space $T_q(M)$ is naturally isomorphic to the tangent space $T_p(M)$ by the translation $q - p$, for an arbitrary manifold, there is no relationship between $T_p(M)$ and $T_q(M)$ when $p \neq q$.

The second way of defining tangent vectors has the advantage that it makes it easier to define tangent bundles (see Section 9.1).

Definition 7.13. (Tangent Vectors, Version 2) Given any C^k-manifold, M, of dimension n, with $k \geq 1$, for any $p \in M$, consider the triples, (U, φ, u), where (U, φ) is any chart at p and u is any vector in $\mathbb{R}^n$. Say that two such triples (U, φ, u) and (V, ψ, v) are *equivalent* iff

$$(\psi \circ \varphi^{-1})'_{\varphi(p)}(u) = v.$$

See Figure 7.10. A *tangent vector* to M at p is an equivalence class of triples, $[(U, \varphi, u)]$, for the above equivalence relation.

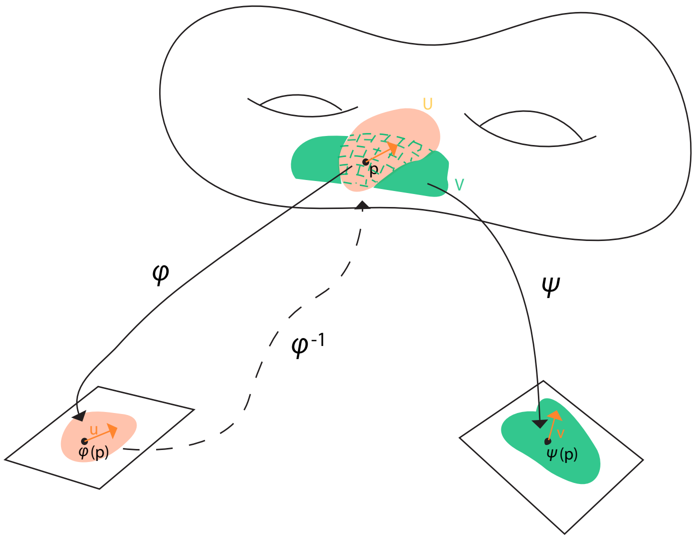


Figure 7.10: Two equivalent tangent vector u and v.

The intuition behind Definition 7.13 is quite clear: The vector u is considered as a tangent vector to $\mathbb{R}^n$ at $\varphi(p)$. If (U, φ) is a chart on M at p, we can define a natural bijection $\theta_{U,\varphi,p}\colon \mathbb{R}^n \to T_p(M)$ between $\mathbb{R}^n$ and $T_p(M)$, as follows: For any $u \in \mathbb{R}^n$,

$$\theta_{U,\varphi,p}\colon u \mapsto [(U, \varphi, u)].$$

As for Version 1 of tangent vectors, we can use the bijection $\theta_{U,\varphi,p}$ to transfer the vector space structure on $\mathbb{R}^n$ to T_pM so that $\theta_{U,\varphi,p}$ becomes a linear isomorphism. Given a chart (U, φ), for simplicity of notation if we denote the equivalence class of the triple (U, φ, u) by $[u]$, we set

$$\begin{aligned} [u] + [v] &= \theta_{U,\varphi,p}(\theta^{-1}_{U,\varphi,p}([u]) + \theta^{-1}_{U,\varphi,p}([v])) \\ \lambda[u] &= \theta_{U,\varphi,p}(\lambda\theta^{-1}_{U,\varphi,p}([u])). \end{aligned}$$

Since the equivalence between triples (U, φ, u) and (V, ψ, v) is given by

$$(\psi \circ \varphi^{-1})'_{\varphi(p)}(u) = v,$$

we have

$$\theta^{-1}_{V,\psi,p} = (\psi \circ \varphi^{-1})'_{\varphi(p)} \circ \theta^{-1}_{U,\varphi,p},$$

so the vector space structure on T_pM does not depend on the choice of chart at p.

The equivalence of this definition with the definition in terms of curves (Definition 7.12) is easy to prove.

Proposition 7.2. *Let M be any C^k-manifold of dimension n, with $k \geq 1$. For every $p \in M$, for every chart, (U, φ), at p, if $x = [\gamma]$ is any tangent vector (Version 1) given by some equivalence class of C^1-curves $\gamma\colon (-\epsilon, +\epsilon) \to M$ through p (i.e., $p = \gamma(0)$), then the map*

$$x \mapsto [(U, \varphi, (\varphi \circ \gamma)'(0))]$$

is an isomorphism between $T_p(M)$-Version 1 and $T_p(M)$-Version 2.

Proof. If σ is another curve equivalent to γ, then $(\varphi \circ \gamma)'(0) = (\varphi \circ \sigma)'(0)$, so the map is well-defined. It is clearly injective. As for surjectivity, define the curve γ_u on M through p by

$$\gamma_u(t) = \varphi^{-1}(\varphi(p) + tu);$$

see Figure 7.11. Then, $(\varphi \circ \gamma_u)(t) = \varphi(p) + tu$ and

$$(\varphi \circ \gamma_u)'(0) = u,$$

as desired. □

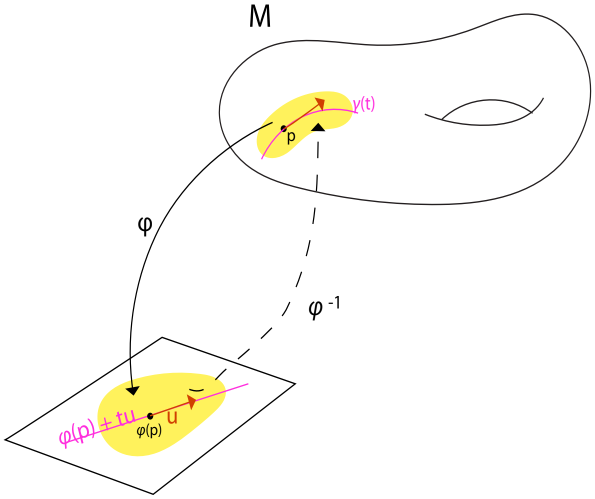


Figure 7.11: The tangent vector u is in one-to-one correspondence with the line through $\varphi(p)$ with direction u.

7.3 Tangent Vectors as Derivations

One of the defects of the above definitions of a tangent vector is that it has no clear relation to the C^k-differential structure of M. In particular, the definition does not seem to have anything to do with the functions defined locally at p. There is another way to define tangent vectors that reveals this connection more clearly. Moreover, such a definition is more intrinsic, i.e., does not refer explicitly to charts. Our presentation of this second approach is heavily inspired by Schwartz [104] (Chapter 3, Section 9) but also by Warner [114] and Serre [105] (Chapter III, Sections 7 and 8).

As a first step, consider the following: Let (U, φ) be a chart at $p \in M$ (where M is a C^k-manifold of dimension n, with $k \geq 1$) and let $x_i = pr_i \circ \varphi$, the ith local coordinate $(1 \leq i \leq n)$. For any real-valued function f defined on $p \in U$, set

$$\left(\frac{\partial}{\partial x_i}\right)_p f = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)}, \qquad 1 \leq i \leq n.$$

(Here, $(\partial g/\partial X_i)|_y$ denotes the partial derivative of a function $g\colon \mathbb{R}^n \to \mathbb{R}$ with respect to the ith coordinate, evaluated at y.)

We would expect that the function that maps f to the above value is a linear map on the set of functions defined locally at p, but there is technical difficulty: The set of real-valued functions defined locally at p is **not** a vector space! To see this, observe that if f is defined on an open $p \in U$ and g is defined on a different open $p \in V$, then we do not know how to define $f + g$. The problem is that we need to identify functions that agree on a smaller open subset. This leads to the notion of *germs*.

Definition 7.14. Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$, a *locally defined function at* p is a pair (U, f), where U is an open subset of M containing p and f is a real-valued function defined on U. Two locally defined functions (U, f) and (V, g) at p are *equivalent* iff there is some open subset $W \subseteq U \cap V$ containing p, so that

$$f \restriction W = g \restriction W.$$

The equivalence class of a locally defined function at p, denoted $[f]$ or $\mathbf{f}$, is called a *germ at* p.

One should check that the relation of Definition 7.14 is indeed an equivalence relation. Of course, the value at p of all the functions f in any germ $\mathbf{f}$, is $f(p)$. Thus, we set $\mathbf{f}(p) = f(p)$, for any $f \in \mathbf{f}$.

For example, for every $a \in (-1, 1)$, the locally defined functions $(\mathbb{R} - \{1\}, 1/(1-x))$ and $((-1, 1), \sum_{n=0}^{\infty} x^n)$ at a are equivalent.

We can define addition of germs, multiplication of a germ by a scalar, and multiplication of germs as follows. If (U, f) and (V, g) are two locally defined functions at p, we define $(U \cap V, f + g)$, $(U \cap V, fg)$ and $(U, \lambda f)$ as the locally defined functions at p given by $(f+g)(q) = f(q) + g(q)$ and $(fg)(q) = f(q)g(q)$ for all $q \in U \cap V$, and $(\lambda f)(q) = \lambda f(q)$ for all $q \in U$, with $\lambda \in \mathbb{R}$. Then, if $\mathbf{f} = [f]$ and $\mathbf{g} = [g]$ are two germs at p, we define

$$\begin{aligned} [f] + [g] &= [f + g] \\ \lambda[f] &= [\lambda f] \\ [f][g] &= [fg]. \end{aligned}$$

However, we have to check that these definitions make sense, that is, that they don't depend on the choice of representatives chosen in the equivalence classes $[f]$ and $[g]$. Let us give the details of this verification for the sum of two germs, $[f]$ and $[g]$.

We need to check that for any locally defined functions (U_1, f_1), (U_2, f_2), (V_1, g_1), and (V_2, g_2), at p, if (U_1, f_1) and (U_2, f_2) are equivalent and if (V_1, g_1) and (V_2, g_2) are equivalent, then $(U_1 \cap V_1, f_1 + g_1)$ and $(U_2 \cap V_2, f_2 + g_2)$ are equivalent. However, as (U_1, f_1) and (U_2, f_2) are equivalent, there is some $W_1 \subseteq U_1 \cap U_2$ so that $f_1 \restriction W_1 = f_2 \restriction W_1$ and as (V_1, g_1) and (V_2, g_2) are equivalent, there is some $W_2 \subseteq V_1 \cap V_2$ so that $g_1 \restriction W_2 = g_2 \restriction W_2$. Then, observe that $(f_1 + g_1) \restriction (W_1 \cap W_2) = (f_2 + g_2) \restriction (W_1 \cap W_2)$, which means that $[f_1 + g_1] = [f_2 + g_2]$. Therefore, $[f + g]$ does not depend on the representatives chosen in the equivalence classes $[f]$ and $[g]$ and it makes sense to set

$$[f] + [g] = [f + g].$$

We can proceed in a similar fashion to define $\lambda[f]$ and $[f][g]$. Therefore, the germs at p form a ring.

Definition 7.15. Given a C^k-manifold M, the commutative ring of germs of C^k-functions at p is denoted $\mathcal{O}_{M,p}^{(k)}$. When $k = \infty$, we usually drop the superscript ∞.

Remark: Most readers will most likely be puzzled by the notation $\mathcal{O}_{M,p}^{(k)}$. In fact, it is standard in algebraic geometry, but it is not as commonly used in differential geometry. For any open subset U of a manifold M, the ring $\mathcal{C}^k(U)$ of C^k-functions on U is also denoted $\mathcal{O}_M^{(k)}(U)$ (certainly by people with an algebraic geometry bent!). Then it turns out that the map $U \mapsto \mathcal{O}_M^{(k)}(U)$ is a *sheaf*, denoted $\mathcal{O}_M^{(k)}$, and the ring $\mathcal{O}_{M,p}^{(k)}$ is the *stalk* of the sheaf $\mathcal{O}_M^{(k)}$ at p. Such rings are called *local rings*. Roughly speaking, all the "local" information about M at p is contained in the local ring $\mathcal{O}_{M,p}^{(k)}$. (This is to be taken with a grain of salt. In the C^k-case where $k < \infty$, we also need the "stationary germs," as we will see shortly.)

Now that we have a rigorous way of dealing with functions locally defined at p, observe that the map

$$v_i \colon f \mapsto \left(\frac{\partial}{\partial x_i}\right)_p f$$

yields the same value for all functions f in a germ $\mathbf{f}$ at p. Furthermore, the above map is linear on $\mathcal{O}_{M,p}^{(k)}$. More is true:

(1) For any two functions f, g locally defined at p, we have

$$\left(\frac{\partial}{\partial x_i}\right)_p (fg) = \left(\left(\frac{\partial}{\partial x_i}\right)_p f\right) g(p) + f(p) \left(\frac{\partial}{\partial x_i}\right)_p g.$$

(2) If $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, then

$$\left(\frac{\partial}{\partial x_i}\right)_p f = 0.$$

The first property says that v_i is a *point-derivation*; it is also known as the *Leibniz property*. As to the second property, when $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, we say that f *is stationary at* p.

It is easy to check (using the chain rule) that being stationary at p does not depend on the chart (U, φ) at p or on the function chosen in a germ $\mathbf{f}$. Therefore, the notion of a stationary germ makes sense.

Definition 7.16. We say that a germ $\mathbf{f}$ at $p \in M$ is a *stationary germ* iff $(f \circ \varphi^{-1})'(\varphi(p)) = 0$ for some chart (U, φ), at p and some function f in the germ $\mathbf{f}$. The C^k-stationary germs form a subring of $\mathcal{O}^{(k)}_{M,p}$ (but not an ideal) denoted $\mathcal{S}^{(k)}_{M,p}$.

Remarkably, it turns out that the set of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ is isomorphic to the tangent space $T_p(M)$. First we prove that this space has $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ as a basis.

Proposition 7.3. *Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n functions $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ defined on $\mathcal{O}^{(k)}_{M,p}$ by*

$$\left(\frac{\partial}{\partial x_i}\right)_p f = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} \qquad 1 \leq i \leq n,$$

are linear forms that vanish on $\mathcal{S}^{(k)}_{M,p}$. Every linear form L on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$ can be expressed in a unique way as

$$L = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p,$$

where $\lambda_i \in \mathbb{R}$. Therefore, the linear forms

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$$

form a basis of the vector space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$.

Proof. The first part of the proposition is trivial by definition of $\left(\frac{\partial}{\partial x_i}\right)_p f$, since for a stationary germ $\mathbf{f}$, we have $(f \circ \varphi^{-1})'(\varphi(p)) = 0$.

Next assume that L is a linear form on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$. For any function (U, f) locally defined at p, consider the function (U, g) locally defined at p given by

$$g(q) = f(q) - \sum_{i=1}^{n} (pr_i \circ \varphi)(q) \left(\frac{\partial}{\partial x_i}\right)_p f, \quad q \in U.$$

Observe that the germ of g is stationary at p. Indeed, if we let $X = \varphi(q)$, then $q = \varphi^{-1}(X)$, and we can write

$$\begin{aligned}(g \circ \varphi^{-1})(X) &= (f \circ \varphi^{-1})(X) - \sum_{i=1}^{n} pr_i(X) \left(\frac{\partial}{\partial x_i}\right)_p f \\ &= (f \circ \varphi^{-1})(X_1 \ldots, X_n) - \sum_{i=1}^{n} X_i \left(\frac{\partial}{\partial x_i}\right)_p f.\end{aligned}$$

By definition it follows that

$$\left.\frac{\partial(g \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} = \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} - \left(\frac{\partial}{\partial x_i}\right)_p f = 0.$$

But then as L vanishes on stationary germs, and the germ of

$$g = f - \sum_{i=1}^{n} (pr_i \circ \varphi) \left(\frac{\partial}{\partial x_i}\right)_p f$$

is stationary at p, we have $L(g) = 0$, so

$$L(f) = \sum_{i=1}^{n} L(pr_i \circ \varphi) \left(\frac{\partial}{\partial x_i}\right)_p f,$$

as desired. We still have to prove linear independence. If

$$\sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p = 0,$$

then if we apply this relation to the functions $x_i = pr_i \circ \varphi$, as

$$\left(\frac{\partial}{\partial x_i}\right)_p x_j = \delta_{ij},$$

we get $\lambda_i = 0$, for $i = 1, \ldots, n$. □

To define our third version of tangent vectors, we need to define point-derivations.

Definition 7.17. Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$, a *derivation at p in M* or *point-derivation on $\mathcal{O}^{(k)}_{M,p}$* is a linear form v on $\mathcal{O}^{(k)}_{M,p}$, such that

$$v(\mathbf{fg}) = v(\mathbf{f})\mathbf{g}(p) + \mathbf{f}(p)v(\mathbf{g}),$$

for all germs $\mathbf{f}, \mathbf{g} \in \mathcal{O}^{(k)}_{M,p}$. The above is called the *Leibniz property*. Let $\mathcal{D}^{(k)}_p(M)$ denote the set of point-derivations on $\mathcal{O}^{(k)}_{M,p}$.

As expected, point-derivations vanish on constant functions.

Proposition 7.4. *Every point-derivation v on $\mathcal{O}_{M,p}^{(k)}$ vanishes on germs of constant functions.*

Proof. If $\mathbf{g}$ is a germ of a constant function at p, then there is some $\lambda \in \mathbb{R}$ so that $g = \lambda$ (a constant function with value λ) for all $g \in \mathbf{g}$. Since v is linear,

$$v(\mathbf{g}) = v(\lambda \mathbf{1}) = \lambda v(\mathbf{1}),$$

where $\mathbf{1}$ is the germ of constant functions with value 1, so we just have to show that $v(\mathbf{1}) = 0$. However, because $\mathbf{1} = \mathbf{1} \cdot \mathbf{1}$ and v is a point-derivation, we get

$$\begin{aligned} v(\mathbf{1}) &= v(\mathbf{1} \cdot \mathbf{1}) \\ &= v(\mathbf{1})\mathbf{1}(p) + \mathbf{1}(p)v(\mathbf{1}) \\ &= v(\mathbf{1})1 + 1v(\mathbf{1}) = 2v(\mathbf{1}) \end{aligned}$$

from which we conclude that $v(\mathbf{1}) = 0$, as claimed. □

Recall that we observed earlier that the $\left(\frac{\partial}{\partial x_i}\right)_p$ are point-derivations at p. Therefore, we have

Proposition 7.5. *Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$, the linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$ are exactly the point-derivations on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$.*

Proof. By Proposition 7.3,

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$$

form a basis of the linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$. Since each $\left(\frac{\partial}{\partial x_i}\right)_p$ is a also a point-derivation at p, the result follows. □

Remark: Proposition 7.5 says that any linear form on $\mathcal{O}_{M,p}^{(k)}$ that vanishes on $\mathcal{S}_{M,p}^{(k)}$ belongs to $\mathcal{D}_p^{(k)}(M)$, the set of point-derivations on $\mathcal{O}_{M,p}^{(k)}$. However, in general, when $k \neq \infty$, a point-derivation on $\mathcal{O}_{M,p}^{(k)}$ does *not* necessarily vanish on $\mathcal{S}_{M,p}^{(k)}$. We will see in Proposition 7.9 that this is true for $k = \infty$.

Here is now our third definition of a tangent vector.

Definition 7.18. (Tangent Vectors, Version 3) Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$, a *tangent vector to M at p* is any point-derivation on $\mathcal{O}_{M,p}^{(k)}$ that vanishes on $\mathcal{S}_{M,p}^{(k)}$, the subspace of stationary germs.

Let us consider the simple case where $M = \mathbb{R}$. In this case, for every $x \in \mathbb{R}$, the tangent space $T_x(\mathbb{R})$ is a one-dimensional vector space isomorphic to $\mathbb{R}$ and $\left(\frac{\partial}{\partial t}\right)_x = \frac{d}{dt}\Big|_x$ is a basis vector of $T_x(\mathbb{R})$. For every C^k-function f locally defined at x, we have

$$\left(\frac{\partial}{\partial t}\right)_x f = \frac{df}{dt}\bigg|_x = f'(x).$$

Thus, $\left(\frac{\partial}{\partial t}\right)_x$ is: compute the derivative of a function at x.

We now prove the equivalence of Version 1 and Version 3 of a tangent vector.

Proposition 7.6. *Let M be any C^k-manifold of dimension n, with $k \geq 1$. For any $p \in M$, let u be any tangent vector (Version 1) given by some equivalence class of C^1-curves $\gamma\colon (-\epsilon, +\epsilon) \to M$ through p (i.e., $p = \gamma(0)$). Then the map L_u defined on $\mathcal{O}^{(k)}_{M,p}$ by*

$$L_u(\mathbf{f}) = (f \circ \gamma)'(0)$$

is a point-derivation that vanishes on $\mathcal{S}^{(k)}_{M,p}$. Furthermore, the map $u \mapsto L_u$ defined above is an isomorphism between $T_p(M)$ and the space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$.

Proof. (After L. Schwartz) Clearly, $L_u(\mathbf{f})$ does not depend on the representative f chosen in the germ $\mathbf{f}$. If γ and σ are equivalent curves defining u, then $(\varphi \circ \sigma)'(0) = (\varphi \circ \gamma)'(0)$, so from the chain rule we get

$$(f \circ \sigma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \sigma)'(0)) = (f \circ \varphi^{-1})'(\varphi(p))((\varphi \circ \gamma)'(0)) = (f \circ \gamma)'(0),$$

which shows that $L_u(\mathbf{f})$ does not depend on the curve γ defining u. If $\mathbf{f}$ is a stationary germ, then pick any chart (U, φ) at p, and let $\psi = \varphi \circ \gamma$. We have

$$L_u(\mathbf{f}) = (f \circ \gamma)'(0) = ((f \circ \varphi^{-1}) \circ (\varphi \circ \gamma))'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\psi'(0)) = 0,$$

since $(f \circ \varphi^{-1})'(\varphi(p)) = 0$, as $\mathbf{f}$ is a stationary germ. The definition of L_u makes it clear that L_u is a point-derivation at p. If $u \neq v$ are two distinct tangent vectors, then there exist some curves γ and σ through p so that

$$(\varphi \circ \gamma)'(0) \neq (\varphi \circ \sigma)'(0).$$

Thus, there is some i, with $1 \leq i \leq n$, so that if we let $f = pr_i \circ \varphi$, then

$$(f \circ \gamma)'(0) \neq (f \circ \sigma)'(0),$$

and so, $L_u \neq L_v$. This proves that the map $u \mapsto L_u$ is injective.

For surjectivity, recall that every linear map L on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$ can be uniquely expressed as

$$L = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p.$$

Define the curve γ on M through p by

$$\gamma(t) = \varphi^{-1}(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

for t in a small open interval containing 0. See Figure 7.11. Then we have

$$f(\gamma(t)) = (f \circ \varphi^{-1})(\varphi(p) + t(\lambda_1, \ldots, \lambda_n)),$$

and by the chain rule we get

$$(f \circ \gamma)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(\lambda_1, \ldots, \lambda_n) = \sum_{i=1}^{n} \lambda_i \, \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} = L(\mathbf{f}).$$

This proves that $T_p(M)$ is isomorphic to the space of linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$. □

We show in the next section that the the space of linear forms on $\mathcal{O}_{M,p}^{(k)}$ that vanish on $\mathcal{S}_{M,p}^{(k)}$ is isomorphic to $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^*$ (the dual of the quotient space $\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$).

Even though this is just a restatement of Proposition 7.3, we state the following proposition because of its practical usefulness:

Proposition 7.7. *Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n tangent vectors*

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$$

form a basis of T_pM.

When M is a smooth manifold, things get a little simpler. Indeed, it turns out that in this case, every point-derivation vanishes on stationary germs. To prove this, we recall the following result from calculus (see Warner [114]):

Proposition 7.8. *If $g \colon \mathbb{R}^n \to \mathbb{R}$ is a C^k-function ($k \geq 2$) on a convex open U about $p \in \mathbb{R}^n$, then for every $q \in U$, we have*

$$g(q) = g(p) + \sum_{i=1}^{n} \left.\frac{\partial g}{\partial X_i}\right|_p (q_i - p_i) + \sum_{i,j=1}^{n} (q_i - p_i)(q_j - p_j) \int_0^1 (1-t) \left.\frac{\partial^2 g}{\partial X_i \partial X_j}\right|_{(1-t)p+tq} dt.$$

In particular, if $g \in C^\infty(U)$, then the integral as a function of q is C^∞.

Proposition 7.9. *Let M be any C^∞-manifold of dimension n. For any $p \in M$, any point-derivation on $\mathcal{O}_{M,p}^{(\infty)}$ vanishes on $\mathcal{S}_{M,p}^{(\infty)}$, the ring of stationary germs. Consequently, $T_p(M) = \mathcal{D}_p^{(\infty)}(M)$.*

Proof. Pick some chart (U, φ) at p, where $\varphi(U)$ is convex (for instance, an open ball) and let $\mathbf{f}$ be any stationary germ. If we apply Proposition 7.8 to $f \circ \varphi^{-1}$ (for any $f \in \mathbf{f}$) and then compose $f \circ \varphi^{-1}$ with φ, we get

$$f(q) = f(p) + \sum_{i=1}^{n} \left.\frac{\partial(f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)} (x_i(q) - x_i(p)) + \sum_{i,j=1}^{n} (x_i(q) - x_i(p))(x_j(q) - x_j(p))h,$$

near p, where h is C^∞ and $x_i = pr_i \circ \varphi$. Since $\mathbf{f}$ is a stationary germ, this yields

$$f(q) = f(p) + \sum_{i,j=1}^{n} (x_i(q) - x_i(p))(x_j(q) - x_j(p))h.$$

If v is any point-derivation, since $f(p)$ is constant, Proposition 7.4 implies $v(f(p)) = 0$, and we get

$$\begin{aligned} v(f) = v(f(p)) + \sum_{i,j=1}^{n} \Big[&(x_i(q) - x_i(p))(p)(x_j(q) - x_j(p))(p)v(h) \\ &+ (x_i(q) - x_i(p))(p)v(x_j(q) - x_j(p))h(p) + v(x_i(q) - x_i(p))(x_j(q) - x_j(p))(p)h(p)\Big] = 0, \end{aligned}$$

where the three terms in the summand vanish since

$$(x_i(q) - x_i(p))(p) = x_i(p) - x_i(p) = 0 = x_j(p) - x_j(p) = (x_j(q) - x_j(p))(p).$$

We conclude that v vanishes on stationary germs. □

Proposition 7.9 shows that in the case of a smooth manifold, in Definition 7.18, we can omit the requirement that point-derivations vanish on stationary germs, since this is automatic.

Remark: In the case of smooth manifolds ($k = \infty$) some authors, including Morita [87] (Chapter 1, Definition 1.32) and O'Neil [91] (Chapter 1, Definition 9), define derivations as linear derivations with domain $\mathcal{C}^\infty(M)$, the set of all smooth funtions on the entire manifold, M. This definition is simpler in the sense that it does not require the definition of the notion of germ but it is not local, because it is not obvious that if v is a point-derivation at p, then $v(f) = v(g)$ whenever $f, g \in \mathcal{C}^\infty(M)$ agree locally at p. In fact, if two smooth locally defined functions agree near p it may not be possible to extend both of them to the whole of M. However, it can be proved that this property is local because on smooth manifolds, "bump functions" exist (see Section 10.1, Proposition 10.2). Unfortunately, this argument breaks down for C^k-manifolds with $k < \infty$ and in this case the ring of germs at p can't be avoided.

7.4 Tangent and Cotangent Spaces Revisited ⊛

The space of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ turns out to be isomorphic to the dual of the quotient space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, and this fact shows that the dual $(T_pM)^*$ of the tangent space T_pM, called the *cotangent space* to M at p, can be viewed as the quotient space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. This provides a fairly intrinsic definition of the cotangent space to M at p. For notational simplicity, we write T^*_pM instead of $(T_pM)^*$. This section is quite technical and can be safely skipped upon first (or second!) reading.

Let us refresh the reader's memory and review quotient vector spaces. If E is a vector space, the set of all linear forms $f\colon E\to\mathbb{R}$ on E is a vector space called the *dual* of E and denoted by E^*. If $H\subseteq E$ is any subspace of E, we define the equivalence relation $\sim$ so that for all $u,v\in E$,

$$u\sim v \quad \text{iff} \quad u-v\in H.$$

Every equivalence class $[u]$, is equal to the subset $u+H=\{u+h\mid h\in H\}$, called a *coset*, and the set of equivalence classes E/H modulo $\sim$ is a vector space under the operations

$$\begin{aligned}[u]+[v]&=[u+v]\\ \lambda[u]&=[\lambda u].\end{aligned}$$

The space E/H is called the *quotient of E by H* or for short, a *quotient space*.

Denote by $\mathcal{L}(E/H)$ the set of linear forms $f\colon E\to\mathbb{R}$ that vanish on H (this means that for every $f\in\mathcal{L}(E/H)$, we have $f(h)=0$ for all $h\in H$). The following proposition plays a crucial role.

Proposition 7.10. *Given a vector space E and a subspace H of E, there is an isomorphism*

$$\mathcal{L}(E/H)\cong(E/H)^*$$

between the set $\mathcal{L}(E/H)$ of linear forms $f\colon E\to\mathbb{R}$ that vanish on H and the dual of the quotient space E/H.

Proof. To see this, define the map $f\mapsto\widehat{f}$ from $\mathcal{L}(E/H)$ to $(E/H)^*$ as follows: for any $f\in\mathcal{L}(E/H)$,

$$\widehat{f}([u])=f(u),\quad [u]\in E/H.$$

This function is well-defined because it does not depend on the representative u, chosen in the equivalence class $[u]$. Indeed, if $v\sim u$, then $v=u+h$ some $h\in H$ and so

$$f(v)=f(u+h)=f(u)+f(h)=f(u),$$

since $f(h)=0$ for all $h\in H$. The formula $\widehat{f}([u])=f(u)$ makes it obvious that $\widehat{f}$ is linear since f is linear. The mapping $f\mapsto\widehat{f}$ is injective. This is because if $\widehat{f_1}=\widehat{f_2}$, then

$$\widehat{f_1}([u])=\widehat{f_2}([u])$$

for all $u \in E$, and because $\widehat{f_1}([u]) = f_1(u)$ and $\widehat{f_2}([u]) = f_2(u)$, we get $f_1(u) = f_2(u)$ for all $u \in E$, that is, $f_1 = f_2$. The mapping $f \mapsto \widehat{f}$ is surjective because given any linear form $\varphi \in (E/H)^*$, if we define f by

$$f(u) = \varphi([u])$$

for all $u \in E$, then f is linear, vanishes on H and clearly, $\widehat{f} = \varphi$. Therefore, we have the isomorphism,

$$\mathcal{L}(E/H) \cong (E/H)^*,$$

as claimed. $\square$

As a consequence of Proposition 7.10 the subspace of linear forms on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$ is isomorphic to the dual $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ of the space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, we see that the linear forms

$$\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$$

also form a basis of $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$.

There is a conceptually clearer way to define a canonical isomorphism between $T_p(M)$ and the dual of $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ in terms of a nondegenerate pairing between $T_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. This pairing is described by Serre in [105] (Chapter III, Section 8) for analytic manifolds and can be adapted to our situation.

Define the map $\omega\colon T_p(M) \times \mathcal{O}^{(k)}_{M,p} \to \mathbb{R}$, so that

$$\omega([\gamma], \mathbf{f}) = (f \circ \gamma)'(0),$$

for all $[\gamma] \in T_p(M)$ and all $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$ (with $f \in \mathbf{f}$). It is easy to check that the above expression does not depend on the representatives chosen in the equivalences classes $[\gamma]$, and $\mathbf{f}$ and that ω is bilinear. However, as defined, ω is degenerate because $\omega([\gamma], \mathbf{f}) = 0$ if $\mathbf{f}$ is a stationary germ. Thus, we are led to consider the pairing with domain $T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})$ given by

$$\omega([\gamma], [\mathbf{f}]) = (f \circ \gamma)'(0),$$

where $[\gamma] \in T_p(M)$ and $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, which we also denote $\omega\colon T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}) \to \mathbb{R}$. Then the following result holds.

Proposition 7.11. *The map $\omega\colon T_p(M) \times (\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}) \to \mathbb{R}$ defined so that*

$$\omega([\gamma], [\mathbf{f}]) = (f \circ \gamma)'(0),$$

for all $[\gamma] \in T_p(M)$ and all $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, is a nondegenerate pairing (with $f \in \mathbf{f}$). Consequently, there is a canonical isomorphism between $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and a canonical isomorphism between $T^*_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$.

Proof. This is basically a replay of the proof of Proposition 7.6. First assume that given some $[\gamma] \in T_p(M)$, we have $\omega([\gamma], [\mathbf{f}]) = 0$ for all $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. Pick a chart (U, φ), with $p \in U$ and let $x_i = pr_i \circ \varphi$. Then, the $\mathbf{x_i}$'s are not stationary germs, since $x_i \circ \varphi^{-1} = pr_i \circ \varphi \circ \varphi^{-1} = pr_i$ and $(pr_i)'(0) = pr_i$ (because pr_i is a linear form). By hypothesis, $\omega([\gamma], [\mathbf{x_i}]) = 0$ for $i = 1, \ldots, n$, which means that

$$(x_i \circ \gamma)'(0) = (pr_i \circ \varphi \circ \gamma)'(0) = 0$$

for $i = 1, \ldots, n$, namely, $pr_i((\varphi \circ \gamma)'(0)) = 0$ for $i = 1, \ldots, n$; that is,

$$(\varphi \circ \gamma)'(0) = 0_n,$$

proving that $[\gamma] = 0$.

Next assume that given some $[\mathbf{f}] \in \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$, we have $\omega([\gamma], [\mathbf{f}]) = 0$ for all $[\gamma] \in T_p(M)$. Again pick a chart (U, φ). For every $z \in \mathbb{R}^n$, we have the curve γ_z given by

$$\gamma_z(t) = \varphi^{-1}(\varphi(p) + tz)$$

for all t in a small open interval containing 0. See Figure 7.11. By hypothesis,

$$\omega([\gamma_z], [\mathbf{f}]) = (f \circ \gamma_z)'(0) = (f \circ \varphi^{-1})'(\varphi(p))(z) = 0$$

for all $z \in \mathbb{R}^n$, which means that

$$(f \circ \varphi^{-1})'(\varphi(p)) = 0.$$

But then, $\mathbf{f}$ is a stationary germ and so, $[\mathbf{f}] = 0$. Therefore, we proved that ω is a nondegenerate pairing. Since $T_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ have finite dimension n, it follows that there is are canonical isomorphisms between $T_p(M)$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and between $T^*_p(M)$ and $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. □

In view of Proposition 7.11, we can identify $T_p(M)$ with $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $T^*_p(M)$ with $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$.

Remark: Also recall that if E is a finite dimensional space, the map $\mathrm{eval}_E \colon E \to E^{**}$ defined so that,

$$\mathrm{eval}_E(v)(f) = f(v), \quad \text{for all } v \in E \text{ and for all } f \in E^*,$$

is a linear isomorphism.

Observe that we can view $\omega(u, \mathbf{f}) = \omega([\gamma], [\mathbf{f}])$ as the result of computing the directional derivative of the locally defined function $f \in \mathbf{f}$ in the direction u (given by a curve γ). Proposition 7.11 suggests the following definition:

Definition 7.19. (Tangent and Cotangent Spaces, Version 4) Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$, the *tangent space at* p denoted $T_p(M)$ is the space of point-derivations on $\mathcal{O}^{(k)}_{M,p}$ that vanish on $\mathcal{S}^{(k)}_{M,p}$. Thus, $T_p(M)$ can be identified with $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$, the dual of the quotient space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$. The space $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ is called the *cotangent space at* p; it is isomorphic to the dual $T^*_p(M)$, of $T_p(M)$. (For simplicity of notation we also denote $T_p(M)$ by T_pM and $T^*_p(M)$ by T^*_pM.)

We can consider any C^k-function f on some open subset U of M as a representative of the germ $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$, so the image of f in $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ under the canonical projection of $\mathcal{O}^{(k)}_{M,p}$ onto $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ makes sense. Observe that if $x_i = pr_i \circ \varphi$, as

$$\left(\frac{\partial}{\partial x_i}\right)_p x_j = \delta_{i,j},$$

the images of $x_1, \ldots, x_n$ in $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ form the dual basis of the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$.

Definition 7.20. Given any C^k-function f on U, we denote the image of f in $T^*_p(M) = \mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ by df_p. This is the *differential of* f *at* p.

Using the isomorphism between $\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p}$ and $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^{**}$ described above, df_p corresponds to the linear map in $T^*_p(M)$ defined by

$$df_p(v) = v(\mathbf{f}),$$

for all $v \in T_p(M)$. With this notation, we see that $(dx_1)_p, \ldots, (dx_n)_p$ is a basis of $T^*_p(M)$, and this basis is dual to the basis $\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p$ of $T_p(M)$. For simplicity of notation, we often omit the subscript p unless confusion arises.

Remark: Strictly speaking, a tangent vector $v \in T_p(M)$ is defined on the space of germs $\mathcal{O}^{(k)}_{M,p}$, at p. However, it is often convenient to define v on C^k-functions $f \in \mathcal{C}^k(U)$, where U is some open subset containing p. This is easy: set

$$v(f) = v(\mathbf{f}).$$

Given any chart (U, φ) at p, since v can be written in a unique way as

$$v = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p,$$

we get

$$v(f) = \sum_{i=1}^{n} \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p f.$$

This shows that $v(f)$ is the *directional derivative of f in the direction v*. The directional derivative, $v(f)$, is also denoted $v[f]$.

It is also possible to define $T_p(M)$ just in terms of $\mathcal{O}^{(\infty)}_{M,p}$, and we get a fifth definition of T_pM.

Definition 7.21. Let $\mathfrak{m}_{M,p} \subseteq \mathcal{O}^{(\infty)}_{M,p}$ be the ideal of germs that vanish at p. We also have the ideal $\mathfrak{m}^2_{M,p}$, which consists of all finite sums of products of two elements in $\mathfrak{m}_{M,p}$.

It turns out that $T^*_p(M)$ is isomorphic to $\mathfrak{m}_{M,p}/\mathfrak{m}^2_{M,p}$ (see Warner [114], Lemma 1.16).

Definition 7.22. Let $\mathfrak{m}^{(k)}_{M,p} \subseteq \mathcal{O}^{(k)}_{M,p}$ denote the ideal of C^k-germs that vanish at p and $\mathfrak{s}^{(k)}_{M,p} \subseteq \mathcal{S}^{(k)}_{M,p}$ denote the ideal of stationary C^k-germs that vanish at p.

Adapting Warner's argument, we can prove the following proposition:

Proposition 7.12. *We have the inclusion, $(\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}$ and the isomorphism*

$$(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^* \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*.$$

As a consequence, $T_p(M) \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ and $T^*_p(M) \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$.

Proof. Given any two germs, $\mathbf{f}, \mathbf{g} \in \mathfrak{m}^{(k)}_{M,p}$, for any two locally defined functions, $f \in \mathbf{f}$ and $g \in \mathbf{g}$, since $f(p) = g(p) = 0$, for any chart (U, φ) with $p \in U$, by definition of the product fg of two functions, for any $q \in M$ near p, we have

$$\begin{aligned}(fg \circ \varphi^{-1})(q) &= (fg)(\varphi^{-1}(q))\\ &= f(\varphi^{-1}(q))g(\varphi^{-1}(q))\\ &= (f \circ \varphi^{-1})(q)(g \circ \varphi^{-1})(q),\end{aligned}$$

so

$$fg \circ \varphi^{-1} = (f \circ \varphi^{-1})(g \circ \varphi^{-1}),$$

and by the product rule for derivatives, we get

$$(fg \circ \varphi^{-1})'(0) = (f \circ \varphi^{-1})'(0)(g \circ \varphi^{-1})(0) + (f \circ \varphi^{-1})(0)(g \circ \varphi^{-1})'(0) = 0,$$

because $(g \circ \varphi^{-1})(0) = g(\varphi^{-1}(0)) = g(p) = 0$ and $(f \circ \varphi^{-1})(0) = f(\varphi^{-1}(0)) = f(p) = 0$. Therefore, fg is stationary at p and since $fg(p) = 0$, we have $\mathbf{fg} \in \mathfrak{s}^{(k)}_{M,p}$, which implies the inclusion $(\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}$.

Now the key point is that any constant germ is stationary, since the derivative of a constant function is zero. Consequently, if v is a linear form on $\mathcal{O}^{(k)}_{M,p}$ vanishing on $\mathcal{S}^{(k)}_{M,p}$, then

$$v(\mathbf{f}) = v(\mathbf{f} - \mathbf{f}(\mathbf{p})),$$

for all $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$, where $\mathbf{f}(\mathbf{p})$ denotes the germ of constant functions with value $\mathbf{f}(p)$. We use this fact to define two functions between $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ which are mutual inverses.

The map from $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ to $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ is restriction to $\mathfrak{m}^{(k)}_{M,p}$: every linear form v on $\mathcal{O}^{(k)}_{M,p}$ vanishing on $\mathcal{S}^{(k)}_{M,p}$ yields a linear form on $\mathfrak{m}^{(k)}_{M,p}$ that vanishes on $\mathfrak{s}^{(k)}_{M,p}$.

Conversely, for any linear form ℓ on $\mathfrak{m}^{(k)}_{M,p}$ vanishing on $\mathfrak{s}^{(k)}_{M,p}$, define the function v_ℓ so that

$$v_\ell(\mathbf{f}) = \ell(\mathbf{f} - \mathbf{f}(\mathbf{p})),$$

for any germ $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$. Since ℓ is linear, it is clear that v_ℓ is also linear. If f is stationary at p, then $f - f(p)$ is also stationary at p because the derivative of a constant is zero. Obviously, $f - f(p)$ vanishes at p. It follows that v_ℓ vanishes on stationary germs at p.

Using the fact that $v(\mathbf{f}) = v(\mathbf{f} - \mathbf{f}(\mathbf{p}))$, it is easy to check that the above maps between $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ and $(\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ are mutual inverses, establishing the desired isomorphism. Because $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ is finite-dimensional, we also have the isomorphism

$$\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p} \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$$

which yields the isomorphims $T_p(M) \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$ and $T^*_p(M) \cong \mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p}$. □

When $k = \infty$, Proposition 7.8 shows that every stationary germ that vanishes at p belongs to $\mathfrak{m}^2_{M,p}$. Therefore, when $k = \infty$, we have $\mathfrak{s}^{(\infty)}_{M,p} = \mathfrak{m}^2_{M,p}$ and so, we obtain the result quoted above (from Warner):

$$T^*_p(M) = \mathcal{O}^{(\infty)}_{M,p}/\mathcal{S}^{(\infty)}_{M,p} \cong \mathfrak{m}_{M,p}/\mathfrak{m}^2_{M,p}.$$

Remarks:

(1) The isomorphism

$$(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^* \cong (\mathfrak{m}^{(k)}_{M,p}/\mathfrak{s}^{(k)}_{M,p})^*$$

yields another proof that the linear forms in $(\mathcal{O}^{(k)}_{M,p}/\mathcal{S}^{(k)}_{M,p})^*$ are point-derivations, using the argument from Warner [114] (Lemma 1.16). It is enough to prove that every linear form of the form v_ℓ is a point-derivation. Indeed, if ℓ is a linear form on $\mathfrak{m}^{(k)}_{M,p}$ vanishing on $\mathfrak{s}^{(k)}_{M,p}$, we have

$$\begin{aligned} v_\ell(\mathbf{fg}) &= \ell(\mathbf{fg} - \mathbf{f}(\mathbf{p})\mathbf{g}(\mathbf{p})) \\ &= \ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p})) + (\mathbf{f} - \mathbf{f}(\mathbf{p}))\mathbf{g}(\mathbf{p}) + \mathbf{f}(\mathbf{p})(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) \\ &= \ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) + \ell(\mathbf{f} - \mathbf{f}(\mathbf{p}))\mathbf{g}(p) + \mathbf{f}(p)\ell(\mathbf{g} - \mathbf{g}(\mathbf{p})) \\ &= v_\ell(\mathbf{f})\mathbf{g}(p) + \mathbf{f}(p)v_\ell(\mathbf{g}), \end{aligned}$$

using the fact that $\ell\big((\mathbf{f} - \mathbf{f}(\mathbf{p}))(\mathbf{g} - \mathbf{g}(\mathbf{p}))\big) = 0$ since $(\mathfrak{m}^{(k)}_{M,p})^2 \subseteq \mathfrak{s}^{(k)}_{M,p}$ and ℓ vanishes on $\mathfrak{s}^{(k)}_{M,p}$, which proves that v_ℓ is a point-derivation.

(2) The ideal $\mathfrak{m}^{(k)}_{M,p}$ is in fact the unique maximal ideal of $\mathcal{O}^{(k)}_{M,p}$. This is because if $\mathbf{f} \in \mathcal{O}^{(k)}_{M,p}$ does not vanish at p, then $\mathbf{1/f}$ belongs to $\mathcal{O}^{(k)}_{M,p}$ (because if f does not vanish at p, then by continuity, f does not vanish in some open subset containing p, for all $f \in \mathbf{f}$), and any proper ideal containing $\mathfrak{m}^{(k)}_{M,p}$ and $\mathbf{f}$ would be equal to $\mathcal{O}^{(k)}_{M,p}$, which is absurd. Thus, $\mathcal{O}^{(k)}_{M,p}$ is a local ring (in the sense of commutative algebra) called the *local ring of germs of C^k-functions at p*. These rings play a crucial role in algebraic geometry.

(3) Using the map $\mathbf{f} \mapsto \mathbf{f} - \mathbf{f}(\mathbf{p})$, it is easy to see that

$$\mathcal{O}^{(k)}_{M,p} \cong \mathbb{R} \oplus \mathfrak{m}^{(k)}_{M,p} \quad \text{and} \quad \mathcal{S}^{(k)}_{M,p} \cong \mathbb{R} \oplus \mathfrak{s}^{(k)}_{M,p}.$$

7.5 Tangent Maps

After having explored thoroughly the notion of tangent vector, we show how a C^k-map $h\colon M \to N$, between C^k manifolds, induces a linear map $dh_p\colon T_p(M) \to T_{h(p)}(N)$, for every $p \in M$. We find it convenient to use Version 3 of the definition of a tangent vector. Let $u \in T_p(M)$ be a point-derivation on $\mathcal{O}^{(k)}_{M,p}$ that vanishes on $\mathcal{S}^{(k)}_{M,p}$. We would like $dh_p(u)$ to be a point-derivation on $\mathcal{O}^{(k)}_{N,h(p)}$ that vanishes on $\mathcal{S}^{(k)}_{N,h(p)}$. For every germ $\mathbf{g} \in \mathcal{O}^{(k)}_{N,h(p)}$, if $g \in \mathbf{g}$ is any locally defined function at $h(p)$, it is clear that $g \circ h$ is locally defined at p and is C^k, and that if $g_1, g_2 \in \mathbf{g}$ then $g_1 \circ h$ and $g_2 \circ h$ are equivalent. The germ of all locally defined functions at p of the form $g \circ h$, with $g \in \mathbf{g}$, will be denoted $\mathbf{g} \circ h$. We set

$$dh_p(u)(\mathbf{g}) = u(\mathbf{g} \circ h).$$

In any chart (U, φ) at p, if $u = \sum_{i=1}^n \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p$, then

$$dh_p(u)(\mathbf{g}) = \sum_{i=1}^n \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p g \circ h$$

for any $g \in \mathbf{g}$. Moreover, if $\mathbf{g}$ is a stationary germ at $h(p)$, then for some chart (V, ψ) on N at $q = h(p)$, we have $(g \circ \psi^{-1})'(\psi(q)) = 0$ and, for any chart (U, φ) at p on M, we use the chain rule to obtain

$$(g \circ h \circ \varphi^{-1})'(\varphi(p)) = (g \circ \psi^{-1})'(\psi(q))((\psi \circ h \circ \varphi^{-1})'(\varphi(p))) = 0,$$

which means that $\mathbf{g} \circ h$ is stationary at p. Therefore, $dh_p(u) \in T_{h(p)}(N)$. It is also clear that dh_p is a linear map. We summarize all this in the following definition.

Definition 7.23. (Using Version 3 of a tangent vector) Given any two C^k-manifolds M and N, of dimension m and n respectively, for any C^k-map $h\colon M \to N$ and for every $p \in M$, the *differential of h at p* or *tangent map* $dh_p\colon T_p(M) \to T_{h(p)}(N)$ (also denoted $T_p h\colon T_p(M) \to T_{h(p)}(N)$), is the linear map defined so that

$$dh_p(u)(\mathbf{g}) = T_p h(u)(\mathbf{g}) = u(\mathbf{g} \circ h)$$

for every $u \in T_p(M)$ and every germ $\mathbf{g} \in \mathcal{O}^{(k)}_{N,h(p)}$. The linear map dh_p $(= T_ph)$ is sometimes denoted h'_p or D_ph. See Figure 7.12.

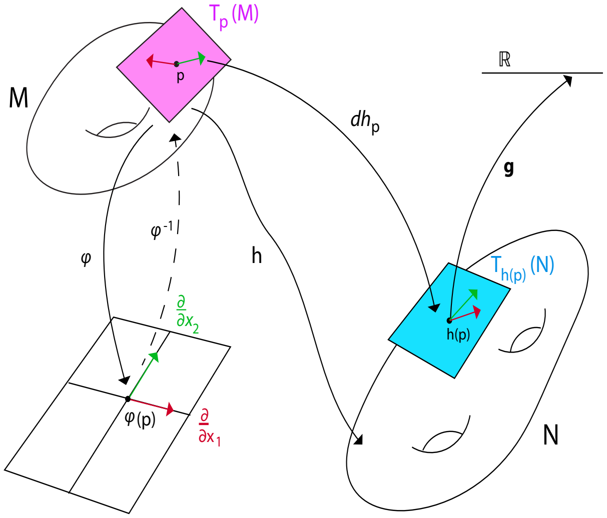


Figure 7.12: The tangent map $dh_p(u)(\mathbf{g}) = \sum_{i=1}^n \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p g \circ h$.

The chain rule is easily generalized to manifolds.

Proposition 7.13. *Given any two C^k-maps $f\colon M \to N$ and $g\colon N \to P$ between smooth C^k-manifolds, for any $p \in M$, we have*

$$d(g \circ f)_p = dg_{f(p)} \circ df_p.$$

In the special case where $N = \mathbb{R}$, a C^k-map between the manifolds M and $\mathbb{R}$ is just a C^k-function on M. It is interesting to see what T_pf is explicitly. Since $N = \mathbb{R}$, germs (of functions on $\mathbb{R}$) at $t_0 = f(p)$ are just germs of C^k-functions $g\colon \mathbb{R} \to \mathbb{R}$ locally defined at t_0. Then for any $u \in T_p(M)$ and every germ $\mathbf{g}$ at t_0,

$$T_pf(u)(\mathbf{g}) = u(\mathbf{g} \circ f).$$

If we pick a chart (U, φ) on M at p, we know that the $\left(\frac{\partial}{\partial x_i}\right)_p$ form a basis of $T_p(M)$, with $1 \leq i \leq n$. Therefore, it is enough to figure out what $T_pf(u)(\mathbf{g})$ is when $u = \left(\frac{\partial}{\partial x_i}\right)_p$. In this case,

$$T_pf\left(\left(\frac{\partial}{\partial x_i}\right)_p\right)(\mathbf{g}) = \left(\frac{\partial}{\partial x_i}\right)_p g \circ f = \left.\frac{\partial(g \circ f \circ \varphi^{-1})}{\partial X_i}\right|_{\varphi(p)}.$$

Using the chain rule, we find that

$$T_pf\left(\left(\frac{\partial}{\partial x_i}\right)_p\right)(\mathbf{g}) = \left(\frac{\partial}{\partial x_i}\right)_p f\,\frac{dg}{dt}\bigg|_{t_0}.$$

Therefore, we have

$$T_pf(u) = u(\mathbf{f})\,\frac{d}{dt}\bigg|_{t_0}.$$

This shows that we can identify T_pf with the linear form in $T_p^*(M)$ defined by

$$df_p(u) = u(\mathbf{f}), \qquad u \in T_pM,$$

by identifying $T_{t_0}\mathbb{R}$ with $\mathbb{R}$. This is consistent with our previous definition of df_p as the image of f in $T_p^*(M) = \mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)}$ (as $T_p(M)$ is isomorphic to $(\mathcal{O}_{M,p}^{(k)}/\mathcal{S}_{M,p}^{(k)})^*$).

Again, even though this is just a restatement of facts we already showed, we state the following proposition because of its practical usefulness.

Proposition 7.14. *Given any C^k-manifold M of dimension n, with $k \geq 1$, for any $p \in M$ and any chart (U, φ) at p, the n linear maps*

$$(dx_1)_p, \ldots, (dx_n)_p$$

form a basis of T_p^*M, where $(dx_i)_p$, the differential of x_i at p, is identified with the linear form in T_p^*M such that $(dx_i)_p(v) = v(\mathbf{x_i})$, for every $v \in T_pM$ (by identifying $T_\lambda\mathbb{R}$ with $\mathbb{R}$).

In preparation for the definition of the flow of a vector field (which will be needed to define the exponential map in Lie group theory), we need to define the tangent vector to a curve on a manifold. Given a C^k-curve $\gamma\colon (a,b) \to M$ on a C^k-manifold M, for any $t_0 \in (a,b)$, we would like to define the tangent vector to the curve γ at t_0 as a tangent vector to M at $p = \gamma(t_0)$. We do this as follows: Recall that $\frac{d}{dt}\big|_{t_0}$ is a basis vector of $T_{t_0}(\mathbb{R}) = \mathbb{R}$.

Definition 7.24. The *tangent vector to the curve γ at t_0*, denoted $\dot{\gamma}(t_0)$ (or $\gamma'(t_0)$, or $\frac{d\gamma}{dt}(t_0)$), is given by

$$\dot{\gamma}(t_0) = d\gamma_{t_0}\left(\frac{d}{dt}\bigg|_{t_0}\right) = \left(\frac{\partial}{\partial t}\right)_{t_0}\gamma.$$

We find it necessary to define curves (in a manifold) whose domain is not an open interval. A map $\gamma\colon [a,b] \to M$, is a *C^k-curve in M* if it is the restriction of some C^k-curve $\widetilde{\gamma}\colon (a-\epsilon, b+\epsilon) \to M$, for some (small) $\epsilon > 0$. Note that for such a curve (if $k \geq 1$) the tangent vector $\dot{\gamma}(t)$ is defined for all $t \in [a,b]$. A continuous curve $\gamma\colon [a,b] \to M$ is *piecewise* C^k iff there a sequence $a_0 = a, a_1, \ldots, a_m = b$, so that the restriction γ_i of γ to each $[a_i, a_{i+1}]$ is a C^k-curve, for $i = 0, \ldots, m-1$. This implies that $\gamma_i'(a_{i+1})$ and $\gamma_{i+1}'(a_{i+1})$ are defined for

$i = 0, \ldots, m-1$, but there may be a jump in the tangent vector to γ at a_{i+1}, that is, we may have $\gamma_i'(a_{i+1}) \neq \gamma_{i+1}'(a_{i+1})$.

Sometimes, especially in the case of a linear Lie group, it is more convenient to define the tangent map in terms of Version 1 of a tangent vector. Given any C^k-map $h\colon M \to N$, for every $p \in M$, for any two curves γ_1 and γ_2 such that $\gamma_1(0) = \gamma_2(0) = p$, if γ_1 and γ_2 are equivalent, then for any chart $\varphi\colon U \to \Omega_1$ at p, we have $(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0)$, and since f is C^k, for some (in fact, any) chart $\psi\colon V \to \Omega_2$ at $q = h(p)$, the map $\psi \circ h \circ \varphi^{-1}$ is C^k, so

$$\begin{aligned}(\psi \circ h \circ \gamma_1)'(0) &= (\psi \circ h \circ \varphi^{-1})'_{\varphi(p)}((\varphi \circ \gamma_1)'(0))\\ &= (\psi \circ h \circ \varphi^{-1})'_{\varphi(p)}((\varphi \circ \gamma_2)'(0))\\ &= (\psi \circ h \circ \gamma_2)'(0),\end{aligned}$$

which shows that $h \circ \gamma_1$ and $h \circ \gamma_2$ are equivalent. As a consequence, for every equivalence class $u = [\gamma]$ of curves through p in M, all curves of the form $h \circ \gamma$ (with $\gamma \in u$) through $h(p)$ in N belong to the same equivalence class, and can make the following definition.

Definition 7.25. (Using Version 1 of a tangent vector) Given any two C^k-manifolds M and N, of dimension m and n respectively, for any C^k-map $h\colon M \to N$ and for every $p \in M$, the *differential of h at p* or *tangent map dh_p*: $T_p(M) \to T_{h(p)}(N)$ (also denoted $T_ph\colon T_p(M) \to T_{h(p)}(N)$), is the linear map defined such that for every equivalence class $u = [\gamma]$ of curves γ in M with $\gamma(0) = p$,

$$dh_p(u) = T_ph(u) = v,$$

where v is the equivalence class of all curves through $h(p)$ in N of the form $h \circ \gamma$, with $\gamma \in u$. See Figure 7.13.

If M is a manifold in $\mathbb{R}^{N_1}$ and N is a manifold in $\mathbb{R}^{N_2}$ (for some $N_1, N_2 \geq 1$), then $\gamma'(0) \in \mathbb{R}^{N_1}$ and $(h \circ \gamma)'(0) \in \mathbb{R}^{N_2}$, so in this case the definition of $dh_p = T_ph$ is just Definition 3.9; namely, for any curve γ in M such that $\gamma(0) = p$ and $\gamma'(0) = u$,

$$dh_p(u) = T_ph(u) = (h \circ \gamma)'(0).$$

Example 7.6. For example, consider the linear Lie group $\mathbf{SO}(3)$, pick any vector $u \in \mathbb{R}^3$, and let $f\colon \mathbf{SO}(3) \to \mathbb{R}^3$ be given by

$$f(R) = Ru, \quad R \in \mathbf{SO}(3).$$

To compute $df_R\colon T_R\mathbf{SO}(3) \to T_{Ru}\mathbb{R}^3$, since $T_R\mathbf{SO}(3) = R\mathfrak{so}(3)$ and $T_{Ru}\mathbb{R}^3 = \mathbb{R}^3$, pick any tangent vector $RB \in R\mathfrak{so}(3) = T_R\mathbf{SO}(3)$ (where B is any 3×3 skew symmetric matrix), let $\gamma(t) = Re^{tB}$ be the curve through R such that $\gamma'(0) = RB$, and compute

$$df_R(RB) = (f(\gamma(t)))'(0) = (Re^{tB}u)'(0) = RBu.$$

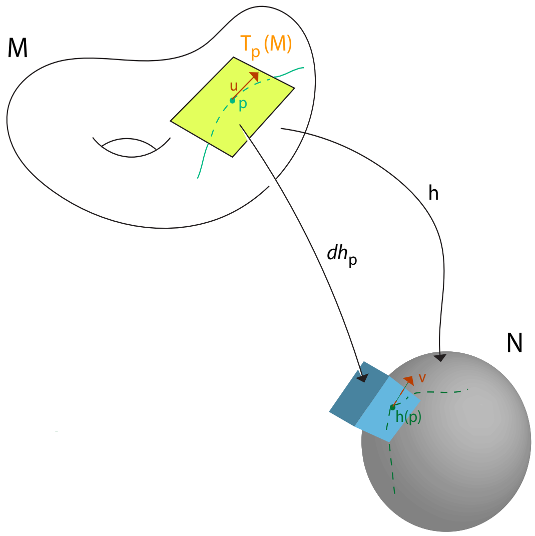


Figure 7.13: The tangent map $dh_p(u) = v$ defined via equivalent curves.

Therefore, we see that

$$df_R(X) = Xu, \quad X \in T_R\mathbf{SO}(3) = R\mathfrak{so}(3).$$

If we express the skew symmetric matrix $B \in \mathfrak{so}(3)$ as $B = \omega_\times$ for some vector $\omega \in \mathbb{R}^3$, then we have

$$df_R(R\omega_\times) = R\omega_\times u = R(\omega \times u).$$

Using the isomorphism of the Lie algebras $(\mathbb{R}^3, \times)$ and $\mathfrak{so}(3)$, the tangent map df_R is given by

$$df_R(R\omega) = R(\omega \times u).$$

Here is another example inspired by an optimization problem investigated by Taylor and Kriegman.

Example 7.7. Pick any two vectors $u, v \in \mathbb{R}^3$, and let $f\colon \mathbf{SO}(3) \to \mathbb{R}$ be the function given by

$$f(R) = (u^\top Rv)^2.$$

To compute $df_R\colon T_R\mathbf{SO}(3) \to T_{f(R)}\mathbb{R}$, since $T_R\mathbf{SO}(3) = R\mathfrak{so}(3)$ and $T_{f(R)}\mathbb{R} = \mathbb{R}$, again pick any tangent vector $RB \in R\mathfrak{so}(3) = T_R\mathbf{SO}(3)$ (where B is any 3×3 skew symmetric matrix), let $\gamma(t) = Re^{tB}$ be the curve through R such that $\gamma'(0) = RB$, and compute via the product rule

$$\begin{aligned} df_R(RB) &= (f(\gamma(t)))'(0) \\ &= ((u^\top Re^{tB}v)^2)'(0) \\ &= u^\top RBvu^\top Rv + u^\top Rvu^\top RBv \\ &= 2u^\top RBvu^\top Rv, \end{aligned}$$

where the last equality used the observation that $u^\top RBv$ and $u^\top Rv$ are real numbers. Therefore,

$$df_R(X) = 2u^\top Xvu^\top Rv, \quad X \in R\mathfrak{so}(3).$$

Unlike the case of functions defined on vector spaces, in order to define the gradient of f, a function defined on $\mathbf{SO}(3)$, a "nonflat" manifold, we need to pick a Riemannian metric on $\mathbf{SO}(3)$. We will explain how to do this in Chapter 13.

7.6 Submanifolds, Immersions, Embeddings

Although the notion of submanifold is intuitively rather clear, technically, it is a bit tricky. In fact, the reader may have noticed that many different definitions appear in books and that it is not obvious at first glance that these definitions are equivalent. What is important is that a submanifold M of a given manifold N has the topology induced by N but also that the charts of M are somehow induced by those of N.

Given m, n, with $0 \leq m \leq n$, we can view $\mathbb{R}^m$ as a subspace of $\mathbb{R}^n$ using the inclusion

$$\mathbb{R}^m \cong \mathbb{R}^m \times \{\underbrace{(0, \ldots, 0)}_{n-m}\} \hookrightarrow \mathbb{R}^m \times \mathbb{R}^{n-m} = \mathbb{R}^n, \quad (x_1, \ldots, x_m) \mapsto (x_1, \ldots, x_m, \underbrace{0, \ldots, 0}_{n-m}).$$

Definition 7.26. Given a C^k-manifold N of dimension n, a subset M of N is an *m-dimensional submanifold of N* (where $0 \leq m \leq n$) iff for every point $p \in M$, there is a chart (U, φ) of N (in the maximal atlas for N), with $p \in U$, so that

$$\varphi(U \cap M) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}).$$

(We write $0_{n-m} = \underbrace{(0, \ldots, 0)}_{n-m}$.)

The subset $U \cap M$ of Definition 7.26 is sometimes called a *slice* of (U, φ) and we say that (U, φ) is *adapted to M* (See O'Neill [91] or Warner [114]).

Other authors, including Warner [114], use the term submanifold in a broader sense than us and they use the word *embedded submanifold* for what is defined in Definition 7.26.

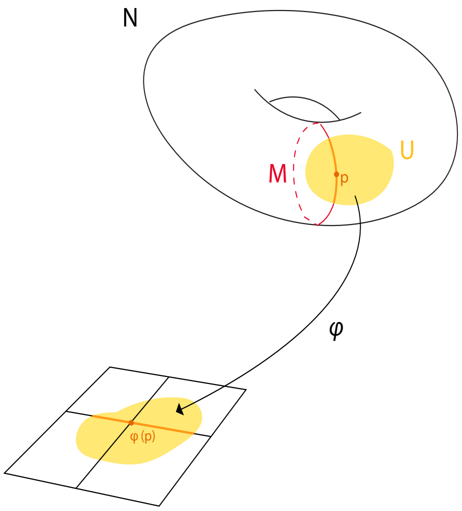


Figure 7.14: The red circle M is a 1-dimensional submanifold of the torus N.

The following proposition has an almost trivial proof but it justifies the use of the word submanifold.

Proposition 7.15. *Given a C^k-manifold N of dimension n, for any submanifold M of N of dimension $m \leq n$, the family of pairs $(U \cap M, \varphi \restriction U \cap M)$, where (U, φ) ranges over the charts over any atlas for N, is an atlas for M, where M is given the subspace topology. Therefore, M inherits the structure of a C^k-manifold.*

In fact, every chart on M arises from a chart on N in the following precise sense.

Proposition 7.16. *Given a C^k-manifold N of dimension n and a submanifold M of N of dimension $m \leq n$, for any $p \in M$ and any chart (W, η) of M at p, there is some chart (U, φ) of N at p, so that*

$$\varphi(U \cap M) = \varphi(U) \cap (\mathbb{R}^m \times \{0_{n-m}\}) \quad \textit{and} \quad \varphi \restriction U \cap M = \eta \restriction U \cap M,$$

where $p \in U \cap M \subseteq W$.

Proof. See Berger and Gostiaux [15] (Chapter 2). □

It is also useful to define more general kinds of "submanifolds."

Definition 7.27. Let $h\colon M \to N$ be a C^k-map of manifolds.

(a) The map h is an *immersion* of M into N iff dh_p is injective for all $p \in M$.

(b) The set $h(M)$ is an *immersed submanifold* of N iff h is an injective immersion.

(c) The map h is an *embedding* of M into N iff h is an injective immersion such that the induced map, $M \longrightarrow h(M)$, is a homeomorphism, where $h(M)$ is given the subspace topology (equivalently, h is an open map from M into $h(M)$ with the subspace topology). We say that $h(M)$ (with the subspace topology) is an *embedded submanifold* of N.

(d) The map h is a *submersion* of M into N iff dh_p is surjective for all $p \in M$.

Again, we warn our readers that certain authors (such as Warner [114]) call $h(M)$, in (b), a submanifold of N! We prefer the terminology *immersed submanifold*.

The notion of immersed submanifold arises naturally in the framework of Lie groups. Indeed, the fundamental correspondence between Lie groups and Lie algebras involves Lie subgroups that are not necessarily closed. But, as we will see later, subgroups of Lie groups that are also submanifolds are always closed. It is thus necessary to have a more inclusive notion of submanifold for Lie groups and the concept of immersed submanifold is just what's needed.

Immersions of $\mathbb{R}$ into $\mathbb{R}^3$ are parametric curves and immersions of $\mathbb{R}^2$ into $\mathbb{R}^3$ are parametric surfaces. These have been extensively studied, for example, see DoCarmo [38], Berger and Gostiaux [15], or Gallier [48].

Immersions (i.e., subsets of the form $h(M)$, where h is an immersion) are generally neither injective immersions (i.e., subsets of the form $h(M)$, where h is an injective immersion) nor embeddings (or submanifolds). For example, immersions can have self-intersections, as the plane curve (nodal cubic) shown in Figure 7.15 and given by $x = t^2 - 1; y = t(t^2 - 1)$.

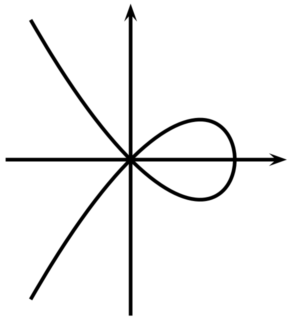


Figure 7.15: A nodal cubic; an immersion, but not an immersed submanifold.

Note that the cuspidal cubic, $t \mapsto (t^2, t^3)$, (see Figure 7.4), is an injective map, but it is not an immersion since its derivative at the origin is zero.

Injective immersions are generally not embeddings (or submanifolds) because $h(M)$ may not be homeomorphic to M. An example is given by the lemniscate of Bernoulli shown in Figure 7.16, an injective immersion of $\mathbb{R}$ into $\mathbb{R}^2$:

$$\begin{aligned} x &= \frac{t(1+t^2)}{1+t^4}, \\ y &= \frac{t(1-t^2)}{1+t^4}. \end{aligned}$$

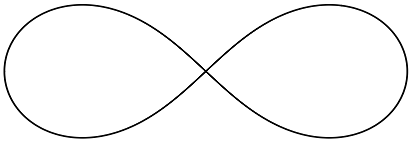


Figure 7.16: Lemniscate of Bernoulli; an immersed submanifold, but not an embedding.

When $t = 0$, the curve passes through the origin. When $t \mapsto -\infty$, the curve tends to the origin from the left and from above, and when $t \mapsto +\infty$, the curve tends tends to the origin from the right and from below. Therefore, the inverse of the map defining the lemniscate of Bernoulli is not continuous at the origin.

Another interesting example is the immersion of $\mathbb{R}$ into the 2-torus, $T^2 = S^1 \times S^1 \subseteq \mathbb{R}^4$, given by

$$t \mapsto (\cos t, \sin t, \cos ct, \sin ct),$$

where $c \in \mathbb{R}$. One can show that the image of $\mathbb{R}$ under this immersion is closed in T^2 iff c is rational. Moreover, the image of this immersion is dense in T^2 but not closed iff c is irrational. The above example can be adapted to the torus in $\mathbb{R}^3$: One can show that the immersion given by

$$t \mapsto ((2+\cos t)\cos(\sqrt{2}\,t), (2+\cos t)\sin(\sqrt{2}\,t), \sin t),$$

is dense but not closed in the torus (in $\mathbb{R}^3$) given by

$$(s,t) \mapsto ((2+\cos s)\cos t, (2+\cos s)\sin t, \sin s),$$

where $s, t \in \mathbb{R}$.

There is, however, a close relationship between submanifolds and embeddings.

Proposition 7.17. *If M is a submanifold of N, then the inclusion map $j\colon M \to N$ is an embedding. Conversely, if $h\colon M \to N$ is an embedding, then $h(M)$ with the subspace topology is a submanifold of N and h is a diffeomorphism between M and $h(M)$.*

Proof. See O'Neill [91] (Chapter 1) or Berger and Gostiaux [15] (Chapter 2). □

In summary, **embedded submanifolds and (our) submanifolds coincide**. Some authors refer to spaces of the form $h(M)$, where h is an injective immersion, as *immersed submanifolds* and we have adopted this terminology. However, in general, an immersed submanifold is *not* a submanifold. One case where this holds is when M is compact, since then, a bijective continuous map is a homeomorphism. For yet a notion of submanifold intermediate between immersed submanifolds and (our) submanifolds, see Sharpe [107] (Chapter 1).

7.7 Problems

Problem 7.1. Prove that the collection $\widetilde{\mathcal{A}}$ of all charts compatible with an atlas $\mathcal{A}$ is a maximal atlas in the equivalence class of atlases compatible with $\mathcal{A}$.

Problem 7.2. Prove that every C^k-manifold ($k = 0, \ldots, \infty$) is locally connected and locally compact.

Problem 7.3. Consider a C^k n-atlas on a set M as defined in Definition 7.6. Give M the topology in which the open sets are arbitrary unions of domains of charts U_i, more precisely, the U_i's of the maximal atlas defining the differentiable structure on M.

(1) Check that the axioms of a topology are verified, and M is indeed a topological space with this topology.

(2) Check that that when M is equipped with the above topology, then the maps $\varphi_i\colon U_i \to \varphi_i(U_i)$ are homeomorphisms, so M is a manifold according to Definition 7.4.

Problem 7.4. Referring to Example 7.1, show that on the overlap $U_N \cap U_S = S^n - \{N, S\}$, the transition maps

$$\mathcal{I} = \sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1}$$

defined on $\varphi_N(U_N \cap U_S) = \varphi_S(U_N \cap U_S) = \mathbb{R}^n - \{0\}$, are given by

$$(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2}\,(x_1, \ldots, x_n);$$

that is, the inversion $\mathcal{I}$ of center $O = (0, \ldots, 0)$ and power 1.

Problem 7.5. In Example 7.4, check that the transition map $\varphi_T \circ \varphi_S^{-1}$ from $\varphi_S(U_S \cap U_T)$ to $\varphi_T(U_S \cap U_T)$ is given by

$$M \mapsto (P_3 + P_4 M)(P_1 + P_2 M)^{-1},$$

where

$$\begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix} = P_T^{-1} P_S$$

is the matrix of the permutation $\pi_T^{-1} \circ \pi_S$ and M is an $(n-k) \times k$ matrix.

Problem 7.6. Referring to Example 7.4, prove that the collection of k-planes represented by matrices in U_S is precisely the set of k-planes W supplementary to the $(n-k)$-plane spanned by the canonical basis vectors $e_{j_{k+1}}, \ldots, e_{j_n}$ (i.e., $\mathrm{span}(W \cup \{e_{j_{k+1}}, \ldots, e_{j_n}\}) = \mathbb{R}^n$, where $S = \{i_1, \ldots, i_k\}$ and $\{j_{k+1}, \ldots, j_n\} = \{1, \ldots, n\} - S$).

Problem 7.7. Prove that the condition of Definition 7.7 does not depend on the choice of charts.

Problem 7.8. Check that the relation of Definition 7.14 is an equivalence relation.

Problem 7.9. Prove that the operations $\lambda[f]$ and $[f][g]$ are well defined on germs of C^k-functions at a point p of a manifold.

Problem 7.10. Check that being stationary at a point p does not depend on the chart (U, φ) at p or on the function chosen in the germ $\mathbf{f}$.

Problem 7.11. Consider the immersion of $\mathbb{R}$ in the torus $T^2 = S^1 \times S^1 \subseteq \mathbb{R}^4$, given by

$$t \mapsto (\cos t, \sin t, \cos ct, \sin ct),$$

where $c \in \mathbb{R}$. Show that the image of $\mathbb{R}$ under this immersion is closed in T^2 iff c is rational. Moreover, show that the image of this immersion is dense in T^2 but not closed iff c is irrational.

Problem 7.12. Show that the immersion given by

$$t \mapsto ((2 + \cos t)\cos(\sqrt{2}\, t), (2 + \cos t)\sin(\sqrt{2}\, t), \sin t),$$

is dense but not closed in the torus (in $\mathbb{R}^3$) given by

$$(s, t) \mapsto ((2 + \cos s)\cos t, (2 + \cos s)\sin t, \sin s),$$

where $s, t \in \mathbb{R}$.

Problem 7.13. Prove that if N is a compact manifold, then an injective immersion $f\colon N \to M$ is an embedding.

Problem 7.14. Let $f\colon M \to N$ be a map of smooth manifolds. A point, $p \in M$, is called a *critical point (of f)* iff df_p is *not* surjective and a point $q \in N$ is called a *critical value (of f)* iff $q = f(p)$, for some critical point, $p \in M$. A point $p \in M$ is a *regular point (of f)* iff p is not critical, i.e., df_p is surjective, and a point $q \in N$ is a *regular value (of f)* iff it is not a critical value. In particular, any $q \in N - f(M)$ is a regular value and $q \in f(M)$ is a regular value iff *every* $p \in f^{-1}(q)$ is a regular point (but, in contrast, q is a critical value iff *some* $p \in f^{-1}(q)$ is critical).

(a) Prove that for every regular value, $q \in f(M)$, the preimage $Z = f^{-1}(q)$ is a manifold of dimension $\dim(M) - \dim(N)$.

Hint. Pick any $p \in f^{-1}(q)$ and some parametrizations φ at p and ψ at q, with $\varphi(0) = p$ and $\psi(0) = q$, and consider $h = \psi^{-1} \circ f \circ \varphi$. Prove that dh_0 is surjective and then apply Lemma 3.5.

(b) Under the same assumptions as (a), prove that for every point $p \in Z = f^{-1}(q)$, the tangent space, T_pZ, is the kernel of $df_p \colon T_pM \to T_qN$.

(c) If $X, Z \subseteq \mathbb{R}^N$ are manifolds and $Z \subseteq X$, we say that Z *is a submanifold of* X. Assume there is a smooth function, $g \colon X \to \mathbb{R}^k$, and that $0 \in \mathbb{R}^k$ is a regular value of g. Then, by (a), $Z = g^{-1}(0)$ is a submanifold of X of dimension $\dim(X) - k$. Let $g = (g_1, \ldots, g_k)$, with each g_i a function, $g_i \colon X \to \mathbb{R}$. Prove that for any $p \in X$, dg_p is surjective iff the linear forms, $(dg_i)_p \colon T_pX \to \mathbb{R}$, are linearly independent. In this case, we say that $g_1, \ldots, g_k$ are *independent at* p. We also say that Z is *cut out* by $g_1, \ldots, g_k$ when

$$Z = \{p \in X \mid g_1(p) = 0, \ldots, g_k(p) = 0\}$$

with $g_1, \ldots, g_k$ independent for all $p \in Z$.

Let $f \colon X \to Y$ be a smooth maps of manifolds and let $q \in f(X)$ be a regular value. Prove that $Z = f^{-1}(q)$ is a submanifold of X cut out by $k = \dim(X) - \dim(Y)$ independent functions.

Hint. Pick some parametrization, ψ, at q, so that $\psi(0) = q$ and check that 0 is a regular value of $g = \psi^{-1} \circ f$, so that $g_1, \ldots, g_k$ work.

(d) Now, assume Z is a submanifold of X. Prove that locally, Z is cut out by independent functions. This means that if $k = \dim(X) - \dim(Z)$, the *codimension* of Z in X, then for every $z \in Z$, there are k independent functions, $g_1, \ldots, g_k$, defined on some open subset, $W \subseteq X$, with $z \in W$, so that $Z \cap W$ is the common zero set of the g_i's.

Hint. Apply Lemma 3.4 to the immersion $Z \longrightarrow X$.

(e) We would like to generalize our result in (a) to the more general situation where we have a smooth map, $f \colon X \to Y$, but this time, we have a submanifold, $Z \subseteq Y$ and we are investigating whether $f^{-1}(Z)$ is a submanifold of X. In particular, if X is also a submanifold of Y and f is the inclusion of X into Y, then $f^{-1}(Z) = X \cap Z$.

Convince yourself that, in general, the intersection of two submanifolds is *not* a submanifold. Try examples involving curves and surfaces and you will see how bad the situation can be. What is needed is a notion generalizing that of a regular value, and this turns out to be the notion of transversality.

We say that f *is transversal to* Z iff

$$df_p(T_pX) + T_{f(p)}Z = T_{f(p)}Y,$$

for all $p \in f^{-1}(Z)$. (Recall, if U and V are subspaces of a vector space, E, then $U + V$ is the subspace $U + V = \{u + v \in E \mid u \in U,\, v \in V\}$). In particular, if f is the inclusion of X into Y, the transversality condition is

$$T_pX + T_pZ = T_pY,$$

for all $p \in X \cap Z$.

Draw several examples of transversal intersections to understand better this concept. Prove that if f is transversal to Z, then $f^{-1}(Z)$ is a submanifold of X of codimension equal to $\dim(Y) - \dim(Z)$.

Hint. The set $f^{-1}(Z)$ is a manifold iff for every $p \in f^{-1}(Z)$, there is some open subset, $U \subseteq X$, with $p \in U$, and $f^{-1}(Z) \cap U$ is a manifold. First, use (d) to assert that locally near $q = f(p)$, Z is cut out by $k = \dim(Y) - \dim(Z)$ independent functions, $g_1, \ldots, g_k$, so that locally near p, the preimage $f^{-1}(Z)$ is cut out by $g_1 \circ f, \ldots, g_k \circ f$. If we let $g = (g_1, \ldots, g_k)$, it is a submersion and the issue is to prove that 0 is a regular value of $g \circ f$ in order to apply (a). Show that transversality is just what's needed to show that 0 is a regular value of $g \circ f$.

(f) With the same assumptions as in (g) (f is transversal to Z), if $W = f^{-1}(Z)$, prove that for every $p \in W$,

$$T_pW = (df_p)^{-1}(T_{f(p)}Z),$$

the preimage of $T_{f(p)}Z$ by $df_p \colon T_pX \to T_{f(p)}Y$. In particular, if f is the inclusion of X into Y, then

$$T_p(X \cap Z) = T_pX \cap T_pZ.$$

(g) Let $X, Z \subseteq Y$ be two submanifolds of Y, with X compact, Z closed, $\dim(X) + \dim(Z) = \dim(Y)$ and X transversal to Z. Prove that $X \cap Z$ consists of a finite set of points.

Problem 7.15. Show that a smooth map $f \colon M \to \mathbb{R}^m$ from a compact manifold M to $\mathbb{R}^m$ has some critical point.

Chapter 8

Construction of Manifolds From Gluing Data ⊛

8.1 Sets of Gluing Data for Manifolds

The definition of a manifold given in Chapter 7 assumes that the underlying set M is already known. However, there are situations where we only have some indirect information about the overlap of the domains U_i of the local charts defining our manifold M in terms of the transition functions

$$\varphi_i^j = \varphi_{ji}\colon \varphi_i(U_i \cap U_j) \to \varphi_j(U_i \cap U_j),$$

but where M itself is not known. For example, this situation happens when trying to construct a surface approximating a 3D-mesh. If we let $\Omega_{ij} = \varphi_i(U_i \cap U_j)$ and $\Omega_{ji} = \varphi_j(U_i \cap U_j)$, then φ_{ji} can be viewed as a "gluing map"

$$\varphi_{ji}\colon \Omega_{ij} \to \Omega_{ji}$$

between two open subsets of Ω_i and Ω_j, respectively.

For technical reasons, it is desirable to assume that the images $\Omega_i = \varphi_i(U_i)$ and $\Omega_j = \varphi_j(U_j)$ of distinct charts are disjoint, but this can always be achieved for manifolds. Indeed, the map

$$\beta\colon (x_1, \ldots, x_n) \mapsto \left(\frac{x_1}{\sqrt{1 + \sum_{i=1}^n x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 + \sum_{i=1}^n x_i^2}} \right)$$

is a smooth diffeomorphism from $\mathbb{R}^n$ to the open unit ball $B(0, 1)$, with inverse given by

$$\beta^{-1}\colon (x_1, \ldots, x_n) \mapsto \left(\frac{x_1}{\sqrt{1 - \sum_{i=1}^n x_i^2}}, \ldots, \frac{x_n}{\sqrt{1 - \sum_{i=1}^n x_i^2}} \right).$$

Since M has a countable basis, using compositions of β with suitable translations, we can make sure that the Ω_i's are mapped diffeomorphically to disjoint open subsets of $\mathbb{R}^n$.

Remarkably, manifolds can be constructed using the "gluing process" alluded to above from what is often called sets of "gluing data." In this chapter we are going to describe this construction and prove its correctness in detail, provided some mild assumptions on the gluing data. It turns out that this procedure for building manifolds can be made practical. Indeed, it is the basis of a class of new methods for approximating 3D meshes by smooth surfaces, see Siqueira, Xu and Gallier [108].

Some care must be exercised to ensure that the space obtained by gluing the pieces Ω_{ij} and Ω_{ji} is Hausdorff. Some care must also be exercised in formulating the consistency conditions relating the φ_{ji}'s (the so-called "cocycle condition"). This is because the traditional condition (for example, in bundle theory) has to do with triple overlaps of the $U_i = \varphi_i^{-1}(\Omega_i)$ on the manifold M, but in our situation, we do not have M nor the parametrization maps $\theta_i = \varphi_i^{-1}$, and the cocycle condition on the φ_{ji}'s has to be stated in terms of the Ω_i's and the Ω_{ji}'s.

Note that if the Ω_{ij} arise from the charts of a manifold, then nonempty triple intersections $U_i \cap U_j \cap U_k$ of domains of charts have images $\varphi_i(U_i \cap U_j \cap U_k)$ in Ω_i, $\varphi_j(U_i \cap U_j \cap U_k)$ in Ω_j, and $\varphi_k(U_i \cap U_j \cap U_k)$ in Ω_k, and since the φ_i's are bijective maps, we get

$$\varphi_i(U_i \cap U_j \cap U_k) = \varphi_i(U_i \cap U_j \cap U_i \cap U_k) = \varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k) = \Omega_{ij} \cap \Omega_{ik},$$

and similarly

$$\varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk}, \quad \varphi_k(U_i \cap U_j \cap U_k) = \Omega_{ki} \cap \Omega_{kj},$$

and these sets are related. Indeed, we have

$$\begin{aligned}\varphi_{ji}(\Omega_{ij} \cap \Omega_{ik}) &= \varphi_j \circ \varphi_i^{-1}(\varphi_i(U_i \cap U_j) \cap \varphi_i(U_i \cap U_k)) \\ &= \varphi_j(U_i \cap U_j \cap U_k) = \Omega_{ji} \cap \Omega_{jk},\end{aligned}$$

and similar equations relating the other "triple intersections." In particular,

$$\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik},$$

which implies that

$$\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) \subseteq \Omega_{ik}.$$

This is important, because $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$ is the domain of $\varphi_{kj} \circ \varphi_{ji}$ and Ω_{ik} is the domain of φ_{ki}, so the condition $\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$ implies that the domain of $\varphi_{kj} \circ \varphi_{ji}$ is a subset of the domain of φ_{ki}. See Figure 8.1. The definition of gluing data given by Grimm and Hughes [53, 54] misses the above condition.

Finding an easily testable necessary and sufficient criterion for the Hausdorff condition appears to be a very difficult problem. We propose a necessary and sufficient condition, but it is not easily testable in general. If M is a manifold, then observe that difficulties may arise when we want to separate two distinct point $p, q \in M$ such that p and q neither belong to the same open $\theta_i(\Omega_i)$, (recalling that $\theta_i = \varphi_i^{-1}$), nor to two disjoint opens $\theta_i(\Omega_i)$ and $\theta_j(\Omega_j)$,

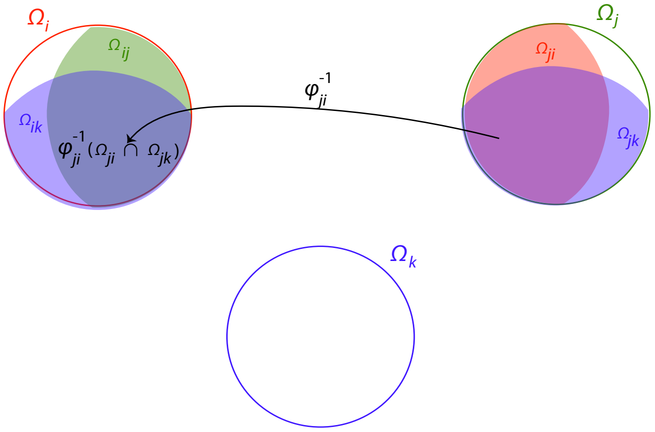


Figure 8.1: The domain of φ_{ki} is the blue region in the red circle. A subset of this domain is $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$, namely the pull back of the intersection of the blue and red regions from the green circle.

but instead to the boundary points in $(\partial(\theta_i(\Omega_{ij})) \cap \theta_i(\Omega_i)) \cup (\partial(\theta_j(\Omega_{ji})) \cap \theta_j(\Omega_j))$. In this case, there are some disjoint open subsets U_p and U_q of M with $p \in U_p$ and $q \in U_q$, and we get two disjoint open subsets $V_x = \theta_i^{-1}(U_p) = \varphi_i(U_p) \subseteq \Omega_i$ and $V_y = \theta_j^{-1}(U_q) \subseteq \Omega_j$ with $\theta_i(x) = p$, $\theta_j(y) = q$, and such that $x \in \partial(\Omega_{ij}) \cap \Omega_i$, $y \in \partial(\Omega_{ji}) \cap \Omega_j$, and no point in $V_y \cap \Omega_{ji}$ is the image of any point in $V_x \cap \Omega_{ij}$ by φ_{ji}. See Figure 8.2. Since V_x and V_y are open, we may assume that they are open balls. This necessary condition turns out to be also sufficient.

With the above motivations in mind, here is the definition of sets of gluing data.

Definition 8.1. Let n be an integer with $n \geq 1$ and let k be either an integer with $k \geq 1$ or $k = \infty$. A *set of gluing data* is a triple $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$ satisfying the following properties, where I is a (nonempty) countable set:

(1) For every $i \in I$, the set Ω_i is a nonempty open subset of $\mathbb{R}^n$ called a *parametrization domain*, for short, *p-domain*, and the Ω_i are pairwise disjoint (*i.e.*, $\Omega_i \cap \Omega_j = \emptyset$ for all $i \neq j$).

(2) For every pair $(i, j) \in I \times I$, the set Ω_{ij} is an open subset of Ω_i. Furthermore, $\Omega_{ii} = \Omega_i$ and $\Omega_{ij} \neq \emptyset$ iff $\Omega_{ji} \neq \emptyset$. Each nonempty Ω_{ij} (with $i \neq j$) is called a *gluing domain*.

(3) If we let
$$K = \{(i, j) \in I \times I \mid \Omega_{ij} \neq \emptyset\},$$
then $\varphi_{ji} \colon \Omega_{ij} \to \Omega_{ji}$ is a C^k bijection for every $(i, j) \in K$ called a *transition function* (or *gluing function*) and the following condition holds:

   (a) $\varphi_{ii} = \mathrm{id}_{\Omega_i}$, for all $i \in I$.

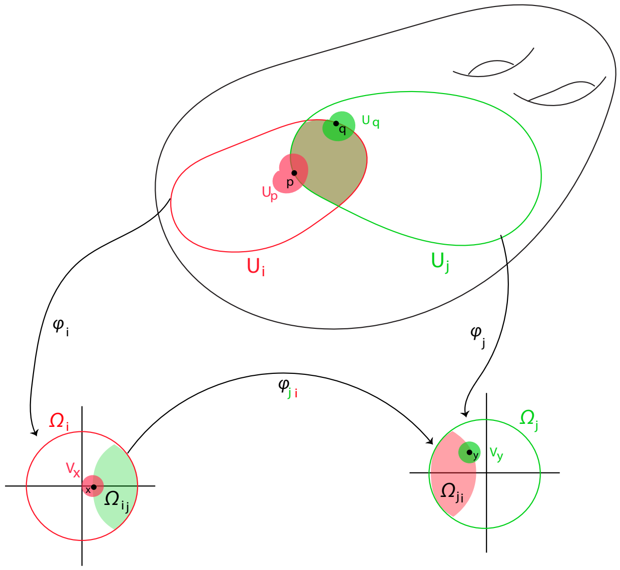


Figure 8.2: A schematic illustration of how to separate boundary points.

(b) $\varphi_{ij} = \varphi_{ji}^{-1}$, for all $(i, j) \in K$.

(c) For all i, j, k, if $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$, then $\varphi_{ij}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik}$, and $\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x)$, for all $x \in \Omega_{ij} \cap \Omega_{ik}$.

Condition (c) is called the *cocycle* condition. See Figure 8.3.

(4) For every pair $(i, j) \in K$, with $i \neq j$, for every $x \in \partial(\Omega_{ij}) \cap \Omega_i$ and every $y \in \partial(\Omega_{ji}) \cap \Omega_j$, there are open balls V_x and V_y centered at x and y so that no point of $V_y \cap \Omega_{ji}$ is the image of any point of $V_x \cap \Omega_{ij}$ by φ_{ji}. See Figure 8.2.

Remarks.

(1) In practical applications, the index set I is of course finite and the open subsets Ω_i may have special properties (for example, connected; open simplicies, *etc.*).

(2) We are only interested in the Ω_{ij}'s that are nonempty, but empty Ω_{ij}'s do arise in proofs and constructions, and this is why our definition allows them.

(3) Observe that $\Omega_{ij} \subseteq \Omega_i$ and $\Omega_{ji} \subseteq \Omega_j$. If $i \neq j$, as Ω_i and Ω_j are disjoint, so are Ω_{ij} and Ω_{ij}.

(4) The cocycle Condition (c) may seem overly complicated but it is actually needed to guarantee the transitivity of the relation $\sim$ defined in the proof of Proposition 8.1. Flawed versions of Condition (c) appear in the literature; see the discussion after the

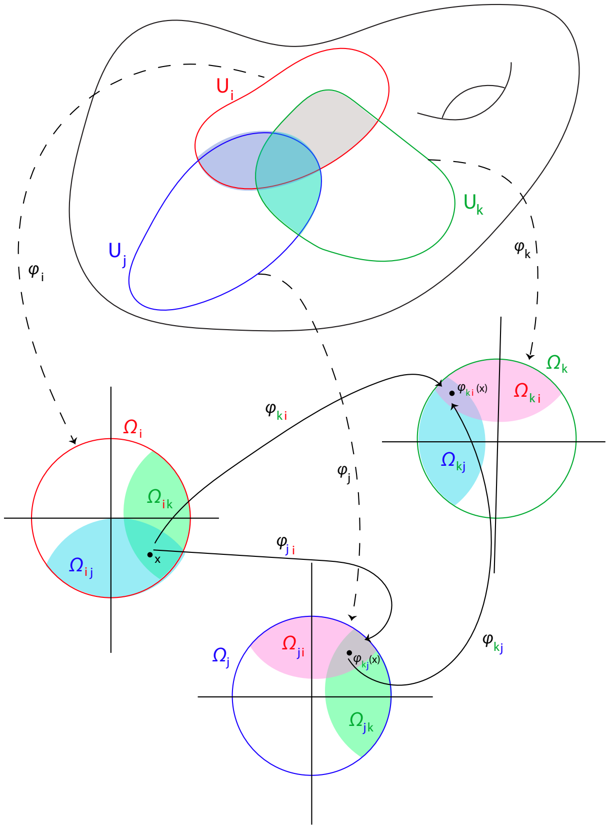


Figure 8.3: A schematic illustration of the cocycle condition.

proof of Proposition 8.1. The problem is that $\varphi_{kj}\circ\varphi_{ji}$ is a partial function whose domain $\varphi_{ji}^{-1}(\Omega_{ji}\cap\Omega_{jk})$ is not necessarily related to the domain Ω_{ik} of φ_{ki}. To ensure transitivity of $\sim$, we must assert that whenever the composition $\varphi_{kj}\circ\varphi_{ji}$ has a nonempty domain, this domain is contained in the domain Ω_{ik} of φ_{ki}, and that $\varphi_{kj}\circ\varphi_{ji}$ and φ_{ki} agree in $\varphi_{ji}^{-1}(\Omega_{ji}\cap\Omega_{jk})$.

Since the φ_{ji} are bijections, it turns out that Condition (c) implies Conditions (a) and (b). To get (a), set $i=j=k$. Then Condition (b) follows from (a) and (c) by setting $k=i$.

(5) If M is a C^k manifold (including $k=\infty$), then using the notation of our introduction, it is easy to check that the open sets Ω_i, Ω_{ij} and the gluing functions φ_{ji} satisfy the conditions of Definition 8.1 (provided that we fix the charts so that the images of distinct charts are disjoint). Proposition 8.1 will show that a manifold can be reconstructed from a set of gluing data.

The idea of defining gluing data for manifolds is not new. André Weil introduced this idea to define abstract algebraic varieties by gluing irreducible affine sets in his book [115] published in 1946. The same idea is well-known in bundle theory and can be found in standard texts such as Steenrod [109], Bott and Tu [18], Morita [87] and Wells [117].

The beauty of the idea is that it allows the reconstruction of a manifold M without having prior knowledge of the topology of this manifold (that is, without having explicitly the underlying topological space M) by gluing open subets of $\mathbb{R}^n$ (the Ω_i's) according to prescribed gluing instructions (namely, glue Ω_i and Ω_j by identifying Ω_{ij} and Ω_{ji} using φ_{ji}). This method of specifying a manifold separates clearly the local structure of the manifold (given by the Ω_i's) from its global structure which is specified by the gluing functions. Furthermore, this method ensures that the resulting manifold is C^k (even for $k = \infty$) with no extra effort since the gluing functions φ_{ji} are assumed to be C^k.

Grimm and Hughes [53, 54] appear to be the first to have realized the power of this latter property for practical applications, and we wish to emphasize that this is a very significant discovery. However, Grimm [53] uses a condition stronger than our Condition (4) to ensure that the resulting space is Hausdorff. The cocycle condition in Grimm and Hughes [53, 54] is also not strong enough to ensure transitivity of the relation $\sim$. We will come back to these points after the proof of Proposition 8.1.

Working with overlaps of *open subsets* of the parameter domain makes it much easier to enforce smoothness conditions compared to the traditional approach with splines where the parameter domain is subdivided into *closed* regions, and where enforcing smoothness along boundaries is much more difficult.

Let us show that a set of gluing data defines a C^k manifold in a natural way.

Proposition 8.1. *For every set of gluing data $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, there is an n-dimensional C^k manifold $M_{\mathcal{G}}$ whose transition functions are the φ_{ji}'s.*

Proof. Define the binary relation $\sim$ on the disjoint union $\coprod_{i\in I}\Omega_i$ of the open sets Ω_i as follows: For all $x, y \in \coprod_{i\in I}\Omega_i$,

$$x \sim y \quad \text{iff} \quad (\exists (i,j) \in K)(x \in \Omega_{ij}, y \in \Omega_{ji}, y = \varphi_{ji}(x)).$$

We claim that $\sim$ is an equivalence relation. This follows easily from the cocycle condition. Clearly Condition 3a of Definition 8.1 ensures reflexivity, while Condition 3b ensures symmetry. To check transitivity, assume that $x \sim y$ and $y \sim z$. Then there are some i, j, k such that (i) $x \in \Omega_{ij}$, $y \in \Omega_{ji} \cap \Omega_{jk}$, $z \in \Omega_{kj}$, and (ii) $y = \varphi_{ji}(x)$ and $z = \varphi_{kj}(y)$. Consequently, $\Omega_{ji} \cap \Omega_{jk} \neq \emptyset$ and $x \in \varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk})$, so by 3c, we get $\varphi_{ji}^{-1}(\Omega_{ji} \cap \Omega_{jk}) = \Omega_{ij} \cap \Omega_{ik} \subseteq \Omega_{ik}$. So, $\varphi_{ki}(x)$ is defined and by 3c again, $\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x) = z$, i.e., $x \sim z$, as desired. See Figure 8.4.

Since $\sim$ is an equivalence relation, let

$$M_{\mathcal{G}} = \left(\coprod_{i\in I}\Omega_i\right)/\sim$$

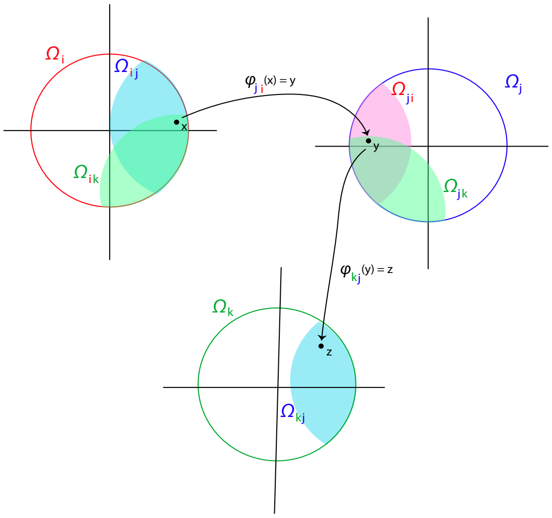


Figure 8.4: A schematic illustration transitivity, where $x \sim y$ and $y \sim z$ implies $x \sim z$.

be the quotient set and let $p\colon \coprod_{i\in I}\Omega_i \to M_{\mathcal{G}}$ be the quotient map, with $p(x) = [x]$, where $[x]$ denotes the equivalence class of x. Also, for every $i \in I$, let $\mathrm{in}_i\colon \Omega_i \to \coprod_{i\in I}\Omega_i$ be the natural injection and let

$$\tau_i = p \circ \mathrm{in}_i\colon \Omega_i \to M_{\mathcal{G}}\,.$$

Note that if $x \sim y$ and $x \neq y$, then $i \neq j$, as $\varphi_{ii} = \mathrm{id}$. But then, as $x \in \Omega_{ij} \subseteq \Omega_i$, $y \in \Omega_{ji} \subseteq \Omega_j$ and $\Omega_i \cap \Omega_j = \emptyset$ when $i \neq j$, if $x \sim y$ and $x, y \in \Omega_i$, then $x = y$. As a consequence we conclude that every τ_i is injective. We give $M_{\mathcal{G}}$ the largest topology that makes the bijections, $\tau_i\colon \Omega_i \to \tau_i(\Omega_i)$, into homeomorphisms. Then, if we let $U_i = \tau_i(\Omega_i)$ and $\varphi_i = \tau_i^{-1}$, it is immediately verified that the (U_i, φ_i) are charts and that this collection of charts forms a C^k atlas for $M_{\mathcal{G}}$. As there are countably many charts, $M_{\mathcal{G}}$ is second-countable.

To prove that the topology is Hausdorff, we first prove the following:

Claim. For all $(i, j) \in I \times I$, we have $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i, j) \in K$ and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})\,.$$

Assume that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ and let $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$. Observe that $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$ iff $z \sim x$ and $z \sim y$, for some $x \in \Omega_i$ and some $y \in \Omega_j$. Consequently, $x \sim y$, which implies that $(i, j) \in K$, $x \in \Omega_{ij}$ and $y \in \Omega_{ji}$. We have $[z] \in \tau_i(\Omega_{ij})$ iff $z \sim x$, for some $x \in \Omega_{ij}$. Then either $i = j$ and $z = x$ or $i \neq j$ and $z \in \Omega_{ji}$, which shows that $[z] \in \tau_j(\Omega_{ji})$, and consequently we get $\tau_i(\Omega_{ij}) \subseteq \tau_j(\Omega_{ji})$. Since the same argument applies by interchanging i and j, we have that $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, for all $(i, j) \in K$. Furthermore, because $\Omega_{ij} \subseteq \Omega_i$, $\Omega_{ji} \subseteq \Omega_j$, and $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, for all $(i, j) \in K$, we also have that $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}) \subseteq \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, for all $(i, j) \in K$. See Figure 8.5.

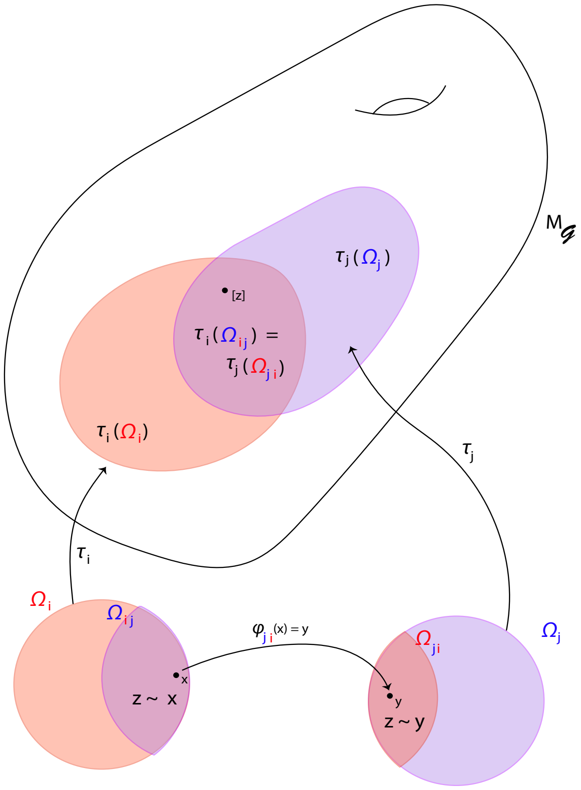


Figure 8.5: A schematic illustration of $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, where $M_{\mathcal{G}}$ is depicted as a torus.

For the reverse inclusion, if $[z] \in \tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, then we know that there is some $x \in \Omega_{ij}$ and some $y \in \Omega_{ji}$ such that $z \sim x$ and $z \sim y$, so $[z] = [x] \in \tau_i(\Omega_{ij})$ and $[z] = [y] \in \tau_j(\Omega_{ji})$, and then we get

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \subseteq \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

This proves that if $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$, then $(i, j) \in K$ and

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

Finally, assume that $(i, j) \in K$. Then, for any $x \in \Omega_{ij} \subseteq \Omega_i$, we have $y = \varphi_{ji}(x) \in \Omega_{ji} \subseteq \Omega_j$ and $x \sim y$, so that $\tau_i(x) = \tau_j(y)$, which proves that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$. So, our claim is true, and we can use it.

We now prove that the topology of $M_{\mathcal{G}}$ is Hausdorff. Pick $[x], [y] \in M_{\mathcal{G}}$ with $[x] \neq [y]$, for some $x \in \Omega_i$ and some $y \in \Omega_j$. Either $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \emptyset$, in which case, as τ_i and τ_j are homeomorphisms, $[x]$ and $[y]$ belong to the two disjoint open sets $\tau_i(\Omega_i)$ and $\tau_j(\Omega_j)$. If not, then by the claim, $(i, j) \in K$ and

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

There are several cases to consider:

1. If $i = j$ then x and y can be separated by disjoint opens V_x and V_y, and as τ_i is a homeomorphism, $[x]$ and $[y]$ are separated by the disjoint open subsets $\tau_i(V_x)$ and $\tau_j(V_y)$.

2. If $i \neq j$, $x \in \Omega_i - \overline{\Omega_{ij}}$ and $y \in \Omega_j - \overline{\Omega_{ji}}$, then $\tau_i(\Omega_i - \overline{\Omega_{ij}})$ and $\tau_j(\Omega_j - \overline{\Omega_{ji}})$ are disjoint open subsets separating $[x]$ and $[y]$, where $\overline{\Omega_{ij}}$ and $\overline{\Omega_{ji}}$ are the closures of Ω_{ij} and Ω_{ji}, respectively. See Figure 8.6.

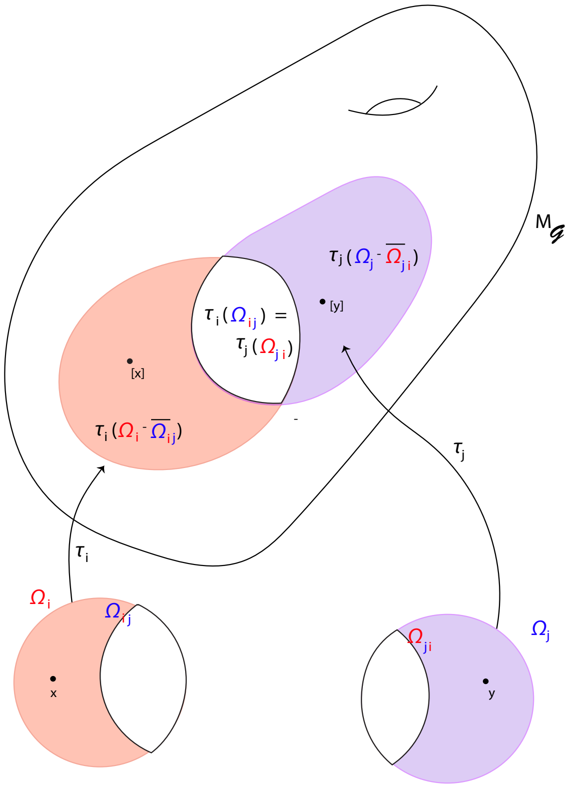


Figure 8.6: The separation of $[x]$ and $[y]$ when $x \in \Omega_i - \overline{\Omega_{ij}}$ and $y \in \Omega_j - \overline{\Omega_{ji}}$.

3. If $i \neq j$, $x \in \Omega_{ij}$ and $y \in \Omega_{ji}$, as $[x] \neq [y]$ and $y \sim \varphi_{ij}(y)$, then $x \neq \varphi_{ij}(y)$. We can separate x and $\varphi_{ij}(y)$ by disjoint open subsets V_x and V_y, and $[x]$ and $[y] = [\varphi_{ij}(y)]$ are separated by the disjoint open subsets $\tau_i(V_x)$ and $\tau_i(V_{\varphi_{ij}(y)})$. See Figure 8.7.

4. If $i \neq j$, $x \in \partial(\Omega_{ij}) \cap \Omega_i$ and $y \in \partial(\Omega_{ji}) \cap \Omega_j$, then we use Condition 4 of Definition 8.1. This condition yields two disjoint open subsets V_x and V_y with $x \in V_x$ and $y \in V_y$, such that no point of $V_x \cap \Omega_{ij}$ is equivalent to any point of $V_y \cap \Omega_{ji}$, and so $\tau_i(V_x)$ and $\tau_j(V_y)$ are disjoint open subsets separating $[x]$ and $[y]$. See Figure 8.2.

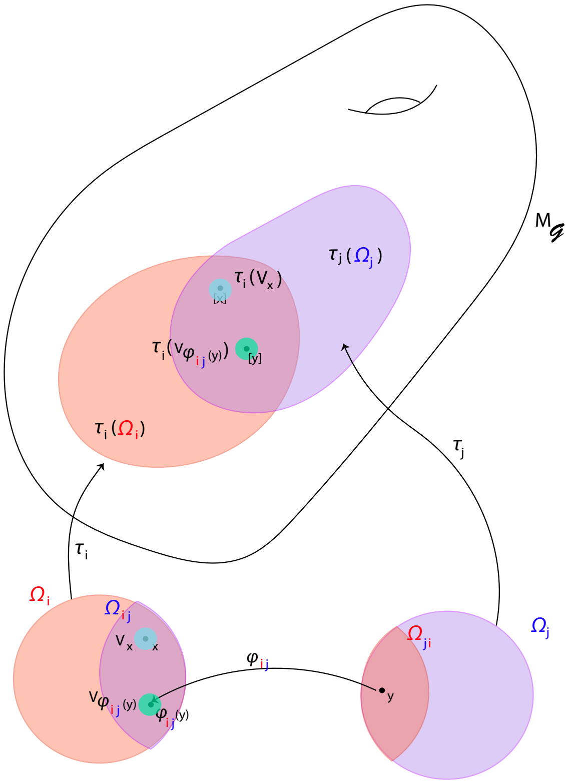


Figure 8.7: The separation of $[x]$ and $[y]$ when $x \neq \varphi_{ij}(y)$.

Therefore, the topology of $M_{\mathcal{G}}$ is Hausdorff and $M_{\mathcal{G}}$ is indeed a manifold. Finally, it is trivial to verify that the transition maps of $M_{\mathcal{G}}$ are the original gluing functions φ_{ij}, since $\varphi_i = \tau_i^{-1}$ and $\varphi_{ji} = \varphi_j \circ \varphi_i^{-1}$. □

It should be noted that as nice as it is, Proposition 8.1 is a theoretical construction that yields an "abstract" manifold, but does not yield any information as to the geometry of this manifold. Furthermore, the resulting manifold may not be orientable or compact, even if we start with a finite set of p-domains.

Here is an example showing that if Condition (4) of Definition 8.1 is omitted then we may get non-Hausdorff spaces. Cindy Grimm uses a similar example in her dissertation [53] (Appendix C2, page 126), but her presentation is somewhat confusing because her Ω_1 and Ω_2 appear to be two disjoint copies of the real line in $\mathbb{R}^2$, but these are not open in $\mathbb{R}^2$!

Let $\Omega_1 = (-3, -1)$, $\Omega_2 = (1, 3)$, $\Omega_{12} = (-3, -2)$, $\Omega_{21} = (1, 2)$ and $\varphi_{21}(x) = x + 4$. The resulting space M is a curve looking like a "fork," and the problem is that the images of -2 and 2 in M, which are distinct points of M, cannot be separated. See Figure 8.8. Indeed, the images of any two open intervals $(-2 - \epsilon, -2 + \epsilon)$ and $(2 - \eta, 2 + \eta)$ (for $\epsilon, \eta > 0$) always intersect, since $(-2 - \min(\epsilon, \eta), -2)$ and $(2 - \min(\epsilon, \eta), 2)$ are identified. Clearly Condition (4) fails.

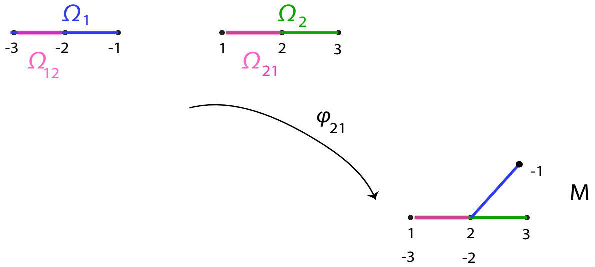


Figure 8.8: The fork construction M.

Cindy Grimm [53] (page 40) uses a condition stronger than our Condition (4) to ensure that the quotient, $M_{\mathcal{G}}$ is Hausdorff; namely, that for all $(i, j) \in K$ with $i \neq j$, the quotient $(\Omega_i \coprod \Omega_j)/\sim$ should be embeddable in $\mathbb{R}^n$. This is a rather strong condition that prevents obtaining a 2-sphere by gluing two open discs in $\mathbb{R}^2$ along an annulus (see Grimm [53], Appendix C2, page 126).

Remark: Readers familiar with fibre bundles may wonder why the cocycle Condition (3c) of Definition 8.1 is more arcane than the corresponding definition found in bundle theory. The reason is that if $\pi\colon E \to B$ is a (smooth or C^k) fibre bundle with fibre, F, then there is some open cover, (U_α), of the base space, B, and for every index, α, there is a *local trivialization map*, namely a diffeomorphism,

$$\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times F,$$

such that

$$\pi = p_1 \circ \varphi_\alpha,$$

where $p_1\colon U_\alpha \times F \to U_\alpha$ is the projection onto U_α. Whenever $U_\alpha \cap U_\beta \neq \emptyset$, we have a map

$$\varphi_\alpha \circ \varphi_\beta^{-1}\colon (U_\alpha \cap U_\beta) \times F \to (U_\alpha \cap U_\beta) \times F,$$

and because $\pi = p_1 \circ \varphi_\alpha$ for all α, there is a map,

$$g_{\beta\,\alpha}\colon U_\alpha \cap U_\beta \to \mathrm{Diff}(F),$$

where $\mathrm{Diff}(F)$ denotes the group of diffeomorphisms of the fibre, F, such that

$$\varphi_\alpha \circ \varphi_\beta^{-1}(b, p) = (b, g_{\beta\,\alpha}(b)(p)),$$

for all $b \in U_\alpha \cap U_\beta$ and all $p \in F$. The maps, $g_{\beta\,\alpha}$, are the *transition maps* of the bundle. Observe that for *all* $b \in U_\alpha \cap U_\beta$, the maps, $g_{\beta\,\alpha}(b)$, have *the same domain* and *the same* *range*, F. So, whenever $U_\alpha \cap U_\beta \cap U_\gamma \neq \emptyset$, for all $b \in U_\alpha \cap U_\beta \cap U_\gamma$, the maps $g_{\beta\,\alpha}$, $g_{\gamma\,\beta}$ and $g_{\gamma\,\alpha}$ have the same domain and the same range. Consequently, in this case, the cocycle condition can be simply stated as

$$g_{\gamma\,\alpha} = g_{\gamma\,\beta} \circ g_{\beta\,\alpha}\,,$$

without taking any precautions about the domains of these maps. However, in our situation (a manifold), the transition maps are of the form $\varphi_{ji}\colon \Omega_{ij} \to \Omega_{ji}$, where the Ω_{ij} are various unrelated open subsets of $\mathbb{R}^n$, and so, the composite map, $\varphi_{kj} \circ \varphi_{ji}$ only makes sense on a subset of Ω_{ij} (the domain of φ_{ji}). However, this subset need not be contained in the domain of φ_{ki}. So in order to avoid the extra complications we saw before, the constraints in Condition (3c) of Definition 8.1 must be imposed. In reconstructing a fibre bundle from B and the transition maps $g_{\beta\,\alpha}$, we use the $g_{\beta\,\alpha}$ to glue the spaces $U_\alpha \times F$ and $U_\beta \times F$ along $(U_\alpha \cap U_\beta) \times F$, where two points (a, p) and (b, q) in $(U_\alpha \cap U_\beta) \times F$ are identified iff $a = b$ and $q = g_{\beta\,\alpha}(a)(p)$. In reconstructing a manifold from a set of gluing data, we glue the open sets Ω_i and Ω_j along Ω_{ij} and Ω_{ji}, which are identified using the maps, φ_{ji}.

Grimm uses the following cocycle condition in [53] (page 40) and [54] (page 361):

(c') For all $x \in \Omega_{ij} \cap \Omega_{ik}$,
$$\varphi_{ki}(x) = \varphi_{kj} \circ \varphi_{ji}(x).$$

This condition is not strong enough to imply transitivity of the relation $\sim$, as shown by the following counter-example:

Let $\Omega_1 = (0, 3)$, $\Omega_2 = (4, 5)$, $\Omega_3 = (6, 9)$, $\Omega_{12} = (0, 1)$, $\Omega_{13} = (2, 3)$, $\Omega_{21} = \Omega_{23} = (4, 5)$, $\Omega_{32} = (8, 9)$, $\Omega_{31} = (6, 7)$, $\varphi_{21}(x) = x + 4$, $\varphi_{32}(x) = x + 4$ and $\varphi_{31}(x) = x + 4$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x + 8$, for all $x \in \Omega_{12}$, but $\Omega_{12} \cap \Omega_{13} = \emptyset$. Thus, $0.5 \sim 4.5 \sim 8.5$, and if the relation $\sim$ was transitive, then we would conclude that $0.5 \sim 8.5$. However, the definition of the relation $\sim$ requires that $\varphi_{31}(0.5)$ be defined, which is not the case. Therefore, the relation $\sim$ is not transitive. See Figure 8.9. The problem is that because $\Omega_{12} \cap \Omega_{13} = \emptyset$, Condition ($c'$) holds vacuously, but it is not strong enough to ensure that $\varphi_{31}(0.5)$ is defined.

Here is another counter-example in which $\Omega_{12} \cap \Omega_{13} \neq \emptyset$, using a disconnected open Ω_2.

Let $\Omega_1 = (0, 3)$, $\Omega_2 = (4, 5) \cup (6, 7)$, $\Omega_3 = (8, 11)$, $\Omega_{12} = (0, 1) \cup (2, 3)$, $\Omega_{13} = (2, 3)$, $\Omega_{21} = \Omega_{23} = (4, 5) \cup (6, 7)$, $\Omega_{32} = (8, 9) \cup (10, 11)$, $\Omega_{31} = (8, 9)$, $\varphi_{21}(x) = x + 4$, $\varphi_{32}(x) = x + 2$ on $(6, 7)$, $\varphi_{32}(x) = x + 6$ on $(4, 5)$, $\varphi_{31}(x) = x + 6$.

Note that the pairwise gluings yield Hausdorff spaces. Obviously, $\varphi_{32} \circ \varphi_{21}(x) = x + 6 = \varphi_{31}(x)$ for all $x \in \Omega_{12} \cap \Omega_{13} = (2, 3)$. Thus, $0.5 \sim 4.5 \sim 10.5$, but $0.5 \not\sim 10.5$ since $\varphi_{31}(0.5)$ is undefined. See Figure 8.10. This time Condition (c') holds and is nontrivial since $\Omega_{12} \cap \Omega_{13} = (2, 3)$, but it is not strong enough to ensure that $\varphi_{31}(0.5)$ is defined.

It is possible to give a construction, in the case of a surface, which builds a compact manifold whose geometry is "close" to the geometry of a prescribed 3D-mesh (see Siqueira, Xu

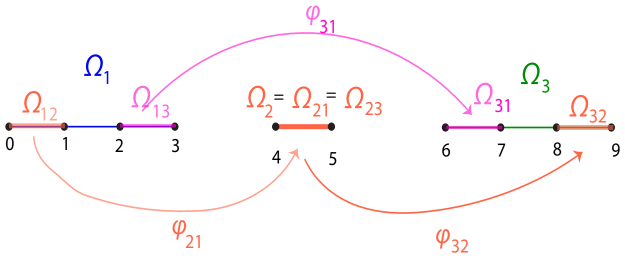


Figure 8.9: A counter-example to Condition (c'). Note $\varphi_{31} \neq \varphi_{32} \circ \varphi_{21}$ since these partial functions have different domains.

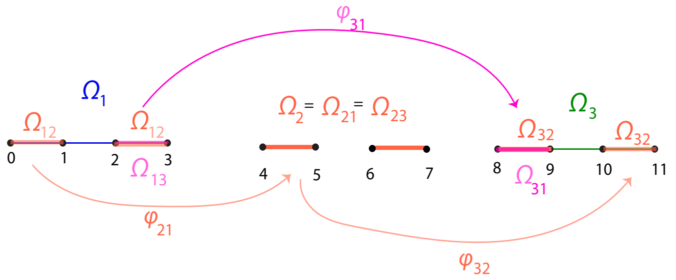


Figure 8.10: Another counter-example to Condition (c'). Once again $\varphi_{31} \neq \varphi_{32} \circ \varphi_{21}$ since these partial functions have different domains.

and Gallier [108]). Actually, we are not able to guarantee, in general, that the parametrization functions θ_i that we obtain are injective, but we are not aware of any algorithm that achieves this.

Given a set of gluing data, $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$, it is natural to consider the collection of manifolds M parametrized by maps $\theta_i \colon \Omega_i \to M$ whose domains are the Ω_i's and whose transitions functions are given by the φ_{ji}; that is, such that

$$\varphi_{ji} = \theta_j^{-1} \circ \theta_i.$$

We will say that such manifolds are *induced* by the set of gluing data $\mathcal{G}$.

The proof of Proposition 8.1 shows that the parametrization maps τ_i satisfy the property: $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i, j) \in K$, and if so

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

Furthermore, they also satisfy the consistency condition:

$$\tau_i = \tau_j \circ \varphi_{ji},$$

for all $(i, j) \in K$. If M is a manifold induced by the set of gluing data $\mathcal{G}$, because the θ_i's are injective and $\varphi_{ji} = \theta_j^{-1} \circ \theta_i$, the two properties stated above for the τ_i's also hold for the θ_i's. We will see in Section 8.2 that the manifold $M_{\mathcal{G}}$ is a "universal" manifold induced by $\mathcal{G}$, in the sense that every manifold induced by $\mathcal{G}$ is the image of $M_{\mathcal{G}}$ by some C^k map.

Interestingly, it is possible to characterize when two manifolds induced by sets of gluing data sharing the same sets of Ω_i's and Ω_{ij}'s are isomorphic in terms of a condition on their transition functions.

Proposition 8.2. *Given two sets of gluing data $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$ and $\mathcal{G}' = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi'_{ji})_{(i,j)\in K})$ over the same sets of Ω_i's and Ω_{ij}'s, for any two manifolds M and M' such that M is induced by $\mathcal{G}$ and M' is induced by $\mathcal{G}'$, where M and M' are given by families of parametrizations $(\Omega_i, \theta_i)_{i\in I}$ and $(\Omega_i, \theta'_i)_{i\in I}$ respectively, if $f \colon M \to M'$ is a C^k isomorphism, then there are C^k bijections $\rho_i \colon W_{ij} \to W'_{ij}$ for some open subsets $W_{ij}, W'_{ij} \subseteq \Omega_i$, such that*

$$\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \qquad \textit{for all} \quad x \in W'_{ij},$$

with $\varphi_{ji} = \theta_j^{-1} \circ \theta_i$ and $\varphi'_{ji} = \theta_j'^{-1} \circ \theta'_i$. Furthermore, $\rho_i = (\theta_i'^{-1} \circ f \circ \theta_i) \restriction W_{ij}$, and if $\theta_i'^{-1} \circ f \circ \theta_i$ is a bijection from Ω_i to itself and $\theta_i'^{-1} \circ f \circ \theta_i(\Omega_{ij}) = \Omega_{ij}$, for all i, j, then $W_{ij} = W'_{ij} = \Omega_i$. See Figure 8.11.

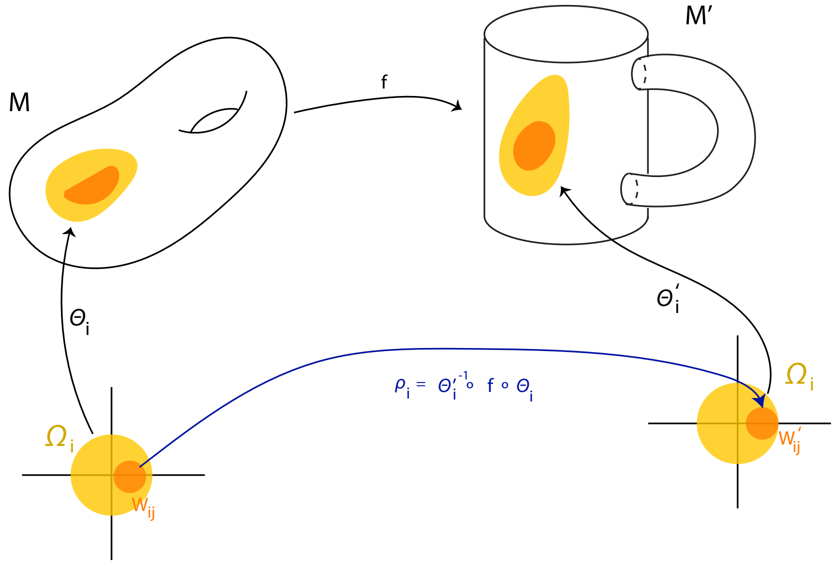


Figure 8.11: The construction of ρ_i between the diffeomorphic manifolds M and M'.

Proof. The composition $\theta_i'^{-1} \circ f \circ \theta_i$ is actually a partial function with domain

$$\mathrm{dom}(\theta_i'^{-1} \circ f \circ \theta_i) = \{x \in \Omega_i \mid \theta_i(x) \in f^{-1} \circ \theta_i'(\Omega_i)\},$$

and its "inverse" $\theta_i^{-1} \circ f^{-1} \circ \theta_i'$ is a partial function with domain

$$\mathrm{dom}(\theta_i^{-1} \circ f^{-1} \circ \theta_i') = \{x \in \Omega_i \mid \theta_i'(x) \in f \circ \theta_i(\Omega_i)\}.$$

The composition $\theta_j'^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'$ is also a partial function, and we let

$$W_{ij} = \Omega_{ij} \cap \mathrm{dom}(\theta_j'^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i'), \qquad \rho_i = (\theta_i'^{-1} \circ f \circ \theta_i) \restriction W_{ij}$$

and $W'_{ij} = \rho_i(W_{ij})$. Observe that $\theta_j \circ \varphi_{ji} = \theta_j \circ \theta_j^{-1} \circ \theta_i = \theta_i$, that is,

$$\theta_i = \theta_j \circ \varphi_{ji}.$$

Using this, on W_{ij} we get

$$\begin{aligned} \rho_j \circ \varphi_{ji} \circ \rho_i^{-1} &= \theta_j'^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ (\theta_i'^{-1} \circ f \circ \theta_i)^{-1} \\ &= \theta_j'^{-1} \circ f \circ \theta_j \circ \varphi_{ji} \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i' \\ &= \theta_j'^{-1} \circ f \circ \theta_i \circ \theta_i^{-1} \circ f^{-1} \circ \theta_i' \\ &= \theta_j'^{-1} \circ \theta_i' = \varphi'_{ji}, \end{aligned}$$

as claimed. The last part of the proposition is clear. □

Proposition 8.2 suggests defining a notion of equivalence on sets of gluing data which yields a converse of this proposition.

Definition 8.2. Two sets of gluing data $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}(\varphi_{ji})_{(i,j)\in K})$ and $\mathcal{G}' = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}(\varphi'_{ji})_{(i,j)\in K})$ over the same sets of Ω_i's and Ω_{ij}'s are *equivalent* iff there is a family of C^k bijections $(\rho_i \colon \Omega_i \to \Omega_i)_{i\in I}$, such that $\rho_i(\Omega_{ij}) = \Omega_{ij}$ and

$$\varphi'_{ji}(x) = \rho_j \circ \varphi_{ji} \circ \rho_i^{-1}(x), \qquad \text{for all} \quad x \in \Omega_{ij},$$

for all i, j. See Figure 8.12.

Here is the converse of Proposition 8.2. It is actually nicer than Proposition 8.2, because we can take $W_{ij} = W'_{ij} = \Omega_i$.

Proposition 8.3. *If two sets of gluing data $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}(\varphi_{ji})_{(i,j)\in K})$ and $\mathcal{G}' = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}(\varphi'_{ji})_{(i,j)\in K})$ are equivalent, then there is a C^k isomorphism $f \colon M_{\mathcal{G}} \to M_{\mathcal{G}'}$ between the manifolds induced by $\mathcal{G}$ and $\mathcal{G}'$. Furthermore, $f \circ \tau_i = \tau'_i \circ \rho_i$, for all $i \in I$.*

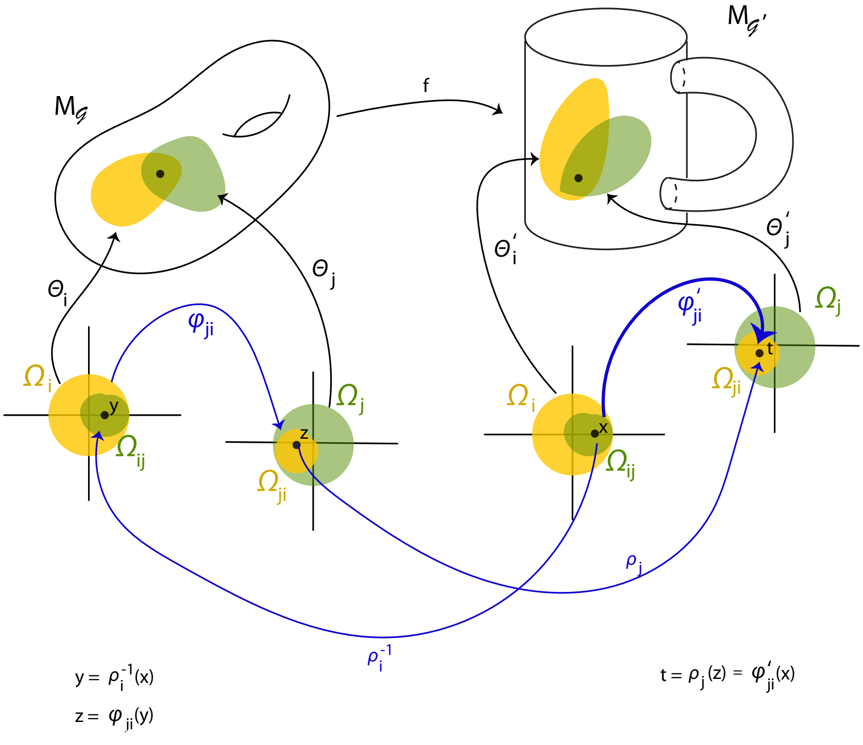


Figure 8.12: The equivalence between the two sets of gluing data $\mathcal{G}$ and $\mathcal{G}'$.

Proof. Let $f_i \colon \tau_i(\Omega_i) \to \tau'_i(\Omega_i)$ be the C^k bijection given by

$$f_i = \tau'_i \circ \rho_i \circ \tau_i^{-1},$$

where the $\rho_i \colon \Omega_i \to \Omega_i$'s are the maps giving the equivalence of $\mathcal{G}$ and $\mathcal{G}'$. If we prove that f_i and f_j agree on the overlap $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$, then the f_i patch and yield a C^k isomorphism $f \colon M_{\mathcal{G}} \to M_{\mathcal{G}'}$. The conditions of Proposition 8.2 imply that

$$\varphi'_{ji} \circ \rho_i = \rho_j \circ \varphi_{ji},$$

and we know that

$$\tau'_i = \tau'_j \circ \varphi'_{ji}.$$

Consequently, for every $[x] \in \tau_j(\Omega_{ji}) = \tau_i(\Omega_{ij})$ with $x \in \Omega_{ij}$, we have

$$\begin{aligned} f_j([x]) &= \tau_j' \circ \rho_j \circ \tau_j^{-1}([x]) \\ &= \tau_j' \circ \rho_j \circ \tau_j^{-1}([\varphi_{ji}(x)]) \\ &= \tau_j' \circ \rho_j \circ \varphi_{ji}(x) \\ &= \tau_j' \circ \varphi_{ji}' \circ \rho_i(x) \\ &= \tau_i' \circ \rho_i(x) \\ &= \tau_i' \circ \rho_i \circ \tau_i^{-1}([x]) \\ &= f_i([x]), \end{aligned}$$

which shows that f_i and f_j agree on $\tau_i(\Omega_i) \cap \tau_j(\Omega_j)$, as claimed. □

In the next section we describe a class of spaces that can be defined by gluing data and parametrization functions θ_i that are not necessarily injective. Roughly speaking, the gluing data specify the topology and the parametrizations define the geometry of the space. Such spaces have more structure than spaces defined parametrically but they are not quite manifolds. Yet they arise naturally in practice and they are the basis of efficient implementations of very good approximations of 3D meshes.

8.2 Parametric Pseudo-Manifolds

In practice it is often desirable to specify some n-dimensional geometric shape as a subset of $\mathbb{R}^d$ (usually for $d = 3$) in terms of parametrizations which are functions θ_i from some subset of $\mathbb{R}^n$ into $\mathbb{R}^d$ (usually, $n = 2$). For "open" shapes, this is reasonably well understood, but dealing with a "closed" shape is a lot more difficult because the parametrized pieces should overlap as smoothly as possible, and this is hard to achieve. Furthermore, in practice, the parametrization functions θ_i may not be injective. Proposition 8.1 suggests various ways of defining such geometric shapes. For the lack of a better term, we will call these shapes, *parametric pseudo-manifolds*.

Definition 8.3. Let n, k, d be three integers with $d > n \geq 1$ and $k \geq 1$ or $k = \infty$. A *parametric C^k pseudo-manifold of dimension n in $\mathbb{R}^d$* is a pair $\mathcal{M} = (\mathcal{G}, (\theta_i)_{i \in I})$, where $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j) \in I \times I}, (\varphi_{ji})_{(i,j) \in K})$ is a set of gluing data for some finite set I, and each θ_i is a C^k function $\theta_i \colon \Omega_i \to \mathbb{R}^d$ called a *parametrization*, such that the following property holds:

(C) For all $(i, j) \in K$, we have
$$\theta_i = \theta_j \circ \varphi_{ji}.$$

For short we use terminology *parametric pseudo-manifold*. The subset $M \subseteq \mathbb{R}^d$ given by

$$M = \bigcup_{i \in I} \theta_i(\Omega_i)$$

is called the *image* of the parametric pseudo-manifold $\mathcal{M}$. When $n = 2$ and $d = 3$, we say that $\mathcal{M}$ is a *parametric pseudo-surface*.

Condition (C) obviously implies that

$$\theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}),$$

for all $(i, j) \in K$. Consequently, θ_i and θ_j are consistent parametrizations of the overlap $\theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji})$. The shape M is covered by pieces $U_i = \theta_i(\Omega_i)$ not necessarily open, with each U_i parametrized by θ_i, and where the overlapping pieces $U_i \cap U_j$, are parametrized consistently. The local structure of M is given by the θ_i's, and the global structure is given by the gluing data. We recover a manifold if we require the θ_i to be bijective and to satisfy the following additional conditions:

(C') For all $(i, j) \in K$,

$$\theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \theta_i(\Omega_{ij}) = \theta_j(\Omega_{ji}).$$

(C") For all $(i, j) \notin K$,

$$\theta_i(\Omega_i) \cap \theta_j(\Omega_j) = \emptyset.$$

Even if the θ_i's are not injective, Properties (C') and (C") would be desirable since they guarantee that $\theta_i(\Omega_i - \Omega_{ij})$ and $\theta_j(\Omega_j - \Omega_{ji})$ are parametrized uniquely. Unfortunately, these properties are difficult to enforce. Observe that any manifold induced by $\mathcal{G}$ is the image of a parametric pseudo-manifold.

Although this is an abuse of language, it is more convenient to call M a parametric pseudo-manifold, or even a *pseudo-manifold*.

We can also show that the parametric pseudo-manifold M is the image in $\mathbb{R}^d$ of the abstract manifold $M_{\mathcal{G}}$.

Proposition 8.4. *Let $\mathcal{M} = (\mathcal{G}, (\theta_i)_{i\in I})$ be parametric C^k pseudo-manifold of dimension n in $\mathbb{R}^d$, where $\mathcal{G} = ((\Omega_i)_{\in I}, (\Omega_{ij})_{(i,j)\in I\times I}, (\varphi_{ji})_{(i,j)\in K})$ is a set of gluing data for some finite set I. Then the parametrization maps θ_i induce a surjective map $\Theta\colon M_{\mathcal{G}} \to M$ from the abstract manifold $M_{\mathcal{G}}$ specified by $\mathcal{G}$ to the image $M \subseteq \mathbb{R}^d$ of the parametric pseudo-manifold $\mathcal{M}$, and the following property holds: for every Ω_i,*

$$\theta_i = \Theta \circ \tau_i,$$

where the $\tau_i\colon \Omega_i \to M_{\mathcal{G}}$ are the parametrization maps of the manifold $M_{\mathcal{G}}$ (see Proposition 8.1). In particular, every manifold M induced by the gluing data $\mathcal{G}$ is the image of $M_{\mathcal{G}}$ by a map $\Theta\colon M_{\mathcal{G}} \to M$.

Proof. Recall that

$$M_{\mathcal{G}} = \Big(\coprod_{i\in I} \Omega_i\Big) / \sim,$$

where $\sim$ is the equivalence relation defined so that, for all $x, y \in \coprod_{i\in I} \Omega_i$,

$$x \sim y \quad \text{iff} \quad (\exists (i,j) \in K)(x \in \Omega_{ij},\ y \in \Omega_{ji},\ y = \varphi_{ji}(x)).$$

The proof of Proposition 8.1 also showed that $\tau_i(\Omega_i) \cap \tau_j(\Omega_j) \neq \emptyset$ iff $(i,j) \in K$, and if so,

$$\tau_i(\Omega_i) \cap \tau_j(\Omega_j) = \tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji}).$$

In particular,

$$\tau_i(\Omega_i - \Omega_{ij}) \cap \tau_j(\Omega_j - \Omega_{ji}) = \emptyset$$

for all $(i,j) \in I \times I$ ($\Omega_{ij} = \Omega_{ji} = \emptyset$ when $(i,j) \notin K$). These properties with the fact that the τ_i's are injections show that for all $(i,j) \notin K$, we can define $\Theta_i \colon \tau_i(\Omega_i) \to \mathbb{R}^d$ and $\Theta_j \colon \tau_i(\Omega_j) \to \mathbb{R}^d$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_i \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_j.$$

For $(i,j) \in K$, as the the τ_i's are injections we can define $\Theta_i \colon \tau_i(\Omega_i - \Omega_{ij}) \to \mathbb{R}^d$ and $\Theta_j \colon \tau_i(\Omega_j - \Omega_{ji}) \to \mathbb{R}^d$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_i - \Omega_{ij} \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_j - \Omega_{ji}.$$

It remains to define Θ_i on $\tau_i(\Omega_{ij})$ and Θ_j on $\tau_j(\Omega_{ji})$ in such a way that they agree on $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$. However, Condition (C) in Definition 8.3 says that for all $x \in \Omega_{ij}$,

$$\theta_i(x) = \theta_j(\varphi_{ji}(x)).$$

Consequently, if we define Θ_i on $\tau_i(\Omega_{ij})$ and Θ_j on $\tau_j(\Omega_{ji})$ by

$$\Theta_i([x]) = \theta_i(x),\ x \in \Omega_{ij}, \qquad \Theta_j([y]) = \theta_j(y),\ y \in \Omega_{ji},$$

as $x \sim \varphi_{ji}(x)$, we have

$$\Theta_i([x]) = \theta_i(x) = \theta_j(\varphi_{ji}(x)) = \Theta_j([\varphi_{ji}(x)]) = \Theta_j([x]),$$

which means that Θ_i and Θ_j agree on $\tau_i(\Omega_{ij}) = \tau_j(\Omega_{ji})$. But then the functions Θ_i agree whenever their domains overlap, and so they patch to yield a function Θ with domain $M_{\mathcal{G}}$ and image M. By construction, $\theta_i = \Theta \circ \tau_i$, and as a manifold induced by $\mathcal{G}$ is a parametric pseudo-manifold, the last statement is obvious. □

The function $\Theta \colon M_{\mathcal{G}} \to M$ given by Proposition 8.4 shows how the parametric pseudo-manifold M differs from the abstract manifold $M_{\mathcal{G}}$. As we said before, a practical method for approximating 3D meshes based on parametric pseudo surfaces is described in Siqueira, Xu and Gallier [108].

Chapter 9

Vector Fields, Lie Derivatives, Integral Curves, Flows

Our goal in this chapter is to generalize the concept of a vector field to manifolds and to promote some standard results about ordinary differential equations to manifolds.

9.1 Tangent and Cotangent Bundles

Let M be a C^k-manifold (with $k \geq 2$). Roughly speaking, a vector field on M is the assignment $p \mapsto X(p)$, of a tangent vector $X(p) \in T_p(M)$, to a point $p \in M$. Generally, we would like such assignments to have some smoothness properties when p varies in M, for example, to be C^l, for some l related to k. If the collection $T(M)$ of all tangent spaces $T_p(M)$ was a C^l-manifold, then it would be very easy to define what we mean by a C^l-vector field: we would simply require the map $X\colon M \to T(M)$ to be C^l.

If M is a C^k-manifold of dimension n, then we can indeed make $T(M)$ into a C^{k-1}-manifold of dimension $2n$, and we now sketch this construction.

We find it most convenient to use Version 2 of the definition of tangent vectors, i.e., as equivalence classes of triples (U, φ, x), where (U, φ) is a chart at p and $x \in \mathbb{R}^n$. Recall that (U, φ, x) and (V, ψ, y) are equivalent iff

$$(\psi \circ \varphi^{-1})'_{\varphi(p)}(x) = y.$$

First we let $T(M)$ be the disjoint union of the tangent spaces $T_p(M)$, for all $p \in M$. Formally,

$$T(M) = \{(p, v) \mid p \in M, v \in T_p(M)\}.$$

See Figure 9.1.

There is a *natural projection*

$$\pi\colon T(M) \to M, \quad \text{with} \quad \pi(p, v) = p.$$

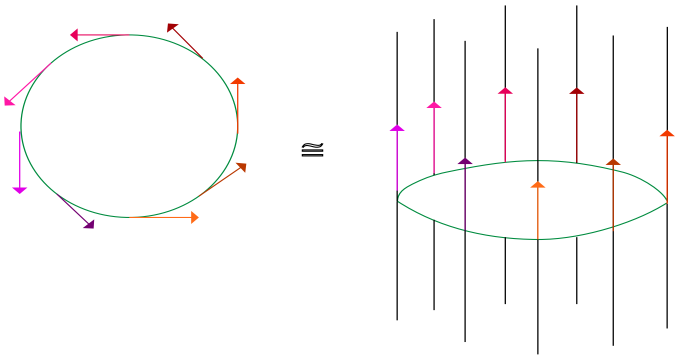


Figure 9.1: The tangent bundle of S^1.

We still have to give $T(M)$ a topology and to define a C^{k-1}-atlas. For every chart (U, φ) of M (with U open in M), we define the function $\widetilde{\varphi}\colon \pi^{-1}(U) \to \mathbb{R}^{2n}$, by

$$\widetilde{\varphi}(p, v) = (\varphi(p), \theta^{-1}_{U,\varphi,p}(v)),$$

where $(p, v) \in \pi^{-1}(U)$ and $\theta_{U,\varphi,p}$ is the isomorphism between $\mathbb{R}^n$ and $T_p(M)$ described just after Definition 7.13. It is obvious that $\widetilde{\varphi}$ is a bijection between $\pi^{-1}(U)$ and $\varphi(U) \times \mathbb{R}^n$, an open subset of $\mathbb{R}^{2n}$. See Figure 9.2.

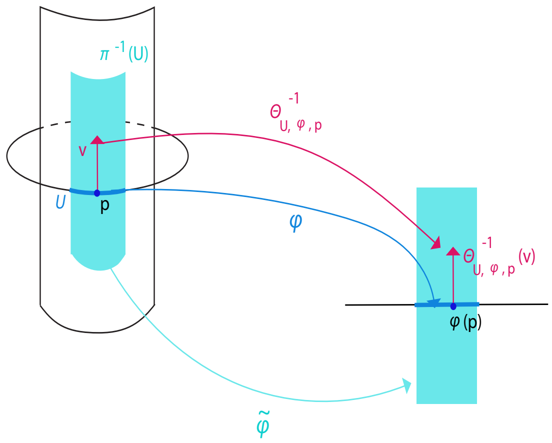


Figure 9.2: A chart for $T(S^1)$.

We give $T(M)$ the weakest topology that makes all the $\widetilde{\varphi}$ continuous, i.e., we take the collection of subsets of the form $\widetilde{\varphi}^{-1}(W)$, where W is any open subset of $\varphi(U) \times \mathbb{R}^n$, as a basis of the topology of $T(M)$. One may check that $T(M)$ is Hausdorff and second-countable in this topology. If (U, φ) and (V, ψ) are two overlapping charts of M, then the definition of the equivalence relation on triples (U, φ, x) and (V, ψ, y) immediately implies that

$$\theta^{-1}_{(V,\psi,p)} \circ \theta_{(U,\varphi,p)} = (\psi \circ \varphi^{-1})'_z$$

for all $p \in U \cap V$, with $z = \varphi(p)$, so the transition map,

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1} \colon \varphi(U \cap V) \times \mathbb{R}^n \longrightarrow \psi(U \cap V) \times \mathbb{R}^n$$

is given by

$$\widetilde{\psi} \circ \widetilde{\varphi}^{-1}(z, x) = (\psi \circ \varphi^{-1}(z), (\psi \circ \varphi^{-1})'_z(x)), \qquad (z, x) \in \varphi(U \cap V) \times \mathbb{R}^n.$$

It is clear that $\widetilde{\psi} \circ \widetilde{\varphi}^{-1}$ is a C^{k-1}-map.

Definition 9.1. The space $T(M)$ resulting from the previous construction from a C^k $(k \geq 2)$ manifold M is a C^{k-1}-manifold of dimension $2n$ called the *tangent bundle* of M.

Remark: Even if the manifold M is naturally embedded in $\mathbb{R}^N$ (for some $N \geq n = \dim(M)$), it is not at all obvious how to view the tangent bundle $T(M)$ as embedded in $\mathbb{R}^{N'}$, for some suitable N'. Hence, we see that the definition of an abstract manifold is unavoidable.

A similar construction can be carried out for the cotangent bundle. In this case, we let $T^*(M)$ be the disjoint union of the cotangent spaces $T^*_p(M)$, that is,

$$T^*(M) = \{(p, \omega) \mid p \in M, \omega \in T^*_p(M)\}.$$

We also have a natural projection $\pi \colon T^*(M) \to M$ with $\pi(p, \omega) = p$, and we can define charts in several ways. One method used by Warner [114] goes as follows: for any chart, (U, φ), on M, we define the function,
$\widetilde{\varphi} \colon \pi^{-1}(U) \to \mathbb{R}^{2n}$, by

$$\widetilde{\varphi}(p, \omega) = \left(\varphi(p), \omega\left(\left(\frac{\partial}{\partial x_1}\right)_p\right), \ldots, \omega\left(\left(\frac{\partial}{\partial x_n}\right)_p\right)\right),$$

where $(p, \omega) \in \pi^{-1}(U)$ and the $\left(\frac{\partial}{\partial x_i}\right)_p$ are the basis of $T_p(M)$ associated with the chart (U, φ). Again, one can make $T^*(M)$ into a C^{k-1}-manifold of dimension $2n$, called the *cotangent bundle* We leave the details as an exercise to the reader (or look at Berger and Gostiaux [15]).

Another way of obtaining the manifold structure of $T^*(M)$ is as follows. For each chart (U, φ) on M, we obtain a chart

$$\widetilde{\varphi}^* \colon \pi^{-1}(U) \to \varphi(U) \times \mathbb{R}^n \subseteq \mathbb{R}^{2n}$$

on $T^*(M)$ given by

$$\widetilde{\varphi}^*(p,\omega) = (\varphi(p), \theta^*_{U,\varphi,p}(\omega))$$

for all $(p,\omega) \in \pi^{-1}(U)$, where

$$\theta^*_{U,\varphi,p} = \iota \circ \theta^\top_{U,\varphi,p} \colon T^*_p(M) \to \mathbb{R}^n.$$

Here, $\theta^\top_{U,\varphi,p} \colon T^*_p(M) \to (\mathbb{R}^n)^*$ is obtained by dualizing the map, $\theta_{U,\varphi,p} \colon \mathbb{R}^n \to T_p(M)$ and $\iota \colon (\mathbb{R}^n)^* \to \mathbb{R}^n$ is the isomorphism induced by the canonical basis $(e_1, \ldots, e_n)$ of $\mathbb{R}^n$ and its dual basis. Recall that the transpose $\theta^\top_{U,\varphi,p}$ of the linear map $\theta_{U,\varphi,p}$ is given by $\theta^\top_{U,\varphi,p}(\omega) = \omega \circ \theta_{U,\varphi,p}$, for every linear form $\omega \colon T_p(M) \to \mathbb{R}$ in $T^*_p(M)$.

For simplicity of notation, we also use the notation TM for $T(M)$ (resp. T^*M for $T^*(M)$).

Observe that for every chart (U,φ) on M, there is a bijection

$$\tau_U \colon \pi^{-1}(U) \to U \times \mathbb{R}^n,$$

given by

$$\tau_U(p,v) = (p, \theta^{-1}_{U,\varphi,p}(v)).$$

Clearly, $pr_1 \circ \tau_U = \pi$ on $\pi^{-1}(U)$, as illustrated by the following commutative diagram.

$$\begin{tikzcd}
\pi^{-1}(U) \arrow[rr, "\tau_U"] \arrow[dr, "\pi"'] & & U \times \mathbb{R}^n \arrow[dl, "pr_1"] \\
& U &
\end{tikzcd}$$

Thus locally, that is over U, the bundle $T(M)$ looks like the product manifold $U \times \mathbb{R}^n$. We say that $T(M)$ is *locally trivial* (over U) and we call τ_U a *trivializing map*. For any $p \in M$, the vector space $\pi^{-1}(p) = \{p\} \times T_p(M) \cong T_p(M)$ is called the *fibre above p*. Observe that the restriction of τ_U to $\pi^{-1}(p)$ is a linear isomorphism between $\{p\} \times T_p(M) \cong T_p(M)$ and $\{p\} \times \mathbb{R}^n \cong \mathbb{R}^n$, for any $p \in M$. Furthermore, for any two overlapping charts (U,φ) and (V,ψ), there is a function $g_{UV} \colon U \cap V \to \mathbf{GL}(n,\mathbb{R})$ such that

$$(\tau_U \circ \tau_V^{-1})(p,x) = (p, g_{UV}(p)(x))$$

for all $p \in U \cap V$ and all $x \in \mathbb{R}^n$, with $g_{UV}(p)$ given by

$$g_{UV}(p) = (\varphi \circ \psi^{-1})'_{\psi(p)}.$$

Obviously, $g_{UV}(p)$ is a linear isomorphism of $\mathbb{R}^n$ for all $p \in U \cap V$. The maps $g_{UV}(p)$ are called the *transition functions* of the tangent bundle.

For example, if $M = S^n$, the n-sphere in $\mathbb{R}^{n+1}$, we have two charts given by the stereographic projection (U_N, σ_N) from the north pole, and the stereographic projection (U_S, σ_S)

from the south pole (with $U_N = S^n - \{N\}$ and $U_S = S^n - \{S\}$), and on the overlap, $U_N \cap U_S = S^n - \{N, S\}$, the transition maps

$$\mathcal{I} = \sigma_S \circ \sigma_N^{-1} = \sigma_N \circ \sigma_S^{-1}$$

defined on $\varphi_N(U_N \cap U_S) = \varphi_S(U_N \cap U_S) = \mathbb{R}^n - \{0\}$, are given by

$$(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2} (x_1, \ldots, x_n);$$

that is, the inversion $\mathcal{I}$ of center $O = (0, \ldots, 0)$ and power 1. We leave it as an exercise to prove that for every point $u \in \mathbb{R}^n - \{0\}$, we have

$$d\mathcal{I}_u(h) = \|u\|^{-2} \left(h - 2\frac{\langle u, h\rangle}{\|u\|^2}\, u \right),$$

the composition of the hyperplane reflection about the hyperplane $u^\perp \subseteq \mathbb{R}^n$ with the magnification of center O and ratio $\|u\|^{-2}$. (Hint: Write $\mathcal{I}(u) = u/\|u\|^2$ and compute $\mathcal{I}(u+h) - \mathcal{I}(u)$.) This is a *similarity transformation*. Therefore, the transition function g_{NS} (defined on $U_N \cap U_S$) of the tangent bundle TS^n is given by

$$g_{NS}(p)(h) = \|\sigma_S(p)\|^{-2} \left(h - 2\frac{\langle \sigma_S(p), h\rangle}{\|\sigma_S(p)\|^2}\, \sigma_S(p) \right).$$

All these ingredients are part of being a *vector bundle*. For more on bundles, Lang [75], Gallot, Hulin and Lafontaine [49], Lafontaine [72], or Bott and Tu [18].

When $M = \mathbb{R}^n$, observe that $T(M) = M \times \mathbb{R}^n = \mathbb{R}^n \times \mathbb{R}^n$, i.e., the bundle $T(M)$ is (globally) trivial.

Given a C^k-map $h\colon M \to N$ between two C^k-manifolds, we can define the function $dh\colon T(M) \to T(N)$ (also denoted Th, or h_*, or Dh), by setting

$$dh(u) = dh_p(u), \quad \text{iff} \quad u \in T_p(M).$$

We leave the next proposition as an exercise to the reader. (A proof can be found in Berger and Gostiaux [15].)

Proposition 9.1. *Given a C^k-map $h\colon M \to N$ between two C^k-manifolds M and N (with $k \geq 1$), the map $dh\colon T(M) \to T(N)$ is a C^{k-1} map.*

We are now ready to define vector fields.

9.2 Vector Fields, Lie Derivative

In Section 11.3 we introduced the notion of a vector field in $\mathbb{R}^n$. We now generalize the notion of a vector field to a manifold. Let M be a C^{k+1} manifold. A C^k-vector field on M is an assignment $p \mapsto X(p)$ of a tangent vector $X(p) \in T_p(M)$ to a point $p \in M$, so that $X(p)$ varies in a C^k-fashion in terms of p. This notion is captured rigorously by the following definition.

Definition 9.2. Let M be a C^{k+1} manifold, with $k \geq 1$. For any open subset U of M, a *vector field on U* is any *section X of $T(M)$ over U*, that is, any function $X\colon U \to T(M)$, such that $\pi \circ X = \mathrm{id}_U$ (i.e., $X(p) \in T_p(M)$, for every $p \in U$). We also say that X is a *lifting of U into $T(M)$*. We say that X is a *C^k-vector field on U* iff X is a section over U and a C^k-map. The set of C^k-vector fields over U is denoted $\Gamma^{(k)}(U, T(M))$; see Figure 9.3. Given a curve, $\gamma\colon [a,b] \to M$, a *vector field X along γ* is any section of $T(M)$ over γ, i.e., a C^k-function $X\colon [a,b] \to T(M)$, such that $\pi \circ X = \gamma$. We also say that X *lifts γ into $T(M)$*.

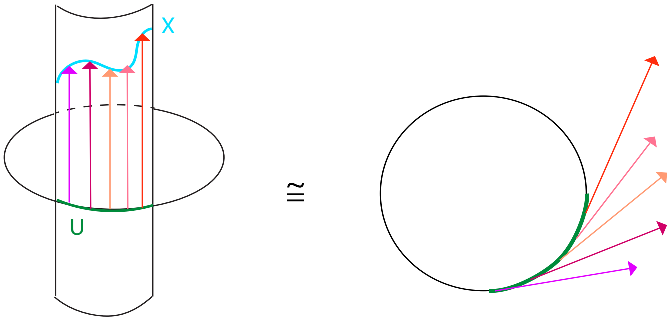


Figure 9.3: A vector field on S^1 represented as the section X in $T(S^1)$.

Clearly, $\Gamma^{(k)}(U, T(M))$ is a real vector space. For short, the space $\Gamma^{(k)}(M, T(M))$ is also denoted by $\Gamma^{(k)}(T(M))$ (or $\mathfrak{X}^{(k)}(M)$, or even $\Gamma(T(M))$ or $\mathfrak{X}(M)$).

Remark: We can also define a *C^j-vector field on U* as a section, X, over U which is a C^j-map, where $0 \leq j \leq k$. Then we have the vector space $\Gamma^{(j)}(U, T(M))$, *etc.*

If $M = \mathbb{R}^n$ and U is an open subset of M, then $T(M) = \mathbb{R}^n \times \mathbb{R}^n$ and a section of $T(M)$ over U is simply a function, X, such that

$$X(p) = (p, u), \quad \text{with} \quad u \in \mathbb{R}^n,$$

for all $p \in U$. In other words, X is defined by a function, $f\colon U \to \mathbb{R}^n$ (namely, $f(p) = u$). This corresponds to the "old" definition of a vector field in the more basic case where the manifold, M, is just $\mathbb{R}^n$.

For any vector field $X \in \Gamma^{(k)}(U, T(M))$ and for any $p \in U$, we have $X(p) = (p, v)$ for some $v \in T_p(M)$, and it is convenient to denote the vector v by X_p so that $X(p) = (p, X_p)$.

In fact, in most situations it is convenient to identify $X(p)$ with $X_p \in T_p(M)$, and we will do so from now on. This amounts to identifying the isomorphic vector spaces $\{p\} \times T_p(M)$ and $T_p(M)$, which we always do. Let us illustrate the advantage of this convention with the next definition.

Given any C^k-function $f \in \mathcal{C}^k(U)$ and a vector field $X \in \Gamma^{(k)}(U, T(M))$, we define the vector field fX by

$$(fX)_p = f(p)X_p, \quad p \in U.$$

Obviously, $fX \in \Gamma^{(k)}(U, T(M))$, which shows that $\Gamma^{(k)}(U, T(M))$ is also a $\mathcal{C}^k(U)$-module.

Definition 9.3. For any chart (U, φ) on M it is easy to check that the map

$$p \mapsto \left(\frac{\partial}{\partial x_i}\right)_p, \quad p \in U,$$

is a C^k-vector field on U (with $1 \leq i \leq n$). This vector field is denoted $\left(\frac{\partial}{\partial x_i}\right)$ or $\frac{\partial}{\partial x_i}$.

Definition 9.4. Let M be a C^{k+1} manifold and let X be a C^k vector field on M. If U is any open subset of M and f is any function in $\mathcal{C}^k(U)$, then the *Lie derivative of f with respect to X*, denoted $X(f)$ or $L_X f$, is the function on U given by

$$X(f)(p) = X_p(f) = X_p(\mathbf{f}), \qquad p \in U.$$

In particular, if (U, φ) is any chart at p and $X_p = \sum_{i=1}^n \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p$, then

$$X_p(f) = \sum_{i=1}^n \lambda_i \left(\frac{\partial}{\partial x_i}\right)_p f.$$

Observe that

$$X(f)(p) = df_p(X_p),$$

where df_p is identified with the linear form in $T_p^*(M)$ defined by

$$df_p(v) = v(\mathbf{f}), \qquad v \in T_pM,$$

by identifying $T_{t_0}\mathbb{R}$ with $\mathbb{R}$ (see the discussion following Proposition 7.13). The Lie derivative, $L_X f$, is also denoted $X[f]$.

As a special case, when (U, φ) is a chart on M, the vector field, $\frac{\partial}{\partial x_i}$, just defined above induces the function

$$p \mapsto \left(\frac{\partial}{\partial x_i}\right)_p f, \quad p \in U,$$

denoted $\frac{\partial}{\partial x_i}(f)$ or $\left(\frac{\partial}{\partial x_i}\right) f$.

It is easy to check that $X(f) \in \mathcal{C}^{k-1}(U)$. As a consequence, every vector field $X \in \Gamma^{(k)}(U, T(M))$ induces a linear map,

$$L_X \colon \mathcal{C}^k(U) \longrightarrow \mathcal{C}^{k-1}(U),$$

given by $f \mapsto X(f)$. It is immediate to check that L_X has the Leibniz property, i.e.,

$$L_X(fg) = L_X(f)g + fL_X(g).$$

Linear maps with this property are called *derivations*. Thus, we see that every vector field induces some kind of differential operator, namely, a linear derivation. Unfortunately, not every linear derivation of the above type arises from a vector field, although this turns out to be true in the smooth case i.e., when $k = \infty$ (for a proof, see Gallot, Hulin and Lafontaine [49] or Lafontaine [72]).

In the rest of this section, unless stated otherwise, we assume that $k \geq 1$. The following easy proposition holds (c.f. Warner [114]).

Proposition 9.2. *Let X be a vector field on the C^{k+1}-manifold M, of dimension n. Then the following are equivalent:*

(a) X is C^k.

(b) If (U, φ) is a chart on M and if $f_1, \ldots, f_n$ are the functions on U uniquely defined by

$$X \restriction U = \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i},$$

then each f_i is a C^k-map.

(c) Whenever U is open in M and $f \in \mathcal{C}^k(U)$, then $X(f) \in \mathcal{C}^{k-1}(U)$.

Given any two C^k-vector field X, Y on M, for any function $f \in \mathcal{C}^k(M)$, we defined above the function $X(f)$ and $Y(f)$. Thus, we can form $X(Y(f))$ (resp. $Y(X(f))$), which are in $\mathcal{C}^{k-2}(M)$. Unfortunately, even in the smooth case, there is generally *no* vector field Z such that

$$Z(f) = X(Y(f)), \quad \text{for all } f \in \mathcal{C}^k(M).$$

This is because $X(Y(f))$ (and $Y(X(f))$) involve second-order derivatives. However, if we consider $X(Y(f)) - Y(X(f))$, then second-order derivatives cancel out and there is a unique vector field inducing the above differential operator. Intuitively, $XY - YX$ measures the "failure of X and Y to commute."

Proposition 9.3. *Given any C^{k+1}-manifold M, of dimension n, for any two C^k-vector fields X, Y on M, there is a unique C^{k-1}-vector field $[X, Y]$, such that*

$$[X, Y](f) = X(Y(f)) - Y(X(f)), \quad \textit{for all} \quad f \in \mathcal{C}^{k-1}(M).$$

Proof. First we prove uniqueness. For this it is enough to prove that $[X, Y]$ is uniquely defined on $\mathcal{C}^k(U)$, where (U, φ) is a chart over U. For this chart, we know that

$$X = \sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i} \quad \text{and} \quad Y = \sum_{i=1}^{n} Y_i \frac{\partial}{\partial x_i},$$

where $X_i, Y_i \in \mathcal{C}^k(U)$. Then for any $f \in \mathcal{C}^k(M)$, we have

$$\begin{aligned} X(Y(f)) &= X\left(\sum_{j=1}^{n} Y_j \frac{\partial}{\partial x_j}(f)\right) = \sum_{i,j=1}^{n} X_i \frac{\partial}{\partial x_i}(Y_j) \frac{\partial}{\partial x_j}(f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j}(f) \\ Y(X(f)) &= Y\left(\sum_{i=1}^{n} X_i \frac{\partial}{\partial x_i}(f)\right) = \sum_{i,j=1}^{n} Y_j \frac{\partial}{\partial x_j}(X_i) \frac{\partial}{\partial x_i}(f) + \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i}(f). \end{aligned}$$

However, as $f \in \mathcal{C}^k(M)$, with $k \geq 2$, we have

$$\sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_j \partial x_i}(f) = \sum_{i,j=1}^{n} X_i Y_j \frac{\partial^2}{\partial x_i \partial x_j}(f),$$

and we deduce that

$$X(Y(f)) - Y(X(f)) = \sum_{i,j=1}^{n} \left(X_i \frac{\partial}{\partial x_i}(Y_j) - Y_i \frac{\partial}{\partial x_i}(X_j)\right) \frac{\partial}{\partial x_j}(f).$$

This proves that $[X, Y] = XY - YX$ is uniquely defined on U and that it is C^{k-1}. Thus, if $[X, Y]$ exists, it is unique.

To prove existence, we use the above expression to define $[X, Y]_U$, locally on U, for every chart, (U, φ). On any overlap, $U \cap V$, by the uniqueness property that we just proved, $[X, Y]_U$ and $[X, Y]_V$ must agree. Then we can define the vector field $[X, Y]$ as follows: for every chart (U, φ), the restriction $[X, Y]$ to U is equal to $[X, Y]_U$. This well defined because whenever two charts with domains U and V overlap, we know that $[X, Y]_U = [X, Y]_V$ agree. Therefore, $[X, Y]$ is a C^{k-1}-vector field defined on the whole of M. □

Definition 9.5. Given any C^{k+1}-manifold M, of dimension n, for any two C^k-vector fields X, Y on M, the *Lie bracket* $[X, Y]$ of X and Y, is the C^{k-1} vector field defined so that

$$[X, Y](f) = X(Y(f)) - Y(X(f)), \quad \text{for all} \quad f \in \mathcal{C}^{k-1}(M).$$

An an example in $\mathbb{R}^3$, if X and Y are the two vector fields,

$$X = \frac{\partial}{\partial x} + y \frac{\partial}{\partial z} \quad \text{and} \quad Y = \frac{\partial}{\partial y},$$

then to compute $[X, Y]$, set $g = Y(f) = \frac{\partial f}{\partial y}$ and observe that

$$X(Y(f)) = X(g) = \frac{\partial g}{\partial x} + y\frac{\partial g}{\partial z} = \frac{\partial^2 f}{\partial x \partial y} + y\frac{\partial^2 f}{\partial z \partial y}.$$

Next set $h = X(f) = \frac{\partial f}{\partial x} + y\frac{\partial f}{\partial z}$ and calculate

$$Y(X(f)) = Y(h) = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial z}\right) = \frac{\partial^2 f}{\partial y \partial x} + \frac{\partial f}{\partial z} + y\frac{\partial^2 f}{\partial y \partial z}.$$

Then

$$\begin{aligned}[X, Y](f) &= X(Y(f)) - Y(X(f)) \\ &= \frac{\partial^2 f}{\partial x \partial y} + y\frac{\partial^2 f}{\partial z \partial y} - \frac{\partial^2 f}{\partial y \partial x} - \frac{\partial f}{\partial z} - y\frac{\partial^2 f}{\partial y \partial z} \\ &= -\frac{\partial f}{\partial z}.\end{aligned}$$

Hence

$$[X, Y] = -\frac{\partial}{\partial z}.$$

We also have the following simple proposition whose proof is left as an exercise (or, see Do Carmo [39]).

Proposition 9.4. *Given any C^{k+1}-manifold M, of dimension n, for any C^k-vector fields X, Y, Z on M, for all $f, g \in \mathcal{C}^k(M)$, we have:*

(a) $[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0$ *(Jacobi identity).*

(b) $[X, X] = 0$.

(c) $[fX, gY] = fg[X, Y] + fX(g)Y - gY(f)X$.

(d) $[-, -]$ *is bilinear.*

As a consequence, for smooth manifolds ($k = \infty$), the space of vector fields $\Gamma^{(\infty)}(T(M))$ is a vector space equipped with a bilinear operation $[-, -]$ that satisfies the Jacobi identity. This makes $\Gamma^{(\infty)}(T(M))$ a *Lie algebra*.

Let $h\colon M \to N$ be a *diffeomorphism* between two manifolds. Then vector fields can be transported from N to M and conversely.

Definition 9.6. Let $h\colon M \to N$ be a diffeomorphism between two C^{k+1}-manifolds. For every C^k-vector field Y on N, the *pull-back of Y along h* is the vector field h^*Y on M, given by

$$(h^*Y)_p = dh^{-1}_{h(p)}(Y_{h(p)}), \qquad p \in M.$$

See Figure 9.4. For every C^k-vector field X on M, the *push-forward of X along h* is the vector field h_*X on N, given by

$$h_*X = (h^{-1})^*X,$$

that is, for every $p \in M$, $(h_*X)_{h(p)} = dh_p(X_p)$, or equivalently,

$$(h_*X)_q = dh_{h^{-1}(q)}(X_{h^{-1}(q)}), \qquad q \in N.$$

See Figure 9.5.

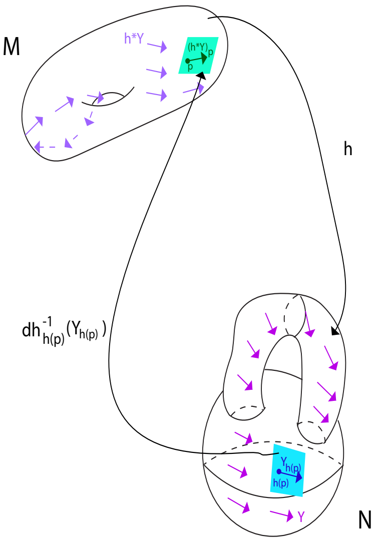


Figure 9.4: The pull-back of the vector field Y.

We have the following result.

Proposition 9.5. *For any diffeomorphism $h\colon M \to N$, for every C^k vector field X on M, we have*

$$L_{h_*X}f = L_X(f \circ h) \circ h^{-1},$$

for any function $f \in C^k(N)$.

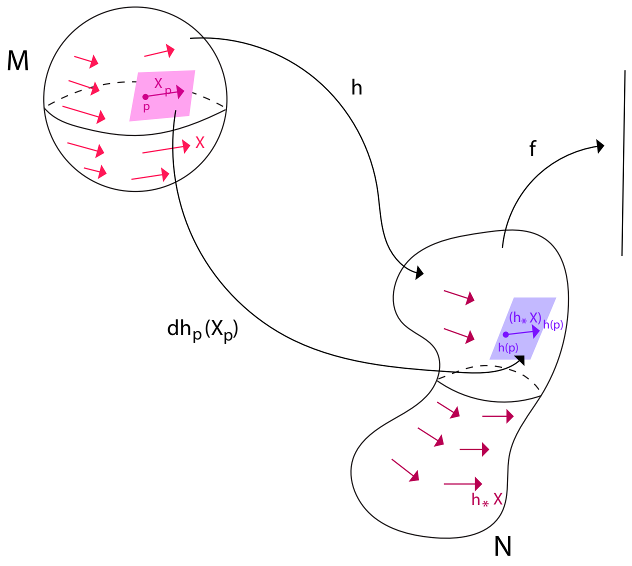


Figure 9.5: The push-forward of the vector field X.

Proof. We have

$$
\begin{aligned}
(L_{h_*X}f)_{h(p)} &= (h_*X)_{h(p)}(f) && \text{by Definition 9.4}\\
&= df_{h(p)}\left((h_*X)_{h(p)}\right), && \text{by remark after Definition 9.4}\\
&= df_{h(p)}\left(dh_p(X_p)\right), && \text{by Definition 9.6}\\
&= d(f\circ h)_p(X_p), && \text{by the chain rule}\\
&= d(f\circ h)_{h^{-1}(q)}(X_{h^{-1}(q)}), && p = h^{-1}(q)\\
&= X_{h^{-1}(q)}(f\circ h), && \text{by remark after Definition 9.4}\\
&= (L_X(f\circ h))_{h^{-1}(q)}
\end{aligned}
$$

as claimed. □

One more notion will be needed when we deal with Lie algebras.

Definition 9.7. Let $h\colon M \to N$ be a C^{k+1}-map of manifolds. If X is a C^k-vector field on M and Y is a C^k-vector field on N, we say that *X and Y are h-related* iff

$$dh \circ X = Y \circ h.$$

The basic result about h-related vector fields is:

Proposition 9.6. *Let $h\colon M \to N$ be a C^{k+1}-map of manifolds, let X and Y be C^k-vector fields on M and let X_1, Y_1 be C^k-vector fields on N. If X is h-related to X_1 and Y is h-related to Y_1, then $[X, Y]$ is h-related to $[X_1, Y_1]$.*

Proof. Basically, one needs to unwind the definitions; see Warner [114] (Chapter 1, Proposition 1.55). □

There is another way to characterize when two vector fields are h-related which is often more convenient than the definition. Recall that $X_p f = df_p(X_p)$ for any $f \in C^k(M)$ and any C^k-vector field X on M.

Proposition 9.7. *Let $h\colon M \to N$ be a C^{k+1}-map of manifolds, and let X be a C^k-vector fields on M and Y be a C^k-vector field on N. Then X and Y are h-related iff*

$$X(g \circ h) = Y g \circ h$$

for all $g \in C^k(N)$.

Proof. We have the following sequence of equivalences

$$\begin{aligned} X(g \circ h) &= Y g \circ h \\ X_p(g \circ h) &= (Y g)_{h(p)} && \text{for all } p \in M \\ d(g \circ h)_p(X_p) &= dg_{h(p)}(Y_{h(p)}) \\ dg_{h(p)}(dh_p(X_p)) &= dg_{h(p)}(Y_{h(p)}) && \text{by the chain rule} \\ (dh \circ X)_p g &= Y_{h(p)} g, \end{aligned}$$

and the last equation says that $dh \circ X = Y \circ h$, which means that X and Y are h-related. □

Since by definition of $h_* X$, we have

$$(h_* X)_{h(p)} = dh_p(X_p),$$

the vector fields X and $h_* X$ are h-related, and the proof of Proposition 9.7 shows that X are Y are h-related iff

$$(h_* X)_{h(p)} g = Y_{h(p)} g \quad \text{for all } g \in C^k(N),$$

for short, $h_* X = Y$.

Proposition 9.7 can also be used to prove Proposition 9.6; see Tu [112] (Chapter 14, Proposition 14.19). Here is the proof:

Proof of Proposition 9.6. For every function $g \in C^k(N)$, we have

$$\begin{aligned} [X, Y](g \circ h) &= XY(g \circ h) - YX(g \circ h) && \text{by definition of } [X, Y] \\ &= X((Y_1 g) \circ h) - Y((X_1 g) \circ h) && \text{by Proposition 9.7} \\ &= (X_1 Y_1 g) \circ h - (Y_1 X_1 g) \circ h && \text{by Proposition 9.7} \\ &= ((X_1 Y_1 - Y_1 X_1) g) \circ h \\ &= ([X_1, Y_1] g) \circ h. \end{aligned}$$

By Proposition 9.7 again, this shows that $[X, Y]$ and $[X_1, Y_1]$ are h-related. □

As a corollary of Proposition 9.6, for any two vector fields X, Y on M, the vector fields $[X, Y]$ and $[h_*X, h_*Y]$ are h-related, which means that

$$h_*[X, Y] = [h_*X, h_*Y];$$

that is,

$$dh_p([X, Y]_p) = [dh_p(X_p), dh_p(Y_p)].$$

9.3 Integral Curves, Flow of a Vector Field, One-Parameter Groups of Diffeomorphisms

We begin with integral curves and (local) flows of vector fields on a manifold.

Definition 9.8. Let X be a C^{k-1} vector field on a C^k-manifold M $(k \geq 2)$, and let p_0 be a point on M. An *integral curve (or trajectory) for X with initial condition* p_0 is a C^{k-1}-curve $\gamma\colon I \to M$, so that

$$\dot{\gamma}(t) = X_{\gamma(t)}{}^1 \quad \text{for all } t \in I, \quad \text{and} \quad \gamma(0) = p_0,$$

where $I = (a, b) \subseteq \mathbb{R}$ is an open interval containing 0. See Figure 11.7.

What Definition 9.8 says is that an integral curve γ with initial condition p_0 is a curve on the manifold M passing through p_0, and such that for every point $p = \gamma(t)$ on this curve, the tangent vector to this curve at p, that is $\dot{\gamma}(t)$, coincides with the value X_p of the vector field X at p.

Given a vector field X as above, and a point $p_0 \in M$, is there an integral curve through p_0? Is such a curve unique? If so, how large is the open interval I? We provide some answers to the above questions below.

Definition 9.9. Let X be a C^{k-1} vector field on a C^k-manifold M $(k \geq 2)$, and let p_0 be a point on M. A *local flow for X at* p_0 is a map

$$\varphi\colon J \times U \to M,$$

where $J \subseteq \mathbb{R}$ is an open interval containing 0 and U is an open subset of M containing p_0, so that for every $p \in U$, the curve $t \mapsto \varphi(t, p)$ is an integral curve of X with initial condition p. See Figure 11.8.

Thus, a local flow for X is a family of integral curves for all points in some small open set around p_0 such that these curves all have the same domain J, independently of the initial condition $p \in U$.

[1] Recall our convention: if X is a vector field on M, then for every point $q \in M$ we identify $X(q) = (q, X_q)$ and X_q.

The following theorem is the main existence theorem of local flows. This is a promoted version of a similar theorem in the classical theory of ODE's in the case where M is an open subset of $\mathbb{R}^n$. For a full account of this theory, see Lang [75] or Berger and Gostiaux [15].

Theorem 9.8. *(Existence of a local flow) Let X be a C^{k-1} vector field on a C^k-manifold M ($k \geq 2$), and let p_0 be a point on M. There is an open interval $J \subseteq \mathbb{R}$ containing 0 and an open subset $U \subseteq M$ containing p_0, so that there is a unique local flow $\varphi\colon J \times U \to M$ for X at p_0. What this means is that if $\varphi_1\colon J \times U \to M$ and $\varphi_2\colon J \times U \to M$ are both local flows with domain $J \times U$, then $\varphi_1 = \varphi_2$. Furthermore, φ is C^{k-1}.*

We know that for any initial condition p_0, there is some integral curve through p_0. However, there could be two (or more) integral curves $\gamma_1\colon I_1 \to M$ and $\gamma_2\colon I_2 \to M$ with initial condition p_0. This leads to the natural question: How do γ_1 and γ_2 differ on $I_1 \cap I_2$? The next proposition shows they don't!

Proposition 9.9. *Let X be a C^{k-1} vector field on a C^k-manifold M ($k \geq 2$), and let p_0 be a point on M. If $\gamma_1\colon I_1 \to M$ and $\gamma_2\colon I_2 \to M$ are any two integral curves both with initial condition p_0, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$. See Figure 9.6.*

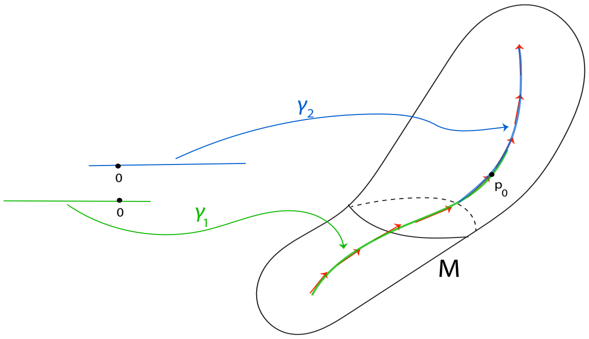


Figure 9.6: Two integral curves, γ_1 and γ_2, with initial condition p_0, which agree on the domain overlap $I_1 \cap I_2$.

Proof. Let $Q = \{t \in I_1 \cap I_2 \mid \gamma_1(t) = \gamma_2(t)\}$. Since $\gamma_1(0) = \gamma_2(0) = p_0$, the set Q is nonempty. If we show that Q is both closed and open in $I_1 \cap I_2$, as $I_1 \cap I_2$ is connected since it is an open interval of $\mathbb{R}$, we will be able to conclude that $Q = I_1 \cap I_2$.

Since by definition, a manifold is Hausdorff, it is a standard fact in topology that the diagonal $\Delta = \{(p, p) \mid p \in M\} \subseteq M \times M$ is closed, and since

$$Q = I_1 \cap I_2 \cap (\gamma_1, \gamma_2)^{-1}(\Delta)$$

where $(\gamma_1, \gamma_2)\colon I_1 \cap I_2 \to M \times M$ is the curve given by $(\gamma_1, \gamma_2)(t) = (\gamma_1(t), \gamma_2(t))$, and since γ_1 and γ_2 are continuous, we see that Q is closed in $I_1 \cap I_2$.

Pick any $u \in Q$ and consider the curves β_1 and β_2 given by

$$\beta_1(t) = \gamma_1(t+u) \quad \text{and} \quad \beta_2(t) = \gamma_2(t+u),$$

where $t \in I_1 - u$ in the first case, and $t \in I_2 - u$ in the second. (Here, if $I = (a, b)$, we have $I - u = (a - u, b - u)$.) Observe that

$$\dot{\beta}_1(t) = \dot{\gamma}_1(t+u) = X(\gamma_1(t+u)) = X(\beta_1(t)),$$

and similarly $\dot{\beta}_2(t) = X(\beta_2(t))$. We also have

$$\beta_1(0) = \gamma_1(u) = \gamma_2(u) = \beta_2(0) = q,$$

since $u \in Q$ (where $\gamma_1(u) = \gamma_2(u)$). Thus, $\beta_1\colon (I_1 - u) \to M$ and $\beta_2\colon (I_2 - u) \to M$ are two integral curves with the same initial condition q. By Theorem 9.8, the uniqueness of local flow implies that there is some open interval $\widetilde{I} \subseteq I_1 \cap I_2 - u$, such that $\beta_1 = \beta_2$ on $\widetilde{I}$. Consequently, γ_1 and γ_2 agree on $\widetilde{I} + u$, an open subset of Q, proving that Q is indeed open in $I_1 \cap I_2$. □

Proposition 9.9 implies the important fact that there is a *unique maximal* integral curve with initial condition p. Indeed, if $\{\gamma_j\colon I_j \to M\}_{j \in K}$ is the family of all integral curves with initial condition p (for some big index set K), if we let $I(p) = \bigcup_{j \in K} I_j$, we can define a curve $\gamma_p\colon I(p) \to M$ so that

$$\gamma_p(t) = \gamma_j(t), \quad \text{if} \quad t \in I_j.$$

Since γ_j and γ_l agree on $I_j \cap I_l$ for all $j, l \in K$, the curve γ_p is indeed well defined, and it is clearly an integral curve with initial condition p with the largest possible domain (the open interval, $I(p)$).

Definition 9.10. The curve γ_p defined above is called the *maximal integral curve with initial condition* p, and it is also denoted by $\gamma(p, t)$. The domain of γ_p is $I(p)$.

Note that Proposition 9.9 implies that any two distinct integral curves are disjoint, i.e., do not intersect each other.

Consider the vector field in $\mathbb{R}^2$ given by

$$X_{(x,y)} = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$$

shown in Figure 9.7. If we write $\gamma(t) = (x(t), y(t))$, the differential equation $\dot{\gamma}(t) = X(\gamma(t))$ is expressed by

$$\begin{aligned} x'(t) &= -y(t) \\ y'(t) &= x(t), \end{aligned}$$

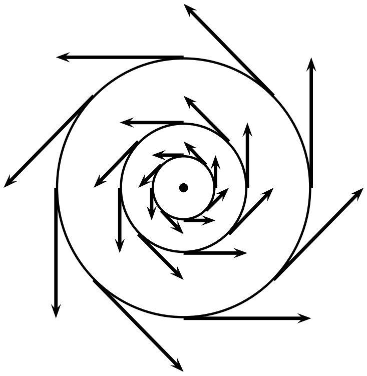


Figure 9.7: A vector field in $\mathbb{R}^2$.

or in matrix form,

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

If we write $X = \binom{x}{y}$ and $A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, then the above equation is written as

$$X' = AX.$$

Now as

$$e^{tA} = I + \frac{A}{1!}t + \frac{A^2}{2!}t^2 + \cdots + \frac{A^n}{n!}t^n + \cdots ,$$

we get

$$\frac{d}{dt}(e^{tA}) = A + \frac{A^2}{1!}t + \frac{A^3}{2!}t^2 + \cdots + \frac{A^n}{(n-1)!}t^{n-1} + \cdots = Ae^{tA},$$

so we see that $e^{tA}p$ is a solution of the ODE $X' = AX$ with initial condition $X = p$, and by uniqueness, $X = e^{tA}p$ is *the* solution of our ODE starting at $X = p$. Thus, our integral curve γ_p through $p = \binom{x_0}{y_0}$ is the circle given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}.$$

Observe that $I(p) = \mathbb{R}$, for every $p \in \mathbb{R}^2$.

If we delete the points $(-1, 0)$ and $(1, 0)$ on the x-axis, then for every point p_0 not on the unit circle S^1 (given by $x^2 + y^2 = 1$), the maximal integral curve through p_0 is the circle of center O through p_0, as before. However, for every point p_0 on the open upper half unit circle S^1_+ , the maximal integral curve through p_0 is S^1_+, and for every point p_0 on the open lower half unit circle S^1_-, the maximal integral curve through p_0 is S^1_-. In both cases, the domain of the integral curve is an open interval properly contained in $\mathbb{R}$. This example shows that it may not be possible to extend the domain of an integral curve to the entire real line.

Here is one more example of a vector field on $M = \mathbb{R}$ that has integral curves not defined on the whole of $\mathbb{R}$. Let X be the vector field on $\mathbb{R}$ given by

$$X_{(x)} = (1 + x^2)\frac{\partial}{\partial x}.$$

By solving the differential equation $\gamma'(t) = x'(t) = 1 + x^2$, it is easy to see that the maximal integral curve with initial condition $p_0 = 0$ is the curve $\gamma\colon (-\pi/2, \pi/2) \to \mathbb{R}$ given by

$$\gamma(t) = \tan t.$$

The following interesting question now arises. Given any $p_0 \in M$, if $\gamma_{p_0}\colon I(p_0) \to M$ is the maximal integral curve with initial condition p_0, and for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$, then there is a maximal integral curve $\gamma_{p_1}\colon I(p_1) \to M$ with initial condition p_1; what is the relationship between γ_{p_0} and γ_{p_1}, if any? The answer is given by

Proposition 9.10. *Let X be a C^{k-1} vector field on a C^k-manifold M $(k \geq 2)$, and let p_0 be a point on M. If $\gamma_{p_0}\colon I(p_0) \to M$ is the maximal integral curve with initial condition p_0, for any $t_1 \in I(p_0)$, if $p_1 = \gamma_{p_0}(t_1) \in M$ and $\gamma_{p_1}\colon I(p_1) \to M$ is the maximal integral curve with initial condition p_1, then*

$$I(p_1) = I(p_0) - t_1 \quad \text{and} \quad \gamma_{p_1}(t) = \gamma_{\gamma_{p_0}(t_1)}(t) = \gamma_{p_0}(t + t_1), \quad \text{for all } t \in I(p_0) - t_1.$$

See Figure 9.8.

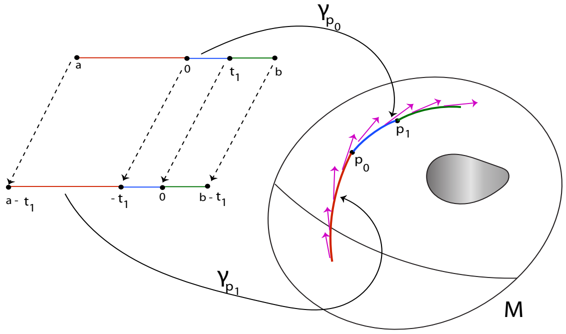


Figure 9.8: The integral curve γ_{p_1} is a reparametrization of γ_{p_0}.

Proof. Let $\gamma(t)$ be the curve given by

$$\gamma(t) = \gamma_{p_0}(t + t_1), \quad \text{for all } \ t \in I(p_0) - t_1.$$

Clearly γ is defined on $I(p_0) - t_1$, and

$$\dot{\gamma}(t) = \dot{\gamma}_{p_0}(t + t_1) = X(\gamma_{p_0}(t + t_1)) = X(\gamma(t))$$

and $\gamma(0) = \gamma_{p_0}(t_1) = p_1$. Thus, γ is an integal curve defined on $I(p_0) - t_1$ with initial condition p_1. If γ was defined on an interval $\widetilde{I} \supseteq I(p_0) - t_1$ with $\widetilde{I} \neq I(p_0) - t_1$, then γ_{p_0} would be defined on $\widetilde{I} + t_1 \supset I(p_0)$, an interval strictly bigger than $I(p_0)$, contradicting the maximality of $I(p_0)$. Therefore, $I(p_0) - t_1 = I(p_1)$. □

Proposition 9.10 says that the traces $\gamma_{p_0}(I(p_0))$ and $\gamma_{p_1}(I(p_1))$ in M of the maximal integral curves γ_{p_0} and γ_{p_1} are identical; they only differ by a simple reparametrization $(u = t + t_1)$.

It is useful to restate Proposition 9.10 by changing point of view. So far, we have been focusing on integral curves: given any $p_0 \in M$, we let t vary in $I(p_0)$ and get an integral curve γ_{p_0} with domain $I(p_0)$. Instead of holding $p_0 \in M$ fixed, we can hold $t \in \mathbb{R}$ fixed and consider the set

$$\mathcal{D}_t(X) = \{p \in M \mid t \in I(p)\},$$

the set of points such that it is possible to "travel for t units of time from p" along the maximal integral curve γ_p with initial condition p (It is possible that $\mathcal{D}_t(X) = \emptyset$). By definition, if $\mathcal{D}_t(X) \neq \emptyset$, the point $\gamma_p(t)$ is well defined, and so we obtain a map
$\Phi_t^X \colon \mathcal{D}_t(X) \to M$ with domain $\mathcal{D}_t(X)$, given by

$$\Phi_t^X(p) = \gamma_p(t).$$

The above suggests the following definition.

Definition 9.11. Let X be a C^{k-1} vector field on a C^k-manifold M $(k \geq 2)$. For any $t \in \mathbb{R}$, let

$$\mathcal{D}_t(X) = \{p \in M \mid t \in I(p)\} \quad \text{and} \quad \mathcal{D}(X) = \{(t, p) \in \mathbb{R} \times M \mid t \in I(p)\},$$

and let $\Phi^X \colon \mathcal{D}(X) \to M$ be the map given by

$$\Phi^X(t, p) = \gamma_p(t).$$

The map Φ^X is called the *(global) flow of* X, and $\mathcal{D}(X)$ is called its *domain of definition*. For any $t \in \mathbb{R}$ such that $\mathcal{D}_t(X) \neq \emptyset$, the map $p \in \mathcal{D}_t(X) \mapsto \Phi^X(t, p) = \gamma_p(t)$ is denoted by Φ_t^X (i.e., $\Phi_t^X(p) = \Phi^X(t, p) = \gamma_p(t)$).

Observe that

$$\mathcal{D}(X) = \bigcup_{p\in M} (I(p) \times \{p\}).$$

Also, using the Φ_t^X notation, the property of Proposition 9.10 reads

$$\Phi_s^X \circ \Phi_t^X = \Phi_{s+t}^X, \tag{$*$}$$

whenever both sides of the equation make sense. Indeed, the above says

$$\Phi_s^X(\Phi_t^X(p)) = \Phi_s^X(\gamma_p(t)) = \gamma_{\gamma_p(t)}(s) = \gamma_p(s+t) = \Phi_{s+t}^X(p).$$

Using the above property, we can easily show that the Φ_t^X are invertible. In fact, the inverse of Φ_t^X is Φ_{-t}^X. First, note that

$$\mathcal{D}_0(X) = M \quad \text{and} \quad \Phi_0^X = \mathrm{id},$$

because, by definition, $\Phi_0^X(p) = \gamma_p(0) = p$, for every $p \in M$. Then, $(*)$ implies that

$$\Phi_t^X \circ \Phi_{-t}^X = \Phi_{t+(-t)}^X = \Phi_0^X = \mathrm{id},$$

which shows that $\Phi_t^X \colon \mathcal{D}_t(X) \to \mathcal{D}_{-t}(X)$ and $\Phi_{-t}^X \colon \mathcal{D}_{-t}(X) \to \mathcal{D}_t(X)$ are inverse of each other. Moreover, each Φ_t^X is a C^{k-1}-diffeomorphism. We summarize in the following proposition some additional properties of the domains $\mathcal{D}(X)$, $\mathcal{D}_t(X)$ and the maps Φ_t^X. (For a proof, see Lang [75] or Warner [114].)

Theorem 9.11. *Let X be a C^{k-1} vector field on a C^k-manifold M ($k \geq 2$). The following properties hold:*

(a) For every $t \in \mathbb{R}$, if $\mathcal{D}_t(X) \neq \emptyset$, then $\mathcal{D}_t(X)$ is open (this is trivially true if $\mathcal{D}_t(X) = \emptyset$).

(b) The domain $\mathcal{D}(X)$ of the flow Φ^X is open, and the flow is a C^{k-1} map $\Phi^X \colon \mathcal{D}(X) \to M$.

(c) Each $\Phi_t^X \colon \mathcal{D}_t(X) \to \mathcal{D}_{-t}(X)$ is a C^{k-1}-diffeomorphism with inverse Φ_{-t}^X.

(d) For all $s, t \in \mathbb{R}$, the domain of definition of $\Phi_s^X \circ \Phi_t^X$ is contained but generally not equal to $\mathcal{D}_{s+t}(X)$. However, $\mathrm{dom}(\Phi_s^X \circ \Phi_t^X) = \mathcal{D}_{s+t}(X)$ *if s and t have the same sign. Moreover, on* $\mathrm{dom}(\Phi_s^X \circ \Phi_t^X)$*, we have*

$$\Phi_s^X \circ \Phi_t^X = \Phi_{s+t}^X.$$

Remarks:

(1) We may omit the superscript X and write Φ instead of Φ^X if no confusion arises.

(2) The reason for using the terminology flow in referring to the map Φ^X can be clarified as follows. For any t such that $\mathcal{D}_t(X) \neq \emptyset$, every integral curve γ_p with initial condition $p \in \mathcal{D}_t(X)$ is defined on some open interval containing $[0, t]$, and we can picture these curves as "flow lines" along which the points p flow (travel) for a time interval t. Then, $\Phi^X(t, p)$ is the point reached by "flowing" for the amount of time t on the integral curve γ_p (through p) starting from p. Intuitively, we can imagine the flow of a fluid through M, and the vector field X is the field of velocities of the flowing particles.

Given a vector field X as above, it may happen that $\mathcal{D}_t(X) = M$, for all $t \in \mathbb{R}$.

Definition 9.12. When $\mathcal{D}(X) = \mathbb{R} \times M$, we say that the vector field X is *complete*. Then the Φ_t^X are diffeomorphisms of M, and they form a group under composition. The family $\{\Phi_t^X\}_{t\in\mathbb{R}}$ a called a *1-parameter group of X*.

If the vector field X is complete, then Φ^X induces a group homomorphism $(\mathbb{R}, +) \longrightarrow \mathrm{Diff}(M)$, from the additive group $\mathbb{R}$ to the group of C^{k-1}-diffeomorphisms of M.

By abuse of language, even when it is **not** the case that $\mathcal{D}_t(X) = M$ for all t, the family $\{\Phi_t^X\}_{t\in\mathbb{R}}$ is called a *local 1-parameter group generated by X*, even though it is **not** a group, because the composition $\Phi_s^X \circ \Phi_t^X$ may not be defined.

If we go back to the vector field in $\mathbb{R}^2$ given by

$$X = -y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y},$$

since the integral curve $\gamma_p(t)$, through $p = \binom{x_0}{x_0}$ is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},$$

the global flow associated with X is given by

$$\Phi^X(t, p) = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} p,$$

and each diffeomorphism Φ_t^X is the rotation

$$\Phi_t^X = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

The 1-parameter group $\{\Phi_t^X\}_{t\in\mathbb{R}}$ generated by X is the group of rotations in the plane, $\mathbf{SO}(2)$.

More generally, if B is an $n \times n$ invertible matrix that has a real logarithm A (that is, if $e^A = B$), then the matrix A defines a vector field X in $\mathbb{R}^n$, with

$$X = \sum_{i,j=1}^{n} (a_{ij}x_j)\frac{\partial}{\partial x_i},$$

whose integral curves are of the form

$$\gamma_p(t) = e^{tA}p,$$

and we have

$$\gamma_p(1) = Bp.$$

The one-parameter group $\{\Phi_t^X\}_{t\in\mathbb{R}}$ generated by X is given by $\{e^{tA}\}_{t\in\mathbb{R}}$.

When M is compact, it turns out that every vector field is complete, a nice and useful fact.

Proposition 9.12. *Let X be a C^{k-1} vector field on a C^k-manifold M ($k \geq 2$). If M is compact, then X is complete, which means that $\mathcal{D}(X) = \mathbb{R} \times M$. Moreover, the map $t \mapsto \Phi_t^X$ is a homomorphism from the additive group $\mathbb{R}$ to the group* Diff(M) *of (C^{k-1}) diffeomorphisms of M.*

Proof. Pick any $p \in M$. By Theorem 9.8, there is a local flow $\varphi_p \colon J(p) \times U(p) \to M$, where $J(p) \subseteq \mathbb{R}$ is an open interval containing 0 and $U(p)$ is an open subset of M containing p, so that for all $q \in U(p)$, the map $t \mapsto \varphi(t, q)$ is an integral curve with initial condition q (where $t \in J(p)$). Thus, we have $J(p) \times U(p) \subseteq \mathcal{D}(X)$. Now, the $U(p)$'s form an open cover of M, and since M is compact, we can extract a finite subcover $\bigcup_{q\in F} U(q) = M$, for some finite subset $F \subseteq M$. But then, we can find $\epsilon > 0$ so that $(-\epsilon, +\epsilon) \subseteq J(q)$, for all $q \in F$ and for all $t \in (-\epsilon, +\epsilon)$, and for all $p \in M$, if γ_p is the maximal integral curve with initial condition p, then $(-\epsilon, +\epsilon) \subseteq I(p)$.

For any $t \in (-\epsilon, +\epsilon)$, consider the integral curve $\gamma_{\gamma_p(t)}$, with initial condition $\gamma_p(t)$. This curve is well defined for all $t \in (-\epsilon, +\epsilon)$, and by Proposition 9.10 we have

$$\gamma_{\gamma_p(t)}(t) = \gamma_p(t + t) = \gamma_p(2t),$$

which shows that γ_p is in fact defined for all $t \in (-2\epsilon, +2\epsilon)$. By induction we see that

$$(-2^n\epsilon, +2^n\epsilon) \subseteq I(p),$$

for all $n \geq 0$, which proves that $I(p) = \mathbb{R}$. As this holds for all $p \in M$, we conclude that $\mathcal{D}(X) = \mathbb{R} \times M$. □

Remarks:

(1) The proof of Proposition 9.12 also applies when X is a vector field with compact support (this means that the closure of the set $\{p \in M \mid X(p) \neq 0\}$ is compact).

(2) If $h \colon M \to N$ is a diffeomorphism and X is a vector field on M, then it can be shown that the local 1-parameter group associated with the vector field h_*X is

$$\{h \circ \Phi_t^X \circ h^{-1}\}_{t\in\mathbb{R}}.$$

A point $p \in M$ where a vector field vanishes (i.e., $X(p) = 0$) is called a *critical point of* X. Critical points play a major role in the study of vector fields, in differential topology (e.g., the celebrated Poincaré–Hopf index theorem), and especially in Morse theory, but we won't go into this here (curious readers should consult Milnor [81], Guillemin and Pollack [55] or DoCarmo [38], which contains an informal but very clear presentation of the Poincaré–Hopf index theorem). Another famous theorem about vector fields says that every smooth vector field on a sphere of even dimension (S^{2n}) must vanish in at least one point (the so-called "hairy-ball theorem." A proof of this result can be found in Milnor [83]. On S^2, it says that you can't comb your hair without having a singularity somewhere. Try it, it's true!).

Let us just observe that if an integral curve γ passes through a critical point p, then γ is reduced to the point p; that is, $\gamma(t) = p$, for all t. Indeed, such a curve is an integral curve with initial condition p. By the uniqueness property, it is the only one. Then we see that if a maximal integral curve is defined on the whole of $\mathbb{R}$, either it is injective (it has no self-intersection), or it is simply periodic (i.e., there is some $T > 0$ so that $\gamma(t + T) = \gamma(t)$, for all $t \in \mathbb{R}$ and γ is injective on $[0, T)$), or it is reduced to a single point.

We conclude this section with the definition of the Lie derivative of a vector field with respect to another vector field.

Say we have two vector fields X and Y on M. For any $p \in M$, we can flow along the integral curve of X with initial condition p to $\Phi_t(p)$ (for t small enough) and then evaluate Y there, getting $Y(\Phi_t(p))$. Now, this vector belongs to the tangent space $T_{\Phi_t(p)}(M)$, but $Y(p) \in T_p(M)$. So, to "compare" $Y(\Phi_t(p))$ and $Y(p)$, we bring back $Y(\Phi_t(p))$ to $T_p(M)$ by applying the tangent map $d\Phi_{-t}$ at $\Phi_t(p)$ to $Y(\Phi_t(p))$. (Note that to alleviate the notation, we use the slight abuse of notation $d\Phi_{-t}$ instead of $d(\Phi_{-t})_{\Phi_t(p)}$.) We can then form the difference $d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)$, divide by t, and consider the limit as t goes to 0.

Definition 9.13. Let M be a C^{k+1} manifold. Given any two C^k vector fields X and Y on M, for every $p \in M$, the *Lie derivative of* Y *with respect to* X *at* p denoted $(L_X\, Y)_p$, is given by

$$(L_X\, Y)_p = \lim_{t \longrightarrow 0} \frac{d\Phi_{-t}(Y(\Phi_t(p))) - Y(p)}{t} = \frac{d}{dt}\left(d\Phi_{-t}(Y(\Phi_t(p)))\right)\bigg|_{t=0}.$$

It can be shown that $(L_X\, Y)_p$ is our old friend the Lie bracket; that is,

$$(L_X\, Y)_p = [X, Y]_p.$$

For a proof, see Warner [114] (Chapter 2, Proposition 2.25) or O'Neill [91] (Chapter 1, Proposition 58).

In terms of Definition 9.6, observe that

$$(L_X\, Y)_p = \lim_{t \longrightarrow 0} \frac{\big((\Phi_{-t})_* Y\big)(p) - Y(p)}{t} = \lim_{t \longrightarrow 0} \frac{\big(\Phi_t^* Y\big)(p) - Y(p)}{t} = \frac{d}{dt}\big(\Phi_t^* Y\big)(p)\bigg|_{t=0},$$

since $(\Phi_{-t})^{-1} = \Phi_t$.

Next we discuss the application of vector fields and integral curves to the blending of locally affine transformations, known as Log-Euclidean polyaffine transformations, as presented in Arsigny, Commowick, Pennec and Ayache [7].

9.4 Log-Euclidean Polyaffine Transformations

The registration of medical images is an important and difficult problem. The work described in Arsigny, Commowick, Pennec and Ayache [7] (and Arsigny's thesis [6]) makes an orginal and valuable contribution to this problem by describing a method for parametrizing a class of non-rigid deformations with a small number of degrees of freedom. After a global affine alignment, this sort of parametrization allows a finer local registration with very smooth transformations. This type of parametrization is particularly well adpated to the registration of histological slices, see Arsigny, Pennec and Ayache [9].

The goal is to fuse some affine or rigid transformations in such a way that the resulting transformation is invertible and smooth. The direct approach which consists in blending N global affine or rigid transformations $T_1, \ldots, T_N$ using weights $w_1, \ldots, w_N$ does not work, because the resulting transformation

$$T = \sum_{i=1}^{N} w_i T_i$$

is not necessarily invertible. The purpose of the weights is to define the domain of influence in space of each T_i.

The key idea is to associate to each rigid (or affine) transformation T of $\mathbb{R}^n$ a vector field V on $\mathbb{R}^n$ viewed as a manifold, and to view T as the diffeomorphism Φ_1^V corresponding to the time $t = 1$, where Φ_t^V is the global flow associated with V. In other words, T is the result of integrating an ODE

$$X' = V(X, t),$$

starting with some initial condition X_0, and $T = X(1)$.

It would be highly desirable if the vector field V did not depend on the time parameter, and this is indeed possible for a large class of affine transformations, which is one of the nice contributions of the work of Arsigny, Commowick, Pennec and Ayache [7]. Recall that an affine transformation $X \mapsto LX + v$ (where L is an $n \times n$ matrix and $X, v \in \mathbb{R}^n$) can be conveniently represented as a linear transformation from $\mathbb{R}^{n+1}$ to itself if we write

$$\begin{pmatrix} X \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} L & v \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}.$$

Then the ODE with constant coefficients

$$X' = LX + v$$

can be written

$$\begin{pmatrix} X' \\ 0 \end{pmatrix} = \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix},$$

and for every initial condition $X = X_0$, its unique solution is given by

$$\begin{pmatrix} X(t) \\ 1 \end{pmatrix} = \exp\left(t \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} X_0 \\ 1 \end{pmatrix}.$$

Therefore, if we can find reasonable conditions on matrices $T = \begin{pmatrix} M & u \\ 0 & 1 \end{pmatrix}$ to ensure that they have a unique real logarithm

$$\log(T) = \begin{pmatrix} L & v \\ 0 & 0 \end{pmatrix},$$

then we will be able to associate a vector field $V(X) = LX + v$ to T in such a way that T is recovered by integrating the ODE $X' = LX + v$. Furthermore, given N transformations $T_1, \ldots, T_N$ such that $\log(T_1), \ldots, \log(T_N)$ are uniquely defined, we can fuse $T_1, \ldots, T_N$ at the *infinitesimal level* by defining the ODE obtained by blending the vector fields $V_1, \ldots, V_N$ associated with $T_1, \ldots, T_N$ (with $V_i(X) = L_i X + v_i$), namely

$$V(X) = \sum_{i=1}^{N} w_i(X)(L_i X + v_i).$$

Then it is easy to see that the ODE

$$X' = V(X)$$

has a unique solution for every $X = X_0$ defined for all t, and the fused transformation is just $T = X(1)$. Thus, the fused vector field

$$V(X) = \sum_{i=1}^{N} w_i(X)(L_i X + v_i)$$

yields a one-parameter group of diffeomorphisms Φ_t. Each transformation Φ_t is smooth and invertible, and is called a *Log-Euclidean polyaffine transformation*, for short, *LEPT*. Of course, we have the equation

$$\Phi_{s+t} = \Phi_s \circ \Phi_t,$$

for all $s, t \in \mathbb{R}$, so in particular, the inverse of Φ_t is Φ_{-t}. We can also interpret Φ_s as $(\Phi_1)^s$, which will yield a fast method for computing Φ_s. Observe that when the weight are scalars, the one-parameter group is given by

$$\begin{pmatrix} \Phi_t(X) \\ 1 \end{pmatrix} = \exp\left(t \sum_{i=1}^{N} w_i \begin{pmatrix} L_i & v_i \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} X \\ 1 \end{pmatrix},$$

which is the Log-Euclidean mean of the affine transformations T_i's (w.r.t. the weights w_i).

Fortunately, there is a sufficient condition for a real matrix to have a unique real logarithm and this condition is not too restrictive in practice.

Recall that $\mathcal{E}(n)$ denotes the set of all real matrices whose eigenvalues $\lambda + i\mu$ lie in the horizontal strip determined by the condition $-\pi < \mu < \pi$. We have the following version of Theorem 2.2.

Theorem 9.13. *The image* $\exp(\mathcal{E}(n))$ *of* $\mathcal{E}(n)$ *by the exponential map is the set of real invertible matrices with no negative eigenvalues and* $\exp\colon \mathcal{E}(n) \to \exp(\mathcal{E}(n))$ *is a bijection.*

Theorem 9.13 is stated in Kenney and Laub [65] without proof. Instead, Kenney and Laub cite DePrima and Johnson [35] for a proof, but this latter paper deals with complex matrices and does not contain a proof of our result either. The injectivity part of Theorem 9.13 can be found in Mmeimné and Testard [86], Chapter 3, Theorem 3.8.4.

In fact, $\exp\colon \mathcal{E}(n) \to \exp(\mathcal{E}(n))$ is a diffeomorphism, a result proved in Bourbaki [19]; see Chapter III, Section 6.9, Proposition 17 and Theorem 6. Curious readers should read Gallier [47] for the full story.

For any matrix $A \in \exp(\mathcal{E}(n))$, we refer to the unique matrix $X \in \mathcal{E}(n)$ such that $e^X = A$ as the *principal logarithm* of A, and we denote it as $\log A$.

Observe that if T is an affine transformation given in matrix form by

$$T = \begin{pmatrix} M & t \\ 0 & 1 \end{pmatrix},$$

since the eigenvalues of T are those of M plus the eigenvalue 1, the matrix T has no negative eigenvalues iff M has no negative eigenvalues, and thus the principal logarithm of T exists iff the principal logarithm of M exists.

It is proved in Arsigny, Commowick, Pennec and Ayache that LEPT's are affine invariant; see [7], Section 2.3. This shows that LEPT's are produced by a truly geometric kind of blending, since the result does not depend at all on the choice of the coordinate system.

In the next section, we describe a fast method for computing due to Arsigny, Commowick, Pennec and Ayache [7].

9.5 Fast Polyaffine Transforms

Recall that since LEPT's are members of the one-parameter group $(\Phi_t)_{t\in\mathbb{R}}$, we have

$$\Phi_{2t} = \Phi_{t+t} = \Phi_t^2,$$

and thus,

$$\Phi_1 = (\Phi_{1/2^N})^{2^N}.$$

Observe the formal analogy of the above formula with the formula

$$\exp(M) = \exp\left(\frac{M}{2^N}\right)^{2^N}$$

for computing the exponential of a matrix M by the *scaling and squaring method*.

It turns out that the "scaling and squaring method" is one of the most efficient methods for computing the exponential of a matrix; see Kenney and Laub [65] and Higham [59]. The key idea is that $\exp(M)$ is easy to compute if M is close zero, since in this case, one can use a few terms of the exponential series, or better a Padé approximant (see Higham [59]). The scaling and squaring method for computing the exponential of a matrix M can be sketched as follows:

1. *Scaling Step*: Divide M by a factor 2^N, so that $\frac{M}{2^N}$ is close enough to zero.
2. *Exponentiation Step*: Compute $\exp\left(\frac{M}{2^N}\right)$ with high precision, for example using a Padé approximant.
3. *Squaring Step*: Square $\exp\left(\frac{M}{2^N}\right)$ repeatedly N times to obtain $\exp\left(\frac{M}{2^N}\right)^{2^N}$, a very accurate approximation of e^M.

There is also a so-called *inverse scaling and squaring method* to compute efficiently the principal logarithm of a real matrix; see Cheng, Higham, Kenney and Laub [30].

Arsigny, Commowick, Pennec and Ayache made the very astute observation that the scaling and squaring method can be adapted to compute LEPT's very efficiently [7]. This method called *fast polyaffine transform* computes the values of a Log-Euclidean polyaffine transformation $T = \Phi_1$ at the vertices of a regular n-dimensional grid (in practice, for $n = 2$ or $n = 3$). Recall that T is obtained by integrating an ODE $X' = V(X)$, where the vector field V is obtained by blending the vector fields associated with some affine transformations $T_1, \ldots, T_n$, having a principal logarithm.

Here are the three steps of the **fast polyaffine transform**:

1. *Scaling Step*: Divide the vector field V by a factor 2^N, so that $\frac{V}{2^N}$ is close enough to zero.
2. *Exponentiation Step*: Compute $\Phi_{1/2^N}$, using some adequate numerical integration method.
3. *Squaring Step*: Compose $\Phi_{1/2^N}$ with itself recursively N times to obtain an accurate approximation of $T = \Phi_1$.

Of course, one has to provide practical methods to achieve Step 2 and Step 3. Several methods to achieve Step 2 and Step 3 are proposed in Arsigny, Commowick, Pennec and Ayache [7]. One also has to worry about boundary effects, but this problem can be alleviated too, using bounding boxes. At this point, the reader is urged to read the full paper [7] for complete details and beautiful pictures illustrating the use of LEPT's in medical imaging.

For more details regarding the LEPT, including the Log-Euclidean framework for locally rigid or affine deformation, the reader should read Arsigny, Commowick, Pennec and Ayache [7].

9.6 Problems

Problem 9.1. The inversion $\mathcal{I}$ (in $\mathbb{R}^n$) of center $O = (0, \ldots, 0)$ and power 1 is given by

$$(x_1, \ldots, x_n) \mapsto \frac{1}{\sum_{i=1}^n x_i^2} (x_1, \ldots, x_n).$$

Prove that for every point $u \in \mathbb{R}^n - \{0\}$, we have

$$d\mathcal{I}_u(h) = \|u\|^{-2} \left(h - 2\frac{\langle u, h\rangle}{\|u\|^2}\, u \right).$$

Problem 9.2. Check that $\Gamma^{(k)}(U, T(M))$ is a real vector space.

Problem 9.3. Prove Proposition 9.6.

Problem 9.4. Check that L_X has the Leibniz property, that is,

$$L_X(fg) = L_X(f)g + fL_X(g).$$

Problem 9.5. Let X be the vector field on $\mathbb{R}$ given by

$$X_{(x)} = (1 + x^2)\frac{\partial}{\partial x}.$$

Prove that maximal integral curve with initial condition $p_0 = 0$ is the curve $\gamma\colon (-\pi/2, \pi/2) \to \mathbb{R}$ given by

$$\gamma(t) = \tan t.$$

Problem 9.6. If B is an $n \times n$ invertible matrix that has a real logarithm A (that is, if $e^A = B$), then the matrix A defines a vector field X in $\mathbb{R}^n$, with

$$X = \sum_{i,j=1}^n (a_{ij}x_j)\frac{\partial}{\partial x_i}.$$

Show that the integral curves are of the form

$$\gamma_p(t) = e^{tA}p,$$

and that

$$\gamma_p(1) = Bp.$$

Show that the one-parameter group $\{\Phi_t^X\}_{t\in\mathbb{R}}$ generated by X is given by $\{e^{tA}\}_{t\in\mathbb{R}}$.

Problem 9.7. Prove that two smooth vector fields X and Y on a smooth manifold M are equal if and only if for every smooth function f on M, we have $Xf = Yf$.

Problem 9.8. Let $x_1, y_1, \ldots, x_n, y_n$ denote the standard coordinates on $\mathbb{R}^{2n}$. Prove that the vector field X defined on the sphere S^{2n-1} of equation $\sum_{i=1}^n (x_i^2 + y_i^2) = 1$ given by

$$X = \sum_{i=1}^n -y_i \frac{\partial}{\partial x_i} + \sum_{i=1}^n x_i \frac{\partial}{\partial y_i}$$

is a nowhere-vanishing smooth vector field.

On the other hand, it can be shown that every continuous vector field on an even-dimensional sphere $S^{2n} \subseteq \mathbb{R}^{2n+1}$ must vanish at some point. A proof of this result can be found in Milnor [83].

Problem 9.9. Let M be a smooth manifold and let $f\colon M \to \mathbb{R}$ be a smooth function on M. Recall that a crititical point $p \in M$ of f is a point such that $df_p = 0$.

(1) Prove that if p is a critical point of f then there exists a function $H\colon T_pM \times T_pM \to \mathbb{R}$ (called a *Hessian*) such that

$$H(X_p, Y_p) = X_p(Yf) = Y_p(Xf)$$

for all smooth vector fields $X, Y \in \mathfrak{X}(M)$.

(2) Prove that H is bilinear, symmetric, and satisfies

$$H\left(\left.\frac{\partial}{\partial x_i}\right|_p, \left.\frac{\partial}{\partial x_j}\right|_p\right) = \frac{\partial^2 f}{\partial x_i \partial x_j}(p),$$

relative to a coordinate system.

(3) Prove that

$$H(v, v) = \frac{d^2(f \circ \alpha)}{ds^2}(0)$$

for every curve α through p such that $\alpha'(0) = v$.

Problem 9.10. Read Arsigny, Commowick, Pennec and Ayache [7] and implement the method for fusing affine transformations described in Section 2 of that paper.

Chapter 10

Partitions of Unity, Covering Maps ⊛

This chapter contains a selection of technical tools. It is preparatory for best understanding certain proofs which occur in the remaining chapters.

10.1 Partitions of Unity

To study manifolds, it is often necessary to construct various objects such as functions, vector fields, Riemannian metrics, volume forms, *etc.*, by gluing together items constructed on the domains of charts. Partitions of unity are a crucial technical tool in this gluing process.

The first step is to define "bump functions" (also called plateau functions). For any $r > 0$, we denote by $B(r)$ the open ball

$$B(r) = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 < r\},$$

and by $\overline{B(r)} = \{(x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_1^2 + \cdots + x_n^2 \leq r\}$ its closure.

Proposition 10.1. *There is a smooth function $b\colon \mathbb{R}^n \to \mathbb{R}$, so that*

$$b(x) = \begin{cases} 1 & \textit{if } x \in \overline{B(1)} \\ 0 & \textit{if } x \in \mathbb{R}^n - B(2). \end{cases}$$

See Figures 10.1 and 10.2.

Proof. There are many ways to construct such a function. We can proceed as follows. Consider the function $h\colon \mathbb{R} \to \mathbb{R}$ given by

$$h(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

It is easy to show that h is C^∞ (but **not** analytic!). For details, see Section 1.1 of Tu [112]. Define $b\colon \mathbb{R}^n \to \mathbb{R}$ by

$$b(x_1, \ldots, x_n) = \frac{h((4 - x_1^2 - \cdots - x_n^2)/3)}{h((4 - x_1^2 - \cdots - x_n^2)/3) + h((x_1^2 + \cdots + x_n^2 - 1)/3)}.$$

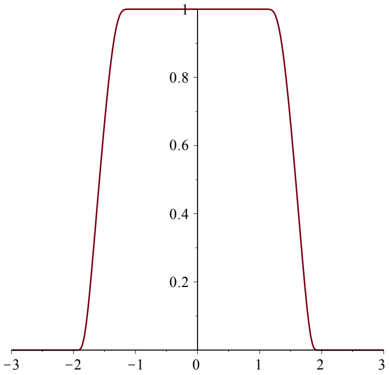


Figure 10.1: The graph of $b\colon \mathbb{R} \to \mathbb{R}$ used in Proposition 10.1.

It is immediately verified that b satisfies the required conditions. $\square$

Remark: The function obtained by omitting the factor $1/3$ also yields a smooth bump function, but it looks a little different; its cross-section by a plane through the x_{n+1}-axis has four inflection points instead of two. See Figures 10.4 and 10.5.

Definition 10.1. Given a topological space X, for any function $f\colon X \to \mathbb{R}$, the *support of f*, denoted $\operatorname{supp} f$, is the closed set

$$\operatorname{supp} f = \overline{\{x \in X \mid f(x) \neq 0\}}.$$

Proposition 10.1 yields the following useful technical result, which says that a smooth partial function f defined on an open subset U of a smooth manifold M can be extended to the whole of M as a function $\widetilde{f}$ that vanishes outside U and agrees with f on some compact subset contained in U.

Proposition 10.2. *Let M be a smooth manifold. For any open subset $U \subseteq M$, any $p \in U$ and any smooth function $f\colon U \to \mathbb{R}$, there exist an open subset V with $p \in V$ and a smooth function $\widetilde{f}\colon M \to \mathbb{R}$ defined on the whole of M, so that $\overline{V}$ is compact,*

$$\overline{V} \subseteq U, \qquad \operatorname{supp} \widetilde{f} \subseteq U,$$

and

$$\widetilde{f}(q) = f(q), \qquad \textit{for all} \quad q \in \overline{V}.$$

Proof. Using a scaling function, it is easy to find a chart (W, φ) at p so that $W \subseteq U$, $B(3) \subseteq \varphi(W)$, and $\varphi(p) = 0$. Let $\widetilde{b} = b \circ \varphi$, where b is the function given by Proposition 10.1. Then $\widetilde{b}$ is a smooth function on W with support $\varphi^{-1}(\overline{B(2)}) \subseteq W$. We can extend $\widetilde{b}$ outside W, by setting it to be 0, and we get a smooth function on the whole M. If we let

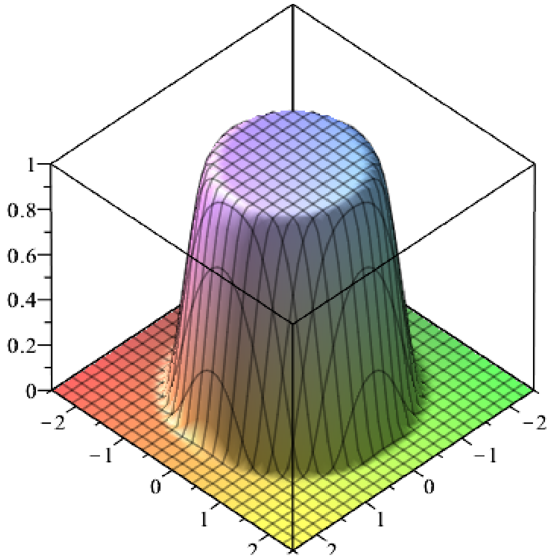


Figure 10.2: The graph of $b\colon \mathbb{R}^2 \to \mathbb{R}$ used in Proposition 10.1.

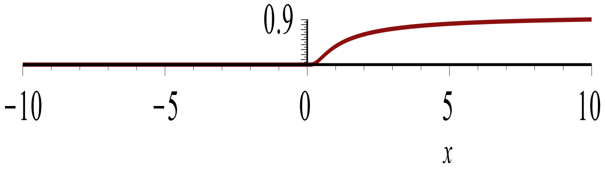


Figure 10.3: The graph of $h(x)$ used in Proposition 10.1.

$V = \varphi^{-1}(B(1))$, then V is an open subset around p, $\overline{V} = \varphi^{-1}(\overline{B(1)}) \subseteq W$ is compact, and clearly, $\widetilde{b} = 1$ on $\overline{V}$. Therefore, if we set

$$\widetilde{f}(q) = \begin{cases} \widetilde{b}(q)f(q) & \text{if } q \in W \\ 0 & \text{if } q \in M - W, \end{cases}$$

we see that $\widetilde{f}$ satisfies the required properties. □

Definition 10.2. If X is a (Hausdorff) topological space, a family $\{U_\alpha\}_{\alpha \in I}$ of subsets U_α of X is a *cover* (or *covering*) of X iff $X = \bigcup_{\alpha \in I} U_\alpha$. A cover $\{U_\alpha\}_{\alpha \in I}$ such that each U_α is open is an *open cover*. If $\{U_\alpha\}_{\alpha \in I}$ is a cover of X, for any subset $J \subseteq I$, the subfamily $\{U_\alpha\}_{\alpha \in J}$ is a *subcover* of $\{U_\alpha\}_{\alpha \in I}$ if $X = \bigcup_{\alpha \in J} U_\alpha$, i.e., $\{U_\alpha\}_{\alpha \in J}$ is still a cover of X. Given a cover $\{U_\beta\}_{\beta \in J}$, we say that a family $\{V_\alpha\}_{\alpha \in I}$ is a *refinement* of $\{U_\beta\}_{\beta \in J}$ if it is a cover and if there is a function $h\colon I \to J$ so that $V_\alpha \subseteq U_{h(\alpha)}$, for all $\alpha \in I$. See Figure 10.6.

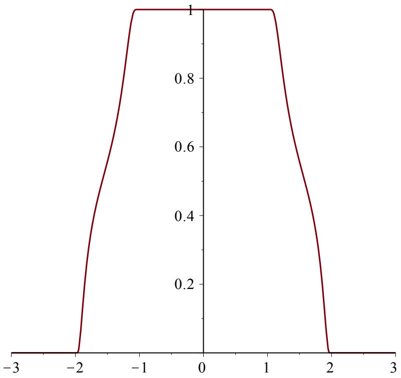


Figure 10.4: The graph of $b\colon \mathbb{R} \to \mathbb{R}$ with 1/3 omitted.

Definition 10.3. A family $\{U_\alpha\}_{\alpha\in I}$ of subsets of X is *locally finite* iff for every point $p \in X$, there is some open subset U with $p \in U$, so that $U \cap U_\alpha \neq \emptyset$ for only finitely many $\alpha \in I$. See Figure 10.7. A space X is *paracompact* iff every open cover has an open locally finite refinement.

Remark: Recall that a space X is *compact* iff it is Hausdorff and if every open cover has a *finite* subcover. Thus, the notion of paracompactness (due to Jean Dieudonné) is a generalization of the notion of compactness.

Definition 10.4. A topological space X is *second-countable* if it has a countable basis; that is, if there is a countable family of open subsets $\{U_i\}_{i\geq 1}$, so that every open subset of X is the union of some of the U_i's. A topological space X is *locally compact* iff it is Hausdorff, and for every $a \in X$, there is some compact subset K and some open subset U, with $a \in U$ and $U \subseteq K$.

As we will see shortly, every locally compact and second-countable topological space is paracompact.

The following fact is important.

Proposition 10.3. *Every manifold (even not second-countable) is locally compact.*

Proof. For every $p \in M$, if we pick a chart (U, φ) around p, then $\varphi(U) = \Omega$ for some open $\Omega \subseteq \mathbb{R}^n$ ($n = \dim M$). So, we can pick a small closed ball $\overline{B(q,\epsilon)} \subseteq \Omega$ of center $q = \varphi(p)$ and radius ϵ, and as φ is a homeomorphism, we see that

$$p \in \varphi^{-1}(B(q,\epsilon/2)) \subseteq \varphi^{-1}(\overline{B(q,\epsilon)}),$$

where $\varphi^{-1}(\overline{B(q,\epsilon)})$ is compact and $\varphi^{-1}(B(q,\epsilon/2))$ is open. $\square$

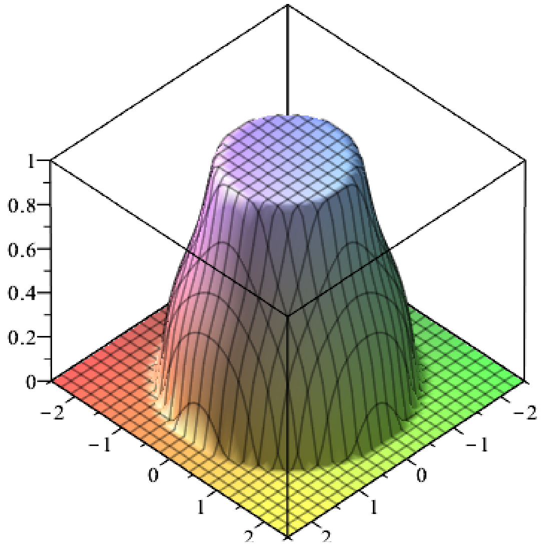


Figure 10.5: The graph of $b\colon \mathbb{R}^2 \to \mathbb{R}$ with 1/3 omitted.

Finally we define partitions of unity.

Definition 10.5. Let M be a (smooth) manifold. A *partition of unity on* M is a family $\{f_i\}_{i\in I}$ of smooth functions on M (the index set I may be uncountable), such that:

(a) The family of supports $\{\operatorname{supp} f_i\}_{i\in I}$ is locally finite.

(b) For all $i \in I$ and all $p \in M$, we have $0 \leq f_i(p) \leq 1$, and

$$\sum_{i\in I} f_i(p) = 1, \quad \text{for every } p \in M.$$

Note that Condition (b) implies that for every $p \in M$, there must be some $i \in I$ such that $f_i(p) > 0$. Thus $\{\operatorname{supp} f_i\}_{i\in I}$ is a cover of M. If $\{U_\alpha\}_{\alpha\in J}$ is a cover of M, we say that the partition of unity $\{f_i\}_{i\in I}$ is *subordinate* to the cover $\{U_\alpha\}_{\alpha\in J}$ if $\{\operatorname{supp} f_i\}_{i\in I}$ is a refinement of $\{U_\alpha\}_{\alpha\in J}$. When $I = J$ and $\operatorname{supp} f_i \subseteq U_i$, we say that $\{f_i\}_{i\in I}$ is *subordinate* to $\{U_\alpha\}_{\alpha\in I}$ *with the same index set as the partition of unity.*

In Definition 10.5, by Condition (a), for every $p \in M$, there is some open set U with $p \in U$, and U meets only finitely many of the supports $\operatorname{supp} f_i$. So $f_i(p) \neq 0$ for only finitely many $i \in I$, and the infinite sum $\sum_{i\in I} f_i(p)$ is well defined.

Proposition 10.4. *Let X be a topological space which is second-countable and locally compact (thus, also Hausdorff). Then X is paracompact. Moreover, every open cover has a countable, locally finite refinement consisting of open sets with compact closures.*

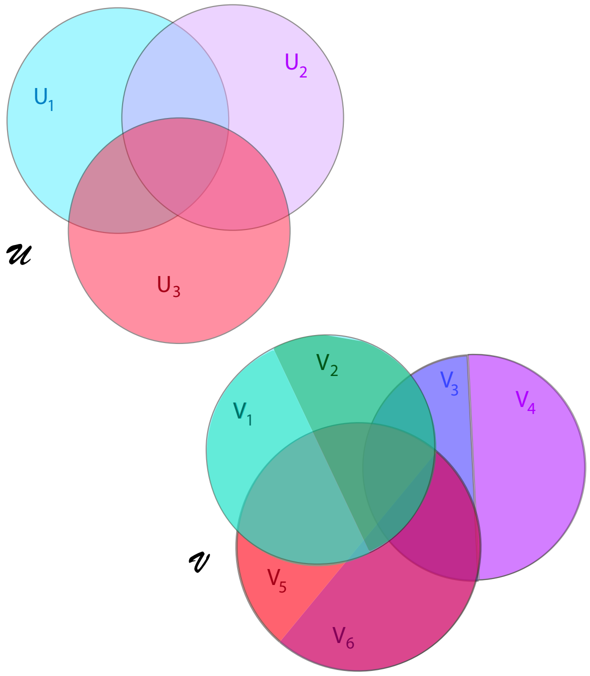


Figure 10.6: Let $\mathcal{U} = U_1 \cup U_2 \cup U_3$. Let $\mathcal{V} = V_1 \cup V_2 \cup V_3 \cup V_4 \cup V_5 \cup V_6$. Then V is a refinement of U with $h\colon \{1,2,3,4,5,6\} \to \{1,2,3\}$ where $h(1) = 1$, $h(2) = 1$, $h(3) = 2$, $h(4) = 2$, $h(5) = 3$, $h(6) = 3$ since $V_1 \subseteq U_1$, $V_2 \subseteq U_1$, $V_3 \subseteq U_2$, $V_4 \subseteq U_2$, $V_5 \subseteq U_3$, $V_6 \subseteq U_3$.

Proof. The proof is quite technical, but since this is an important result, we reproduce Warner's proof for the reader's convenience (Warner [114], Lemma 1.9).

The first step is to construct a sequence of open sets G_i, such that

1. $\overline{G}_i$ is compact,
2. $\overline{G}_i \subseteq G_{i+1}$,
3. $X = \bigcup_{i=1}^{\infty} G_i$.

As X is second-countable, there is a countable basis of open sets $\{U_i\}_{i\geq 1}$ for X. Since X is locally compact, we can find a subfamily of $\{U_i\}_{i\geq 1}$ consisting of open sets with compact closures such that this subfamily is also a basis of X. Therefore, we may assume that we start with a countable basis $\{U_i\}_{i\geq 1}$ of open sets with compact closures. Set $G_1 = U_1$, and assume inductively that

$$G_k = U_1 \cup \cdots \cup U_{j_k}.$$

Since $\overline{G}_k$ is compact, it is covered by finitely many of the U_j's. So, let j_{k+1} be the smallest integer greater than j_k so that

$$\overline{G}_k \subseteq U_1 \cup \cdots \cup U_{j_{k+1}},$$

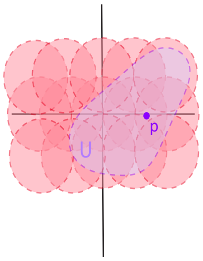


Figure 10.7: Let $X = \mathbb{R}^2$ and $\{U_\alpha\}_{\alpha \in I}$ be the open cover of pink unit disks centered at lattice points (p, q), where $p, q, \in \mathbb{Z}$. For any point $p \in \mathbb{R}^2$, there exists a purple open set U containing p which intersects only finitely many of the pink disks.

and set

$$G_{k+1} = U_1 \cup \cdots \cup U_{j_{k+1}}.$$

See Figure 10.8.

Obviously, the family $\{G_i\}_{i \geq 1}$ satisfies Conditions (1)–(3).

Let $\{U_\alpha\}_{\alpha \in I}$ be an arbitrary open cover of M. For any $i \geq 3$, the set $\overline{G}_i - G_{i-1}$ is compact and contained in the open $G_{i+1} - \overline{G}_{i-2}$. See Figure 10.9. For each $i \geq 3$, choose a finite subcover of the open cover $\{U_\alpha \cap (G_{i+1} - \overline{G}_{i-2})\}_{\alpha \in I}$ of $\overline{G}_i - G_{i-1}$, and choose a finite subcover of the open cover $\{U_\alpha \cap G_3\}_{\alpha \in I}$ of the compact set $\overline{G}_2$. We leave it to the reader to check that this family of open sets is indeed a countable, locally finite refinement of the original open cover $\{U_\alpha\}_{\alpha \in I}$ and consists of open sets with compact closures. □

Remarks:

1. Proposition 10.4 implies that a second-countable, locally compact (Hausdorff) topological space is the union of countably many compact subsets. Thus, X is *countable at infinity*, a notion that we already encountered in Proposition 4.11 and Theorem 4.14. The reason for this odd terminology is that in the Alexandroff one-point compactification of X, the family of open subsets containing the point at infinity (ω) has a countable basis of open sets. (The open subsets containing ω are of the form $(X - K) \cup \{\omega\}$, where K is compact.)

2. A manifold that is countable at infinity has a countable open cover by domains of charts. This is because, if $M = \bigcup_{i \geq 1} K_i$, where the $K_i \subseteq M$ are compact, then for any open cover of M by domains of charts, for every K_i, we can extract a finite subcover, and the union of these finite subcovers is a countable open cover of M by domains of charts. But then, since for every chart (U_i, φ_i), the map φ_i is a homeomorphism onto some open subset of $\mathbb{R}^n$, which is second-countable, so we deduce easily that M

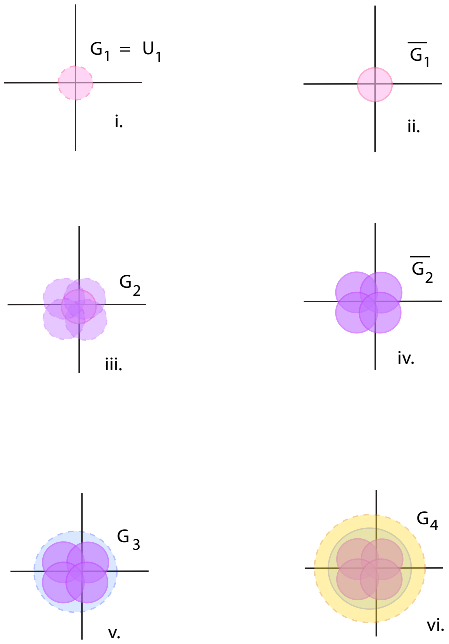


Figure 10.8: The construction of $\{G_i\}_{i=1}^4$ for $X = \mathbb{R}^2$ and $\{U_i\}_{i\geq 1}$, the open disks with rational radius centered at points with rational coordinates.

is second-countable. Thus, for manifolds, second-countable is equivalent to countable at infinity.

We can now prove the main theorem stating the existence of partitions of unity. *Recall that we are assuming that our manifolds are Hausdorff and second-countable.*

Theorem 10.5. *Let M be a smooth manifold and let $\{U_\alpha\}_{\alpha\in I}$ be an open cover for M. Then there is a countable partition of unity $\{f_i\}_{i\geq 1}$ subordinate to the cover $\{U_\alpha\}_{\alpha\in I}$, and the support* $\mathrm{supp}\, f_i$ *of each f_i is compact. If one does not require compact supports, then there is a partition of unity $\{f_\alpha\}_{\alpha\in I}$ subordinate to the cover $\{U_\alpha\}_{\alpha\in I}$ with at most countably many of the f_α not identically zero. (In the second case,* $\mathrm{supp}\, f_\alpha \subseteq U_\alpha$*.)*

Proof. Again, we reproduce Warner's proof (Warner [114], Theorem 1.11). As our manifolds are second-countable, Hausdorff and locally compact, from the proof of Proposition 10.4, we have the sequence of open subsets $\{G_i\}_{i\geq 1}$, and we set $G_0 = \emptyset$. For any $p \in M$, let i_p be the largest integer such that $p \in M - \overline{G}_{i_p}$. Choose an α_p such that $p \in U_{\alpha_p}$; we can find a chart

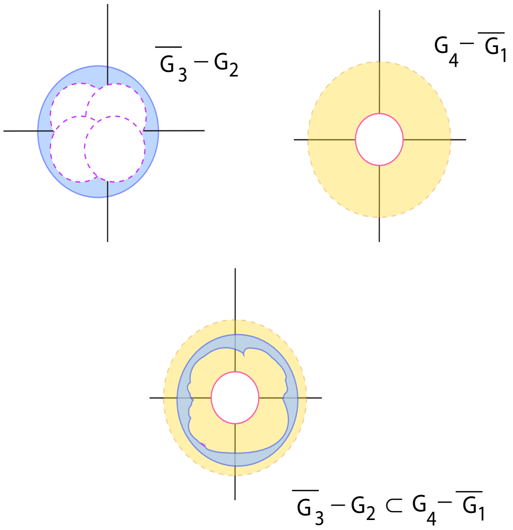


Figure 10.9: The illustration $\overline{G}_3 - G_2 \subset G_4 - \overline{G}_1$, where $\{G_i\}_{i=1}^4$ is illustrated in Figure 10.8.

(U, φ) centered at p such that $U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G}_{i_p})$ and such that $\overline{B(2)} \subseteq \varphi(U)$. Define

$$\psi_p = \begin{cases} b \circ \varphi & \text{on } U \\ 0 & \text{on } M - U, \end{cases}$$

where b is the bump function defined just before Proposition 10.1. Then, ψ_p is a smooth function on M which has value 1 on some open subset W_p containing p and has compact support lying in $U \subseteq U_{\alpha_p} \cap (G_{i_p+2} - \overline{G}_{i_p})$. For each $i \geq 1$, choose a finite set of points $p \in M$, whose corresponding open W_p cover $\overline{G}_i - G_{i-1}$. Order the corresponding ψ_p functions in a sequence ψ_j, $j = 1, 2, \ldots$. The supports of the ψ_j form a locally finite family of subsets of M. Thus, the function

$$\psi = \sum_{j=1}^{\infty} \psi_j$$

is well-defined on M and smooth. Moreover, $\psi(p) > 0$ for each $p \in M$. For each $i \geq 1$, set

$$f_i = \frac{\psi_i}{\psi}.$$

Then the family $\{f_i\}_{i \geq 1}$ is a partition of unity subordinate to the cover $\{U_\alpha\}_{\alpha \in I}$, and supp f_i is compact for all $i \geq 1$. When we don't require compact support, if we let f_α be identically zero if no f_i has support in U_α and otherwise let f_α be the sum of the f_i with support in U_α, then we obtain a partition of unity subordinate to $\{U_\alpha\}_{\alpha \in I}$ with at most countably many of the f_α not identically zero. We must have $\operatorname{supp} f_\alpha \subseteq U_\alpha$, because for any locally finite family of closed sets $\{F_\beta\}_{\beta \in J}$, we have $\overline{\bigcup_{\beta \in J} F_\beta} = \bigcup_{\beta \in J} F_\beta$. □

We close this section by stating a famous theorem of Whitney whose proof uses partitions of unity.

Theorem 10.6. *(Whitney, 1935) Any smooth manifold (Hausdorff and second-countable) M of dimension n is diffeomorphic to a closed submanifold of $\mathbb{R}^{2n+1}$.*

For a proof, see Hirsch [61], Chapter 2, Section 2, Theorem 2.14.

10.2 Covering Maps and Universal Covering Manifolds

Covering maps are an important technical tool in algebraic topology, and more generally in geometry. This brief section only gives some basic definitions and states a few major facts. Appendix A of O'Neill [91] gives a review of definitions and main results about covering manifolds. Expositions including full details can be found in Hatcher [57], Greenberg [52], Munkres [89], Fulton [45], and Massey [78, 79] (the most extensive).

We begin with covering maps.

Definition 10.6. A map $\pi\colon M \to N$ between two smooth manifolds is a *covering map* (or *cover*) iff

(1) The map π is smooth and surjective.

(2) For any $q \in N$, there is some open subset $V \subseteq N$ so that $q \in V$ and

$$\pi^{-1}(V) = \bigcup_{i \in I} U_i,$$

where the U_i are pairwise disjoint open subsets $U_i \subseteq M$, and $\pi\colon U_i \to V$ is a diffeomorphism for every $i \in I$. We say that V is *evenly covered*.

The manifold M is called a *covering manifold* of N. See Figure 10.10.

It is useful to note that a covering map $\pi\colon M \to N$ is a local diffeomorphism (which means that $d\pi_p\colon T_pM \to T_{\pi(p)}N$ is a bijective linear map for every $p \in M$). Indeed, given any $p \in M$, if $q = \pi(p)$, then there is some open subset $V \subseteq N$ containing q so that V is evenly covered by a family of disjoint open subsets $\{U_i\}_{i\in I}$, with each $U_i \subseteq M$ diffeomorphic to V under π. As $p \in U_i$ for some i, we have a diffeomorphism $\pi \restriction U_i\colon U_i \longrightarrow V$, as required.

Definition 10.7. A *homomorphism* of coverings $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$ is a smooth map $\phi\colon M_1 \to M_2$, so that

$$\pi_1 = \pi_2 \circ \phi;$$

that is, the following diagram commutes.

$$\begin{tikzcd}
M_1 \arrow[rr, "\phi"] \arrow[dr, "\pi_1"'] & & M_2 \arrow[dl, "\pi_2"] \\
& N &
\end{tikzcd}$$

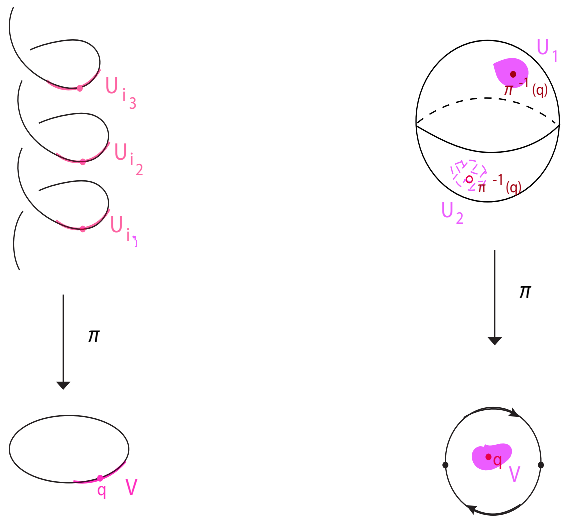


Figure 10.10: Two examples of a covering map. The left illustration is $\pi\colon \mathbb{R} \to S^1$ with $\pi(t) = (\cos(2\pi t), \sin(2\pi t))$, while the right illustration is the 2-fold antipodal covering of $\mathbb{RP}^2$ by S^2.

We say that the coverings $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$ are *equivalent* iff there is a homomorphism $\phi\colon M_1 \to M_2$ between the two coverings, and ϕ is a diffeomorphism.

As usual, the inverse image $\pi^{-1}(q)$ of any element $q \in N$ is called the *fibre over* q, the space N is called the *base*, and M is called the *covering space*. As π is a covering map, each fibre is a discrete space. Note that a homomorphism maps each fibre $\pi_1^{-1}(q)$ in M_1 to the fibre $\pi_2^{-1}(\phi(q))$ in M_2, for every $q \in M_1$.

Proposition 10.7. *Let $\pi\colon M \to N$ be a covering map. If N is connected, then all fibres $\pi^{-1}(q)$ have the same cardinality for all $q \in N$. Furthermore, if $\pi^{-1}(q)$ is not finite, then it is countably infinite.*

Proof. Pick any point, $p \in N$. We claim that the set

$$S = \{q \in N \mid |\pi^{-1}(q)| = |\pi^{-1}(p)|\}$$

is open and closed.

If $q \in S$, then there is some open subset V with $q \in V$, so that $\pi^{-1}(V)$ is evenly covered by some family $\{U_i\}_{i\in I}$ of disjoint open subsets U_i, each diffeomorphic to V under π. Then every $s \in V$ must have a unique preimage in each U_i, so

$$|I| = |\pi^{-1}(s)|, \qquad \text{for all } s \in V.$$

However, as $q \in S$, $|\pi^{-1}(q)| = |\pi^{-1}(p)|$, so

$$|I| = |\pi^{-1}(p)| = |\pi^{-1}(s)|, \qquad \text{for all } s \in V,$$

and thus, $V \subseteq S$. Therefore, S is open. Similarly the complement of S is open. As N is connected, $S = N$.

Since M is a manifold, it is second-countable, that is every open subset can be written as some countable union of open subsets. But then, every family $\{U_i\}_{i\in I}$ of pairwise disjoint open subsets forming an even cover must be countable, and since $|I|$ is the common cardinality of all the fibres, every fibre is countable. $\square$

When the common cardinality of fibres is finite, it is called the *multiplicity* of the covering (or the number of *sheets*).

For any integer, $n > 0$, the map $z \mapsto z^n$ from the unit circle $S^1 = \mathbf{U}(1)$ to itself is a covering with n sheets. The map,

$$t\colon \mapsto (\cos(2\pi t), \sin(2\pi t)),$$

is a covering $\mathbb{R} \to S^1$, with infinitely many sheets.

Definition 10.8. Let $\pi\colon M \to N$ be a covering map, and let P be a Hausdorff topological space. For any map $\phi\colon P \to N$, a *lift of ϕ through π* is a map $\widetilde{\phi}\colon P \to M$ so that

$$\phi = \pi \circ \widetilde{\phi},$$

as in the following commutative diagram.

$$\begin{tikzcd}
 & M \arrow[d, "\pi"] \\
P \arrow[ur, "\widetilde{\phi}"] \arrow[r, "\phi"'] & N
\end{tikzcd}$$

The crucial property of covering manifolds is that curves in N can be lifted to M, in a unique way.

We would like to state three propositions regarding covering spaces. However, two of these propositions use the notion of a simply connected manifold. Intuitively, a manifold is simply connected if it has no "holes." More precisely, a manifold is simply connected if it has a trivial fundamental group. Those readers familiar with the fundamental group may proceed directly to Proposition 10.12 as we now provide a brief review of the fundamental group construction based on Sections 5.1 and 5.2 of Armstrong [5].

A fundamental group is a homotopic loop group. Therefore, given topological spaces X and Y, we need to define a homotopy between two continuous functions $f\colon X \to Y$ and $g\colon X \to Y$.

Definition 10.9. Let X and Y be topological spaces, $f\colon X \to Y$ and $g\colon X \to Y$ be two continuous functions, and let $I = [0,1]$. We say that f *is homotopic to* g if there exists a continuous function $F\colon X \times I \to Y$ (where $X \times I$ is given the product topology) such that $F(x,0) = f(x)$ and $F(x,1) = g(x)$ for all $x \in X$. The map F is a *homotopy from* f *to* g, and this is denoted $f \sim_F g$. If f and g agree on $A \subseteq X$, i.e. $f(a) = g(a)$ whenever $a \in A$, we say f *is homotopic to* g *relative* A, and this is denoted $f \sim_F g$ rel A.

A homotopy provides a means of continuously deforming f into g through a family $\{f_t\}$ of continuous functions $f_t\colon X \to Y$ where $t \in [0,1]$ and $f_0(x) = f(x)$ and $f_1(x) = g(x)$ for all $x \in X$. For example, let D be the unit disk in $\mathbb{R}^2$ and consider two continuous functions $f\colon I \to D$ and $g\colon I \to D$. Then $f \sim_F g$ via the straight line homotopy $F\colon I \times I \to D$, where $F(x,t) = (1-t)f(x) + tg(x)$.

Proposition 10.8. *Let X and Y be topological spaces and let $A \subseteq X$. Homotopy (or homotopy rel A) is an equivalence relation on the set of all continuous functions from X to Y.*

The next two propositions show that homotopy behaves well with respect to composition.

Proposition 10.9. *Let X, Y, and Z be topological spaces and let $A \subseteq X$. For any continuous functions $f\colon X \to Y$, $g\colon X \to Y$, and $h\colon Y \to Z$, if $f \sim_F g$ rel A, then $h \circ f \sim_{h\circ F} h \circ g$ rel A as maps from X to Z.*

$$\begin{tikzcd} X \arrow[r, shift left, "f"] \arrow[r, shift right, "g"'] & Y \arrow[r, "h"] & Z. \end{tikzcd}$$

Proposition 10.10. *Let X, Y, and Z be topological spaces and let $B \subseteq Y$. For any continuous functions $f\colon X \to Y$, $g\colon Y \to Z$, and $h\colon Y \to Z$, if $g \sim_G h$ rel B, then $g \circ f \sim_F h \circ f$ rel $f^{-1}B$, where $F(x,t) = G(f(x),t)$.*

$$\begin{tikzcd} X \arrow[r, "f"] & Y \arrow[r, shift left, "g"] \arrow[r, shift right, "h"'] & Z. \end{tikzcd}$$

In order to define the fundamental group of a topological space X, we recall the definition of a loop.

Definition 10.10. Let X be a topological space, p be a point in X, and let $I = [0,1]$. We say α is a *loop based at* $p = \alpha(0)$ if α is a continuous map $\alpha\colon I \to X$ with $\alpha(0) = \alpha(1)$.

Given a topological space X, choose a point $p \in X$ and form S, the set of all loops in X based at p. By applying Proposition 10.8, we know that the relation of homotopy relative to $\{0,1\}$ is an equivalence relation on S. This leads to the following definition.

Definition 10.11. Let X be a topological space, p be a point in X, and let α be a loop in X based at p. The set of all loops homotopic to α relative to $\{0,1\}$ is *the homotopy class of* α and is denoted $\langle\alpha\rangle$.

Definition 10.12. Given two loops α and β in a topological space X based at p, the *product* $\alpha \cdot \beta$ is a loop in X based at p defined by

$$\alpha \cdot \beta(t) = \begin{cases} \alpha(2t), & 0 \leq t \leq \frac{1}{2} \\ \beta(2t-1), & \frac{1}{2} < t \leq 1. \end{cases}$$

The product of loops gives rise to the product of homotopy classes where

$$\langle \alpha \rangle \cdot \langle \beta \rangle = \langle \alpha \cdot \beta \rangle.$$

We leave it the reader to check that the multiplication of homotopy classes is well defined and associative, namely $\langle \alpha \cdot \beta \rangle \cdot \langle \gamma \rangle = \langle \alpha \rangle \cdot \langle \beta \cdot \gamma \rangle$ whenever α, β, and γ are loops in X based at p.

Let $\langle e \rangle$ be the homotopy class of the constant loop in X based at p, and define the inverse of $\langle \alpha \rangle$ as $\langle \alpha \rangle^{-1} = \langle \alpha^{-1} \rangle$, where $\alpha^{-1}(t) = \alpha(1-t)$. With these conventions, the product operation between homotopy classes gives rise to a group. In particular,

Proposition 10.11. *Let X be a topological space and let p be a point in X. The set of homotopy classes of loops in X based at p is a group with multiplication given by $\langle \alpha \rangle \cdot \langle \beta \rangle = \langle \alpha \cdot \beta \rangle$*

Definition 10.13. Let X be a topological space and p a point in X. The group of homotopy classes of loops in X based at p is the *fundamental group of X based at p*, and is denoted by $\pi_1(X, p)$.

If we assume X is path connected, we can show that $\pi_1(X, p) \cong \pi_1(X, q)$ for any points p and q in X. Therefore, when X is path connected, we simply write $\pi_1(X)$.

Definition 10.14. If X is path connected topological space and $\pi_1(X) = \langle e \rangle$, (which is also denoted as $\pi_1(X) = (0)$), we say X is *simply connected.*

In other words, every loop in X can be shrunk in a continuous manner within X to its basepoint. Examples of simply connected spaces include R^n and S^n whenever $n \geq 2$. On the other hand, the torus and the circle are not simply connected. See Figures 10.11 and 10.12.

We now state without proof the following results regarding covering spaces.

Proposition 10.12. *If $\pi \colon M \to N$ is a covering map, then for every smooth curve $\alpha \colon I \to N$ in N (with $0 \in I$) and for any point $q \in M$ such that $\pi(q) = \alpha(0)$, there is a unique smooth curve $\widetilde{\alpha} \colon I \to M$ lifting α through π such that $\widetilde{\alpha}(0) = q$. See Figure 10.13.*

Proposition 10.13. *Let $\pi \colon M \to N$ be a covering map and let $\phi \colon P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \phi(p_0) = r_0$, the following properties hold:*

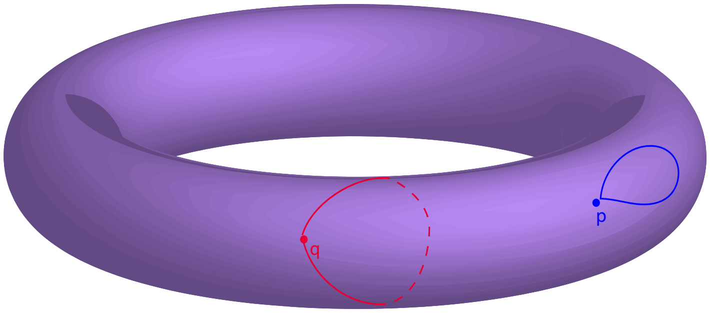


Figure 10.11: The torus is not simply connected. The loop at p is homotopic to a point, but the loop at q is not.

(1) If P is connected then there is at most one lift $\widetilde{\phi}\colon P \to M$ of ϕ through π such that $\widetilde{\phi}(p_0) = q_0$.

(2) If P is simply connected, then such a lift exists.

$$\begin{tikzcd}
 & M \ni q_0 \arrow[d, "\pi"] \\
p_0 \in P \arrow[ur, "\widetilde{\phi}"] \arrow[r, "\phi"'] & N \ni r_0
\end{tikzcd}$$

Theorem 10.14. *Every connected manifold M possesses a simply connected covering map $\pi\colon \widetilde{M} \to M$; that is, with $\widetilde{M}$ simply connected. Any two simply connected coverings of N are equivalent.*

Definition 10.15. In view of Theorem 10.14, it is legitimate to speak of *the* simply connected cover $\widetilde{M}$ of M, also called *universal covering* (or *cover*) of M.

Given any point $p \in M$, let $\pi_1(M, p)$ denote the fundamental group of M with basepoint p. See Definition 10.13. If $\phi\colon M \to N$ is a smooth map, for any $p \in M$, if we write $q = \phi(p)$, then we have an induced group homomorphism

$$\phi_*\colon \pi_1(M, p) \to \pi_1(N, q).$$

Proposition 10.15. *If $\pi\colon M \to N$ is a covering map, for every $p \in M$, if $q = \pi(p)$, then the induced homomorphism $\pi_*\colon \pi_1(M, p) \to \pi_1(N, q)$ is injective.*

The next proposition is a stronger version of Part (1) of Proposition 10.13.

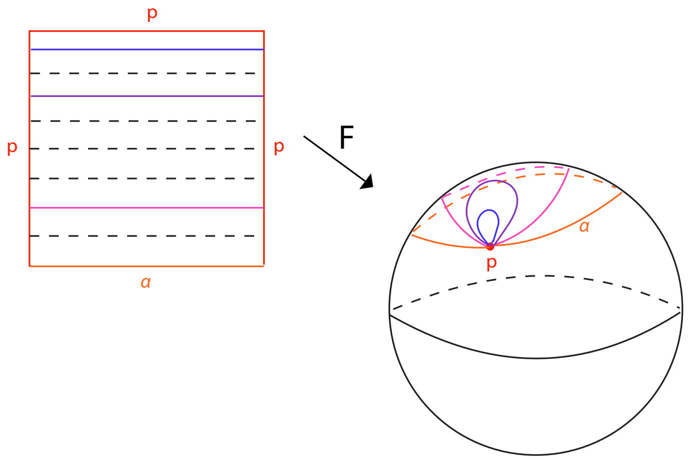


Figure 10.12: The unit sphere S^2 is simply connected since every loop can be continuously deformed to a point. This deformation is represented by the map $F\colon I \times I \to S^2$ where $F(x,0) = \alpha$ and $F(x,1) = p$.

Proposition 10.16. *Let $\pi\colon M \to N$ be a covering map and let $\phi\colon P \to N$ be a smooth map. For any $p_0 \in P$, any $q_0 \in M$ and any $r_0 \in N$ with $\pi(q_0) = \phi(p_0) = r_0$, if P is connected, then a lift $\widetilde{\phi}\colon P \to M$ of ϕ such that $\widetilde{\phi}(p_0) = q_0$ exists iff*

$$\phi_*(\pi_1(P,p_0)) \subseteq \pi_*(\pi_1(M,q_0)),$$

as illustrated in the diagram below.

$$\begin{tikzcd}
& M \arrow[d, "\pi"] \\
P \arrow[ur, dashed, "\widetilde{\phi}"] \arrow[r, "\phi"'] & N
\end{tikzcd}$$

iff

$$\begin{tikzcd}
& \pi_1(M,q_0) \arrow[d, "\pi_*"] \\
\pi_1(P,p_0) \arrow[ur, dashed] \arrow[r, "\phi_*"'] & \pi_1(N,r_0)
\end{tikzcd}$$

Basic Assumption: For any covering $\pi\colon M \to N$, if N is connected then we also assume that M is connected.

Using Proposition 10.15, we get

Proposition 10.17. *If $\pi\colon M \to N$ is a covering map and N is simply connected, then π is a diffeomorphism (recall that M is connected); thus, M is diffeomorphic to the universal cover $\widetilde{N}$, of N.*

Proof. Pick any $p \in M$ and let $q = \pi(p)$. As N is simply connected, $\pi_1(N,q) = (0)$. By Proposition 10.15, since $\pi_*\colon \pi_1(M,p) \to \pi_1(N,q)$ is injective, $\pi_1(M,p) = (0)$, so M is simply

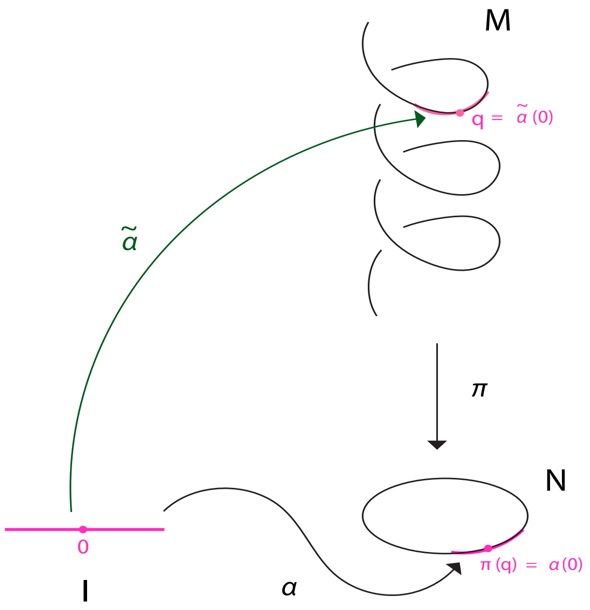


Figure 10.13: The lift of a curve α when $\pi\colon \mathbb{R} \to S^1$ is $\pi(t) = (\cos(2\pi t), \sin(2\pi t))$.

connected (by hypothesis, M is connected). But then, by Theorem 10.14, M and N are diffeomorphic. □

The following proposition shows that the universal covering of a space covers every other covering of that space. This justifies the terminology "universal covering."

Proposition 10.18. *Say $\pi_1\colon M_1 \to N$ and $\pi_2\colon M_2 \to N$ are two coverings of N, with N connected. Every homomorphism $\phi\colon M_1 \to M_2$ between these two coverings is a covering map.*

$$\begin{tikzcd} M_1 \arrow[rr, "\phi"] \arrow[dr, "\pi_1"'] & & M_2 \arrow[dl, "\pi_2"] . \\ & N & \end{tikzcd}$$

As a consequence, if $\pi\colon \widetilde{N} \to N$ is a universal covering of N, then for every covering $\pi'\colon M \to N$ of N, there is a covering $\phi\colon \widetilde{N} \to M$ of M.

The notion of deck-transformation group of a covering is also useful because it yields a way to compute the fundamental group of the base space.

Definition 10.16. If $\pi\colon M \to N$ is a covering map, a *deck-transformation* is any diffeomorphism $\phi\colon M \to M$ such that $\pi = \pi \circ \phi$; that is, the following diagram commutes.

$$\begin{tikzcd}
M \arrow[rr, "\phi"] \arrow[dr, "\pi"'] & & M \arrow[dl, "\pi"] \\
& N &
\end{tikzcd}$$

Note that deck-transformations are just automorphisms of the covering map. The commutative diagram of Definition 10.16 means that a deck transformation permutes every fibre. It is immediately verified that the set of deck transformations of a covering map is a group under composition denoted Γ_π (or simply Γ), called the *deck-transformation group* of the covering.

Observe that any deck transformation ϕ is a lift of π through π. Consequently, if M is connected, by Proposition 10.13 (1), every deck-transformation is determined by its value at a single point. So, the deck-transformations are determined by their action on each point of any fixed fibre $\pi^{-1}(q)$, with $q \in N$. Since the fibre $\pi^{-1}(q)$ is countable, Γ is also countable, that is, a discrete Lie group. Moreover, if M is compact, as each fibre $\pi^{-1}(q)$ is compact and discrete, it must be finite and so, the deck-transformation group is also finite.

The following proposition gives a useful method for determining the fundamental group of a manifold.

Proposition 10.19. *If $\pi\colon \widetilde{M} \to M$ is the universal covering of a connected manifold M, then the deck-transformation group $\widetilde{\Gamma}$ is isomorphic to the fundamental group $\pi_1(M)$ of M.*

Remark: When $\pi\colon \widetilde{M} \to M$ is the universal covering of M, it can be shown that the group $\widetilde{\Gamma}$ acts simply and transitively on every fibre $\pi^{-1}(q)$. This means that for any two elements $x, y \in \pi^{-1}(q)$, there is a unique deck-transformation $\phi \in \widetilde{\Gamma}$ such that $\phi(x) = y$. So, there is a bijection between $\pi_1(M) \cong \widetilde{\Gamma}$ and the fibre $\pi^{-1}(q)$.

Proposition 10.14 together with previous observations implies that if the universal cover of a connected (compact) manifold is compact, then M has a finite fundamental group. We will use this fact later, in particular in the proof of Myers' Theorem.

10.3 Problems

Problem 10.1. Consider the function $h\colon \mathbb{R} \to \mathbb{R}$ given by

$$h(x) = \begin{cases} e^{-1/x} & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

Show by induction that for all $k \geq 0$ and all $x > 0$, the kth derivative $h^{(k)}(x)$ is of the form $P_{2k}(1/x)e^{-1/x}$ for some polynomial $P_{2k}(X)$ of degree $2k$ in X. Prove that h is smooth on $\mathbb{R}$ and that $h^{(k)}(0) = 0$ for all $k \geq 0$.

Problem 10.2. Define $b\colon \mathbb{R}^n \to \mathbb{R}$ by

$$b(x_1, \ldots, x_n) = \frac{h((4 - x_1^2 - \cdots - x_n^2)/3)}{h((4 - x_1^2 - \cdots - x_n^2)/3) + h((x_1^2 + \cdots + x_n^2 - 1)/3)}.$$

Verify that b satisfies the conditions of Proposition 10.1.

Problem 10.3. Let $(A_i)_{i\in I}$ be a locally finite family of subsets of a topological space X. Show that every compact set K in X has a neighborhood W that intersects only finitely many of the A_i.

Problem 10.4. Let $(A_i)_{i\in I}$ be a locally finite family of subsets of a topological space X. Show that

$$\overline{\bigcup_{i\in I} A_i} = \bigcup_{i\in I} \overline{A}_i.$$

Note that the inclusion

$$\bigcup_{i\in I} \overline{A}_i \subseteq \overline{\bigcup_{i\in I} A_i}$$

holds for any family $(A_i)_{i\in I}$ of subsets of X, but the inclusion

$$\overline{\bigcup_{i\in I} A_i} \subseteq \bigcup_{i\in I} \overline{A}_i$$

is not true in general. For example, let $A_i = [0, 1 - 1/i]$, for $i \in \mathbb{N} - \{0\}$.

Problem 10.5. Provide the details at the end of the proof of Proposition 10.4.

Problem 10.6. Check that the supports of the functions ψ_j constructed during the proof of Theorem 10.5 form a locally finite family.

Problem 10.7. Check that the multiplication of homotopy classes given after Definition 10.11 is well defined and associative. Show that setting $\langle\alpha\rangle^{-1} = \langle\alpha^{-1}\rangle$ defines an inverse with respect to multiplication of homotopy classes.

Problem 10.8. Prove Proposition 10.8.

Problem 10.9. Prove Propositions 10.9 and 10.10.

Problem 10.10. Prove that if X is path connected, then $\pi_1(X, p) \cong \pi_1(X, q)$ for any points p and q in X.

Chapter 11

Basic Analysis: Review of Series and Derivatives

The goal of Chapter 3 is to define *embedded submanifolds* and linear Lie groups. Before doing this, we believe that some readers might appreciate a review of the basic properties of power series involving matrix coefficients and a review of the notion of the *derivative* of a function between two normed vector spaces. Those readers familiar with these concepts may proceed directly to Chapter 3.

11.1 Series and Power Series of Matrices

Since a number of important functions on matrices are defined by power series, in particular the exponential, we review quickly some basic notions about series in a complete normed vector space.

Given a normed vector space $(E, \| \|)$, a *series* is an infinite sum $\sum_{k=0}^{\infty} a_k$ of elements $a_k \in E$. We denote by S_n the partial sum of the first $n+1$ elements,

$$S_n = \sum_{k=0}^{n} a_k.$$

Definition 11.1. We say that the series $\sum_{k=0}^{\infty} a_k$ *converges* to the limit $a \in E$ if the sequence (S_n) converges to a, i.e., given any $\epsilon > 0$, there exists a positive integer N such that for all $n \geq N$,

$$\|S_n - a\| < \epsilon.$$

In this case, we say that the series is *convergent*. We say that the series $\sum_{k=0}^{\infty} a_k$ *converges absolutely* if the series of norms $\sum_{k=0}^{\infty} \|a_k\|$ is convergent.

If the series $\sum_{k=0}^{\infty} a_k$ converges to a, since for all m, n with $m > n$ we have

$$\sum_{k=0}^{m} a_k - S_n = \sum_{k=0}^{m} a_k - \sum_{k=0}^{n} a_k = \sum_{k=n+1}^{m} a_k,$$

if we let m go to infinity (with n fixed), we see that the series $\sum_{k=n+1}^{\infty} a_k$ converges and that

$$a - S_n = \sum_{k=n+1}^{\infty} a_k.$$

To intuitively understand Definition 11.1, think of (a_n) as a long string or "snake" of vector entries. We subdivide this snake into head, body, and tail by choosing $m > n \geq 0$ and writing

$$\sum_{k=0}^{\infty} a_k = H + B + T,$$

where

$$\begin{aligned} H &= \sum_{k=0}^{n} a_k = a_0 + a_1 + \cdots + a_n, \\ B &= \sum_{k=n+1}^{m} a_k = a_{n+1} + a_{n+2} + \cdots + a_m, \\ T &= \sum_{k=m+1}^{\infty} a_k = a_{m+1} + a_{m+2} + \ldots \quad . \end{aligned}$$

Note H stands for head, B stands for body, and T stands for tail. The convergence of $\sum_{k=0}^{\infty} a_k$ means T is arbitrarily small whenever m is "large enough". See Figure 11.1.

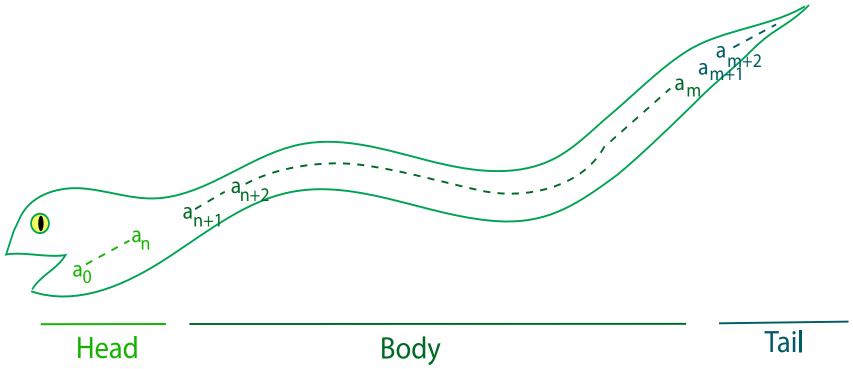


Figure 11.1: The "snake" view of the sequence (a_n).

In particular, we have the following useful proposition.

Proposition 11.1. *If $\sum_{k=0}^{\infty} a_k$ converges, then $\lim_{k\mapsto\infty} a_k = \lim_{k\mapsto\infty} \|a_k\| = 0$. Given $N \geq 0$ and a fixed positive value s, if $\|a_k\| > s > 0$ infinitely many times whenever $k \geq N$, then $\sum_{k=0}^{\infty} a_k$ diverges.*

The "belly" of the snake may be characterized in terms of a Cauchy sequence.

Definition 11.2. Given a normed vector space, E, we say that a sequence, (a_n), with $a_n \in E$, is a *Cauchy sequence* iff for every $\epsilon > 0$, there is some $N > 0$ so that for all $m, n \geq N$,

$$\|a_n - a_m\| < \epsilon.$$

Definition 11.3. A normed vector space, E, is *complete* iff every Cauchy sequence converges. A complete normed vector space is also called a *Banach space*, after Stefan Banach (1892-1945).

There are series that are convergent but not absolutely convergent; for example, the series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}.$$

If E is complete, the converse is an enormously useful result.

Proposition 11.2. *Assume $(E, \| \, \|)$ is a complete normed vector space. If a series $\sum_{k=0}^{\infty} a_k$ is absolutely convergent, then it is convergent.*

Proof. If $\sum_{k=0}^{\infty} a_k$ is absolutely convergent, then we prove that the sequence (S_m) is a Cauchy sequence; that is, for every $\epsilon > 0$, there is some $p > 0$ such that for all $n \geq m \geq p$,

$$\|S_n - S_m\| \leq \epsilon.$$

Observe that

$$\|S_n - S_m\| = \|a_{m+1} + \cdots + a_n\| \leq \|a_{m+1}\| + \cdots + \|a_n\|,$$

and since the sequence $\sum_{k=0}^{\infty} \|a_k\|$ converges, it satisfies Cauchy's criterion. Thus, the sequence (S_m) also satisfies Cauchy's criterion, and since E is a complete vector space, the sequence (S_m) converges. □

Remark: It can be shown that if $(E, \| \, \|)$ is a normed vector space such that every absolutely convergent series is also convergent, then E must be complete (see Schwartz [103]).

An important corollary of absolute convergence is that if the terms in series $\sum_{k=0}^{\infty} a_k$ are rearranged, then the resulting series is still absolutely convergent, and has the same sum. More precisely, let σ be any permutation (bijection) of the natural numbers. The series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is called a *rearrangement* of the original series. The following result can be shown (see Schwartz [103]).

Proposition 11.3. *Assume $(E, \| \|)$ is a normed vector space. If a series $\sum_{k=0}^{\infty} a_k$ is convergent as well as absolutely convergent, then for every permutation σ of $\mathbb{N}$, the series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is convergent and absolutely convergent, and its sum is equal to the sum of the original series:*

$$\sum_{k=0}^{\infty} a_{\sigma(k)} = \sum_{k=0}^{\infty} a_k.$$

In particular, if $(E, \| \|)$ is a complete normed vector space, then Proposition 11.3 holds. A series $\sum_{k=0}^{\infty} a_k$ is said to be *unconditionally convergent* (or *commutatively convergent*) if the series $\sum_{k=0}^{\infty} a_{\sigma(k)}$ is convergent for every permutation σ of $\mathbb{N}$, and if all these rearrangements have the same sum. It can be shown that if E has finite dimension, then a series is absolutely convergent iff it is unconditionally convergent. However, this is false if E has infinite dimension (but hard to prove).

If $E = \mathbb{C}$, there are several conditions that imply the absolute convergence of a series. In the rest of this section we omit most proofs, details of which can be found in introductory analysis books such as Apostol [4] and Schwartz [103].

The *ratio test* is the following test. Suppose there is some $N > 0$ such that $a_n \neq 0$ for all $n \geq N$, and either

$$r = \lim_{n \mapsto \infty} \left| \frac{a_{n+1}}{a_n} \right|$$

exists, or the sequence of ratios diverges to infinity, in which case we write $r = \infty$. Then, if $0 \leq r < 1$, the series $\sum_{k=0}^{n} a_k$ converges absolutely, else if $1 < r \leq \infty$, the series diverges.

If (r_n) is a sequence of real numbers, recall that

$$\limsup_{n \mapsto \infty} r_n = \lim_{n \mapsto \infty} \sup_{k \geq n} \{r_k\}.$$

If $r_n \geq 0$ for all n, then either the sequence (r_n) is unbounded, in which case $\sup_{k \geq n}\{r_k\}$ is infinite for all n and $\limsup_{n \mapsto \infty} r_n = \infty$, or the sequence (r_n) is bounded, and since $\sup_{k \geq n+1}\{r_k\} \leq \sup_{k \geq n}\{r_k\}$, the sequence $(\sup_{k \geq n})_{n \geq 0}$ is nonincreasing and bounded from below by 0, so $\limsup_{n \mapsto \infty} r_n = r$ exists and is finite. In this case, it is easy to see that r is characterized as follows:

For every $\epsilon > 0$, there is some $N \in \mathbb{N}$ such that $r_n < r + \epsilon$ for all $n \geq N$, and $r_n > r - \epsilon$ for infinitely many n.

The notion of $\limsup_{n \mapsto \infty} r_n$ may also be characterized in terms of limits of subsequences. Take the family of all subsequences $\{(r_{n_j})\}$ of (r_n). Consider the set, L, of all possible limits of these subsequences. Then $\limsup_{n \mapsto \infty} r_n$ is the largest element (possibly infinity) of L. For example if $(r_n) = (1, -1, 1, -1, \ldots)$, then $L = \{-1, 1\}$ and $\limsup_{n \mapsto \infty} r_n = 1$.

The *root test* is this. Let

$$r = \limsup_{n \mapsto \infty} |a_n|^{1/n}$$

if the limit exists (is finite), else write $r = \infty$. Then, if $0 \leq r < 1$, the series $\sum_{k=0}^{n} a_k$ converges absolutely, else if $1 < r \leq \infty$, the series diverges.

The root test also applies if $(E, \| \, \|)$ is a complete normed vector space by replacing $|a_n|$ by $\|a_n\|$. Let $\sum_{k\geq 0}^{\infty} a_k$ be a series of elements $a_k \in E$ and let

$$r = \limsup_{n \mapsto \infty} \|a_n\|^{1/n}$$

if the limit exists (is finite), else write $r = \infty$. Then, if $0 \leq r < 1$, the series $\sum_{k=0}^{n} a_k$ converges absolutely, else if $1 < r \leq \infty$, the series diverges.

A *power series* with coefficients $a_k \in \mathbb{C}$ in the indeterminate z is a formal expression $f(z)$ of the form

$$f(z) = \sum_{k=0}^{\infty} a_k z^k,$$

For any fixed value $z \in \mathbb{C}$, the series $f(z)$ may or may not converge. It always converges for $z = 0$, since $f(0) = a_0$. A fundamental fact about power series is that they have a *radius of convergence*.

Proposition 11.4. *Given any power series*

$$f(z) = \sum_{k=0}^{\infty} a_k z^k,$$

there is a nonnegative real R, possibly infinite, called the **radius of convergence** *of the power series, such that if $|z| < R$, then $f(z)$ converges absolutely, else if $|z| > R$, then $f(z)$ diverges. Moreover (Hadamard), we have*

$$R = \frac{1}{\limsup_{n\mapsto\infty} |a_n|^{1/n}}.$$

Note that Proposition 11.4 does not say anything about the behavior of the power series for boundary values, that is, values of z such that $|z| = R$.

Proof. Given $\sum_{n=0}^{\infty} A_n$, where (A_n) is an arbitrary sequence of complex numbers, note that $\sum_{n=0}^{\infty} |A_n| = \sum_{n=0}^{\infty} \left[|A_n|^{\frac{1}{n}} \right]^n$. If $\limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}} < 1$, then $\sum_{n=0}^{\infty} A_n$ converges absolutely. To see why this is the case, observe that the definition of $\limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}}$ implies that given $\epsilon > 0$, there exists $N(\epsilon)$ such that

$$|A_n|^{\frac{1}{n}} \leq \limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}} + \epsilon, \qquad \text{whenever } n > N(\epsilon).$$

Choose ϵ small enough so that

$$|A_n|^{\frac{1}{n}} \leq \limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}} + \epsilon \leq r_1 < 1.$$

Then

$$\sum_{n=N(\epsilon)+1}^{\infty} |A_n| \leq \sum_{n=N(\epsilon)+1} \left[|A_n|^{\frac{1}{n}}\right]^n \leq \sum_{n=N(\epsilon)+1}^{\infty} r_1^n = \frac{r_1^{N(\epsilon)+1}}{1-r_1},$$

and an application of the comparison test implies that $\sum_{n=0}^{\infty} A_n$ converges absolutely. It is then a matter of setting $A_n = a_n z^n$ and requiring that

$$\limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}} = |z| \limsup_{n\mapsto\infty} |a_n|^{\frac{1}{n}} < 1.$$

If $\limsup_{n\mapsto\infty} |A_n|^{\frac{1}{n}} = |z| \limsup_{n\mapsto\infty} |a_n|^{\frac{1}{n}} > 1$, the definition of $\limsup_{n\mapsto\infty} |a_n|^{\frac{1}{n}}$ implies that

$$1 < |z|[\limsup_{n\mapsto\infty} |a_n|^{\frac{1}{n}} - \epsilon], \qquad \text{for infinitely many } n.$$

Then Proposition 11.1 implies that $\sum_{n=0}^{\infty} A_n = \sum_{n=0}^{\infty} a_n z^n$ diverges. □

Even though the ratio test does not apply to every power series, it provides a useful way of computing the radius of convergence of a power series.

Proposition 11.5. *Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ be a power series with coefficients $a_k \in \mathbb{C}$. Suppose there is some $N > 0$ such that $a_n \neq 0$ for all $n \geq N$, and either*

$$R = \lim_{n\mapsto\infty} \left|\frac{a_n}{a_{n+1}}\right|$$

exists, or the sequence on the righthand side diverges to infinity, in which case we write $R = \infty$. Then the power series $\sum_{k=0}^{\infty} a_k z^k$ has radius of convergence R.

For example, for the power series

$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!},$$

we have

$$\left|\frac{a_k}{a_{k+1}}\right| = \frac{(k+1)!}{k!} = k+1,$$

whose limit is ∞, so the exponential is defined for all $z \in \mathbb{C}$; its radius of convergence is ∞. For the power series

$$f(z) = \sum_{k=0}^{\infty} \frac{z^k}{(k+1)!},$$

we have

$$\left|\frac{a_k}{a_{k+1}}\right| = \frac{(k+2)!}{(k+1)!} = k+2,$$

so $f(z)$ also has infinite radius of convergence.

For the power series

$$\log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k},$$

if $k \geq 1$ we have

$$\left| \frac{a_k}{a_{k+1}} \right| = \frac{k+1}{k}$$

whose limit is 1, so $\log(1+x)$ has radius of convergence 1. For $x = 1$, the series converges to $\log(2)$, but for $x = -1$, the series diverges.

Power series behave very well with respect to term by term differentiation and term by term integration.

Proposition 11.6. *Suppose the power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$ (with complex coefficients) has radius of convergence $R > 0$. Then, $f'(z)$ exists if $|z| < R$, the power series $\sum_{k=1}^{\infty} k a_k z^{k-1}$ has radius of convergence R, and*

$$f'(z) = \sum_{k=1}^{\infty} k a_k z^{k-1}.$$

Proposition 11.7. *Suppose the power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$ (with complex coefficients) has radius of convergence $R > 0$. Then $F(z) = \int_0^z f(t)\, dt$ exists if $|z| < R$, the power series $\sum_{k=0}^{\infty} \frac{a_k}{k+1} z^{k+1}$ has radius of convergence R, and*

$$F(z) = \sum_{k=0}^{\infty} \frac{a_k}{k+1} z^{k+1}.$$

Let us now assume that $f(z) = \sum_{k=0}^{\infty} a_k z^k$ is a power series with coefficients $a_k \in \mathbb{C}$, and that its radius of convergence is R. Given any matrix $A \in \mathrm{M}_n(\mathbb{C})$ we can form the power series obtained by substituting A for z,

$$f(A) = \sum_{k=0}^{\infty} a_k A^k.$$

Let $\| \ \|$ be any matrix norm on $\mathrm{M}_n(\mathbb{C})$. Then the following proposition regarding the convergence of the power series $f(A)$ holds.

Proposition 11.8. *Let $f(z) = \sum_{k=1}^{\infty} a_k z^k$ be a power series with complex coefficients, write R for its radius of convergence, and assume that $R > 0$. For every ρ such that $0 < \rho < R$, the series $f(A) = \sum_{k=1}^{\infty} a_k A^k$ is absolutely convergent for all $A \in \mathrm{M}_n(\mathbb{C})$ such that $\|A\| \leq \rho$. Furthermore, f is continuous on the open ball $B(R) = \{A \in \mathrm{M}_n(\mathbb{C}) \mid \|A\| < R\}$.*

Note that unlike the case where $A \in \mathbb{C}$, if $\|A\| > R$, we cannot claim that the series $f(A)$ diverges. This has to do with the fact that even for the operator norm we may have $\|A^n\| < \|A\|^n$, a fact which should be contrasted to situation in $\mathbb{C}$ where $|a|^n = |a^n|$. We leave it as an exercise to find an example of a series and a matrix A with $\|A\| > R$, and yet $f(A)$ converges. Hint: Consider A to be nilpotent, i.e. $A \neq 0$ but $A^k = 0$ for some positive integer k.

As an application of Proposition 11.8, the exponential power series

$$e^A = \exp(A) = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

is absolutely convergent for all $A \in \mathrm{M}_n(\mathbb{C})$, and continuous everywhere. Proposition 11.8 also implies that the series

$$\log(I + A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}$$

is absolutely convergent if $\|A\| < 1$.

Next, let us consider the generalization of the notion of a power series $f(t) = \sum_{k=1}^{\infty} a_k t^k$ of a complex variable t, where the coefficients a_k belong to a complete normed vector space $(F, \|\ \|)$. Then it is easy to see that Proposition 11.4 generalizes to this situation.

Proposition 11.9. *Let $(F, \|\ \|)$ be a complete normed vector space. Given any power series*

$$f(t) = \sum_{k=0}^{\infty} a_k t^k,$$

with $t \in \mathbb{C}$ and $a_k \in F$, there is a nonnegative real R, possibly infinite, called the **radius of convergence** *of the power series, such that if $|t| < R$, then $f(t)$ converges absolutely, else if $|t| > R$, then $f(t)$ diverges. Moreover, we have*

$$R = \frac{1}{\limsup_{n \mapsto \infty} \|a_n\|^{1/n}}.$$

Propositions 11.6 and 11.7 also holds in this more general setting and the proofs are the same.

Proposition 11.10. *Let $(F, \|\ \|)$ be a complete normed vector space. Suppose the power series $f(t) = \sum_{k=0}^{\infty} a_k t^k$ (with coefficients $a_k \in F$) has radius of convergence R. Then, $f'(t)$ exists if $|t| < R$, the power series $\sum_{k=1}^{\infty} k a_k t^{k-1}$ has radius of convergence R, and*

$$f'(t) = \sum_{k=1}^{\infty} k a_k t^{k-1}.$$

Proposition 11.11. *Let $(F, \| \|)$ be a complete normed vector space. Suppose the power series $f(t) = \sum_{k=0}^{\infty} a_k t^k$ (with coefficients $a_k \in F$) has radius of convergence $R > 0$. Then $F(t) = \int_0^t f(z)\, dz$ exists if $|t| < R$, the power series $\sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}$ has radius of convergence R, and*

$$F(t) = \sum_{k=0}^{\infty} \frac{a_k}{k+1} t^{k+1}.$$

So far we have considered series as individual entities. We end this section with a discussion on ways to combine pairs of series through addition, multiplication, and composition. Given a complete normed vector space $(E, \| \|)$, if $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ are two series with $a_k, b_k \in E$, we can form the series $\sum_{k=0}^{\infty}(a_k + b_k)$ whose kth terms is $a_k + b_k$, and for any scalar λ, the series $\sum_{k=0}^{\infty} \lambda a_k$, whose kth terms is λa_k. It is easy to see that if $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ are absolutely convergent with sums A and B, respectively, then the series $\sum_{k=0}^{\infty}(a_k + b_k)$ and $\sum_{k=0}^{\infty} \lambda a_k$ are absolutely convergent, and their sums are given by

$$\sum_{k=0}^{\infty}(a_k + b_k) = A + B = \sum_{k=0}^{\infty} a_k + \sum_{k=0}^{\infty} b_k$$

$$\sum_{k=0}^{\infty} \lambda a_k = \lambda A = \lambda \sum_{k=0}^{\infty} a_k.$$

If $f(z) = \sum_{k=0}^{\infty} a_k z^k$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ are two power series with $a_k, b_k \in E$, we can form the power series $h(z) = \sum_{k=0}^{\infty}(a_k + b_k) z^k$, and for any scalar λ, the power series $s(z) = \sum_{k=0}^{\infty} \lambda a_k z^k$. We can show easily that if $f(z)$ has radius of convergence $R(f)$ and $g(z)$ has radius of convergence $R(g)$, then $h(z)$ has radius of convergence $\geq \min(R(f), R(g))$, and for every z such that $|z| < \min(R(f), R(g))$, we have

$$h(z) = f(z) + g(z).$$

Furthermore, $s(z)$ has radius of convergence $\geq R(f)$, and for every z such that $|z| < R(f)$, we have

$$s(z) = \lambda f(z).$$

The above also applies to power series $f(A) = \sum_{k=0}^{\infty} a_k A^k$ and $g(A) = \sum_{k=0}^{\infty} b_k A^k$ with matrix argument $A \in \mathrm{M}_n(\mathbb{C})$, with $|z|$ replaced by $\|A\|$.

Let us now consider the product of two series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ where $a_k, b_k \in \mathbb{C}$. The *Cauchy product* of these two series is the series $\sum_{k=0}^{\infty} c_k$, where

$$c_k = \sum_{i=0}^{k} a_i b_{k-i} \quad k \in \mathbb{N}.$$

The following result can be shown (for example, see Cartan [27]).

Proposition 11.12. *Let $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ be two series with coefficients $a_k, b_k \in \mathbb{C}$. If both series converge absolutely to limits A and B, respectively, then their Cauchy product $\sum_{k=0}^{\infty} c_k$, converges absolutely, and if C is the limit of the Cauchy product, then $C = AB$.*

Next, if $f(z) = \sum_{k=0}^{\infty} a_k z^k$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ are two power series with coefficients $a_k, b_k \in \mathbb{C}$, the product of the power series $f(z)$ and $g(z)$ is the power series $h(z) = \sum_{k=0}^{\infty} c_k z^k$ where c_k is the Cauchy product

$$c_k = \sum_{i=0}^{k} a_i b_{k-i} \quad k \in \mathbb{N}.$$

Proposition 11.13. *Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ be two series with coefficients $a_k, b_k \in \mathbb{C}$. If both series have a radius of convergence $\geq \rho$, then their Cauchy product $h(z) = \sum_{k=0}^{\infty} c_k z^k$ has radius of convergence $\geq \rho$. Furthermore, for all z, if $|z| < \rho$, then*

$$h(z) = f(z)g(z).$$

Proposition 11.13 still holds for power series $f(A) = \sum_{k=0}^{\infty} a_k A^k$ and $g(A) = \sum_{k=0}^{\infty} b_k A^k$ with matrix argument $A \in \mathrm{M}_n(\mathbb{R})$, with $|z| < \rho$ replaced by $\|A\| < \rho$.

Finally, let us consider the substitution of power series. Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ be two series with coefficients $a_k, b_k \in \mathbb{C}$, and assume that $a_0 = 0$. Then if we substitute $f(z)$ for z in $g(z)$, we get an expression

$$g(f(z)) = \sum_{k=0}^{\infty} b_k \left(\sum_{n=0}^{\infty} a_n z^n \right)^k,$$

and because $a_0 = 0$, when we expand the powers, there are only finitely many terms involving any monomial z^m, since for $k > m$, the power $\left(\sum_{n=0}^{\infty} a_n z^n \right)^k$ has no terms of degree less than m. Thus, we can regroup the terms of $g(f(z))$ involving each monomial z^m, and the resulting power series is denoted by $(g \circ f)(z)$. We have the following result (for example, see Cartan [27]).

Proposition 11.14. *Let $f(z) = \sum_{k=0}^{\infty} a_k z^k$ and $g(z) = \sum_{k=0}^{\infty} b_k z^k$ be two power series with coefficients $a_k, b_k \in \mathbb{C}$, and write $R(f)$ for the radius of convergence of $f(z)$ and $R(g)$ for the radius of convergence of $g(z)$. If $R(f) > 0$, $R(g) > 0$, and $a_0 = 0$, then for any $r > 0$ chosen so that $\sum_{k=1}^{\infty} |a_k| r^k < R(g)$, the following hold:*

1. *The radius of convergence $R(h)$ of $h(z) = (g \circ f)(z)$ is at least r.*

2. *For every z, if $|z| \leq r$, then $|f(z)| < R(g)$, and*

$$h(z) = g(f(z)).$$

Proposition 11.14 still holds for power series $f(A) = \sum_{k=0}^{\infty} a_k A^k$ and $g(A) = \sum_{k=0}^{\infty} b_k A^k$ with matrix argument $A \in \mathrm{M}_n(\mathbb{C})$, with $|z| \leq r$ replaced by $\|A\| \leq r$ and $|f(z)| < R(g)$ replaced by $\|f(z)\| < R(g)$.

As an application of Proposition 11.14, (see Cartan [27]) note that the formal power series

$$E(A) = \sum_{k=1}^{\infty} \frac{A^k}{k!}$$

and

$$L(A) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{A^k}{k}$$

are mutual inverses; that is,

$$E(L(A)) = A, \quad L(E(A)) = A, \quad \text{for all } A.$$

Observe that $E(A) = e^A - I = \exp(A) - I$ and $L(A) = \log(I + A)$. It follows that

$$\begin{aligned} \log(\exp(A)) &= A \quad \text{for all } A \text{ with } \|A\| < \log(2) \\ \exp(\log(I + A)) &= I + A \quad \text{for all } A \text{ with } \|A\| < 1. \end{aligned}$$

11.2 The Derivative of a Function Between Normed Vector Spaces

In this section we review some basic notions of differential calculus, in particular, the *derivative* of a function $f\colon E \to F$, where E and F are normed vector spaces. In most cases, $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$. However, if we need to deal with infinite dimensional manifolds, then it is necessary to allow E and F to be infinite dimensional. We omit most proofs and refer the reader to standard analysis textbooks such as Lang [74, 73], Munkres [88], Choquet-Bruhat [32] or Schwartz [103, 104], for a complete exposition.

Let E and F be two *normed vector spaces*, let $A \subseteq E$ be some open subset of E, and let $a \in A$ be some element of A. Even though a is a vector, we may also call it a point.

The idea behind the derivative of the function f at a is that it is a *linear approximation* of f in a small open set around a. The difficulty is to make sense of the quotient

$$\frac{f(a + h) - f(a)}{h}$$

where h is a vector. We circumvent this difficulty in two stages.

A first possibility is to consider the *directional derivative* with respect to a vector $u \neq 0$ in E.

We can consider the vector $f(a+tu)-f(a)$, where $t \in \mathbb{R}$ (or $t \in \mathbb{C}$). Now,

$$\frac{f(a+tu)-f(a)}{t}$$

makes sense.

The idea is that in E, the points of the form $a+tu$ for t in some small interval $[-\epsilon, +\epsilon]$ in $\mathbb{R}$ form a line segment $[r, s]$ in A containing a, and that the image of this line segment defines a small curve segment on $f(A)$. This curve segment is defined by the map $t \mapsto f(a+tu)$, from $[r, s]$ to F, and the directional derivative $\mathrm{D}_u f(a)$ defines the direction of the tangent line at a to this curve. See Figure 11.2.

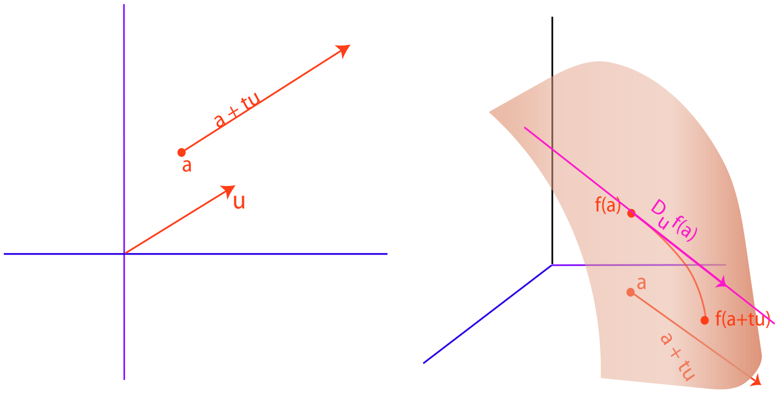


Figure 11.2: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the peach surface in $\mathbb{R}^3$, and $t \mapsto f(a+tu)$ is the embedded orange curve connecting $f(a)$ to $f(a+tu)$. Then $\mathrm{D}_u f(a)$ is the slope of the pink tangent line in the direction of u.

Definition 11.4. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, for any $u \neq 0$ in E, the *directional derivative of f at a w.r.t. the vector u*, denoted by $\mathrm{D}_u f(a)$, is the limit (if it exists)

$$\lim_{t \to 0,\, t \in U} \frac{f(a+tu)-f(a)}{t},$$

where $U = \{t \in \mathbb{R} \mid a+tu \in A,\, t \neq 0\}$ (or $U = \{t \in \mathbb{C} \mid a+tu \in A,\, t \neq 0\}$).

Since the map $t \mapsto a+tu$ is continuous, and since $A - \{a\}$ is open, the inverse image U of $A - \{a\}$ under the above map is open, and the definition of the limit in Definition 11.4 makes sense.

Remark: Since the notion of limit is purely topological, the existence and value of a directional derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

The directional derivative is sometimes called the *Gâteaux derivative*.

In the special case where $E = \mathbb{R}$, $F = \mathbb{R}$ and we let $u = 1$ (i.e., the real number 1, viewed as a vector), it is immediately verified that $\mathrm{D}_1 f(a) = f'(a)$. When $E = \mathbb{R}$ (or $E = \mathbb{C}$) and F is any normed vector space, the derivative $\mathrm{D}_1 f(a)$, also denoted by $f'(a)$, provides a suitable generalization of the notion of derivative.

However, when E has dimension ≥ 2, directional derivatives present a serious problem, which is that their definition is not sufficiently uniform. Indeed, there is no reason to believe that the directional derivatives w.r.t. all nonzero vectors u share something in common. As a consequence, a function can have all directional derivatives at a, and yet not be continuous at a. Two functions may have all directional derivatives in some open sets, and yet their composition may not. Thus we introduce a more uniform notion.

Given two normed vector spaces E and F, recall that a linear map $f\colon E \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u)\| \leq C\,\|u\| \quad \text{for all } u \in E.$$

The set of *continuous* linear maps from E to F is a vector space denoted $\mathcal{L}(E; F)$, and the set of *all* linear maps from E to F is a vector space denoted by $\mathrm{Hom}(E, F)$. If E is finite-dimensional, then $\mathcal{L}(E; F) = \mathrm{Hom}(E, F)$, but if E is infinite-dimensional, then there may be linear maps that are *not* continuous, and in general the space $\mathcal{L}(E; F)$ is a proper subspace of $\mathrm{Hom}(E, F)$.

Definition 11.5. Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, we say that f is *differentiable at* $a \in A$ if there is a continuous linear map, $L\colon E \to F$, and a function, $\epsilon(h)$, such that

$$f(a + h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for every $a + h \in A$, where

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0,$$

with $U = \{h \in E \mid a + h \in A,\ h \neq 0\}$. The linear map L is denoted by $\mathrm{D}f(a)$, or $\mathrm{D}f_a$, or $df(a)$, or df_a, or $f'(a)$, and it is called the *Fréchet derivative*, or *total derivative*, or *derivative*, or *total differential*, or *differential*, of f at a. See Figure 11.3.

Since the map $h \mapsto a + h$ from E to E is continuous, and since A is open in E, the inverse image U of $A - \{a\}$ under the above map is open in E, and it makes sense to say that

$$\lim_{h \to 0,\, h \in U} \epsilon(h) = 0.$$

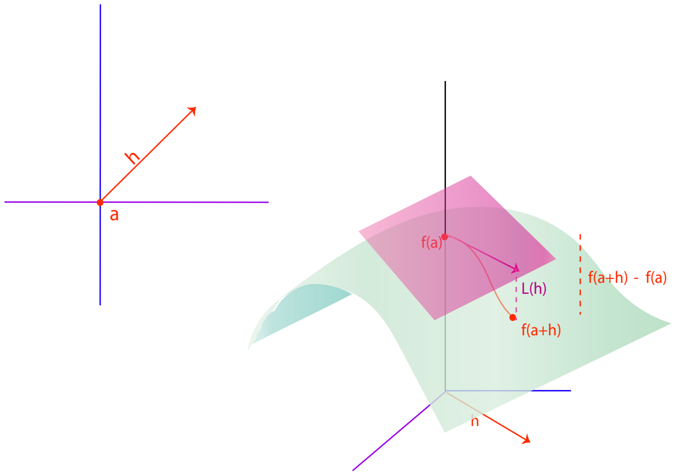


Figure 11.3: Let $f\colon \mathbb{R}^2 \to \mathbb{R}$. The graph of f is the green surface in $\mathbb{R}^3$. The linear map $L = \mathrm{D}f(a)$ is the pink tangent plane. For any vector $h \in \mathbb{R}^2$, $L(h)$ is approximately equal to $f(a+h) - f(a)$. Note that $L(h)$ is also the direction tangent to the curve $t \mapsto f(a+tu)$.

Note that for every $h \in U$, since $h \neq 0$, $\epsilon(h)$ is uniquely determined since

$$\epsilon(h) = \frac{f(a+h) - f(a) - L(h)}{\|h\|},$$

and the value $\epsilon(0)$ plays absolutely no role in this definition. It does no harm to assume that $\epsilon(0) = 0$, and we will assume this from now on.

Remark: Since the notion of limit is purely topological, the existence and value of a derivative is independent of the choice of norms in E and F, as long as they are equivalent norms.

The following proposition shows that our new definition is consistent with the definition of the directional derivative and that *the continuous linear map L is unique*, if it exists.

Proposition 11.15. *Let E and F be two normed vector spaces, let A be a nonempty open subset of E, and let $f\colon A \to F$ be any function. For any $a \in A$, if $\mathrm{D}f(a)$ is defined, then f is continuous at a and f has a directional derivative $\mathrm{D}_u f(a)$ for every $u \neq 0$ in E. Furthermore,*

$$\mathrm{D}_u f(a) = \mathrm{D}f(a)(u)$$

and thus, $\mathrm{D}f(a)$ is uniquely defined.

Proof. If $L = \mathrm{D}f(a)$ exists, then for any nonzero vector $u \in E$, because A is open, for any $t \in \mathbb{R} - \{0\}$ (or $t \in \mathbb{C} - \{0\}$) small enough, $a + tu \in A$, so

$$\begin{aligned} f(a+tu) &= f(a) + L(tu) + \epsilon(tu)\|tu\| \\ &= f(a) + tL(u) + |t|\epsilon(tu)\|u\| \end{aligned}$$

which implies that

$$L(u) = \frac{f(a+tu) - f(a)}{t} - \frac{|t|}{t}\epsilon(tu)\|u\|,$$

and since $\lim_{t\mapsto 0} \epsilon(tu) = 0$, we deduce that

$$L(u) = \mathrm{D}f(a)(u) = \mathrm{D}_u f(a).$$

Because

$$f(a+h) = f(a) + L(h) + \epsilon(h)\|h\|$$

for all h such that $\|h\|$ is small enough, L is continuous, and $\lim_{h\mapsto 0} \epsilon(h)\|h\| = 0$, we have $\lim_{h\mapsto 0} f(a+h) = f(a)$, that is, f is continuous at a. □

Observe that the uniqueness of $\mathrm{D}f(a)$ follows from Proposition 11.15. Also when E is of finite dimension, it is easily shown that every linear map is continuous and this assumption is then redundant.

As an example, consider the map $f\colon \mathrm{M}_n(\mathbb{R}) \to \mathrm{M}_n(\mathbb{R})$ given by

$$f(A) = A^\top A - I,$$

where $\mathrm{M}_n(\mathbb{R})$ denotes the vector space of all $n \times n$ matrices with real entries equipped with any matrix norm, since they are all equivalent; for example, pick the Frobenius norm $\|A\|_F = \sqrt{\mathrm{tr}(A^\top A)}$. We claim that

$$Df(A)(H) = A^\top H + H^\top A, \quad \text{for all } A \text{ and } H \text{ in } \mathrm{M}_n(\mathbb{R}).$$

We have

$$\begin{aligned} f(A+H) - f(A) - (A^\top H + H^\top A) &= (A+H)^\top (A+H) - I - (A^\top A - I) - A^\top H - H^\top A \\ &= A^\top A + A^\top H + H^\top A + H^\top H - A^\top A - A^\top H - H^\top A \\ &= H^\top H. \end{aligned}$$

It follows that

$$\epsilon(H) = \frac{f(A+H) - f(A) - (A^\top H + H^\top A)}{\|H\|} = \frac{H^\top H}{\|H\|},$$

and since our norm is the Frobenius norm,

$$\|\epsilon(H)\| = \left\|\frac{H^\top H}{\|H\|}\right\| \leq \frac{\left\|H^\top\right\| \|H\|}{\|H\|} = \left\|H^\top\right\| = \|H\|,$$

so

$$\lim_{H\mapsto 0} \epsilon(H) = 0,$$

and we conclude that

$$Df(A)(H) = A^\top H + H^\top A.$$

If $\mathrm{D}f(a)$ exists for every $a \in A$, we get a map $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$, called the *derivative of f on A*, and also denoted by df. Here $\mathcal{L}(E; F)$ denotes the vector space of continuous linear maps from E to F.

We now consider a number of standard results about derivatives. A function $f\colon E \to F$ is said to be *affine* if there is some linear map $\overrightarrow{f}\colon E \to F$ and some fixed vector $c \in F$, such that

$$f(u) = \overrightarrow{f}(u) + c$$

for all $u \in E$. We call $\overrightarrow{f}$ the *linear map associated with f*.

Proposition 11.16. *Given two normed vector spaces E and F, if $f\colon E \to F$ is a constant function, then $\mathrm{D}f(a) = 0$, for every $a \in E$. If $f\colon E \to F$ is a continuous affine map, then $\mathrm{D}f(a) = \overrightarrow{f}$, for every $a \in E$, where $\overrightarrow{f}$ denotes the linear map associated with f.*

Proposition 11.17. *Given a normed vector space E and a normed vector space F, for any two functions $f, g\colon E \to F$, for every $a \in E$, if $\mathrm{D}f(a)$ and $\mathrm{D}g(a)$ exist, then $\mathrm{D}(f + g)(a)$ and $\mathrm{D}(\lambda f)(a)$ exist, and*

$$\begin{aligned} \mathrm{D}(f + g)(a) &= \mathrm{D}f(a) + \mathrm{D}g(a), \\ \mathrm{D}(\lambda f)(a) &= \lambda \mathrm{D}f(a). \end{aligned}$$

Given two normed vector spaces $(E_1, \| \|_1)$ and $(E_2, \| \|_2)$, there are three natural and equivalent norms that can be used to make $E_1 \times E_2$ into a normed vector space:

1. $\|(u_1, u_2)\|_1 = \|u_1\|_1 + \|u_2\|_2$.

2. $\|(u_1, u_2)\|_2 = (\|u_1\|_1^2 + \|u_2\|_2^2)^{1/2}$.

3. $\|(u_1, u_2)\|_\infty = \max(\|u_1\|_1, \|u_2\|_2)$.

We usually pick the first norm. If E_1, E_2, and F are three normed vector spaces, recall that a bilinear map $f\colon E_1 \times E_2 \to F$ is *continuous* iff there is some constant $C \geq 0$ such that

$$\|f(u_1, u_2)\| \leq C \|u_1\|_1 \|u_2\|_2 \quad \text{for all } u_1 \in E_1 \text{ and all } u_2 \in E_2.$$

Proposition 11.18. *Given three normed vector spaces E_1, E_2, and F, for any continuous bilinear map $f\colon E_1 \times E_2 \to F$, for every $(a, b) \in E_1 \times E_2$, $\mathrm{D}f(a, b)$ exists, and for every $u \in E_1$ and $v \in E_2$,*

$$\mathrm{D}f(a, b)(u, v) = f(u, b) + f(a, v).$$

Proof. Since f is bilinear, a simple computation implies that

$$\begin{aligned} f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v)) &= f(a+u,b+v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(a+u,b) + f(a+u,v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(a,b) + f(u,b) + f(a,v) + f(u,v) - f(a,b) - f(u,b) - f(a,v) \\ &= f(u,v). \end{aligned}$$

We define

$$\epsilon(u,v) = \frac{f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))}{\|(u,v)\|_1},$$

and observe that the continuity of f implies

$$\begin{aligned} \|f((a,b)+(u,v)) - f(a,b) - (f(u,b)+f(a,v))\| &= \|f(u,v)\| \\ &\leq C\,\|u\|_1\,\|v\|_2 \leq C\,(\|u\|_1 + \|v\|_2)^2 . \end{aligned}$$

Hence

$$\|\epsilon(u,v)\| = \left\|\frac{f(u,v)}{\|(u,v)\|_1}\right\| = \frac{\|f(u,v)\|}{\|(u,v)\|_1} \leq \frac{C\,(\|u\|_1 + \|v\|_2)^2}{\|u\|_1 + \|v\|_2} = C\,(\|u\|_1 + \|v\|_2) = C\,\|(u,v)\|_1 ,$$

which in turn implies $\lim_{(u,v)\mapsto(0,0)} \epsilon(u,v) = 0$. □

We now state the very useful *chain rule*.

Theorem 11.19. *Given three normed vector spaces E, F, and G, let A be an open set in E, and let B an open set in F. For any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(f(a))$ exists, then $\mathrm{D}(g \circ f)(a)$ exists, and*

$$\mathrm{D}(g \circ f)(a) = \mathrm{D}g(f(a)) \circ \mathrm{D}f(a).$$

Theorem 11.19 has many interesting consequences. We mention one corollary.

Proposition 11.20. *Given two normed vector spaces E and F, let A be some open subset in E, let B be some open subset in F, let $f\colon A \to B$ be a bijection from A to B, and assume that $\mathrm{D}f$ exists on A and that $\mathrm{D}f^{-1}$ exists on B. Then for every $a \in A$,*

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

Proposition 11.20 has the remarkable consequence that the two vector spaces E and F have the same dimension. In other words, a local property, the existence of a bijection f between an open set A of E and an open set B of F, such that f is differentiable on A and f^{-1} is differentiable on B, implies a global property, that the two vector spaces E and F have the same dimension. Let us mention two more rules about derivatives that are used all the time.

Let $\iota\colon \mathbf{GL}(n, \mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ be the function (inversion) defined on invertible $n \times n$ matrices by $\iota(A) = A^{-1}$. Then we have

$$d\iota_A(H) = -A^{-1}HA^{-1},$$

for all $A \in \mathbf{GL}(n, \mathbb{C})$ and for all $H \in \mathrm{M}_n(\mathbb{C})$.

To prove the preceding line observe that for H with sufficiently small norm, we have

$$\begin{aligned}\iota(A+H) - \iota(A) + A^{-1}HA^{-1} &= (A+H)^{-1} - A^{-1} + A^{-1}HA^{-1}\\ &= (A+H)^{-1}[I - (A+H)A^{-1} + (A+H)A^{-1}HA^{-1}]\\ &= (A+H)^{-1}[I - I - HA^{-1} + HA^{-1} + HA^{-1}HA^{-1}]\\ &= (A+H)^{-1}HA^{-1}HA^{-1}.\end{aligned}$$

Consequently, we get

$$\epsilon(H) = \frac{\iota(A+H) - \iota(A) + A^{-1}HA^{-1}}{\|H\|} = \frac{(A+H)^{-1}HA^{-1}HA^{-1}}{\|H\|},$$

and since

$$\left\|(A+H)^{-1}HA^{-1}HA^{-1}\right\| \leq \|H\|^2 \left\|A^{-1}\right\|^2 \left\|(A+H)^{-1}\right\|,$$

it is clear that $\lim_{H\mapsto 0} \epsilon(H) = 0$, which proves that

$$d\iota_A(H) = -A^{-1}HA^{-1}.$$

In particular, if $A = I$, then $d\iota_I(H) = -H$.

Next, if $f\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ and $g\colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ are differentiable matrix functions, then

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathrm{M}_n(\mathbb{C})$. This is known as the *product rule*.

When E is of finite dimension n, for any basis, $(u_1, \ldots, u_n)$, of E, we can define the directional derivatives with respect to the vectors in the basis $(u_1, \ldots, u_n)$ (actually, we can also do it for an infinite basis). This way we obtain the definition of partial derivatives, as follows:

Definition 11.6. For any two normed spaces E and F, if E is of finite dimension n, for every basis $(u_1, \ldots, u_n)$ for E, for every $a \in E$, for every function $f\colon E \to F$, the directional derivatives $\mathrm{D}_{u_j}f(a)$ (if they exist) are called the *partial derivatives of f with respect to the basis* $(u_1, \ldots, u_n)$. The partial derivative $\mathrm{D}_{u_j}f(a)$ is also denoted by $\partial_j f(a)$, or $\dfrac{\partial f}{\partial x_j}(a)$.

The notation $\frac{\partial f}{\partial x_j}(a)$ for a partial derivative, although customary and going back to Leibniz, is a “logical obscenity.” Indeed, the variable x_j really has nothing to do with the formal definition. This is just another of these situations where tradition is just too hard to overthrow!

If both E and F are of finite dimension, for any basis $(u_1, \ldots, u_n)$ of E and any basis $(v_1, \ldots, v_m)$ of F, every function $f\colon E \to F$ is determined by m functions $f_i\colon E \to \mathbb{R}$ (or $f_i\colon E \to \mathbb{C}$), where

$$f(x) = f_1(x)v_1 + \cdots + f_m(x)v_m,$$

for every $x \in E$. Then we get

$$\mathrm{D}f(a)(u_j) = \mathrm{D}f_1(a)(u_j)v_1 + \cdots + \mathrm{D}f_i(a)(u_j)v_i + \cdots + \mathrm{D}f_m(a)(u_j)v_m,$$

that is,

$$\mathrm{D}f(a)(u_j) = \partial_j f_1(a)v_1 + \cdots + \partial_j f_i(a)v_i + \cdots + \partial_j f_m(a)v_m.$$

Since the j-th column of the $m \times n$-matrix representing $\mathrm{D}f(a)$ w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots ,v_m)$ is equal to the components of the vector $\mathrm{D}f(a)(u_j)$ over the basis $(v_1, \ldots, v_m)$, the linear map $\mathrm{D}f(a)$ is determined by the $m \times n$-matrix
$J(f)(a) = (\partial_j f_i(a))$, or $J(f)(a) = \left(\frac{\partial f_i}{\partial x_j}(a)\right)$:

$$J(f)(a) = \begin{pmatrix} \partial_1 f_1(a) & \partial_2 f_1(a) & \ldots & \partial_n f_1(a) \\ \partial_1 f_2(a) & \partial_2 f_2(a) & \ldots & \partial_n f_2(a) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_m(a) & \partial_2 f_m(a) & \ldots & \partial_n f_m(a) \end{pmatrix}$$

or

$$J(f)(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \ldots & \frac{\partial f_1}{\partial x_n}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \ldots & \frac{\partial f_2}{\partial x_n}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \frac{\partial f_m}{\partial x_2}(a) & \ldots & \frac{\partial f_m}{\partial x_n}(a) \end{pmatrix}.$$

This matrix is called the *Jacobian matrix* of $\mathrm{D}f$ at a. When $m = n$, the determinant, $\det(J(f)(a))$, of $J(f)(a)$ is called the *Jacobian* of $\mathrm{D}f(a)$.

We know that this determinant only depends on $\mathrm{D}f(a)$, and not on specific bases. However, partial derivatives give a means for computing it.

When $E = \mathbb{R}^n$ and $F = \mathbb{R}^m$, for any function $f\colon \mathbb{R}^n \to \mathbb{R}^m$, it is easy to compute the partial derivatives $\frac{\partial f_i}{\partial x_j}(a)$. We simply treat the function $f_i\colon \mathbb{R}^n \to \mathbb{R}$ as a function of its j-th argument, leaving the others fixed, and compute the derivative as the usual derivative.

Example 11.1. For example, consider the function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$, defined by

$$f(r, \theta) = (r\cos\theta, r\sin\theta).$$

Then we have

$$J(f)(r,\theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

and the Jacobian (determinant) has value $\det(J(f)(r,\theta)) = r$.

In the case where $E = \mathbb{R}$ (or $E = \mathbb{C}$), for any function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), the Jacobian matrix of $\mathrm{D}f(a)$ is a column vector. In fact, this column vector is just $\mathrm{D}_1 f(a)$. Then for every $\lambda \in \mathbb{R}$ (or $\lambda \in \mathbb{C}$), $\mathrm{D}f(a)(\lambda) = \lambda \mathrm{D}_1 f(a)$. This case is sufficiently important to warrant a definition.

Definition 11.7. Given a function $f\colon \mathbb{R} \to F$ (or $f\colon \mathbb{C} \to F$), where F is a normed space, the vector

$$\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$$

is called the *vector derivative or velocity vector (in the real case)* at a. We usually identify $\mathrm{D}f(a)$ with its Jacobian matrix $\mathrm{D}_1 f(a)$, which is the column vector corresponding to $\mathrm{D}_1 f(a)$. By abuse of notation, we also let $\mathrm{D}f(a)$ denote the vector $\mathrm{D}f(a)(1) = \mathrm{D}_1 f(a)$.

When $E = \mathbb{R}$, the physical interpretation is that f defines a (parametric) curve that is the trajectory of some particle moving in $\mathbb{R}^m$ as a function of time, and the vector $\mathrm{D}_1 f(a)$ is the *velocity* of the moving particle $f(t)$ at $t = a$. See Figure 11.4.

Example 11.2.

1. When $A = (0, 1)$ and $F = \mathbb{R}^3$, a function $f\colon (0,1) \to \mathbb{R}^3$ defines a (parametric) curve in $\mathbb{R}^3$. If $f = (f_1, f_2, f_3)$, its Jacobian matrix at $a \in \mathbb{R}$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f_1}{\partial t}(a) \\ \dfrac{\partial f_2}{\partial t}(a) \\ \dfrac{\partial f_3}{\partial t}(a) \end{pmatrix}.$$

See Figure 11.4.

The velocity vectors $J(f)(a) = \begin{pmatrix} -\sin(t) \\ \cos(t) \\ 1 \end{pmatrix}$ are represented by the blue arrows.

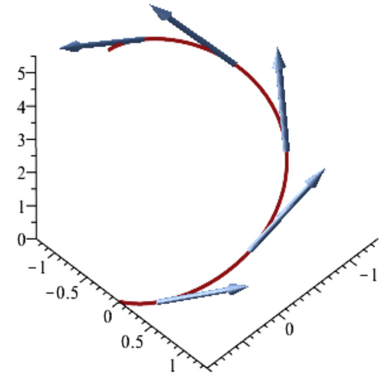


Figure 11.4: The red space curve $f(t) = (\cos(t), \sin(t), t)$.

2. When $E = \mathbb{R}^2$ and $F = \mathbb{R}^3$, a function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ defines a parametric surface. Letting $\varphi = (f, g, h)$, its Jacobian matrix at $a \in \mathbb{R}^2$ is

$$J(\varphi)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial u}(a) & \dfrac{\partial f}{\partial v}(a) \\ \dfrac{\partial g}{\partial u}(a) & \dfrac{\partial g}{\partial v}(a) \\ \dfrac{\partial h}{\partial u}(a) & \dfrac{\partial h}{\partial v}(a) \end{pmatrix}.$$

   See Figure 11.5. The Jacobian matrix is $J(f)(a) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 2u & 2v \end{pmatrix}$. The first column is the vector tangent to the pink u-direction curve, while the second column is the vector tangent to the blue v-direction curve.

3. When $E = \mathbb{R}^3$ and $F = \mathbb{R}$, for a function $f\colon \mathbb{R}^3 \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^3$ is

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x}(a) & \dfrac{\partial f}{\partial y}(a) & \dfrac{\partial f}{\partial z}(a) \end{pmatrix}.$$

More generally, when $f\colon \mathbb{R}^n \to \mathbb{R}$, the Jacobian matrix at $a \in \mathbb{R}^n$ is the row vector

$$J(f)(a) = \begin{pmatrix} \dfrac{\partial f}{\partial x_1}(a) & \cdots & \dfrac{\partial f}{\partial x_n}(a) \end{pmatrix}.$$

Its transpose is a column vector called the *gradient* of f at a, denoted by $\mathrm{grad} f(a)$ or $\nabla f(a)$. Then given any $v \in \mathbb{R}^n$, note that

$$\mathrm{D}f(a)(v) = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n = \mathrm{grad} f(a) \cdot v,$$

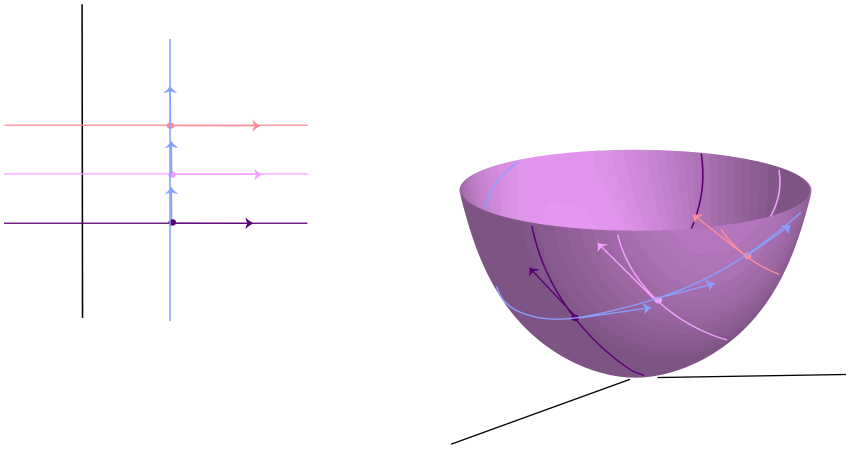


Figure 11.5: The parametric surface $x = u, y = v, z = u^2 + v^2$.

the scalar product of $\mathrm{grad} f(a)$ and v.

When E, F, and G have finite dimensions, where $(u_1, \ldots, u_p)$ is a basis for E, $(v_1, \ldots, v_n)$ is a basis for F, and $(w_1, \ldots, w_m)$ is a basis for G, if A is an open subset of E, B is an open subset of F, for any functions $f\colon A \to F$ and $g\colon B \to G$, such that $f(A) \subseteq B$, for any $a \in A$, letting $b = f(a)$, and $h = g \circ f$, if $\mathrm{D}f(a)$ exists and $\mathrm{D}g(b)$ exists, by Theorem 11.19, the Jacobian matrix $J(h)(a) = J(g \circ f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(w_1, \ldots, w_m)$ is the product of the Jacobian matrices $J(g)(b)$ w.r.t. the bases $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_m)$, and $J(f)(a)$ w.r.t. the bases $(u_1, \ldots, u_p)$ and $(v_1, \ldots, v_n)$:

$$J(h)(a) = \begin{pmatrix} \frac{\partial g_1}{\partial y_1}(b) & \frac{\partial g_1}{\partial y_2}(b) & \ldots & \frac{\partial g_1}{\partial y_n}(b) \\ \frac{\partial g_2}{\partial y_1}(b) & \frac{\partial g_2}{\partial y_2}(b) & \ldots & \frac{\partial g_2}{\partial y_n}(b) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial y_1}(b) & \frac{\partial g_m}{\partial y_2}(b) & \ldots & \frac{\partial g_m}{\partial y_n}(b) \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \ldots & \frac{\partial f_1}{\partial x_p}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \ldots & \frac{\partial f_2}{\partial x_p}(a) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(a) & \frac{\partial f_n}{\partial x_2}(a) & \ldots & \frac{\partial f_n}{\partial x_p}(a) \end{pmatrix}.$$

Thus we have the familiar formula

$$\frac{\partial h_i}{\partial x_j}(a) = \sum_{k=1}^{n} \frac{\partial g_i}{\partial y_k}(b) \frac{\partial f_k}{\partial x_j}(a).$$

Given two normed vector spaces E and F of finite dimension, given an open subset A of E, if a function $f\colon A \to F$ is differentiable at $a \in A$, then its Jacobian matrix is well defined.

One should be warned that the converse is false. There are functions such that all the partial derivatives exist at some $a \in A$, but yet, the function is not differentiable at a, and not even continuous at a.

However, there are sufficient conditions on the partial derivatives for $\mathrm{D}f(a)$ to exist, namely, continuity of the partial derivatives. If f is differentiable on A, then f defines a function $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$. It turns out that the continuity of the partial derivatives on A is a necessary and sufficient condition for $\mathrm{D}f$ to exist and to be continuous on A. To prove this, we need an important result known as the mean value theorem.

If E is a vector space (over $\mathbb{R}$ or $\mathbb{C}$), given any two points $a, b \in E$, the *closed segment* $[a, b]$ is the set of all points $a + \lambda(b - a)$, where $0 \leq \lambda \leq 1$, $\lambda \in \mathbb{R}$, and the *open segment* (a, b) is the set of all points $a + \lambda(b - a)$, where $0 < \lambda < 1$, $\lambda \in \mathbb{R}$. The following result is known as the *mean value theorem.*

Proposition 11.21. *Let E and F be two normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a continuous function on A. Given any $a \in A$ and any $h \neq 0$ in E, if the closed segment $[a, a + h]$ is contained in A, if $f\colon A \to F$ is differentiable at every point of the open segment $(a, a + h)$, and if*

$$\sup_{x \in (a, a+h)} \|\mathrm{D}f(x)\| \leq M$$

for some $M \geq 0$, then

$$\|f(a + h) - f(a)\| \leq M\|h\|.$$

As a corollary, if $L\colon E \to F$ is a continuous linear map, then

$$\|f(a + h) - f(a) - L(h)\| \leq M\|h\|,$$

where $M = \sup_{x \in (a, a+h)} \|\mathrm{D}f(x) - L\|$.

A very useful result which is proved using the mean value theorem is the proposition below.

Proposition 11.22. *Let $f\colon A \to F$ be any function between two normed vector spaces E and F, where A is an open subset of E. If A is connected and if $\mathrm{D}f(a) = 0$ for all $a \in A$, then f is a constant function on A.*

The mean value theorem also implies the following important result.

Theorem 11.23. *Given two normed vector spaces E and F, where E is of finite dimension n and where $(u_1, \ldots, u_n)$ is a basis of E, given any open subset A of E, given any function $f\colon A \to F$, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$ is defined and continuous on A iff every partial derivative $\partial_j f$ (or $\dfrac{\partial f}{\partial x_j}$) is defined and continuous on A, for all j, $1 \leq j \leq n$. As a corollary, if F is of finite dimension m, and $(v_1, \ldots, v_m)$ is a basis of F, the derivative $\mathrm{D}f\colon A \to \mathcal{L}(E; F)$ is defined and continuous on A iff every partial derivative $\partial_j f_i$ $\left(\text{or } \dfrac{\partial f_i}{\partial x_j}\right)$ is defined and continuous on A, for all i, j, $1 \leq i \leq m$, $1 \leq j \leq n$.*

Definition 11.8. Given two normed vector spaces E and F, and an open subset A of E, we say that a function $f\colon A \to F$ is a *C^0-function on A* if f is continuous on A. We say that $f\colon A \to F$ is a *C^1-function on A* if $\mathrm{D}f$ exists and is continuous on A.

Let E and F be two normed vector spaces, let $U \subseteq E$ be an open subset of E and let $f\colon E \to F$ be a function such that $Df(a)$ exists for all $a \in U$. If $Df(a)$ is injective for all $a \in U$, we say that f is an *immersion* (on U) and if $Df(a)$ is surjective for all $a \in U$, we say that f is a *submersion* (on U).

When E and F are finite dimensional with $\dim(E) = n$ and $\dim(F) = m$, if $m \geq n$, then f is an immersion iff the Jacobian matrix, $J(f)(a)$, has full rank n for all $a \in E$ and if $n \geq m$, then f is a submersion iff the Jacobian matrix, $J(f)(a)$, has full rank m for all $a \in E$.

For example, $f\colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (\cos(t), \sin(t))$ is an immersion since $J(f)(t) = \begin{pmatrix} -\sin(t) \\ \cos(t) \end{pmatrix}$ has rank 1 for all t. On the other hand, $f\colon \mathbb{R} \to \mathbb{R}^2$ defined by $f(t) = (t^2, t^2)$ is not an immersion since $J(f)(t) = \begin{pmatrix} 2t \\ 2t \end{pmatrix}$ vanishes at $t = 0$. See Figure 11.6. An example of a submersion is given by the projection map $f\colon \mathbb{R}^2 \to \mathbb{R}$, where $f(x, y) = x$, since $J(f)(x, y) = \begin{pmatrix} 1 & 0 \end{pmatrix}$.

A very important theorem is the inverse function theorem. In order for this theorem to hold for infinite dimensional spaces, it is necessary to assume that our normed vector spaces are complete. Fortunately, $\mathbb{R}, \mathbb{C}$, and every finite dimensional (real or complex) normed vector space is complete. A real (resp. complex) vector space, E, is a real (resp. complex) *Hilbert space* if it is complete as a normed space with the norm $\|u\| = \sqrt{\langle u, u\rangle}$ induced by its Euclidean (resp. Hermitian) inner product (of course, positive definite).

Definition 11.9. Given two topological spaces E and F and an open subset A of E, we say that a function $f\colon A \to F$ is a *local homeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a one-to-one, onto, continuous function from U to $V = f(U)$ which has continuous inverse $f^{-1}\colon V \to U$. If B is an open subset of F, we say that $f\colon A \to F$ is a *(global) homeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$.

If E and F are normed vector spaces, we say that $f\colon A \to F$ is a *local diffeomorphism from A to F* if for every $a \in A$, there is an open set $U \subseteq A$ containing a and an open set V containing $f(a)$ such that f is a bijection from U to V, f is a C^1-function on U, and f^{-1} is a C^1-function on $V = f(U)$. We say that $f\colon A \to F$ is a *(global) diffeomorphism from A to B* if f is a homeomorphism from A to $B = f(A)$, f is a C^1-function on A, and f^{-1} is a C^1-function on B.

Note that a local diffeomorphism is a local homeomorphism. As a consequence of Proposition 11.20, if f is a diffeomorphism on A, then $\mathrm{D}f(a)$ is a linear isomorphism for every $a \in A$.

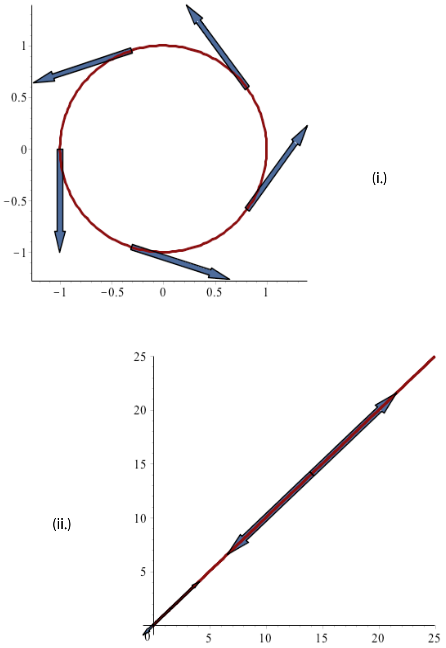


Figure 11.6: Figure $(i.)$ is the immersion of $\mathbb{R}$ into $\mathbb{R}^2$ given by $f(t) = (\cos(t), \sin(t))$. Figure $(ii.)$, the parametric curve $f(t) = (t^2, t^2)$, is not an immersion since the tangent vanishes at the origin.

Theorem 11.24. *(Inverse Function Theorem) Let E and F be complete normed vector spaces, let A be an open subset of E, and let $f\colon A \to F$ be a C^1-function on A. The following properties hold:*

(1) For every $a \in A$, if $\mathrm{D}f(a)$ is a linear isomorphism (which means that both $\mathrm{D}f(a)$ and $(\mathrm{D}f(a))^{-1}$ are linear and continuous),[1] then there exist some open subset $U \subseteq A$ containing a, and some open subset V of F containing $f(a)$, such that f is a diffeomorphism from U to $V = f(U)$. Furthermore,

$$\mathrm{D}f^{-1}(f(a)) = (\mathrm{D}f(a))^{-1}.$$

For every neighborhood N of a, the image $f(N)$ of N is a neighborhood of $f(a)$, and for every open ball $U \subseteq A$ of center a, the image $f(U)$ of U contains some open ball of center $f(a)$.

(2) If $\mathrm{D}f(a)$ is invertible for every $a \in A$, then $B = f(A)$ is an open subset of F, and f is a local diffeomorphism from A to B. Furthermore, if f is injective, then f is a diffeomorphism from A to B.

[1] Actually, since E and F are Banach spaces, by the Open Mapping Theorem, it is sufficient to assume that $\mathrm{D}f(a)$ is continuous and bijective; see Lang [73].

Proofs of the inverse function theorem can be found in Lang [73], Abraham and Marsden [1], Schwartz [104], and Cartan [28]. Part (1) of Theorem 11.24 is often referred to as the "(local) inverse function theorem." It plays an important role in the study of manifolds and (ordinary) differential equations.

If E and F are both of finite dimension, the case where $\mathrm{D}f(a)$ is just injective or just surjective is also important for defining manifolds, using implicit definitions.

Suppose as before that $f\colon A \to F$ is a function from some open subset A of E, with E and F two normed vector spaces. If $\mathrm{D}f\colon A \to \mathcal{L}(E;F)$ exists for all $a \in A$, then we can consider taking the derivative $\mathrm{DD}f(a)$ of $\mathrm{D}f$ at a. If it exists, $\mathrm{DD}f(a)$ is a continuous linear map in $\mathcal{L}(E;\mathcal{L}(E;F))$, and we denote $\mathrm{DD}f(a)$ as $\mathrm{D}^2f(a)$. It is known that the vector space $\mathcal{L}(E;\mathcal{L}(E;F))$ is isomorphic to the vector space of continuous bilinear maps $\mathcal{L}_2(E^2;F)$, so we can view $\mathrm{D}^2f(a)$ as a bilinear map in $\mathcal{L}_2(E^2;F)$. It is also known by Schwarz's lemma that $\mathrm{D}^2f(a)$ is symmetric (partial derivatives commute; see Schwartz [104]). Therefore, for every $a \in A$, where it exists, $\mathrm{D}^2f(a)$ belongs to the space $\mathcal{S}\mathrm{ym}_2(E^2;F)$ of continuous symmetric bilinear maps from E^2 to F. If E has finite dimension n and $F = \mathbb{R}$, with respect to any basis $(e_1,\ldots,e_n)$ of E, $\mathrm{D}^2f(a)(u,v)$ is the value of the quadratic form $u^\top\mathrm{Hess}f(a)v$, where

$$\mathrm{Hess}f(a) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(a)\right)$$

is the *Hessian matrix* of f at a.

By induction, if $\mathrm{D}^m f\colon A \to \mathcal{S}\mathrm{ym}_m(E^m;F)$ exists for $m \geq 1$, where $\mathcal{S}\mathrm{ym}_m(E^m;F)$ denotes the vector space of continuous symmetric multilinear maps from E^m to F, and if $\mathrm{DD}^m f(a)$ exists for all $a \in A$, we obtain the $(m+1)$th derivative $\mathrm{D}^{m+1}f$ of f, and $\mathrm{D}^{m+1}f \in \mathcal{S}\mathrm{ym}_{m+1}(E^{m+1};F)$, where $\mathcal{S}\mathrm{ym}_{m+1}(E^{m+1};F)$ is the vector space of continuous symmetric multilinear maps from E^{m+1} to F.

For any $m \geq 1$, we say that the map $f\colon A \to F$ is a *C^m-function* (or simply that *f is C^m*) if $\mathrm{D}f, \mathrm{D}^2f, \ldots, \mathrm{D}^m f$ exist and are continuous on A.

We say that f is C^∞ or *smooth* if $\mathrm{D}^m f$ exists and is continuous on A for all $m \geq 1$. If E has finite dimension n, it can be shown that f is smooth iff all of its partial derivatives

$$\frac{\partial^m f}{\partial x_{i_1} \cdots \partial x_{i_m}}(a)$$

are defined and continuous for all $a \in A$, all $m \geq 1$, and all $i_1, \ldots, i_m \in \{1, \ldots, n\}$.

The function $f\colon A \to F$ is a *C^m-diffeomorphism* between A and $B = f(A)$ if f is a bijection from A to B and if f and f^{-1} are C^m. Similarly, f is a *smooth diffeomorphism* between A and $B = f(A)$ if f is a bijection from A to B and if f and f^{-1} are smooth.

11.3 Linear Vector Fields and the Exponential

Definition 11.10. Given some open subset A of $\mathbb{R}^n$, a *vector field* X on A is a function $X\colon A \to \mathbb{R}^n$, which assigns to every point $p \in A$ a vector $X(p) \in \mathbb{R}^n$.

Usually we assume that X is at least C^1 on A. For example, if $f\colon \mathbb{R}^2 \to \mathbb{R}$ is $f(x, y) = \cos(xy^2)$, the gradient vector field X is $(-y^2 \sin(xy^2), -2xy \sin(xy^2)) = (X_1, X_2)$. Note that

$$\frac{\partial X_1}{\partial y} = -2y \sin(xy^2) - 2xy^3 \cos(xy^2) = \frac{\partial X_2}{\partial x}.$$

This example is easily generalized to $\mathbb{R}^n$. In particular, if $f\colon A \to \mathbb{R}$ is a C^1 function, then its gradient defines a vector field X; namely, $p \mapsto \mathrm{grad}\, f(p)$. In general, if f is C^2, then its second partials commute; that is,

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(p) = \frac{\partial^2 f}{\partial x_j \partial x_i}(p), \quad 1 \leq i, j \leq n,$$

so this gradient vector field $X = (X_1, \ldots, X_n)$ has a very special property:

$$\frac{\partial X_i}{\partial x_j} = \frac{\partial X_j}{\partial x_i}, \quad 1 \leq i, j \leq n.$$

This is a necessary condition for a vector field to be the gradient of some function, but not a sufficient condition in general. The existence of such a function depends on the topological shape of the domain A. Understanding what are sufficient conditions to answer the above question led to the development of differential forms and cohomology.

Definition 11.11. Given a vector field $X\colon A \to \mathbb{R}^n$, for any point $p_0 \in A$, a C^1 curve $\gamma\colon (-\epsilon, \epsilon) \to \mathbb{R}^n$ (with $\epsilon > 0$) is an *integral curve for X with initial condition p_0* if $\gamma(0) = p_0$, and

$$\gamma'(t) = X(\gamma(t)) \quad \text{for all } t \in (-\epsilon, \epsilon).$$

An integral curve has the property that for every time $t \in (-\epsilon, \epsilon)$, the tangent vector $\gamma'(t)$ to the curve γ at the point $\gamma(t)$ coincides with the vector $X(\gamma(t))$ given by the vector field at the point $\gamma(t)$. See Figure 11.7.

Definition 11.12. Given a C^1 vector field $X\colon A \to \mathbb{R}^n$, for any point $p_0 \in A$, a *local flow for X at p_0* is a function

$$\varphi\colon J \times U \to \mathbb{R}^n,$$

where $J \subseteq \mathbb{R}$ is an open interval containing 0 and U is an open subset of A containing p_0, so that for every $p \in U$, the curve $t \mapsto \varphi(t, p)$ is an integral curve of X with initial condition p. See Figure 11.8

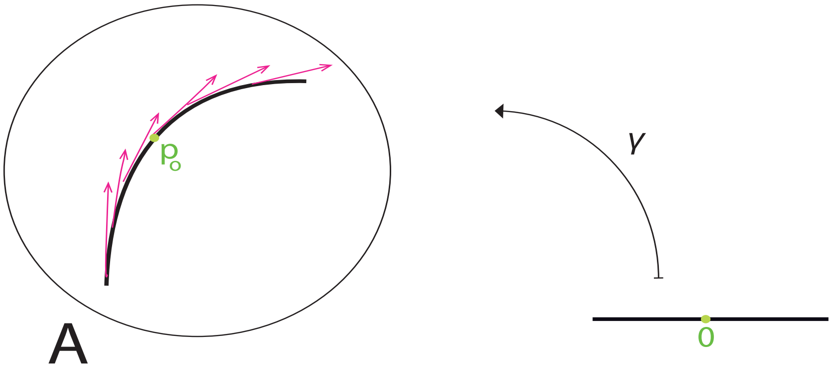


Figure 11.7: An integral curve in $\mathbb{R}^2$.

The theory of ODE tells us that if X is C^1, then for every $p_0 \in A$, there is a pair (J, U) as above such that there is a *unique* C^1 local flow $\varphi\colon J \times U \to \mathbb{R}^n$ for X at p_0.

Let us now consider the special class of vector fields induced by matrices in $\mathrm{M}_n(\mathbb{R})$. For any matrix $A \in \mathrm{M}_n(\mathbb{R})$, let X_A be the vector field given by

$$X_A(p) = Ap \quad \text{for all } p \in \mathbb{R}^n.$$

Such vector fields are obviously C^1 (in fact, C^∞).

The vector field induced by the matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

is shown in Figure 11.9. Integral curves are circles of center $(0, 0)$.

It turns out that the local flows of X_A are global, in the sense that $J = \mathbb{R}$ and $U = \mathbb{R}^n$, and that they are given by the matrix exponential. The proof of this fact relies on the observation that the map $f\colon t \mapsto e^{tA}$, where A is any matrix $A \in \mathrm{M}_n(\mathbb{C})$, is represented by a power series with infinite radius of convergence. An application of Propositions 11.9 and 11.10 to this power series implies that

$$f'(t) = \sum_{k=1}^{\infty} k \frac{t^{k-1}A^k}{k!} = A \sum_{k=1}^{\infty} \frac{t^{k-1}A^{k-1}}{(k-1)!} = Ae^{tA}.$$

Note that

$$Ae^{tA} = e^{tA}A.$$

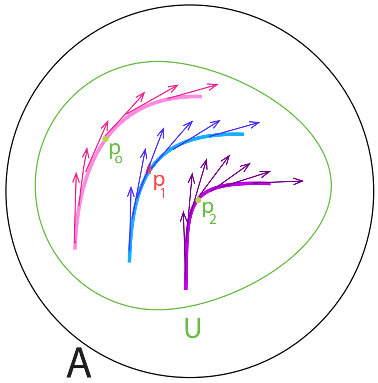


Figure 11.8: A portion of local flow $\varphi\colon J \times U \to \mathbb{R}^2$. If p is fixed and t varies, the flow moves along one of the colored curves. If t is fixed and p varies, p acts as a parameter for the individually colored curves.

Proposition 11.25. *For any matrix $A \in \mathrm{M}_n(\mathbb{R})$, for any $p_0 \in \mathbb{R}^n$, there is a unique local flow $\varphi\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ for the vector field X_A given by*

$$\varphi(t, p) = e^{tA}p,$$

for all $t \in \mathbb{R}$ and all $p \in \mathbb{R}^n$.

Proof. For any $p \in \mathbb{R}^n$, write $\gamma_p(t) = \varphi(t, p)$. We claim that $\gamma_p(t) = e^{tA}p$ is the unique integral curve for X_A with initial condition p.

We have

$$\gamma_p'(t) = (e^{tA}p)'(t) = Ae^{tA}p = A\gamma_p(t) = X_A(\gamma_p(t)),$$

which shows that γ_p is an integral curve for X_A with initial condition p.

Say θ is another integral curve for X_A with initial condition p. Let us compute the derivative of the function $t \mapsto e^{-tA}\theta(t)$. Using the product rule and the fact that $\theta'(t) = X_A(\theta(t)) = A\theta(t)$, we have

$$\begin{aligned}(e^{-tA}\theta)'(t) &= (e^{-tA})'(t)\theta(t) + e^{-tA}\theta'(t)\\ &= e^{-tA}(-A)\theta(t) + e^{-tA}A\theta(t)\\ &= -e^{-tA}A\theta(t) + e^{-tA}A\theta(t) = 0.\end{aligned}$$

Therefore, by Proposition 11.22, the function $t \mapsto e^{-tA}\theta(t)$ is constant on $\mathbb{R}$. Furthermore, since $\theta(0) = p$, its value is p, so

$$e^{-tA}\theta(t) = p \quad \text{for all } t \in \mathbb{R}.$$

Therefore, $\theta(t) = e^{tA}p = \gamma_p(t)$, establishing uniqueness. □

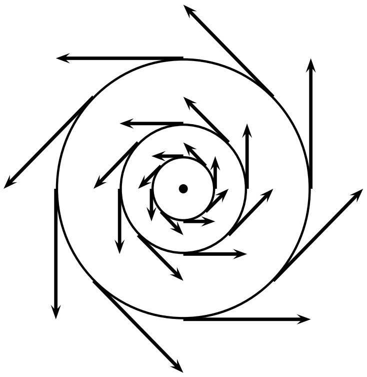


Figure 11.9: A vector field in $\mathbb{R}^2$.

For t fixed, the map $\Phi_t \colon p \mapsto e^{tA}p$ is a smooth diffeomorphism of $\mathbb{R}^n$ (with inverse given by e^{-tA}). We can think of Φ_t as the map which, given any p, moves p along the integral curve γ_p from p to $\gamma_p(t) = e^{tA}p$. For the vector field of Figure 11.9, each Φ_t is the rotation

$$e^{tA} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

The map $\Phi \colon \mathbb{R} \to \mathrm{Diff}(\mathbb{R}^n)$ is a group homomorphism, because

$$\Phi_s \circ \Phi_t = e^{sA}e^{tA}p = e^{(s+t)A}p = \Phi_{s+t} \quad \text{for all } s, t \in \mathbb{R}.$$

Observe that $\Phi_t(p) = \varphi(t, p)$. If we hold p fixed, we obtain the integral curve with initial condition p, which is also called a *flow line* of the local flow. If we hold t fixed, we obtain a smooth diffeomorphism of $\mathbb{R}^n$ (moving p to $\varphi(t, p)$). The family $\{\Phi_t\}_{t\in\mathbb{R}}$ is called the 1*-parameter group generated by* X_A, and Φ is called the *(global) flow generated by* X_A.

In the case of 2×2 matrices, it is possible to describe explicitly the shape of all integral curves; see Rossmann [98] (Section 1.1).

11.4 Problems

Problem 11.1. Prove Proposition 11.3.

Problem 11.2.

(i) Let $(E, \| \, \|)$ be a finite dimensional normed vector space. Prove that a a series is absolutely convergence iff it is unconditionally convergent.

(ii) (Advanced) Now assume E has infinite dimension. Find a counter example to Part (i), namely find a series that is unconditionally convergent yet not absolutely convergent.

Problem 11.3. Let $f(z) = \sum_{k=1}^{\infty} a_k z^k$ be a power series with complex coefficients and radius of convergence R, with $R > 0$. Find an example of a series and a matrix A with $\|A\| > R$ such that $f(A)$ converges.

Problem 11.4. Prove Propositions 11.12 and 11.13.

Problem 11.5. Prove Proposition 11.14.

Problem 11.6. Let E be a finite dimensional normed vector space and let F be a (possibly infinite) normed vector space. Show that every linear map $f \colon E \to F$ is continuous.

Problem 11.7. Prove Propositions 11.16 and 11.17.

Problem 11.8. Prove Theorem 11.19.

Problem 11.9. Prove the following: if $f \colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ and $g \colon \mathrm{M}_n(\mathbb{C}) \to \mathrm{M}_n(\mathbb{C})$ are differentiable matrix functions, then

$$d(fg)_A(B) = df_A(B)g(A) + f(A)dg_A(B),$$

for all $A, B \in \mathrm{M}_n(\mathbb{C})$.

Problem 11.10. Consider the function $f \colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(0,0) = 0$ and

$$f(x,y) = \frac{x^2 y}{x^4 + y^2} \quad \text{if } (x,y) \neq (0,0).$$

Show that the partial derivatives of the function f exist at $a = (0,0) \in \mathbb{R}^2$, but yet, f is not differentiable at a, and not even continuous at a.

Chapter 12

A Review of Point Set Topology

This chapter contains a review of the topological concepts necessary for studying differential geometry and includes the following material:

1. The definition of a topological space in terms of open sets;
2. The definition of a basis for a topology;
3. The definition of the subspace topology;
4. The definition of the product topology;
5. The definition of continuity and notion of a homeomorphism;
6. The definition of a limit of a sequence;
7. The definition of connectivity and path-wise connectivity;
8. The definition of compactness;
9. The definition of the quotient topology.

Readers familiar with this material may proceed to Chapter 3.

12.1 Topological Spaces

We begin with the notion of a topological space.

Definition 12.1. Given a set E, a *topology on* E *(or a topological structure on* E*)*, is defined as a family $\mathcal{O}$ of subsets of E, called *open sets*, which satisfy the following three properties:

(1) For every finite family $(U_i)_{1\leq i\leq n}$ of sets $U_i \in \mathcal{O}$, we have $U_1 \cap \cdots \cap U_n \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under finite intersections.

(2) For every arbitrary family $(U_i)_{i\in I}$ of sets $U_i \in \mathcal{O}$, we have $\bigcup_{i\in I} U_i \in \mathcal{O}$, i.e., $\mathcal{O}$ is closed under arbitrary unions.

(3) $\emptyset \in \mathcal{O}$, and $E \in \mathcal{O}$, i.e., $\emptyset$ and E belong to $\mathcal{O}$.

A set E together with a topology $\mathcal{O}$ on E is called a *topological space*. Given a topological space $(E, \mathcal{O})$, a subset F of E is a *closed set* if $F = E - U$ for some open set $U \in \mathcal{O}$, i.e., F is the complement of some open set.

By taking complements, we can state properties of the closed sets dual to those of Definition 12.1. Thus, $\emptyset$ and E are closed sets, and the closed sets are closed under finite unions and arbitrary intersections.

It is possible that an open set is also a closed set. For example, $\emptyset$ and E are both open and closed. When a topological space contains a proper nonempty subset U which is both open and closed, the space E is said to be *disconnected*.

The reader is probably familiar with a certain class of topological spaces known as metric spaces. Recall that a *metric space* is a set E together with a function $d\colon E \times E \to \mathbb{R}_+$, called a *metric, or distance*, assigning a nonnegative real number $d(x, y)$ to any two points $x, y \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(D1) $d(x, y) = d(y, x)$. (symmetry)

(D2) $d(x, y) \geq 0$, and $d(x, y) = 0$ iff $x = y$. (positivity)

(D3) $d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

For example, let $E = \mathbb{R}^n$ (or $E = \mathbb{C}^n$). We have the *Euclidean metric*

$$d_2(x, y) = \left(|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2\right)^{\frac{1}{2}}.$$

This particular metric is called the *Euclidean norm*, $\|x - y\|_2$, where a *norm on E* is a function $\|\ \|\colon E \to \mathbb{R}_+$, assigning a nonnegative real number $\|u\|$ to any vector $u \in E$, and satisfying the following conditions for all $x, y, z \in E$:

(N1) $\|x\| \geq 0$, and $\|x\| = 0$ iff $x = 0$. (positivity)

(N2) $\|\lambda x\| = |\lambda|\,\|x\|$. (scaling)

(N3) $\|x + y\| \leq \|x\| + \|y\|$. (triangle inequality)

Given a metric space E with metric d, for every $a \in E$, for every $\rho \in \mathbb{R}$, with $\rho > 0$, the set

$$B(a, \rho) = \{x \in E \mid d(a, x) \leq \rho\}$$

is called the *closed ball of center a and radius ρ*, the set

$$B_0(a,\rho) = \{x \in E \mid d(a,\, x) < \rho\}$$

is called the *open ball of center a and radius ρ*, and the set

$$S(a,\, \rho) = \{x \in E \mid d(a,\, x) = \rho\}$$

is called the *sphere of center a and radius ρ*. It should be noted that ρ is finite (i.e., not $+\infty$). Clearly, $B(a,\rho) = B_0(a,\rho) \cup S(a,\, \rho)$. Furthermore, any metric space E is a topological space with $\mathcal{O}$ being the family of arbitrary unions of open balls. See Figure 12.1.

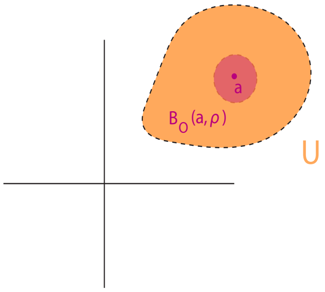


Figure 12.1: An open set U in $E = \mathbb{R}^2$ under the standard Euclidean metric. Any point in the peach set U is surrounded by a small raspberry open ball $B_0(a,\rho)$ which lies within U.

One should be careful that, in general, the family of open sets is not closed under infinite intersections. For example, in $\mathbb{R}$ under the metric $|x-y|$, letting $U_n = (-1/n,\, +1/n)$, each U_n is open, but $\bigcap_n U_n = \{0\}$, which is not open.

A topological space $(E, \mathcal{O})$ is said to satisfy the *Hausdorff separation axiom (or T_2-separation axiom)* if for any two distinct points $a \neq b$ in E, there exist two open sets U_a and U_b such that, $a \in U_a$, $b \in U_b$, and $U_a \cap U_b = \emptyset$. When the T_2-separation axiom is satisfied, we also say that $(E, \mathcal{O})$ is a *Hausdorff space*. See Figure 12.2.

Any metric space is a topological Hausdorff space. Similarly, any normed vector space is a topological Hausdorff space, the family of open sets being the family of arbitrary unions of open balls. The topology $\mathcal{O}$ consisting of all subsets of E is called the *discrete topology*.

Remark: Most (if not all) spaces used in analysis are Hausdorff spaces. Intuitively, the Hausdorff separation axiom says that there are enough "small" open sets. Without this axiom, some counter-intuitive behaviors may arise. For example, a sequence may have more than one limit point (or a compact set may not be closed).

It is also worth noting that the Hausdorff separation axiom implies that for every $a \in E$, the set $\{a\}$ is closed. Indeed, if $x \in E - \{a\}$, then $x \neq a$, and so there exist open sets U_a

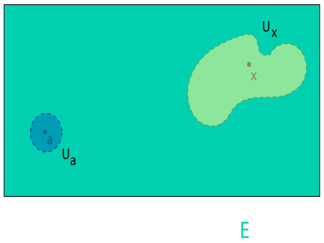


Figure 12.2: A schematic illustration of the Hausdorff separation property.

and U_x such that $a \in U_a$, $x \in U_x$, and $U_a \cap U_x = \emptyset$. Thus, for every $x \in E - \{a\}$, there is an open set U_x containing x and contained in $E - \{a\}$, showing by (O3) that $E - \{a\}$ is open, and thus that the set $\{a\}$ is closed.

Given a topological space, $(E, \mathcal{O})$, given any subset A of E, since $E \in \mathcal{O}$ and E is a closed set, the family $\mathcal{C}_A = \{F \mid A \subseteq F,\ F \text{ a closed set}\}$ of closed sets containing A is nonempty, and since any arbitrary intersection of closed sets is a closed set, the intersection $\bigcap \mathcal{C}_A$ of the sets in the family $\mathcal{C}_A$ is the smallest closed set containing A. By a similar reasoning, the union of all the open subsets contained in A is the largest open set contained in A.

Definition 12.2. Given a topological space $(E, \mathcal{O})$, for any subset A of E, the smallest closed set containing A is denoted by $\overline{A}$, and is called the *closure or adherence of* A. See Figure 12.3. A subset A of E is *dense in* E if $\overline{A} = E$. The largest open set contained in A is denoted by $\mathring{A}$, and is called the *interior of* A. See Figure 12.4. The set $\mathrm{Fr}\, A = \overline{A} \cap \overline{E - A}$ is called the *boundary (or frontier) of* A. See Figure 12.5. We also denote the boundary of A by ∂A.

Remark: The notation $\overline{A}$ for the closure of a subset A of E is somewhat unfortunate, since $\overline{A}$ is often used to denote the set complement of A in E. Still, we prefer it to more cumbersome notations such as $\mathrm{clo}(A)$, and we denote the complement of A in E by $E - A$ (or sometimes, A^c).

By definition, it is clear that a subset A of E is closed iff $A = \overline{A}$. The set $\mathbb{Q}$ of rationals is dense in $\mathbb{R}$. It is easily shown that $\overline{A} = \mathring{A} \cup \partial A$ and $\mathring{A} \cap \partial A = \emptyset$. Another useful characterization of $\overline{A}$ is given by the following proposition. Since this a review chapter, we will not provide proofs of the theorems and propositions and instead refer the reader to Massey [78, 79], Armstrong [5], and Munkres [89].

Proposition 12.1. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, the closure $\overline{A}$ of A is the set of all points $x \in E$ such that for every open set U containing x, $U \cap A \neq \emptyset$. See Figure 12.6.*

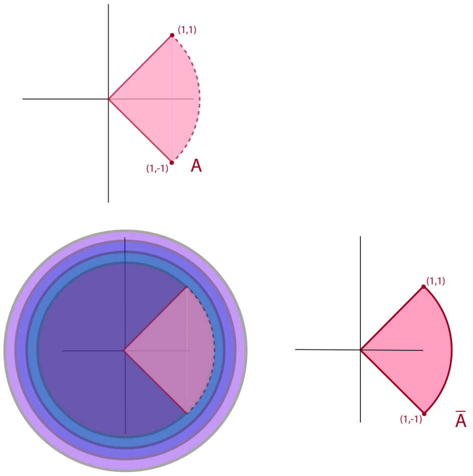


Figure 12.3: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The closure of A is obtained by the intersection of A with the closed unit ball.

Often it is necessary to consider a subset A of a topological space E, and to view the subset A as a topological space. The following definition shows how to define a topology on a subset.

Definition 12.3. Given a topological space $(E, \mathcal{O})$, given any subset A of E, the *subspace topology on* A *induced by* $\mathcal{O}$ is the family $\mathcal{U}$ of open sets defined such that

$$\mathcal{U} = \{U \cap A \mid U \in \mathcal{O}\}$$

is the family of all subsets of A obtained as the intersection of any open set in $\mathcal{O}$ with A. We say that $(A, \mathcal{U})$ has the *subspace topology*. If (E, d) is a metric space, the restriction $d_A \colon A \times A \to \mathbb{R}_+$ of the metric d to A is called the *subspace metric*.

For example, if $E = \mathbb{R}^n$ and d is the Euclidean metric, we obtain the subspace topology on the closed n-cube

$$\{(x_1, \ldots, x_n) \in E \mid a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}.$$

See Figure 12.7.

One should realize that every open set $U \in \mathcal{O}$ which is entirely contained in A is also in the family $\mathcal{U}$, but $\mathcal{U}$ may contain open sets that are not in $\mathcal{O}$. For example, if $E = \mathbb{R}$ with $|x - y|$, and $A = [a, b]$, then sets of the form $[a, c)$, with $a < c < b$ belong to $\mathcal{U}$, but they are not open sets for $\mathbb{R}$ under $|x - y|$. However, there is agreement in the following situation.

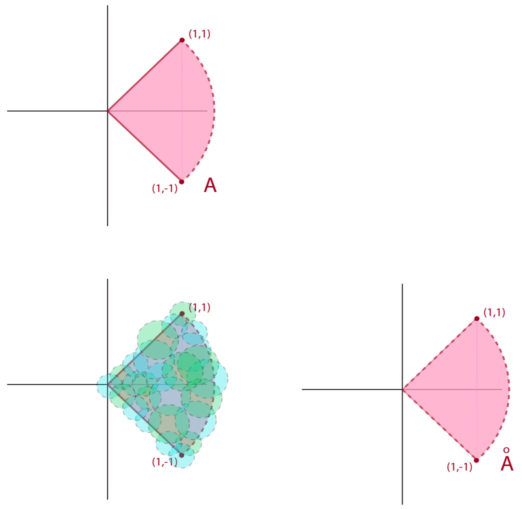


Figure 12.4: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The interior of A is obtained by covering A with small open balls.

Proposition 12.2. *Given a topological space $(E, \mathcal{O})$, given any subset A of E, if $\mathcal{U}$ is the subspace topology, then the following properties hold.*

(1) If A is an open set $A \in \mathcal{O}$, then every open set $U \in \mathcal{U}$ is an open set $U \in \mathcal{O}$.

(2) If A is a closed set in E, then every closed set w.r.t. the subspace topology is a closed set w.r.t. $\mathcal{O}$.

The concept of product topology is also useful.

Definition 12.4. Given n topological spaces $(E_i, \mathcal{O}_i)$, the *product topology on* $E_1 \times \cdots \times E_n$ is the family $\mathcal{P}$ of subsets of $E_1 \times \cdots \times E_n$ defined as follows: if

$$\mathcal{B} = \{U_1 \times \cdots \times U_n \mid U_i \in \mathcal{O}_i,\ 1 \leq i \leq n\},$$

then $\mathcal{P}$ is the family consisting of arbitrary unions of sets in $\mathcal{B}$, including $\emptyset$. The set, $E_1 \times \cdots \times E_n$, when given the product topology, is called the *product space*. See Figure 12.8.

It can be verified that when $E_i = \mathbb{R}$, with the standard topology induced by $|x - y|$, the product topology on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm. This equality between the two topologies suggestion the following definition.

Definition 12.5. Two metrics d and d' on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E (i.e., they define the same family $\mathcal{O}$ of open sets). Similarly, two norms $\| \; \|$ and $\| \; \|'$ on a space E are *equivalent* if they induce the same topology $\mathcal{O}$ on E.

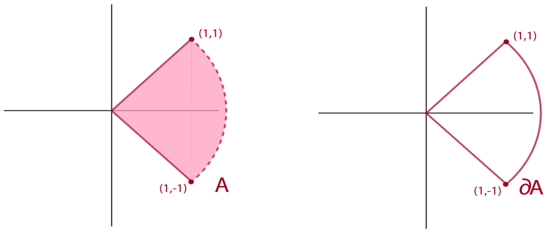


Figure 12.5: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The subset A is the section $B_0(1)$ in the first and fourth quadrants bound by the lines $y = x$ and $y = -x$. The boundary of A is $\overline{A} \cap \overset{\circ}{A}$.

Given a topological space $(E, \mathcal{O})$, it is often useful, as in Definition 12.4, to define the topology $\mathcal{O}$ in terms of a subfamily $\mathcal{B}$ of subsets of E.

Definition 12.6. We say that a family $\mathcal{B}$ of subsets of E is a *basis for the topology* $\mathcal{O}$, if $\mathcal{B}$ is a subset of $\mathcal{O}$, and if every open set U in $\mathcal{O}$ can be obtained as some union (possibly infinite) of sets in $\mathcal{B}$ (agreeing that the empty union is the empty set). A *subbasis for* $\mathcal{O}$ is a family $\mathcal{S}$ of subsets of E, such that the family $\mathcal{B}$ of all finite intersections of sets in $\mathcal{S}$ (including E itself, in case of the empty intersection) is a basis of $\mathcal{O}$.

For example, given any metric space (E, d), $\mathcal{B} = \{B_0(a, \rho)\}$. In particular, if $d = \| \, \|_2$, the open intervals form a basis for $\mathbb{R}$, while the open disks form a basis for $\mathbb{R}^2$. The open rectangles also form a basis for $\mathbb{R}^2$ with the standard topology. See Figure 12.9.

It is immediately verified that if a family $\mathcal{B} = (U_i)_{i \in I}$ is a basis for the topology of $(E, \mathcal{O})$, then $E = \bigcup_{i \in I} U_i$, and the intersection of any two sets $U_i, U_j \in \mathcal{B}$ is the union of some sets in the family $\mathcal{B}$ (again, agreeing that the empty union is the empty set). Conversely, a family $\mathcal{B}$ with these properties is the basis of the topology obtained by forming arbitrary unions of sets in $\mathcal{B}$.

The following proposition gives useful criteria for determining whether a family of open subsets is a basis of a topological space.

Proposition 12.3. *Given a topological space $(E, \mathcal{O})$ and a family $\mathcal{B}$ of open subsets in $\mathcal{O}$ the following properties hold:*

(1) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff for every open set $U \in \mathcal{O}$ and every $x \in U$, there is some $B \in \mathcal{B}$ such that $x \in B$ and $B \subseteq U$. See Figure 12.10.

(2) The family $\mathcal{B}$ is a basis for the topology $\mathcal{O}$ iff

(a) For every $x \in E$, there is some $B \in \mathcal{B}$ such that $x \in B$.

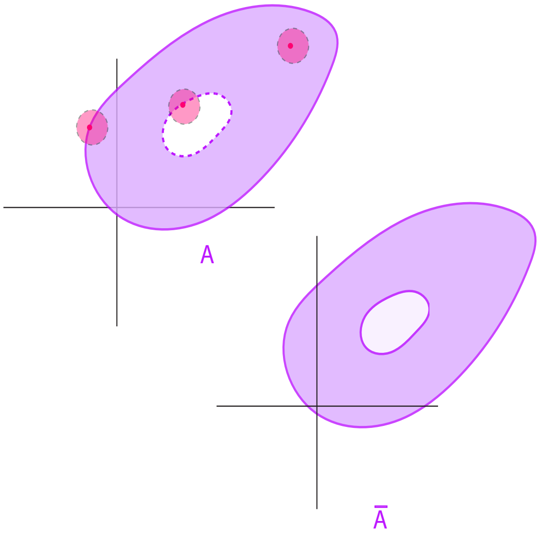


Figure 12.6: The topological space $(E, \mathcal{O})$ is $\mathbb{R}^2$ with topology induced by the Euclidean metric. The purple subset A is illustrated with three red points, each in its closure since the open ball centered at each point has nontrivial intersection with A.

(b) For any two open subsets, $B_1, B_2 \in \mathcal{B}$, for every $x \in E$, if $x \in B_1 \cap B_2$, then there is some $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$. See Figure 12.11.

We now consider the fundamental property of continuity.

12.2 Continuous Functions, Limits

Definition 12.7. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f \colon E \to F$ be a function. For every $a \in E$, we say that f *is continuous at* a, if for every open set $V \in \mathcal{O}_F$ containing $f(a)$, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U) \subseteq V$. We say that f *is continuous* if it is continuous at every $a \in E$.

Define a *neighborhood of* $a \in E$ as any subset N of E containing some open set $O \in \mathcal{O}$ such that $a \in O$.

It is easy to see that Definition 12.7 is equivalent to the following statements.

Proposition 12.4. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f \colon E \to F$ be a function. For every $a \in E$, the function f is continuous at $a \in E$ iff for every neighborhood N of $f(a) \in F$, then $f^{-1}(N)$ is a neighborhood of a. The function f is continuous on E iff $f^{-1}(V)$ is an open set in $\mathcal{O}_E$ for every open set $V \in \mathcal{O}_F$.*

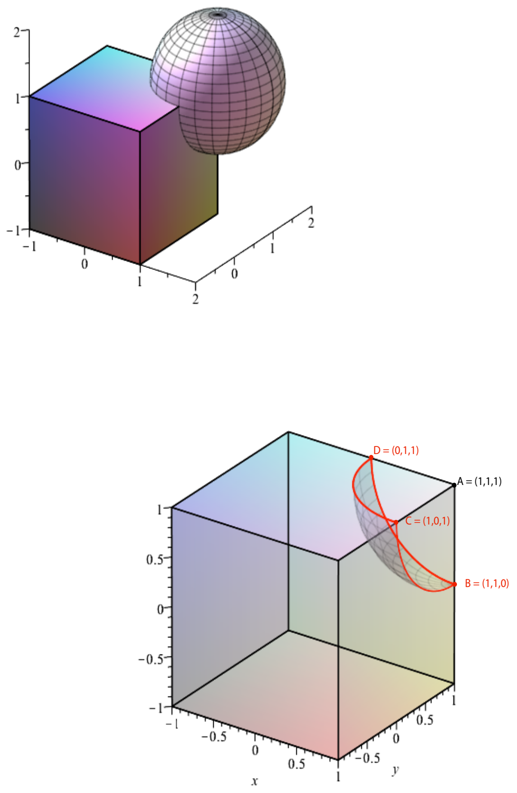


Figure 12.7: An example of an open set in the subspace topology for $\{(x, y, z) \in \mathbb{R}^3 \mid -1 \leq x \leq 1, -1 \leq y \leq 1, -1 \leq z \leq 1\}$. The open set is the corner region $ABCD$ and is obtained by intersection the cube $B_0((1, 1, 1), 1)$.

If E and F are metric spaces defined by metrics d_E and d_F, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } d_E(a, x) \leq \eta, \text{ then } d_F(f(a), f(x)) \leq \epsilon.$$

Similarly, if E and F are normed vector spaces defined by norms $\| \, \|_E$ and $\| \, \|_F$, we can show easily that f is continuous at a iff

for every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in E$,

$$\text{if } \|x - a\|_E \leq \eta, \text{ then } \|f(x) - f(a)\|_F \leq \epsilon.$$

It is worth noting that continuity is a topological notion, in the sense that equivalent metrics (or equivalent norms) define exactly the same notion of continuity.

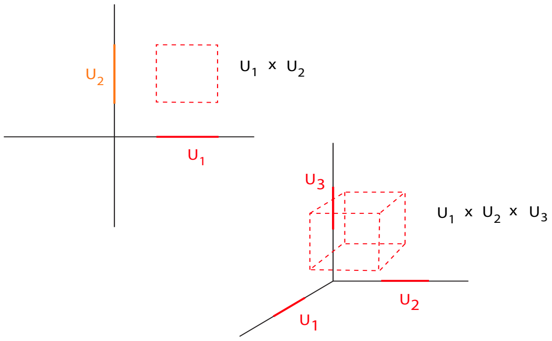


Figure 12.8: Examples of open sets in the product topology for $\mathbb{R}^2$ and $\mathbb{R}^3$ induced by the Euclidean metric.

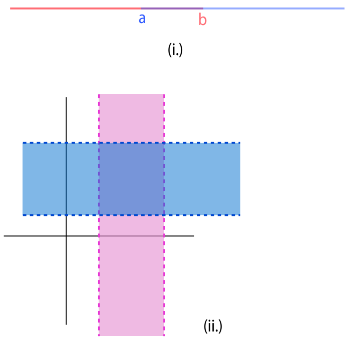


Figure 12.9: Figure $(i.)$ shows that the set of infinite open intervals forms a subbasis for $\mathbb{R}$. Figure $(ii.)$ shows that the infinite open strips form a subbasis for $\mathbb{R}^2$.

If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, and $f \colon E \to F$ is a function, for every nonempty subset $A \subseteq E$ of E, we say that f *is continuous on* A if the restriction of f to A is continuous with respect to $(A, \mathcal{U})$ and $(F, \mathcal{O}_F)$, where $\mathcal{U}$ is the subspace topology induced by $\mathcal{O}_E$ on A.

Given a product $E_1 \times \cdots \times E_n$ of topological spaces, as usual, we let $\pi_i \colon E_1 \times \cdots \times E_n \to E_i$ be the projection function such that, $\pi_i(x_1, \ldots, x_n) = x_i$. It is immediately verified that each π_i is continuous. In fact, it can be shown that the product topology is the smallest topology on $E_1 \times \cdots \times E_n$ for which each π_i is continuous.

Given a topological space $(E, \mathcal{O})$, we say that a point $a \in E$ is *isolated* if $\{a\}$ is an open set in $\mathcal{O}$. If $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ are topological spaces, any function $f \colon E \to F$ is continuous at every isolated point $a \in E$. In the discrete topology, every point is isolated.

The following proposition is easily shown.

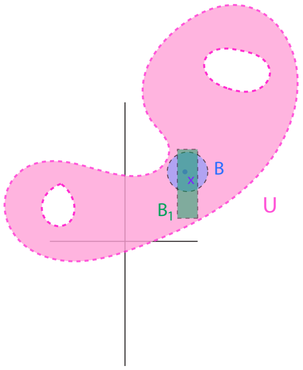


Figure 12.10: Given an open subset U of $\mathbb{R}^2$ and $x \in U$, there exists an open ball B containing x with $B \subset U$. There also exists an open rectangle B_1 containing x with $B_1 \subset U$.

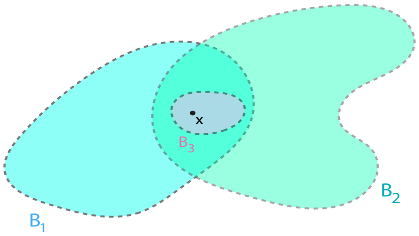


Figure 12.11: A schematic illustration of Condition (b) in Proposition 12.3.

Proposition 12.5. *Given topological spaces $(E, \mathcal{O}_E)$, $(F, \mathcal{O}_F)$, and $(G, \mathcal{O}_G)$, and two functions $f \colon E \to F$ and $g \colon F \to G$, if f is continuous at $a \in E$ and g is continuous at $f(a) \in F$, then $g \circ f \colon E \to G$ is continuous at $a \in E$. Given n topological spaces $(F_i, \mathcal{O}_i)$, for every function $f \colon E \to F_1 \times \cdots \times F_n$, f is continuous at $a \in E$ iff every $f_i \colon E \to F_i$ is continuous at a, where $f_i = \pi_i \circ f$.*

One can also show that in a metric space (E, d), $d \colon E \times E \to \mathbb{R}$ is continuous, where $E \times E$ has the product topology, and that for a normed vector space $(E, \| \, \|)$, the norm $\| \, \| \colon E \to \mathbb{R}$ is continuous.

Given a function $f \colon E_1 \times \cdots \times E_n \to F$, we can fix $n - 1$ of the arguments, say $a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n$, and view f as a function of the remaining argument,

$$x_i \mapsto f(a_1, \ldots, a_{i-1}, x_i, a_{i+1}, \ldots, a_n),$$

where $x_i \in E_i$. If f is continuous, it is clear that each f_i is continuous.

One should be careful that the converse is false! For example, consider the function $f \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, defined such that,

$$f(x, y) = \frac{xy}{x^2 + y^2} \quad \text{if } (x, y) \neq (0, 0), \quad \text{and} \quad f(0, 0) = 0.$$

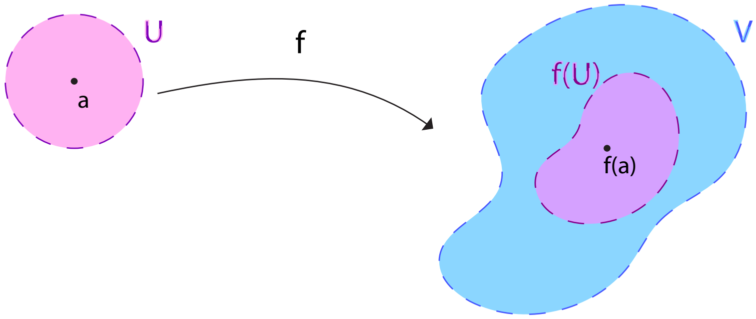


Figure 12.12: A schematic illustration of Definition 12.7

The function f is continuous on $\mathbb{R} \times \mathbb{R} - \{(0,0)\}$, but on the line $y = mx$, with $m \neq 0$, we have $f(x,y) = \frac{m}{1+m^2} \neq 0$, and thus, on this line, $f(x,y)$ does not approach 0 when (x,y) approaches $(0,0)$. See Figure 12.13.

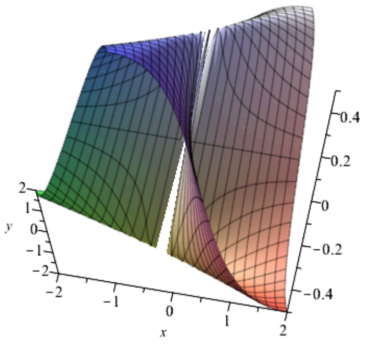


Figure 12.13: The graph of $f(x,y) = \frac{xy}{x^2+y^2}$ for $(x,y) \neq (0,0)$. The bottom of this graph, which shows the approach along the line $y = -x$, does not have a z value of 0.

The following proposition is useful for showing that real-valued functions are continuous.

Proposition 12.6. *If E is a topological space, and $(\mathbb{R}, |x-y|)$ is the reals under the standard topology, for any two functions $f \colon E \to \mathbb{R}$ and $g \colon E \to \mathbb{R}$, for any $a \in E$, for any $\lambda \in \mathbb{R}$, if f and g are continuous at a, then $f+g$, λf, $f \cdot g$, are continuous at a, and f/g is continuous at a if $g(a) \neq 0$.*

Remark: Proposition 12.6 is true if $\mathbb{R}$ is replaced with $\mathbb{C}$, where the $\mathbb{C}$ has the topology induced by the Euclidean norm on $\mathbb{R}^2$.

Using Proposition 12.6, we can show easily that every real or complex polynomial function is continuous.

The notion of isomorphism of topological spaces is defined as follows.

Definition 12.8. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, and let $f\colon E \to F$ be a function. We say that *f is a homeomorphism between E and F* if f is bijective, and both $f\colon E \to F$ and $f^{-1}\colon F \to E$ are continuous.

One should be careful that a bijective continuous function $f\colon E \to F$ is not necessarily an homeomorphism. For example, if $E = \mathbb{R}$ with the discrete topology, and $F = \mathbb{R}$ with the standard topology, the identity is not a homeomorphism.

We now introduce the concept of limit of a sequence. Given any set E, a *sequence* is any function $x\colon \mathbb{N} \to E$, usually denoted by $(x_n)_{n\in\mathbb{N}}$, or $(x_n)_{n\geq 0}$, or even by (x_n).

Definition 12.9. Given a topological space, $(E, \mathcal{O})$, we say that *a sequence $(x_n)_{n\in\mathbb{N}}$ converges to some $a \in E$* if for every open set U containing a, there is some $n_0 \geq 0$, such that, $x_n \in U$, for all $n \geq n_0$. We also say that *a is a limit of* $(x_n)_{n\in\mathbb{N}}$. See Figure 12.14.

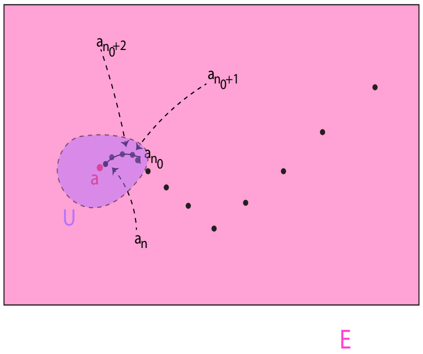


Figure 12.14: A schematic illustration of Definition 12.9.

When E is a metric space with metric d, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $d(x_n, a) \leq \epsilon$, for all $n \geq n_0$.

When E is a normed vector space with norm $\| \|$, it is easy to show that this is equivalent to the fact that,

for every $\epsilon > 0$, there is some $n_0 \geq 0$, such that, $\|x_n - a\| \leq \epsilon$, for all $n \geq n_0$.

The following proposition shows the importance of the Hausdorff separation axiom.

Proposition 12.7. *Given a topological space $(E, \mathcal{O})$, if the Hausdorff separation axiom holds, then every sequence has at most one limit.*

It is worth noting that the notion of limit is topological, in the sense that a sequence converge to a limit b iff it converges to the same limit b in any equivalent metric (and similarly for equivalent norms).

We still need one more concept of limit for functions.

Definition 12.10. Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be topological spaces, let A be some nonempty subset of E, and let $f \colon A \to F$ be a function. For any $a \in \overline{A}$ and any $b \in F$, we say that $f(x)$ *approaches b as x approaches a with values in A* if for every open set $V \in \mathcal{O}_F$ containing b, there is some open set $U \in \mathcal{O}_E$ containing a, such that, $f(U \cap A) \subseteq V$. See Figure 12.15. This is denoted by

$$\lim_{x \to a, x \in A} f(x) = b.$$

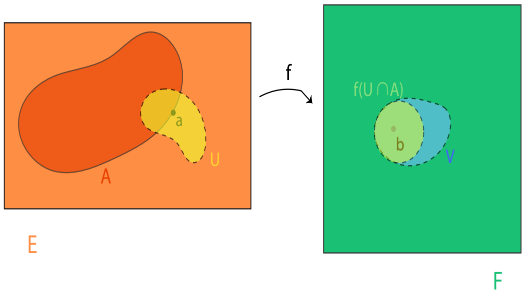


Figure 12.15: A schematic illustration of Definition 12.10.

First, note that by Proposition 12.1, since $a \in \overline{A}$, for every open set U containing a, we have $U \cap A \neq \emptyset$, and the definition is nontrivial. Also, even if $a \in A$, the value $f(a)$ of f at a plays no role in this definition. When E and F are metric space with metrics d_E and d_F, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } d_E(x, a) \leq \eta, \text{ then } d_F(f(x), b) \leq \epsilon.$$

When E and F are normed vector spaces with norms $\| \, \|_E$ and $\| \, \|_F$, it can be shown easily that the definition can be stated as follows:

For every $\epsilon > 0$, there is some $\eta > 0$, such that, for every $x \in A$,

$$\text{if } \|x - a\|_E \leq \eta, \text{ then } \|f(x) - b\|_F \leq \epsilon.$$

We have the following result relating continuity at a point and the previous notion.

Proposition 12.8. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f \colon E \to F$ be a function. For any $a \in E$, the function f is continuous at a iff $f(x)$ approaches $f(a)$ when x approaches a (with values in E).*

Another important proposition relating the notion of convergence of a sequence to continuity is stated without proof.

Proposition 12.9. *Let $(E, \mathcal{O}_E)$ and $(F, \mathcal{O}_F)$ be two topological spaces, and let $f \colon E \to F$ be a function.*

(1) If f is continuous, then for every sequence $(x_n)_{n \in \mathbb{N}}$ in E, if (x_n) converges to a, then $(f(x_n))$ converges to $f(a)$.

(2) If E is a metric space, and $(f(x_n))$ converges to $f(a)$ whenever (x_n) converges to a, for every sequence $(x_n)_{n \in \mathbb{N}}$ in E, then f is continuous.

We now turn to connectivity properties of topological spaces.

12.3 Connected Sets

Connectivity properties of topological spaces play a very important role in understanding the topology of surfaces.

Definition 12.11. A topological space $(E, \mathcal{O})$ is *connected* if the only subsets of E that are both open and closed are the empty set and E itself. Equivalently, $(E, \mathcal{O})$ is connected if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty open sets U, V, or if E cannot be written as the union $E = U \cup V$ of two disjoint nonempty closed sets. A topological space $(E, \mathcal{O})$ is *disconnected* if is not connected.

Definition 12.12. A subset $S \subseteq E$ is *connected* if it is connected in the subspace topology on S induced by $(E, \mathcal{O})$. Otherwise the subset S is *disconnected* which means there exits open subsets G and H of X such that S is the disjoint union of the two nonempty subsets $S \cap H$ and $S \cap G$. See Figure 12.16. A connected open set is called a *region* and a closed set is a *closed region* if its interior is a connected (open) set.

Most readers have an intuitive notion of the meaning of connectivity, namely that the space E is in "one piece." In particular, the following standard proposition characterizing the connected subsets of $\mathbb{R}$ can be found in most topology texts (for example, Munkres [89], Schwartz [103]).

Proposition 12.10. *A subset of the real line, $\mathbb{R}$, is connected iff it is an interval, i.e., of the form $[a, b]$, $(a, b]$, where $a = -\infty$ is possible, $[a, b)$, where $b = +\infty$ is possible, or (a, b), where $a = -\infty$ or $b = +\infty$ is possible.*

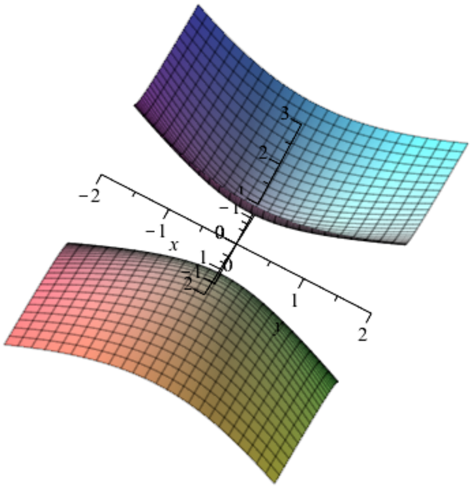


Figure 12.16: The graph of $z^2 - x^2 - y^2 = 1$ is disconnected in $\mathbb{R}^3$. Let $G = \{(x, y, z) | z > 0\}$ and $H = \{(x, y, z) | z < 0\}$.

A characterization of the connected subsets of $\mathbb{R}^n$ is harder and requires the notion of arcwise connectedness which we discuss at the end of this section.

One of the most important properties of connected sets is that they are preserved by continuous maps.

Proposition 12.11. *Given any continuous map $f \colon E \to F$, if $A \subseteq E$ is connected, then $f(A)$ is connected.*

An important corollary of Proposition 12.11 is that for every continuous function $f \colon E \to \mathbb{R}$, where E is a connected space, $f(E)$ is an interval. Indeed, this follows from Proposition 12.10. Thus, if f takes the values a and b where $a < b$, then f takes all values $c \in [a, b]$. This is property is the *intermediate value theorem.*

Here are two more properties of connected subsets.

Lemma 12.12. *Given a topological space, E, for any family, $(A_i)_{i \in I}$, of (nonempty) connected subsets of E, if $A_i \cap A_j \neq \emptyset$ for all $i, j \in I$, then the union $A = \bigcup_{i \in I} A_i$ of the family $(A_i)_{i \in I}$ is also connected.*

In particular, the above lemma applies when the connected sets in a family $(A_i)_{i \in I}$ have a point in common.

Lemma 12.13. *If A is a connected subset of a topological space, E, then for every subset, B, such that $A \subseteq B \subseteq \overline{A}$, where $\overline{A}$ is the closure of A in E, the set B is connected.*

In particular, Lemma 12.13 shows that if A is a connected subset, then its closure, $\overline{A}$, is also connected.

Connectivity provides a equivalence relation among the points of E.

Definition 12.13. Given a topological space, $(E, \mathcal{O})$, we say that two points $a, b \in E$ are *connected* if there is some connected subset A of E such that $a \in A$ and $b \in A$.

An application of Lemma 12.12 verifies that "a and b are connected in E" is an equivalence relation. The above equivalence relation defines a partition of E into nonempty disjoint *connected components*. The following proposition, proven via Lemmas 12.12 and 12.13, provides a way of constructing the connected components of E.

Proposition 12.14. *Given any topological space, E, for any $a \in E$, the connected component containing a is the largest connected set containing a. The connected components of E are closed.*

The connected components are the "pieces" of E. Intuitively, if a space is not connected, it is possible to define a continuous function which is constant on disjoint connected components and which takes possibly distinct values on disjoint components. This can be stated in terms of the concept of a locally constant function.

Definition 12.14. Given two topological spaces X and Y, a function $f \colon X \to Y$ is *locally constant* if for every $x \in X$, there is an open set $U \subseteq X$ such that $x \in U$ and f is constant on U.

We claim that a locally constant function is continuous. In fact, we will prove that $f^{-1}(V)$ is open for every subset, $V \subseteq Y$ (not just for an open set V). It is enough to show that $f^{-1}(y)$ is open for every $y \in Y$, since for every subset $V \subseteq Y$,

$$f^{-1}(V) = \bigcup_{y \in V} f^{-1}(y),$$

and open sets are closed under arbitrary unions. However, either $f^{-1}(y) = \emptyset$ if $y \in Y - f(X)$ or f is constant on $U = f^{-1}(y)$ if $y \in f(X)$ (with value y), and since f is locally constant, for every $x \in U$, there is some open set, $W \subseteq X$, such that $x \in W$ and f is constant on W, which implies that $f(w) = y$ for all $w \in W$ and thus, that $W \subseteq U$, showing that U is a union of open sets and thus, is open. The following proposition shows that a space is connected iff every locally constant function is constant.

Proposition 12.15. *A topological space is connected iff every locally constant function is constant. See Figure 12.17.*

The notion of a locally connected space is also useful.

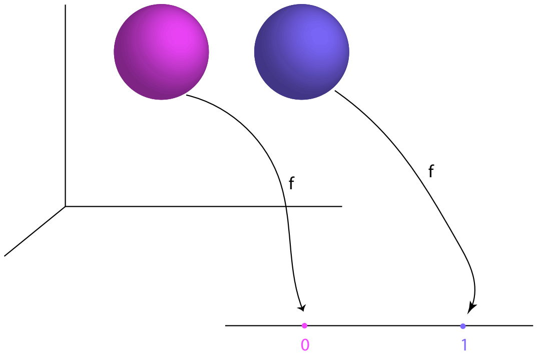


Figure 12.17: An example of a locally constant, but not constant, real-valued function f over the disconnected set consisting of the disjoint union of the two solid balls. On the pink ball, f is 0, while on the purple ball, f is 1.

Definition 12.15. A topological space, $(E, \mathcal{O})$, is *locally connected* if for every $a \in E$, for every neighborhood V of a, there is a connected neighborhood U of a such that $U \subseteq V$. See Figure 12.18

As we shall see in a moment, it would be equivalent to require that E has a basis of connected open sets.

There are connected spaces that are not locally connected and there are locally connected spaces that are not connected. The two properties are independent. For example, let X be a set with the discrete topology. Since $\{x\}$ is open for every $x \in X$, the topological space X is locally connected. However, if $|X| > 1$, then X, with the discrete topology, is not connected. On the other hand, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the y-axis, for which $-1 \leq y \leq 1$, is connected, but not locally connected. The open disk centered at $(0,1)$ with radius $\frac{1}{4}$ does not contain a connected neighborhood of $(0,1)$. See Figure 12.19.

Proposition 12.16. *A topological space, E, is locally connected iff for every open subset A of E, the connected components of A are open.*

Proposition 12.16 shows that in a locally connected space, the connected open sets form a basis for the topology. It is easily seen that $\mathbb{R}^n$ is locally connected. Manifolds are also locally connected.

Another very important property of surfaces and more generally, manifolds, is to be arcwise connected. The intuition is that any two points can be joined by a continuous arc of curve. This is formalized as follows.

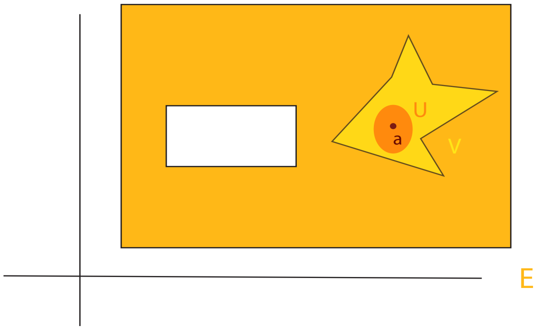


Figure 12.18: The topological space E, which is homeomorphic to an annulus, is locally connected since each point is surrounded by a small disk contained in E.

Definition 12.16. Given a topological space, $(E, \mathcal{O})$, an *arc (or path)* is a continuous map $\gamma\colon [a, b] \to E$, where $[a, b]$ is index[sub]$\gamma\colon [a, b]] \to E$a closed interval of the real line, $\mathbb{R}$. The point $\gamma(a)$ is the *initial point* of the arc and the point $\gamma(b)$ is the *terminal point* of the arc. We say that γ *is an arc joining* $\gamma(a)$ *and* $\gamma(b)$. See Figure 12.20. An arc is a *closed curve* if $\gamma(a) = \gamma(b)$. The set $\gamma([a, b])$ is the *trace* of the arc γ.

Typically, $a = 0$ and $b = 1$.

One should not confuse an arc $\gamma\colon [a, b] \to E$ with its trace. For example, γ could be constant, and thus, its trace reduced to a single point.

An arc is a *Jordan arc* if γ is a homeomorphism onto its trace. An arc $\gamma\colon [a, b] \to E$ is a *Jordan curve* if $\gamma(a) = \gamma(b)$ and γ is injective on $[a, b)$. Since $[a, b]$ is connected, by Proposition 12.11, the trace $\gamma([a, b])$ of an arc is a connected subset of E.

Given two arcs $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$ such that $\gamma(1) = \delta(0)$, we can form a new arc defined as follows:

Definition 12.17. Given two arcs, $\gamma\colon [0, 1] \to E$ and $\delta\colon [0, 1] \to E$, such that $\gamma(1) = \delta(0)$, we can form their *composition (or product)*, $\gamma\delta$, defined such that

$$\gamma\delta(t) = \begin{cases} \gamma(2t) & \text{if } 0 \leq t \leq 1/2; \\ \delta(2t - 1) & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

The *inverse* γ^{-1} *of the arc* γ is the arc defined such that $\gamma^{-1}(t) = \gamma(1 - t)$, for all $t \in [0, 1]$.

It is trivially verified that Definition 12.17 yields continuous arcs.

Definition 12.18. A topological space, E, is *arcwise connected* if for any two points $a, b \in E$, there is an arc $\gamma\colon [0, 1] \to E$ joining a and b, such that $\gamma(0) = a$ and $\gamma(1) = b$. A

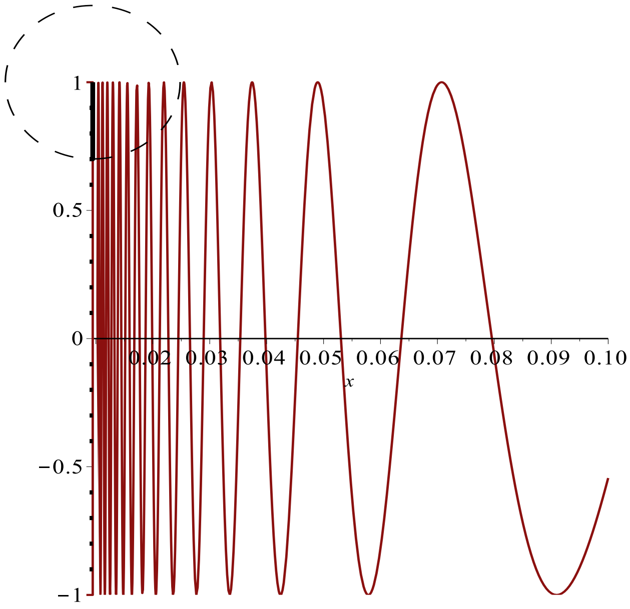


Figure 12.19: Let S be the graph of $f(x) = \sin(1/x)$ union the y-axis between -1 and 1. This space is connected, but not locally connected.

topological space, E, is *locally arcwise connected* if for every $a \in E$, for every neighborhood V of a, there is an arcwise connected neighborhood U of a such that $U \subseteq V$. See Figure 12.20.

The space $\mathbb{R}^n$ is locally arcwise connected, since for any open ball, any two points in this ball are joined by a line segment. Manifolds and surfaces are also locally arcwise connected. Proposition 12.11 also applies to arcwise connectedness. The following theorem is crucial to the theory of manifolds and surfaces.

Theorem 12.17. *If a topological space, E, is arcwise connected, then it is connected. If a topological space, E, is connected and locally arcwise connected, then E is arcwise connected.*

If E is locally arcwise connected, the above argument shows that the connected components of E are arcwise connected.

It is not true that a connected space is arcwise connected. For example, the space consisting of the graph of the function

$$f(x) = \sin(1/x),$$

where $x > 0$, together with the portion of the y-axis, for which $-1 \leq y \leq 1$, is connected, but not arcwise connected. See Figure 12.19.

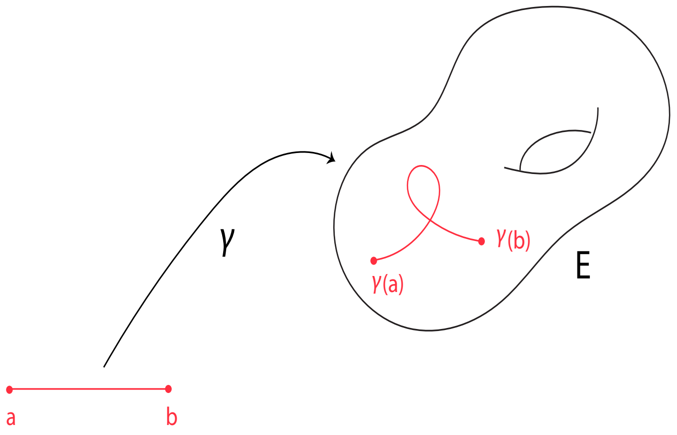


Figure 12.20: Let E be the torus with subspace topology induced from $\mathbb{R}^3$ with red arc $\gamma([a, b])$. The torus is both arcwise connected and locally arcwise connected.

A trivial modification of the proof of Theorem 12.17 shows that in a normed vector space, E, a connected open set is arcwise connected by polygonal lines (arcs consisting of line segments). This is because in every open ball, any two points are connected by a line segment. Furthermore, if E is finite dimensional, these polygonal lines can be forced to be parallel to basis vectors.

We conclude this section with the following theorem regarding the connectivity of product spaces.

Theorem 12.18. *Let X and Y be topological spaces. The product space $X \times Y$ is connected if and only if X and Y are connected.*

Remark: Theorem 12.18 can be extended to the set $\{X_i\}_{i=1}^n$, where n is a positive integer, $n \geq 2$.

We now consider compactness.

12.4 Compact Sets

The property of compactness is very important in topology and analysis. We provide a quick review geared towards the study of manifolds and for details, we refer the reader to Munkres [89], Schwartz [103]. In this section we will need to assume that the topological spaces are Hausdorff spaces. This is not a luxury, as many of the results are false otherwise.

There are various equivalent ways of defining compactness. For our purposes, the most convenient way involves the notion of open cover.

Definition 12.19. Given a topological space, E, for any subset A of E, an *open cover, $(U_i)_{i\in I}$ of* A, is a family of open subsets of E such that $A \subseteq \bigcup_{i\in I} U_i$. An *open subcover* of an open cover, $(U_i)_{i\in I}$ of A, is any subfamily, $(U_j)_{j\in J}$, which is an open cover of A, with $J \subseteq I$. An open cover, $(U_i)_{i\in I}$ of A, is *finite* if I is finite. See Figure 12.21.

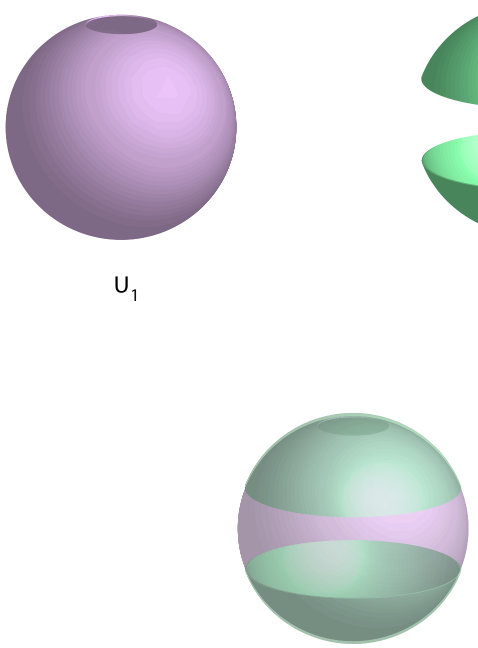


Figure 12.21: An open cover of S^2 using two open sets induced by the Euclidean topology of $\mathbb{R}^3$.

Definition 12.20. The topological space, E, is *compact* if it is Hausdorff and for every open cover, $(U_i)_{i\in I}$ of E, there is a finite open subcover $(U_j)_{j\in J}$ of E. Given any subset A of E, we say that A is *compact* if it is compact with respect to the subspace topology. We say that A is *relatively compact* if its closure $\overline{A}$ is compact.

It is immediately verified that a subset, A, of E is compact in the subspace topology relative to A iff for every open cover, $(U_i)_{i\in I}$ of A by open subsets of E, there is a finite open subcover $(U_j)_{j\in J}$ of A. The property that every open cover contains a finite open subcover is often called the *Heine-Borel-Lebesgue* property. By considering complements, a Hausdorff space is compact iff for every family, $(F_i)_{i\in I}$ of closed sets, if $\bigcap_{i\in I} F_i = \emptyset$, then $\bigcap_{j\in J} F_j = \emptyset$ for some finite subset J of I.

Definition 12.20 requires that a compact space be Hausdorff. There are books in which a compact space is not necessarily required to be Hausdorff. Following Schwartz, we prefer calling such a space *quasi-compact*.

Another equivalent and useful characterization can be given in terms of families having the finite intersection property. A family $(F_i)_{i\in I}$ of sets has the *finite intersection property* if $\bigcap_{j\in J} F_j \neq \emptyset$ for every finite subset J of I. We have the following proposition.

Proposition 12.19. *A topological Hausdorff space, E, is compact iff for every family $(F_i)_{i\in I}$ of closed sets having the finite intersection property, then $\bigcap_{i\in I} F_i \neq \emptyset$.*

Another useful consequence of compactness is as follows. For any family $(F_i)_{i\in I}$ of closed sets such that $F_{i+1} \subseteq F_i$ for all $i \in I$, if $\bigcap_{i\in I} F_i = \emptyset$, then $F_i = \emptyset$ for some $i \in I$. Indeed, there must be some finite subset J of I such that $\bigcap_{j\in J} F_j = \emptyset$, and since $F_{i+1} \subseteq F_i$ for all $i \in I$, we must have $F_j = \emptyset$ for the smallest F_j in $(F_j)_{j\in J}$. Using this fact, we note that $\mathbb{R}$ is *not* compact. Indeed, the family of closed sets, $([n, +\infty))_{n\geq 0}$, is decreasing and has an empty intersection.

Given a metric space, if we define a *bounded subset* to be a subset that can be enclosed in some closed ball (of finite radius), then any nonbounded subset of a metric space is not compact. However, a closed interval $[a, b]$ of the real line is compact, and by extension every closed set, $[a_1, b_1] \times \cdots \times [a_m, b_m]$, when considered as a subspace of $\mathbb{R}^m$, is compact.

The following two propositions give very important properties of the compact sets, and they only hold for Hausdorff spaces.

Proposition 12.20. *Given a topological Hausdorff space, E, for every compact subset, A, and every point b not in A, there exist disjoint open sets, U and V, such that $A \subseteq U$ and $b \in V$. See Figure 12.22. As a consequence, every compact subset is closed.*

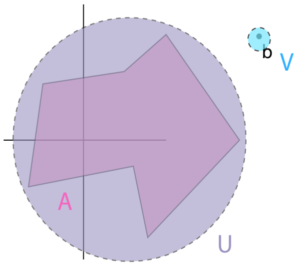


Figure 12.22: The compact set of $\mathbb{R}^2$, A, is separated by any point in its complement.

Proposition 12.21. *Given a topological Hausdorff space, E, for every pair of compact disjoint subsets, A and B, there exist disjoint open sets, U and V, such that $A \subseteq U$ and $B \subseteq V$.*

The following proposition shows that in a compact topological space, every closed set is compact.

Proposition 12.22. *Given a compact topological space, E, every closed set is compact.*

Remark: Proposition 12.22 also holds for quasi-compact spaces, i.e., the Hausdorff separation property is not needed.

Putting Proposition 12.21 and Proposition 12.22 together, we note that if X is compact, then for every pair of disjoint closed, sets A and B, there exist disjoint open sets, U and V, such that $A \subseteq U$ and $B \subseteq V$. We say that X is a *normal* space.

Proposition 12.23. *Given a compact topological space, E, for every $a \in E$, and for every neighborhood V of a, there exists a compact neighborhood U of a such that $U \subseteq V$. See Figure 12.23.*

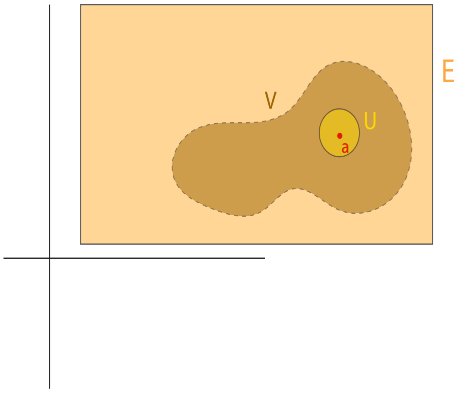


Figure 12.23: Let E be the peach square of $\mathbb{R}^2$. Each point of E is contained in a compact neighborhood U, in this case the small closed yellow disk.

It can be shown that in a normed vector space of finite dimension, a subset is compact iff it is closed and bounded. This is what we use to show that $\mathbf{SO}(n)$ is compact in $\mathbb{R}^{n^2}$. For $\mathbb{R}^n$, the proof is simple.

In a normed vector space of infinite dimension, there are closed and bounded sets that are not compact!

Another crucial property of compactness is that it is preserved under continuity.

Proposition 12.24. *Let E be a topological space and let F be a topological Hausdorff space. For every compact subset, A of E, and for every continuous map, $f\colon E \to F$, the subspace $f(A)$ is compact.*

As a corollary of Proposition 12.24, if E is compact, F is Hausdorff, and $f\colon E \to F$ is continuous and bijective, then f is a homeomorphism. Indeed, it is enough to show that f^{-1} is continuous, which is equivalent to showing that f maps closed sets to closed sets. However, closed sets are compact and Proposition 12.24 shows that compact sets are mapped to compact sets, which, by Proposition 12.20, are closed.

It can also be shown that if E is a compact nonempty space and $f\colon E \to \mathbb{R}$ is a continuous function, then there are points $a, b \in E$ such that $f(a)$ is the minimum of $f(E)$ and $f(b)$ is the maximum of $f(E)$. Indeed, $f(E)$ is a compact subset of $\mathbb{R}$ and thus, a closed and bounded set which contains its greatest lower bound and its least upper bound.

Another useful notion is that of local compactness. Indeed, manifolds and surfaces are locally compact.

Definition 12.21. A topological space, E, is *locally compact* if it is Hausdorff and for every $a \in E$, there is some compact neighborhood, K, of a. See Figure 12.23.

From Proposition 12.23, every compact space is locally compact but the converse is false. For example, the real line $\mathbb{R}$, which is not compact, is locally compact since each $x \in \mathbb{R}$, given any neighborhood N of x, there exist $\epsilon > 0$ such that $x \in [x - \epsilon, x + \epsilon] \subseteq N$. Furthermore, it can be shown that a normed vector space of finite dimension is locally compact.

Proposition 12.25. *Given a locally compact topological space, E, for every $a \in E$, and for every neighborhood N of a, there exists a compact neighborhood U of a such that $U \subseteq N$.*

Finally, in studying surfaces and manifolds, an important property is the existence of a countable basis for the topology.

Definition 12.22. A topological space E is called *second-countable* if there is a countable basis for its topology, i.e., if there is a countable family $(U_i)_{i \geq 0}$ of open sets such that every open set of E is a union of open sets U_i.

It is easily seen that $\mathbb{R}^n$ is second-countable and more generally, that every normed vector space of finite dimension is second-countable. We have the following property regarding second-countablility.

Proposition 12.26. *Given a second-countable topological space E, every open cover $(U_i)_{i \in I}$ of E contains some countable subcover.*

As an immediate corollary of Proposition 12.26, a locally connected second-countable space has countably many connected components.

In second-countable Hausdorff spaces, compactness can be characterized in terms of accumulation points (this is also true for metric spaces).

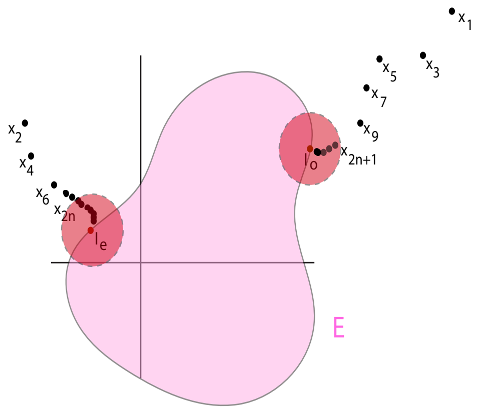


Figure 12.24: The space E is the closed, bounded pink subset of $\mathbb{R}^2$. The sequence (x_n) has two accumulation points, one for the subsequence (x_{2n+1}) and one for (x_{2n}).

Definition 12.23. Given a topological Hausdorff space, E, and given any sequence (x_n) of points in E, a point, $l \in E$, is an *accumulation point (or cluster point)* of the sequence (x_n) if every open set, U, containing l contains x_n for infinitely many n. See Figure 12.24.

Clearly, if l is a limit of the sequence (x_n), then it is an accumulation point, since every open set, U, containing a contains all x_n except for finitely many n. The following proposition provides another characterization of an accumulation point.

Proposition 12.27. *Given a second-countable topological Hausdorff space, E, a point, l, is an accumulation point of the sequence (x_n) iff l is the limit of some subsequence (x_{n_k}),of (x_n).*

Remark: Proposition 12.27 also holds for metric spaces.

As an illustration of Proposition 12.27, let $E = \mathbb{R}$ and let (x_n) be the sequence $(1, -1, 1, -1, \ldots)$. This sequence has two accumulation points, namely 1 and -1 since $(x_{2n+1}) = (1)$ and $(x_{2n}) = (-1)$.

The next proposition relates the existence of accumulation points to the notion of compactness.

Proposition 12.28. *A second-countable topological Hausdorff space, E, is compact iff every sequence (x_n) has some accumulation point.*

Remark: It should be noted that the proof showing that if E is compact, then every sequence has some accumulation point, holds for any arbitrary compact space (the proof does not use a countable basis for the topology). The converse also holds for metric spaces.

Closely related to Proposition 12.28 is the Bolzano-Weierstrass property which states that an infinite subset of a compact space has a limit point.

We end this section with a result about the product of compact spaces. But first we state the following proposition.

Proposition 12.29. *Let X and Y be topological spaces. The product space $X \times Y$ is a Hausdorff space iff X and Y are Hausdorff spaces.*

Remark: Proposition 12.29 is true for finite set of topological spaces, $\{X_i\}_{i=1}^n$, with $n \geq 2$.

Proposition 12.30. *Let $\{X_i\}_{i=1}^n$ be a family of topological spaces. The product space $X_1 \times \cdots \times X_n$ is compact iff X_i is compact for all $1 \leq i \leq n$.*

12.5 Quotient Spaces

In the final section of this chapter we discuss a topological construction, the quotient space, which plays important role in the study of orbifolds and homogenous manifolds. For example, real projective spaces and Grassmannians are obtained this way. In this situation, the natural topology on the quotient object is the quotient topology, but unfortunately, even if the original space is Hausdorff, the quotient topology may not be. Therefore, it is useful to have criteria that insure that a quotient topology is Hausdorff (or second-countable). We will present two criteria. First, let us review the notion of quotient topology. For more details, consult Munkres [89], Massey [78, 79], Armstrong [5], or Tu [112].

Definition 12.24. Given any topological space X and any set Y, for any surjective function $f \colon X \to Y$, we define the *quotient topology on Y determined by f* (also called the *identification topology on Y determined by f*), by requiring a subset V of Y to be open if $f^{-1}(V)$ is an open set in X. Given an equivalence relation R on a topological space X, if $\pi \colon X \to X/R$ is the projection sending every $x \in X$ to its equivalence class $[x]$ in X/R, the space X/R equipped with the quotient topology determined by π is called the *quotient space of X modulo R*. Thus, a set V of equivalence classes in X/R is open iff $\pi^{-1}(V)$ is open in X, which is equivalent to the fact that $\bigcup_{[x]\in V}[x]$ is open in X.

It is immediately verified that Definition 12.24 defines topologies and that $f \colon X \to Y$ and $\pi \colon X \to X/R$ are continuous when Y and X/R are given these quotient topologies.

To intuitively understand the quotient space construction, start with a topological space X, and form a partition $\mathcal{R}$ of X, where $\mathcal{R}$ is a collection of pairwise disjoint nonempty subsets whose union is X. The elements of $\mathcal{R}$ are subsets of X. This partition $\mathcal{R}$ defines the equivalent relation R, where $x \sim_R y$ iff x and y are in the same element of $\mathcal{R}$. Define a new topological space Y as follows. The points of Y are elements $\mathcal{R}$, and Y is formed by "gluing" together equivalent points of X into a single point. In other words, Y is homeomorphic to X/R and if $\pi\colon X \to Y$ maps each point in X to the subset of $\mathcal{R}$ which contains it, the topology of Y is the largest for which π is continuous.

We demonstrate this construction by building a cylinder as a quotient of the rectangle $Q = [0,2] \times [0,1]$. The partition $\mathcal{R} = \cup_{i\in I}\mathcal{R}_i$ of Q is defined as follows:

i. $\mathcal{R}_{(x,y)} = \{(x,y)\}$ where $0 < x < 2$ and $0 \le y \le 1$.

ii. $\mathcal{R}_y = \{(0,y),(2,y)\}$ where $0 \le y \le 1$

Each $\mathcal{R}_i$ is a point in Y and the function $\pi\colon Q \to Y$ maps (x,y) to the $\mathcal{R}_i$ which contains it. The map π "glues" together the left and right vertical edges of Q and forms a cylinder. See Figure 12.25.

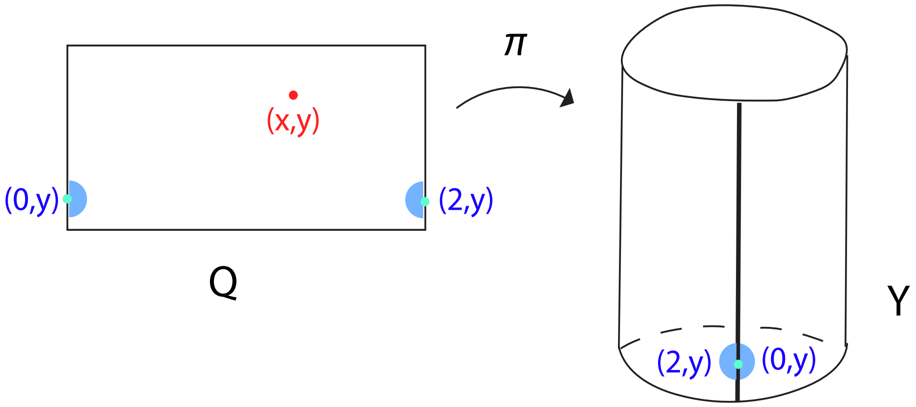


Figure 12.25: Constructing a cylinder as a quotient of a rectangle.

A similar construction creates a Möbius strip as a quotient of $Q = [0,2] \times [0,1]$. This time the partition $\mathcal{R} = \cup_{i\in I}\mathcal{R}_i$ of Q is

i. $\mathcal{R}_{(x,y)} = \{(x,y)\}$ where $0 < x < 2$ and $0 \le y \le 1$,

ii. $\mathcal{R}_y = \{(0,y),(2,1-y)\}$ where $0 \le y \le 1$.

This time the map $\pi\colon Q \to Y$ "glues" the left and right vertical edges with a twist and forms a Möbius strip. See Figure 12.26.

We can also build a torus as quotient of the unit square $S = [0,1] \times [0,1]$ by giving S the following partition $\mathcal{R} = \cup_{i\in I}\mathcal{R}_i$:

i. $\mathcal{R}_{(0,0)} = \{(0,0),(0,1),(1,1),(1,0)\}$.

ii. $\mathcal{R}_y = \{(0,y),(1,y)\}$ for $0 < y < 1$.

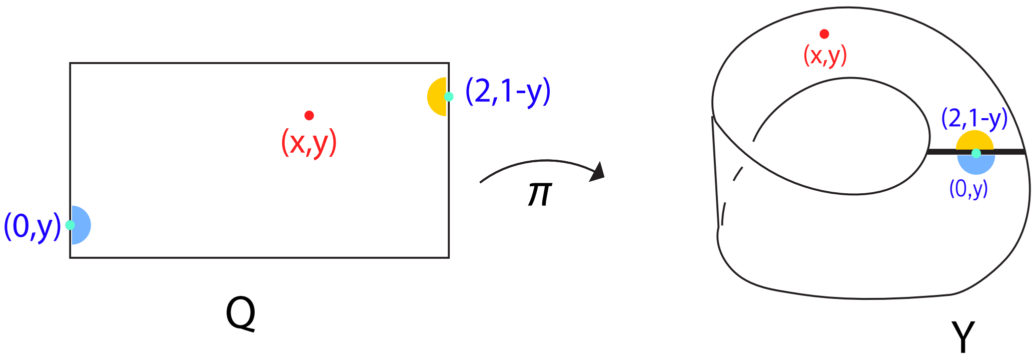


Figure 12.26: Constructing a Möbius strip as a quotient of a rectangle.

iii. $\mathcal{R}_x = \{(x,0),(x,1)\}$ for $0 < x < 1$.

iv. $\mathcal{R}_{(x,y)} = \{(x,y)\}$ for $0 < x < 1$ and $0 < y < 1$.

Once again each $\mathcal{R}_i$ is a point in Y and the function $\pi\colon Q \to Y$ maps (x,y) to the equivalence class $\mathcal{R}_i$ containing it. Geometrically π takes S, glues together the left and right edges to form a cylinder, then glues together the top and bottom of the cylinder to form the torus. See Figure 12.27.

Although we visualized the proceeding three quotients spaces in $\mathbb{R}^3$, the quotient construction, namely $\pi\colon Q \to Y$, is abstract and independent of any pictorial representation.

One should be careful that if X and Y are topological spaces and $f\colon X \to Y$ is a continuous surjective map, Y *does not* necessarily have the quotient topology determined by f. Indeed, it may not be true that a subset V of Y is open when $f^{-1}(V)$ is open. However, this will be true in two important cases.

Definition 12.25. A continuous map $f\colon X \to Y$ is an *open map* (or simply *open*) if $f(U)$ is open in Y whenever U is open in X, and similarly, $f\colon X \to Y$ is a *closed map* (or simply *closed*) if $f(F)$ is closed in Y whenever F is closed in X.

Then Y has the quotient topology induced by the continuous surjective map f if either f is open or f is closed. Indeed, if f is open, then assuming that $f^{-1}(V)$ is open in X, we have $f(f^{-1}(V)) = V$ open in Y. Now, since $f^{-1}(Y - B) = X - f^{-1}(B)$, for any subset B of Y, a subset V of Y is open in the quotient topology iff $f^{-1}(Y - V)$ is closed in X. From this, we can deduce that if f is a closed map, then V is open in Y iff $f^{-1}(V)$ is open in X.

Unfortunately, the Hausdorff separation property is not necessarily preserved under quotient. Nevertheless, it is preserved in some special important cases.

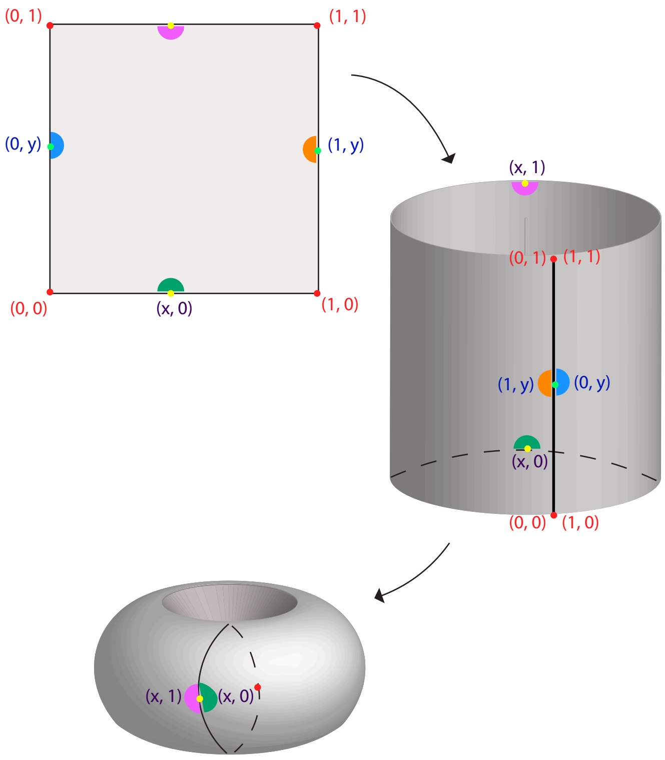


Figure 12.27: Constructing a torus as a quotient of a square.

Proposition 12.31. *Let X and Y be topological spaces, let $f\colon X \to Y$ be a continuous surjective map, and assume that X is compact and that Y has the quotient topology determined by f. Then Y is Hausdorff iff f is a closed map.*

Proof. Because X is compact, Proposition 12.22 implies that every closed set F in X is compact. An application of Proposition 12.24 shows that $f(F)$ is also compact. Since Y is Hausdorff, Proposition 12.20 tells us that $f(F)$ is closed, and we conclude that f is a closed map.

For the converse we use the fact that in a Hausdorff space E, if A and B are compact disjoint subsets of E, then there exist two disjoint open sets U and V such that $A \subseteq U$ and $B \subseteq V$. See Proposition 12.21.

Since X is Hausdorff, every set $\{a\}$ consisting of a single element $a \in X$ is closed, and since f is a closed map, $\{f(a)\}$ is also closed in Y. Since f is surjective, every set $\{b\}$ consisting of a single element $b \in Y$ is closed. If $b_1, b_2 \in Y$ and $b_1 \neq b_2$, since $\{b_1\}$ and $\{b_2\}$ are closed in Y and f is continuous, the sets $f^{-1}(b_1)$ and $f^{-1}(b_2)$ are closed in X, thus compact, and by the fact stated above, there exists some disjoint open sets U_1 and U_2 such that $f^{-1}(b_1) \subseteq U_1$ and $f^{-1}(b_2) \subseteq U_2$. Since f is closed, the sets $f(X - U_1)$ and $f(X - U_2)$

are closed, and thus the sets

$$\begin{aligned} V_1 &= Y - f(X - U_1) \\ V_2 &= Y - f(X - U_2) \end{aligned}$$

are open, and it is immediately verified that $V_1 \cap V_2 = \emptyset$, $b_1 \in V_1$, and $b_2 \in V_2$. This proves that Y is Hausdorff. $\square$

Under the hypotheses of Proposition 12.31, it is easy to show that Y is Hausdorff iff the set

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Another simple criterion uses continuous open maps. The following proposition is proved in Massey [78] (Appendix A, Proposition 5.3).

Proposition 12.32. *Let $f\colon X \to Y$ be a surjective continuous map between topological spaces. If f is an open map, then Y is Hausdorff iff the set*

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Note that the hypothesis of Proposition 12.32 implies that Y has the quotient topology determined by f.

The following special case of Proposition 12.32 is discussed in Tu [112] (Section 7.5, Theorem 7.8). Given a topological space X and an equivalence relation R on X, we say that R is *open* if the projection map $\pi\colon X \to X/R$ is an open map, where X/R is equipped with the quotient topology. Then, if R is an open equivalence relation on X, the topological space X/R is Hausdorff iff R is closed in $X \times X$.

The following proposition, also from Tu [112] (Section 7.5, Theorem 7.9), yields a sufficient condition for second-countability.

Proposition 12.33. *If X is a topological space and R is an open equivalence relation on X, then for any basis $\{B_\alpha\}$ for the topology of X, the family $\{\pi(B_\alpha)\}$ is a basis for the topology of X/R, where $\pi\colon X \to X/R$ is the projection map. Consequently, if X is second-countable, then so is X/R.*

Examples of quotient spaces, such as the Grassmannian and Stiefel manifolds, are discussed in Chapter 4, since their definitions require the notion of a group acting on a set.

12.6 Problems

Problem 12.1. A topological space E is said to be *trivial* if the only open sets are E and $\emptyset$. Construct a nontrivial topological space $(E, \mathcal{O})$ which does not satisfy the Hausdorff separation axiom.

Problem 12.2. Prove Proposition 12.1.

Problem 12.3. Prove Proposition 12.2.

Problem 12.4. Let $E_i = \mathbb{R}$, with the standard topology induced by $|x - y|$. Show that the product topology on $\mathbb{R}^n$ is the standard topology induced by the Euclidean norm.

Problem 12.5. Prove Proposition 12.3.

Problem 12.6. Prove Proposition 12.4.

Problem 12.7. Let $E = E_1 \times \cdots \times E_n$ be the set theoretical product of topological spaces and let $\pi_i \colon E_1 \times \cdots \times E_n \to E_i$ be the projection function $\pi_i(x_1, \ldots, x_n) = x_i$. Show that the product topology on E is the smallest topology, (i.e. the topology with the least amount of open sets), which ensures that each π_i is continuous.

Problem 12.8. Prove Propositions 12.5 and 12.6.

Problem 12.9. Prove Propositions 12.8 and 12.9.

Problem 12.10. Prove Proposition 12.11.

Problem 12.11.

(i) Prove Lemmas 12.12 and 12.13.

(ii) Given a topological space $(E, \mathcal{O})$, show that "a and b are connected in E" is an equivalence relation.

(iii) Prove Proposition 12.14.

Problem 12.12. Prove Proposition 12.15.

Problem 12.13. (Advanced) Prove Theorem 12.17.

Problem 12.14. Let $\{X_i\}_{i=1}^n$ be a family of topological spaces. Show that $X_1 \times \cdots \times X_n$ is connected if and only if each X_i is connected.

Problem 12.15. Prove Proposition 12.19.

Problem 12.16.

(i) Prove Propositions 12.20 and 12.21.

(ii) Construct a nontrivial Hausdorff space E with a compact subset $A \subset E$ which is not closed.

(iii) Prove Proposition 12.22.

Problem 12.17.

(i) Prove that if E is a normed vector space of finite dimension, then a A subset of E is compact iff it is closed and bounded.

(ii) (Advanced) Find an infinite dimensional normed vector space E and a subset $A \subseteq E$, such that A is closed and bounded, yet not compact.

Problem 12.18. Prove Proposition 12.24.

Problem 12.19. Prove that any finite dimensional normed vector space is second-countable.

Problem 12.20.

(i) Prove Proposition 12.27.

(ii) Prove Proposition 12.27 when E is a metric space.

Problem 12.21.

(i) Prove Proposition 12.28.

(ii) (Advanced) Prove Proposition 12.28 when E is a metric space.

Problem 12.22. Prove the Bolzano-Weierstrass property, namely that an infinite subset of a compact space has a limit point.

Problem 12.23. Let X be a Hausdorff topological space. Construct an example of a quotient space X/R which is not Hausdorff.

Problem 12.24. Prove that under the hypotheses of Proposition 12.31, Y is Hausdorff iff the set

$$\{(x_1, x_2) \in X \times X \mid f(x_1) = f(x_2)\}$$

is closed in $X \times X$.

Problem 12.25. Prove Proposition 12.33.

Part II

Riemannian Geometry, Lie Groups, Homogeneous Spaces

Chapter 13

Riemannian Metrics, Riemannian Manifolds

Fortunately, the rich theory of vector spaces endowed with a Euclidean inner product can, to a great extent, be lifted to the tangent bundle of a manifold. The idea is to equip the tangent space T_pM at p to the manifold M with an inner product $\langle -, - \rangle_p$, in such a way that these inner products vary smoothly as p varies on M. It is then possible to define the length of a curve segment on a M and to define the distance between two points on M.

In Section 13.1, we define the notion of local (and global) frame. Using frames, we obtain a criterion for the tangent bundle TM of a smooth manifold M to be trivial (that is, isomorphic to $M \times \mathbb{R}^n$).

Riemannian metrics and Riemannian manifolds are defined in Section 13.2, where several examples are given. The generalization of the notion of the gradient of a function defined on a smooth manifold requires a metric. We define the gradient of a function on a Riemannian manifold. We conclude by defining local isometries, isometries, and the isometry group $\mathrm{Isom}(M, g)$ of a Riemannian manifold (M, g).

13.1 Frames

Definition 13.1. Let M be an n-dimensional smooth manifold. For any open subset $U \subseteq M$, an n-tuple of vector fields $(X_1, \ldots, X_n)$ over U is called a *frame over U* iff $(X_1(p), \ldots, X_n(p))$ is a basis of the tangent space T_pM, for every $p \in U$. If $U = M$, then the X_i are global sections and $(X_1, \ldots, X_n)$ is called a *frame* (of M).

The notion of a frame is due to Élie Cartan who (after Darboux) made extensive use of them under the name of *moving frame* (and the *moving frame method*). Cartan's terminology is intuitively clear. As a point p moves in U, the frame $(X_1(p), \ldots, X_n(p))$ moves from fibre to fibre. Physicists refer to a frame as a choice of *local gauge*.

If $\dim(M) = n$, then for every chart (U, φ), since $d\varphi^{-1}_{\varphi(p)} \colon \mathbb{R}^n \to T_pM$ is a bijection for every $p \in U$, the n-tuple of vector fields $(X_1, \ldots, X_n)$, with $X_i(p) = d\varphi^{-1}_{\varphi(p)}(e_i)$, is a frame of TM over U, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. See Figure 13.1.

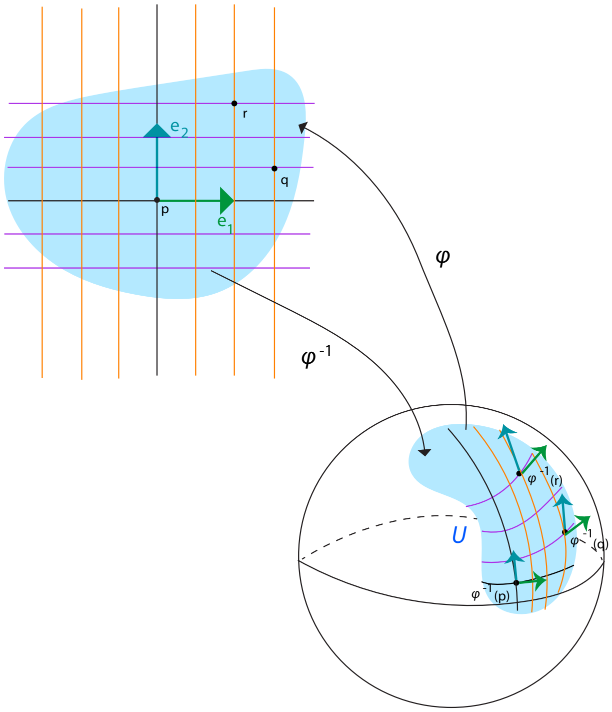


Figure 13.1: A frame on S^2.

The following proposition tells us when the tangent bundle is trivial (that is, isomorphic to the product $M \times \mathbb{R}^n$).

Proposition 13.1. *The tangent bundle TM of a smooth n-dimensional manifold M is trivial iff it possesses a frame of global sections (vector fields defined on M).*

As an illustration of Proposition 13.1 we can prove that the tangent bundle TS^1 of the circle is trivial. Indeed, we can find a section that is everywhere nonzero, *i.e.* a non-vanishing vector field, namely

$$X(\cos\theta, \sin\theta) = (-\sin\theta, \cos\theta).$$

The reader should try proving that TS^3 is also trivial (use the quaternions).

However, TS^2 is nontrivial, although this not so easy to prove. More generally, it can be shown that TS^n is nontrivial for all even $n \geq 2$. It can even be shown that S^1, S^3 and S^7 are the only spheres whose tangent bundle is trivial. This is a deep theorem and its proof is hard.

Remark: A manifold M such that its tangent bundle TM is trivial is called *parallelizable*.

We now define Riemannian metrics and Riemannian manifolds.

13.2 Riemannian Metrics

Definition 13.2. Given a smooth n-dimensional manifold M, a *Riemannian metric on M (or TM)* is a family $(\langle -, - \rangle_p)_{p\in M}$ of inner products on each tangent space T_pM, such that $\langle -, - \rangle_p$ depends smoothly on p, which means that for every chart $\varphi_\alpha \colon U_\alpha \to \mathbb{R}^n$, for every frame $(X_1, \ldots, X_n)$ on U_α, the maps

$$p \mapsto \langle X_i(p), X_j(p)\rangle_p, \qquad p \in U_\alpha,\ 1 \leq i, j \leq n,$$

are smooth. A smooth manifold M with a Riemannian metric is called a *Riemannian manifold*.

If $\dim(M) = n$, then for every chart (U, φ), we have the frame $(X_1, \ldots, X_n)$ over U, with $X_i(p) = d\varphi^{-1}_{\varphi(p)}(e_i)$, where $(e_1, \ldots, e_n)$ is the canonical basis of $\mathbb{R}^n$. Since every vector field over U is a linear combination $\sum_{i=1}^n f_i X_i$, for some smooth functions $f_i \colon U \to \mathbb{R}$, the condition of Definition 13.2 is equivalent to the fact that the maps

$$p \mapsto \langle d\varphi^{-1}_{\varphi(p)}(e_i), d\varphi^{-1}_{\varphi(p)}(e_j)\rangle_p, \qquad p \in U,\ 1 \leq i, j \leq n,$$

are smooth. If we let $x = \varphi(p)$, the above condition says that the maps

$$x \mapsto \langle d\varphi_x^{-1}(e_i), d\varphi_x^{-1}(e_j)\rangle_{\varphi^{-1}(x)} = \left\langle \left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p \right\rangle, \quad x \in \varphi(U),\ 1 \leq i, j \leq n,$$

are smooth.

If M is a Riemannian manifold, the metric on TM is often denoted $g = (g_p)_{p\in M}$. In a chart, using local coordinates, we often use the notation $g = \sum_{ij} g_{ij} dx_i \otimes dx_j$, or simply $g = \sum_{ij} g_{ij} dx_i dx_j$, where

$$g_{ij}(p) = \left\langle \left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p \right\rangle_p.$$

For every $p \in U$, the matrix $(g_{ij}(p))$ is symmetric, positive definite.

The standard Euclidean metric on $\mathbb{R}^n$, namely

$$g = dx_1^2 + \cdots + dx_n^2,$$

makes $\mathbb{R}^n$ into a Riemannian manifold. Then every submanifold M of $\mathbb{R}^n$ inherits a metric by restricting the Euclidean metric to M.

For example, the sphere S^{n-1} inherits a metric that makes S^{n-1} into a Riemannian manifold. It is instructive to find the local expression of this metric for S^2 in spherical coordinates. We can parametrize the sphere S^2 in terms of two angles θ (the *colatitude*) and φ (the *longitude*) as follows:

$$\begin{aligned} x &= \sin\theta\cos\varphi \\ y &= \sin\theta\sin\varphi \\ z &= \cos\theta. \end{aligned}$$

See Figure 13.2.

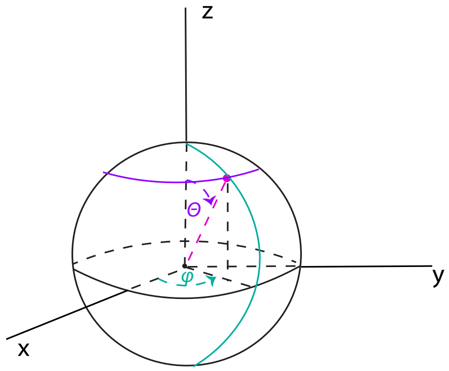


Figure 13.2: The spherical coordinates of S^2.

In order for the above to be a parametrization, we need to restrict its domain to $V = \{(\theta, \varphi) \mid 0 < \theta < \pi, 0 < \varphi < 2\pi\}$. Then the semicircle from the north pole to the south pole lying in the xz-plane is omitted from the sphere. In order to cover the whole sphere, we need another parametrization obained by choosing the axes in a suitable fashion; for example, to omit the semicircle in the xy-plane from $(0, 1, 0)$ to $(0, -1, 0)$ and with $x \leq 0$.

To compute the matrix giving the Riemannian metric in this chart, we need to compute a basis $(u(\theta, \varphi), v(\theta, \varphi))$ of the the tangent plane T_pS^2 at $p = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$. We can use

$$\begin{aligned} u(\theta, \varphi) &= \frac{\partial p}{\partial \theta} = (\cos\theta\cos\varphi, \cos\theta\sin\varphi, -\sin\theta) \\ v(\theta, \varphi) &= \frac{\partial p}{\partial \varphi} = (-\sin\theta\sin\varphi, \sin\theta\cos\varphi, 0), \end{aligned}$$

and we find that

$$\begin{aligned} \langle u(\theta, \varphi), u(\theta, \varphi)\rangle &= 1 \\ \langle u(\theta, \varphi), v(\theta, \varphi)\rangle &= 0 \\ \langle v(\theta, \varphi), v(\theta, \varphi)\rangle &= \sin^2\theta, \end{aligned}$$

so the metric on T_pS^2 w.r.t. the basis $(u(\theta,\varphi), v(\theta,\varphi))$ is given by the matrix

$$g_p = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}.$$

Thus, for any tangent vector

$$w = au(\theta,\varphi) + bv(\theta,\varphi), \quad a, b \in \mathbb{R},$$

we have

$$g_p(w,w) = a^2 + \sin^2\theta\, b^2.$$

A nontrivial example of a Riemannian manifold is the *Poincaré upper half-space*, namely, the set $H = \{(x,y) \in \mathbb{R}^2 \mid y > 0\}$ equipped with the metric

$$g = \frac{dx^2 + dy^2}{y^2}.$$

Consider the Lie group $\mathbf{SO}(n)$. We know from Section 7.2 that its tangent space at the identity $T_I\mathbf{SO}(n)$ is the vector space $\mathfrak{so}(n)$ of $n \times n$ skew symmetric matrices, and that the tangent space $T_Q\mathbf{SO}(n)$ to $\mathbf{SO}(n)$ at Q is isomorphic to

$$Q\mathfrak{so}(n) = \{QB \mid B \in \mathfrak{so}(n)\}.$$

(It is also isomorphic to $\mathfrak{so}(n)Q = \{BQ \mid B \in \mathfrak{so}(n)\}$.) If we give $\mathfrak{so}(n)$ the inner product

$$\langle B_1, B_2\rangle = \mathrm{tr}(B_1^\top B_2) = -\mathrm{tr}(B_1B_2),$$

the inner product on $T_Q\mathbf{SO}(n)$ is given by

$$\langle QB_1, QB_2\rangle = \mathrm{tr}((QB_1)^\top QB_2) = \mathrm{tr}(B_1^\top Q^\top QB_2) = \mathrm{tr}(B_1^\top B_2).$$

We will see in Chapter 15 that the length $L(\gamma)$ of the curve segment γ from I to e^B given by $t \mapsto e^{tB}$ (with $B \in \mathfrak{so}(n)$) is given by

$$L(\gamma) = \Big(\mathrm{tr}(-B^2)\Big)^{\frac{1}{2}}.$$

More generally, given any Lie group G, any inner product $\langle -,-\rangle$ on its Lie algebra $\mathfrak{g}$ induces by left translation an inner product $\langle -,-\rangle_g$ on T_gG for every $g \in G$, and this yields a Riemannian metric on G (which happens to be left-invariant; see Chapter 20).

Going back to the second example of Section 7.5, where we computed the differential df_R of the function $f\colon \mathbf{SO}(3) \to \mathbb{R}$ given by

$$f(R) = (u^\top Rv)^2, \qquad u, v \in \mathbb{R}^3$$

we found that

$$df_R(X) = 2u^\top X v u^\top R v, \quad X \in R\mathfrak{so}(3).$$

Since each tangent space $T_R\mathbf{SO}(3)$ is a Euclidean space under the inner product defined above, by duality, there is a unique vector $Y \in T_R\mathbf{SO}(3)$ defining the linear form df_R; that is,

$$\langle Y, X\rangle = df_R(X), \quad \text{for all } X \in T_R\mathbf{SO}(3).$$

By definition, the vector Y is the *gradient of f at R*, denoted $(\mathrm{grad}(f))_R$. The gradient of f at R is given by

$$(\mathrm{grad}(f))_R = u^\top R v R(R^\top u v^\top - v u^\top R)$$

since

$$\begin{aligned}
\langle (\mathrm{grad}(f))_R, X\rangle &= \mathrm{tr}((\mathrm{grad}(f))_R^\top X)\\
&= u^\top R v \; \mathrm{tr}((R^\top u v^\top - v u^\top R)^\top R^\top X)\\
&= u^\top R v \; \mathrm{tr}((v u^\top R - R^\top u v^\top) R^\top X)\\
&= u^\top R v(\mathrm{tr}(v u^\top X) - \mathrm{tr}(R^\top u v^\top R^\top X)), \qquad \text{since } RR^\top = I\\
&= u^\top R v(\mathrm{tr}(u^\top X v) - \mathrm{tr}(R^\top u v^\top R^\top X))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) - \mathrm{tr}(R^\top u v^\top R^\top R B)), \qquad X = RB \text{ with } B^\top = -B\\
&= u^\top R v(\mathrm{tr}(u^\top X v) - \mathrm{tr}(R^\top u v^\top B))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) - \mathrm{tr}((R^\top u v^\top B)^\top))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) + \mathrm{tr}(B v u^\top R))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) + \mathrm{tr}(v u^\top R B))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) + \mathrm{tr}(v u^\top X))\\
&= u^\top R v(\mathrm{tr}(u^\top X v) + \mathrm{tr}(u^\top X v))\\
&= 2u^\top X v u^\top R v, \qquad \text{since } u^\top X v \in \mathbb{R}\\
&= df_R(X).
\end{aligned}$$

More generally, the notion of gradient is defined as follows.

Definition 13.3. If $(M, \langle -, -\rangle)$ is a smooth manifold with a Riemannian metric and if $f\colon M \to \mathbb{R}$ is a smooth function on M, then the unique smooth vector field $\mathrm{grad}(f)$ defined such that

$$\langle (\mathrm{grad}(f))_p, u\rangle_p = df_p(u), \quad \text{for all } p \in M \text{ and all } u \in T_pM$$

is called the *gradient of f*.

It is usually complicated to find the gradient of a function.

If (U, φ) is a chart of M, with $p \in M$, and if

$$\left(\left(\frac{\partial}{\partial x_1}\right)_p, \ldots, \left(\frac{\partial}{\partial x_n}\right)_p\right)$$

denotes the basis of T_pM induced by φ, the local expression of the metric g at p is given by the $n \times n$ matrix $(g_{ij})_p$, with

$$(g_{ij})_p = g_p\left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_j}\right)_p\right).$$

The inverse is denoted by $(g^{ij})_p$. We often omit the subscript p and observe that for every function $f \in C^\infty(M)$,

$$\operatorname{grad} f = \sum_{ij} g^{ij} \frac{\partial f}{\partial x_j} \frac{\partial}{\partial x_i}.$$

A way to obtain a metric on a manifold N, is to pull-back the metric g on another manifold M along a local diffeomorphism $\varphi\colon N \to M$.

Definition 13.4. Recall that φ is a local diffeomorphism iff

$$d\varphi_p\colon T_pN \to T_{\varphi(p)}M$$

is a bijective linear map for every $p \in N$. Given any metric g on M, if φ is a local diffeomorphism, we define the *pull-back metric* φ^*g on N induced by g as follows. For all $p \in N$, for all $u, v \in T_pN$,

$$(\varphi^*g)_p(u, v) = g_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)).$$

We need to check that $(\varphi^*g)_p$ is an inner product, which is very easy since $d\varphi_p$ is a linear isomorphism.

The local diffeomorphism φ between the two Riemannian manifolds (N, φ^*g) and (M, g) has the special property that it is metric-preserving. Such maps are called local isometries, as defined below.

Definition 13.5. Given two Riemannian manifolds (M_1, g_1) and (M_2, g_2), a *local isometry* is a smooth map $\varphi\colon M_1 \to M_2$, such that $d\varphi_p\colon T_pM_1 \to T_{\varphi(p)}M_2$ is an isometry between the Euclidean spaces $(T_pM_1, (g_1)_p)$ and $(T_{\varphi(p)}M_2, (g_2)_{\varphi(p)})$, for every $p \in M_1$; that is,

$$(g_1)_p(u, v) = (g_2)_{\varphi(p)}(d\varphi_p(u), d\varphi_p(v)),$$

for all $u, v \in T_pM_1$, or equivalently, $\varphi^*g_2 = g_1$. Moreover, φ is an *isometry* iff it is a local isometry and a diffeomorphism.

An interesting example of the notion of isometry arises in machine learning, namely with respect to the *multinomial manifold*.

Example 13.1. Let Δ^n_+ be the standard open simplex

$$\Delta^n_+ = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1 + \cdots + x_{n+1} = 1,\ x_i > 0\}.$$

This is an open submanifold of the hyperplane of equation $x_1 + \cdots + x_{n+1} = 1$, which is itself a submanifold of $\mathbb{R}^{n+1}$. The manifold Δ^n_+ is diffeomorphic to the positive quadrant of the unit sphere in $\mathbb{R}^{n+1}$ given by

$$S^n_+ = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1,\ x_i > 0\}.$$

See Figure 13.3.

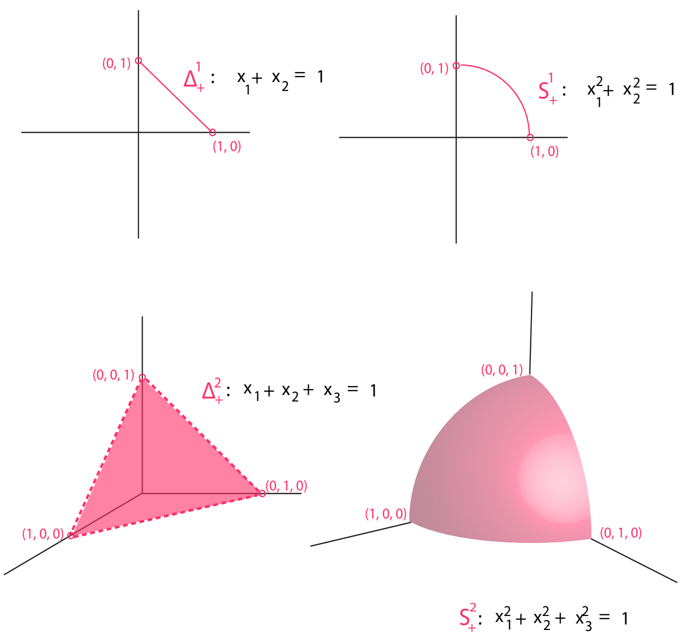

(0, 1)
Δ^1_+: $x_1 + x_2 = 1$
(1, 0)
(0, 1)
S^1_+: $x_1^2 + x_2^2 = 1$
(1, 0)
(0, 0, 1)
Δ^2_+: $x_1 + x_2 + x_3 = 1$
(0, 1, 0)
(1, 0, 0)
(0, 0, 1)
(1, 0, 0)
(0, 1, 0)
S^2_+: $x_1^2 + x_2^2 + x_3^2 = 1$

Figure 13.3: The open simplexes Δ^1_+ and Δ^2_+ along with the diffeomorphic S^1_+ and S^2_+.

The maps $\varphi\colon S^n_+ \to \Delta^n_+$ and $\psi\colon \Delta^n_+ \to S^n_+$ given by

$$\begin{aligned}\varphi(x_1 \ldots, x_{n+1}) &= (x_1^2, \ldots, x_{n+1}^2)\\ \psi(x_1 \ldots, x_{n+1}) &= (\sqrt{x_1}, \ldots, \sqrt{x_{n+1}})\end{aligned}$$

are clearly inverse diffeomorphisms. The map $\varphi\colon S^n_+ \to \Delta^n_+$ is often called the *real moment map*. For any $x \in S^n_+$, the tangent space $T_xS^n_+$ is given by

$$T_xS^n_+ = \{u \in \mathbb{R}^{n+1} \mid \langle x, u\rangle = 0\} = \{u \in \mathbb{R}^{n+1} \mid x_1u_1 + \cdots + x_{n+1}u_{n+1} = 0\},$$

where $\langle -, - \rangle$ is the standard Euclidean inner product in $\mathbb{R}^{n+1}$, and for any $x \in \Delta^n_+$, the tangent space $T_x\Delta^n_+$ is given by

$$T_x\Delta^n_+ = \{u \in \mathbb{R}^{n+1} \mid u_1 + \cdots + u_{n+1} = 0\}.$$

It is easily verified that the derivative $d\varphi_x$ of φ at $x \in S^n_+$ is given by

$$d\varphi_x(u_1, \ldots, u_{n+1}) = 2(x_1u_1, \ldots, x_{n+1}u_{n+1}).$$

As a consequence, if we give Δ^n_+ the Riemannian metric defined by

$$\langle u, v\rangle^F_x = \frac{1}{4}\sum_{i=1}^{n+1}\frac{u_iv_i}{x_i}, \quad x \in \Delta^n_+,$$

then we have

$$\begin{aligned}\langle d\varphi_x(u), d\varphi_x(v)\rangle^F_{\varphi(x)} &= \langle 2(x_1u_1, \ldots, x_{n+1}u_{n+1}), 2(x_1v_1, \ldots, x_{n+1}v_{n+1})\rangle^F_{(x_1^2,\ldots,x_{n+1}^2)} \\ &= \frac{1}{4}\sum_{i=1}^{n+1}\frac{2x_iu_i2x_iv_i}{x_i^2} \\ &= \sum_{i=1}^{n+1} u_iv_i = \langle u, v\rangle.\end{aligned}$$

Therefore, φ is an isometry between the Riemannian manifold $(S^n_+, \langle -, -\rangle)$ (equipped with the restriction of the Euclidean metric of $\mathbb{R}^{n+1}$) to the manifold $(\Delta^n_+, \langle -, -\rangle^F)$ equipped with the metric

$$\langle u, v\rangle^F_x = \frac{1}{4}\sum_{i=1}^{n+1}\frac{u_iv_i}{x_i} = \frac{1}{4}\sum_{i=1}^{n+1} x_i\frac{u_i}{x_i}\,\frac{v_i}{x_i} = \frac{1}{4}\sum_{i=1}^{n+1} x_i\frac{d(\log x_i)}{dx_i}\,\frac{d(\log x_i)}{dx_i}u_iv_i, \quad x \in \Delta^n_+,$$

known as the *Fisher information metric* (actually, one fourth of the Fisher information metric). The above shows that the Fisher information metric is the pullback of the Euclidean metric on S^n_+ along the inverse ψ of the real moment map φ. In machine learning the manifold $(\Delta^n_+, \langle -, -\rangle^F)$ is called the *multinomial manifold*. Unfortunately, it is often denoted by $\mathbb{P}^n$, which clashes with the standard notation for projective space.

The isometries of a Riemannian manifold (M, g) form a group $\mathrm{Isom}(M, g)$, called the *isometry group of* (M, g). An important theorem of Myers and Steenrod asserts that the isometry group $\mathrm{Isom}(M, g)$ is a Lie group.

Given a map $\varphi\colon M_1 \to M_2$ and a metric g_1 on M_1, in general, φ does not induce any metric on M_2. However, if φ has some extra properties, it does induce a metric on M_2. This is the case when M_2 arises from M_1 as a quotient induced by some group of isometries of M_1. For more on this, see Gallot, Hulin and Lafontaine [49] (Chapter 2, Section 2.A), and Chapter 22.

Because a manifold is *paracompact* (see Section 10.1), a Riemannian metric always exists on M. This is a consequence of the existence of partitions of unity (see Theorem 10.5).

Theorem 13.2. *Every smooth manifold admits a Riemannian metric.*

Theorem 13.2 is proved in Gallot, Hulin, Lafontaine [49] (Chapter 2, Theorem 2.2), using a partition of unity.

Except in special simple cases (vector spaces, the spheres S^d) it is hard to define *explicitly* Riemannian metrics on a manifold. However, there are two important classes of manifolds for which the problem of defining metrics (with some natural properties) basically reduces to simple linear algebra:

(1) Lie groups.

(2) Reductive homogeneous spaces.

Metrics on Lie groups are investigated in Chapter 20, and metrics on reductive homogeneous spaces are investigated in Chapter 22.

13.3 Problems

Problem 13.1. Prove that the tangent bundle TS^3 of the sphere $S^3 \subseteq \mathbb{R}^4$ is a trivial bundle.

Hint. Consider the matrix

$$\begin{pmatrix} a & -b & -c & -d \\ b & a & -d & c \\ c & d & a & -b \\ d & -c & b & a \end{pmatrix}.$$

Problem 13.2. Let (M_1, g_1) and (M_2, g_2) be two Riemannian manifolds. We know that the product $M_1 \times M_2$ can be made into a manifold (see Example 7.5), and let $\pi_1 \colon M_1 \times M_2 \to M_1$ and $\pi_2 \colon M_1 \times M_2 \to M_2$ be the natural projections. Define a Riemannian metric g on $M_1 \times M_2$ called the *product metric* on $M_1 \times M_2$ as follows: for all $(p, q) \in M_1 \times M_2$ and all $(u, v) \in T_{(p,q)}(M_1 \times M_2)$, we have

$$g_{(p,q)}(u, v) = (g_1)_p(d\pi_1(u), d\pi_1(v)) + (g_2)_q(d\pi_2(u), d\pi_2(v)).$$

Check that g is indeed a Riemannian metric on $M_1 \times M_2$.

In the case of the torus $T^n = S^1 \times \cdots \times S^1$ with the metric on the circle $S^1 \subseteq \mathbb{R}^2$ induced by the metric on $\mathbb{R}^2$, we say that (T^n, g) is a *flat torus*.

Problem 13.3. Determine the metric on the patch on the sphere S^2 that omits the semicircle in the xy-plane.

Problem 13.4. Consider the surjective map $\pi \colon \mathbb{R}^n \to T^n$ given by

$$\pi(x_1, \ldots, x_n) = (e^{ix_1}, \ldots, e^{ix_n}),$$

viewing S^1 as $S^1 = \{e^{ix} \mid 0 \leq x < 2\pi\}$ and with $T^n = S^1 \times \cdots \times S^1$.

(1) Find a Riemannian metric on T^n such that π becomes a local isometry, and prove that with this metric T^n is isomorphic to the flat torus of Problem 13.2.

(2) A *lattice* Γ in $\mathbb{R}^n$ is a set of vectors

$$\Gamma = \{k_1 u_1 + \cdots + k_n u_n \mid k_1, \ldots, k_n \in \mathbb{Z}\}$$

where $(u_1, \ldots, u_n)$ is a basis of $\mathbb{R}^n$. Since Γ is an abelian subgroup of $\mathbb{R}^n$, the quotient space (group) $\mathbb{R}^n/\Gamma$ is well defined.

Define the map $p\colon \mathbb{R}^n \to T^n$ by

$$p\,(x_1 u_1 + \cdots + x_n u_n) = (e^{2i\pi x_1}, \ldots, e^{2i\pi x_n}).$$

Prove that p is surjective and constant on Γ, so that p induces a continuous bijective map $\widehat{p}\colon \mathbb{R}^n/\Gamma \to T^n$. Prove that $\widehat{h}$ is a homeomorphism.

Remark: The space $\mathbb{R}^n/\Gamma$ can be equipped with a metric g_1 so that the projection map $\pi\colon \mathbb{R}^n \to \mathbb{R}^n/\Gamma$ is a Riemannian covering map. Then, we obtain a metric g_2 on T^n that makes T^n into a flat torus, and $\widehat{p}$ is an isometry between $\mathbb{R}^n/\Gamma$ and a flat torus; see Gallot, Hulin, Lafontaine [49] (Chapter 2, Example 2.22).

(3) Show that the map $F\colon T^2 \to \mathbb{R}^4$ given by

$$F(\theta_1, \theta_2) = \frac{1}{2}(\cos\theta_1, \sin\theta_1, \cos\theta_2, \sin\theta_2)$$

is an injective immersion and a local isometry.

Problem 13.5. Consider the polar coordinate system $(r, \theta) \in \mathbb{R}^+ \times (-\pi, +\pi)$ on $\mathbb{R}^2 - \{(x, 0) \mid x \leq 0\}$, with $x = r\cos\theta$ and $y = r\sin\theta$.

Prove that restriction of the Euclidean metric to $\mathbb{R}^2 - \{(x, 0) \mid x \leq 0\}$ is given by

$$g = (dr)^2 + r^2(d\theta)^2.$$

Problem 13.6. Let $\gamma\colon (a, b) \to \mathbb{R}^2$ be a regular injective smooth curve given parametrically by

$$\gamma(t) = (r(t), z(t)),$$

where $r(t) > 0$ and $\gamma'(t) \neq 0$ for all t such that $a < t < b$. By rotating this curve around the z-axis we get a cylindrical surface S that can be represented parametrically as

$$(t, \theta) \mapsto (S(t, \theta) = (r(t)\cos\theta, r(t)\sin\theta, z(t)).$$

Show that the metric on S induced by the Euclidean metric on $\mathbb{R}^3$ is given by

$$g = \left(\left(\frac{dr}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2\right)(dt)^2 + r^2(d\theta)^2.$$

Show that if the curve is parametrized by arc length, then

$$g = (dt)^2 + r^2(d\theta)^2.$$

Problem 13.7. The *hyperbolic space* $\mathcal{H}_n^+(1)$ (see Definition 5.3) is defined in terms of the *Lorentz innner product* $\langle -, - \rangle_1$ on $\mathbb{R}^{n+1}$, given by

$$\langle (x_1, \ldots, x_{n+1}), (y_1, \ldots, y_{n+1}) \rangle_1 = -x_1 y_1 + \sum_{i=2}^{n+1} x_i y_i.$$

By definition, $\mathcal{H}_n^+(1)$, written simply H^n, is given by

$$H^n = \{x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_1 = -1,\ x_1 > 0\}.$$

Given any point $p = (x_1, \ldots, x_{n+1}) \in H^n$, show that the restriction of $\langle -, - \rangle_1$ to $T_p H^n$ is positive, definite, which means that it is a metric on $T_p H^n$.

Hint. See Section 16.2.

Problem 13.8. There are other isometric models of H^n that are perhaps intuitively easier to grasp but for which the metric is more complicated. There is a map $\mathrm{PD}\colon B^n \to H^n$ where $B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ is the open unit ball in $\mathbb{R}^n$, given by

$$\mathrm{PD}(x) = \left(\frac{1 + \|x\|^2}{1 - \|x\|^2}, \frac{2x}{1 - \|x\|^2} \right).$$

(1) Check that $\langle \mathrm{PD}(x), \mathrm{PD}(x) \rangle_1 = -1$ and that PD is bijective and an isometry. Prove that the pull-back metric $g_{\mathrm{PD}} = \mathrm{PD}^* g_H$ on B^n is given by

$$g_{\mathrm{PD}} = \frac{4}{(1 - \|x\|^2)^2} \, (dx_1^2 + \cdots + dx_n^2).$$

The metric g_{PD} is called the *conformal disc metric*, and the Riemannian manifold (B^n, g_{PD}) is called the *Poincaré disc model* or *conformal disc model*. The metric g_{PD} is proportional to the Euclidean metric, and thus angles are preserved under the map PD.

(2) Another model is the *Poincaré half-plane model* $\{x \in \mathbb{R}^n \mid x_1 > 0\}$, with the metric

$$g_{\mathrm{PH}} = \frac{1}{x_1^2} \, (dx_1^2 + \cdots + dx_n^2).$$

Let h be the inversion of $\mathbb{R}^n$ with center $t = (-1, 0, \ldots, 0)$ given by

$$h(x) = t + \frac{2(x - t)}{\|x - t\|^2}.$$

Prove that h is a diffeomorphism onto the half space $\{x \in \mathbb{R}^n \mid x_1 > 0\}$.

(3) Prove that

$$h^* g_{\mathrm{PD}} = g_{\mathrm{PH}} = \frac{1}{x_1^2} \, (dx_1^2 + \cdots + dx_n^2).$$

Problem 13.9. Prove Theorem 13.2 using a partition of unity argument.

Chapter 14

Connections on Manifolds

Given a manifold M, in general, for any two points $p, q \in M$, there is no "natural" isomorphism between the tangent spaces T_pM and T_qM. Given a curve $c\colon [0,1] \to M$ on M, as $c(t)$ moves on M, how does the tangent space $T_{c(t)}M$ change as $c(t)$ moves?

If $M = \mathbb{R}^n$, then the spaces $T_{c(t)}\mathbb{R}^n$ are canonically isomorphic to $\mathbb{R}^n$, and any vector $v \in T_{c(0)}\mathbb{R}^n \cong \mathbb{R}^n$ is simply moved along c by *parallel transport*; that is, at $c(t)$, the tangent vector v also belongs to $T_{c(t)}\mathbb{R}^n$. However, if M is curved, for example a sphere, then it is not obvious how to "parallel transport" a tangent vector at $c(0)$ along a curve c. A way to achieve this is to define the notion of *parallel vector field* along a curve, and this can be defined in terms of the notion of *covariant derivative* of a vector field.

In Section 14.1, we define the general notion of a *connection* on a manifold M as a function $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ defined on vector fields and satisfying some properties that make it a generalization of the notion of covariant derivative on a surface. We show that $(\nabla_X Y)(p)$ only depends on the value of X at p and on the value of Y in a neighborhood of p.

In Section 14.2, we show that the notion of covariant derivative is well-defined for vector fields along a curve. Given a vector field X along a curve γ, this covariant derivative is denoted by DX/dt. We then define the crucial notion of a vector field *parallel along a curve* γ, which means that $DX/dt(s) = 0$ for all s (in the domain of γ). As a consequence, we can define the notion of *parallel transport* of a vector along a curve.

The notion of a connection on a manifold *does not* assume that the manifold is equipped with a Riemannian metric. In Section 14.3, we consider connections having additional properties, such as being compatible with a Riemannian metric or being torsion-free. Then we have a phenomenon called by some people the "miracle" of Riemannian geometry, namely that for every Riemannian manifold, there is a *unique* connection which is torsion-free and compatible with the metric. Furthermore, this connection is determined by an implicit formula known as the *Koszul formula.* Such a connection is called the *Levi-Civita connection.* We conclude this section with some properties of connections compatible with a metric, in particular about parallel vectors fields along a curve.

14.1 Connections on Manifolds

Given any two vector fields X and Y defined on some open subset $U \subseteq \mathbb{R}^3$, for every $p \in U$, the *directional derivative $D_X Y(p)$ of Y with respect to X* is defined by

$$D_X Y(p) = \lim_{t \to 0} \frac{Y(p + tX(p)) - Y(p)}{t}.$$

See Figure 14.1.

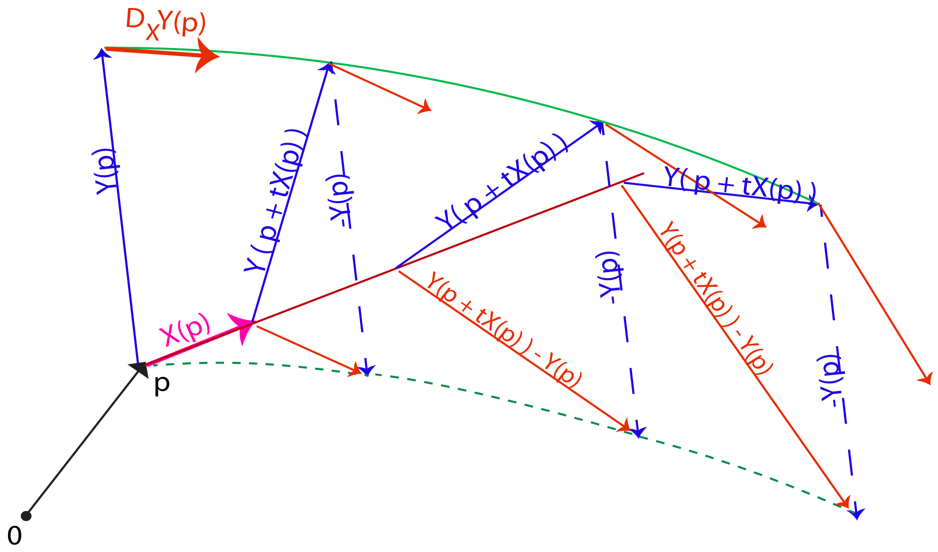


Figure 14.1: The directional derivative of the blue vector field $Y(p)$ in the direction of X.

Observe that the above is the directional derivative of the function $p \mapsto Y(p)$ as given in Definition 11.4, except that the direction vector $X(p)$ varies with p.

If $f \colon U \to \mathbb{R}$ is a differentiable function on U, for every $p \in U$, the *directional derivative $X[f](p)$ (or $X(f)(p)$) of f with respect to X* is defined by

$$X[f](p) = \lim_{t \to 0} \frac{f(p + tX(p)) - f(p)}{t}.$$

Again, this is Definition 11.4, except that the direction vector $X(p)$ varies with p. We know that $X[f](p) = df_p(X(p))$.

It is easily shown that $D_X Y(p)$ is $\mathbb{R}$-bilinear in X and Y, is $C^\infty(U)$-linear in X, and satisfies the Leibniz derivation rule with respect to Y; that is:

Proposition 14.1. *If X and Y are vector fields from U to $\mathbb{R}^3$ that are differentiable on some open subset U of $\mathbb{R}^3$, then their directional derivatives satisfy the following properties:*

$$\begin{aligned} D_{X_1+X_2}Y(p) &= D_{X_1}Y(p) + D_{X_2}Y(p) \\ D_{fX}Y(p) &= f(p)D_XY(p) \\ D_X(Y_1+Y_2)(p) &= D_XY_1(p) + D_XY_2(p) \\ D_X(fY)(p) &= X[f](p)Y(p) + f(p)D_XY(p), \end{aligned}$$

for all $X, X_1, X_2, Y, Y_1, Y_2 \in \mathfrak{X}(U)$ and all $f \in C^\infty(U)$.

Proof. By definition we have

$$\begin{aligned} D_X(Y_1+Y_2)(p) &= \lim_{t\to 0} \frac{(Y_1+Y_2)(p+tX(p)) - (Y_1+Y_2)(p)}{t} \\ &= \lim_{t\to 0} \frac{Y_1(p+tX(p)) - Y_1(p)}{t} + \lim_{t\to 0} \frac{Y_2(p+tX(p)) - Y_2(p)}{t} \\ &= D_XY_1(p) + D_XY_2(p). \end{aligned}$$

Since $Y\colon U \to \mathbb{R}^3$ is assumed to be differentiable, by Proposition 11.15, we have $D_XY(p) = dY_p(X(p))$, so by linearity of dY_p, we have

$$D_{X_1+X_2}Y(p) = dY_p(X_1(p) + X_2(p)) = dY_p(X_1(p)) + dY_p(X_2(p)) = D_{X_1}Y(p) + D_{X_2}Y(p).$$

The definition also implies

$$\begin{aligned} D_X(fY)(p) &= \lim_{t\to 0} \frac{fY(p+tX(p)) - fY(p)}{t} \\ &= \lim_{t\to 0} \frac{f(p+tX(p))Y(p+tX(p)) - f(p)Y(p)}{t} \\ &= \lim_{t\to 0} \frac{f(p+tX(p))Y(p+tX(p)) - f(p)Y(p+tX(p))}{t} \\ &\qquad + \lim_{t\to 0} \frac{f(p)Y(p+tX(p)) - f(p)Y(p)}{t} \\ &= X[f](p)Y(p) + f(p)D_XY(p). \end{aligned}$$

It remains to prove $D_{fX}Y(p) = f(p)D_XY(p)$. If $f(p) = 0$, this trivially true. So assume $f(p) \neq 0$. Then

$$\begin{aligned} D_{fX}Y(p) = f(p) \lim_{t\to 0} \frac{Y(p+tfX(p)) - Y(p)}{tf(p)} &= f(p) \lim_{t\to 0} \frac{Y(p+tf(p)X(p)) - Y(p)}{tf(p)} \\ = f(p) \lim_{u\to 0} \frac{Y(p+uX(p)) - Y(p)}{u} &= f(p)D_XY(p), \end{aligned}$$

as claimed. □

Now assume that M is a surface in $\mathbb{R}^3$. If X and Y are two vector fields defined on some open subset $U \subseteq \mathbb{R}^3$, and if there is some open subset $W \subseteq M$ of the surface M such that $X(p), Y(p) \in T_pM$ for all $p \in W$, for every $p \in W$, the directional derivative $D_XY(p)$ makes sense and it has an orthogonal decomposition

$$D_XY(p) = \nabla_XY(p) + (D_n)_XY(p),$$

where its *horizontal (or tangential) component* is $\nabla_XY(p) \in T_pM$, and its normal component is $(D_n)_XY(p)$. See Figure 14.2.

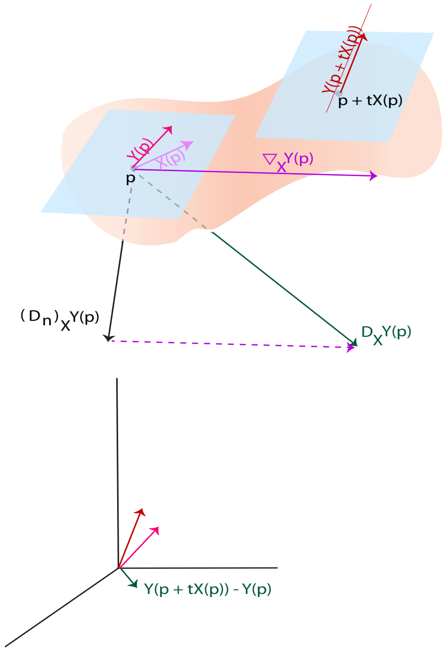


Figure 14.2: The orthogonal decomposition of $D_XY(p)$ for the peach surface M.

The component $\nabla_XY(p)$ is the *covariant derivative* of Y with respect to $X \in T_pM$, and it allows us to define the covariant derivative of a vector field $Y \in \mathfrak{X}(U)$ with respect to a vector field $X \in \mathfrak{X}(M)$ on M. We easily check that ∇_XY satisfies the four equations of Proposition 14.1.

In particular, Y may be a vector field associated with a curve $c\colon [0,1] \to M$. A *vector field along a curve* c is a vector field Y such that $Y(c(t)) \in T_{c(t)}M$, for all $t \in [0,1]$. We also write $Y(t)$ for $Y(c(t))$. Then we say that Y *is parallel along* c iff $\nabla_{c'(t)}Y = 0$ along c.

The notion of *parallel transport* on a surface can be defined using parallel vector fields along curves. Let p, q be any two points on the surface M, and assume there is a curve $c\colon [0,1] \to M$ joining $p = c(0)$ to $q = c(1)$. Then using the uniqueness and existence theorem for ordinary differential equations, it can be shown that for any initial tangent vector $Y_0 \in T_pM$, there is a unique parallel vector field Y along c, with $Y(0) = Y_0$. If we set $Y_1 = Y(1)$, we obtain a linear map $Y_0 \mapsto Y_1$ from T_pM to T_qM which is also an isometry.

As a summary, given a surface M, if we can define a notion of covariant derivative $\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$ satisfying the properties of Proposition 14.1, then we can define the notion of parallel vector field along a curve, and the notion of parallel transport, which yields a natural way of relating two tangent spaces T_pM and T_qM, using curves joining p and q.

This can be generalized to manifolds using the notion of connection. We will see that the notion of connection induces the notion of curvature. Moreover, if M has a Riemannian metric, we will see that this metric induces a unique connection with two extra properties (the *Levi-Civita* connection).

Definition 14.1. Let M be a smooth manifold. A *connection* on M is a $\mathbb{R}$-bilinear map

$$\nabla\colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M),$$

where we write $\nabla_X Y$ for $\nabla(X, Y)$, such that the following two conditions hold:

$$\begin{aligned} \nabla_{fX} Y &= f\nabla_X Y \\ \nabla_X(fY) &= X[f]Y + f\nabla_X Y, \end{aligned}$$

for all $X, Y \in \mathfrak{X}(M)$ and all $f \in C^\infty(M)$. The vector field $\nabla_X Y$ is called the *covariant derivative of Y with respect to X*.

A connection on M is also known as an *affine connection* on M. The following proposition gives the first of two basic property of ∇.

Proposition 14.2. *Let M be a smooth manifold and let ∇ be a connection on M. For every open subset $U \subseteq M$, for every vector field $Y \in \mathfrak{X}(M)$, if $Y \equiv 0$ on U, then $\nabla_X Y \equiv 0$ on U for all $X \in \mathfrak{X}(M)$.*

The property of ∇ stated in Proposition 14.2 is characteristic of a *local operator*.

Proposition 14.2 is proved in Milnor and Stasheff [85] (Appendix C). Proposition 14.2 implies that a connection ∇ on M restricts to a connection $\nabla \upharpoonright U$ on every open subset $U \subseteq M$.

The second basic property of ∇ is that $(\nabla_X Y)(p)$ only depends on $X(p)$.

Proposition 14.3. *For any two vector fields $X, Y \in \mathfrak{X}(M)$, if $X(p) = Y(p)$ for some $p \in M$, then*

$$(\nabla_X Z)(p) = (\nabla_Y Z)(p) \qquad \textit{for every } Z \in \mathfrak{X}(M).$$

A proof of Proposition 14.3 is given in O'Neil [91] (Chapter 2, Lemma 3). This proposition is extremely useful. Although the definition of $(\nabla_X Y)(p)$ requires the vector fields X and Y to be *globally defined* on M, to compute $(\nabla_X Y)(p)$, it is enough to know $u = X(p)$.

Consequently, for any $p \in M$, the covariant derivative $(\nabla_u Y)(p)$ is well defined for any tangent vector $u \in T_pM$ and any vector field Y defined on some open subset $U \subseteq M$, with $p \in U$.

Observe that on U, the n-tuple of vector fields $\left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right)$ is a local frame.

Definition 14.2. We have

$$\nabla_{\frac{\partial}{\partial x_i}} \left(\frac{\partial}{\partial x_j}\right) = \sum_{k=1}^{n} \Gamma_{ij}^{k} \frac{\partial}{\partial x_k},$$

for some unique smooth functions Γ_{ij}^k defined on U, called the *Christoffel symbols*.

Definition 14.3. We say that a connection ∇ is *flat* on U iff

$$\nabla_X \left(\frac{\partial}{\partial x_i}\right) = 0, \quad \text{for all} \quad X \in \mathfrak{X}(U),\ 1 \leq i \leq n.$$

Proposition 14.4. *Every smooth manifold M possesses a connection.*

Proof. We can find a family of charts $(U_\alpha, \varphi_\alpha)$ such that $\{U_\alpha\}_\alpha$ is a locally finite open cover of M. If (f_α) is a partition of unity subordinate to the cover $\{U_\alpha\}_\alpha$ and if ∇^α is the flat connection on U_α, then it is immediately verified that

$$\nabla = \sum_{\alpha} f_\alpha \nabla^\alpha$$

is a connection on M. □

Remark: A connection on TM can be viewed as a linear map

$$\nabla\colon \mathfrak{X}(M) \longrightarrow \mathrm{Hom}_{C^\infty(M)}(\mathfrak{X}(M), \mathfrak{X}(M)),$$

such that, for any fixed $Y \in \mathfrak{X}(M)$, the map $\nabla Y\colon X \mapsto \nabla_X Y$ is $C^\infty(M)$-linear, which implies that ∇Y is a $(1,1)$ tensor.

As for Riemannian metrics, except in special simple cases (vector spaces, the spheres S^d) it is hard to define *explicitly* connections on a manifold. However, there are two important classes of manifolds for which the problem of defining connections (with some natural properties) basically reduces to simple linear algebra:

(1) Lie groups.

(2) Reductive homogeneous spaces.

Connections on Lie groups are investigated in Chapter 20, and connections on reductive homogeneous spaces are investigated in Chapter 22.

14.2 Parallel Transport

The notion of connection yields the notion of parallel transport. First, we need to define the covariant derivative of a vector field along a curve.

Definition 14.4. Let M be a smooth manifold and let $\gamma\colon [a,b] \to M$ be a smooth curve in M. A *smooth vector field along the curve* γ is a smooth map $X\colon [a,b] \to TM$, such that $\pi(X(t)) = \gamma(t)$, for all $t \in [a,b]$ $(X(t) \in T_{\gamma(t)}M)$. See Figure 14.3.

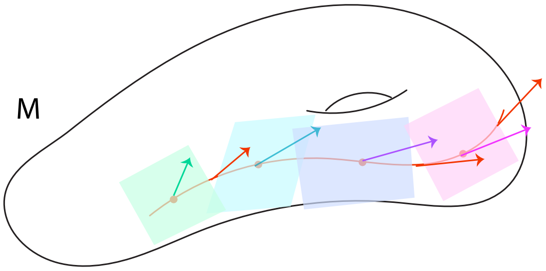


Figure 14.3: A smooth vector field along the orange curve γ.

Recall that the curve $\gamma\colon [a,b] \to M$ is smooth iff γ is the restriction to $[a,b]$ of a smooth curve on some open interval containing $[a,b]$.

Since a vector X field along a curve γ does not necessarily extend to an open subset of M (for example, if the image of γ is dense in M; see Problem 7.11), the covariant derivative $(\nabla_{\gamma'(t_0)} X)_{\gamma(t_0)}$ may *not be defined*, so we need a proposition showing that the covariant derivative of a vector field along a curve makes sense. Roughly, this is analogous to the difference between uniform continuity and continuity.

Proposition 14.5. *Let M be a smooth manifold, let ∇ be a connection on M and $\gamma\colon [a,b] \to M$ be a smooth curve in M. There is a $\mathbb{R}$-linear map D/dt, defined on the vector space of smooth vector fields X along γ, which satisfies the following conditions.*

(1) For any smooth function $f\colon [a,b] \to \mathbb{R}$,

$$\frac{D(fX)}{dt} = \frac{df}{dt} X + f\frac{DX}{dt}.$$

(2) If X is induced by a vector field $Z \in \mathfrak{X}(M)$, that is $X(t_0) = Z(\gamma(t_0))$ for all $t_0 \in [a,b]$, then $\dfrac{DX}{dt}(t_0) = (\nabla_{\gamma'(t_0)} Z)_{\gamma(t_0)}$.

Proof. Since $\gamma([a,b])$ is compact, it can be covered by a finite number of open subsets U_α, such that $(U_\alpha, \varphi_\alpha)$ is a chart. Thus, we may assume that $\gamma\colon [a,b] \to U$ for some chart (U, φ). As $\varphi \circ \gamma\colon [a,b] \to \mathbb{R}^n$, we can write

$$\varphi \circ \gamma(t) = (u_1(t), \ldots, u_n(t)),$$

where each $u_i = pr_i \circ \varphi \circ \gamma$ is smooth. By applying the chain rule it is easy to see that

$$\gamma'(t_0) = \sum_{i=1}^n \frac{du_i}{dt} \left(\frac{\partial}{\partial x_i}\right)_{\gamma(t_0)}.$$

If $(s_1, \ldots, s_n)$ is a frame over U, we can write

$$X(t) = \sum_{i=1}^n X_i(t) s_i(\gamma(t)),$$

for some smooth functions X_i. For every $t \in [a,b]$, each vector field s_j over U can be extended to a vector field on M whose restriction to some open subset containing $\gamma(t)$ agrees with s_j, so the $\mathbb{R}$-linearity of ∇, along with Conditions (1) and (2), imply that

$$\begin{aligned}\frac{DX}{dt} &= \nabla_{\gamma'(t)} X(t) = \nabla_{\gamma'(t)} \sum_{j=1}^n X_j(t) s_j(\gamma(t)) = \sum_{j=1}^n \nabla_{\gamma'(t)} \left(X_j(t) s_j(\gamma(t))\right) \\ &= \sum_{j=1}^n \left(\frac{dX_j}{dt} s_j(\gamma(t)) + X_j(t) \nabla_{\gamma'(t)}(s_j(\gamma(t)))\right).\end{aligned}$$

Since

$$\gamma'(t) = \sum_{i=1}^n \frac{du_i}{dt} \left(\frac{\partial}{\partial x_i}\right)_{\gamma(t)},$$

there exist some smooth functions Γ_{ij}^k (generally different from the Christoffel symbols) so that

$$\begin{aligned}\nabla_{\gamma'(t)}(s_j(\gamma(t))) &= \nabla_{\sum_{i=1}^n \frac{du_i}{dt}\left(\frac{\partial}{\partial x_i}\right)_{\gamma(t)}} (s_j(\gamma(t))) \\ &= \nabla_{\frac{du_1}{dt}\left(\frac{\partial}{\partial x_1}\right)_{\gamma(t)}} (s_j(\gamma(t))) + \cdots + \nabla_{\frac{du_n}{dt}\left(\frac{\partial}{\partial x_n}\right)_{\gamma(t)}} (s_j(\gamma(t))) \\ &= \frac{du_1}{dt} \nabla_{\left(\frac{\partial}{\partial x_1}\right)_{\gamma(t)}} (s_j(\gamma(t))) + \cdots + \frac{du_n}{dt} \nabla_{\left(\frac{\partial}{\partial x_n}\right)_{\gamma(t)}} (s_j(\gamma(t))) \\ &= \sum_{i=1}^n \frac{du_i}{dt} \nabla_{\frac{\partial}{\partial x_i}} (s_j(\gamma(t))) \\ &= \sum_{i,k} \frac{du_i}{dt} \Gamma_{ij}^k s_k(\gamma(t)).\end{aligned}$$

It follows that

$$\frac{DX}{dt} = \sum_{k=1}^{n} \left(\frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j \right) s_k(\gamma(t)).$$

Conversely, the above expression defines a linear operator D/dt, and it is easy to check that it satisfies Conditions (1) and (2). □

Definition 14.5. The operator D/dt is often called *covariant derivative along* γ and it is also denoted by $\nabla_{\gamma'(t)}$ or simply $\nabla_{\gamma'}$.

The use of the notation $\nabla_{\gamma'(t)}$ instead of D/dt is quite unfortunate, since D/dt is applied to a vector field only *locally defined* along a curve, whereas $\nabla_{\gamma'(t)}$ is applied to a vector field *globally defined* on M (or a least, on some open subset of M). This is another of these notational conventions that we have to live with.

Definition 14.6. Let M be a smooth manifold and let ∇ be a connection on M. For every curve $\gamma\colon [a,b] \to M$ in M, a vector field X along γ is *parallel (along γ)* iff

$$\frac{DX}{dt}(s) = 0 \quad \text{for all } s \in [a,b].$$

If M was embedded in $\mathbb{R}^d$ for some d, then to say that X is parallel along γ would mean that the directional derivative $(D_{\gamma'}X)(\gamma(t))$ is normal to $T_{\gamma(t)}M$. See Figure 14.4.

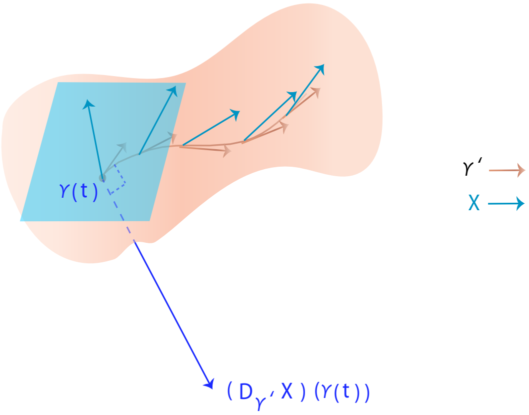


Figure 14.4: The real vector field X is parallel to the curve γ since $(D_{\gamma'}X)(\gamma(t))$ is perpendicular to the tangent plane $T_{\gamma(t)}M$.

The following proposition can be shown using the existence and uniqueness of solutions of ODE's (in our case, linear ODE's).

Proposition 14.6. *Let M be a smooth manifold and let ∇ be a connection on M. For every C^1 curve $\gamma\colon [a,b] \to M$ in M, for every $t \in [a,b]$ and every $v \in T_{\gamma(t)}M$, there is a unique parallel vector field X along γ such that $X(t) = v$.*

Proof. For the proof of Proposition 14.6, it is sufficient to consider the portions of the curve γ contained in some chart. In such a chart (U, φ), as in the proof of Proposition 14.5, using a local frame $(s_1, \ldots, s_n)$ over U, we have

$$\frac{DX}{dt} = \sum_{k=1}^{n} \left(\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^{k} \, \frac{du_i}{dt} \, X_j \right) s_k(\gamma(t)),$$

with $u_i = pr_i \circ \varphi \circ \gamma$. Consequently, X is parallel along our portion of γ iff the system of linear ODE's in the unknowns X_k,

$$\frac{dX_k}{dt} + \sum_{ij} \Gamma_{ij}^{k} \, \frac{du_i}{dt} \, X_j = 0, \qquad k = 1, \ldots, n,$$

is satisfied. □

Remark: Proposition 14.6 can be extended to piecewise C^1 curves.

Definition 14.7. Let M be a smooth manifold and let ∇ be a connection on M. For every curve $\gamma\colon [a,b] \to M$ in M, for every $t \in [a,b]$, the *parallel transport from $\gamma(a)$ to $\gamma(t)$ along γ* is the linear map from $T_{\gamma(a)}M$ to $T_{\gamma(t)}M$ which associates to any $v \in T_{\gamma(a)}M$ the vector $X_v(t) \in T_{\gamma(t)}M$, where X_v is the unique parallel vector field along γ with $X_v(a) = v$. See Figure 14.5.

The following proposition is an immediate consequence of properties of linear ODE's.

Proposition 14.7. *Let M be a smooth manifold and let ∇ be a connection on M. For every C^1 curve $\gamma\colon [a,b] \to M$ in M, the parallel transport along γ defines for every $t \in [a,b]$ a linear isomorphism $P_\gamma\colon T_{\gamma(a)}M \to T_{\gamma(t)}M$, between the tangent spaces $T_{\gamma(a)}M$ and $T_{\gamma(t)}M$.*

In particular, if γ is a closed curve, that is if $\gamma(a) = \gamma(b) = p$, we obtain a linear isomorphism P_γ of the tangent space T_pM, called the *holonomy of γ*. The *holonomy group of ∇ based at p*, denoted $\mathrm{Hol}_p(\nabla)$, is the subgroup of $\mathrm{GL}(n, \mathbb{R})$ (where n is the dimension of the manifold M) given by

$$\mathrm{Hol}_p(\nabla) = \{P_\gamma \in \mathrm{GL}(n, \mathbb{R}) \mid \gamma \text{ is a closed curve based at } p\}.$$

If M is connected, then $\mathrm{Hol}_p(\nabla)$ depends on the basepoint $p \in M$ up to conjugation, and so $\mathrm{Hol}_p(\nabla)$ and $\mathrm{Hol}_q(\nabla)$ are isomorphic for all $p, q \in M$. In this case, it makes sense to talk about the *holonomy group of* ∇. By abuse of language, we call $\mathrm{Hol}_p(\nabla)$ the *holonomy group of M*.

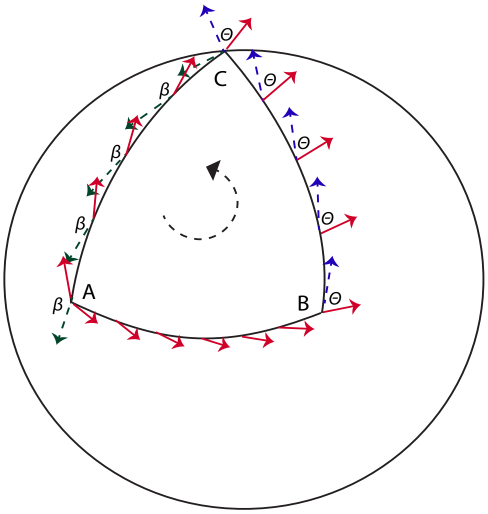


Figure 14.5: The parallel transport of the red vector field around the spherical triangle ABC.

14.3 Connections Compatible with a Metric; Levi-Civita Connections

If a Riemannian manifold M has a metric, then it is natural to define when a connection ∇ on M is compatible with the metric.

Given any two vector fields $Y, Z \in \mathfrak{X}(M)$, the smooth function $\langle Y, Z\rangle$ is defined by

$$\langle Y, Z\rangle(p) = \langle Y_p, Z_p\rangle_p,$$

for all $p \in M$.

Definition 14.8. Given any metric $\langle -, -\rangle$ on a smooth manifold M, a connection ∇ on M is *compatible with the metric*, for short, a *metric connection*, iff

$$X(\langle Y, Z\rangle) = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle,$$

for all vector fields, $X, Y, Z \in \mathfrak{X}(M)$.

Proposition 14.8. *Let M be a Riemannian manifold with a metric $\langle -, -\rangle$. Then M possesses metric connections.*

Proof. For every chart $(U_\alpha, \varphi_\alpha)$, we use the Gram-Schmidt procedure to obtain an orthonormal frame over U_α and we let ∇^α be the flat connection over U_α. By construction, ∇^α is compatible with the metric. We finish the argument by using a partition of unity, leaving the details to the reader. □

We know from Proposition 14.8 that metric connections on TM exist. However, there are *many metric connections* on TM and none of them seems more relevant than the others.

It is remarkable that if we require a certain kind of symmetry on a metric connection, then it is uniquely determined. Such a connection is known as the *Levi-Civita connection*. The Levi-Civita connection can be characterized in several equivalent ways, a rather simple way involving the notion of torsion of a connection.

There are two *error terms* associated with a connection. The first one is the *curvature*

$$R(X, Y) = \nabla_{[X,Y]} + \nabla_Y \nabla_X - \nabla_X \nabla_Y.$$

The second natural error term is the *torsion* $T(X, Y)$ of the connection ∇, given by

$$T(X, Y) = \nabla_X Y - \nabla_Y X - [X, Y],$$

which measures the failure of the connection to behave like the Lie bracket.

Proposition 14.9. *(Levi-Civita, Version 1) Let M be any Riemannian manifold. There is a unique, metric, torsion-free connection ∇ on M; that is, a connection satisfying the conditions:*

$$\begin{aligned} X(\langle Y, Z\rangle) &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ \nabla_X Y - \nabla_Y X &= [X, Y], \end{aligned}$$

for all vector fields, $X, Y, Z \in \mathfrak{X}(M)$. This connection is called the Levi-Civita connection (or canonical connection) on M. Furthermore, this connection is determined by the Koszul formula

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle. \end{aligned}$$

Proof. First we prove uniqueness. Since our metric is a non-degenerate bilinear form, it suffices to prove the Koszul formula. As our connection is compatible with the metric, we have

$$\begin{aligned} X(\langle Y, Z\rangle) &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ Y(\langle X, Z\rangle) &= \langle \nabla_Y X, Z\rangle + \langle X, \nabla_Y Z\rangle \\ -Z(\langle X, Y\rangle) &= -\langle \nabla_Z X, Y\rangle - \langle X, \nabla_Z Y\rangle. \end{aligned}$$

Adding up the above equations and using the fact that the torsion is zero gives us

$$\begin{aligned} X(\langle Y, Z\rangle) \;+\; & Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ =\; & \langle Y, \nabla_X Z - \nabla_Z X\rangle + \langle X, \nabla_Y Z - \nabla_Z Y\rangle + \langle Z, \nabla_X Y + \nabla_Y X\rangle \\ =\; & \langle Y, \nabla_X Z - \nabla_Z X\rangle + \langle X, \nabla_Y Z - \nabla_Z Y\rangle \\ & + \langle Z, \nabla_Y X - \nabla_X Y\rangle + \langle Z, \nabla_X Y + \nabla_X Y\rangle \\ =\; & \langle Y, [X, Z]\rangle + \langle X, [Y, Z]\rangle + \langle Z, [Y, X]\rangle + 2\langle Z, \nabla_X Y\rangle, \end{aligned}$$

which yields the Koszul formula.

Next we prove existence. We begin by checking that the right-hand side of the Koszul formula is $C^\infty(M)$-linear in Z, for X and Y fixed. But then, the linear map $Z \mapsto \langle \nabla_X Y, Z\rangle$ induces a one-form and $\nabla_X Y$ is the vector field corresponding to it *via* the non-degenerate pairing. It remains to check that ∇ satisfies the properties of a connection, which it a bit tedious (for example, see Kuhnel [71], Chapter 5, Section D). □

In the simple case where $M = \mathbb{R}^n$ and the metric is the Euclidean inner product on $\mathbb{R}^n$, any two smooth vector fields X, Y can be written as

$$X = \sum_{i=1}^n f_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^n g_i \frac{\partial}{\partial x_i},$$

for some smooth functions f_i, g_i, and they can be viewed as smooth functions $X, Y \colon \mathbb{R}^n \to \mathbb{R}^n$. Then it is easy to verify that the Levi-Civita connection is given by

$$(\nabla_X Y)(p) = dY_p(X(p)), \quad p \in \mathbb{R}^n,$$

because the right-hand side satisfies all the conditions of Proposition 14.9, and there is a unique such connection. Thus, the Levi-Civita connection induced by the Euclidean inner product on $\mathbb{R}^n$ is the flat connection.

Remark: In a chart (U, φ), recall that $g_{ij} = \langle \frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\rangle$. If we set

$$\partial_k g_{ij} = \frac{\partial}{\partial x_k}(g_{ij}),$$

then it can be shown that the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^k_{ij} = \frac{1}{2}\sum_{l=1}^n g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),$$

where (g^{kl}) is the inverse of the matrix (g_{kl}), and the Γ^k_{ij} are defined by

$$\nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_j} = \sum_{k=1}^n \Gamma^k_{ij} \frac{\partial}{\partial x_k};$$

see Definition 14.2. For example, suppose we take the polar coordinate parameterization of the plane given by

$$x = r\cos\theta \qquad y = r\sin\theta,$$

with $0 < \theta < 2\pi$ and $r > 0$. For any $p = (r\cos\theta, r\sin\theta)$, a basis for the tangent plane $T_p\mathbb{R}^2$ is

$$\begin{aligned} \frac{\partial p}{\partial r} &= (\cos\theta, \sin\theta) \\ \frac{\partial p}{\partial \theta} &= (-r\sin\theta, r\cos\theta). \end{aligned}$$

Since

$$\begin{aligned} \langle \frac{\partial p}{\partial r}, \frac{\partial p}{\partial r} \rangle &= 1 \\ \langle \frac{\partial p}{\partial r}, \frac{\partial p}{\partial \theta} \rangle &= 0 \\ \langle \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \theta} \rangle &= r^2, \end{aligned}$$

we discover that

$$g = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \qquad g^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{r^2} \end{pmatrix}.$$

By associating r with 1 and θ with 2, we discover that

$$\Gamma^r_{\theta\theta} = \Gamma^1_{22} = -r,$$

since

$$\begin{aligned} \Gamma^1_{22} &= \frac{1}{2} \sum_{l=1}^{2} g^{1l}(\partial_2 g_{2l} + \partial_2 g_{2l} - \partial_l g_{22}) \\ &= \frac{1}{2} \left[g^{11}(2\partial_2 g_{21} - \partial_1 g_{22}) + g^{12}(2\partial_2 g_{22} - \partial_2 g_{22}) \right] \\ &= -\frac{1}{2} g^{11} \partial_1 g_{22} = -\frac{1}{2} \frac{\partial}{\partial r} g_{22} \\ &= -\frac{1}{2} \frac{\partial}{\partial r} r^2 = -r. \end{aligned}$$

Similar calculations show that

$$\begin{aligned} \Gamma^r_{r\theta} &= \Gamma^1_{12} = \Gamma^1_{21} = 0 \\ \Gamma^\theta_{r\theta} &= \Gamma^2_{12} = \Gamma^2_{21} = \frac{1}{r} \\ \Gamma^r_{rr} &= \Gamma^1_{11} = 0 \\ \Gamma^\theta_{rr} &= \Gamma^1_{22} = 0 \\ \Gamma^\theta_{\theta\theta} &= \Gamma^2_{22} = 0. \end{aligned}$$

Since

$$\nabla_{\frac{\partial}{\partial x_i}}\left(\frac{\partial}{\partial x_j}\right) = \sum_{k=1}^{n} \Gamma_{ij}^{k}\,\frac{\partial}{\partial x_k},$$

we explicitly calculate the Levi-Civita connection as

$$\begin{aligned}
\nabla_{\frac{\partial}{\partial r}}\left(\frac{\partial}{\partial r}\right) &= \sum_{k=1}^{2} \Gamma_{11}^{k}\frac{\partial}{\partial x_k} = 0\\
\nabla_{\frac{\partial}{\partial r}}\left(\frac{\partial}{\partial \theta}\right) &= \sum_{k=1}^{2} \Gamma_{12}^{k}\frac{\partial}{\partial x_k} = \frac{1}{r}\frac{\partial}{\partial \theta}\\
\nabla_{\frac{\partial}{\partial \theta}}\left(\frac{\partial}{\partial r}\right) &= \sum_{k=1}^{2} \Gamma_{21}^{k}\frac{\partial}{\partial x_k} = \frac{1}{r}\frac{\partial}{\partial \theta}\\
\nabla_{\frac{\partial}{\partial \theta}}\left(\frac{\partial}{\partial \theta}\right) &= \sum_{k=1}^{2} \Gamma_{22}^{k}\frac{\partial}{\partial x_k} = -r\frac{\partial}{\partial r}.
\end{aligned}$$

It can be shown that a connection is torsion-free iff

$$\Gamma_{ij}^{k} = \Gamma_{ji}^{k}, \qquad \text{for all } i, j, k.$$

We conclude this section with various useful facts about torsion-free or metric connections. First, there is a nice characterization for the Levi-Civita connection induced by a Riemannian manifold over a submanifold.

Proposition 14.10. *Let M be any Riemannian manifold and let N be any submanifold of M equipped with the induced metric. If ∇^M and ∇^N are the Levi-Civita connections on M and N, respectively, induced by the metric on M, then for any two vector field X and Y in $\mathfrak{X}(M)$ with $X(p), Y(p) \in T_pN$, for all $p \in N$, we have*

$$\nabla_X^N Y = (\nabla_X^M Y)^{\parallel},$$

where $(\nabla_X^M Y)^{\parallel}(p)$ is the orthogonal projection of $\nabla_X^M Y(p)$ onto T_pN, for every $p \in N$.

In particular, if γ is a curve on a surface $M \subseteq \mathbb{R}^3$, then a vector field $X(t)$ along γ is parallel iff $X'(t)$ is normal to the tangent plane $T_{\gamma(t)}M$. See Figure 14.4.

If ∇ is a metric connection, then we can say more about the parallel transport along a curve. Recall from Section 14.2, Definition 14.6, that a vector field X along a curve γ is parallel iff

$$\frac{DX}{dt} = 0.$$

The following proposition will be needed:

Proposition 14.11. *Given any Riemannian manifold M and any metric connection ∇ on M, for every curve $\gamma\colon [a,b] \to M$ on M, if X and Y are two vector fields along γ, then*

$$\frac{d}{dt}\langle X(\gamma(t)), Y(\gamma(t))\rangle = \left\langle \frac{DX}{dt}, Y(\gamma(t))\right\rangle + \left\langle X(\gamma(t)), \frac{DY}{dt}\right\rangle.$$

Proof. Since

$$\frac{d}{dt}\langle X(\gamma(t)), Y(\gamma(t))\rangle = d\langle X, Y\rangle_{\gamma(t)}(\gamma'(t)) = \gamma'(t)\langle X, Y\rangle_{\gamma(t)},$$

it would be tempting to apply directly the equation

$$Z(\langle X, Y\rangle) = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle$$

asserting the compatibility of the connection with the metric, but this is wrong because the above equation applies to vectors fields X, Y defined on the whole of M (or at least on some open subset of M), and yet in our situation X and Y are only defined along the curve γ, and in general, such vector fields cannot be extended to an open subset of M. This subtle point seems to have been overlooked in several of the classical texts. Note that Milnor [81] circumvents this difficulty by defining compatibility in a different way (which turns out to be equivalent to the notion used here). Our way out is to use charts, as in the proof of Proposition 14.5; this is the proof method used by O'Neill [91] and Gallot, Hulin and Lafontaine [49] (Chapter 2), although they leave computations to the reader.

We may assume that $\gamma\colon [a,b] \to U$ for some chart (U, φ). Then, if $(s_1, \ldots, s_n)$ is a frame above U, we can write

$$\begin{aligned} X(\gamma(t)) &= \sum_{i=1}^n X_i(t) s_i(\gamma(t)) \\ Y(\gamma(t)) &= \sum_{k=1}^n Y_k(t) s_k(\gamma(t)), \end{aligned}$$

and as in the proof of Proposition 14.5, we have

$$\begin{aligned} \frac{DX}{dt} &= \sum_{j=1}^n \left(\frac{dX_j}{dt}\, s_j(\gamma(t)) + X_j(t) \nabla_{\gamma'(t)}(s_j(\gamma(t))) \right) \\ \frac{DY}{dt} &= \sum_{l=1}^n \left(\frac{dY_l}{dt}\, s_l(\gamma(t)) + Y_l(t) \nabla_{\gamma'(t)}(s_l(\gamma(t))) \right). \end{aligned}$$

It follows that

$$\begin{aligned}\left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle &= \sum_{j,k=1}^{n} \frac{dX_j}{dt} Y_k(t) \langle s_j(\gamma(t)), s_k(\gamma(t))\rangle \\ &\quad + \sum_{j,k=1}^{n} X_j(t) Y_k(t) \langle \nabla_{\gamma'(t)} s_j(\gamma(t)), s_k(\gamma(t))\rangle \\ &\quad + \sum_{i,l=1}^{n} X_i(t) \frac{dY_l}{dt} \langle s_i(\gamma(t)), s_l(\gamma(t))\rangle \\ &\quad + \sum_{i,l=1}^{n} X_i(t) Y_l(t) \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_l(\gamma(t))\rangle,\end{aligned}$$

so

$$\begin{aligned}\left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle &= \sum_{i,k=1}^{n} \left(\frac{dX_i}{dt} Y_k(t) + X_i(t) \frac{dY_k}{dt} \right) \langle s_i(\gamma(t)), s_k(\gamma(t))\rangle \\ &\quad + \sum_{i,k=1}^{n} X_i(t) Y_k(t) \left(\langle \nabla_{\gamma'(t)} s_i(\gamma(t)), s_k(\gamma(t))\rangle + \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_k(\gamma(t))\rangle \right).\end{aligned}$$

On the other hand, the compatibility of the connection with the metric implies that

$$\langle \nabla_{\gamma'(t)} s_i(\gamma(t)), s_k(\gamma(t))\rangle + \langle s_i(\gamma(t)), \nabla_{\gamma'(t)} s_k(\gamma(t))\rangle = \gamma'(t)\langle s_i, s_k\rangle_{\gamma(t)} = \frac{d}{dt}\langle s_i(\gamma(t)), s_k(\gamma(t))\rangle,$$

and thus we have

$$\begin{aligned}\left\langle \frac{DX}{dt}, Y \right\rangle + \left\langle X, \frac{DY}{dt} \right\rangle &= \sum_{i,k=1}^{n} \left(\frac{dX_i}{dt} Y_k(t) + X_i(t) \frac{dY_k}{dt} \right) \langle s_i(\gamma(t)), s_k(\gamma(t))\rangle \\ &\quad + \sum_{i,k=1}^{n} X_i(t) Y_k(t) \frac{d}{dt} \langle s_i(\gamma(t)), s_k(\gamma(t))\rangle \\ &= \frac{d}{dt} \left(\sum_{i,k=1}^{n} X_i(t) Y_k(t) \langle s_i(\gamma(t)), s_k(\gamma(t))\rangle \right) \\ &= \frac{d}{dt} \langle X(\gamma(t)), Y(\gamma(t))\rangle,\end{aligned}$$

as claimed. □

Using Proposition 14.11 we get

Proposition 14.12. *Given any Riemannian manifold M and any metric connection ∇ on M, for every curve $\gamma\colon [a,b] \to M$ on M, if X and Y are two vector fields along γ that are parallel, then*

$$\langle X, Y\rangle = C,$$

for some constant C. In particular, $\|X(t)\|$ is constant. Furthermore, the linear isomorphism $P_\gamma\colon T_{\gamma(a)} \to T_{\gamma(b)}$ is an isometry.

Proof. From Proposition 14.11, we have

$$\frac{d}{dt}\left\langle X(\gamma(t)), Y(\gamma(t))\right\rangle = \left\langle \frac{DX}{dt}, Y(\gamma(t))\right\rangle + \left\langle X(\gamma(t)), \frac{DY}{dt}\right\rangle.$$

As X and Y are parallel along γ, we have $DX/dt = 0$ and $DY/dt = 0$, so

$$\frac{d}{dt}\left\langle X(\gamma(t)), Y(\gamma(t))\right\rangle = 0,$$

which shows that $\langle X(\gamma(t)), Y(\gamma(t))\rangle$ is constant. Therefore, for all $v, w \in T_{\gamma(a)}$, if X and Y are the unique vector fields parallel along γ such that $X(\gamma(a)) = v$ and $Y(\gamma(a)) = w$ given by Proposition 14.6, we have

$$\langle P_\gamma(v), P_\gamma(w)\rangle = \langle X(\gamma(b)), Y(\gamma(b))\rangle = \langle X(\gamma(a)), Y(\gamma(a))\rangle = \langle v, w\rangle,$$

which proves that P_γ is an isometry. □

In particular, Proposition 14.12 shows that the holonomy group $\mathrm{Hol}_p(\nabla)$ based at p is a subgroup of $\mathbf{O}(n)$.

14.4 Problems

Problem 14.1. Prove that in a chart (U, φ), with ith local coordinate $x_i = pr_i \circ \varphi$, for any two vector fields X and Y, we have

$$\nabla_X Y = \sum_{i=1}^n \left(\sum_{j=1}^n X_j \frac{\partial Y_i}{\partial x_j} + \sum_{j,k=1}^n \Gamma^i_{jk} X_j Y_k\right) \frac{\partial}{\partial x_i},$$

where the Γ^i_{jk} are the Christofffel symbols of Definition 14.2.

Problem 14.2. Check that the expression $\nabla = \sum_\alpha f_\alpha \nabla^\alpha$ constructed in Proposition 14.4 is indeed a connection.

Problem 14.3. Let M be a smooth manifold and let $\gamma\colon [a,b] \to M$ be a smooth curve in M. For any chart (U, φ), let $u_i = pr_i \circ \varphi \circ \gamma$. For any vector field X on M, show that

$$\frac{DX}{dt} = \sum_{k=1}^n \left(\frac{dX_k}{dt} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} X_j\right) \frac{\partial}{\partial x_k}.$$

Problem 14.4. Fill in the details of Proposition 14.8.

Problem 14.5. Check that the vector field $\nabla_X Y$ constructed at the end of the proof of Proposition 14.9 is indeed a connection.

Problem 14.6. Given a Riemannian manifold (M, g), in a chart (U, φ), if we set

$$\partial_k g_{ij} = \frac{\partial}{\partial x_k}(g_{ij}),$$

then show that the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^k_{ij} = \frac{1}{2}\sum_{l=1}^{n} g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),$$

where (g^{kl}) is the inverse of the matrix (g_{kl}) representing the metric g in U.

Problem 14.7. In a Riemannian manifold, prove that a connection is torsion-free iff

$$\Gamma^k_{ij} = \Gamma^k_{ji}, \qquad \text{for all } i, j, k.$$

Problem 14.8. Prove Proposition 14.10.

Problem 14.9. Give an example of a connection defined on $\mathbb{R}^3$ which is compatible with the Euclidean metric but is not torsion-free.

Hint. Define this connection in terms of the Christoffel symbols.

Problem 14.10. Let $M \subseteq \mathbb{R}^3$ be a surface with the Riemannian metric induced from $\mathbb{R}^3$. Let $c\colon (0,1) \to M$ be a differentiable curve on M and let V be a vector field tangent to M along c, which can be viewed as a smooth function $V\colon (0,1) \to \mathbb{R}^3$ with $V(t) \in T_{c(t)}M$.

(1) Prove that V is parallel iff dV/dt is perpendicular to $T_{c(t)}M \subseteq \mathbb{R}^3$, where dV/dt is the usual derivative of $V\colon (0,1) \to \mathbb{R}^3$.

(2) If $S^2 \subseteq \mathbb{R}^3$ is the unit sphere of $\mathbb{R}^3$, show that the velocity field along great circles, parametrized by arc length, is a parallel field.

Problem 14.11. Let X and Y be differentiable vector fields on a Riemannian manifold M. For any $p \in M$, let $c\colon I \to M$ be an integral curve through p, so that $c(0) = p$ and $dc/dt = X(c(t))$. Prove that the Levi-Civita connection of M is given by

$$(\nabla_X Y)(p) = \frac{d}{dt}(P^{-1}_{c,0,t}(Y(c(t))))\bigg|_{t=0},$$

where $P_{c,0,t}\colon T_{c(0)} \to T_{c(t)}$ is the parallel transport along c from 0 to t. This shows that the Levi-Civita connection can be obtained from the concept of parallelism.

Problem 14.12. Complete the partition of unity argument in the proof of Proposition 14.8.

Chapter 15

Geodesics on Riemannian Manifolds

If (M, g) is a Riemannian manifold, then the concept of *length* makes sense for any piecewise smooth (in fact, C^1) curve on M. It is then possible to define the structure of a metric space on M, where $d(p, q)$ is the greatest lower bound of the length of all curves joining p and q. Curves on M which locally yield the shortest distance between two points are of great interest. These curves, called *geodesics*, play an important role and the goal of this chapter is to study some of their properties.

In Section 15.1, we define geodesics and prove some of their basic properties, in particular the fact that they always exist locally. Note that the notion of geodesic only requires a connection on a manifold, since by definition, a geodesic is a curve γ such that γ' is parallel along γ, that is

$$\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = 0,$$

where $\frac{D}{dt}$ be the covariant derivative along γ, also denoted $\nabla_{\gamma'}$ (see Proposition 14.5 and Definition 14.5). Thus, geodesics can be defined in manifolds that are not endowed with a Riemannian metric. However, most useful properties of geodesics involve metric notions, and their proofs use the fact that the connection on the manifold is compatible with the metric and torsion-free. For this reason, *we usually assume that we are dealing with Riemannian manifolds equipped with the Levi-Civita connection.*

For every point $p \in M$ on a manifold M, using geodesics through p we can define the *exponential map* $\exp_p$, which maps a neighborhood of 0 in the tangent space T_pM back into M. This provides nice parametrizations of the manifold; see Section 15.2. The exponential map is a very useful technical tool because it establishes a precise link between the linearization of a manifold by its tangent spaces and the manifold itself. In particular, manifolds for which the exponential map is defined for all $p \in M$ and all $v \in T_pM$ can be studied in more depth; see Section 15.3. Such manifolds are called *complete*. A fundamental theorem about complete manifolds is the theorem of Hopf and Rinow, which we prove in full.

Geodesics are locally distance minimizing, but in general they fail to be distance minimizing if they extend too far. This phenomenon is captured by the subtle notion of *cut locus*, which we define and study briefly. In Section 15.4, we also discuss briefly various notions of convexity induced by geodesics.

In Section 15.5, we define the Hessian of a function defined on a Riemannian manifold, and show how the Hessian can be computed using geodesics.

Geodesics between two points p and q turn out to be critical points of the energy functional on the path space $\Omega(p, q)$, the space of all piecewise smooth curves from p to q. This is an infinite dimensional manifold consisting of functions (curves), so in order to define what it means for a curve ω in $\Omega(p, q)$ to be a critical point of a function F defined on $\Omega(p, q)$, we introduce the notion of *variation* (of a curve). Then it is possible to obtain a formula giving the derivative $dE(\widetilde{\alpha}(u))/du \mid_{u=0}$ of the energy function E (with $E(\omega) = \int_0^1 \|\omega'(t)\|^2\, dt$) applied to a variation $\widetilde{\alpha}$ of a curve ω (the *first variation formula*); see Section 15.6. It turns out that a curve ω is a geodesic iff it is a critical point of the energy function (that is, $dE(\widetilde{\alpha}(u))/du \mid_{u=0}= 0$ for all variations of ω). This result provides a fruitful link with the calculus of variations.

Among the many presentations of this subject, in our opinion, Milnor's account [81] (Part II, Section 11) is still one of the best, certainly by its clarity and elegance. We acknowledge that our presentation was heavily inspired by this beautiful work. We also relied heavily on Gallot, Hulin and Lafontaine [49] (Chapter 2), Do Carmo [39], O'Neill [91], Kuhnel [71], and class notes by Pierre Pansu
(see http://www.math.u-psud.fr/%7Epansu/web_dea/resume_dea_04.html in http://www.math.u-psud.fr˜pansu/). Another reference that is remarkable by its clarity and the completeness of its coverage is Postnikov [96].

15.1 Geodesics, Local Existence and Uniqueness

Recall the following definitions regarding curves.

Definition 15.1. Given any smooth manifold M, a *smooth parametric curve* (for short, *curve*) on M is a smooth map $\gamma\colon I \to M$, where I is some open interval of $\mathbb{R}$. For a closed interval $[a, b] \subseteq \mathbb{R}$, a map $\gamma\colon [a, b] \to M$ is a *smooth curve from* $p = \gamma(a)$ *to* $q = \gamma(b)$ iff γ can be extended to a smooth curve $\widetilde{\gamma}\colon (a - \epsilon, b + \epsilon) \to M$, for some $\epsilon > 0$. Given any two points $p, q \in M$, a continuous map $\gamma\colon [a, b] \to M$ is a *piecewise smooth curve from p to q* iff

(1) There is a sequence $a = t_0 < t_1 < \cdots < t_{k-1} < t_k = b$ of numbers $t_i \in \mathbb{R}$, so that each map $\gamma_i = \gamma \restriction [t_i, t_{i+1}]$, called a *curve segment*, is a smooth curve for $i = 0, \ldots, k - 1$.

(2) $\gamma(a) = p$ and $\gamma(b) = q$.

The set of all piecewise smooth curves from p to q is denoted by $\Omega(M; p, q)$, or briefly by $\Omega(p, q)$ (or even by Ω, when p and q are understood).

The set $\Omega(M; p, q)$ is an important object, sometimes called the *path space* of M (from p to q). Unfortunately it is an infinite-dimensional manifold, which makes it hard to investigate its properties.

Observe that at any junction point $\gamma_{i-1}(t_i) = \gamma_i(t_i)$, there may be a jump in the velocity vector of γ. We let $\gamma'((t_i)_+) = \gamma_i'(t_i)$ and $\gamma'((t_i)_-) = \gamma_{i-1}'(t_i)$.

Definition 15.2. Let (M, g) be a Riemannian manifold. Given any $p \in M$, for every $v \in T_pM$, the *(Riemannian) norm* of v, denoted $\|v\|$, is defined by

$$\|v\| = \sqrt{g_p(v, v)}.$$

The Riemannian inner product $g_p(u, v)$ of two tangent vectors $u, v \in T_pM$ will also be denoted by $\langle u, v\rangle_p$, or simply $\langle u, v\rangle$.

Definition 15.3. Let (M, g) be a Riemannian manifold. Given any curve $\gamma \in \Omega(M; p, q)$, the *length* $L(\gamma)$ of γ is defined by

$$L(\gamma) = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \|\gamma'(t)\| \; dt = \sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \sqrt{g(\gamma'(t), \gamma'(t))}\, dt.$$

It is easy to see that $L(\gamma)$ is unchanged by a monotone reparametrization (that is, a map $h\colon [a, b] \to [c, d]$ whose derivative h' has a constant sign).

Now let M be any smooth manifold equipped with an arbitrary connection ∇. For every curve γ on M, recall that $\frac{D}{dt}$ is the associated covariant derivative along γ, also denoted $\nabla_{\gamma'}$ (see Proposition 14.5 and Definition 14.5).

Definition 15.4. Let M be any smooth manifold equipped with a connection ∇. A curve $\gamma\colon I \to M$ (where $I \subseteq \mathbb{R}$ is any interval) is a *geodesic* iff $\gamma'(t)$ is parallel along γ; that is, iff

$$\frac{D\gamma'}{dt} = \nabla_{\gamma'}\gamma' = 0.$$

Observe that the notion of geodesic only requires a connection on a manifold, and that geodesics can be defined in manifolds that are *not endowed with a Riemannian metric*. However, most useful properties of geodesics involve metric notions, and their proofs use the fact that the connection on the manifold is compatible with the metric and torsion-free. *Therefore, from now on, we assume unless otherwise specified that our Riemannian manifold (M, g) is equipped with the Levi-Civita connection.*

If M was embedded in $\mathbb{R}^d$, a geodesic would be a curve γ such that the acceleration vector $\gamma'' = \frac{D\gamma'}{dt}$ is normal to $T_{\gamma(t)}M$.

Since our connection is compatible with the metric, by Proposition 14.12 implies that for a geodesic γ, $\|\gamma'(t)\| = \sqrt{g(\gamma'(t), \gamma'(t))}$ is constant, say $\|\gamma'(t)\| = c$. If we define the *arc-length* function $s(t)$ relative to a, where a is any chosen point in I, by

$$s(t) = \int_a^t \sqrt{g(\gamma'(t), \gamma'(t))}\, dt = c(t-a), \qquad t \in I,$$

we conclude that for a geodesic $\gamma(t)$, the parameter t is an affine function of the arc-length. When $c = 1$, which can be achieved by an affine reparametrization, we say that the geodesic is *normalized.*

The geodesics in $\mathbb{R}^n$ are the straight lines parametrized by constant velocity. The geodesics of the 2-sphere are the great circles, parametrized by arc-length. The geodesics of the Poincaré half-plane are the half-lines $x = a$ and the half-circles centered on the x-axis. The geodesics of an ellipsoid are quite fascinating. They can be completely characterized, and they are parametrized by elliptic functions (see Hilbert and Cohn-Vossen [60], Chapter 4, Section and Berger and Gostiaux [15], Section 10.4.9.5).

In a local chart (U, φ), since a geodesic is characterized by the fact that its velocity vector field $\gamma'(t)$ along γ is parallel, by Proposition 14.6, it is the solution of the following system of second-order ODE's in the unknowns u_k:

$$\frac{d^2 u_k}{dt^2} + \sum_{ij} \Gamma^k_{ij} \frac{du_i}{dt} \frac{du_j}{dt} = 0, \qquad k = 1, \ldots, n, \tag{$*$}$$

with $u_i = pr_i \circ \varphi \circ \gamma$ $(n = \dim(M))$.

The standard existence and uniqueness results for ODE's can be used to prove the following proposition (see O'Neill [91], Chapter 3):

Proposition 15.1. *Let (M, g) be a Riemannian manifold. For every point $p \in M$ and every tangent vector $v \in T_pM$, there is some interval $(-\eta, \eta)$ and a unique geodesic*

$$\gamma_v \colon (-\eta, \eta) \to M,$$

satisfying the conditions

$$\gamma_v(0) = p, \qquad \gamma_v'(0) = v.$$

From a practical point of view, Proposition 15.1 is useless. In general, for an arbitrary manifold M, it is impossible to solve explicitly the second-order equations $(*)$; even for familiar manifolds it is very hard to solve explicitly the second-order equations $(*)$. Riemannian covering maps and Riemannian submersions are notions that can be used for finding geodesics; see Chapter 17. In the case of a Lie group with a bi-invariant metric, geodesics can be described explicitly; see Chapter 20. Geodesics can also be described explicitly for certain classes of reductive homogeneous manifolds; see Chapter 22.

The following proposition is used to prove that every geodesic is contained in a unique maximal geodesic (i.e., with largest possible domain). For a proof, see O'Neill [91] (Chapter 3) or Petersen [93] (Chapter 5, Section 2, Lemma 7).

Proposition 15.2. *For any two geodesics $\gamma_1\colon I_1 \to M$ and $\gamma_2\colon I_2 \to M$, if $\gamma_1(a) = \gamma_2(a)$ and $\gamma_1'(a) = \gamma_2'(a)$ for some $a \in I_1 \cap I_2$, then $\gamma_1 = \gamma_2$ on $I_1 \cap I_2$.*

Remark: It is easy to check that Propositions 15.1 and 15.2 hold for any smooth manifold equipped with an arbitrary connection.

Propositions 15.1 and 15.2 imply the following definition:

Definition 15.5. Let M be a smooth manifold equipped with an arbitrary connection. For every $p \in M$ and every $v \in T_pM$, there is a unique geodesic, denoted γ_v, such that $\gamma(0) = p$, $\gamma'(0) = v$, and the domain of γ is the largest possible, that is, cannot be extended. We call γ_v a *maximal geodesic* (with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v$).

Observe that the system of differential equations satisfied by geodesics has the following homogeneity property: If $t \mapsto \gamma(t)$ is a solution of the above system, then for every constant c, the curve $t \mapsto \gamma(ct)$ is also a solution of the system. We can use this fact together with standard existence and uniqueness results for ODE's to prove the proposition below.

Proposition 15.3. *Let (M, g) be a Riemannian manifold. For every point $p_0 \in M$, there is an open subset $U \subseteq M$, with $p_0 \in U$, and some $\epsilon > 0$, so that for every $p \in U$ and every tangent vector $v \in T_pM$, with $\|v\| < \epsilon$, there is a unique geodesic*

$$\gamma_v\colon (-2, 2) \to M$$

satisfying the conditions

$$\gamma_v(0) = p, \qquad \gamma_v'(0) = v.$$

Proof. We follow Milnor [81] (Part II, Section 10, Proposition 10.2). By a standard theorem about the existence and uniqueness of solutions of ODE's, for every $p_0 \in M$, there is some open subset U of M containing p_0, and some numbers $\epsilon_1 > 0$ and $\epsilon_2 > 0$, such that for every $p \in M$ and every $v \in T_pM$ with $\|v\| < \epsilon_1$, there is a unique geodesic $\widetilde{\gamma}_v\colon (-2\epsilon_2, 2\epsilon_2) \to M$ such that $\widetilde{\gamma}_v(0) = p$ and $\widetilde{\gamma}_v'(0) = v$. Let $\eta = 2\epsilon_2$. For any constant $c \neq 0$, the curve $t \mapsto \widetilde{\gamma}_v(ct)$ is a geodesic defined on $(-\eta/c, \eta/c)$ (or $(\eta/c, -\eta/c)$ if $c < 0$) such that $\widetilde{\gamma}'(0) = cv$. Thus,

$$\widetilde{\gamma}_v(ct) = \widetilde{\gamma}_{cv}(t), \qquad ct \in (-\eta, \eta).$$

Pick $\epsilon > 0$ so that $\epsilon < \epsilon_1\epsilon_2$. Then, if $\|v\| < \epsilon$ and $|t| < 2$, note that

$$\|v/\epsilon_2\| < \epsilon_1 \quad \text{and} \quad |\epsilon_2 t| < 2\epsilon_2.$$

Hence, we can define the geodesic γ_v by

$$\gamma_v(t) = \widetilde{\gamma}_{v/\epsilon_2}(\epsilon_2 t), \quad \|v\| < \epsilon,\ |t| < 2,$$

and we have $\gamma_v(0) = p$ and $\gamma_v'(0) = v$, which concludes the proof. □

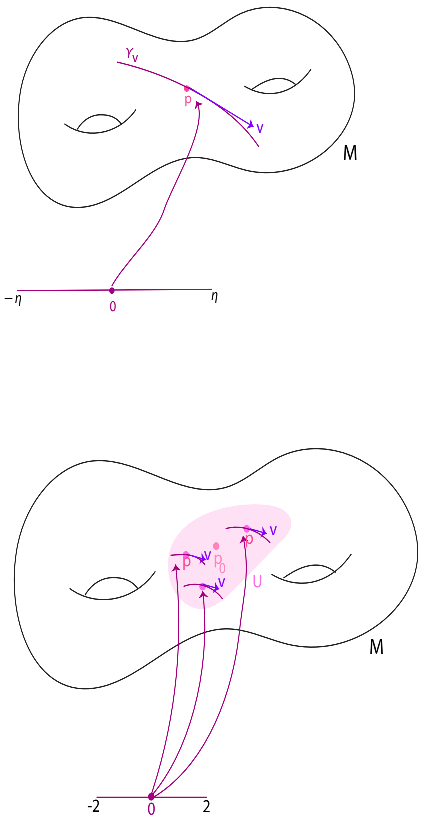


Figure 15.1: The top figure illustrates Proposition 15.1 for the torus M while the bottom figure illustrates Proposition 15.3.

A major difference between Proposition 15.1 and Proposition 15.3 is that Proposition 15.1 yields for any $p \in M$ and any $v \in T_pM$ a *single* geodesic $\gamma_v \colon (-\eta, \eta) \to M$ such that $\gamma_v(0) = p$ and $\gamma_p'(0) = v$, but Proposition 15.3 yields a *family* of geodesics $\gamma_v \colon (-2, 2) \to M$ such that $\gamma_v(0) = p$ and $\gamma_p'(0) = v$, with the *same domain*, for every p in some small enough open subset U, and for small enough $v \in T_pM$. See Figure 15.1.

Remark: Proposition 15.3 holds for a Riemannian manifold equipped with an arbitrary connection.

15.2 The Exponential Map

The idea behind the exponential map is to parametrize a smooth manifold M locally near any $p \in M$ in terms of a map from the tangent space T_pM to the manifold, this map being defined in terms of geodesics.

Definition 15.6. Let M be a smooth manifold equipped with some arbitrary connection. For every $p \in M$, let $\mathcal{D}(p)$ (or simply, $\mathcal{D}$) be the open subset of T_pM given by

$$\mathcal{D}(p) = \{v \in T_pM \mid \gamma_v(1) \quad \text{is defined}\},$$

where γ_v is the unique maximal geodesic with initial conditions $\gamma_v(0) = p$ and $\gamma_v'(0) = v$. The *exponential map* is the map $\exp_p\colon \mathcal{D}(p) \to M$ given by

$$\exp_p(v) = \gamma_v(1).$$

It is easy to see that $\mathcal{D}(p)$ is *star-shaped* (with respect to p), which means that if $w \in \mathcal{D}(p)$, then the line segment $\{tw \mid 0 \leq t \leq 1\}$ is contained in $\mathcal{D}(p)$. See Figure 15.2.

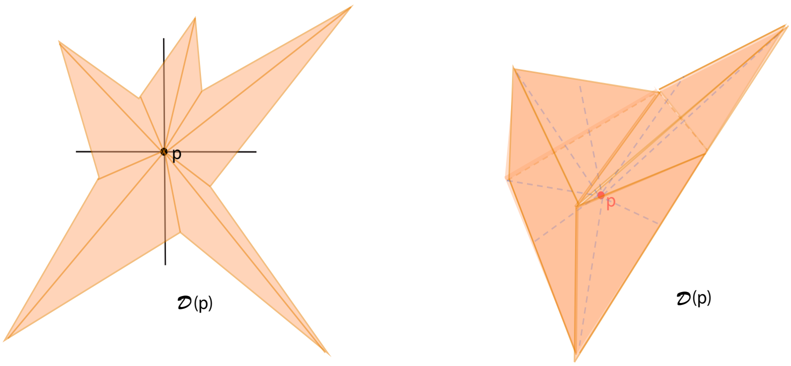


Figure 15.2: The left figure is a star-shaped region in $\mathbb{R}^2$ (with respect to p), while the right figure is a star-shaped region in $\mathbb{R}^3$ (with respect to p). Both regions contain line segments radiating from p.

In view of the fact that if $\gamma_v\colon (-\eta, \eta) \to M$ is a geodesic through p with initial velocity v, then for any $c \neq 0$,

$$\gamma_v(ct) = \gamma_{cv}(t), \quad ct \in (-\eta, \eta),$$

we have

$$\exp_p(tv) = \gamma_{tv}(1) = \gamma_v(t), \quad tv \in \mathcal{D}(p),$$

so the curve

$$t \mapsto \exp_p(tv), \qquad tv \in \mathcal{D}(p),$$

is the geodesic γ_v through p such that $\gamma_v'(0) = v$. Such geodesics are called *radial geodesics*.

In a Riemannian manifold with the Levi-Civita connection, the point $\exp_p(tv)$ is obtained by running along the geodesic γ_v an arc length equal to $t\,\|v\|$, starting from p. If the tangent vector tv at p is a flexible wire, the exponential map wraps the wire along the geodesic curve without stretching its length. See Figure 15.3.

In general, $\mathcal{D}(p)$ is a proper subset of T_pM. For example, if U is a bounded open subset of $\mathbb{R}^n$, since we can identify T_pU with $\mathbb{R}^n$ for all $p \in U$, then $\mathcal{D}(p) \subseteq U$, for all $p \in U$.

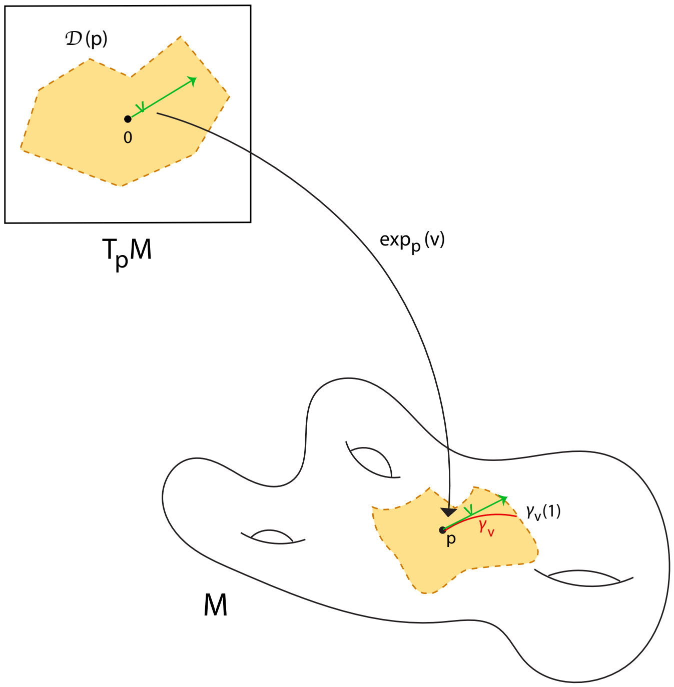


Figure 15.3: The image of v under $\exp_p$.

Definition 15.7. A smooth manifold M equipped with an arbitrary connection is *geodesically complete* iff $\mathcal{D}(p) = T_pM$ for all $p \in M$; that is, the exponential $\exp_p(v)$ is defined for all $p \in M$ and for all $v \in T_pM$.

Equivalently, (M, g) is geodesically complete iff every geodesic can be extended indefinitely.

Geodesically complete Riemannian manifolds (with the Levi-Civita connection) have nice properties, some of which will be investigated later.

Proposition 15.4. *Let M be a Riemannian manifold. For any $p \in M$ we have $d(\exp_p)_0 = \mathrm{id}_{T_pM}$.*

Proof. For every $v \in \mathcal{D}(p)$, the map $t \mapsto \exp_p(tv)$ is the geodesic γ_v, and

$$\frac{d}{dt}(\gamma_v(t))|_{t=0} = v = \frac{d}{dt}(\exp_p(tv))|_{t=0} = d(\exp_p)_0(v). \qquad \square$$

It follows from the inverse function theorem that $\exp_p$ is a diffeomorphism from some open ball in T_pM centered at 0 to M.

By using the curve $t \mapsto (t+1)v$ passing through v in T_pM and with initial velocity $v \in T_v(T_pM) \approx T_pM$, we get

$$d(\exp_p)_v(v) = \frac{d}{dt}(\gamma_v(t+1))|_{t=0} = \gamma_v'(1).$$

The following stronger proposition plays a crucial role in the proof of the Hopf-Rinow Theorem; see Theorem 15.16.

Proposition 15.5. *Let (M, g) be a Riemannian manifold. For every point $p \in M$, there is an open subset $W \subseteq M$, with $p \in W$, and a number $\epsilon > 0$, so that:*

(1) Any two points q_1, q_2 of W are joined by a unique geodesic of length $< \epsilon$.

(2) This geodesic depends smoothly upon q_1 and q_2; that is, if $t \mapsto \exp_{q_1}(tv)$ is the geodesic joining q_1 and q_2 $(0 \leq t \leq 1)$, then $v \in T_{q_1}M$ depends smoothly on (q_1, q_2).

(3) For every $q \in W$, the map $\exp_q$ is a diffeomorphism from the open ball $B(0,\epsilon) \subseteq T_qM$ to its image $U_q = \exp_q(B(0,\epsilon)) \subseteq M$, with $W \subseteq U_q$ and U_q open.

Proof. We follow Milnor [81] (Chapter II, Section 10, Lemma 10.3). Let

$$\mathcal{U} = \{(q, v) \in TM \mid q \in U, v \in T_qM, \|v\| < \epsilon_1\},$$

where the open subset U of M and ϵ_1 are given by Proposition 15.3, for the point $p_0 = p \in M$. Then we can define the map $\Phi\colon \mathcal{U} \to M \times M$ by

$$\Phi(q, v) = (q, \exp_q(v)).$$

We claim that $d\Phi_{(p,0)}$ is invertible, which implies that Φ is a local diffeomorphism near $(p, 0)$. If we pick a chart (V, φ) at p, then we have the chart $(V \times V, \varphi \times \varphi)$ at $(p, p) = \Phi(p, 0)$ in $M \times M$, and since

$$d(\exp_p)_0 = \mathrm{id},$$

it is easy to check that in the basis of $T_pM \times T_pM$ consisting of the pairs

$$\left(\left(\frac{\partial}{\partial x_1}\right)_p, 0\right), \ldots, \left(\left(\frac{\partial}{\partial x_n}\right)_p, 0\right), \left(0, \left(\frac{\partial}{\partial x_1}\right)_p\right), \ldots, \left(0, \left(\frac{\partial}{\partial x_n}\right)_p\right),$$

we have

$$\begin{aligned} d\Phi_{(p,0)}\left(\left(\frac{\partial}{\partial x_i}\right)_p, 0\right) &= \left(\left(\frac{\partial}{\partial x_i}\right)_p, \left(\frac{\partial}{\partial x_i}\right)_p\right) \\ d\Phi_{(p,0)}\left(0, \left(\frac{\partial}{\partial x_i}\right)_p\right) &= \left(0, \left(\frac{\partial}{\partial x_i}\right)_p\right), \end{aligned}$$

so the Jacobian matrix of $\Phi_{(p,0)}$ is equal to

$$\begin{pmatrix} I & 0 \\ I & I \end{pmatrix}.$$

By the inverse function theorem, there is an open subset $\mathcal{U}'$ contained in $\mathcal{U}$ with $(p, 0) \in \mathcal{U}'$ and an open subset $\mathcal{W}'$ of $M \times M$ containing (p, p) such that Φ is a diffeomorphism between $\mathcal{U}'$ and $\mathcal{W}'$. We may assume that there is some open subset U' of U containing p and some $\epsilon > 0$ such that $\epsilon < \epsilon_1$ and

$$\mathcal{U}' = \{(q, v) \mid q \in U', v \in T_qM, \|v\| < \epsilon\} = \bigcup_{q \in U'} \{q\} \times B(0, \epsilon).$$

Now, if we choose a smaller open subset W containing p such that $W \times W \subseteq \mathcal{W}'$, because Φ is a diffeomorphism on $\mathcal{U}'$, we have

$$\{q\} \times W \subseteq \Phi(\{q\} \times B(0, \epsilon)),$$

for all $q \in W$. From the definition of Φ, we have $W \subseteq \exp_q(B(0, \epsilon))$, and $\exp_q$ is a diffeomorphism on $B(0, \epsilon) \subseteq T_qM$, which proves Part (3).

Given any two points $q_1, q_2 \in W$, since Φ is a diffeomorphism between $\mathcal{U}'$ and $\mathcal{W}'$ with $W \times W \subseteq \mathcal{W}'$, there is a unique $v \in T_{q_1}M$ such that $\|v\| < \epsilon$ and $\Phi(q_1, v) = (q_1, q_2)$; that is, $\exp_{q_1}(v) = q_2$, which means that $t \mapsto \exp_{q_1}(tv)$ is the unique geodesic from q_1 to q_2, which proves (1).

Finally, since $(q_1, v) = \Phi^{-1}(q_1, q_2)$ and Φ is a diffeomorphism, Part (2) holds. □

Remark: Except for the part of Statement (1) about the length of geodesics having length $< \epsilon$, Proposition 15.5 holds for a Riemannian manifold equipped with an arbitrary connection.

Definition 15.8. Let (M, g) be a Riemannian manifold. For any $q \in M$, an open neighborhood of q of the form $U_q = \exp_q(B(0, \epsilon))$ where $\exp_q$ is a diffeomorphism from the open ball $B(0, \epsilon)$ onto U_q, is called a *normal neighborhood*.

Remark: The proof of the previous proposition can be sharpened to prove that for any $p \in M$, there is some $\beta > 0$ such that any two points $q_1, q_2 \in \exp(B(0, \beta))$, there is a unique geodesic from q_1 to q_2 that stays within $\exp(B(0, \beta))$; see Do Carmo [39] (Chapter 3, Proposition 4.2). We say that $\exp(B(0, \beta))$ is *strongly convex*. The least upper bound of these β is called the *convexity radius* at p.

Definition 15.9. Let (M, g) be a Riemannian manifold. For every point $p \in M$, the *injectivity radius of M at p*, denoted $i(p)$, is the least upper bound of the numbers $r > 0$ such that $\exp_p$ is a diffeomorphism on the open ball $B(0, r) \subseteq T_pM$. The *injectivity radius $i(M)$ of M* is the greatest lower bound of the numbers $i(p)$, where $p \in M$.

Definition 15.10. Let (M, g) be a Riemannian manifold. For every $p \in M$, we get a chart (U_p, φ), where $U_p = \exp_p(B(0, i(p)))$ and $\varphi = \exp^{-1}$, called a *normal chart*. If we pick any orthonormal basis $(e_1, \ldots, e_n)$ of T_pM, then the x_i's, with $x_i = pr_i \circ \exp^{-1}$ and pr_i the projection onto $\mathbb{R}e_i$, are called *normal coordinates* at p (here, $n = \dim(M)$).

Normal coordinates are defined up to an isometry of T_pM. The following proposition shows that Riemannian metrics do not admit any local invariants of order one. The proof is left as an exercise.

Proposition 15.6. *Let (M, g) be a Riemannian manifold. For every point $p \in M$, in normal coordinates at p,*

$$g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)_p = \delta_{ij} \qquad \text{and} \qquad \Gamma^k_{ij}(p) = 0.$$

The need to consider vector fields along a surface and the partial derivatives of such vector fields arise in several proofs to be presented shortly.

Definition 15.11. If $\alpha\colon U \to M$ is a parametrized surface, where M is a smooth manifold and U is some open subset of $\mathbb{R}^2$, we say that a vector field $V \in \mathfrak{X}(M)$ is a *vector field along* α iff $V(x, y) \in T_{\alpha(x,y)}M$, for all $(x, y) \in U$.

For any smooth vector field V along α, we also define the covariant derivatives $DV/\partial x$ and $DV/\partial y$ as follows. For each fixed y_0, if we restrict V to the curve

$$x \mapsto \alpha(x, y_0)$$

we obtain a vector field V_{y_0} along this curve, and we set

$$\frac{DV}{\partial x}(x, y_0) = \frac{DV_{y_0}}{dx}.$$

Then we let y_0 vary so that $(x, y_0) \in U$, and this yields $DV/\partial x$. We define $DV/\partial y$ is a similar manner, using a fixed x_0. The following technical result will be used several times.

Proposition 15.7. *For any smooth vector field V along a surface $\alpha\colon U \to M$, for any torsion-free connection on M, we have*

$$\frac{D}{\partial y}\frac{\partial \alpha}{\partial x} = \frac{D}{\partial x}\frac{\partial \alpha}{\partial y}.$$

The above equation is checked in a coordinate system. The details of the computation are given in Do Carmo [39] (Chapter 3, Lemma 3.4).

For the next proposition known as *Gauss Lemma*, we need to define *polar coordinates* on T_pM. If $n = \dim(M)$, observe that the map $(0, \infty) \times S^{n-1} \longrightarrow T_pM - \{0\}$ given by

$$(r, v) \mapsto rv, \qquad r > 0,\ v \in S^{n-1}$$

is a diffeomorphism, where S^{n-1} is the sphere of radius $r = 1$ in T_pM. Then the map $(0, i(p)) \times S^{n-1} \longrightarrow U_p - \{p\} \subset M$ given by

$$(r, v) \mapsto \exp_p(rv), \qquad 0 < r < i(p),\ v \in S^{n-1}$$

is also a diffeomorphism.

Proposition 15.8. *(Gauss Lemma) Let (M, g) be a Riemannian manifold. For every point $p \in M$, the images $\exp_p(S(0, r))$ of the spheres $S(0, r) \subseteq T_pM$ centered at 0 by the exponential map $\exp_p$ are orthogonal to the radial geodesics $r \mapsto \exp_p(rv)$ through p for all $r < i(p)$, with $v \in S^{n-1}$. This means that for any differentiable curve $t \mapsto v(t)$ on the unit sphere S^{n-1}, the corresponding curve on M*

$$t \mapsto \exp_p(rv(t)) \quad \text{with } r \text{ fixed},$$

is orthogonal to the radial geodesic

$$r \mapsto \exp_p(rv(t)) \quad \text{with } t \text{ fixed } (0 < r < i(p)).$$

See Figure 15.4. Furthermore, in polar coordinates, the pull-back metric $\exp^* g$ induced on T_pM is of the form

$$\exp^* g = dr^2 + g_r,$$

where g_r is a metric on the unit sphere S^{n-1}, with the property that g_r/r^2 converges to the standard metric on S^{n-1} (induced by $\mathbb{R}^n$) when r goes to zero (here, $n = \dim(M)$).

Proof sketch. We follow Milnor; see [81], Chapter II, Section 10. Pick any curve $t \mapsto v(t)$ on the unit sphere S^{n-1}. The first statement can be restated in terms of the parametrized surface

$$f(r, t) = \exp_p(rv(t));$$

we must prove that

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0,$$

for all (r, t). However, as we are using the Levi-Civita connection, which is compatible with the metric, we have

$$\frac{\partial}{\partial r} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = \left\langle \frac{D}{\partial r} \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle + \left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial r} \frac{\partial f}{\partial t} \right\rangle. \tag{$\dagger$}$$

The first expression on the right-hand side of (†) is zero since the curves

$$r \mapsto f(r, t)$$

are geodesics. For the second expression, first observe that

$$\left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial t} \frac{\partial f}{\partial r} \right\rangle = \frac{1}{2} \frac{\partial}{\partial t} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial r} \right\rangle = 0,$$

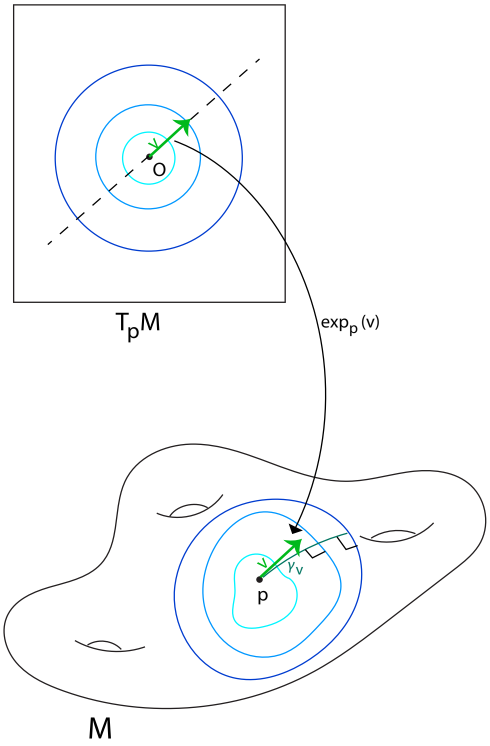


Figure 15.4: An illustration of the Gauss lemma for a two-dimensional manifold.

since $1 = \|v(t)\| = \|\partial f/\partial r\|$, since the velocity vector of a geodesic has constant norm (this fact was noted just after Definition 15.4). Next, note that if we can prove that

$$\frac{D}{\partial t}\frac{\partial f}{\partial r} = \frac{D}{\partial r}\frac{\partial f}{\partial t},$$

then

$$0 = \left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial t}\frac{\partial f}{\partial r} \right\rangle = \left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial r}\frac{\partial f}{\partial t} \right\rangle,$$

so the second expression on the right-hand side of (†) is also zero. Since the Levi-Civita connection is torsion-free the equation

$$\frac{D}{\partial t}\frac{\partial f}{\partial r} = \frac{D}{\partial r}\frac{\partial f}{\partial t}$$

follows from Proposition 15.7.

Since the right-hand side of (†) is zero,

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle$$

is independent of r. But, for $r = 0$, we have

$$f(0, t) = \exp_p(0) = p,$$

hence

$$\frac{\partial f}{\partial t}(0, t) = 0$$

and thus,

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0$$

for all r, t, which concludes the proof of the first statement.

The orthogonality of $\partial f/\partial r$ and $\partial f/\partial t$ implies that the pullback metric $\exp^* g$ induced on T_pM is of the form $\exp^* g = dr^2 + g_r$, where g_r is a metric on the unit sphere S^{n-1}. For the proof that g_r/r^2 converges to the standard metric on S^{n-1}, see Pansu's class notes, Chapter 3, Section 3.5. □

Observe that the proof of Gauss Lemma (Proposition 15.8) uses the fact that the connection is compatible with the metric and torsion-free.

Remark: If $v(t)$ is a curve on S^{n-1} such that $v(0) = v$ and $v'(0) = w_N$ (with $\|v\| < i(p)$), then since $f(r, t) = \exp_p(rv(t))$

$$\frac{\partial f}{\partial r}(1, 0) = (d\exp_p)_v(v), \quad \frac{\partial f}{\partial t}(1, 0) = (d\exp_p)_v(w_N),$$

and Gauss lemma can be stated as

$$\langle (d\exp_p)_v(v), (d\exp_p)_v(w_N)\rangle = \langle v, w_N\rangle = 0.$$

This is how Gauss lemma is stated in Do Carmo [39] (Chapter 3, Lemma 3.5).

Remark: There is also another version of "Gauss lemma" whose proof uses Jacobi fields (see Gallot, Hulin and Lafontaine [49], Chapter 3, Lemma 3.70).

Proposition 15.9. *(Gauss Lemma) Given any point $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$\langle d(\exp_p)_{tv}(u), d(\exp_p)_{tv}(v)\rangle = \langle u, v\rangle, \qquad 0 \leq t \leq 1.$$

The next three results use the fact that the connection is compatible with the metric and torsion-free. Consider any piecewise smooth curve

$$\omega \colon [a, b] \to U_p - \{p\} \subset M.$$

We can write each point $\omega(t)$ uniquely as

$$\omega(t) = \exp_p(r(t)v(t)),$$

with $0 < r(t) < i(p)$, $v(t) \in T_pM$ and $\|v(t)\| = 1$.

Proposition 15.10. *Let (M, g) be a Riemannian manifold. We have*

$$\int_a^b \|\omega'(t)\| \; dt \geq |r(b) - r(a)|,$$

where equality holds only if the function r is monotone and the function v is constant. Thus, the shortest path joining two concentric spherical shells $\exp_p(S(0, r(a)))$ and $\exp_p(S(0, r(b)))$ is a radial geodesic.

Proof. (After Milnor, see [81], Chapter II, Section 10.) Again, let $f(r, t) = \exp_p(rv(t))$ so that $\omega(t) = f(r(t), t)$. Then,

$$\frac{d\omega}{dt} = \frac{\partial f}{\partial r} r'(t) + \frac{\partial f}{\partial t}.$$

The proof of the previous proposition showed that the two vectors on the right-hand side are orthogonal and since $\|\partial f/\partial r\| = 1$, this gives

$$\left\|\frac{d\omega}{dt}\right\|^2 = |r'(t)|^2 + \left\|\frac{\partial f}{\partial t}\right\|^2 \geq |r'(t)|^2$$

where equality holds only if $\partial f/\partial t = 0$; hence only if $v'(t) = 0$. Thus,

$$\int_a^b \left\|\frac{d\omega}{dt}\right\| \; dt \geq \int_a^b |r'(t)| dt \geq |r(b) - r(a)|$$

where equality holds only if $r(t)$ is monotone and $v(t)$ is constant. □

We now get the following important result from Proposition 15.8 and Proposition 15.10, namely that geodesics are locally lengthwise minimizing curves.

Theorem 15.11. *Let (M, g) be a Riemannian manifold. Let W and ϵ be as in Proposition 15.5 and let $\gamma\colon [0, 1] \to M$ be the geodesic of length $< \epsilon$ joining two points q_1, q_2 of W. For any other piecewise smooth path ω joining q_1 and q_2, we have*

$$\int_0^1 \|\gamma'(t)\| \, dt \leq \int_0^1 \|\omega'(t)\| \, dt,$$

where equality holds only if the images $\omega([0, 1])$ and $\gamma([0, 1])$ coincide. Thus, γ is the shortest path from q_1 to q_2.

Proof. (After Milnor, see [81], Chapter II, Section 10.) Consider any piecewise smooth path ω from $q_1 = \gamma(0)$ to some point

$$q_2 = \exp_{q_1}(rv) \in U_{q_1},$$

where $0 < r < \epsilon$ and $\|v\| = 1$. Then for any δ with $0 < \delta < r$, the path ω must contain a segment joining the spherical shell of radius δ to the spherical shell of radius r, and lying between these two shells. The length of this segment will be at least $r - \delta$; hence if we let δ go to zero, the length of ω will be at least r. If $\omega([0, 1]) \neq \gamma([0, 1])$, we easily obtain a strict inequality. □

Here is an important consequence of Theorem 15.11.

Corollary 15.12. *Let (M, g) be a Riemannian manifold. If $\omega\colon [0, b] \to M$ is any curve parametrized by arc-length and ω has length less than or equal to the length of any other curve from $\omega(0)$ to $\omega(b)$, then ω is a geodesic.*

Proof. Consider any segment of ω lying within an open set W as above, and having length $< \epsilon$. By Theorem 15.11, this segment must be a geodesic. Hence, the entire curve is a geodesic. □

Corollary 15.12 together with the fact that isometries preserve geodesics can be used to determine the geodesics in various spaces, for example in the Poincaré half-plane.

Definition 15.12. Let (M, g) be a Riemannian manifold. A geodesic $\gamma\colon [a, b] \to M$ is *minimal* iff its length is less than or equal to the length of any other piecewise smooth curve joining its endpoints.

Theorem 15.11 asserts that any sufficiently small segment of a geodesic is minimal. On the other hand, a long geodesic may not be minimal. For example, a great circle arc on the unit sphere is a geodesic. If such an arc has length greater than π, then it is not minimal. This is illustrated by the magenta equatorial geodesic connecting points a and b of Figure 15.5 (i.). Minimal geodesics are generally not unique. For example, any two antipodal points on a sphere are joined by an infinite number of minimal geodesics. Figure 15.5 (ii.) illustrates five geodesics connecting the antipodal points a and b.

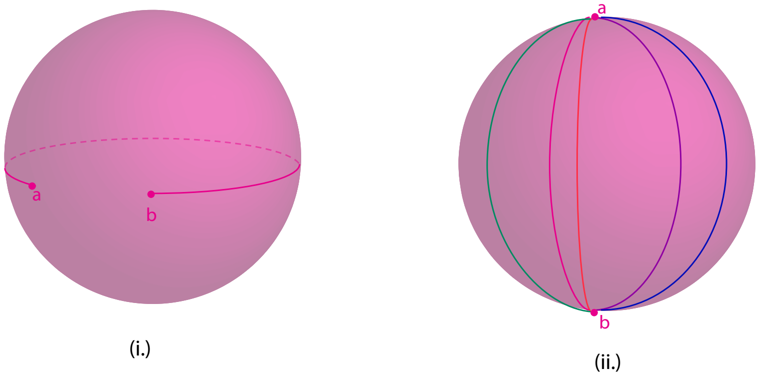


Figure 15.5: Examples of geodesics, i.e. arcs of great circles, on S^2.

A *broken geodesic* is a piecewise smooth curve as in Definition 15.1, where each curve segment is a geodesic.

Proposition 15.13. *A Riemannian manifold (M, g) is connected iff any two points of M can be joined by a broken geodesic.*

Proof. Assume M is connected, pick any $p \in M$, and let $S_p \subseteq M$ be the set of all points that can be connected to p by a broken geodesic. For any $q \in M$, choose a normal neighborhood U of q. If $q \in S_p$, then it is clear that $U \subseteq S_p$. On the other hand, if $q \notin S_p$, then $U \subseteq M - S_p$. Therefore, $S_p \neq \emptyset$ is open and closed, so $S_p = M$. The converse is obvious. □

Remark: Proposition 15.13 holds for a smooth manifold equipped with any connection.

In general, if M is connected, then it is not true that any two points are joined by a geodesic. However, this will be the case if M is a geodesically complete Riemannian manifold equipped with the Levi-Civita connection, as we will see in the next section.

Next we will see that a Riemannian metric induces a distance on the manifold whose induced topology agrees with the original metric.

15.3 Complete Riemannian Manifolds, the Hopf-Rinow Theorem and the Cut Locus

Every connected Riemannian manifold (M, g) is a metric space in a natural way. Furthermore, M is a complete metric space iff M is geodesically complete. In this section, we explore briefly some properties of complete Riemannian manifolds equipped with the Levi-Civita connection.

Proposition 15.14. *Let (M, g) be a connected Riemannian manifold. For any two points $p, q \in M$, let $d(p, q)$ be the greatest lower bound of the lengths of all piecewise smooth curves joining p to q. Then d is a metric on M, and the topology of the metric space (M, d) coincides with the original topology of M.*

A proof of the above proposition can be found in Gallot, Hulin and Lafontaine [49] (Chapter 2, Proposition 2.91) or O'Neill [91] (Chapter 5, Proposition 18).

The distance d is often called the *Riemannian distance* on M. For any $p \in M$ and any $\epsilon > 0$, the *metric ball of center p and radius ϵ* is the subset $B_\epsilon(p) \subseteq M$ given by

$$B_\epsilon(p) = \{q \in M \mid d(p, q) < \epsilon\}.$$

The next proposition follows easily from Proposition 15.5 (Milnor [81], Section 10, Corollary 10.8).

Proposition 15.15. *Let (M, g) be a connected Riemannian manifold. For any compact subset $K \subseteq M$, there is a number $\delta > 0$ so that any two points $p, q \in K$ with distance $d(p, q) < \delta$ are joined by a unique geodesic of length less than δ. Furthermore, this geodesic is minimal and depends smoothly on its endpoints.*

Recall from Definition 15.7 that (M, g) is geodesically complete iff the exponential map $v \mapsto \exp_p(v)$ is defined for all $p \in M$ and for all $v \in T_pM$. We now prove the following important theorem due to Hopf and Rinow (1931).

Theorem 15.16. *(Hopf-Rinow) Let (M, g) be a connected Riemannian manifold. If there is a point $p \in M$ such that $\exp_p$ is defined on the entire tangent space T_pM, then any point $q \in M$ can be joined to p by a minimal geodesic. As a consequence, if M is geodesically complete, then any two points of M can be joined by a minimal geodesic.*

Proof. We follow Milnor's proof in [81], Chapter 10, Theorem 10.9. Pick any two points $p, q \in M$ and let $r = d(p, q)$. By Proposition 15.5, there is some open subset $W \subseteq M$, with $p \in W$, and some $\epsilon > 0$, such that the exponential map is a diffeomorphism between the open ball $B(0, \epsilon)$ and its image $U_p = \exp_p(B(0, \epsilon))$. For $\delta < \min(\epsilon, r)$, let $S = \exp_p(S(0, \delta))$, where $S(0, \delta)$ is the sphere of radius δ. By Proposition 15.5, there is a unique geodesic from p to any point $s \in S$, and since the length of this geodesic is δ, we have $d(p, s) = \delta$ for all $s \in S$. Since $S \subseteq U_p$ is compact, there is some point

$$p_0 = \exp_p(\delta v), \qquad \text{with } \|v\| = 1,$$

on S for which the distance to q is minimized. We will prove that

$$\exp_p(rv) = q,$$

which will imply that the geodesic γ given by $\gamma(t) = \exp_p(tv)$ is actually a minimal geodesic from p to q (with $t \in [0, r]$). Here we use the fact that the exponential $\exp_p$ is defined everywhere on T_pM. See Figure 15.6.

The proof amounts to showing that a point which moves along the geodesic γ must get closer and closer to q. In fact, for each $t \in [\delta, r]$, we prove

$$d(\gamma(t), q) = r - t. \qquad (*_t)$$

We get the proof by setting $t = r$.

First we prove $(*_\delta)$. Every path from p to q must pass through S, because $\gamma([0, r])$ is a connected set which must intersect the boundary S of $\exp_p(B(0, \delta))$. Otherwise, since p is in the interior of $\exp_p(B(0, \delta))$ and q is in the exterior of $\exp_p(B(0, \delta))$, the subset $\gamma([0, r])$ would intersect the interior and the exterior of $\exp_p(B(0, \delta))$, contradicting the fact that $\gamma([0, r])$ is connected. By the choice of p_0 as a point on S minimizing the distance from S to q, we have

$$r = d(p, q) = \min_{s \in S}\{d(p, s) + d(s, q)\} = \delta + \min_{s \in S}\{d(s, q)\} = \delta + d(p_0, q).$$

Therefore, $d(p_0, q) = r - \delta$, and since $p_0 = \gamma(\delta)$, this proves $(*_\delta)$.

Define $t_0 \in [\delta, r]$ by

$$t_0 = \sup\{t \in [\delta, r] \mid d(\gamma(t), q) = r - t\}.$$

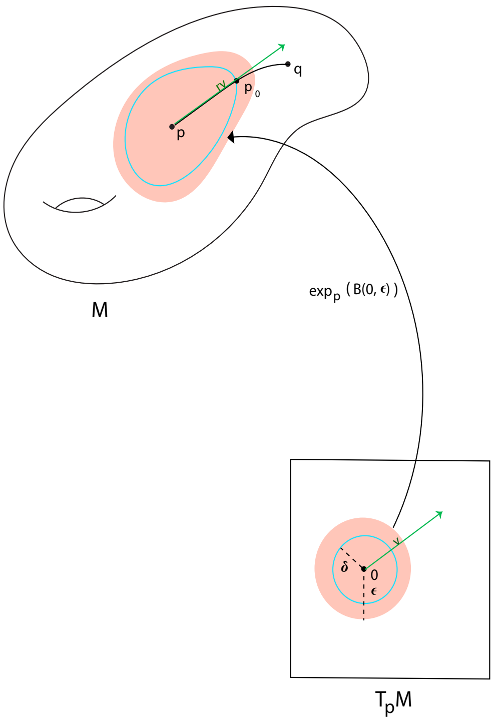


Figure 15.6: An illustration of the first paragraph in the proof of Theorem 15.16.

As the set $\{t \in [\delta, r] \mid d(\gamma(t), q) = r - t\}$ is closed because the curve γ and the distance function d are continuous, it contains its upper bound t_0, so the equation $(*_{t_0})$ also holds. We claim that if $t_0 < r$, then we obtain a contradiction.

As we did with p, we reapply Proposition 15.5 to find some small $\delta' > 0$ so that if $S' = \exp_{\gamma(t_0)}(B(0, \delta'))$, then there is some point p_0' on S' with minimum distance from q and p_0' is joined to $\gamma(t_0)$ by a minimal geodesic. See Figure 15.7.

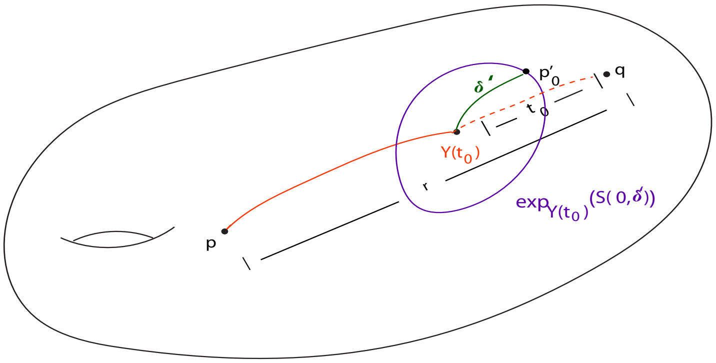


Figure 15.7: The construction of p_0' in Theorem 15.16.

We have

$$r - t_0 = d(\gamma(t_0), q) = \min_{s \in S'}\{d(\gamma(t_0), s) + d(s, q)\} = \delta' + \min_{s \in S'}\{d(s, q)\} = \delta' + d(p_0', q),$$

hence

$$d(p_0', q) = r - t_0 - \delta'. \tag{$\dagger$}$$

We claim that $p_0' = \gamma(t_0 + \delta')$.

By the triangle inequality and using (†) (recall that $d(p, q) = r$), we have

$$d(p, p_0') \geq d(p, q) - d(p_0', q) = t_0 + \delta'.$$

But a path of length precisely $t_0 + \delta'$ from p to p_0' is obtained by following γ from p to $\gamma(t_0)$, and then following a minimal geodesic from $\gamma(t_0)$ to p_0'. Since this broken geodesic has minimal length, by Corollary 15.12, it is a genuine (unbroken) geodesic, and so it coincides with γ. But then, as $p_0' = \gamma(t_0 + \delta')$, equality (†) becomes $(*_{t_0+\delta'})$, namely

$$d(\gamma(t_0 + \delta'), q) = r - (t_0 + \delta'),$$

contradicting the maximality of t_0. Therefore, we must have $t_0 = r$, and $q = \exp_p(rv)$, as desired. □

Remark: Theorem 15.16 is proved in nearly every book on Riemannian geometry. Among those, we mention Gallot, Hulin and Lafontaine [49] (Chapter 2, Theorem 2.103), Do Carmo [39] (Chapter 7, Theorem 2.8), and O'Neill [91] (Chapter 5, Lemma 24). Since the proof of Theorem 15.16 makes crucial use of Corollary 15.12, which itself relies on the fact that the connection is symmetric and torsion-free, Theorem 15.16 only holds for the Levi-Civita connection.

Theorem 15.16 implies the following result (often known as the *Hopf-Rinow Theorem*).

Theorem 15.17. *Let (M, g) be a connected, Riemannian manifold. The following statements are equivalent:*

(1) The manifold (M, g) is geodesically complete; that is, for every $p \in M$, every geodesic through p can be extended to a geodesic defined on all of $\mathbb{R}$.

(2) For every point $p \in M$, the map $\exp_p$ is defined on the entire tangent space T_pM.

(3) There is a point $p \in M$, such that $\exp_p$ is defined on the entire tangent space T_pM.

(4) Any closed and bounded subset of the metric space (M, d) is compact.

(5) The metric space (M, d) is complete (that is, every Cauchy sequence converges).

Proof. Proofs of Theorem 15.17 can be found in Gallot, Hulin and Lafontaine [49] (Chapter 2, Corollary 2.105), Do Carmo [39] (Chapter 7, Theorem 2.8), and O'Neill [91] (Chapter 5, Theorem 21).

The implications (1) $\Rightarrow$ (2) and (2) $\Rightarrow$ (3) are obvious. We prove the implication (3) $\Rightarrow$ (4) as follows. Let A be a closed and bounded subset of M. Since A is bounded it is contained in a metric ball B with center p. By Theorem 15.16, (3) implies that there is a minimal geodesic from p to any point in B, so there is an open ball $B(0, r) \subseteq T_pM$ such that $B \subseteq \exp_p(\overline{B(0, r)})$. Since $\exp_p$ is continuous and $\overline{B(0, r)}$ is compact, $\exp_p(\overline{B(0, r)})$ is compact. Since A is closed and $A \subseteq B \subseteq \exp_p(\overline{B(0, r)})$, with $\exp_p(\overline{B(0, r)})$ compact, A itself is compact.

Assume that (4) holds. If (p_m) is any Cauchy sequence in M, then it is a bounded subset, hence contained in a compact ball. Thus the sequence (p_m) contains a convergent subsequence, and since it is a Cauchy sequence, it converges. Therefore the implication (4) $\Rightarrow$ (5) holds.

Finally, we prove the implication (5) $\Rightarrow$ (1). Let $\gamma\colon I \to M$ be a geodesic in M parametrized by arc length. If we prove that I is open and closed, then $I = \mathbb{R}$ and we are done. The fact that I is open follows from Proposition 15.1. Next let (t_n) be a sequence of elements of I converging to some number t. We would like to prove that $t \in I$. Since

$$d(\gamma(t_i), \gamma(t_j)) \leq |t_i - t_j|,$$

the sequence $(\gamma(t_n))$ is a Cauchy sequence, so by (5) it converges to some element $q \in M$. Let W be the open subset containing q and let ϵ given by Proposition 15.5, so that any geodesic starting from any point in W is defined on $(-\epsilon, \epsilon)$. By chosing n large enough so that $|t_n - t| < \epsilon/2$ and $\gamma(t_n) \in W$, we see that the geodesic γ is defined up to $t + \epsilon/2$, so $t \in I$, as desired. Therefore I is closed, and the proof is complete. □

In view of Theorem 15.17, a connected Riemannian manifold (M, g) is geodesically complete iff the metric space (M, d) is complete. We will refer simply to M as a *complete Riemannian manifold* (it is understood that M is connected). Also, by (4), every compact Riemannian manifold is complete. If we remove any point p from a Riemannian manifold M, then $M - \{p\}$ is not complete, since every geodesic that formerly went through p yields a geodesic that can't be extended.

Definition 15.13. Let (M, g) be a complete Riemannian manifold. Given any point $p \in M$, let $\mathcal{U}_p \subseteq T_pM$ be the subset consisting of all $v \in T_pM$ such that the geodesic

$$t \mapsto \exp_p(tv)$$

is a minimal geodesic up to $t = 1 + \epsilon$, for some $\epsilon > 0$. The left-over part $M - \exp_p(\mathcal{U}_p)$ (if nonempty) is actually equal to $\exp_p(\partial\mathcal{U}_p)$, and it is an important subset of M called the *cut locus of* p.

Remark: It can be shown that the subset $\mathcal{U}_p$ is open and star-shaped, and it turns out that $\exp_p$ is a diffeomorphism from $\mathcal{U}_p$ onto its image $\exp_p(\mathcal{U}_p)$ in M.

The following proposition is needed to establish properties of the cut locus.

Proposition 15.18. *Let (M, g) be a complete Riemannian manifold. For any geodesic $\gamma\colon [0, a] \to M$ from $p = \gamma(0)$ to $q = \gamma(a)$, the following properties hold:*

(i) If there is no geodesic shorter than γ between p and q, then γ is minimal on $[0, a]$.

(ii) If there is another geodesic of the same length as γ between p and q, then γ is no longer minimal on any larger interval, $[0, a + \epsilon]$.

(iii) If γ is minimal on any interval I, then γ is also minimal on any subinterval of I.

Proof. Part (iii) is an immediate consequence of the triangle inequality. As M is complete, by the Hopf-Rinow Theorem, (Theorem 15.16), there is a minimal geodesic from p to q, so γ must be minimal too. This proves Part (i). For Part (ii), assume that ω is another geodesic from p to q of the same length as γ and that γ is defined in $[0, a+\epsilon]$ some some $\epsilon > 0$. Since γ and ω are assumed to be distinct curves, the curve $\varphi\colon [0, a + \epsilon] \to M$ given by

$$\varphi(t) = \begin{cases} \omega(t) & 0 \leq t \leq a \\ \gamma(t) & a \leq t \leq a + \epsilon \end{cases}$$

is not smooth at $t = a$, since otherwise Proposition 15.1 implies that γ and ω would be equal on their common domain; in particular, Proposition 15.1 implies there is a unique geodesic through q with initial condition $v = \gamma'(a) = \omega'(a)$. Pick ϵ' so that $0 < \epsilon' < \min\{\epsilon, a\}$, and consider the points $q_1 = \varphi(a - \epsilon')$ and $q_2 = \varphi(a + \epsilon')$. By Hopf-Rinow's theorem, there is a minimal geodesic ψ from q_1 to q_2, and since the portion of φ from q_1 to q_2 is not smooth, the length of ψ is strictly smaller than the length of the segment of φ from q_1 to q_2. But then, the curve $\widetilde{\varphi}$ obtained by concatenating the segment of ω from p to q_1 and ψ from q_1 to q_2 is strictly shorter that the curve obtained by concatenating the curve segment ω from p to q with the curve segment γ from q to q_2. See Figure 15.8.

However, the length of the curve segment ω from p to q is equal to length of the curve segment γ from p to q. This proves that $\widetilde{\varphi}$ from p to q_2 is strictly shorter than γ from p to q_2, so γ is no longer minimal beyond q. □

Again, assume (M, g) is a complete Riemannian manifold and let $p \in M$ be any point. For every $v \in T_pM$, let

$$I_v = \{s \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic} \quad t \mapsto \exp_p(tv) \quad \text{is minimal on } [0, s]\}.$$

It is easy to see that I_v is a closed interval, so $I_v = [0, \rho(v)]$ (with $\rho(v)$ possibly infinite). It can be shown that if $w = \lambda v$, then $\rho(v) = \lambda\rho(w)$, so we can restrict our attention to unit vectors v. It can also be shown that the map $\rho\colon S^{n-1} \to \mathbb{R}$ is continuous, where S^{n-1} is

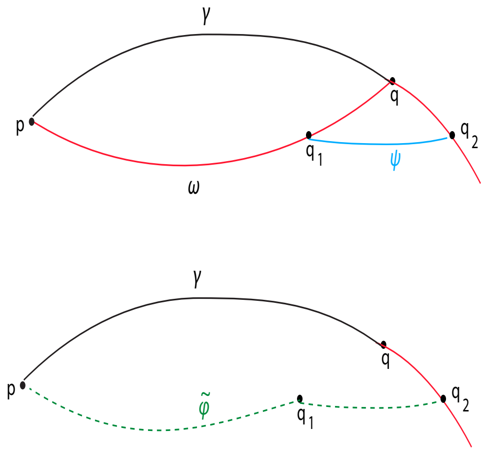


Figure 15.8: The geodesics ω, γ, ψ, and the path $\widetilde{\varphi}$ used in the proof of Proposition 15.18.

the unit sphere of center 0 in T_pM, and that $\rho(v)$ is bounded below by a strictly positive number.

By using $\rho(v)$, we are able to restate Definition 15.13 as follows:

Definition 15.14. Let (M, g) be a complete Riemannian manifold and let $p \in M$ be any point. Define $\mathcal{U}_p$ by

$$\mathcal{U}_p = \left\{ v \in T_pM \;\middle|\; \rho\left(\frac{v}{\|v\|}\right) > \|v\| \right\} = \{v \in T_pM \mid \rho(v) > 1\},$$

and the *cut locus* of p by

$$\mathrm{Cut}(p) = \exp_p(\partial\mathcal{U}_p) = \{\exp_p(\rho(v)v) \mid v \in S^{n-1}\}.$$

The set $\mathcal{U}_p$ is open and star-shaped. The boundary $\partial\mathcal{U}_p$ of $\mathcal{U}_p$ in T_pM is sometimes called the *tangential cut locus* of p and is denoted $\widetilde{\mathrm{Cut}}(p)$.

Remark: The cut locus was first introduced for convex surfaces by Poincaré (1905) under the name *ligne de partage*. According to Do Carmo [39] (Chapter 13, Section 2), for Riemannian manifolds, the cut locus was introduced by J.H.C. Whitehead (1935). But it was Klingenberg (1959) who revived the interest in the cut locus and showed its usefuleness.

Proposition 15.19. *Let (M, g) be a complete Riemannian manifold. For any point $p \in M$, the sets $\exp_p(\mathcal{U}_p)$ and $\mathrm{Cut}(p)$ are disjoint and*

$$M = \exp_p(\mathcal{U}_p) \cup \mathrm{Cut}(p).$$

Proof. From the Hopf-Rinow Theorem, (Theorem 15.16), for every $q \in M$, there is a minimal geodesic $t \mapsto \exp_p(vt)$ such that $\exp_p(v) = q$. This shows that $\rho(v) \geq 1$, so $v \in \overline{\mathcal{U}_p}$ and

$$M = \exp_p(\mathcal{U}_p) \cup \mathrm{Cut}(p).$$

It remains to show that this is a disjoint union. Assume $q \in \exp_p(\mathcal{U}_p) \cap \mathrm{Cut}(p)$. Since $q \in \exp_p(\mathcal{U}_p)$, there is a geodesic γ such that $\gamma(0) = p$, $\gamma(a) = q$, and γ is minimal on $[0, a + \epsilon]$, for some $\epsilon > 0$. On the other hand, as $q \in \mathrm{Cut}(p)$, there is some geodesic $\widetilde{\gamma}$ with $\widetilde{\gamma}(0) = p$, $\widetilde{\gamma}(b) = q$, $\widetilde{\gamma}$ minimal on $[0, b]$, but $\widetilde{\gamma}$ not minimal after b. As γ and $\widetilde{\gamma}$ are both minimal from p to q, they have the same length from p to q. But then, as γ and $\widetilde{\gamma}$ are distinct, by Proposition 15.18 (ii), the geodesic γ can't be minimal after q, a contradiction. $\square$

We can now restate Definition 15.9 as follows:

Definition 15.15. Let (M, g) be a complete Riemannian manifold and let $p \in M$ be any point. The *injectivity radius $i(p)$ of M at p* is equal to the distance from p to the cut locus of p:

$$i(p) = d(p, \mathrm{Cut}(p)) = \inf_{q \in \mathrm{Cut}(p)} d(p, q).$$

Consequently, the *injectivity radius $i(M)$ of M* is given by

$$i(M) = \inf_{p \in M} d(p, \mathrm{Cut}(p)).$$

If M is compact, it can be shown that $i(M) > 0$. It can also be shown using Jacobi fields that $\exp_p$ is a diffeomorphism from $\mathcal{U}_p$ onto its image $\exp_p(\mathcal{U}_p)$. Thus, $\exp_p(\mathcal{U}_p)$ is diffeomorphic to an open ball in $\mathbb{R}^n$ (where $n = \dim(M)$) and the cut locus is closed. Hence, the manifold M is obtained by gluing together an open n-ball onto the cut locus of a point. In some sense the topology of M is "contained" in its cut locus.

Given any sphere S^{n-1}, the cut locus of any point p is its antipodal point $\{-p\}$. For more examples, consult Gallot, Hulin and Lafontaine [49] (Chapter 2, Section 2C7), Do Carmo [39] (Chapter 13, Section 2) or Berger [14] (Chapter 6). In general, the cut locus is very hard to compute. In fact, even for an ellipsoid, the determination of the cut locus of an arbitrary point was a matter of conjecture for a long time. This conjecture was finally settled around 2011.

15.4 Convexity, Convexity Radius

Proposition 15.5 shows that if (M, g) is a Riemannian manifold, then for every point $p \in M$, there is an open subset $W \subseteq M$ with $p \in W$ and a number $\epsilon > 0$, so that any two points q_1, q_2 of W are joined by a unique geodesic of length $< \epsilon$. However, there is no guarantee that this unique geodesic between q_1 and q_2 stays inside W. Intuitively this says that W may not be convex.

The notion of convexity can be generalized to Riemannian manifolds, but there are some subtleties. In this short section we review various definitions of convexity found in the literature and state one basic result. Following Sakai [100] (Chapter IV, Section 5), we make the following definition.

Definition 15.16. Let $C \subseteq M$ be a nonempty subset of some Riemannian manifold M.

(1) The set C is called *strongly convex* iff for any two points $p, q \in C$, there exists a unique minimal geodesic γ from p to q in M and γ is contained in C.

(2) If for every point $p \in \overline{C}$, there is some $\epsilon(p) > 0$ so that $C \cap B_{\epsilon(p)}(p)$ is strongly convex, then we say that C is *locally convex* (where $B_{\epsilon(p)}(p)$ is the metric ball of center p and radius $\epsilon(p)$).

(3) The set C is called *totally convex* iff for any two points $p, q \in C$, all geodesics from p to q in M are contained in C.

It is clear that if C is strongly convex or totally convex, then C is locally convex. If M is complete and any two points are joined by a unique geodesic, then the three conditions of Definition 15.16 are equivalent.

Definition 15.17. For any $p \in M$, the *convexity radius at* p, denoted $r(p)$, is the least upper bound of the numbers $r > 0$ such that for any metric ball $B_\epsilon(q)$, if $B_\epsilon(q) \subseteq B_r(p)$, then $B_\epsilon(q)$ is strongly convex and every geodesic contained in $B_r(p)$ is a minimal geodesic joining its endpoints. The *convexity radius of* M, $r(M)$, is the greatest lower bound of the set $\{r(p) \mid p \in M\}$.

Note that it is possible that $r(M) = 0$ if M is not compact.

The following proposition proved in Sakai [100] (Chapter IV, Section 5, Theorem 5.3) shows that a metric ball with sufficiently small radius is strongly convex.

Proposition 15.20. *If M is a Riemannian manifold, then $r(p) > 0$ for every $p \in M$, and the map $p \mapsto r(p) \in \mathbb{R}_+ \cup \{\infty\}$ is continuous. Furthermore, if $r(p) = \infty$ for some $p \in M$, then $r(q) = \infty$ for all $q \in M$.*

That $r(p) > 0$ is also proved in Do Carmo [39] (Chapter 3, Section 4, Proposition 4.2). More can be said about the structure of connected locally convex subsets of M; see Sakai [100] (Chapter IV, Section 5).

Remark: The following facts are stated in Berger [14] (Chapter 6):

(1) If M is compact, then the convexity radius $r(M)$ is strictly positive.

(2) $r(M) \leq \frac{1}{2} i(M)$, where $i(M)$ is the injectivity radius of M.

Berger also points out that if M is compact, then the existence of a finite cover by convex balls can used to triangulate M. This method was proposed by Hermann Karcher (see Berger [14], Chapter 3, Note 3.4.5.3).

Besides the notion of the gradient of a function, there is also the notion of Hessian. Now that we have geodesics at our disposal, we also have a method to compute the Hessian, a task which is generally quite complex.

15.5 Hessian of a Function on a Riemannian Manifold

Given a smooth function $f\colon M \to \mathbb{R}$ on a Riemannian manifold M, recall from Definition 13.3 that the gradient $\mathrm{grad}\, f$ of f is the vector field uniquely defined by the condition

$$\langle (\mathrm{grad}\, f)_p, u\rangle_p = df_p(u) = u(f), \quad \text{for all } u \in T_pM \text{ and all } p \in M.$$

Definition 15.18. The *Hessian* $\mathrm{Hess}(f)$ (or $\nabla^2(f)$) of a function $f \in C^\infty(M)$ is defined by

$$\mathrm{Hess}(f)(X, Y) = X(Y(f)) - (\nabla_X Y)(f) = X(df(Y)) - df(\nabla_X Y),$$

for all vector fields $X, Y \in \mathfrak{X}(M)$.

Since ∇ is torsion-free, we get $\nabla_X Y(f) - \nabla_Y X(f) = [X, Y](f) = X(Y(f)) - Y(X(f))$, which in turn implies

$$\mathrm{Hess}(f)(X, Y) = X(Y(f)) - (\nabla_X Y)(f) = Y(X(f)) - (\nabla_Y X)(f) = \mathrm{Hess}(f)(Y, X),$$

which means that the Hessian is *symmetric*.

Proposition 15.21. *The Hessian is given by the equation*

$$\mathrm{Hess}(f)(X, Y) = \langle \nabla_X(\mathrm{grad}\, f), Y\rangle, \quad X, Y \in \mathfrak{X}(M).$$

Proof. We have

$$\begin{aligned} X(Y(f)) &= X(df(Y)) \\ &= X(\langle \mathrm{grad}\, f, Y\rangle) \\ &= \langle \nabla_X(\mathrm{grad}\, f), Y\rangle + \langle \mathrm{grad}\, f, \nabla_X Y\rangle \\ &= \langle \nabla_X(\mathrm{grad}\, f), Y\rangle + (\nabla_X Y)(f) \end{aligned}$$

which yields

$$\langle \nabla_X(\mathrm{grad}\, f), Y\rangle = X(Y(f)) - (\nabla_X Y)(f) = \mathrm{Hess}(f)(X, Y),$$

as claimed. □

In the simple case where $M = \mathbb{R}^n$ and the metric is the usual Euclidean inner product on $\mathbb{R}^n$, we can easily compute the Hessian of a function $f\colon \mathbb{R}^n \to \mathbb{R}$. For any two vector fields

$$X = \sum_{i=1}^n x_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{i=1}^n y_i \frac{\partial}{\partial x_i},$$

with $x_i, y_i \in \mathbb{R}$, we have $\nabla_X Y = dY(X) = 0$ (x_i, y_i are constants and the Levi-Civita connection induced by the Euclidean inner product is the flat connection), so $\mathrm{Hess}(f)(X, Y) = X(Y(f))$ and if we write $x^\top = (x_1, \ldots, x_n)^\top$ and $y^\top = (y_1, \ldots, y_n)^\top$, it is easy to see that

$$\mathrm{Hess}(f)_p(X, Y) = x^\top H_p\, y,$$

where H_p is the matrix

$$H_p = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(p)\right),$$

the usual Hessian matrix of the function f at p.

In the general case of a Riemanian manifold $(M, \langle -, -\rangle)$, given any function $f \in C^\infty(M)$, for any $p \in M$ and for any $u \in T_pM$, the value of the Hessian $\mathrm{Hess}(f)_p(u, u)$ can be computed using geodesics.

Proposition 15.22. *For any geodesic $\gamma\colon [0, \epsilon] \to M$ such that $\gamma(0) = p$ and $\gamma'(0) = u$, we have*

$$\mathrm{Hess}(f)_p(u, u) = \left.\frac{d^2}{dt^2} f(\gamma(t))\right|_{t=0}.$$

Proof. We have

$$\mathrm{Hess}(f)_p(u, u) = \gamma'(\gamma'(f)) - (\nabla_{\gamma'}\gamma')(f) = \gamma'(\gamma'(f)),$$

since $\nabla_{\gamma'}\gamma' = 0$ because γ is a geodesic, and

$$\gamma'(\gamma'(f)) = \gamma'(df(\gamma')) = \gamma'\left(\left.\frac{d}{dt} f(\gamma(t))\right|_{t=0}\right) = \left.\frac{d^2}{dt^2} f(\gamma(t))\right|_{t=0}.$$

Therefore, we have

$$\mathrm{Hess}(f)_p(u, u) = \left.\frac{d^2}{dt^2} f(\gamma(t))\right|_{t=0},$$

as claimed. □

Since the Hessian is a symmetric bilinear form, we obtain $\mathrm{Hess}(f)_p(u, v)$ by polarization; that is,

$$\mathrm{Hess}(f)_p(u, v) = \frac{1}{2}(\mathrm{Hess}(f)_p(u + v, u + v) - \mathrm{Hess}(f)_p(u, u) - \mathrm{Hess}(f)_p(v, v)).$$

Let us find the Hessian of the function $f\colon \mathbf{SO}(3) \to \mathbb{R}$ defined in the second example of Section 7.5, with

$$f(R) = (u^\top Rv)^2.$$

We found that

$$df_R(X) = 2u^\top Xvu^\top Rv, \quad X \in R\mathfrak{so}(3)$$

and that the gradient is given by

$$(\mathrm{grad}(f))_R = u^\top RvR(R^\top uv^\top - vu^\top R).$$

To compute the Hessian, we use the curve $\gamma(t) = Re^{tB}$, where $B \in \mathfrak{so}(3)$. Indeed, it can be shown (see Section 20.3, Proposition 20.20) that the metric induced by the inner product

$$\langle B_1, B_2\rangle = \mathrm{tr}(B_1^\top B_2) = -\mathrm{tr}(B_1 B_2)$$

on $\mathfrak{so}(n)$ is bi-invariant, and so the curve γ is a geodesic.

First we compute

$$\begin{aligned} (f(\gamma(t)))'(t) &= ((u^\top Re^{tB}v)^2)'(t) \\ &= 2u^\top Re^{tB}vu^\top RBe^{tB}v, \end{aligned}$$

and then

$$\begin{aligned} \mathrm{Hess}(f)_R(RB, RB) &= (f(\gamma(t)))''(0) \\ &= (2u^\top Re^{tB}vu^\top RBe^{tB}v)'(0) \\ &= 2u^\top RBvu^\top RBv + 2u^\top Rvu^\top RBBv \\ &= 2u^\top RBvu^\top RBv + 2u^\top Rvu^\top RBR^\top RBv. \end{aligned}$$

By polarization, we obtain

$$\mathrm{Hess}(f)_R(X, Y) = 2u^\top Xvu^\top Yv + u^\top Rvu^\top XR^\top Yv + u^\top Rvu^\top YR^\top Xv,$$

with $X, Y \in R\mathfrak{so}(3)$.

15.6 The Calculus of Variations Applied to Geodesics; The First Variation Formula

In this section, we consider a Riemannian manifold (M, g) equipped with the Levi-Civita connection. The path space $\Omega(p, q)$ was introduced in Definition 15.1. It is an "infinite dimensional" manifold. By analogy with finite dimensional manifolds, we define a kind of tangent space to $\Omega(p, q)$ at a "point" ω. In this section, it is convenient to assume that paths in $\Omega(p, q)$ are parametrized over the interval $[0, 1]$.

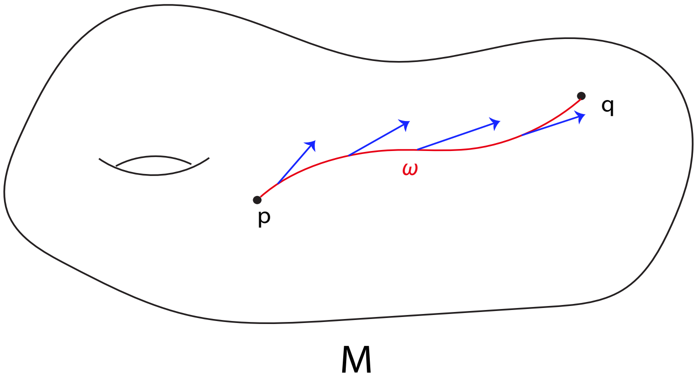


Figure 15.9: The point ω in $\Omega(p,q)$ and its associated tangent vector, the blue vector field. Each blue vector is contained in a tangent space for $\omega(t)$.

Definition 15.19. For every "point" $\omega \in \Omega(p,q)$, we define the "*tangent space*" $T_\omega\Omega(p,q)$ to $\Omega(p,q)$ at ω, as the space of all piecewise smooth vector fields W along ω (see Definition 14.4), for which $W(0) = W(1) = 0$. See Figure 15.9.

If $F\colon \Omega(p,q) \to \mathbb{R}$ is a real-valued function on $\Omega(p,q)$, it is natural to ask what the induced "tangent map"

$$dF_\omega\colon T_\omega\Omega(p,q) \to \mathbb{R},$$

should mean (here, we are identifying $T_{F(\omega)}\mathbb{R}$ with $\mathbb{R}$). Observe that $\Omega(p,q)$ is not even a topological space so the answer is far from obvious!

In the case where $f\colon M \to \mathbb{R}$ is a function on a manifold, there are various equivalent ways to define df, one of which involves curves. For every $v \in T_pM$, if $\alpha\colon (-\epsilon,\epsilon) \to M$ is a curve such that $\alpha(0) = p$ and $\alpha'(0) = v$, then we know that

$$df_p(v) = \left.\frac{d(f(\alpha(t)))}{dt}\right|_{t=0}.$$

We may think of α as a small *variation* of p. Recall that p is a *critical point* of f iff $df_p(v) = 0$, for all $v \in T_pM$.

Rather than attempting to define dF_ω (which requires some conditions on F), we will mimic what we did with functions on manifolds and define what is a *critical path* of a function $F\colon \Omega(p,q) \to \mathbb{R}$, using the notion of *variation*. Now geodesics from p to q are special paths in $\Omega(p,q)$, and they turn out to be the critical paths of the *energy function*

$$E_a^b(\omega) = \int_a^b \|\omega'(t)\|^2\, dt,$$

where $\omega \in \Omega(p,q)$, and $0 \le a < b \le 1$.

Definition 15.20. Given any path $\omega \in \Omega(p, q)$, a *variation of* ω *(keeping endpoints fixed)* is a function $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$, for some $\epsilon > 0$, such that:

(1) $\widetilde{\alpha}(0) = \omega$

(2) There is a subdivision $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$ of $[0, 1]$ so that the map

$$\alpha\colon (-\epsilon, \epsilon) \times [0, 1] \to M$$

defined by $\alpha(u, t) = \widetilde{\alpha}(u)(t)$ is smooth on each strip $(-\epsilon, \epsilon)\times[t_i, t_{i+1}]$, for $i = 0, \ldots, k-1$.

See Figure 15.10. If U is an open subset of $\mathbb{R}^n$ containing the origin and if we replace $(-\epsilon, \epsilon)$ by U in the above, then $\widetilde{\alpha}\colon U \to \Omega(p, q)$ is called an *n-parameter variation* of ω.

The function α is also called a *variation* of ω. Since each $\widetilde{\alpha}(u)$ belongs to $\Omega(p, q)$, note that

$$\alpha(u, 0) = p, \quad \alpha(u, 1) = q, \quad \text{for all } u \in (-\epsilon, \epsilon).$$

The function $\widetilde{\alpha}$ may be considered as a “smooth path” in $\Omega(p, q)$, since for every $u \in (-\epsilon, \epsilon)$, the map $\widetilde{\alpha}(u)$ is a curve in $\Omega(p, q)$ called a *curve in the variation (or longitudinal curve of the variation).*

Definition 15.21. Let $\omega \in \Omega(p, q)$, and let $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$ be a variation of ω as defined in Definition 15.20. The “tangent vector” $\frac{d\widetilde{\alpha}}{du}(0) \in T_\omega\Omega(p, q)$ is defined to be the vector field W along ω given by

$$W_t = \frac{\partial \alpha}{\partial u}(u, t)\bigg|_{u=0}.$$

By definition,

$$\frac{d\widetilde{\alpha}}{du}(0)_t = W_t, \quad t \in [0, 1].$$

Clearly, $W \in T_\omega\Omega(p, q)$. In particular, $W(0) = W(1) = 0$. The vector field W is also called the *variation vector field* associated with the variation α. See Figure 15.10.

Besides the curves in the variation $\widetilde{\alpha}(u)$ (with $u \in (-\epsilon, \epsilon)$), for every $t \in [0, 1]$, we have a curve $\alpha_t\colon (-\epsilon, \epsilon) \to M$, called a *transversal curve of the variation*, defined by

$$\alpha_t(u) = \widetilde{\alpha}(u)(t),$$

and W_t is equal to the velocity vector $\alpha_t'(0)$ at the point $\omega(t) = \alpha_t(0)$. For ϵ sufficiently small, the vector field W_t is an infinitesimal model of the variation $\widetilde{\alpha}$.

Proposition 15.23. *For any $W \in T_\omega\Omega(p, q)$, there is a variation $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$ which satisfies the conditions*

$$\widetilde{\alpha}(0) = \omega, \qquad \frac{d\widetilde{\alpha}}{du}(0) = W.$$

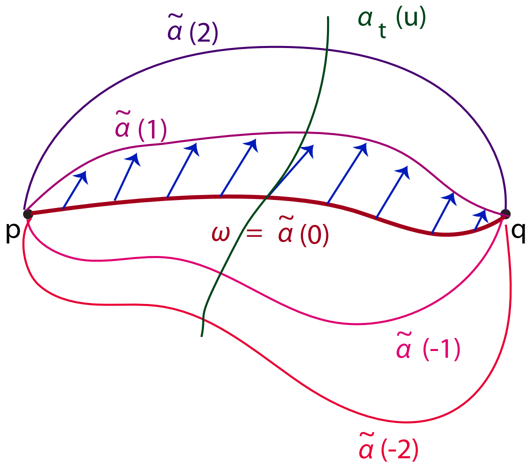

$\widetilde{\alpha}$ (2)
α_t (u)
$\widetilde{\alpha}$ (1)
p
q
$\omega = \widetilde{\alpha}$ (0)
$\widetilde{\alpha}$ (-1)
$\widetilde{\alpha}$ (-2)

Figure 15.10: A variation of ω in $\mathbb{R}^2$ with transversal curve $\alpha_t(u)$. The blue vector field is the variational vector field W_t.

Sketch of the proof. By the compactness of $\omega([0,1])$, it is possible to find a $\delta > 0$ so that $\exp_{\omega(t)}$ is defined for all $t \in [0,1]$ and all $v \in T_{\omega(t)}M$, with $\|v\| < \delta$. Then if

$$N = \max_{t\in[0,1]} \|W_t\|,$$

for any ϵ such that $0 < \epsilon < \frac{\delta}{N}$, it can be shown that

$$\widetilde{\alpha}(u)(t) = \exp_{\omega(t)}(uW_t)$$

works (for details, see Do Carmo [39], Chapter 9, Proposition 2.2). □

As we said earlier, given a function $F\colon \Omega(p,q) \to \mathbb{R}$, we do not attempt to define the differential dF_ω, but instead the notion of critical path.

Definition 15.22. Given a function $F\colon \Omega(p,q) \to \mathbb{R}$, we say that a path $\omega \in \Omega(p,q)$ is a *critical path* for F iff

$$\left.\frac{dF(\widetilde{\alpha}(u))}{du}\right|_{u=0} = 0,$$

for every variation $\widetilde{\alpha}$ of ω (which implies that the derivative $\left.\frac{dF(\widetilde{\alpha}(u))}{du}\right|_{u=0}$ is defined for every variation $\widetilde{\alpha}$ of ω).

For example, if F takes on its minimum on a path ω_0 and if the derivatives $\frac{dF(\widetilde{\alpha}(u))}{du}$ are all defined, then ω_0 is a critical path of F.

We will apply the above to two functions defined on $\Omega(p,q)$.

(1) The *energy function* (also called *action integral*)

$$E_a^b(\omega) = \int_a^b \|\omega'(t)\|^2 \, dt.$$

(We write $E = E_0^1$.)

(2) The *arc-length function*

$$L_a^b(\omega) = \int_a^b \|\omega'(t)\| \, dt.$$

The quantities $E_a^b(\omega)$ and $L_a^b(\omega)$ can be compared as follows: if we apply the Cauchy-Schwarz inequality,

$$\left(\int_a^b f(t)g(t)dt\right)^2 \leq \left(\int_a^b f^2(t)dt\right)\left(\int_a^b g^2(t)dt\right)$$

with $f(t) \equiv 1$ and $g(t) = \|\omega'(t)\|$, we get

$$(L_a^b(\omega))^2 \leq (b-a)E_a^b,$$

where equality holds iff g is constant; that is, iff the parameter t is proportional to arc-length.

Now suppose that there exists a minimal geodesic γ from p to q. Then, using Proposition 15.11 which says that $L(\gamma) \leq L(\omega)$, we get

$$E(\gamma) = L(\gamma)^2 \leq L(\omega)^2 \leq E(\omega),$$

where the equality $L(\gamma)^2 = L(\omega)^2$ holds only if ω is also a minimal geodesic, possibly reparametrized. On the other hand, the equality $L(\omega)^2 = E(\omega)$ can hold only if the parameter is proportional to arc-length along ω. This proves that $E(\gamma) < E(\omega)$ unless ω is also a minimal geodesic. We just proved:

Proposition 15.24. *Let (M, g) be a complete Riemannian manifold. For any two points $p, q \in M$, if $d(p, q) = \delta$, then the energy function $E\colon \Omega(p, q) \to \mathbb{R}$ takes on its minimum δ^2 precisely on the set of minimal geodesics from p to q.*

Next we are going to show that the critical paths of the energy function are exactly the geodesics. For this we need the *first variation formula*.

Let $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p, q)$ be a variation of ω, and let

$$W_t = \left.\frac{\partial \alpha}{\partial u}(u, t)\right|_{u=0}$$

be its associated variation vector field. Furthermore, let

$$V_t = \frac{d\omega}{dt} = \omega'(t),$$

the velocity vector field of ω, and

$$\Delta_t V = V_{t_+} - V_{t_-},$$

the discontinuity in the velocity vector at t, which is nonzero only for $t = t_i$, with $0 < t_i < 1$ (see the definition of $\gamma'((t_i)_+)$ and $\gamma'((t_i)_-)$ just after Definition 15.1). See Figure 15.11.

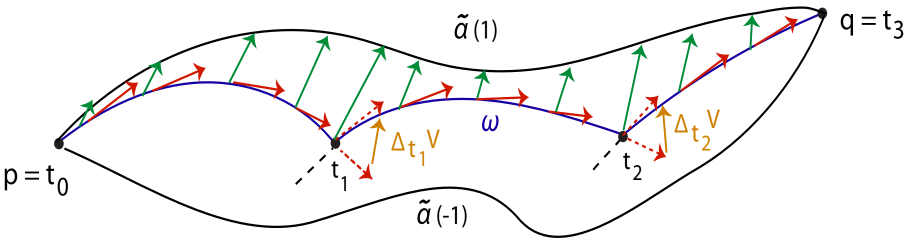


Figure 15.11: The point ω in blue with V_t in red, W_t in green, and $\Delta_t V$ in orange.

Theorem 15.25. *(First Variation Formula) For any path $\omega \in \Omega(p, q)$, we have*

$$\frac{1}{2} \left. \frac{dE(\widetilde{\alpha}(u))}{du} \right|_{u=0} = -\sum_i \langle W_t, \Delta_t V \rangle - \int_0^1 \left\langle W_t, \frac{D}{dt} V_t \right\rangle dt, \tag{$\dagger$}$$

where $\widetilde{\alpha} \colon (-\epsilon, \epsilon) \to \Omega(p, q)$ is any variation of ω.

Proof. (After Milnor, see [81], Chapter II, Section 12, Theorem 12.2.) By Proposition 14.11, we have

$$\frac{\partial}{\partial u} \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle = 2 \left\langle \frac{D}{\partial u} \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle.$$

Therefore,

$$\frac{dE(\widetilde{\alpha}(u))}{du} = \frac{d}{du} \int_0^1 \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle dt = 2 \int_0^1 \left\langle \frac{D}{\partial u} \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t} \right\rangle dt.$$

Now, because we are using the Levi-Civita connection, which is torsion-free, Proposition 15.7 implies that

$$\frac{D}{\partial t} \frac{\partial \alpha}{\partial u} = \frac{D}{\partial u} \frac{\partial \alpha}{\partial t}.$$

Consequently,

$$\frac{dE(\widetilde{\alpha}(u))}{du} = 2 \int_0^1 \left\langle \frac{D}{\partial t} \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t} \right\rangle dt.$$

We can choose $0 = t_0 < t_1 < \cdots < t_k = 1$ so that α is smooth on each strip $(-\epsilon, \epsilon) \times [t_{i-1}, t_i]$. Then we can "integrate by parts" on $[t_{i-1}, t_i]$ as follows. The equation

$$\frac{\partial}{\partial t}\left\langle \frac{\partial\alpha}{\partial u}, \frac{\partial\alpha}{\partial t}\right\rangle = \left\langle \frac{D}{\partial t}\frac{\partial\alpha}{\partial u}, \frac{\partial\alpha}{\partial t}\right\rangle + \left\langle \frac{\partial\alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle$$

implies that

$$\int_{t_{i-1}}^{t_i}\left\langle \frac{D}{\partial t}\frac{\partial\alpha}{\partial u}, \frac{\partial\alpha}{\partial t}\right\rangle dt = \left\langle \frac{\partial\alpha}{\partial u}, \frac{\partial\alpha}{\partial t}\right\rangle\Bigg|_{t=(t_{i-1})_+}^{t=(t_i)_-} - \int_{t_{i-1}}^{t_i}\left\langle \frac{\partial\alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.$$

Adding up these formulae for $i = 1, \ldots k-1$ and using the fact that $\frac{\partial\alpha}{\partial u} = 0$ for $t = 0$ and $t = 1$, we get

$$\frac{1}{2}\frac{dE(\widetilde{\alpha}(u))}{du} = -\sum_{i=1}^{k-1}\left\langle \frac{\partial\alpha}{\partial u}, \Delta_{t_i}\frac{\partial\alpha}{\partial t}\right\rangle - \int_0^1\left\langle \frac{\partial\alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.$$

Setting $u = 0$, we obtain the formula

$$\frac{1}{2}\frac{dE(\widetilde{\alpha}(u))}{du}\bigg|_{u=0} = -\sum_i \langle W_t, \Delta_t V\rangle - \int_0^1\left\langle W_t, \frac{D}{dt}V_t\right\rangle dt,$$

as claimed. □

Remark: The reader will observe that the proof used the fact that the connection is compatible with the metric and torsion-free.

Intuitively, the first term on the right-hand side shows that varying the path ω in the direction of decreasing "kink" tends to decrease E.

The second term shows that varying the curve in the direction of its acceleration vector $\frac{D}{dt}\omega'(t)$ also tends to reduce E.

A geodesic γ (parametrized over $[0, 1]$) is smooth on the entire interval $[0, 1]$ and its acceleration vector $\frac{D}{dt}\gamma'(t)$ is identically zero along γ. This gives us half of

Theorem 15.26. *Let (M, g) be a Riemanian manifold. For any two points $p, q \in M$, a path $\omega \in \Omega(p, q)$ (parametrized over $[0, 1]$) is critical for the energy function E iff ω is a geodesic.*

Proof. From the first variation formula, it is clear that a geodesic is a critical path of E.

Conversely, assume ω is a critical path of E. By Proposition 15.23, there is a variation $\widetilde{\alpha}$ of ω such that its associated variation vector field is equal to

$$W_t = f(t)\frac{D}{dt}\omega'(t),$$

with $f(t)$ smooth and positive except that it vanishes at the t_i's. For this variation, by the first variation formula (†), we get

$$\frac{1}{2}\left.\frac{dE(\widetilde{\alpha}(u))}{du}\right|_{u=0} = -\int_0^1 f(t)\left\langle \frac{D}{dt}\,\omega'(t), \frac{D}{dt}\,\omega'(t)\right\rangle dt.$$

This expression is zero iff

$$\frac{D}{dt}\,\omega'(t) = 0 \qquad \text{on } [0,1].$$

Hence, the restriction of ω to each $[t_i, t_{i+1}]$ is a geodesic.

It remains to prove that ω is smooth on the entire interval $[0,1]$. For this, using Proposition 15.23, pick a variation $\widetilde{\alpha}$ such that

$$W_{t_i} = \Delta_{t_i} V.$$

Then we have

$$\frac{1}{2}\left.\frac{dE(\widetilde{\alpha}(u))}{du}\right|_{u=0} = -\sum_{i=1}^{k}\langle \Delta_{t_i}V, \Delta_{t_i}V\rangle.$$

If the above expression is zero, then $\Delta_{t_i}V = 0$ for $i = 1, \ldots, k-1$, which means that ω is C^1 everywhere on $[0,1]$. By the uniqueness theorem for ODE's, ω must be smooth everywhere on $[0,1]$, and thus, it is an unbroken geodesic. □

Remark: If $\omega \in \Omega(p,q)$ is parametrized by arc-length, then it is easy to prove that

$$\left.\frac{dL(\widetilde{\alpha}(u))}{du}\right|_{u=0} = \frac{1}{2}\left.\frac{dE(\widetilde{\alpha}(u))}{du}\right|_{u=0}.$$

As a consequence, a path $\omega \in \Omega(p,q)$ is critical for the arc-length function L iff it can be reparametrized so that it is a geodesic (see Gallot, Hulin and Lafontaine [49], Chapter 3, Theorem 3.31).

In order to go deeper into the study of geodesics, we need Jacobi fields and the "second variation formula," both involving a curvature term. Therefore, we now proceed with a more thorough study of curvature on Riemannian manifolds.

15.7 Problems

Problem 15.1. Let X be a surface of revolution defined on the open rectangle $(a,b)\times(c,d)$ such that

$$\begin{aligned} x &= f(v)\cos u,\\ y &= f(v)\sin u,\\ z &= g(v). \end{aligned}$$

(i) show that the Riemannian metric is given by

$$g_{11} = f(v)^2, \quad g_{12} = g_{21} = 0, \quad g_{22} = f'(v)^2 + g'(v)^2.$$

From now on, assume that $f'(v)^2 + g'(v)^2 \neq 0$ and that $f(v) \neq 0$. Images by X of the curves $u =$ constant are called *meridians*, and images of the curves $v =$ constant are called *parallels*.

(ii) Show that the Christoffel symbols are given by

$$\begin{aligned}
&\Gamma^1_{11} = 0, \quad \Gamma^2_{11} = -\frac{ff'}{(f')^2 + (g')^2}, \quad && \Gamma^1_{12} = \frac{ff'}{f^2}, \\
&\Gamma^2_{12} = 0, \quad \Gamma^1_{22} = 0, && \Gamma^2_{22} = \frac{f'f'' + g'g''}{(f')^2 + (g')^2}.
\end{aligned}$$

(iii) Show that the equations of the geodesics are

$$\begin{aligned}
u'' + \frac{2ff'}{f^2}\, u'v' &= 0, \\
v'' - \frac{ff'}{(f')^2 + (g')^2}\, (u')^2 + \frac{f'f'' + g'g''}{(f')^2 + (g')^2}\, (v')^2 &= 0.
\end{aligned}$$

Show that the meridians parametrized by arc length are geodesics. Show that a parallel is a geodesic iff it is generated by the rotation of a point of the generating curve where the tangent is parallel to the axis of rotation.

(iv) Show that the first equation of geodesics is equivalent to

$$f^2 u' = c,$$

for some constant c. Since the angle θ, $0 \leq \theta \leq \pi/2$, of a geodesic with a parallel that intersects it is given by

$$\cos\theta = \frac{|X_u \cdot (X_u u' + X_v v')|}{\|X_u\|} = |fu'|,$$

and since $f = r$ is the radius of the parallel at the intersection, show that

$$r\cos\theta = c$$

for some constant $c > 0$. The equation $r\cos\theta = c$ is known as *Clairaut's relation*.

Problem 15.2. Consider the Poincaré half plane $P = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}$, with the Riemannian metric $g = (dx^2 + dy^2)/y^2$.

(1) Compute the Christoffel symbols and write down the equations of the geodesics.

(2) Prove that curves (half-lines) given by $t \mapsto (x_0, e^{\alpha t})$ are geodesics.

(3) Prove that the transformations of the form

$$z \mapsto \frac{\alpha x + \beta}{\gamma z + \delta}, \quad z = x + iy, \quad \alpha\delta - \beta\gamma = 1, \quad \alpha, \beta, \gamma, \delta \in \mathbb{R}$$

are isometries of P.

(4) Prove that the geodesics of P are the half-lines $x = \alpha$ and the half-circles centered on the x-axis,

Problem 15.3. In the proof of Proposition 15.5, check that the Jacobian matrix of $d_{(p,0)}\Phi$ is equal to

$$\begin{pmatrix} I & I \\ I & 0 \end{pmatrix}.$$

Problem 15.4. Prove Proposition 15.6.

Problem 15.5. Prove the equation

$$\frac{D}{\partial t}\frac{\partial f}{\partial r} = \frac{D}{\partial r}\frac{\partial f}{\partial t}$$

of Proposition 15.7.

Problem 15.6. If $v(t)$ is a curve on S^{n-1} such that $v(0) = v$ and $v'(0) = w_N$ (with $\|v\| < i(p)$), then since $f(r,t) = \exp_p(rv(t))$

$$\frac{\partial f}{\partial r}(1,0) = (d\exp_p)_v(v), \quad \frac{\partial f}{\partial t}(1,0) = (d\exp_p)_v(w_N),$$

and Gauss lemma can be stated as

$$\langle (d\exp_p)_v(v), (d\exp_p)_v(w_N)\rangle = \langle v, w_N\rangle = 0.$$

Prove that this statement of Gauss lemma is equivalent to the satement given in Proposition 15.8.

Problem 15.7. Prove that the Poincaré half-plane of Problem 15.2 is complete.

Problem 15.8. Let M be a complete Riemannian manifold and let $N \subseteq M$ be a closed embedded submanifold with the induced Riemannian metric. Prove that N is complete.

Beware that the distance function on N induced by the metric is not in general equal to the Riemannin distance on M.

Problem 15.9. Let M be a complete Riemannian manifold and let $N \subseteq M$ be a closed embedded submanifold. For any point $p \in M - N$, define the distance from p to N as

$$d(p, N) = \inf\{d(p, x) \mid x \in N\}.$$

If $q \in N$ is a point such that $d(p,q) = d(p,N)$ and if γ is any minimizing geodesic from p to q, prove that γ intersects N orthogonally.

Chapter 16

Curvature in Riemannian Manifolds

Since the notion of curvature can be defined for curves and surfaces, it is natural to wonder whether it can be generalized to manifolds of dimension $n \geq 3$. Such a generalization does exist and was first proposed by Riemann. However, Riemann's seminal paper published in 1868 two years after his death only introduced the sectional curvature, and did not contain any proofs or any general methods for computing the sectional curvature. Fifty years or so later, the idea emerged that the *curvature* of a Riemannian manifold M should be viewed as a measure $R(X, Y)Z$ of the extent to which the operator $(X, Y) \mapsto \nabla_X \nabla_Y Z$ is symmetric, where ∇ is a connection on M (where X, Y, Z are vector fields, with Z fixed). It turns out that the operator $R(X, Y)Z$ is $C^\infty(M)$-linear in all of its three arguments, so for all $p \in M$, it defines a trilinear map

$$R_p \colon T_pM \times T_pM \times T_pM \longrightarrow T_pM.$$

The curvature operator R is a rather complicated object, so it is natural to seek a simpler object. Fortunately, there is a simpler object, namely the *sectional curvature* $K(u, v)$, which arises from R through the formula

$$K(u, v) = \langle R(u, v)u, v \rangle,$$

for linearly independent unit vectors u, v. When ∇ is the Levi-Civita connection induced by a Riemannian metric on M, it turns out that the curvature operator R can be recovered from the sectional curvature. Another important notion of curvature is the *Ricci curvature*, $\mathrm{Ric}(x, y)$, which arises as the trace of the linear map $v \mapsto R(x, v)y$. The curvature operator R, sectional curvature, and Ricci curvature are introduced in the first three sections of this chapter.

In Section 15.6, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 15.26). A deeper understanding is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on $\mathbb{R}^n$, it is possible to define a bilinear functional

$$I_\gamma \colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$$

when γ is a critical point of the energy function E (that is, γ is a geodesic). This bilinear form is usually called the *index form*. In order to define the functional I_γ (where γ is a geodesic), we introduce 2-parameter variations, which generalize the variations given by Definition 15.20. Then we derive the *second variation formula*, which gives an expression for the second derivative $\partial^2((E \circ \widetilde{\alpha})/\partial u_1 \partial u_2)(u_1, u_2) \mid_{(0,0)}$, where $\widetilde{\alpha}$ is a 2-variation of a geodesic γ. Remarkably, this expression contains a curvature term $R(V, W_1)V$, where $W_1(t) = (\partial\alpha/\partial u_1)(0, 0, t)$ and $V(t) = \gamma'(t)$. The second variation formula allows us to show that the index form $I(W_1, W_2)$ is well-defined, and symmetric bilinear. When γ is a minimal geodesic, I is positive semi-definite. For any geodesic γ, we define the *index* of

$$I \colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$$

as the maximum dimension of a subspace of $T_\gamma\Omega(p, q)$ on which I is negative definite. Section 16.4 is devoted to the second variation formula and the definition of the index form.

In Section 16.5, we define Jacobi fields and study some of their properties. Given a geodesic $\gamma \in \Omega(p, q)$, a vector field J along γ is a *Jacobi field* iff it satisfies the *Jacobi differential equation*

$$\frac{D^2 J}{dt^2} + R(\gamma', J)\gamma' = 0.$$

We prove that Jacobi fields are exactly the vector fields that belong to the nullspace of the index form I. Jacobi fields also turn out to arise from special variations consisting of geodesics (geodesic variations). We define the notion of *conjugate points* along a geodesic. We show that the derivative of the exponential map is expressible in terms of a Jacobi field and characterize the critical points of the exponential in terms of conjugate points.

Section 16.7 presents some applications of Jacobi fields and the second variation formula to topology. We prove

(1) Hadamard and Cartan's theorems about complete manifolds of non-positive sectional curvature.

(2) Myers' theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

We also state the famous *Morse index theorem.*

In Section 16.8 we revisit the cut locus and prove more properties about it using Jacobi fields.

16.1 The Curvature Tensor

As we said above, if M is a Riemannian manifold and if ∇ is a connection on M, the Riemannian curvature $R(X, Y)Z$ measures the extent to which the operator $(X, Y) \mapsto \nabla_X \nabla_Y Z$ is symmetric (for any fixed Z).

If $(M, \langle -, - \rangle)$ is a Riemannian manifold of dimension n, and if the connection ∇ on M is the flat connection, which means that

$$\nabla_X \left(\frac{\partial}{\partial x_i} \right) = 0, \quad i = 1, \ldots, n,$$

for every chart (U, φ) and all $X \in \mathfrak{X}(U)$, since every vector field Y on U can be written uniquely as

$$Y = \sum_{i=1}^{n} Y_i \frac{\partial}{\partial x_i}$$

for some smooth functions Y_i on U, for every other vector field X on U, because the connection is flat and by the Leibniz property of connections, we have

$$\nabla_X \left(Y_i \frac{\partial}{\partial x_i} \right) = X(Y_i) \frac{\partial}{\partial x_i} + Y_i \nabla_X \left(\frac{\partial}{\partial x_i} \right) = X(Y_i) \frac{\partial}{\partial x_i}.$$

Then it is easy to check that the above implies that

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z,$$

for all $X, Y, Z \in \mathfrak{X}(M)$. Consequently, it is natural to define the deviation of a connection from the flat connection by the quantity

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

for all $X, Y, Z \in \mathfrak{X}(M)$.

Definition 16.1. Let (M, g) be a Riemannian manifold, and let ∇ be any connection on M. The formula

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z, \quad X, Y, Z \in \mathfrak{X}(M),$$

defines a function

$$R \colon \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M)$$

called the *Riemannian curvature* of M.

The Riemannian curvature is a special instance of the notion of curvature of a connection on a vector bundle. This approach is discussed in Morita [87].

The function R is clearly skew-symmetric in X and Y. This function turns out to be $C^\infty(M)$-linear in X, Y, Z.

Proposition 16.1. *Let M be a manifold with any connection ∇. The function*

$$R \colon \mathfrak{X}(M) \times \mathfrak{X}(M) \times \mathfrak{X}(M) \longrightarrow \mathfrak{X}(M)$$

given by

$$R(X, Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z$$

is $C^\infty(M)$-linear in X, Y, Z, and skew-symmetric in X and Y. As a consequence, for any $p \in M$, $(R(X, Y)Z)_p$ depends only on $X(p), Y(p), Z(p)$.

Proof. Let us check $C^\infty(M)$-linearity in Z. Additivity is clear. For any function $f \in C^\infty(M)$, we have

$$\begin{aligned}\nabla_Y \nabla_X (fZ) &= \nabla_Y (X(f)Z + f\nabla_X Z)\\ &= Y(X(f))Z + X(f)\nabla_Y Z + Y(f)\nabla_X Z + f\nabla_Y\nabla_X Z.\end{aligned}$$

It follows that

$$\begin{aligned}\nabla_X \nabla_Y (fZ) - \nabla_Y \nabla_X (fZ) &= X(Y(f))Z + Y(f)\nabla_X Z + X(f)\nabla_Y Z + f\nabla_X\nabla_Y Z\\ &\quad - Y(X(f))Z - X(f)\nabla_Y Z - Y(f)\nabla_X Z - f\nabla_Y\nabla_X Z\\ &= (XY - YX)(f)Z + f(\nabla_X\nabla_Y - \nabla_Y\nabla_X)Z.\end{aligned}$$

Hence

$$\begin{aligned}R(X,Y)(fZ) &= \nabla_X \nabla_Y (fZ) - \nabla_Y \nabla_X (fZ) - \nabla_{[X,Y]}(fZ)\\ &= (XY - YX)(f)Z + f(\nabla_X\nabla_Y - \nabla_Y\nabla_X)Z - [X,Y](f)Z - f\nabla_{[X,Y]}Z\\ &= (XY - YX - [X,Y])(f)Z + f(\nabla_X\nabla_Y - \nabla_Y\nabla_X - \nabla_{[X,Y]})Z\\ &= fR(X,Y)Z.\end{aligned}$$

Let us now check $C^\infty(M)$-linearity in Y. Additivity is clear. For any function $f \in C^\infty(M)$, recall that

$$[X, fY] = X(f)Y + f[X,Y].$$

Then

$$\begin{aligned}R(X,fY)Z &= \nabla_X\nabla_{fY}Z - \nabla_{fY}\nabla_X Z - \nabla_{[X,fY]}Z\\ &= \nabla_X(f\nabla_Y Z) - f\nabla_Y\nabla_X Z - X(f)\nabla_Y Z - f\nabla_{[X,Y]}Z\\ &= X(f)\nabla_Y Z + f\nabla_X\nabla_Y Z - f\nabla_Y\nabla_X Z - X(f)\nabla_Y Z - f\nabla_{[X,Y]}Z\\ &= f(\nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,Y]}Z)\\ &= fR(X,Y)Z.\end{aligned}$$

Since R is skew-symmetric in X and Y, R is also $C^\infty(M)$-linear in X. For any chart (U,φ), we can express the vector fields X, Y, Z uniquely as

$$X = \sum_{i=1}^n X_i \frac{\partial}{\partial x_i}, \quad Y = \sum_{j=1}^n Y_j \frac{\partial}{\partial x_j}, \quad Z = \sum_{k=1}^n Z_k \frac{\partial}{\partial x_k},$$

for some smooth functions $X_i, Y_j, Z_k \in C^\infty(U)$, and by $C^\infty(U)$-linearity, we have

$$\begin{aligned}R(X,Y)Z &= \sum_{i,j,k} R\left(X_i\frac{\partial}{\partial x_i}, Y_j\frac{\partial}{\partial x_j}\right)\left(Z_k\frac{\partial}{\partial x_k}\right)\\ &= \sum_{i,j,k} X_iY_jZ_k R\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right)\left(\frac{\partial}{\partial x_k}\right).\end{aligned}$$

Evaluated at p, we get

$$(R(X,Y)Z)_p = \sum_{i,j,k} X_i(p)Y_j(p)Z_k(q)\left(R\left(\frac{\partial}{\partial x_i},\frac{\partial}{\partial x_j}\right)\left(\frac{\partial}{\partial x_k}\right)\right)_p,$$

an expression that depends only on the values of the functions X_i, Y_j, Z_k at p. □

It follows that R defines for every $p \in M$ a trilinear map

$$R_p\colon T_pM \times T_pM \times T_pM \longrightarrow T_pM.$$

(In fact, R defines a $(1,3)$-tensor.)

If our manifold is a Riemannian manifold $(M, \langle -,- \rangle)$ equipped with a connection, experience shows that it is useful to consider the family of quadrilinear forms (unfortunately!) also denoted R, given by

$$R_p(x,y,z,w) = \langle R_p(x,y)z, w\rangle_p,$$

as well as the expression $R_p(x,y,y,x)$, which, for an orthonormal pair of vectors (x,y), is known as the *sectional curvature* $K_p(x,y)$.

This last expression brings up a dilemma regarding the choice for the sign of R. With our present choice, the sectional curvature $K_p(x,y)$ is given by $K_p(x,y) = R_p(x,y,y,x)$, but many authors define K as $K_p(x,y) = R_p(x,y,x,y)$. Since $R(X,Y)$ is skew-symmetric in X, Y, the latter choice corresponds to using $-R(X,Y)$ instead of $R(X,Y)$, that is, to define $R(X,Y)Z$ by

$$R(X,Y)Z = \nabla_{[X,Y]}Z + \nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z.$$

As pointed out by Milnor [81] (Chapter II, Section 9), the latter choice for the sign of R has the advantage that, in coordinates, the quantity $\langle R(\partial/\partial x_h, \partial/\partial x_i)\partial/\partial x_j, \partial/\partial x_k\rangle$ coincides with the classical Ricci notation, R_{hijk}. Gallot, Hulin and Lafontaine [49] (Chapter 3, Section A.1) give other reasons supporting this choice of sign. Clearly, the choice for the sign of R is mostly a matter of taste and we apologize to those readers who prefer the first choice but we will adopt the second choice advocated by Milnor and others (including O'Neill [91] and Do Carmo [39]), we make the following formal definition.

Definition 16.2. Let $(M, \langle -,- \rangle)$ be a Riemannian manifold equipped with any connection. The *curvature tensor* is the family of trilinear functions $R_p\colon T_pM \times T_pM \times T_pM \to T_pM$ defined by

$$R_p(x,y)z = \nabla_{[X,Y]}Z + \nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z,$$

for every $p \in M$ and for any vector fields $X, Y, Z \in \mathfrak{X}(M)$ such that $x = X(p)$, $y = Y(p)$, and $z = Z(p)$. The family of quadrilinear forms associated with R, also denoted R, is given by

$$R_p(x,y,z,w) = \langle R_p(x,y)z, w\rangle_p,$$

for all $p \in M$ and all $x, y, z, w \in T_pM$.

Following common practice in mathematics, in the interest of keeping notation to a minimum, we often write $R(x, y, z, w)$ instead of $R_p(x, y, z, w)$. Since $x, y, z, w \in T_pM$, this abuse of notation rarely causes confusion.

Remark: The curvature tensor R is indeed a $(1, 3)$-tensor, and the associated family of quadrilinear forms is a $(0, 4)$-tensor.

Locally in a chart, we write

$$R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j} = \sum_l R^l_{jhi}\frac{\partial}{\partial x_l}$$

and

$$R_{hijk} = \left\langle R\left(\frac{\partial}{\partial x_h}, \frac{\partial}{\partial x_i}\right)\frac{\partial}{\partial x_j}, \frac{\partial}{\partial x_k}\right\rangle = \sum_l g_{lk}R^l_{jhi}.$$

The coefficients R^l_{jhi} can be expressed in terms of the Christoffel symbols Γ^k_{ij}, by a rather unfriendly formula; see Gallot, Hulin and Lafontaine [49] (Chapter 3, Section 3.A.3) or O'Neill [91] (Chapter III, Lemma 38). Since we have adopted O'Neill's conventions for the order of the subscripts in R^l_{jhi}, here is the formula from O'Neill:

$$R^l_{jhi} = \partial_i\Gamma^l_{hj} - \partial_h\Gamma^l_{ij} + \sum_m \Gamma^l_{im}\Gamma^m_{hj} - \sum_m \Gamma^l_{hm}\Gamma^m_{ij}.$$

It should be noted that the above formula holds for any connection. However, it may be practically impossible to compute the Christoffel symbols if this connection is not the Levi-Civita connection.

For example, in the case of the sphere S^2, we parametrize as

$$\begin{aligned} x &= \sin\theta\cos\varphi \\ y &= \sin\theta\sin\varphi \\ z &= \cos\theta, \end{aligned}$$

over the domain to $\{(\theta, \varphi) \mid 0 < \theta < \pi, 0 < \varphi < 2\pi\}$. For the basis $(u(\theta, \varphi), v(\theta, \varphi))$ of the the tangent plane T_pS^2 at $p = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$, where

$$\begin{aligned} u(\theta, \varphi) &= \frac{\partial p}{\partial\theta} = (\cos\theta\cos\varphi, \cos\theta\sin\varphi, -\sin\theta) \\ v(\theta, \varphi) &= \frac{\partial p}{\partial\varphi} = (-\sin\theta\sin\varphi, \sin\theta\cos\varphi, 0), \end{aligned}$$

we found that the metric on T_pS^2 is given by the matrix

$$g_p = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix};$$

see Section 13.2. Note that

$$g_p^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sin^2\theta} \end{pmatrix}.$$

Since the Christoffel symbols of the Levi-Civita connection are given by

$$\Gamma^k_{ij} = \frac{1}{2}\sum_{l=1}^{n} g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}),$$

(see Section 14.3), we discover that the only nonzero Christoffel symbols are

$$\begin{aligned}
\Gamma^2_{12} = \Gamma^{\varphi}_{\theta\varphi} = \Gamma^2_{21} = \Gamma^{\varphi}_{\varphi\theta} &= \frac{1}{2}\sum_{l=1}^{2} g^{2l}(\partial_1 g_{2l} + \partial_2 g_{1l} - \partial_l g_{12}) \\
&= \frac{1}{2} g^{22}\partial_1 g_{22} = \frac{1}{2}\left(\frac{1}{\sin^2\theta}\cdot\frac{\partial}{\partial\theta}\sin^2\theta\right) = \frac{\cos\theta}{\sin\theta}, \\
\Gamma^1_{22} = \Gamma^{\theta}_{\varphi\varphi} &= \frac{1}{2}\sum_{l=1}^{2} g^{1l}(\partial_2 g_{2l} + \partial_2 g_{2l} - \partial_l g_{22}) \\
&= -\frac{1}{2}\partial_1 g_{22} = -\frac{\partial}{\partial\theta}\sin^2\theta = -\sin\theta\cos\theta,
\end{aligned}$$

where we have set $\theta \to 1$ and $\varphi \to 2$. The only nonzero Riemann curvature tensor components are

$$\begin{aligned}
R^1_{212} = R^{\theta}_{\varphi\theta\varphi} &= \partial_2\Gamma^1_{12} - \partial_1\Gamma^1_{22} + \sum_{m=1}^{2}\Gamma^1_{2m}\Gamma^m_{12} - \sum_{m=1}^{2}\Gamma^1_{1m}\Gamma^m_{22} \\
&= \frac{\partial}{\partial\theta}(-\sin\theta\cos\theta) + \Gamma^1_{22}\Gamma^2_{12} = -(-\cos^2\theta + \sin^2\theta) + (-\sin\theta\cos\theta)\cdot\frac{\cos\theta}{\sin\theta} \\
&= -\sin^2\theta \\
R^1_{221} = R^{\theta}_{\varphi\varphi\theta} &= \partial_1\Gamma^1_{22} - \partial_2\Gamma^1_{12} + \sum_{m=1}^{2}\Gamma^1_{1m}\Gamma^m_{22} - \sum_{m=1}^{2}\Gamma^1_{2m}\Gamma^m_{12} = -R^1_{212} = \sin^2\theta \\
R^2_{112} = R^{\varphi}_{\theta\theta\varphi} &= \partial_2\Gamma^2_{11} - \partial_1\Gamma^2_{21} + \sum_{m=1}^{2}\Gamma^2_{2m}\Gamma^m_{11} - \sum_{m=1}^{2}\Gamma^2_{1m}\Gamma^m_{21} \\
&= -\frac{\partial}{\partial\theta}\left(\frac{\cos\theta}{\sin\theta}\right) - \Gamma^2_{12}\Gamma^2_{21} = -\frac{\partial}{\partial\theta}\cot\theta - \frac{\cos^2\theta}{\sin^2\theta} = \frac{1-\cos^2\theta}{\sin^2\theta} = 1 \\
R^2_{121} = R^{\varphi}_{\theta\varphi\theta} &= \partial_1\Gamma^2_{21} - \partial_2\Gamma^2_{11} + \sum_{m=1}^{2}\Gamma^2_{1m}\Gamma^m_{21} - \sum_{m=1}^{2}\Gamma^2_{2m}\Gamma^m_{11} = -R^2_{112} = -1,
\end{aligned}$$

while the only nonzero components of the associated quadrilinear form are

$$\begin{aligned}
R_{1221} &= \sum_{l=1}^{2} g_{l1} R^l_{212} = g_{11} R^1_{212} = -\sin^2\theta \\
R_{2121} &= \sum_{l=1}^{2} g_{l1} R^l_{221} = g_{11} R^1_{221} = \sin^2\theta \\
R_{1212} &= \sum_{l=1}^{2} g_{l2} R^l_{112} = g_{22} R^2_{112} = \sin^2\theta \\
R_{2112} &= \sum_{l=1}^{2} g_{l2} R^l_{121} = g_{22} R^2_{121} = -\sin^2\theta.
\end{aligned}$$

When ∇ is the Levi-Civita connection, there is another way of defining the curvature tensor which is useful for comparing second covariant derivatives of one-forms.

For any fixed vector field Z, the map $Y \mapsto \nabla_Y Z$ from $\mathfrak{X}(M)$ to $\mathfrak{X}(M)$ is a $C^\infty(M)$-linear map that we will denote $\nabla_- Z$ (this is a $(1,1)$ tensor).

Definition 16.3. The covariant derivative $\nabla_X \nabla_- Z$ of $\nabla_- Z$ is defined by

$$(\nabla_X(\nabla_- Z))(Y) = \nabla_X(\nabla_Y Z) - (\nabla_{\nabla_X Y})Z.$$

Usually, $(\nabla_X(\nabla_- Z))(Y)$ is denoted by $\nabla^2_{X,Y} Z$, and

$$\nabla^2_{X,Y} Z = \nabla_X(\nabla_Y Z) - \nabla_{\nabla_X Y} Z$$

is called the *second covariant derivative* of Z with respect to X and Y.

Then we have

$$\begin{aligned}
\nabla^2_{Y,X} Z - \nabla^2_{X,Y} Z &= \nabla_Y(\nabla_X Z) - \nabla_{\nabla_Y X} Z - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y} Z \\
&= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{\nabla_X Y - \nabla_Y X} Z \\
&= \nabla_Y(\nabla_X Z) - \nabla_X(\nabla_Y Z) + \nabla_{[X,Y]} Z \\
&= R(X,Y)Z,
\end{aligned}$$

since $\nabla_X Y - \nabla_Y X = [X,Y]$, as the Levi-Civita connection is torsion-free.

Proposition 16.2. *Given a Riemanniain manifold (M,g), if ∇ is the Levi-Civita connection induced by g, then the curvature tensor is given by*

$$R(X,Y)Z = \nabla^2_{Y,X} Z - \nabla^2_{X,Y} Z.$$

We already know that the curvature tensor has some symmetry properties, for example $R(y,x)z = -R(x,y)z$, but when it is induced by the Levi-Civita connection, it has more remarkable properties stated in the next proposition.

Proposition 16.3. *For a Riemannian manifold $(M, \langle -, - \rangle)$ equipped with the Levi-Civita connection, the curvature tensor satisfies the following properties for every $p \in M$ and for all $x, y, z, w \in T_pM$:*

(1) $R(x, y)z = -R(y, x)z$

(2) (First Bianchi Identity) $R(x, y)z + R(y, z)x + R(z, x)y = 0$

(3) $R(x, y, z, w) = -R(x, y, w, z)$

(4) $R(x, y, z, w) = R(z, w, x, y)$.

Proof. The proof of Proposition 16.3 uses the fact that $R_p(x, y)z = R(X, Y)Z$, for any vector fields X, Y, Z such that $x = X(p)$, $y = Y(p)$ and $Z = Z(p)$. In particular, X, Y, Z can be chosen so that their pairwise Lie brackets are zero (choose a coordinate system and give X, Y, Z constant components). Part (1) is already known. Part (2) follows from the fact that the Levi-Civita connection is torsion-free and is equivalent to the Jacobi identity for Lie brackets. In particular

$$\begin{aligned}
&R(x, y)z + R(y, z)x + R(z, x)y \\
&= \nabla_{[X,Y]}Z + \nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z + \nabla_{[Y,Z]}X + \nabla_Z\nabla_Y X - \nabla_Y\nabla_Z X + \nabla_{[Z,X]}Y \\
&\quad + \nabla_X\nabla_Z Y - \nabla_Z\nabla_X Y \\
&= \nabla_{[X,Y]}Z + \nabla_Y(\nabla_X Z - \nabla_Z X) + \nabla_X(\nabla_Z Y - \nabla_Y Z) + \nabla_Z(\nabla_Y X - \nabla_X Y) \\
&\quad + \nabla_{[Y,Z]}X + \nabla_{[Z,X]}Y \\
&= \nabla_{[X,Y]}Z + \nabla_Y[X, Z] + \nabla_X[Z, Y] + \nabla_Z[Y, X] + \nabla_{[Y,Z]}X + \nabla_{[Z,X]}Y \\
&= \nabla_{[X,Y]}Z - \nabla_Z[X, Y] + \nabla_{[Y,Z]}X - \nabla_X[Y, Z] + \nabla_{[Z,X]}Y - \nabla_Y[Z, X] \\
&= [[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0, \qquad \text{by Proposition 9.4.}
\end{aligned}$$

Parts (3) and (4) are a little more tricky. Complete proofs can be found in Milnor [81] (Chapter II, Section 9), O'Neill [91] (Chapter III) and Kuhnel [71] (Chapter 6, Lemma 6.3). □

Part (3) of Proposition 16.3 can be interpreted as the fact that for every $p \in M$ and all $x, y \in T_pM$, the linear map $z \mapsto R(x, y)z$ (from T_pM to itself) is skew-symmetric. Indeed, for all $z, w \in T_pM$, we have

$$\langle R(x, y)z, w\rangle = R(x, y, z, w) = -R(x, y, w, z) = -\langle R(x, y)w, z\rangle = -\langle z, R(x, y)w\rangle.$$

The next proposition will be needed in the proof of the second variation formula. Recall the notion of a vector field along a surface given in Definition 15.11.

Proposition 16.4. *For a Riemannian manifold $(M, \langle -, - \rangle)$ equipped with the Levi-Civita connection, for every parametrized surface $\alpha\colon \mathbb{R}^2 \to M$, for every vector field $V \in \mathfrak{X}(M)$ along α, we have*

$$\frac{D}{\partial y}\frac{D}{\partial x}V - \frac{D}{\partial x}\frac{D}{\partial y}V = R\left(\frac{\partial \alpha}{\partial x}, \frac{\partial \alpha}{\partial y}\right)V.$$

Proof Sketch. This is Lemma 9.2 in Milnor [81] (Chapter II, Section 9.) Express both sides in local coordinates in a chart and make use of the identity

$$\nabla_{\frac{\partial}{\partial x_j}} \nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_k} - \nabla_{\frac{\partial}{\partial x_i}} \nabla_{\frac{\partial}{\partial x_j}} \frac{\partial}{\partial x_k} = R\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) \frac{\partial}{\partial x_k},$$

where this identity used the observation that $\left[\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right] = 0$. A more detailed proof is given in Do Carmo [39] (Chapter 4, Lemma 4.1). □

The curvature tensor is a rather complicated object. Thus, it is quite natural to seek simpler notions of curvature. The sectional curvature is indeed a simpler object, and it turns out that the curvature tensor can be recovered from it.

16.2 Sectional Curvature

Basically, the sectional curvature is the curvature of two-dimensional sections of our manifold. Given any two vectors $u, v \in T_pM$, recall by Cauchy-Schwarz that

$$\langle u, v\rangle_p^2 \leq \langle u, u\rangle_p \langle v, v\rangle_p,$$

with equality iff u and v are linearly dependent. Consequently, if u and v are linearly independent, we have

$$\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2 \neq 0.$$

In this case, we claim that the ratio

$$K_p(u, v) = \frac{R_p(u, v, u, v)}{\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2} = \frac{\langle R_p(u, v)u, v\rangle}{\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2}$$

is independent of the plane Π spanned by u and v.

If (x, y) is another basis of Π, then

$$\begin{aligned} x &= au + bv \\ y &= cu + dv. \end{aligned}$$

After some basic algebraic manipulations involving the symmetric bilinear form $\langle -, -\rangle$, we get

$$\langle x, x\rangle_p \langle y, y\rangle_p - \langle x, y\rangle_p^2 = (ad - bc)^2(\langle u, u\rangle_p \langle v, v\rangle_p - \langle u, v\rangle_p^2).$$

Similarly, the trilinear nature of R_p, along with Properties (1) and (3) given in Proposition 16.3, imply that

$$R_p(x, y, x, y) = \langle R_p(x, y)x, y\rangle_p = (ad - bc)^2 \langle R_p(u, v)u, v\rangle = (ad - bc)^2 R_p(u, v, u, v), \qquad (*)$$

which proves our assertion.

Note that skew-symmetry in w and z in $\langle R(x, y)z, w\rangle$ is crucial to obtain the expression in $(*)$, and this property requires the connection to be compatible with the metric. Thus the sectional curvature is not defined for an arbitrary connection.

Definition 16.4. Let $(M, \langle -, - \rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. For every $p \in T_pM$, for every 2-plane $\Pi \subseteq T_pM$, the *sectional curvature* $K_p(\Pi)$ of Π is given by

$$K_p(\Pi) = K_p(x, y) = \frac{R_p(x, y, x, y)}{\langle x, x\rangle_p \langle y, y\rangle_p - \langle x, y\rangle_p^2}, \tag{†}$$

for any basis (x, y) of Π.

As in the case of the curvature tensor, in order to keep notation to a minimum we often write $K(\Pi)$ instead of $K_p(\Pi)$ (or $K(x, y)$) instead of $K_p(x, y)$). Since $\Pi \subseteq T_pM$ $(x, y \in T_pM)$ for some $p \in M$, this rarely causes confusion.

Let us take a moment to compute the sectional curvature of S^2. By using the notation from Section 16.1 we find that

$$\begin{aligned} K\left(\frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}\right) &= \frac{R\left(\frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}, \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}\right)}{\langle \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \theta}\rangle \langle \frac{\partial p}{\partial \varphi}, \frac{\partial p}{\partial \varphi}\rangle - \langle \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}\rangle^2} \\ &= \frac{R\left(\frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}, \frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}\right)}{\sin^2\theta} = \frac{R_{1212}}{\sin^2\theta} = 1. \end{aligned}$$

Observe that if (x, y) is an orthonormal basis, then the denominator is equal to 1. The expression $R_p(x, y, x, y)$ (the numerator of (†)) is often denoted $\kappa_p(x, y)$. Remarkably, κ_p determines R_p. We denote the function $p \mapsto \kappa_p$ by κ. We state the following proposition without proof:

Proposition 16.5. *Let $(M, \langle -, - \rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. The function κ determines the curvature tensor R. Thus, the knowledge of all the sectional curvatures determines the curvature tensor. Moreover, for all $p \in M$, for all $x, y, w, z \in T_pM$, we have*

$$\begin{aligned} 6\langle R(x, y)z, w\rangle = \kappa(x + w, y + z) &- \kappa(x, y + z) - \kappa(w, y + z) \\ - \ \kappa(y + w, x + z) &+ \kappa(y, x + z) + \kappa(w, x + z) \\ - \ \kappa(x + w, y) &+ \kappa(x, y) + \kappa(w, y) \\ - \ \kappa(x + w, z) &+ \kappa(x, z) + \kappa(w, z) \\ + \ \kappa(y + w, x) &- \kappa(y, x) - \kappa(w, x) \\ + \ \kappa(y + w, z) &- \kappa(y, z) - \kappa(w, z). \end{aligned}$$

For a proof of this formidable equation, see Kuhnel [71] (Chapter 6, Theorem 6.5). A different proof of the above proposition (without an explicit formula) is also given in O'Neill [91] (Chapter III, Corollary 42).

Let

$$R_1(x, y)z = \langle x, z\rangle y - \langle y, z\rangle x.$$

Observe that

$$\langle R_1(x, y)x, y\rangle = \langle \langle x, x\rangle y - \langle x, y\rangle x, y\rangle = \langle x, x\rangle\langle y, y\rangle - \langle x, y\rangle^2,$$

which is the denominator of (†). As a corollary of Proposition 16.5, we get:

Proposition 16.6. *Let $(M, \langle -, -\rangle)$ be any Riemannian manifold equipped with the Levi-Civita connection. If the sectional curvature $K(\Pi)$ does not depend on the plane Π but only on $p \in M$, in the sense that K is a scalar function $K\colon M \to \mathbb{R}$, then*

$$R = K(p)R_1.$$

Proof. By hypothesis,

$$\kappa_p(x, y) = K(p)(\langle x, x\rangle_p\langle y, y\rangle_p - \langle x, y\rangle_p^2),$$

for all x, y. As the right-hand side of the formula in Proposition 16.5 consists of a sum of terms, we see that the right-hand side is equal to $K(p)$ times a similar sum with κ replaced by

$$\langle R_1(x, y)x, y\rangle = \langle x, x\rangle\langle y, y\rangle - \langle x, y\rangle^2,$$

so it is clear that $R = K(p)R_1$. □

In particular, in dimension $n = 2$, the assumption of Proposition 16.6 holds and K is the well-known *Gaussian curvature* for surfaces.

Definition 16.5. A Riemannian manifold $(M, \langle -, -\rangle)$ is said to have *constant (resp. negative, resp. positive) curvature* iff its sectional curvature is constant (resp. negative, resp. positive).

In dimension $n \geq 3$, we have the following somewhat surprising theorem due to F. Schur.

Proposition 16.7. *(F. Schur, 1886) Let $(M, \langle -, -\rangle)$ be a connected Riemannian manifold. If $\dim(M) \geq 3$ and if the sectional curvature $K(\Pi)$ does not depend on the plane $\Pi \subseteq T_pM$ but only on the point $p \in M$, then K is constant (i.e., does not depend on p).*

The proof, which is quite beautiful, can be found in Kuhnel [71] (Chapter 6, Theorem 6.7).

If we replace the metric $g = \langle -, -\rangle$ by the metric $\widetilde{g} = \lambda\langle -, -\rangle$ where $\lambda > 0$ is a constant, some simple calculations show that the Christoffel symbols and the Levi-Civita connection are unchanged, as well as the curvature tensor, but the sectional curvature is changed, with

$$\widetilde{K} = \lambda^{-1}K.$$

As a consequence, if M is a Riemannian manifold of constant curvature, by rescaling the metric, we may assume that either $K = -1$, or $K = 0$, or $K = +1$. Here are standard examples of spaces with constant curvature.

(1) The sphere $S^n \subseteq \mathbb{R}^{n+1}$ with the metric induced by $\mathbb{R}^{n+1}$, where

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid x_1^2 + \cdots + x_{n+1}^2 = 1\}.$$

The sphere S^n has constant sectional curvature $K = +1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}(n+1)$ on S^n is isomorphic to $\mathbf{SO}(n)$. Then it is easy to see that the action of $\mathbf{SO}(n)$ on T_pS^n is transitive on 2-planes and from this, it follows that $K = 1$ (for details, see Gallot, Hulin and Lafontaine [49] (Chapter 3, Proposition 3.14).

(2) Euclidean space $\mathbb{R}^{n+1}$ with its natural Euclidean metric. Of course, $K = 0$.

(3) The *hyperbolic space* $\mathcal{H}_n^+(1)$ from Definition 5.3. Recall that this space is defined in terms of the *Lorentz innner product* $\langle -, - \rangle_1$ on $\mathbb{R}^{n+1}$, given by

$$\langle (x_1, \ldots, x_{n+1}), (y_1, \ldots, y_{n+1}) \rangle_1 = -x_1y_1 + \sum_{i=2}^{n+1} x_iy_i.$$

By definition, $\mathcal{H}_n^+(1)$, written simply H^n, is given by

$$H^n = \{x = (x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} \mid \langle x, x \rangle_1 = -1,\ x_1 > 0\}.$$

Given any point $p = (x_1, \ldots, x_{n+1}) \in H^n$, since a tangent vector at p is defined as $x'(0)$ for any curve $x\colon (-\epsilon, \epsilon) \to H^n$ with $x(0) = p$, we note that

$$\frac{d}{dt}\langle x(t), x(t) \rangle_1 = 2\langle x'(t), x(t) \rangle_1 = \frac{d}{dt}(-1) = 0,$$

which by setting $t = 0$ implies that the set of tangent vectors $u \in T_pH^n$ are given by the equation

$$\langle p, u \rangle_1 = 0;$$

that is, T_pH^n is orthogonal to p with respect to the Lorentz inner-product. Since $p \in H^n$, we have $\langle p, p \rangle_1 = -1$, that is, p is timelike, so by Proposition 5.9, all vectors in T_pH^n are spacelike; that is,

$$\langle u, u \rangle_1 > 0, \qquad \text{for all } u \in T_pH^n,\ u \neq 0.$$

Therefore, the restriction of $\langle -, - \rangle_1$ to T_pH^n is positive, definite, which means that it is a metric on T_pH^n. The space H^n equipped with this metric g_H is called *hyperbolic space* and it has constant curvature $K = -1$. This can be shown by using the fact that the stabilizer of the action of $\mathbf{SO}_0(n, 1)$ on H^n is isomorphic to $\mathbf{SO}(n)$ (see Proposition 5.10). Then it is easy to see that the action of $\mathbf{SO}(n)$ on T_pH^n is transitive on 2-planes and from this, it follows that $K = -1$ (for details, see Gallot, Hulin and Lafontaine [49] (Chapter 3, Proposition 3.14).

There are other isometric models of H^n that are perhaps intuitively easier to grasp but for which the metric is more complicated. For example, there is a map $\mathrm{PD}\colon B^n \to H^n$ where $B^n = \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ is the open unit ball in $\mathbb{R}^n$, given by

$$\mathrm{PD}(x) = \left(\frac{1 + \|x\|^2}{1 - \|x\|^2}, \frac{2x}{1 - \|x\|^2} \right).$$

It is easy to check that $\langle \mathrm{PD}(x), \mathrm{PD}(x) \rangle_1 = -1$ and that PD is bijective and an isometry. One also checks that the pull-back metric $g_{\mathrm{PD}} = \mathrm{PD}^* g_H$ on B^n is given by

$$g_{\mathrm{PD}} = \frac{4}{(1 - \|x\|^2)^2}\,(dx_1^2 + \cdots + dx_n^2).$$

The metric g_{PD} is called the *conformal disc metric*, and the Riemannian manifold (B^n, g_{PD}) is called the *Poincaré disc model* or *conformal disc model*. See Problem 13.8. The metric g_{PD} is proportional to the Euclidean metric, and thus angles are preserved under the map PD. Another model is the *Poincaré half-plane model* $\{x \in \mathbb{R}^n \mid x_1 > 0\}$, with the metric

$$g_{\mathrm{PH}} = \frac{1}{x_1^2}\,(dx_1^2 + \cdots + dx_n^2).$$

We already encountered this space for $n = 2$.

In general, it is practically impossible to find an explicit formula for the sectional curvature of a Riemannian manifold. The spaces S^n, $\mathbb{R}^{n+1}$, and H^n are exceptions. Nice formulae can be given for Lie groups with bi-invariant metrics (see Chapter 20) and for certain kinds of reductive homogeneous manifolds (see Chapter 22).

The metrics for S^n, $\mathbb{R}^{n+1}$, and H^n have a nice expression in polar coordinates, but we prefer to discuss the Ricci curvature next.

16.3 Ricci Curvature

The Ricci tensor is another important notion of curvature. It is mathematically simpler than the sectional curvature (since it is symmetric), and it plays an important role in the theory of gravitation as it occurs in the Einstein field equations. The Ricci tensor is an example of contraction, in this case, the trace of a linear map. Recall that if $f\colon E \to E$ is a linear map from a finite-dimensional Euclidean vector space to itself, given any orthonormal basis $(e_1, \ldots, e_n)$, we have

$$\mathrm{tr}(f) = \sum_{i=1}^{n} \langle f(e_i), e_i \rangle.$$

Definition 16.6. Let $(M, \langle -, - \rangle)$ be a Riemannian manifold (equipped with any connection). The *Ricci curvature* Ric of M is the $(0, 2)$-tensor defined as follows. For every $p \in M$, for all $x, y \in T_pM$, set $\mathrm{Ric}_p(x, y)$ to be the trace of the endomorphism $v \mapsto R_p(x, v)y$. With respect to any orthonormal basis $(e_1, \ldots, e_n)$ of T_pM, we have

$$\mathrm{Ric}_p(x, y) = \sum_{j=1}^{n} \langle R_p(x, e_j)y, e_j\rangle_p = \sum_{j=1}^{n} R_p(x, e_j, y, e_j).$$

The *scalar curvature* S of M is the trace of the Ricci curvature; that is, for every $p \in M$,

$$S(p) = \sum_{i \neq j} R_p(e_i, e_j, e_i, e_j).$$

When the connection on M is the Levi-civita connection, the sectional curvature makes sense and then

$$S(p) = \sum_{i \neq j} R_p(e_i, e_j, e_i, e_j) = \sum_{i \neq j} K_p(e_i, e_j),$$

where $K_p(e_i, e_j)$ denotes the sectional curvature of the plane spanned by e_i, e_j.

In the interest of keeping notation to a minimum, we often write $\mathrm{Ric}(x, y)$ instead of $\mathrm{Ric}_p(x, y)$.

In a chart the Ricci curvature is given by

$$R_{ij} = \mathrm{Ric}\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) = \sum_m R^m_{ijm},$$

and the scalar curvature is given by

$$S(p) = \sum_{i,j} g^{ij} R_{ij},$$

where (g^{ij}) is the inverse of the Riemann metric matrix (g_{ij}). See O'Neill, pp. 87-88 [91]. For S^2, the calculations of Section 16.1 imply that

$$\begin{aligned}
\mathrm{Ric}\left(\frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \varphi}\right) &= R_{12} = \sum_{m=1}^{2} R^m_{12m} = R^1_{121} + R^2_{122} = 0\\
\mathrm{Ric}\left(\frac{\partial p}{\partial \varphi}, \frac{\partial p}{\partial \theta}\right) &= R_{21} = \sum_{m=1}^{2} R^m_{21m} = R^1_{211} + R^2_{212} = 0\\
\mathrm{Ric}\left(\frac{\partial p}{\partial \theta}, \frac{\partial p}{\partial \theta}\right) &= R_{11} = \sum_{m=1}^{2} R^m_{11m} = R^1_{111} + R^1_{112} = 1\\
\mathrm{Ric}\left(\frac{\partial p}{\partial \varphi}, \frac{\partial p}{\partial \varphi}\right) &= R_{22} = \sum_{m=1}^{2} R^m_{22m} = R^1_{221} + R^2_{222} = \sin^2\theta,
\end{aligned}$$

and that

$$\begin{aligned} S(p) &= \sum_{i=1}^{2}\sum_{j=1}^{2} g^{ij}R_{ij} = g^{11}R_{11} + g^{12}R_{12} + g^{21}R_{21} + g^{22}R_{22} \\ &= 1 \cdot 1 + \frac{1}{\sin^2\theta} \cdot \sin^2\theta = 2. \end{aligned}$$

If M is equipped with the Levi-Civita connection, in view of Proposition 16.3 (4), the Ricci curvature is symmetric. The tensor Ric is a $(0,2)$-tensor but it can be interpreted as a $(1,1)$-tensor as follows.

Definition 16.7. Let $(M, \langle -, - \rangle)$ be a Riemannian manifold (equipped with any connection). The $(1,1)$-tensor $\mathrm{Ric}^{\#}_p$ is defined to be

$$\langle \mathrm{Ric}^{\#}_p u, v\rangle_p = \mathrm{Ric}_p(u, v),$$

for all $u, v \in T_pM$.

Proposition 16.8. *Let (M, g) be a Riemannian manifold and let ∇ be any connection on M. If $(e_1, \ldots, e_n)$ is any orthonormal basis of T_pM, we have*

$$\mathrm{Ric}^{\#}_p(u) = \sum_{j=1}^{n} R_p(e_j, u)e_j.$$

Proof. We have

$$\begin{aligned} \mathrm{Ric}_p(u, v) &= \sum_{j=1}^{n} R_p(u, e_j, v, e_j) \\ &= \sum_{j=1}^{n} R_p(e_j, u, e_j, v) \\ &= \sum_{j=1}^{n} \langle R_p(e_j, u)e_j, v\rangle_p, \end{aligned}$$

so

$$\mathrm{Ric}^{\#}_p(u) = \sum_{j=1}^{n} R_p(e_j, u)e_j,$$

as claimed. □

Then it is easy to see that

$$S(p) = \mathrm{tr}(\mathrm{Ric}^{\#}_p).$$

This is why we said (by abuse of language) that S is the trace of Ric.

Observe that in dimension $n = 2$ (with the Levi Civita conection) we get $S(p) = 2K(p)$. Therefore, in dimension 2, the scalar curvature determines the curvature tensor. In dimension $n = 3$, it turns out that the Ricci tensor completely determines the curvature tensor, although this is not obvious. We will come back to this point later.

If the connection is the Levi-Civita connection, since $\mathrm{Ric}(x, y)$ is symmetric, $\mathrm{Ric}(x, x)$ determines $\mathrm{Ric}(x, y)$ completely (Use the polarization identity for a symmetric bilinear form, φ:

$$2\varphi(x, y) = \Phi(x + y) - \Phi(x) - \Phi(y),$$

with $\Phi(x) = \varphi(x, x)$). Observe that for any orthonormal frame $(e_1, \ldots, e_n)$ of T_pM, using the definition of the sectional curvature K, we have

$$\mathrm{Ric}(e_1, e_1) = \sum_{i=1}^{n} \langle R(e_1, e_i)e_1, e_i\rangle = \sum_{i=2}^{n} K(e_1, e_i).$$

Thus, $\mathrm{Ric}(e_1, e_1)$ is the sum of the sectional curvatures of any $n - 1$ orthogonal planes orthogonal to e_1 (a unit vector).

Proposition 16.9. *For a Riemannian manifold with constant sectional curvature (with the Levi-Civita connection), we have*

$$\mathrm{Ric}(x, x) = (n - 1)Kg(x, x), \qquad S = n(n - 1)K,$$

where $g = \langle -, -\rangle$ is the metric on M.

Proof. Indeed, if K is constant, then we know by Proposition 16.6 that $R = KR_1$, and so

$$\begin{aligned}
\mathrm{Ric}(x, x) &= K \sum_{i=1}^{n} g(R_1(x, e_i)x, e_i) \\
&= K \sum_{i=1}^{n} g(\langle x, x\rangle e_i - \langle e_i, x\rangle x, e_i) \\
&= K \sum_{i=1}^{n} (g(e_i, e_i)g(x, x) - g(e_i, x)^2) \\
&= K(ng(x, x) - \sum_{i=1}^{n} g(e_i, x)^2) \\
&= (n - 1)Kg(x, x),
\end{aligned}$$

as claimed. □

Spaces for which the Ricci tensor is proportional to the metric are called Einstein spaces.

Definition 16.8. A Riemannian manifold (M, g) is called an *Einstein space* iff the Ricci curvature is proportional to the metric g; that is:

$$\mathrm{Ric}(x, y) = \lambda g(x, y),$$

for some function $\lambda\colon M \to \mathbb{R}$.

If M is an Einstein space, observe that $S = n\lambda$.

Remark: For any Riemanian manifold (M, g), the quantity

$$G = \mathrm{Ric} - \frac{S}{2} g$$

is called the *Einstein tensor* (or *Einstein gravitation tensor* for space-times spaces). The Einstein tensor plays an important role in the theory of general relativity. For more on this topic, see Kuhnel [71] (Chapters 6 and 8) O'Neill [91] (Chapter 12).

16.4 The Second Variation Formula and the Index Form

As in previous sections, we assume that all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. In Section 15.6, we discovered that the geodesics are exactly the critical paths of the energy functional (Theorem 15.26). For this, we derived the first variation formula (Theorem 15.25). It is not too surprising that a deeper understanding is achieved by investigating the second derivative of the energy functional at a critical path (a geodesic). By analogy with the Hessian of a real-valued function on $\mathbb{R}^n$, it is possible to define a bilinear functional

$$I_\gamma\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$$

when γ is a critical point of the energy function E (that is, γ is a geodesic). This bilinear form is usually called the *index form*. Note that Milnor denotes I_γ by E_{**} and refers to it as the *Hessian* of E, but this is a bit confusing since I_γ is only defined for critical points, whereas the Hessian is defined for all points, critical or not.

Now, if $f\colon M \to \mathbb{R}$ is a real-valued function on a finite-dimensional manifold M and if p is a critical point of f, which means that $df_p = 0$, it turns out that there is a symmetric bilinear map $I_f\colon T_pM \times T_pM \to \mathbb{R}$ such that

$$I_f(X(p), Y(p)) = X_p(Yf) = Y_p(Xf),$$

for all vector fields $X, Y \in \mathfrak{X}(M)$. To show this, observe that for any two vector field X, Y,

$$X_p(Yf) - Y_p(Xf) = ([X, Y])_p(f) = df_p([X, Y]_p) = 0,$$

since p is a critical point, namely $df_p = 0$. It follows that the function $I_f \colon T_pM \times T_pM \to \mathbb{R}$ defined by

$$I_f(X(p), Y(p)) = X_p(Yf)$$

is bilinear and symmetric. Furthermore, $I_f(u, v)$ can be computed as follows: for any $u, v \in T_pM$, for any smooth map $\alpha \colon \mathbb{R}^2 \to M$ such that

$$\alpha(0,0) = p, \quad \frac{\partial \alpha}{\partial x}(0,0) = u, \quad \frac{\partial \alpha}{\partial y}(0,0) = v,$$

we have

$$I_f(u,v) = \left. \frac{\partial^2 (f \circ \alpha)(x,y)}{\partial x \partial y} \right|_{(0,0)} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}(f \circ \alpha) \right)_{(0,0)}.$$

The above suggests that in order to define

$$I_\gamma \colon T_\gamma \Omega(p,q) \times T_\gamma \Omega(p,q) \to \mathbb{R},$$

that is to define $I_\gamma(W_1, W_2)$, where $W_1, W_2 \in T_\gamma \Omega(p,q)$ are vector fields along γ (with $W_1(0) = W_2(0) = 0$ and $W_1(1) = W_2(1) = 0$), we consider 2-*parameter variations*

$$\alpha \colon U \times [0,1] \to M,$$

(see Definition 15.20), where U is an open subset of $\mathbb{R}^2$ with $(0,0) \in U$, such that

$$\alpha(0,0,t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0,0,t) = W_2(t).$$

See Figure 16.1.

Then we set

$$I_\gamma(W_1, W_2) = \left. \frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} \right|_{(0,0)},$$

where $\widetilde{\alpha} \in \Omega(p,q)$ is the path given by

$$\widetilde{\alpha}(u_1, u_2)(t) = \alpha(u_1, u_2, t).$$

For simplicity of notation, the above derivative if often written as $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$.

To prove that $I_\gamma(W_1, W_2)$ is actually well-defined, we need the following result.

Theorem 16.10. *(Second Variation Formula) Let $\alpha \colon U \times [0,1] \to M$ be a 2-parameter variation of a geodesic $\gamma \in \Omega(p,q)$, with variation vector fields $W_1, W_2 \in T_\gamma \Omega(p,q)$ given by*

$$W_1(t) = \frac{\partial \alpha}{\partial u_1}(0,0,t), \quad W_2(t) = \frac{\partial \alpha}{\partial u_2}(0,0,t), \quad \alpha(0,0,t) = \gamma(t).$$

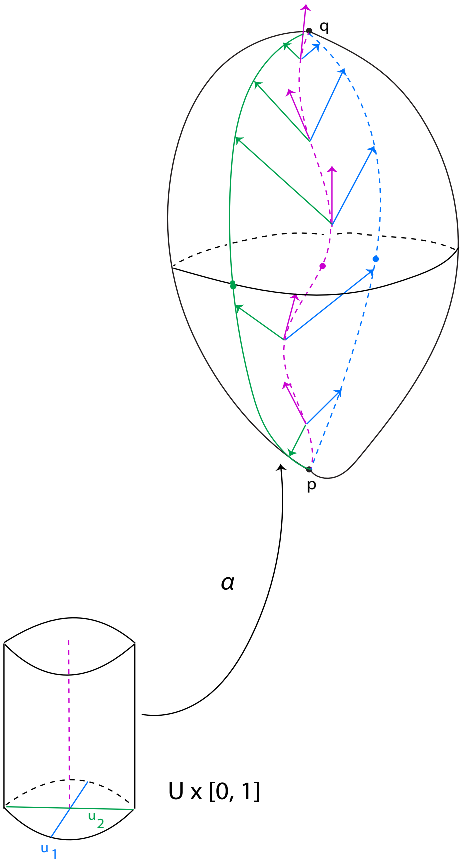


Figure 16.1: A 2-parameter variation α. The pink curve with its associated velocity field is $\alpha(0,0,t) = \gamma(t)$. The blue vector field is $W_1(t)$ while the green vector field is $W_2(t)$.

Then we have the formula

$$\frac{1}{2}\left.\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}\right|_{(0,0)} = -\sum_t \left\langle W_2(t), \Delta_t\frac{DW_1}{dt}\right\rangle - \int_0^1 \left\langle W_2, \frac{D^2W_1}{dt^2} + R(V,W_1)V\right\rangle dt,$$

where $V(t) = \gamma'(t)$ is the velocity field,

$$\Delta_t\frac{DW_1}{dt} = \frac{DW_1}{dt}(t_+) - \frac{DW_1}{dt}(t_-)$$

is the jump in $\frac{DW_1}{dt}$ at one of its finitely many points of discontinuity in $(0,1)$, and E is the energy function on $\Omega(p,q)$.

Proof. (After Milnor, see [81], Chapter II, Section 13, Theorem 13.1.) By the last line in the proof of the first variation formula (Theorem 15.25), we have

$$\frac{1}{2}\frac{\partial E(\widetilde{\alpha}(u_1,u_2))}{\partial u_2} = -\sum_i \left\langle \frac{\partial\alpha}{\partial u_2}, \Delta_t\frac{\partial\alpha}{\partial t}\right\rangle - \int_0^1 \left\langle \frac{\partial\alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.$$

Thus, we get

$$\begin{aligned}\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2} &= -\sum_i\left\langle\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2},\Delta_t\frac{\partial\alpha}{\partial t}\right\rangle-\sum_i\left\langle\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial u_1}\Delta_t\frac{\partial\alpha}{\partial t}\right\rangle\\ &\quad-\int_0^1\left\langle\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt-\int_0^1\left\langle\frac{\partial\alpha}{\partial u_2},\frac{D}{\partial u_1}\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}\right\rangle dt.\end{aligned}$$

Let us evaluate this expression for $(u_1,u_2)=(0,0)$. Since $\gamma=\widetilde{\alpha}(0,0)$ is an unbroken geodesic, we have

$$\Delta_t\frac{\partial\alpha}{\partial t}=0,\qquad\frac{D}{\partial t}\frac{\partial\alpha}{\partial t}=0,$$

so that the first and third term are zero. As

$$\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial t}=\frac{D}{\partial t}\frac{\partial\alpha}{\partial u_1},$$

(see the remark just after Proposition 16.4), we can rewrite the second term and we get

$$\frac{1}{2}\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}(0,0)=-\sum_i\left\langle W_2,\Delta_t\frac{D}{\partial t}W_1\right\rangle-\int_0^1\left\langle W_2,\frac{D}{\partial u_1}\frac{D}{\partial t}V\right\rangle dt.\qquad(*)$$

In order to interchange the operators $\frac{D}{\partial u_1}$ and $\frac{D}{\partial t}$, we need to bring in the curvature tensor. Indeed, by Proposition 16.4, we have

$$\frac{D}{\partial u_1}\frac{D}{\partial t}V-\frac{D}{\partial t}\frac{D}{\partial u_1}V=R\left(\frac{\partial\alpha}{\partial t},\frac{\partial\alpha}{\partial u_1}\right)V=R(V,W_1)V.$$

Together with the equation

$$\frac{D}{\partial u_1}V=\frac{D}{\partial u_1}\frac{\partial\alpha}{\partial t}=\frac{D}{\partial t}\frac{\partial\alpha}{\partial u_1}=\frac{D}{\partial t}W_1,$$

this yields

$$\frac{D}{\partial u_1}\frac{D}{\partial t}V=\frac{D^2W_1}{dt^2}+R(V,W_1)V.$$

Substituting this last expression in $(*)$, we get the second variation formula. □

Theorem 16.10 shows that the expression

$$\left.\frac{\partial^2(E\circ\widetilde{\alpha})(u_1,u_2)}{\partial u_1\partial u_2}\right|_{(0,0)}$$

only depends on the variation fields W_1 and W_2, and thus $I_\gamma(W_1,W_2)$ is actually well-defined. If no confusion arises, we write $I(W_1,W_2)$ for $I_\gamma(W_1,W_2)$.

Proposition 16.11. *Given any geodesic $\gamma \in \Omega(p,q)$, the map $I\colon T_\gamma\Omega(p,q) \times T_\gamma\Omega(p,q) \to \mathbb{R}$ defined so that for all $W_1, W_2 \in T_\gamma\Omega(p,q)$,*

$$I(W_1, W_2) = \left.\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2}\right|_{(0,0)},$$

only depends on W_1 and W_2 and is bilinear and symmetric, where $\alpha\colon U \times [0,1] \to M$ is any 2-parameter variation, with

$$\alpha(0,0,t) = \gamma(t), \quad \frac{\partial \alpha}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial \alpha}{\partial u_2}(0,0,t) = W_2(t).$$

Proof. We already observed that the second variation formula implies that $I(W_1, W_2)$ is well defined. This formula also shows that I is bilinear. As

$$\frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_1 \partial u_2} = \frac{\partial^2 (E \circ \widetilde{\alpha})(u_1, u_2)}{\partial u_2 \partial u_1},$$

I is symmetric (but this is not obvious from the right-hand side of the second variation formula). □

On the diagonal, $I(W, W)$ can be described in terms of a 1-parameter variation of γ. In fact,

$$I(W, W) = \frac{d^2 E(\widetilde{\alpha})}{du^2}(0),$$

where $\widetilde{\alpha}\colon (-\epsilon, \epsilon) \to \Omega(p,q)$ denotes any variation of γ with variation vector field $\frac{d\widetilde{\alpha}}{du}(0)$ equal to W. To prove this equation it is only necessary to introduce the 2-parameter variation

$$\widetilde{\beta}(u_1, u_2) = \widetilde{\alpha}(u_1 + u_2),$$

and to observe that

$$\frac{\partial \widetilde{\beta}}{\partial u_i} = \frac{d\widetilde{\alpha}}{du}, \qquad \frac{\partial^2 (E \circ \widetilde{\beta})}{\partial u_1 \partial u_2} = \frac{d^2 (E \circ \widetilde{\alpha})}{du^2},$$

where $u = u_1 + u_2$.

As an application of the above remark we have the following result.

Proposition 16.12. *If $\gamma \in \Omega(p,q)$ is a minimal geodesic, then the bilinear index form I is positive semi-definite, which means that $I(W, W) \geq 0$ for all $W \in T_\gamma\Omega(p,q)$.*

Proof. The inequality

$$E(\widetilde{\alpha}(u)) \geq E(\gamma) = E(\widetilde{\alpha}(0))$$

from Proposition 15.24 implies that

$$\frac{d^2 E(\widetilde{\alpha})}{du^2}(0) \geq 0,$$

which is exactly what needs to be proved. □

16.5 Jacobi Fields and Conjugate Points

Jacobi fields arise naturally when considering the expression involved under the integral sign in the second variation formula and also when considering the derivative of the exponential. *In this section, all manifolds under consideration are Riemannian manifolds equipped with the Levi-Civita connection.*

Definition 16.9. Let $B\colon E \times E \to \mathbb{R}$ be a symmetric bilinear form defined on some vector space E (possibly infinite dimentional). The *nullspace* of B is the subset $\mathrm{null}(B)$ of E given by

$$\mathrm{null}(B) = \{u \in E \mid B(u, v) = 0, \quad \text{for all } v \in E\}.$$

The *nullity* ν of B is the dimension of its nullspace. The bilinear form B is *nondegenerate* iff $\mathrm{null}(B) = (0)$ iff $\nu = 0$. If U is a subset of E, we say that B is *positive definite* (resp. *negative definite*) *on* U iff $B(u, u) > 0$ (resp. $B(u, u) < 0$) for all $u \in U$, with $u \neq 0$. The *index* of B is the maximum dimension of a subspace of E on which B is negative definite.

We will determine the nullspace of the symmetric bilinear form

$$I\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R},$$

where γ is a geodesic from p to q in some Riemannian manifold M. Now if W is a vector field in $T_\gamma\Omega(p, q)$ and if W satisfies the equation

$$\frac{D^2W}{dt^2} + R(V, W)V = 0, \tag{$*$}$$

where $V(t) = \gamma'(t)$ is the velocity field of the geodesic γ, since W is smooth along γ, (because γ is a geodesic and consists of a single smooth curve), it is obvious from the second variation formula that

$$I(W, W_2) = 0, \quad \text{for all } W_2 \in T_\gamma\Omega(p, q).$$

Therefore, any vector field W vanishing at 0 and 1 and satisfying equation $(*)$ belongs to the nullspace of I. More generally, a vector field (not necessarily vanishing at 0 and 1) satisfying equation $(*)$ is called *Jacobi field*.

Definition 16.10. Given a geodesic $\gamma \in \Omega(p, q)$, a vector field J along γ is a *Jacobi field* iff it satisfies the *Jacobi differential equation*

$$\frac{D^2J}{dt^2} + R(\gamma', J)\gamma' = 0. \tag{J}$$

Note that Definition 16.10 does not require that $J(0) = J(1) = 0$. The equation of Definition 16.10 is a linear second-order differential equation that can be transformed into a more familiar form by picking some *orthonormal parallel vector fields* $X_1, \ldots, X_n$ along γ. To do this, pick any orthonormal basis $(e_1, \ldots, e_n)$ in T_pM, with $e_1 = \gamma'(0)/\left\|\gamma'(0)\right\|$, and use parallel transport along γ to get $X_1, \ldots, X_n$. We can then write $J = \sum_{i=1}^n y_i X_i$, for some smooth functions y_i, and the Jacobi equation becomes the system of second-order linear ODE's

$$\frac{d^2 y_i}{dt^2} + \sum_{j=1}^n R(\gamma', X_j, \gamma', X_i) y_j = 0, \qquad 1 \leq i \leq n. \tag{$*$}$$

As an illustration of how to derive the preceding system of equations, suppose $J = y_1 X_1 + y_2 X_2$. Since $\frac{DX_i}{dt} = 0$ for all i, we find

$$\frac{DJ}{dt} = \frac{dy_1}{dt} X_1 + y_1 \frac{DX_1}{dt} + \frac{dy_2}{dt} X_2 + y_2 \frac{DX_2}{dt} = \frac{dy_1}{dt} X_1 + \frac{dy_2}{dt} X_2,$$

and hence

$$\frac{D^2 J}{dt^2} = \frac{d^2 y_1}{dt^2} X_1 + \frac{dy_1}{dt}\frac{DX_1}{dt} + \frac{d^2 y_2}{dt^2} X_2 + \frac{dy_2}{dt}\frac{DX_2}{dt} = \frac{d^2 y_1}{dt^2} X_1 + \frac{d^2 y_2}{dt^2} X_2.$$

We now compute

$$\begin{aligned} R(\gamma', J)\gamma' &= R(\gamma', y_1 X_1 + y_2 X_2)\gamma' = y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma' \\ &= c_1 X_1 + c_2 X_2, \end{aligned}$$

where the c_i are smooth functions determined as follows. Since $\langle X_1, X_1 \rangle = 1 = \langle X_2, X_2 \rangle$ and $\langle X_1, X_2 \rangle = 0$, we find that

$$\begin{aligned} c_1 &= \langle y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma', X_1 \rangle = y_1 \langle R(\gamma', X_1)\gamma', X_1 \rangle + y_2 \langle R(\gamma', X_2)\gamma, X_1 \rangle \\ &= y_1 R(\gamma', X_1, \gamma', X_1) + y_2 R(\gamma', X_2, \gamma', X_1), \end{aligned}$$

and that

$$\begin{aligned} c_2 &= \langle y_1 R(\gamma', X_1)\gamma' + y_2 R(\gamma', X_2)\gamma', X_2 \rangle = y_1 \langle R(\gamma', X_1)\gamma', X_2 \rangle + y_2 \langle R(\gamma', X_2)\gamma, X_2 \rangle \\ &= y_1 R(\gamma', X_1, \gamma', X_2) + y_2 R(\gamma', X_2, \gamma', X_2). \end{aligned}$$

These calculations show that the coefficient of X_1 is

$$\begin{aligned} \frac{d^2 y_1}{dt^2} + c_1 &= \frac{d^2 y_1}{dt^2} + y_1 R(\gamma', X_1, \gamma', X_1) + y_2 R(\gamma', X_2, \gamma', X_1) \\ &= \frac{d^2 y_1}{dt^2} + \sum_{j=1}^2 R(\gamma', X_j, \gamma', X_1) y_j, \end{aligned}$$

while the coefficient of X_2 is

$$\begin{aligned} \frac{d^2 y_2}{dt^2} + c_2 &= \frac{d^2 y_2}{dt^2} + y_1 R(\gamma', X_1, \gamma', X_2) + y_2 R(\gamma', X_2, \gamma', X_2) \\ &= \frac{d^2 y_2}{dt^2} + \sum_{j=1}^2 R(\gamma', X_j, \gamma', X_2) y_j. \end{aligned}$$

Setting these two coefficients equal to zero gives the systems of equations provided by $(*)$.

By the existence and uniqueness theorem for ODE's, for every pair of vectors $u, v \in T_pM$, there is a unique Jacobi field J so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Since T_pM has dimension n, it follows that the dimension of the space of Jacobi fields along γ is $2n$.

Proposition 16.13. *If $J(0)$ and $\frac{DJ}{dt}(0)$ are orthogonal to $\gamma'(0)$, then $J(t)$ is orthogonal to $\gamma'(t)$ for all $t \in [0, 1]$.*

Proof. Recall that by the remark after Proposition 16.3, the linear map $z \mapsto R(x, y)z$ is skew symmetric. As a consequence, it is a standard fact of linear algebra that $R(x, y)z$ is orthogonal to z; this is because skew-symmetry means that

$$\langle R(x, y)z, z\rangle = -\langle z, R(x, y)z\rangle,$$

which implies that $\langle R(x, y)z, z\rangle = 0$. Since X_1 is obtained by parallel transport along γ starting with $X_1(0)$ collinear to $\gamma'(0)$, the vector $X_1(t)$ is collinear to $\gamma'(t)$, and since $R(\gamma', X_j)\gamma'$ is orthogonal to γ', we have

$$R(\gamma', X_j, \gamma', X_1) = \langle R(\gamma', X_j)\gamma', X_1\rangle = 0.$$

But then the ODE for $J(t) = \sum_{i=1}^n y_i(t)X_i(t)$ given by $(*)$ yields

$$\frac{d^2y_1}{dt^2} = 0.$$

Since

$$J(0) = y_1(0)e_1 + \sum_{j=2}^n y_i(0)e_i = y_1(0)\frac{\gamma'(0)}{\|\gamma'(0)\|} + \sum_{j=2}^n y_i(0)e_i,$$

we find that

$$0 = \langle J(0), \gamma'(0)\rangle = \langle J(0), \|\gamma'(0)\|\, e_1\rangle = \|\gamma'(0)\|\, y_1(0)\langle e_1, e_1\rangle = \|\gamma'(0)\|\, y_1(0),$$

and hence conclude that $y_1(0) = 0$. Since $\frac{DX_i}{dt} = 0$,

$$\frac{DJ}{dt}(0) = \frac{dy_1}{dt}(0)e_1 + \sum_{j=2}^n \frac{dy_j}{dt}(0)e_j = \frac{dy_1}{dt}(0)\frac{\gamma'(0)}{\|\gamma'(0)\|} + \sum_{j=2}^n \frac{dy_j}{dt}(0)e_j,$$

and we again discover that

$$0 = \left\langle \frac{DJ}{dt}(0), \gamma'(0)\right\rangle = \left\langle \frac{DJ}{dt}(0), \|\gamma'(0)\|\, e_1\right\rangle = \|\gamma'(0)\|\, \frac{dy_1(0)}{dt},$$

and conclude that $\frac{dy_1}{dt}(0) = 0$. Because $y_1(0) = 0$ and $\frac{dy_1}{dt}(0) = 0$, the ODE $\frac{d^2y_1}{dt^2} = 0$ implies that $y_1(t) = 0$ for all $t \in [0, 1]$. In other words, $J(t) = \sum_{i=2}^n y_i(t)X_i(t)$, and since $X_2, \ldots, X_n$ are perpendicular to X_1, (which is collinear to γ'), we conclude that $J(t)$ is indeed orthogonal to $\gamma'(t)$ whenever $t \in [0, 1]$. □

Proposition 16.14. *If J is orthogonal to γ, which means that $J(t)$ is orthogonal to $\gamma'(t)$ for all $t \in [0,1]$, then $\frac{DJ}{dt}$ is also orthogonal to γ.*

Proof. Indeed, as γ is a geodesic, $\frac{D\gamma'}{dt} = 0$ and

$$0 = \frac{d}{dt}\left\langle J, \gamma' \right\rangle = \left\langle \frac{DJ}{dt}, \gamma' \right\rangle + \left\langle J, \frac{D\gamma'}{dt} \right\rangle = \left\langle \frac{DJ}{dt}, \gamma' \right\rangle,$$

as claimed. □

In other words, $\frac{DJ}{dt} = \sum_{i=2}^{n} \tilde{y}_i X_i$, where $\tilde{y}_i = \frac{dy_i}{dt}$. In summary, we have shown that the dimension of the space of Jacobi fields normal to γ is $2n-2$, and each such field is of the form $J = \sum_{i=2}^{n} y_i X_i$. These facts prove part of the following proposition.

Proposition 16.15. *If $\gamma \in \Omega(p,q)$ is a geodesic in a Riemannian manifold of dimension n, then the following properties hold.*

(1) For all $u, v \in T_pM$, there is a unique Jacobi fields J so that $J(0) = u$ and $\frac{DJ}{dt}(0) = v$. Consequently, the vector space of Jacobi fields has dimension $2n$.

(2) The subspace of Jacobi fields orthogonal to γ has dimension $2n-2$. The vector fields γ' and $t \mapsto t\gamma'(t)$ are Jacobi fields that form a basis of the subspace of Jacobi fields parallel to γ (that is, such that $J(t)$ is collinear with $\gamma'(t)$, for all $t \in [0,1]$.) See Figure 16.2.

(3) If J is a Jacobi field, then J is orthogonal to γ iff there exist $a, b \in [0,1]$, with $a \neq b$, so that $J(a)$ and $J(b)$ are both orthogonal to γ iff there is some $a \in [0,1]$ so that $J(a)$ and $\frac{DJ}{dt}(a)$ are both orthogonal to γ.

(4) For any two Jacobi fields X, Y along γ, the expression $\langle \nabla_{\gamma'} X, Y\rangle - \langle \nabla_{\gamma'} Y, X\rangle$ is a constant, and if X and Y vanish at some point on γ, then $\langle \nabla_{\gamma'} X, Y\rangle - \langle \nabla_{\gamma'} Y, X\rangle = 0$.

Proof. We already proved (1) and part of (2). If J is parallel to γ, then $J(t) = f(t)\gamma'(t)$ and $R(\gamma', J)\gamma' = fR(\gamma', \gamma')\gamma' = 0$, where the last equality follows from Proposition 16.3 (1). Since $\frac{D\gamma'}{dt} = 0$, we find that

$$\frac{DJ}{dt} = \frac{df}{dt}\gamma'(t) \qquad \frac{D^2 J}{dt^2} = \frac{d^2 f}{dt^2}\gamma'(t),$$

and the Jacobi differential equation of Definition 16.10 implies that

$$\frac{d^2 f}{dt^2} = 0.$$

Therefore,

$$J(t) = (\alpha + \beta t)\gamma'(t).$$

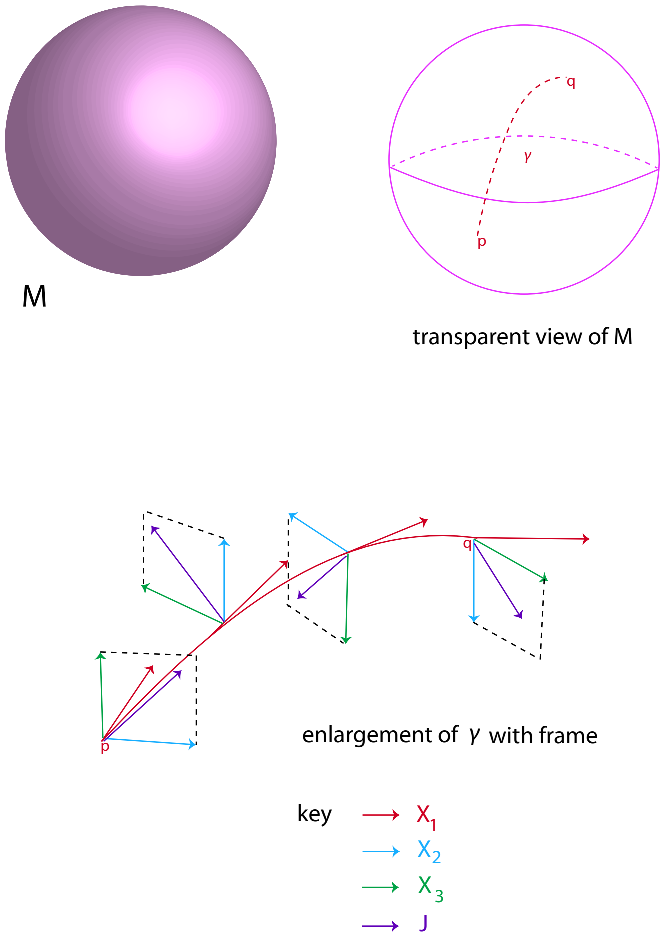


Figure 16.2: An orthogonal Jacobi field J for a three dimensional manifold M. Note that J is in the plane spanned by X_2 and X_3, while X_1 is in the direction of the velocity field.

It is easily shown that γ' and $t \mapsto t\gamma'(t)$ are linearly independent (as vector fields).

To prove (3), using the Jacobi differential equation of Definition 16.10, the fact that $\frac{D\gamma'}{dt} = 0$, and the fact that $R(x, y)z$ is orthogonal to z, observe that

$$\frac{d^2}{dt^2}\langle J, \gamma'\rangle = \left\langle \frac{D^2 J}{dt^2}, \gamma' \right\rangle = -\langle R(\gamma', J)\gamma', \gamma'\rangle = -R(J, \gamma', \gamma', \gamma') = 0.$$

Therefore,

$$\langle J, \gamma'\rangle = \alpha + \beta t$$

and the result follows. For example, if $\langle J(a), \gamma'(a)\rangle = \langle J(b), \gamma'(b)\rangle = 0$ with $a \neq b$, then $\alpha + \beta a = \alpha + \beta b = 0$, which implies $\alpha = \beta = 0$. We leave (4) as an exercise. □

Following Milnor, we will show that the Jacobi fields in $T_\gamma\Omega(p,q)$ are *exactly* the vector fields in the nullspace of the index form I. First, we define the important notion of conjugate points.

Definition 16.11. Let $\gamma \in \Omega(p,q)$ be a geodesic. Two distinct parameter values $a, b \in [0,1]$ with $a < b$ are *conjugate along* γ iff there is some Jacobi field J, not identically zero, such that $J(a) = J(b) = 0$. The dimension k of the space $\mathfrak{J}_{a,b}$ consisting of all such Jacobi fields is called the *multiplicity* (or *order of conjugacy)* of a and b as conjugate parameters. We also say that the points $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$ are *conjugate along* γ.

Remark: As remarked by Milnor and others, as γ may have self-intersections, the above definition is ambiguous if we replace a and b by $p_1 = \gamma(a)$ and $p_2 = \gamma(b)$, even though many authors make this slight abuse. Although it makes sense to say that the points p_1 and p_2 are conjugate, the space of Jacobi fields vanishing at p_1 and p_2 is not well defined. Indeed, if $p_1 = \gamma(a)$ for distinct values of a (or $p_2 = \gamma(b)$ for distinct values of b), then we don't know which of the spaces, $\mathfrak{J}_{a,b}$, to pick. We will say that some points p_1 and p_2 on γ are *conjugate* iff there are parameter values, $a < b$, such that $p_1 = \gamma(a)$, $p_2 = \gamma(b)$, and a and b are conjugate along γ.

However, for the endpoints p and q of the geodesic segment γ, we may assume that $p = \gamma(0)$ and $q = \gamma(1)$, so that when we say that p and q are conjugate we consider the space of Jacobi fields vanishing for $t = 0$ and $t = 1$. This is the definition adopted Gallot, Hulin and Lafontaine [49] (Chapter 3, Section 3E).

In view of Proposition 16.15 (3), the Jacobi fields involved in the definition of conjugate points are orthogonal to γ. The dimension of the space of Jacobi fields such that $J(a) = 0$ is obviously n, since the only remaining parameter determining J is $\frac{dJ}{dt}(a)$. Furthermore, the Jacobi field $t \mapsto (t-a)\gamma'(t)$ vanishes at a but not at b, so the multiplicity of conjugate parameters (points) is at most $n-1$.

For example, if M is a flat manifold, that is if its curvature tensor is identically zero, then the Jacobi equation becomes

$$\frac{D^2 J}{dt^2} = 0.$$

It follows that $J \equiv 0$, and thus, there are no conjugate points. More generally, the Jacobi equation can be solved explicitly for spaces of constant curvature; see Do Carmo [39] (Chapter 5, Example 2.3).

Theorem 16.16. *Let $\gamma \in \Omega(p,q)$ be a geodesic. A vector field $W \in T_\gamma\Omega(p,q)$ belongs to the nullspace of the index form I iff W is a Jacobi field. Hence, I is degenerate if p and q are conjugate. The nullity of I is equal to the multiplicity of p and q.*

Proof. (After Milnor [81], Theorem 14.1). We already observed that a Jacobi field vanishing at 0 and 1 belongs to the nullspace of I.

Conversely, assume that $W_1 \in T_\gamma\Omega(p,q)$ belongs to the nullspace of I. Pick a subdivision $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0,1]$ so that $W_1 \restriction [t_i, t_{i+1}]$ is smooth for all $i = 0, \ldots, k-1$, and let $f\colon [0,1] \to [0,1]$ be a smooth function which vanishes for the parameter values $t_0, \ldots, t_k$ and is strictly positive otherwise. Then if we let

$$W_2(t) = f(t)\left(\frac{D^2 W_1}{dt^2} + R(\gamma', W_1)\gamma'\right)_t,$$

by the second variation formula, we get

$$0 = -\frac{1}{2} I(W_1, W_2) = \sum 0 + \int_0^1 f(t) \left\| \frac{D^2 W_1}{dt^2} + R(\gamma', W_1)\gamma' \right\|^2 dt.$$

Consequently, $W_1 \restriction [t_i, t_{i+1}]$ is a Jacobi field for all $i = 0, \ldots, k-1$.

Now, let $W_2' \in T_\gamma\Omega(p,q)$ be a field such that

$$W_2'(t_i) = \Delta_{t_i} \frac{DW_1}{dt}, \qquad i = 1, \ldots, k-1.$$

We get

$$0 = -\frac{1}{2} I(W_1, W_2') = \sum_{i=1}^{k-1} \left\| \Delta_{t_i} \frac{DW_1}{dt} \right\|^2 + \int_0^1 0\, dt.$$

Hence, $\frac{DW_1}{dt}$ has no jumps. Now, a solution W_1 of the Jacobi equation is completely determined by the vectors $W_1(t_i)$ and $\frac{DW_1}{dt}(t_i)$, so the k Jacobi fields $W_1 \restriction [t_i, t_{i+1}]$ fit together to give a Jacobi field W_1 which is smooth throughout $[0,1]$. □

Theorem 16.16 implies that the nullity of I is finite, since the vector space of Jacobi fields vanishing at 0 and 1 has dimension at most n. In fact, we observed that the dimension of this space is at most $n-1$.

Corollary 16.17. *The nullity ν of I satisfies $0 \leq \nu \leq n-1$, where $n = \dim(M)$.*

As our (connected) Riemannian manifold M is a metric space, (see Proposition 15.14), the path space $\Omega(p,q)$ is also a metric space if we use the metric d^* given by

$$d^*(\omega_1, \omega_2) = \max_t(d(\omega_1(t), \omega_2(t))),$$

where d is the metric on M induced by the Riemannian metric.

Remark: The topology induced by d^* turns out to be the compact open topology on $\Omega(p,q)$.

Theorem 16.18. *Let $\gamma \in \Omega(p,q)$ be a geodesic. Then the following properties hold:*

(1) *If there are no conjugate points to p along γ, then there is some open subset $\mathcal{V}$ of $\Omega(p, q)$, with $\gamma \in \mathcal{V}$, such that*

$$L(\omega) \geq L(\gamma) \quad \textit{and} \quad E(\omega) \geq E(\gamma), \qquad \textit{for all } \omega \in \mathcal{V},$$

with strict inequality when $\omega([0, 1]) \neq \gamma([0, 1])$. We say that γ is a local minimum.

(2) *If there is some $t \in (0, 1)$ such that p and $\gamma(t)$ are conjugate along γ, then there is a fixed endpoints variation α, such that*

$$L(\widetilde{\alpha}(u)) < L(\gamma) \quad \textit{and} \quad E(\widetilde{\alpha}(u)) < E(\gamma), \qquad \textit{for } u \textit{ small enough.}$$

A proof of Theorem 16.18 can be found in Gallot, Hulin and Lafontaine [49] (Chapter 3, Theorem 3.73) or in O'Neill [91] (Chapter 10, Theorem 17 and Remark 18).

16.6 Jacobi Fields and Geodesic Variations

Jacobi fields turn out to be induced by certain kinds of variations called *geodesic variations*.

Definition 16.12. Given a geodesic $\gamma \in \Omega(p, q)$, a *geodesic variation of* γ is a smooth map

$$\alpha \colon (-\epsilon, \epsilon) \times [0, 1] \to M,$$

such that

(1) $\alpha(0, t) = \gamma(t)$, for all $t \in [0, 1]$.

(2) For every $u \in (-\epsilon, \epsilon)$, the curve $\widetilde{\alpha}(u)$ is a *geodesic*, where

$$\widetilde{\alpha}(u)(t) = \alpha(u, t), \qquad t \in [0, 1].$$

Note that the geodesics $\widetilde{\alpha}(u)$ do not necessarily begin at p and end at q, and so a geodesic variation is not a "fixed endpoints" variation. See Figure 16.3.

Proposition 16.19. *If $\alpha \colon (-\epsilon, \epsilon) \times [0, 1] \to M$ is a geodesic variation of $\gamma \in \Omega(p, q)$, then the vector field $W(t) = \frac{\partial \alpha}{\partial u}(0, t)$ is a Jacobi field along γ.*

Proof. As α is a geodesic variation, we have

$$\frac{D}{dt} \frac{\partial \alpha}{\partial t} = 0$$

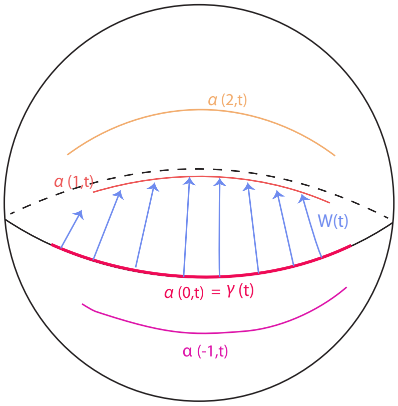


Figure 16.3: A geodesic variation for S^2 with its associated Jacobi field $W(t)$.

identically. Hence, using Proposition 16.4, we have

$$\begin{aligned} 0 &= \frac{D}{\partial u}\frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \\ &= \frac{D}{\partial t}\frac{D}{\partial u}\frac{\partial \alpha}{\partial t} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t} \\ &= \frac{D^2}{\partial t^2}\frac{\partial \alpha}{\partial u} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t}, \end{aligned}$$

where we used the equation

$$\frac{D}{\partial t}\frac{\partial \alpha}{\partial u} = \frac{D}{\partial u}\frac{\partial \alpha}{\partial t}$$

proved in Proposition 15.7. □

For example, on the sphere S^n, for any two antipodal points p and q, rotating the sphere keeping p and q fixed, the variation field along a geodesic γ through p and q (a great circle) is a Jacobi field vanishing at p and q. Rotating in $n-1$ different directions one obtains $n-1$ linearly independent Jacobi fields and thus, p and q are conjugate along γ with multiplicity $n-1$.

Interestingly, the converse of Proposition 16.19 holds.

Proposition 16.20. *For every Jacobi field $W(t)$ along a geodesic $\gamma \in \Omega(p, q)$, there is some geodesic variation $\alpha\colon (-\epsilon, \epsilon) \times [0, 1] \to M$ of γ such that $W(t) = \frac{\partial \alpha}{\partial u}(0, t)$. Furthermore, for every point $\gamma(a)$, there is an open subset U containing $\gamma(a)$ such that the Jacobi fields along a geodesic segment in U are uniquely determined by their values at the endpoints (in U) of the geodesic.*

Proof. (After Milnor, see [81], Chapter III, Lemma 14.4.) We begin by proving the second assertion. By Proposition 15.5 (1), there is an open subset U with $\gamma(0) \in U$, so that any two points of U are joined by a unique minimal geodesic which depends differentially on the endpoints. Suppose that $\gamma(t) \in U$ for $t \in [0, \delta]$. We will construct a Jacobi field W along $\gamma \restriction [0, \delta]$ with arbitrarily prescribed values u at $t = 0$ and v at $t = \delta$. Choose some curve $c_0 \colon (-\epsilon, \epsilon) \to U$ so that $c_0(0) = \gamma(0)$ and $c_0'(0) = u$, and some curve $c_\delta \colon (-\epsilon, \epsilon) \to U$ so that $c_\delta(0) = \gamma(\delta)$ and $c_\delta'(0) = v$. Now define the map

$$\alpha \colon (-\epsilon, \epsilon) \times [0, \delta] \to M$$

by letting $\widetilde{\alpha}(s) \colon [0, \delta] \to M$ be the unique minimal geodesic from $c_0(s)$ to $c_\delta(s)$. It is easily checked that α is a geodesic variation of $\gamma \restriction [0, \delta]$ and that

$$J(t) = \frac{\partial \alpha}{\partial u}(0, t)$$

is a Jacobi field such that $J(0) = u$ and $J(\delta) = v$. See Figure 16.4.

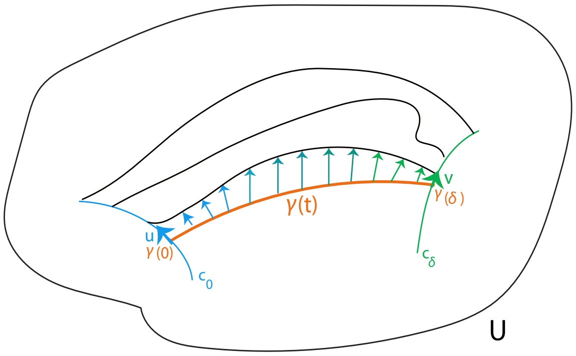


Figure 16.4: The local geodesic variation α with its Jacobi field such that $J(0) = u$ and $J(\delta) = v$.

We claim that every Jacobi field along $\gamma \restriction [0, \delta]$ can be obtained uniquely in this way. If $\mathfrak{J}_\delta$ denotes the vector space of all Jacobi fields along $\gamma \restriction [0, \delta]$, the map $J \mapsto (J(0), J(\delta))$ defines a linear map

$$\ell \colon \mathfrak{J}_\delta \to T_{\gamma(0)}M \times T_{\gamma(\delta)}M.$$

The above argument shows that ℓ is onto. However, both vector spaces have the same dimension $2n$, so ℓ is an isomorphism. Therefore, every Jacobi field in $\mathfrak{J}_\delta$ is determined by its values at $\gamma(0)$ and $\gamma(\delta)$, which is the content of the second assertion.

Now the above argument can be repeated for every point $\gamma(a)$ on γ, so we get an open cover $\{(l_a, r_a)\}$ of $[0, 1]$, such that every Jacobi field along $\gamma \restriction [l_a, r_a]$ is uniquely determined by its endpoints. By compactness of $[0,1]$, the above cover possesses some finite subcover, and we get a geodesic variation α defined on the entire interval $[0,1]$ whose variation field is equal to the original Jacobi field, W. □

Remark: The proof of Proposition 16.20 also shows that there is some open interval $(-\delta, \delta)$ such that if $t \in (-\delta, \delta)$, then $\gamma(t)$ is not conjugate to $\gamma(0)$ along γ. In fact, the Morse index theorem implies that for any geodesic segment, $\gamma\colon [0,1] \to M$, there are only finitely many points which are conjugate to $\gamma(0)$ along γ (see Milnor [81], Part III, Corollary 15.2).

Using Proposition 16.20 it is easy to characterize conjugate points in terms of geodesic variations; see O'Neill [91] (Chapter 10, Proposition 10).

Proposition 16.21. *If $\gamma \in \Omega(p,q)$ is a geodesic, then q is conjugate to p iff there is a geodesic variation α of γ such that every geodesic $\widetilde{\alpha}(u)$ starts from p, the Jacobi field $J(t) = \frac{\partial \alpha}{\partial u}(0,t)$ does not vanish identically, and $J(1) = 0$.*

Jacobi fields, as characterized by Proposition 16.19, can be used to compute the sectional curvature of the sphere S^n and the sectional curvature of hyperbolic space $H^n = \mathcal{H}_n^+(1)$, both equipped with their respective canonical metrics. This requires knowing the geodesics in S^n and H^n. This is done in Section 22.7 for the sphere. The hyperbolic space $H^n = \mathcal{H}_n^+(1)$ is shown to be a symmetric space in Section 22.9, and it would be easy to derive its geodesics by analogy with what we did for the sphere. For the sake of brevity, we will assume without proof that we know these geodesics. The reader may consult Gallot, Hulin and Lafontaine [49] or O'Neill [91] for details.

First we consider the sphere S^n. For any $p \in S^n$, the geodesic from p with initial velocity a unit vector v is

$$\gamma(t) = (\cos t)p + (\sin t)v.$$

Pick some unit vector $u \in T_pM$ orthogonal to v. The variation

$$\alpha(s,t) = (\cos t)p + (\sin t)((\cos s)v + (\sin s)u)$$

is a geodesic variation. We obtain the Jacobi vector field

$$Y(t) = \frac{\partial \alpha}{\partial s}(0,t) = (\sin t)u.$$

Since Y satisfies the Jacobi differential equation, we have

$$Y'' + R(\gamma', Y)\gamma' = 0.$$

But, as $Y(t) = (\sin t)u$, we have

$$Y + Y'' = 0,$$

so

$$R(\gamma', Y)\gamma' = Y,$$

which yields

$$1 = \langle u, u\rangle = \langle R(\gamma', u)\gamma', u\rangle = R(\gamma', u, \gamma', u)$$

since $\langle Y, Y\rangle = (\sin t)^2$ and $R(\gamma', Y, \gamma', Y) = (\sin t)^2 R(\gamma', u, \gamma', u)$. Since $\gamma'(0) = v$, it follows that $R(v, u, v, u) = 1$, which means that the sectional curvature of S^n is constant and equal to 1.

Let us now consider the hyperbolic space H^n. This time the geodesic from p with initial velocity a unit vector v is

$$\gamma(t) = (\cosh t)p + (\sinh t)v.$$

Pick some unit vector $u \in T_pM$ orthogonal to v. The variation

$$\alpha(s, t) = (\cosh t)p + (\sinh t)((\cosh s)v + (\sinh s)u)$$

is a geodesic variation and we obtain the Jacobi vector field

$$Y(t) = \frac{\partial \alpha}{\partial s}(0, t) = (\sinh t)u.$$

This time,

$$Y'' - Y = 0,$$

so the Jacobi equation becomes

$$R(\gamma', Y)\gamma' = -Y.$$

It follows that

$$-1 = -\langle u, u\rangle = \langle R(\gamma', u)\gamma', u\rangle = R(\gamma', u, \gamma', u)$$

and since $\gamma'(0) = v$, we get $R(v, u, v, u) = -1$, which means that the sectional curvature of H^n is constant and equal to -1.

Using the covering map of $\mathbb{RP}^n$ by S^n, it can be shown that $\mathbb{RP}^n$ with the canonical metric also has constant sectional curvature equal to $+1$; see Gallot, Hulin and Lafontaine [49] (Chapter III, section 3.49).

We end this section by exploiting Proposition 16.19 as means to develop intimate connections between Jacobi fields and the differential of the exponential map, and between conjugate points and critical points of the exponential map.

Recall that if $f\colon M \to N$ is a smooth map between manifolds, a point $p \in M$ is a *critical* point of f iff the tangent map at p

$$df_p\colon T_pM \to T_{f(p)}N$$

is not surjective. If M and N have the same dimension, which will be the case for the rest of this section, df_p is not surjective iff it is not injective, so p is a critical point of f iff there is some *nonzero* vector $u \in T_pM$ such that $df_p(u) = 0$.

If $\exp_p\colon T_pM \to M$ is the exponential map, for any $v \in T_pM$ where $\exp_p(v)$ is defined, we have the derivative of $\exp_p$ at v:

$$(d\exp_p)_v\colon T_v(T_pM) \to T_pM.$$

Since T_pM is a finite-dimensional vector space, $T_v(T_pM)$ is isomorphic to T_pM, so we identify $T_v(T_pM)$ with T_pM.

Jacobi fields can be used to compute the derivative of the exponential.

Proposition 16.22. *Given any point $p \in M$, for any vectors $u, v \in T_pM$, if $\exp_p v$ is defined, then*

$$J(t) = (d\exp_p)_{tv}(tu), \qquad 0 \leq t \leq 1,$$

is the unique Jacobi field such that $J(0) = 0$ and $\frac{DJ}{dt}(0) = u$.

Proof. We follow the proof in Gallot, Hulin and Lafontaine [49] (Chapter 3, Corollary 3.46). Another proof can be found in Do Carmo [39] (Chapter 5, Proposition 2.7). Let γ be the geodesic given by $\gamma(t) = \exp_p(tv)$. In T_pM equipped with the inner product g_p, the Jacobi field X along the geodesic $t \mapsto tv$ such that $X(0) = 0$ and $(DX/dt)(0) = u$ is just $X(t) = tu$. This Jacobi field is generated by the variation $H(s,t) = t(v + su)$ since $\frac{\partial H}{\partial s}H(0,t) = tu$; see Proposition 16.19. Because all the curves in this variation are radial geodesics, the variation $\alpha(s,t) = \exp_p H(s,t)$ of γ (in M) is also a geodesic variation, and by Proposition 16.19, the vector field $J(t) = \frac{\partial \alpha}{\partial s}(0,t)$ is a Jacobi vector field. See Figure 16.5.

By the chain rule we have $J(t) = (d\exp_p)_{tv}(tu)$, and since $J(0) = 0$ and $(DJ/dt)(0) = u$, we conclude that

$$J(t) = (d\exp_p)_{tv}(tu)$$

is the unique Jacobi field such that $J(0) = 0$ and $(DJ/dt)(0) = u$. □

Remark: If $u, v \in T_pM$ are orthogonal unit vectors, then $R(u, v, u, v) = K(u, v)$, the sectional curvature of the plane spanned by u and v in T_pM, and for t small enough, we have

$$\|J(t)\| = t - \frac{1}{6}K(u,v)t^3 + o(t^3).$$

(Here, $o(t^3)$ stands for an expression of the form $t^4R(t)$, such that $\lim_{t\mapsto 0} R(t) = 0$.) Intuitively, this formula tells us how fast the geodesics that start from p and are tangent to the plane spanned by u and v spread apart. Locally, for $K(u,v) > 0$ the radial geodesics spread apart less than the rays in T_pM, and for $K(u,v) < 0$ they spread apart more than the rays in T_pM. For more details, see Do Carmo [39] (Chapter 5, Proposition 2.7, Corollary 2.10 and the remark that follows.).

Jacobi fields can also be used to obtain a Taylor expansion for the matrix coefficients g_{ij} representing the metric g in a normal coordinate system near a point $p \in M$.

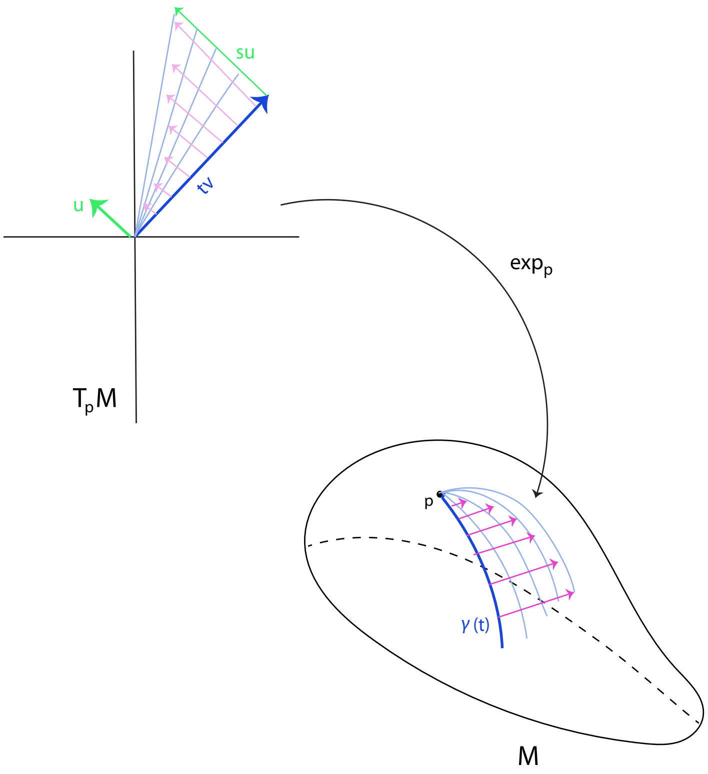


Figure 16.5: The radial geodesic variation and its image under $\exp_p$. Note that $J(t)$ is the dark pink vector field.

Proposition 16.23. *With respect to a normal coordinate system $x = (x_1, \ldots, x_n)$ around a point $p \in M$, the matrix coefficients g_{ij} representing the metric g near 0 are given by*

$$g_{ij}(x_1, \ldots, x_n) = \delta_{ij} + \frac{1}{3}\sum_{k,l} R_{ikjl}(p)\, x_k x_l + o(\|x\|^3).$$

A proof of Proposition 16.23 can be found in Sakai [100] (Chapter II, Section 3, Proposition 3.1). The above formula shows that the deviation of the Riemannian metric on M near p from the canonical Euclidean metric is measured by the curvature coefficients R_{ikjl}.

For any $x \neq 0$, write $x = tu$ with $u = x/\|x\|$ and $t = \|x\|$, where $x = (x_1, \ldots, x_n)$ are local coordinates at p. Proposition 16.23 can used used to give an expression for $\det(g_{ij}(tu))$ in terms of the Ricci curvature $\mathrm{Ric}_p(u, u)$.

Proposition 16.24. *With respect to a normal coordinate system $(x_1, \ldots, x_n) = tu$ with $\|u\| = 1$ around a point $p \in M$, we have*

$$\det(g_{ij}(tu)) = 1 - \frac{1}{3}\mathrm{Ric}_p(u, u)\, t^2 + o(t^3).$$

A proof of Proposition 16.24 can be found in Sakai [100] (Chapter II, Section 3, Lemma 3.5). The above formula shows that the Ricci curvature at p in the direction u is a measure of the deviation of the determinant $\det(g_{ij}(tu))$ to be equal to 1 (as in the case of the canonical Euclidean metric).

We now establish a relationship between conjugate points and critical points of the exponential map. These are points where the exponential is no longer a diffeomorphism.

Proposition 16.25. *Let $\gamma \in \Omega(p, q)$ be a geodesic. The point $r = \gamma(t)$, with $t \in (0, 1]$, is conjugate to p along γ iff $v = t\gamma'(0)$ is a critical point of $\exp_p$. Furthermore, the multiplicity of p and r as conjugate points is equal to the dimension of the kernel of $(d\exp_p)_v$.*

Proof. We follow the proof in Do Carmo [39] (Chapter 5, Proposition 3.5). Other proofs can be found in O'Neill [91] (Chapter 10, Proposition 10), or Milnor [81] (Part III, Theorem 18.1). The point $r = \gamma(t)$ is conjugate to p along γ if there is a non-zero Jacobi field J along γ such that $J(0) = 0$ and $J(t) = 0$. Let $v = \gamma'(0)$ and $w = (DJ/dt)(0)$. From Proposition 16.22, we have

$$J(t) = (d\exp_p)_{tv}(tw), \quad 0 \leq t \leq 1.$$

Observe that J is non-zero iff $(DJ/dt)(0) = w \neq 0$. Therefore, $r = \gamma(t)$ is conjugate to p along γ iff

$$0 = J(t) = (d\exp_p)_{tv}\left(t\frac{DJ}{dt}(0)\right), \quad \frac{DJ}{dt}(0) \neq 0;$$

that is, iff tv is a criticall point of $\exp_p$.

The multiplicity of p and r as conjugate points is equal to the number of linearly independent Jacobi fields $J_1, \ldots, J_k$ such that $J_i(0) = J_i(t) = 0$ for $i = 1, \ldots, k$.

We claim that $J_1, \ldots, J_k$ are linearly independent iff $(DJ_1/dt)(0), \ldots, (DJ_k/dt)(0)$ are linearly independent in T_pM.

Proof of claim. If $(DJ_1/dt)(0), \ldots, (DJ_k/dt)(0)$ are linearly independent, then $J_1, \ldots, J_k$ must be linearly independent since otherwise we would have

$$\lambda_1 J_1 + \cdots + \lambda_k J_k = 0$$

with some $\lambda_i \neq 0$, and by taking the derivative we would obtain a nontrivial dependency among $(DJ_1/dt)(0), \ldots, (DJ_k/dt)(0)$. Conversely, if $J_1, \ldots, J_k$ are linearly independent, then if we could express some $(DJ_i/dt)(0)$ as

$$\frac{DJ_i}{dt}(0) = \sum_{h \neq i} \lambda_h \frac{DJ_h}{dt}(0)$$

with some $\lambda_h \neq 0$, then the Jacobi field

$$J(t) = \sum_{h \neq i} \lambda_h J_h(t)$$

is such that $J(0) = 0$ and $(DJ/dt)(0) = (DJ_i/dt)(0)$, so by uniqueness $J = J_i$, and J_i is a nontrivial combination of the other J_h, a contradiction. □

Since

$$J_i(t) = (d\exp_p)_{tv}\left(t\frac{DJ_i}{dt}(0)\right),$$

we have $J_i(t) = 0$ iff $(DJ_i/dt)(0) \in \mathrm{Ker}\,(d\exp_p)_{tv}$, so the multiplicity of p and r is equal to the dimension of $\mathrm{Ker}\,(d\exp_p)_{tv}$. □

16.7 Topology and Curvature

As before, all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. Jacobi fields and conjugate points are basic tools that can be used to prove many global results of Riemannian geometry. The flavor of these results is that certain constraints on curvature (sectional, Ricci, scalar) have a significant impact on the topology. One may want consider the effect of non-positive curvature, constant curvature, curvature bounded from below by a positive constant, *etc*. This is a vast subject and we highly recommend Berger's Panorama of Riemannian Geometry [14] for a masterly survey. We will content ourselves with three results:

(1) Hadamard and Cartan's theorem about complete manifolds of non-positive sectional curvature.

(2) Myers' theorem about complete manifolds of Ricci curvature bounded from below by a positive number.

(3) The Morse index theorem.

First, on the way to Hadamard and Cartan, we begin with a proposition.

Proposition 16.26. *Let M be a Riemannian manifold such that $\langle R(u,v)u, v\rangle \leq 0$ for all $u, v \in T_pM$ and all $p \in M$, which is equivalent to saying that M has non-positive sectional curvature $K \leq 0$. For every geodesic $\gamma \in \Omega(p,q)$, there are no conjugate points to p along γ. Consequently, the exponential map $\exp_p\colon T_pM \to M$ is a local diffeomorphism for all $p \in M$.*

Proof. Let J be a Jacobi field along γ. Then,

$$\frac{D^2J}{dt^2} + R(\gamma', J)\gamma' = 0,$$

so that by the definition of the sectional curvature,

$$\left\langle \frac{D^2J}{dt^2}, J\right\rangle = -\langle R(\gamma', J)\gamma', J) = -R(\gamma', J, \gamma', J) \geq 0.$$

It follows that

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J\right\rangle = \left\langle \frac{D^2J}{dt^2}, J\right\rangle + \left\|\frac{DJ}{dt}\right\|^2 \geq 0.$$

Thus, the function $t \mapsto \left\langle \frac{DJ}{dt}, J \right\rangle$ is monotonic increasing, and strictly so if $\frac{DJ}{dt} \neq 0$. If J vanishes at both 0 and t, for any given $t \in (0,1]$, then so does $\left\langle \frac{DJ}{dt}, J \right\rangle$, and hence $\left\langle \frac{DJ}{dt}, J \right\rangle$ must vanish throughout the interval $[0,t]$. This implies

$$J(0) = \frac{DJ}{dt}(0) = 0,$$

so that J is identically zero. Therefore, t is not conjugate to 0 along γ. By Proposition 16.25, $d\exp_p$ is nonsingular for all $p \in M$, which implies that $\exp_p$ is a local diffeomorphism. □

Theorem 16.27. *(Hadamard–Cartan) Let M be a complete Riemannian manifold. If M has non-positive sectional curvature $K \leq 0$, then the following hold:*

(1) For every $p \in M$, the map $\exp_p\colon T_pM \to M$ is a Riemannian covering, i.e. $\exp_p$ is a smooth covering and a local isometry.

(2) If M is simply connected, then M is diffeomorphic to $\mathbb{R}^n$, where $n = \dim(M)$; more precisely, $\exp_p\colon T_pM \to M$ is a diffeomorphism for all $p \in M$. Furthermore, any two points on M are joined by a unique minimal geodesic.

Proof. We follow the proof in Sakai [100] (Chapter V, Theorem 4.1).

(1) By Proposition 16.26, the exponential map $\exp_p\colon T_pM \to M$ is a local diffeomorphism for all $p \in M$. Let $\widetilde{g}$ be the pullback metric $\widetilde{g} = (\exp_p)^*g$ on T_pM (where g denotes the metric on M). We claim that $(T_pM, \widetilde{g})$ is complete.

This is because, for every nonzero $u \in T_pM$, the line $t \mapsto tu$ is mapped to the geodesic $t \mapsto \exp_p(tu)$ in M, which is defined for all $t \in \mathbb{R}$ since M is complete, and thus this line is a geodesic in $(T_pM, \widetilde{g})$. Since this holds for all $u \in T_pM$, $(T_pM, \widetilde{g})$ is geodesically complete at 0, so by Hopf-Rinow, (Theorem 15.17), it is complete. But now, by Definition of the pullback metric (see Definition 13.4), $\exp_p\colon T_pM \to M$ is a local isometry, and by Proposition 17.7, it is a Riemannian covering map.

(2) If M is simply connected, then by Proposition 10.17, the covering map $\exp_p\colon T_pM \to M$ is a diffeomorphism (T_pM is connected). Therefore, $\exp_p\colon T_pM \to M$ is a diffeomorphism for all $p \in M$. □

Other proofs of Theorem 16.27 can be found in Do Carmo [39] (Chapter 7, Theorem 3.1), Gallot, Hulin and Lafontaine [49] (Chapter 3, Theorem 3.87), Kobayashi and Nomizu [69] (Chapter VIII, Theorem 8.1) and Milnor [81] (Part III, Theorem 19.2).

Remark: A version of Theorem 16.27 was first proved by Hadamard and then extended by Cartan.

Theorem 16.27 was generalized by Kobayashi, see Kobayashi and Nomizu [69] (Chapter VIII, Remark 2 after Corollary 8.2). Also, it is shown in Milnor [81] that if M is complete, assuming non-positive sectional curvature, then all homotopy groups $\pi_i(M)$ vanish for $i > 1$, and that $\pi_1(M)$ has no element of finite order except the identity. Finally, non-positive sectional curvature implies that the exponential map does not decrease distance (Kobayashi and Nomizu [69], Chapter VIII, Section 8, Lemma 3).

We now turn to manifolds with strictly positive curvature bounded away from zero and to Myers' theorem. The first version of such a theorem was first proved by Bonnet for surfaces with positive sectional curvature bounded away from zero. It was then generalized by Myers in 1941. For these reasons, this theorem is sometimes called the *Bonnet-Myers' theorem*. The proof of Myers theorem involves a beautiful "trick."

Given any metric space X, recall that the *diameter* of X is defined by

$$\mathrm{diam}(X) = \sup\{d(p,q) \mid p,q \in X\}.$$

The diameter of X may be infinite.

Theorem 16.28. *(Myers) Let M be a complete Riemannian manifold of dimension n and assume that*

$$\mathrm{Ric}(u,u) \geq (n-1)/r^2, \qquad \textit{for all unit vectors, } u \in T_pM, \textit{ and for all } p \in M,$$

with $r > 0$. Then,

(1) The diameter of M is bounded by πr and M is compact.

(2) The fundamental group of M is finite.

Proof. (1) Pick any two points $p, q \in M$ and let $d(p,q) = L$. As M is complete, by Hopf-Rinow theorem, (Theorem 15.16), there is a minimal geodesic γ joining p and q, and by Proposition 16.12, the bilinear index form I associated with γ is positive semi-definite, which means that $I(W,W) \geq 0$ for all vector fields $W \in T_\gamma\Omega(p,q)$. Pick an orthonormal basis $(e_1, \ldots, e_n)$ of T_pM, with $e_1 = \gamma'(0)/L$. Using parallel transport, we get a field of orthonormal frames $(X_1, \ldots, X_n)$ along γ, with $X_1(t) = \gamma'(t)/L$. Now comes Myers' beautiful trick. Define new vector fields Y_i along γ, by

$$W_i(t) = \sin(\pi t)X_i(t), \qquad 2 \leq i \leq n.$$

We have

$$\gamma'(t) = LX_1 \quad \text{and} \quad \frac{DX_i}{dt} = 0.$$

Furthermore, observe that

$$\frac{DW_i}{dt} = \pi\cos(\pi t)X_i, \qquad \frac{D^2W_i}{dt^2} = -\pi^2\sin(\pi t)X_i.$$

Then by the second variation formula,

$$\begin{aligned}
\frac{1}{2}\,I(W_i,W_i) &= -\int_0^1 \left\langle W_i, \frac{D^2W_i}{dt^2} + R(\gamma', W_i)\gamma' \right\rangle dt \\
&= -\int_0^1 \left\langle \sin(\pi t)X_i, -\pi^2\sin(\pi t)X_i + R(LX_1, \sin(\pi t)X_i)LX_1 \right\rangle dt \\
&= -\int_0^1 \left\langle \sin(\pi t)X_i, -\pi^2\sin(\pi t)X_i + L^2\sin(\pi t)R(X_1, X_i)X_1 \right\rangle dt \\
&= \int_0^1 (\sin(\pi t))^2(\pi^2 - L^2\left\langle R(X_1, X_i)X_1, X_i\right\rangle)dt,
\end{aligned}$$

for $i = 2, \ldots, n$. Adding up these equations and using the fact that

$$\mathrm{Ric}(X_1(t), X_1(t)) = \sum_{i=2}^{n} \langle R(X_1(t), X_i(t))X_1(t), X_i(t)\rangle,$$

we get

$$\frac{1}{2}\sum_{i=2}^{n} I(W_i, W_i) = \int_0^1 (\sin(\pi t))^2[(n-1)\pi^2 - L^2\,\mathrm{Ric}(X_1(t), X_1(t))]dt.$$

Now by hypothesis,

$$\mathrm{Ric}(X_1(t), X_1(t)) \geq (n-1)/r^2,$$

so

$$0 \leq \frac{1}{2}\sum_{i=2}^{n} I(W_i, W_i) \leq \int_0^1 (\sin(\pi t))^2\left[(n-1)\pi^2 - (n-1)\frac{L^2}{r^2}\right]dt,$$

which implies $\frac{L^2}{r^2} \leq \pi^2$, that is

$$d(p,q) = L \leq \pi r.$$

As the above holds for every pair of points $p, q \in M$, we conclude that

$$\mathrm{diam}(M) \leq \pi r.$$

Since closed and bounded subsets in a complete manifold are compact, M itself must be compact.

(2) Since the universal covering space $\widetilde{M}$ of M has the pullback of the metric on M, this metric satisfies the same assumption on its Ricci curvature as that of M. Therefore, $\widetilde{M}$ is also compact, which implies that the fundamental group $\pi_1(M)$ is finite (see the discussion at the end of Section 10.2). □

Remarks:

(1) The condition on the Ricci curvature cannot be weakened to $\mathrm{Ric}(u, u) > 0$ for all unit vectors. Indeed, the paraboloid of revolution $z = x^2 + y^2$ satisfies the above condition, yet it is not compact.

(2) Theorem 16.28 also holds under the stronger condition that the sectional curvature $K(u, v)$ satisfies

$$K(u, v) \geq (n - 1)/r^2,$$

for all orthonormal vectors, u, v. In this form, it is due to Bonnet (for surfaces).

It would be a pity not to include in this section a beautiful theorem due to Morse. This theorem has to do with the index of $I\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$, which is defined as follows.

Definition 16.13. For any geodesic $\gamma \in \Omega(p, q)$, we define the *index* λ of

$$I\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$$

as the maximum dimension of a subspace of $T_\gamma\Omega(p, q)$ on which I is negative definite.

Proposition 16.12 says that the index of I is zero for a minimal geodesic γ. It turns out that the index of I is finite for any geodesic γ.

Theorem 16.29. *(Morse Index Theorem) Given a geodesic $\gamma \in \Omega(p, q)$, the index λ of the index form $I\colon T_\gamma\Omega(p, q) \times T_\gamma\Omega(p, q) \to \mathbb{R}$ is equal to the number of points $\gamma(t)$, with $0 \leq t \leq 1$, such that $\gamma(t)$ is conjugate to $p = \gamma(0)$ along γ, each such conjugate point counted with its multiplicity. The index λ is always finite.*

As a corollary of Theorem 16.29, we see that there are only finitely many points which are conjugate to $p = \gamma(0)$ along γ.

A proof of Theorem 16.29 can be found in Milnor [81] (Part III, Section 15) and also in Do Carmo [39] (Chapter 11) or Kobayashi and Nomizu [69] (Chapter VIII, Section 6).

A key ingredient of the proof is that the vector space $T_\gamma\Omega(p, q)$ can be split into a direct sum of subspaces mutually orthogonal with respect to I, on one of which (denoted T') I is positive definite. Furthermore, the subspace orthogonal to T' is finite-dimensional. This space is obtained as follows. Since for every point $\gamma(t)$ on γ, there is some open subset U_t containing $\gamma(t)$ such that any two points in U_t are joined by a unique minimal geodesic, by compactness of $[0, 1]$, there is a subdivision $0 = t_0 < t_1 < \cdots < t_k = 1$ of $[0, 1]$ so that $\gamma \restriction [t_i, t_{i+1}]$ lies within an open set where it is a minimal geodesic.

Let $T_\gamma\Omega(t_0, \ldots, t_k) \subseteq T_\gamma\Omega(p, q)$ be the vector space consisting of all vector fields W along γ such that

(1) $W \restriction [t_i, t_{i+1}]$ is a Jacobi field along $\gamma \restriction [t_i, t_{i+1}]$, for $i = 0, \ldots, k - 1$.

(2) $W(0) = W(1) = 0$.

The space $T_\gamma\Omega(t_0, \ldots, t_k) \subseteq T_\gamma\Omega(p, q)$ is a finite-dimensional vector space consisting of broken Jacobi fields. Let $T' \subseteq T_\gamma\Omega(p, q)$ be the vector space consisting of all vector fields $W \in T_\gamma\Omega(p, q)$ for which

$$W(t_i) = 0, \qquad 0 \leq i \leq k.$$

It is not hard to prove that

$$T_\gamma\Omega(p, q) = T_\gamma\Omega(t_0, \ldots, t_k) \oplus T',$$

that $T_\gamma\Omega(t_0, \ldots, t_k)$ and T' are orthogonal *w.r.t* I, and that $I \restriction T'$ is positive definite. The reason why $I(W, W) \geq 0$ for $W \in T'$ is that each segment $\gamma \restriction [t_i, t_{i+1}]$ is a minimal geodesic, which has smaller energy than any other path between its endpoints.

As a consequence, the index (or nullity) of I is equal to the index (or nullity) of I restricted to the finite dimensional vector space $T_\gamma\Omega(t_0, \ldots, t_k)$. This shows that the index is always finite.

In the next section we will use conjugate points to give a more precise characterization of the cut locus.

16.8 Cut Locus and Injectivity Radius: Some Properties

As usual, all our manifolds are Riemannian manifolds equipped with the Levi-Civita connection. We begin by reviewing the definition of the cut locus provided by Definition 15.14 from a slightly different point of view. Let M be a complete Riemannian manifold of dimension n. There is a bundle UM, called the *unit tangent bundle*, such that the fibre at any $p \in M$ is the unit sphere $S^{n-1} \subseteq T_pM$ (check the details). As usual, we let $\pi \colon UM \to M$ denote the projection map which sends every point in the fibre over p to p.

Definition 16.14. The function $\rho \colon UM \to \mathbb{R}$ is defined so that for all $p \in M$, for all $v \in S^{n-1} \subseteq T_pM$,

$$\begin{aligned} \rho(v) &= \sup_{t \in \mathbb{R} \cup \{\infty\}} d(p, \exp_p(tv)) = t \\ &= \sup\{t \in \mathbb{R} \cup \{\infty\} \mid \text{the geodesic} \quad t \mapsto \exp_p(tv) \quad \text{is minimal on } [0, t]\}. \end{aligned}$$

The number $\rho(v)$ is called the *cut value* of v.

It can be shown that ρ is continuous; see Klingenberg [67] (Chapter 2, Lemma 2.1.5).

Definition 16.15. For every $p \in M$, we let

$$\widetilde{\mathrm{Cut}}(p) = \{\rho(v)v \in T_pM \mid v \in UM \cap T_pM,\ \rho(v) \text{ is finite}\}$$

be the *tangential cut locus of* p, and

$$\mathrm{Cut}(p) = \exp_p(\widetilde{\mathrm{Cut}}(p))$$

be the *cut locus of* p. The point $\exp_p(\rho(v)v)$ in M is called the *cut point* of the geodesic $t \mapsto \exp_p(vt)$, and so the cut locus of p is the set of cut points of all the geodesics emanating from p.

Also recall from Definition 15.14 that

$$\mathcal{U}_p = \{v \in T_pM \mid \rho(v) > 1\},$$

and that $\mathcal{U}_p$ is open and star-shaped. It follows from the definitions that

$$\widetilde{\mathrm{Cut}}(p) = \partial \mathcal{U}_p.$$

We also have the following property.

Theorem 16.30. *If M is a complete Riemannian manifold, then for every $p \in M$, the exponential map $\exp_p$ is a diffeomorphism between $\mathcal{U}_p$ and its image $\exp_p(\mathcal{U}_p) = M - \mathrm{Cut}(p)$ in M.*

Proof. The fact that $\exp_p$ is injective on $\mathcal{U}_p$ was shown in Proposition 15.19. Now for any $v \in \mathcal{U}$, as $t \mapsto \exp_p(tv)$ is a minimal geodesic for $t \in [0,1]$, by Theorem 16.18 (2), the point $\exp_p v$ is not conjugate to p, so $d(\exp_p)_v$ is bijective, which implies that $\exp_p$ is a local diffeomorphism. As $\exp_p$ is also injective, it is a diffeomorphism. □

Theorem 16.30 implies that the cut locus is closed.

Remark: In fact, $M - \mathrm{Cut}(p)$ can be retracted homeomorphically onto a ball around p, and $\mathrm{Cut}(p)$ is a deformation retract of $M - \{p\}$.

The following proposition gives a rather nice characterization of the cut locus in terms of minimizing geodesics and conjugate points.

Proposition 16.31. *Let M be a complete Riemannian manifold. For every pair of points $p, q \in M$, the point q belongs to the cut locus of p iff one of the two (not mutually exclusive from each other) properties hold:*

(a) There exist two distinct minimizing geodesics from p to q.

(b) There is a minimizing geodesic γ from p to q, and q is the first conjugate point to p along γ.

A proof of Proposition 16.31 can be found in Do Carmo [39] (Chapter 13, Proposition 2.2) Kobayashi and Nomizu [69] (Chapter VIII, Theorem 7.1) or Klingenberg [67] (Chapter 2, Lemma 2.1.11).

Observe that Proposition 16.31 implies the following symmetry property of the cut locus: $q \in \mathrm{Cut}(p)$ iff $p \in \mathrm{Cut}(q)$; see Do Carmo [39] (Chapter 13, Corollary 2.8). Furthermore, if M is compact, we have

$$p = \bigcap_{q \in \mathrm{Cut}(p)} \mathrm{Cut}(q);$$

see Klingenberg [67] (Chapter 2, Lemma 2.1.11).

Recall from Definition 15.15 the definition of the injectivity radius,

$$i(M) = \inf_{p \in M} d(p, \mathrm{Cut}(p)).$$

Proposition 16.31 admits the following sharpening.

Proposition 16.32. *Let M be a complete Riemannian manifold. For all $p, q \in M$, if $q \in \mathrm{Cut}(p)$, then*

(a) If among the minimizing geodesics from p to q, there is one, say γ, such that q is not conjugate to p along γ, then there is another minimizing geodesic $\omega \neq \gamma$ from p to q.

(b) Suppose $q \in \mathrm{Cut}(p)$ realizes the distance from p to $\mathrm{Cut}(p)$ (i.e. $d(p,q) = d(p, \mathrm{Cut}(p))$). If there are no minimal geodesics from p to q such that q is conjugate to p along this geodesic, then there are exactly two minimizing geodesics γ_1 and γ_2 from p to q, with $\gamma_2'(1) = -\gamma_1'(1)$. Moreover, if $d(p,q) = i(M)$ (the injectivity radius), then γ_1 and γ_2 together form a closed geodesic.

Except for the last statement, Proposition 16.32 is proved in Do Carmo [39] (Chapter 13, Proposition 2.12). The last statement is from Klingenberg [67] (Chapter 2, Lemma 2.1.11).

We conclude this section by stating a classical theorem of Klingenberg about the injectivity radius of a manifold of bounded positive sectional curvature.

Theorem 16.33. *(Klingenberg) Let M be a complete Riemannian manifold and assume that there are some positive constants $K_{\min}$, $K_{\max}$, such that the sectional curvature of K satisfies*

$$0 < K_{\min} \leq K \leq K_{\max}.$$

Then, M is compact, and either

(a) $i(M) \geq \pi/\sqrt{K_{\max}}$, or

(b) There is a closed geodesic γ of minimal length among all closed geodesics in M and such that

$$i(M) = \frac{1}{2} L(\gamma).$$

The proof of Theorem 16.33 is quite hard. A proof using Rauch's comparison theorem can be found in Do Carmo [39] (Chapter 13, Proposition 2.13).

16.9 Problems

Problem 16.1. Let M be a Riemannian manifold with the flat connection ∇. Prove that

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z = \nabla_{[X,Y]} Z,$$

for all $X, Y, Z \in \mathfrak{X}(M)$.

Problem 16.2. Let M be a Riemannian manifold equipped an arbitrary connection. In a chart, show that

$$R^l_{jhi} = \partial_i \Gamma^l_{hj} - \partial_h \Gamma^l_{ij} + \sum_m \Gamma^l_{im} \Gamma^m_{hj} - \sum_m \Gamma^l_{hm} \Gamma^m_{ij}.$$

Hint. See Gallot, Hulin and Lafontaine [49] (Chapter 3, Section 3.A.3) or O'Neill [91] (Chapter III, Lemma 38).

Problem 16.3. Prove Properties (3) and (4) of Proposition 16.3.
Hint. See Milnor [81] (Chapter II, Section 9), O'Neill [91] (Chapter III) or Kuhnel [71] (Chapter 6, Lemma 6.3).

Problem 16.4. Verify Equation $(*)$ of Section 16.2.

Problem 16.5. Prove Proposition 16.5.
Hint. See Kuhnel [71] (Chapter 6, Theorem 6.5).

Problem 16.6. Prove Proposition 16.4; that is, for a Riemannian manifold $(M, \langle -, - \rangle)$ equipped with the Levi-Civita connection, for every parametrized surface $\alpha\colon \mathbb{R}^2 \to M$, for every vector field $V \in \mathfrak{X}(M)$ along α, we have

$$\frac{D}{\partial y}\frac{D}{\partial x} V - \frac{D}{\partial x}\frac{D}{\partial y} V = R\left(\frac{\partial \alpha}{\partial x}, \frac{\partial \alpha}{\partial y}\right) V.$$

Hint. See Do Carmo [39], Chapter 4, Lemma 4.1.

Problem 16.7. Let M be a Riemannian manifold equipped with an arbitrary connection. In a chart, show that the Ricci curvature is given by

$$R_{ij} = \mathrm{Ric}\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j}\right) = \sum_m R^m_{ijm},$$

and that the sectional curvature is given by

$$S(p) = \sum_{i,j} g^{ij} R_{ij}.$$

Hint. See O'Neill, pp. 87-88 [91].

Problem 16.8. Prove Property (4) of Proposition 16.15.

Problem 16.9. Let M be a Riemannian manifold with the following property: for any two points $p, q \in M$, the parallel transport from p to q des not depend on the curve that joins p and q. Prove that the curvature of M is identically zero; that is, $R(X,Y)Z = 0$ for all $X, Y, Z \in \mathfrak{X}(M)$.

Problem 16.10. Let M be a Riemannian manifold of constant sectional curvature K, let $\gamma\colon [0,\ell] \to M$ be a geodesic on M parametrized by arc length, and let J be a Jacobi field along γ normal to γ'.

(1) Prove (using Proposition 16.6) that

$$R(\gamma', J)\gamma' = KJ.$$

Deduce from this that the Jacobi equation can be written as

$$\frac{D^2 J}{dt^2} + KJ = 0.$$

(2) If $w(t)$ is a parallel vector field along γ with $\langle \gamma'(t), w(t)\rangle = 0$ and $\|w(t)\| = 1$, prove that

$$J(t) = \begin{cases} \frac{\sin(t\sqrt{K})}{\sqrt{K}}\, w(t) & \text{if } K > 0 \\ tw(t) & \text{if } K = 0 \\ \frac{\sinh(t\sqrt{-K})}{\sqrt{-K}}\, w(t) & \text{if } K < 0 \end{cases}$$

is a solution of the Jacobi equation with initial condition $J(0) = 0$ and $J'(0) = w(0)$.

Problem 16.11. Given a Riemannian manifold M, for any point $p \in M$, the *conjugate locus* of P, denoted $C(p)$, is the set of all (first) conjugate points to p. Prove that if M has non-positive curvature, then $C(p)$ is empty for every $p \in M$.

Hint. Assume the existence of a non-trivial Jacobi field J along the geodesic $\gamma\colon [0,a] \to M$ with $\gamma(0) = p$, $J(0) = J(a) = 0$. Use the Jacobi equation to show that

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J \right\rangle \geq 0,$$

and then that

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J \right\rangle \equiv 0.$$

Compute

$$\frac{d}{dt}\langle J, J\rangle$$

to conclude that $\|J\|^2 = \text{constant} = 0$, a contradiction.

Problem 16.12. Let M be a Riemannian manifold with constant sectional curvature $b < 0$. Let $\gamma\colon [0, \ell] \to M$ be a geodesic parametrized by arc length, and let $v \in T_{\gamma(\ell)}M$ such that $\langle v, \gamma'(\ell)\rangle = 0$ and $\|v\| = 1$. Show that the Jacobi field J along γ determined by $J(0) = 0$, $J(\ell) = v$ is given by

$$J(t) = \frac{\sinh(t\sqrt{-b})}{\sinh(\ell\sqrt{-b})}\, w(t),$$

where $w(t)$ is the parallel transport along γ of the vector

$$w(0) = \frac{u_0}{\|u_0\|}, \quad u_0 = (d\exp_p)^{-1}_{\ell\gamma'(0)}(v).$$

Here, u_0 is considered as a vector in $T_{\gamma(0)}M$ by the identification $T_{\gamma(0)}M \cong T_{\ell\gamma'(0)}(T_{\gamma(0)}M)$.

Hint. Use Problem 16.3 and Proposition 16.22.

Problem 16.13. Let M be a Riemannian manifold. For any $p \in M$ and any $v \in T_pM$, let $\gamma\colon [0, a] \to M$ be a geodesic with $\gamma(0) = p$ and $\gamma'(0) = v$. For any $w \in T_v(T_pM)$ with $\|w\| = 1$ let J be a Jacobi field along γ given by

$$J(t) = (d\exp_p)_{tv}(tw), \quad 0 \le t \le a.$$

(1) Prove that

$$\nabla_{\gamma'}(R(\gamma', J)\gamma')(0) = R(\gamma', J')\gamma'(0).$$

(2) Prove that the Taylor expansion of $\|J(t)\|^2$ about $t = 0$ is given by

$$\|J(t)\|^2 = t^2 - \frac{1}{3}\langle R(v, w)v, w\rangle t^4 + o(t^4).$$

Conclude that if γ is parametrized by arc length, then

$$\|J(t)\|^2 = t^2 - \frac{1}{3}K(v, w)t^4 + o(t^4).$$

Problem 16.14. Let $f\colon M_1 \to M_2$ be a surjective local diffeomorphism of a manifold M_1 onto a Riemannian manifold M_2. Give M_1 a metric such that f becomes a local isometry. Give an example where M_2 is complete but M_1 is not complete.

Problem 16.15. A Riemannian manifold M is said to be homogenous if given any two points $p, q \in M$ there is an isometry of M which maps p to q. Prove that if M is homogeneous, then it is complete.

Problem 16.16. Let N_1 and N_2 be two closed disjoint submanifolds of a compact Riemannian manifold. Prove that the distance between N_1 and N_2 is assumed by a geodesic γ perpendicular to both N_1 and N_2.

Chapter 17

Isometries, Local Isometries, Riemannian Coverings and Submersions, Killing Vector Fields

The goal of this chapter is to understand the behavior of isometries and local isometries, in particular their action on geodesics. In Section 17.1 we show that isometries preserve the Levi-Civita connection. Local isometries preserve all concepts that are local in nature, such as geodesics, the exponential map, sectional, Ricci, and scalar curvature. In Section 17.2 we define Riemannian covering maps. These are smooth covering maps $\pi\colon M \to N$ that are also local isometries. There is a nice correspondence between the geodesics in M and the geodesics in N. We prove that if M is complete, N is connected, and $\pi\colon M \to N$ is a local isometry, then π is a Riemannian covering. In Section 17.3 we introduce Riemannian submersions. Given a submersion $\pi\colon M \to B$ between two Riemannian manifolds (M, g) and (B, h), for every $b \in B$ in the image of π, the fibre $\pi^{-1}(b)$ is a Riemannian submanifold of M, and for every $p \in \pi^{-1}(b)$, the tangent space T_pM to M at p splits into the two components

$$T_pM = \mathrm{Ker}\, d\pi_p \oplus (\mathrm{Ker}\, d\pi_p)^{\perp},$$

where $\mathcal{V}_p = \mathrm{Ker}\, d\pi_p$ is the *vertical subspace* of T_pM and $\mathcal{H}_p = (\mathrm{Ker}\, d\pi_p)^{\perp}$ (the orthogonal complement of $\mathcal{V}_p$ with respect to the metric g_p on T_pM) is the *horizontal subspace* of T_pM. If the map $d\pi_p$ is an isometry between the horizontal subspace $\mathcal{H}_p$ of T_pM and $T_{\pi(p)}B$ for every p, then π is a *Riemannian submersion.* In this case most of the differential geometry of B can be studied by "lifting" from B to M, and then projecting down to B again. In Section 17.4 we define Killing vector fields. A Killing vector field X satisfies the condition

$$X(\langle Y, Z\rangle) = \langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle,$$

for all $Y, Z \in \mathfrak{X}(M)$. A vector field X is a Killing vector field iff the diffeomorphisms Φ_t induced by the flow Φ of X are isometries (on their domain). Killing vector fields play an important role in the study of reductive homogeneous spaces; see Section 22.4.

17.1 Isometries and Local Isometries

Recall that a *local isometry* between two Riemannian manifolds M and N (necessarily of the same dimension) is a smooth map $\varphi\colon M \to N$ so that

$$\langle (d\varphi)_p(u), (d\varphi_p)(v)\rangle_{\varphi(p)} = \langle u, v\rangle_p,$$

for all $p \in M$ and all $u, v \in T_pM$. See Definition 13.5. An *isometry* is a local isometry and a diffeomorphism.

By the inverse function theorem, if $\varphi\colon M \to N$ is a local isometry, then for every $p \in M$, there is some open subset $U \subseteq M$ with $p \in U$ so that $\varphi \restriction U$ is an isometry between U and $\varphi(U)$.

Also recall by Definition 9.6 that if $\varphi\colon M \to N$ is a diffeomorphism, then for any vector field X on M, the vector field $\varphi_* X$ on N (called the *push-forward* of X) is given by

$$(\varphi_* X)_q = d\varphi_{\varphi^{-1}(q)} X(\varphi^{-1}(q)), \qquad \text{for all } q \in N,$$

or equivalently, by

$$(\varphi_* X)_{\varphi(p)} = d\varphi_p X(p), \qquad \text{for all } p \in M.$$

Proposition 17.1. *For any smooth function $h\colon N \to \mathbb{R}$, for any $q \in N$, we have*

$$(\varphi_* X)(h)_q = X(h \circ \varphi)_{\varphi^{-1}(q)},$$

or equivalently

$$(\varphi_* X)(h)_{\varphi(p)} = X(h \circ \varphi)_p. \qquad (*)$$

See Figure 17.1.

Proof. We have

$$\begin{aligned}(\varphi_* X)(h)_q &= dh_q((\varphi_* X)(q))\\ &= dh_q(d\varphi_{\varphi^{-1}(q)} X(\varphi^{-1}(q)))\\ &= d(h \circ \varphi)_{\varphi^{-1}(q)} X(\varphi^{-1}(q))\\ &= X(h \circ \varphi)_{\varphi^{-1}(q)},\end{aligned}$$

as claimed. □

It is natural to expect that isometries preserve all "natural" Riemannian concepts and this is indeed the case. We begin with the Levi-Civita connection.

Proposition 17.2. *If $\varphi\colon M \to N$ is an isometry, then*

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X}(\varphi_* Y), \qquad \text{for all } X, Y \in \mathfrak{X}(M),$$

where $\nabla_X Y$ is the Levi-Civita connection induced by the metric on M and similarly on N.

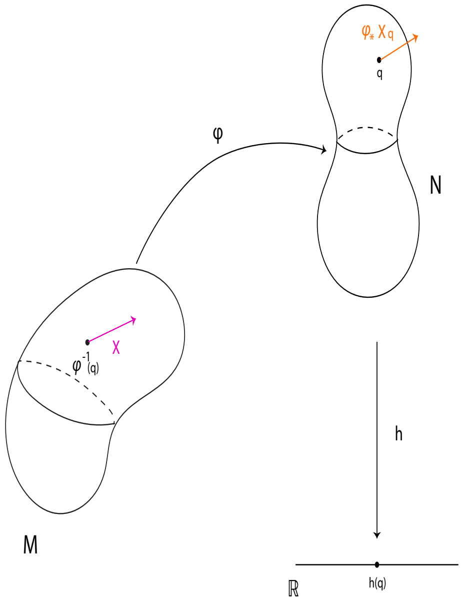


Figure 17.1: The push-forward of vector field X.

Proof. Let $X, Y, Z \in \mathfrak{X}(M)$. A proof can be found in O'Neill [91] (Chapter 3, Proposition 59), but we find it instructive to give a proof using the Koszul formula (Proposition 14.9),

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad - \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle. \end{aligned}$$

We have

$$(\varphi_*(\nabla_X Y))_{\varphi(p)} = d\varphi_p(\nabla_X Y)_p,$$

and as φ is an isometry,

$$\langle d\varphi_p(\nabla_X Y)_p, d\varphi_p Z_p\rangle_{\varphi(p)} = \langle (\nabla_X Y)_p, Z_p\rangle_p, \qquad (**)$$

so Koszul yields

$$\begin{aligned} 2\langle \varphi_*(\nabla_X Y), \varphi_* Z\rangle_{\varphi(p)} &= 2\langle d\varphi_p(\nabla_X Y)_p, d\varphi_p Z_p\rangle_{\varphi(p)} = 2\langle (\nabla_X Y)_p, Z_p\rangle_p \\ &= X(\langle Y, Z\rangle_p) + Y(\langle X, Z\rangle_p) - Z(\langle X, Y\rangle_p) \\ &\quad - \langle Y, [X, Z]\rangle_p - \langle X, [Y, Z]\rangle_p - \langle Z, [Y, X]\rangle_p. \end{aligned}$$

Next we need to compute

$$\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)}.$$

When we plug $\varphi_* X$, $\varphi_* Y$ and $\varphi_* Z$ into the Koszul formula, as φ is an isometry, for the fourth term on the right-hand side we get

$$\begin{aligned}
\langle \varphi_* Y, [\varphi_* X, \varphi_* Z]\rangle_{\varphi(p)} &= \langle d\varphi_p Y_p, [d\varphi_p X_p, d\varphi_p Z_p]\rangle_{\varphi(p)} \\
&= \langle d\varphi_p Y_p, d\varphi_p [X_p, Z_p]\rangle_{\varphi(p)}, \qquad \text{by Proposition 9.6} \\
&= \langle Y_p, [X_p, Z_p]\rangle_p, \qquad \text{by } (**)
\end{aligned}$$

and similarly for the fifth and sixth term on the right-hand side. For the first term on the right-hand side, we get

$$\begin{aligned}
(\varphi_* X)(\langle \varphi_* Y, \varphi_* Z\rangle)_{\varphi(p)} &= (\varphi_* X)(\langle d\varphi_p Y_p, d\varphi_p Z_p\rangle)_{\varphi(p)} \\
&= (\varphi_* X)(\langle Y_p, Z_p\rangle_{\varphi^{-1}(\varphi(p))})_{\varphi(p)}, \qquad \text{by } (**) \\
&= (\varphi_* X)(\langle Y, Z\rangle \circ \varphi^{-1})_{\varphi(p)} \\
&= X(\langle Y, Z\rangle \circ \varphi^{-1} \circ \varphi)_p, \qquad \text{by } (*) \\
&= X(\langle Y, Z\rangle)_p,
\end{aligned}$$

and similarly for the second and third term. Consequently, we get

$$\begin{aligned}
2\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)} &= X(\langle Y, Z\rangle_p) + Y(\langle X, Z\rangle_p) - Z(\langle X, Y\rangle_p) \\
&\quad - \langle Y, [X, Z]\rangle_p - \langle X, [Y, Z]\rangle_p - \langle Z, [Y, X]\rangle_p.
\end{aligned}$$

By comparing right-hand sides, we get

$$2\langle \varphi_*(\nabla_X Y), \varphi_* Z\rangle_{\varphi(p)} = 2\langle \nabla_{\varphi_* X}(\varphi_* Y), \varphi_* Z\rangle_{\varphi(p)}$$

for all X, Y, Z, and as φ is a diffeomorphism, this implies

$$\varphi_*(\nabla_X Y) = \nabla_{\varphi_* X}(\varphi_* Y),$$

as claimed. □

As a corollary of Proposition 17.2, the curvature induced by the connection is preserved; that is

$$\varphi_* R(X, Y)Z = R(\varphi_* X, \varphi_* Y)\varphi_* Z,$$

as well as the parallel transport, the covariant derivative of a vector field along a curve, the exponential map, sectional curvature, Ricci curvature and geodesics.

Actually, all concepts that are local in nature are preserved by local isometries! So, except for the Levi-Civita connection and the Riemann tensor on vector fields, all the above concepts are preserved under local isometries. For the record we state:

Proposition 17.3. *If $\varphi\colon M \to N$ is a local isometry between two Riemannian manifolds equipped with the Levi-Civita connection, then the following concepts are preserved:*

(1) *The covariant derivative of vector fields along a curve γ; that is*

$$d\varphi_{\gamma(t)}\frac{DX}{dt} = \frac{D\varphi_* X}{dt},$$

for any vector field X along γ, with $(\varphi_* X)(t) = d\varphi_{\gamma(t)}X(t)$, for all t.

(2) *Parallel translation along a curve. If P_γ denotes parallel transport along the curve γ (in M) and if $P_{\varphi\circ\gamma}$ denotes parallel transport along the curve $\varphi\circ\gamma$ (in N), then*

$$d\varphi_{\gamma(1)} \circ P_\gamma = P_{\varphi\circ\gamma} \circ d\varphi_{\gamma(0)}.$$

(3) *Geodesics. If γ is a geodesic in M, then $\varphi\circ\gamma$ is a geodesic in N. Thus, if γ_v is the unique geodesic with $\gamma(0) = p$ and $\gamma_v'(0) = v$, then*

$$\varphi \circ \gamma_v = \gamma_{d\varphi_p v},$$

wherever both sides are defined. Note that the domain of $\gamma_{d\varphi_p v}$ may be strictly larger than the domain of γ_v. For example, consider the inclusion of an open disc into $\mathbb{R}^2$.

(4) *Exponential maps. We have*

$$\varphi \circ \exp_p = \exp_{\varphi(p)} \circ d\varphi_p,$$

wherever both sides are defined. See Figure 17.2.

(5) *Riemannian curvature tensor. We have*

$$d\varphi_p R(x, y)z = R(d\varphi_p x, d\varphi_p y)d\varphi_p z, \qquad \text{for all } x, y, z \in T_pM.$$

(6) *Sectional, Ricci, and Scalar curvature. We have*

$$K(d\varphi_p x, d\varphi_p y) = K(x, y)_p,$$

for all linearly independent vectors $x, y \in T_pM$;

$$\mathrm{Ric}(d\varphi_p x, d\varphi_p y) = \mathrm{Ric}(x, y)_p$$

for all $x, y \in T_pM$;

$$S_M = S_N \circ \varphi.$$

where S_M is the scalar curvature on M and S_N is the scalar curvature on N.

A useful property of local isometries is stated below. For a proof, see O'Neill [91] (Chapter 3, Proposition 62):

Proposition 17.4. *Let $\varphi, \psi\colon M \to N$ be two local isometries. If M is connected and if $\varphi(p) = \psi(p)$ and $d\varphi_p = d\psi_p$ for some $p \in M$, then $\varphi = \psi$.*

The idea is to prove that

$$\{p \in M \mid d\varphi_p = d\psi_p\}$$

is both open and closed, and for this, to use the preservation of the exponential under local diffeomorphisms.

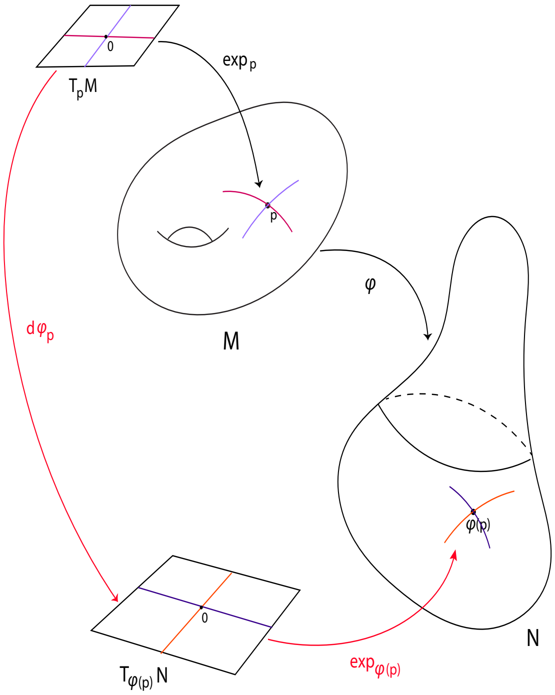


Figure 17.2: An illustration of $\varphi \circ \exp_p = \exp_{\varphi(p)} \circ d\varphi_p$. The composition of the black maps agrees with the composition of the red maps.

17.2 Riemannian Covering Maps

The notion of covering map discussed in Section 10.2 (see Definition 10.6) can be extended to Riemannian manifolds.

Definition 17.1. If M and N are two Riemannian manifold, then a map $\pi\colon M \to N$ is a *Riemannian covering* iff the following conditions hold:

(1) The map π is a smooth covering map.

(2) The map π is a local isometry.

Recall from Section 10.2 that a covering map is a local diffeomorphism. A way to obtain a metric on a manifold M is to pull-back the metric g on a manifold N along a local diffeomorphism $\varphi\colon M \to N$ (see Section 13.2). If φ is a covering map, then it becomes a Riemannian covering map.

Proposition 17.5. *Let $\pi\colon M \to N$ be a smooth covering map. For any Riemannian metric g on N, there is a unique metric $\pi^* g$ on M, so that π is a Riemannian covering.*

Proof. We define the *pull-back metric* $\pi^* g$ on M induced by g as follows. For all $p \in M$, for all $u, v \in T_pM$,

$$(\pi^* g)_p(u, v) = g_{\pi(p)}(d\pi_p(u), d\pi_p(v)).$$

We need to check that $(\pi^* g)_p$ is an inner product, which is very easy since $d\pi_p$ is a linear isomorphism. Our map π between the two Riemannian manifolds $(M, \pi^* g)$ and (N, g) becomes a local isometry. Every metric on M making π a local isometry has to satisfy the equation defining $\pi^* g$, so this metric is unique. □

In general, if $\pi\colon M \to N$ is a smooth covering map, a metric on M does not induce a metric on N such that π is a Riemannian covering. However, if N is obtained from M as a quotient by some suitable group action (by a group G) on M, then the projection $\pi\colon M \to M/G$ is a Riemannian covering.

In the rest of this section we assume that our Riemannian manifolds are equipped with the Levi-Civita connection. Because a Riemannian covering map is a local isometry, we have the following useful result proved in Gallot, Hulin, Lafontaine [49] (Chapter 2, Proposition 2.81).

Proposition 17.6. *Let $\pi\colon M \to N$ be a Riemannian covering. Then the geodesics of (N, h) are the projections of the geodesics of (M, g) (curves of the form $\pi \circ \gamma$, where γ is a geodesic in M), and the geodesics of (M, g) are the liftings of the geodesics of (N, h) (curves γ in M such that $\pi \circ \gamma$ is a geodesic of (N, h)).*

As a corollary of Proposition 17.5 and Theorem 10.14, every connected Riemannian manifold M has a simply connected covering map $\pi\colon \widetilde{M} \to M$, where π is a Riemannian covering. Furthermore, if $\pi\colon M \to N$ is a Riemannian covering and $\varphi\colon P \to N$ is a local isometry, it is easy to see that its lift $\widetilde{\varphi}\colon P \to M$ is also a local isometry. See Proposition 10.13. In particular, the deck-transformations of a Riemannian covering are isometries.

In general a local isometry is not a Riemannian covering. However, this is the case when the source space is complete.

Proposition 17.7. *Let $\pi\colon M \to N$ be a local isometry with N connected. If M is a complete manifold, then π is a Riemannian covering map.*

Proof. We follow the proof in Sakai [100] (Chapter III, Theorem 5.4). Because π is a local isometry, Proposition 17.6 implies that geodesics in M can be projected onto geodesics in N and that geodesics in N can be lifted back to M. The proof makes heavy use of these facts.

First we prove that N is complete. Pick any $p \in M$ and let $q = \pi(p)$. For any geodesic γ_v of N with initial point $q \in N$ and initial direction the unit vector $v \in T_qN$, consider the geodesic $\widetilde{\gamma}_u$ of M with initial point p, and with $u = d\pi_q^{-1}(v) \in T_pM$. As π is a local isometry, it preserves geodesics, so

$$\gamma_v = \pi \circ \widetilde{\gamma}_u,$$

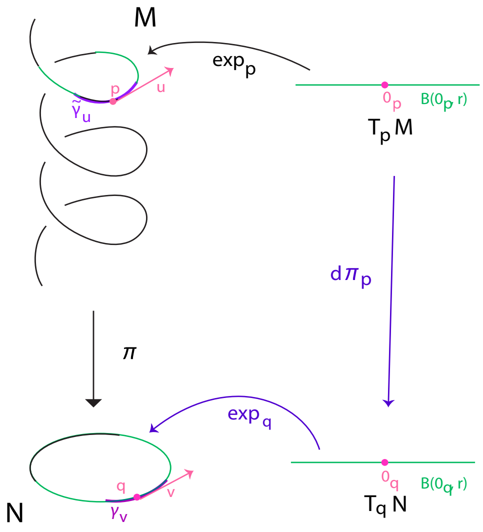


Figure 17.3: An illustration for the completeness of N and that $\pi \restriction B_r(p_i)\colon B_r(p_i) \longrightarrow B_r(q)$ is a diffeomorphism.

and since $\widetilde{\gamma}_u$ is defined on $\mathbb{R}$ because M is complete, so is γ_v. As $\exp_q$ is defined on the whole of T_qN, by Hopf-Rinow, (Theorem 15.17), N is complete. See Figure 17.3.

Next we prove that π is surjective. As N is complete, for any $q_1 \in N$, Theorem 15.16 implies there is a minimal geodesic $\gamma\colon [0, b] \to N$ joining q to q_1 and for the geodesic $\widetilde{\gamma}$ in M emanating from p and with initial direction $d\pi_q^{-1}(\gamma'(0))$, we have $\pi(\widetilde{\gamma}(b)) = \gamma(b) = q_1$, establishing surjectivity.

For any $q \in N$, pick $r > 0$ with $r < i(q)$, where $i(q)$ denotes the injectivity radius of N at q as defined in Definition 15.9, and consider the open metric ball $B_r(q) = \exp_q(B(0_q, r))$ (where $B(0_q, r)$ is the open ball of radius r in T_qN). Let

$$\pi^{-1}(q) = \{p_i\}_{i\in I} \subseteq M.$$

We claim that the following properties hold.

(1) If we write $B_r(p_i) = \exp_{p_i}(B(0_{p_i}, r))$, then each map $\pi \restriction B_r(p_i)\colon B_r(p_i) \longrightarrow B_r(q)$ is a diffeomorphism, in fact an isometry.

(2) $\pi^{-1}(B_r(q)) = \bigcup_{i\in I} B_r(p_i)$.

(3) $B_r(p_i) \cap B_r(p_j) = \emptyset$ whenever $i \neq j$.

It follows from (1), (2) and (3) that $B_r(q)$ is evenly covered by the family of open sets $\{B_r(p_i)\}_{i\in I}$, so π is a covering map.

(1) Since π is a local isometry, Proposition 17.3 (3) and (4) implies π maps geodesics emanating from p_i to geodesics emanating from q, so the following diagram commutes:

$$\begin{tikzcd} B(0_{p_i}, r) \arrow[r, "d\pi_{p_i}"] \arrow[d, "\exp_{p_i}"'] & B(0_q, r) \arrow[d, "\exp_q"] \\ B_r(p_i) \arrow[r, "\pi"'] & B_r(q). \end{tikzcd}$$

See Figure 17.3. Since $\exp_q \circ d\pi_{p_i}$ is a diffeomorphism, $\pi \restriction B_r(p_i)$ must be injective, and since $\exp_{p_i}$ is surjective, so is $\pi \restriction B_r(p_i)$. Then, $\pi \restriction B_r(p_i)$ is a bijection, and as π is a local diffeomorphism, $\pi \restriction B_r(p_i)$ is a diffeomorphism.

(2) Obviously, $\bigcup_{i \in I} B_r(p_i) \subseteq \pi^{-1}(B_r(q))$, by (1). Conversely, pick $p_1 \in \pi^{-1}(B_r(q))$. For $q_1 = \pi(p_1)$, we can write $q_1 = \exp_q v$, for some $v \in B(0_q, r)$, and the map $\gamma(t) = \exp_q(1-t)v$, for $t \in [0, 1]$, is a geodesic in N joining q_1 to q. Then, we have the geodesic $\widetilde{\gamma}$ emanating from p_1 with initial direction $d\pi_{q_1}^{-1}(\gamma'(0))$, and as $\pi \circ \widetilde{\gamma}(1) = \gamma(1) = q$, we have $\widetilde{\gamma}(1) = p_i$ for some α. Since γ has length less than r, we get $p_1 \in B_r(p_i)$.

(3) Suppose $p_1 \in B_r(p_i) \cap B_r(p_j)$. We can pick a minimal geodesic $\widetilde{\gamma}$, in $B_r(p_i)$ (resp $\widetilde{\omega}$ in $B_r(p_j)$) joining p_i to p_1 (resp. joining p_j to p_1). Then the geodesics $\pi \circ \widetilde{\gamma}$ and $\pi \circ \widetilde{\omega}$ are geodesics in $B_r(q)$ from q to $\pi(p_1)$, and their length is less than r. Since $r < i(q)$, these geodesics are minimal so they must coincide. Therefore, $\gamma = \omega$, which implies $i = j$. □

17.3 Riemannian Submersions

Let $\pi \colon M \to B$ be a submersion between two Riemannian manifolds (M, g) and (B, h). For every $b \in B$ in the image of π, the fibre $\pi^{-1}(b)$ is a Riemannian submanifold of M, and for every $p \in \pi^{-1}(b)$, the tangent space $T_p\pi^{-1}(b)$ to $\pi^{-1}(b)$ at p is $\mathrm{Ker}\, d\pi_p$.

Definition 17.2. The tangent space T_pM to M at p splits into the two components

$$T_pM = \mathrm{Ker}\, d\pi_p \oplus (\mathrm{Ker}\, d\pi_p)^{\perp},$$

where $\mathcal{V}_p = \mathrm{Ker}\, d\pi_p$ is the *vertical subspace* of T_pM and $\mathcal{H}_p = (\mathrm{Ker}\, d\pi_p)^{\perp}$ (the orthogonal complement of $\mathcal{V}_p$ with respect to the metric g_p on T_pM) is the *horizontal subspace* of T_pM. Any tangent vector $u \in T_pM$ can be written uniquely as

$$u = u_{\mathcal{H}} + u_{\mathcal{V}},$$

with $u_{\mathcal{H}} \in \mathcal{H}_p$, called the *horizontal component* of u, and $u_{\mathcal{V}} \in \mathcal{V}_p$, called the *vertical component* of u; see Figure 17.4.

A tangent vector $u \in T_pM$ is said to be *horizontal* iff $u \in \mathcal{H}_p$ (equivalently iff $u_{\mathcal{V}} = 0$).

Because π is a submersion, $d\pi_p$ gives a linear isomorphism between $\mathcal{H}_p$ and $T_{\pi(p)}B$. If $d\pi_p$ is an isometry, then most of the differential geometry of B can be studied by "lifting" from B to M.

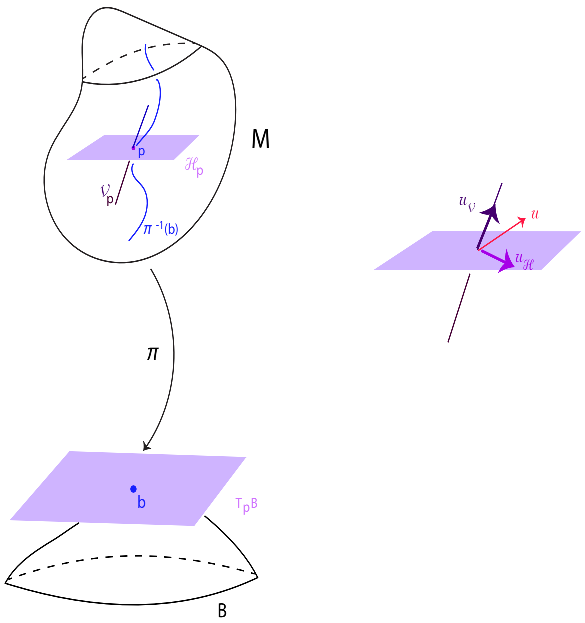


Figure 17.4: An illustration of a Riemannian submersion. Note $\mathcal{H}_p$ is isomorphic to T_bB.

Definition 17.3. A map $\pi \colon M \to B$ between two Riemannian manifolds (M, g) and (B, h) is a *Riemannian submersion* if the following properties hold.

(1) The map π is a smooth submersion.

(2) For every $p \in M$, the map $d\pi_p$ is an isometry between the horizontal subspace $\mathcal{H}_p$ of T_pM and $T_{\pi(p)}B$.

We will see later that Riemannian submersions arise when B is a reductive homogeneous space, or when B is obtained from a free and proper action of a Lie group acting by isometries on B.

Definition 17.4. Let $\pi \colon M \to B$ is a Riemannian submersion which is surjective onto B. Let X be a vector field on B. The unique *horizontal lift* $\overline{X}$ onto M, is defined such that for every $b \in B$ and every $p \in \pi^{-1}(b)$,

$$\overline{X}(p) = (d\pi_p)^{-1}X(b).$$

Since $d\pi_p$ is an isomorphism between $\mathcal{H}_p$ and T_bB, the above condition can be written

$$d\pi \circ \overline{X} = X \circ \pi,$$

which means that $\overline{X}$ and X are π-related (see Definition 9.7).

The following proposition is proved in O'Neill [91] (Chapter 7, Lemma 45) and Gallot, Hulin, Lafontaine [49] (Chapter 2, Proposition 2.109).

Proposition 17.8. *Let $\pi\colon M \to B$ be a Riemannian submersion between two Riemannian manifolds (M, g) and (B, h) equipped with the Levi-Civita connection.*

(1) If γ is a geodesic in M such that $\gamma'(0)$ is a horizontal vector, then γ is horizontal geodesic in M (which means that $\gamma'(t)$ is a horizontal vector for all t), and $c = \pi \circ \gamma$ is a geodesic in B of the same length than γ. See Figure 17.5.

(2) For every $p \in M$, if c is a geodesic in B such that $c(0) = \pi(p)$, then for some ϵ small enough, there is a unique horizontal lift γ of the restriction of c to $[-\epsilon, \epsilon]$, and γ is a geodesic of M.

Furthermore, if $\pi\colon M \to B$ is surjective, then

(3) For any two vector fields $X, Y \in \mathfrak{X}(B)$, we have

(a) $\langle \overline{X}, \overline{Y}\rangle = \langle X, Y\rangle \circ \pi$.

(b) $[\overline{X}, \overline{Y}]_{\mathcal{H}} = \overline{[X, Y]}$.

(c) $(\nabla_{\overline{X}}\overline{Y})_{\mathcal{H}} = \overline{\nabla_X Y}$, where ∇ is the Levi–Civita connection on M.

(4) If M is complete, then B is also complete.

Proof. We prove (1) and (2), following Gallot, Hulin, Lafontaine [49] (Proposition 2.109). We begin with (2). We claim that a Riemannian submersion shortens distance. More precisely, given any two points $p_1, p_2 \in M$,

$$d_B(\pi(p_1), \pi(p_2)) \leq d_M(p_1, p_2),$$

where d_M is the Riemannian distance on M and d_B is the Riemannian distance on B. It suffices to prove that if γ is a curve of M, then $L(\gamma) \geq L(\pi \circ \gamma)$. For any $p \in M$, every tangent vector $u \in T_pM$ can be written uniquely as an orthogonal sum $u = u_{\mathcal{H}} + u_{\mathcal{V}}$, and since $d\pi_p$ is an isometry between $\mathcal{H}_p$ and $T_{\pi(p)}B$, we have

$$\|u\|^2 = \|u_{\mathcal{H}}\|^2 + \|u_{\mathcal{V}}\|^2 \geq \|u_{\mathcal{H}}\|^2 = \|d\pi_p(u_{\mathcal{H}})\|^2 = \|d\pi_p(u)\|^2 .$$

This implies that

$$L(\gamma) = \int_0^1 \|\gamma'(t)\|\, dt \geq \int_0^1 \|(\pi \circ \gamma)'(t)\|\, dt = L(\pi \circ \gamma),$$

as claimed.

For any $p \in M$, let c be a geodesic through $b = \pi(p)$ for $t = 0$. For ϵ small enough, the exponential map $\exp_b$ is a diffeomorphism, so $W = c((-\epsilon, \epsilon))$ is a one-dimensional

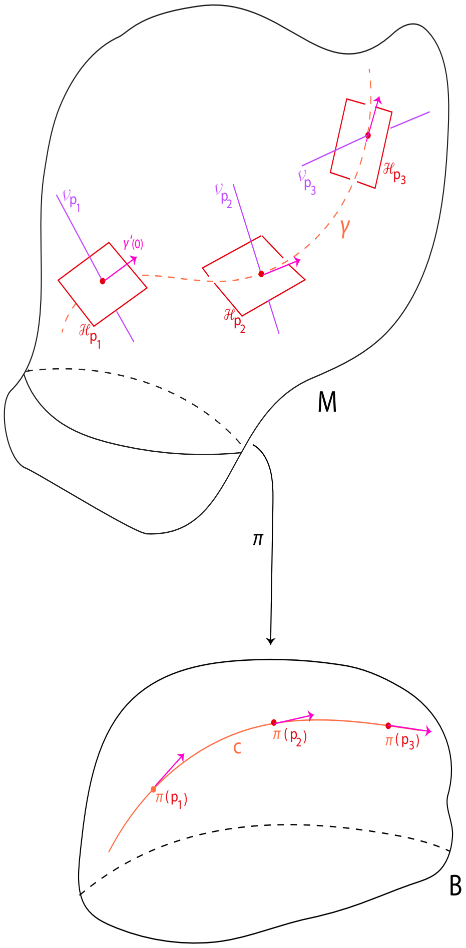


Figure 17.5: An illustration of Part (1), Proposition 17.8. Both γ and c are equal length geodesics in M and B respectively. All the tangent vectors to γ lie in horizontal subspaces.

submanifold of B. Since π is a submersion, $V = \pi^{-1}(W)$ is a submanifold of M. Define a horizontal vector field X on V by

$$X(q) = (d\pi_q)^{-1}(c'(\pi(q))), \quad q \in V,$$

where $d\pi_q$ is the isomorphism between $\mathcal{H}_q$ and $T_{\pi(q)}B$. For any $q \in V$, there is a unique integral curve γ_q through q. In particular, $p \in V$, so the curve γ_p is defined near 0. We claim that it is a geodesic. This is because, first $\left\|\gamma_p'(t)\right\| = \|c'(t)\|$ is a constant, and second, for s small enough, the curve γ_p is locally minimal, that is

$$L(\gamma_p)\mid_{[t,t+s]} = L(c)\mid_{[t,t+s]} = d(c(t), c(t+s)) \leq d(\gamma_p(t), \gamma_p(t+s)).$$

See Figure 17.6.

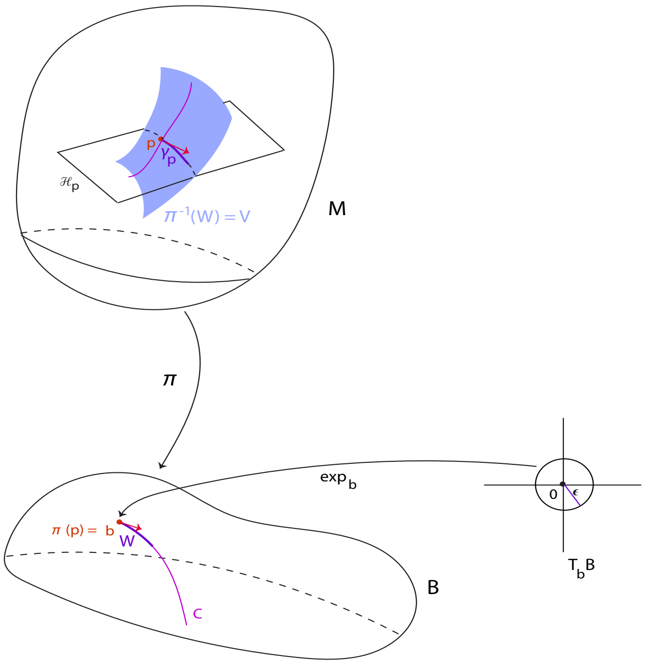


Figure 17.6: A local lift of a geodesic in B to the integral curve γ_p.

We can now prove (1). Let γ be a geodesic through $p = \gamma(0)$ such that $\gamma'(0)$ is a horizontal vector, and write $b = \pi(p)$ and $u = d\pi_p(\gamma'(0))$. Let c be the unique geodesic of B such that $c(0) = b$ and $c'(0) = u$. By (2) we have a horizontal lift $\widetilde{\gamma}$ of c starting at p, and we know it is a geodesic. By construction, $\widetilde{\gamma}'(0) = \gamma'(0)$, so by uniqueness γ and $\widetilde{\gamma}$ coincide on their common domain of definition. It follows that the set of parameters where the geodesic γ is horizontal, and where it is a lift of c is an open subset containing 0. These two conditions being also closed, they must be satisfied on the maximal interval of definition of γ. It is now obvious that $c = \pi \circ \gamma$, a geodesic in B of the same length as γ. □

In (2), we can't expect in general that the whole geodesic c in B can be lifted to M. This is because the manifold (B, h) may be compete but (M, g) may not be. For example, consider the inclusion map $\pi\colon (\mathbb{R}^2 - \{0\}) \to \mathbb{R}^2$, with the canonical Euclidean metrics.

An example of a Riemannian submersion is $\pi\colon S^{2n+1} \to \mathbb{CP}^n$, where S^{2n+1} has the canonical metric and $\mathbb{CP}^n$ has the Fubini–Study metric.

Remark: It shown in Petersen [93] (Chapter 3, Section 5), that the connection $\nabla_{\overline{X}}\overline{Y}$ on M is given by

$$\nabla_{\overline{X}}\overline{Y} = \overline{\nabla_X Y} + \frac{1}{2}[\overline{X}, \overline{Y}]_{\mathcal{V}}.$$

17.4 Isometries and Killing Vector Fields

If X is a vector field on a manifold M, then we saw that we can define the notion of Lie derivative for vector fields ($L_XY = [X, Y]$) and for functions ($L_Xf = X(f)$). It is possible to generalize the notion of Lie derivative to an arbitrary tensor field S; see Gallot, Hullin, Lafontaine [49] (Section 1.F.4). In this section, we only need the following definition.

Definition 17.5. If $S = g$ (the metric tensor), then the *Lie derivative* L_Xg is defined by

$$L_Xg(Y, Z) = X(\langle Y, Z\rangle) - \langle [X, Y], Z\rangle - \langle Y, [X, Z]\rangle,$$

with $X, Y, Z \in \mathfrak{X}(M)$, and where we write $\langle X, Y\rangle$ and $g(X, Y)$ interchangeably.

If Φ_t is an isometry (on its domain), where Φ is the global flow associated with the vector field X, then $\Phi_t^*(g) = g$, and it can be shown that this implies that $L_Xg = 0$. In fact, we have the following result proved in O'Neill [91] (Chapter 9, Proposition 23).

Proposition 17.9. *For any vector field X on a Riemannian manifold (M, g), the diffeomorphisms Φ_t induced by the flow Φ of X are isometries (on their domain) iff $L_Xg = 0$.*

Informally, Proposition 17.9 says that L_Xg measures how much the vector field X changes the metric g.

Definition 17.6. Given a Riemannian manifold (M, g), a vector field X is a *Killing vector field* iff the Lie derivative of the metric vanishes; that is, $L_Xg = 0$.

Killing vector fields play an important role in the study of reductive homogeneous spaces; see Section 22.4. They also interact with the Ricci curvature and play a crucial role in the Bochner technique; see Petersen [93] (Chapter 7).

As the notion of Lie derivative, the notion of covariant derivative ∇_XY of a vector field Y in the direction X can be generalized to tensor fields; see Gallot, Hullin, Lafontaine [49] (Section 2.B.3). In this section, we only need the following definition.

Definition 17.7. The *covariant derivative ∇_Xg of the Riemannian metric g* on a manifold M is given by

$$\nabla_X(g)(Y, Z) = X(\langle Y, Z\rangle) - \langle \nabla_XY, Z\rangle - \langle Y, \nabla_XZ\rangle,$$

for all $X, Y, Z \in \mathfrak{X}(M)$.

Then observe that the connection ∇ on M is compatible with g iff $\nabla_X(g) = 0$ for all X.

Definition 17.8. We define the *covariant derivative ∇X of a vector field X* as the $(1, 1)$-tensor defined so that

$$(\nabla X)(Y) = \nabla_YX$$

for all $X, Y \in \mathfrak{X}(M)$. For every $p \in M$, $(\nabla X)_p$ is defined so that $(\nabla X)_p(u) = \nabla_uX$ for all $u \in T_pM$.

The above facts imply the following proposition.

Proposition 17.10. *Let (M, g) be a Riemannian manifold and let ∇ be the Levi–Civita connection on M induced by g. For every vector field X on M, the following conditions are equivalent.*

(1) X is a Killing vector field; that is, $L_X g = 0$.

(2) $X(\langle Y, Z\rangle) = \langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle$ for all $Y, Z \in \mathfrak{X}(M)$.

(3) $\langle \nabla_Y X, Z\rangle + \langle \nabla_Z X, Y\rangle = 0$ for all $Y, Z \in \mathfrak{X}(M)$; that is, ∇X is skew-adjoint relative to g.

Proof. Since

$$L_X g(Y, Z) = X(\langle Y, Z\rangle) - \langle [X, Y], Z\rangle - \langle Y, [X, Z]\rangle,$$

the equivalence of (1) and (2) is clear.

Since ∇ is the Levi–Civita connection, we have $\nabla_X g = 0$, so

$$X(\langle Y, Z\rangle) - \langle \nabla_X Y, Z\rangle - \langle Y, \nabla_X Z\rangle = 0,$$

which yields

$$\langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle.$$

Since ∇ is also torsion-free we have

$$\begin{aligned} \nabla_X Y - \nabla_Y X &= [X, Y] \\ \nabla_X Z - \nabla_Z X &= [X, Z], \end{aligned}$$

so we get

$$\begin{aligned} \langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle &= \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle \\ &= \langle \nabla_Y X, Z\rangle + \langle Y, \nabla_Z X\rangle + \langle [X, Y], Z\rangle + \langle Y, [X, Z]\rangle, \end{aligned}$$

that is,

$$\langle \nabla_Y X, Z\rangle + \langle \nabla_Z X, Y\rangle = 0.$$

This proves that (2) and (3) are equivalent. □

Condition (3) shows that any parallel vector field is a Killing vector field.

Remark: It can be shown that if γ is any geodesic in M, then the restriction X_γ of any Killing vector field X to γ is a Jacobi field (see Section 16.5), and that $\langle X, \gamma'\rangle$ is constant along γ (see O'Neill [91], Chapter 9, Lemma 26).

17.5 Problems

Problem 17.1. Complete the proof of Proposition 17.4.
Hint. See O'Neill [91] (Chapter 3, Proposition 62).

Problem 17.2. Consider the covering map $p\colon S^n \to \mathbb{RP}^n$ where $\mathbb{RP}^n$ is viewed as the quotient of S^n by the antipodal map. Using Theorem 22.14, it can be shown that there is a Riemannian metric g on $\mathbb{RP}^n$ such that p is a Riemannian submersion (where S^n has the canonical metric induced by $\mathbb{R}^{n+1}$).

(1) Prove that the geodesics of $\mathbb{RP}^n$ are the projections of the geodesics of the sphere S^n. Show that for a geodesic γ on $\mathbb{RP}^n$ we have $\gamma(t+\pi) = \gamma(t)$ for all t.

(2) Prove that $\mathbb{RP}^n$ has constant sectional curvature equal to 1 (use Jacobi fields).

Problem 17.3. Prove Parts (3) and (4) of Proposition 17.8.

Problem 17.4. Let $\pi\colon M \to B$ be a Riemannian submersion between two Riemannian manifolds (M, g) and (B, h) equipped with the Levi-Civita connection. Show that the connection $\nabla_{\overline{X}}\overline{Y}$ on M is given by

$$\nabla_{\overline{X}}\overline{Y} = \overline{\nabla_X Y} + \frac{1}{2}[\overline{X}, \overline{Y}]_{\mathcal{V}}.$$

Hint. See Petersen [93] (Chapter 3, Section 5).

Problem 17.5. Let $p\colon (\widetilde{M}, \widetilde{g}) \to (M, g)$ be a Riemannian submersion. For any orthonormal vector fields X and Y on M with horizontal lifts $\widetilde{X}$ and $\widetilde{Y}$, prove O'Neill's formula:

$$K(X, Y) = K(\widetilde{X}, \widetilde{Y}) + \frac{3}{4}\left\|[\widetilde{X}, \widetilde{Y}]^{\mathcal{V}}\right\|^2.$$

Hint. See Petersen [93], Chapter 3, Section 5.

Problem 17.6. Consider the covering map $p\colon S^{2n+1} \to \mathbb{CP}^n$, viewing $\mathbb{CP}^n$ as the quotient S^{2n+1}/S^1. Using Theorem 22.14, it can be shown that there is a Riemannian metric g on $\mathbb{CP}^n$ such that p is a Riemannian submersion (where S^{2n+1} has the canonical metric induced by $\mathbb{R}^{2n+2}$).

Prove that $\mathbb{CP}^n$ has positive sectional curvature.

Remark: The sectional curvature of $\mathbb{CP}^n$ varies between 1 and 4. See Gallot, Hullin, Lafontaine [49], Chapter III, Section D.

Problem 17.7. Prove Proposition 17.9; that is, for any vector field X on a Riemannian manifold (M, g), the diffeomorphisms Φ_t induced by the flow Φ of X are isometries (on their domain) iff $L_X g = 0$.

Problem 17.8. Prove that if γ is any geodesic in a Riemannian manifold M, then the restriction X_γ of any Killing vector field X to γ is a Jacobi field (see Section 16.5), and $\langle X, \gamma' \rangle$ is constant along γ.

Problem 17.9. Let X be a Killing vector field on a connected Riemannian manifold M. Recall that $(\nabla X)_p$ is defined so that $(\nabla X)_p(u) = \nabla_u X$ for all $u \in T_pM$. Prove that if $X_p = 0$ and $(\nabla X)_p = 0$ for some point $p \in M$, then $X = 0$.

Chapter 18

Lie Groups, Lie Algebras, and the Exponential Map

In Chapter 3 we defined the notion of a Lie group as a certain type of manifold embedded in $\mathbb{R}^N$, for some $N \geq 1$. Now that we have the general concept of a manifold, we can define Lie groups in more generality. If every Lie group was a linear group (a group of matrices), then there would be no need for a more general definition. However, there are Lie groups that are not matrix groups, although it is not a trivial task to exhibit such groups and to prove that they are not matrix groups.

An example of a Lie group which is not a matrix group described in Hall [56] is $G = \mathbb{R} \times \mathbb{R} \times S^1$, with the multiplication given by

$$(x_1, y_1, u_1) \cdot (x_2, y_2, u_2) = (x_1 + x_2, y_1 + y_2, e^{ix_1y_2}u_1u_2).$$

If we define the group H (the Heisenberg group) as the group of 3×3 upper triangular matrices given by

$$H = \left\{ \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} \mid a, b, c \in \mathbb{R} \right\},$$

then it easy to show that the map $\varphi \colon H \to G$ given by

$$\varphi \begin{pmatrix} 1 & a & b \\ 0 & 1 & c \\ 0 & 0 & 1 \end{pmatrix} = (a, c, e^{ib})$$

is a surjective group homomorphism. It is easy to check that the kernel of φ is the discrete group

$$N = \left\{ \begin{pmatrix} 1 & 0 & k2\pi \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mid k \in \mathbb{Z} \right\}.$$

Both groups H and N are matrix groups, yet $G = H/N$ is a Lie group and it can be shown using some representation theory that G is not a matrix group (see Hall [56], Appendix C.3).

Another example of a Lie group that is not a matrix group is obtained by considering the universal cover $\widetilde{\mathbf{SL}(n, \mathbb{R})}$ of $\mathbf{SL}(n, \mathbb{R})$ for $n \geq 2$. The group $\mathbf{SL}(n, \mathbb{R})$ is a matrix group which is not simply connected for $n \geq 2$, and its universal cover $\widetilde{\mathbf{SL}(n, \mathbb{R})}$ is a Lie group which is not a matrix group; see Hall [56] (Appendix C.3) or Ziller [119] (Example 2.22).

Given a Lie group G (not necessarily a matrix group) we begin by defining the Lie bracket on the tangent space $\mathfrak{g} = T_1G$ at the identity in terms of the adjoint representation of G

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

and its derivative at 1, the adjoint representation of $\mathfrak{g}$,

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g});$$

namely, $[u, v] = \mathrm{ad}(u)(v)$.

In Section 18.2, we define left and right invariant vector fields on a Lie group. The map $X \mapsto X(1)$ establishes an isomorphism between the space of left-invariant (resp. right-invariant) vector fields on G and $\mathfrak{g}$. Then by considering integral curves of left-invariant vector fields, we define the generalization of the exponential map $\exp\colon \mathfrak{g} \to G$ to arbitrary Lie groups that are not necessarily matrix groups. We prove some fundamental properties of the exponential map.

In Section 18.3, we revisit homomorphisms of Lie groups and Lie algebras and generalize certain results shown for matrix groups to arbitrary Lie groups. We also define immersed Lie subgroups and (closed) Lie subgroups.

In Section 18.4, we explore the correspondence between Lie groups and Lie algebras and state some of the Lie theorems.

Section 18.5 is devoted to semidirect products of Lie algebras and Lie groups. These are constructions that generalize the notion of direct sum (for Lie algebra) and direct products (for Lie groups). For example, the Lie algebra $\mathfrak{se}(n)$ is the semidirect product of $\mathbb{R}^n$ and $\mathfrak{so}(n)$, and the Lie group $\mathbf{SE}(n)$ is the semidirect product of $\mathbb{R}^n$ and $\mathbf{SO}(n)$.

The notion of universal covering group of a Lie group is described in Section 18.6.

In Section 18.7, we show that the Killing vector fields on a Riemannian manifold M form a Lie algebra. We also describe the relationship between the Lie algebra of complete Killing vector fields and the Lie algebra of the isometry group $\mathrm{Isom}(M)$ of the manifold M.

Besides classic references on Lie groups and Lie algebras, such as Chevalley [31], Knapp [68], Warner [114], Duistermaat and Kolk [43], Bröcker and tom Dieck [24], Sagle and Walde [99], Helgason [58], Serre [106, 105], Kirillov [66], Fulton and Harris [46], and Bourbaki [19], one should be aware of more introductory sources and surveys such as Tapp [111], Kosmann [70], Hall [56], Sattinger and Weaver [102], Carter, Segal and Macdonald [29], Curtis [34], Baker [12], Rossmann [98], Bryant [25], Mneimné and Testard [86] and Arvanitoyeorgos [11].

18.1 Lie Groups and Lie Algebras

We begin our study of Lie groups by generalizing Definition 3.5.

Definition 18.1. A *Lie group* is a nonempty subset G satisfying the following conditions:

(a) G is a group (with identity element denoted e or 1).

(b) G is a smooth manifold.

(c) G is a topological group. In particular, the group operation $\cdot : G \times G \to G$ and the inverse map ${}^{-1}\colon G \to G$ are smooth.

Remark: The smoothness of inversion follows automatically from the smoothness of multiplication. This can be shown by applying the inverse function theorem to the map $(g, h) \mapsto (g, gh)$, from $G \times G$ to $G \times G$.

We have already met a number of Lie groups: $\mathbf{GL}(n, \mathbb{R})$, $\mathbf{GL}(n, \mathbb{C})$, $\mathbf{SL}(n, \mathbb{R})$, $\mathbf{SL}(n, \mathbb{C})$, $\mathbf{O}(n)$, $\mathbf{SO}(n)$, $\mathbf{U}(n)$, $\mathbf{SU}(n)$, $\mathbf{E}(n, \mathbb{R})$, $\mathbf{SO}(n, 1)$. Also, every linear Lie group of $\mathbf{GL}(n, \mathbb{R})$ (see Definition 3.6) is a Lie group.

We saw in the case of linear Lie groups that the tangent space to G at the identity $\mathfrak{g} = T_1 G$ plays a very important role. In particular, this vector space is equipped with a (non-associative) multiplication operation, the Lie bracket, that makes $\mathfrak{g}$ into a Lie algebra. This is again true in this more general setting.

Recall that Lie algebras are defined as follows:

Definition 18.2. A *(real) Lie algebra* $\mathcal{A}$ is a real vector space together with a bilinear map $[\cdot, \cdot]\colon \mathcal{A} \times \mathcal{A} \to \mathcal{A}$, called the *Lie bracket* on $\mathcal{A}$, such that the following two identities hold for all $a, b, c \in \mathcal{A}$:

$$[a,\, a] = 0,$$

and the so-called *Jacobi identity*:

$$[a,\, [b,\, c]] + [c,\, [a,\, b]] + [b,\, [c,\, a]] = 0.$$

It is immediately verified that $[b, a] = -[a, b]$.

For every $a \in \mathcal{A}$, it is customary to define the linear map $\mathrm{ad}(a)\colon \mathcal{A} \to \mathcal{A}$ by

$$\mathrm{ad}(a)(b) = [a, b], \quad b \in \mathcal{A}.$$

The map $\mathrm{ad}(a)$ is also denoted ad_a or $\mathrm{ad}\, a$.

Let us also recall the definition of homomorphisms of Lie groups and Lie algebras.

Definition 18.3. Given two Lie groups G_1 and G_2, a *homomorphism (or map) of Lie groups* is a function $f\colon G_1 \to G_2$ which is a homomorphism of groups, and a smooth map (between the manifolds G_1 and G_2). Given two Lie algebras $\mathcal{A}_1$ and $\mathcal{A}_2$, a *homomorphism (or map) of Lie algebras* is a function $f\colon \mathcal{A}_1 \to \mathcal{A}_2$ which is a linear map between the vector spaces $\mathcal{A}_1$ and $\mathcal{A}_2$, and preserves Lie brackets; that is,

$$f([A,B]) = [f(A), f(B)]$$

for all $A, B \in \mathcal{A}_1$.

An *isomorphism of Lie groups* is a bijective function f such that both f and f^{-1} are maps of Lie groups, and an *isomorphism of Lie algebras* is a bijective function f such that both f and f^{-1} are maps of Lie algebras.

The Lie bracket operation on $\mathfrak{g}$ can be defined in terms of the so-called adjoint representation. Given a Lie group G, for every $a \in G$ we define *left translation* as the map $L_a\colon G \to G$ such that $L_a(b) = ab$ for all $b \in G$, and *right translation* as the map $R_a\colon G \to G$ such that $R_a(b) = ba$ for all $b \in G$. Because multiplication and the inverse maps are smooth, the maps L_a and R_a are diffeomorphisms, and their derivatives play an important role. We also have the inner automorphisms $R_{a^{-1}} \circ L_a = L_a \circ R_{a^{-1}}$, denoted $\mathbf{Ad}_a$. Note that $\mathbf{Ad}_a\colon G \to G$ is defined as

$$\mathbf{Ad}_a(b) = R_{a^{-1}} L_a(b) = aba^{-1}.$$

The derivative

$$d(\mathbf{Ad}_a)_1\colon T_1G \to T_1G$$

of $\mathbf{Ad}_a\colon G \to G$ at 1 is an isomorphism of Lie algebras, and since $T_1G = \mathfrak{g}$, we get a map denoted

$$\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g},$$

where $d(\mathbf{Ad}_a)_1 = \mathrm{Ad}_a$.

Since

$$\mathbf{Ad}_{ab}(c) = abc(ab)^{-1} = abcb^{-1}a^{-1} = \mathbf{Ad}_a(bcb^{-1}) = \mathbf{Ad}_a(\mathbf{Ad}_b(c)),$$

we have $\mathbf{Ad}_{ab} = \mathbf{Ad}_a \circ \mathbf{Ad}_b$, and by taking the derivative at 1, we obtain

$$\mathrm{Ad}_{ab} = d(\mathbf{Ad}_{ab})_1 = d(\mathbf{Ad}_a)_{\mathbf{Ad}_b(1)} \circ d(\mathbf{Ad}_b)_1 = d(\mathbf{Ad}_a)_1 \circ d(\mathbf{Ad}_b)_1 = \mathrm{Ad}_a \circ \mathrm{Ad}_b.$$

It follows that the map $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ given by $a \mapsto \mathrm{Ad}_a$ is a group homomorphism from G to $\mathbf{GL}(\mathfrak{g})$. Furthermore, this map is smooth.

Proposition 18.1. *The map* $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ *is smooth. Thus it is a Lie algebra homomorphism.*

Proof. This fact is shown in Knapp [68] (Chapter 1, Section 10), Warner [114] (Chapter 3, Theorem 3.45), and Lee [76] (Chapter 9, Example 9.3). Knapp's proof use Proposition

18.10(2), which involves the exponential, but fortunately the proof does not depend on the fact that Ad is smooth.

Lee's proof is the most direct. It relies on the fact that the map $C\colon G \times G \to G$ given by $C(a,b) = aba^{-1}$ is smooth, and induces a smooth map $dC\colon T(G \times G) \to T(G)$. For any $a \in G$ and any $u \in \mathfrak{g}$, we can compute $\mathrm{Ad}_a(u)$ using any curve $\gamma\colon (-\epsilon, \epsilon) \to G$ such that $\gamma(0) = 1$ and $\gamma'(0) = u$, as

$$\mathrm{Ad}_a(u) = \frac{d}{dt} C(a, \gamma(t))\bigg|_{t=0} = dC_{(a,1)}(0_a, u).$$

Here we used the isomorphism $T_{(a,1)}(G \times G) \cong T_a G \oplus \mathfrak{g}$. Since dC is smooth, the expression $dC_{(a,1)}(0_a, u)$ is smooth in a. If we pick a basis $(e_1, \ldots, e_n)$ for $\mathfrak{g}$, and if $(e_1^*, \ldots, e_n^*)$ is its canonical dual basis, then setting $u = e_i$, the matrix representing Ad_a has entries $e_j^*(\mathrm{Ad}_a(e_i))$, which are smooth functions. ☐

Definition 18.4. The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

called the *adjoint representation of* G (where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$).

In the case of a Lie linear group, we have verified in Section 3.3 that

$$\mathrm{Ad}(a)(X) = \mathrm{Ad}_a(X) = aXa^{-1}$$

for all $a \in G$ and all $X \in \mathfrak{g}$.

Since $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ is smooth, its derivative $d\mathrm{Ad}_1\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$ exists.

Definition 18.5. The derivative

$$d\mathrm{Ad}_1\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ at 1 is map of Lie algebras, denoted by

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

called the *adjoint representation of* $\mathfrak{g}$.

Recall that Theorem 3.8 immediately implies that the Lie algebra $\mathfrak{gl}(\mathfrak{g})$ of $\mathbf{GL}(\mathfrak{g})$ is the vector space $\mathrm{End}(\mathfrak{g}, \mathfrak{g})$ of all endomorphisms of $\mathfrak{g}$; that is, the vector space of all linear maps on $\mathfrak{g}$.

In the case of a linear Lie group, we verified in Section 3.3 that

$$\mathrm{ad}(A)(B) = [A,\, B] = AB - BA,$$

for all $A, B \in \mathfrak{g}$.

In the case of an abstract Lie group G, since ad is defined, we would like to define the Lie bracket of $\mathfrak{g}$ in terms of ad. This is the key to the definition of the Lie bracket in the case of a general Lie group (not just a linear Lie group).

Definition 18.6. Given a Lie group G, the tangent space $\mathfrak{g} = T_1G$ at the identity with the Lie bracket defined by

$$[u,\, v] = \mathrm{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g}$$

is the *Lie algebra of the Lie group G*. The Lie algebra $\mathfrak{g}$ of a Lie group G is also denoted by $\mathrm{L}(G)$ (for instance, when the notation $\mathfrak{g}$ is already used for something else).

Actually, we have to justify why $\mathfrak{g}$ really is a Lie algebra. For this we have

Proposition 18.2. *Given a Lie group G, the Lie bracket $[u,\, v] = \mathrm{ad}(u)(v)$ of Definition 18.6 satisfies the axioms of a Lie algebra (given in Definition 18.2). Therefore, $\mathfrak{g}$ with this bracket is a Lie algebra.*

Proof. The proof requires Proposition 18.14, but we prefer to defer the proof of this proposition until Section 18.3. Since

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$$

is a Lie group homomorphism, by Proposition 18.14, the map $\mathrm{ad} = d\mathrm{Ad}_1$ is a homomorphism of Lie algebras, $\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$, which means that

$$\mathrm{ad}([u, v]) = [\mathrm{ad}(u), \mathrm{ad}(v)] = \mathrm{ad}(u) \circ \mathrm{ad}(v) - \mathrm{ad}(v) \circ \mathrm{ad}(u), \quad \text{for all } u, v \in \mathfrak{g},$$

since the bracket in $\mathfrak{gl}(\mathfrak{g}) = \mathrm{End}(\mathfrak{g}, \mathfrak{g})$ is just the commutator. Applying the above to $z \in \mathfrak{g}$ gives

$$\begin{aligned}\mathrm{ad}([u, v])(z) &= [[u, v], z]\\ &= \mathrm{ad}(u) \circ \mathrm{ad}(v)(z) - \mathrm{ad}(v) \circ \mathrm{ad}(u)(z)\\ &= \mathrm{ad}(u)[v, z] - \mathrm{ad}(v)[u, z] = [u, [v, z]] - [v, [u, z]],\end{aligned}$$

which is equivalent to the Jacobi identity. We still have to prove that $[u, u] = 0$, or equivalently, that $[v, u] = -[u, v]$. For this, following Duistermaat and Kolk [43] (Chapter 1, Section 1), consider the map

$$F\colon G \times G \longrightarrow G\colon (a, b) \mapsto aba^{-1}b^{-1}.$$

We claim that the derivative of F at $(1, 1)$ is the zero map. This follows using the product rule and chain rule from two facts.

1. The derivative of multiplication in a Lie group $\mu\colon G \times G \to G$ is given by

$$d\mu_{a,b}(u, v) = (dR_b)_a(u) + (dL_a)_b(v),$$

for all $u \in T_aG$ and all $v \in T_bG$. At $(1, 1)$, the above yields

$$d\mu_{1,1}(u, v) = u + v.$$

2. The derivative of the inverse map $\iota\colon G \to G$ is given by

$$d\iota_a(u) = -(dR_{a^{-1}})_1 \circ (dL_{a^{-1}})_a(u) = -(dL_{a^{-1}})_1 \circ (dR_{a^{-1}})_a(u)$$

   for all $u \in T_aG$. At 1, we get

$$d\iota_1(u) = -u.$$

In particular write $F = F_1F_2$, where $F_1\colon G \times G \to G$ is $F_1(a,b) = ab$ and $F_2\colon G \times G \to G$ is $F_2(a,b) = a^{-1}b^{-1} = (ba)^{-1}$. If we let $H\colon G \times G \to G \times G$ be the map given by $H(a,b) = (F_1(a,b), F_2(a,b))$, then $F = \mu \circ H$. The chain rule implies that for all $u, v \in \mathfrak{g} \times \mathfrak{g}$, since $H(1,1) = (F_1(1,1), F_2(1,1)) = (1,1)$,

$$\begin{aligned} dF_{1,1}(u,v) &= (d\mu_{H(1,1)} \circ dH_{(1,1)})(u,v) \\ &= d\mu_{(1,1)}(d(F_1)_{(1,1)}(u,v), d(F_2)_{(1,1)}(u,v)) \\ &= d\mu_{(1,1)}(u+v, -(u+v)) \\ &= u + v + (-u - v) = 0. \end{aligned}$$

Since $dF_{1,1} = 0$, then $(1,1)$ is a critical point of F, and we can adapt the standard reasoning provided at the beginning of Section 16.4 (also see Milnor [81], pages 4-5), to prove that the Hessian $\mathrm{Hess}(F)$ of F is well-defined at $(1,1)$, and is a symmetric bilinear map

$$\mathrm{Hess}(F)_{(1,1)}\colon (\mathfrak{g} \times \mathfrak{g}) \times (\mathfrak{g} \times \mathfrak{g}) \longrightarrow \mathfrak{g}.$$

Furthermore, for any (X_1, Y_1) and $(X_2, Y_2) \in \mathfrak{g} \times \mathfrak{g}$, the value $\mathrm{Hess}(F)_{(1,1)}((X_1, Y_1), (X_2, Y_2))$ of the Hessian can be computed by two successive derivatives, either as

$$(\widetilde{X_1}, \widetilde{Y_1})((\widetilde{X_2}, \widetilde{Y_2})F)_{(1,1)},$$

or as

$$(\widetilde{X_2}, \widetilde{Y_2})((\widetilde{X_1}, \widetilde{Y_1})F)_{(1,1)},$$

where $\widetilde{X_i}$ and $\widetilde{Y_i}$ are smooth vector fields with value X_i and Y_j at 1, which exist by Proposition 10.2. Because of the symmetry property, the above derivatives are independent of the extensions $\widetilde{X_i}$ and $\widetilde{Y_i}$.

The value of the Hessian can also be computed using parametrized surfaces. Indeed, for any smooth surface $(\alpha, \beta)\colon (-\epsilon, \epsilon) \times (-\epsilon, \epsilon) \to G \times G$, such that $\alpha(0,0) = \beta(0,0) = 1$, and

$$\left(\frac{\partial \alpha}{\partial x}(0,0), \frac{\partial \alpha}{\partial y}(0,0)\right) = (X_1, Y_1), \quad \left(\frac{\partial \beta}{\partial x}(0,0), \frac{\partial \beta}{\partial y}(0,0)\right) = (X_2, Y_2),$$

we have

$$\mathrm{Hess}(F)_{(1,1)}((X_1, Y_1), (X_2, Y_2)) = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}(F \circ (\alpha, \beta))\right)_{(0,0)} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}(F \circ (\alpha, \beta))\right)_{(0,0)}.$$

We apply the above to the function F given by $F(a,b) = aba^{-1}b^{-1}$, and to the tangent vectors $(u,0)$ and $(0,v)$. Consider a parametrized surface (α,β) where $\alpha(x,y)$ is independent of y, $\beta(x,y)$ is independent of x, $\alpha(0,0)=\beta(0,0)=1$,

$$\left(\frac{\partial\alpha}{\partial x}(0,0), \frac{\partial\alpha}{\partial y}(0,0)\right) = (u,0), \quad \left(\frac{\partial\beta}{\partial x}(0,0), \frac{\partial\beta}{\partial y}(0,0)\right) = (0,v).$$

First we compute

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}(F\circ(\alpha,\beta))\right)_{(0,0)} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\alpha\beta\alpha^{-1}\beta^{-1}\right)_{(0,0)} = \frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\mu(\mathbf{Ad}_{\alpha}(\beta),\iota(\beta))\right)_{(0,0)}.$$

Using the chain rule, we get

$$\begin{aligned}\left(\frac{\partial}{\partial y}\mu(\mathbf{Ad}_{\alpha(x)}(\beta(y)),\iota(\beta(y)))\right)_{(0,0)} &= d\mu_{(1,1)}(d(\mathbf{Ad}_{\alpha(x)})_1(v), d\iota_1(v))\\ &= \mathrm{Ad}_{\alpha(x)}(v) - v.\end{aligned}$$

Then we obtain

$$\begin{aligned}\frac{\partial}{\partial x}\left(\frac{\partial}{\partial y}\mu(\mathbf{Ad}_{\alpha}(\beta),\iota(\beta))\right)_{(0,0)} &= \frac{\partial}{\partial x}\left(\mathrm{Ad}_{\alpha(x)}(v) - v\right)_{(0,0)}\\ &= \mathrm{ad}_u(v)\\ &= [u,v].\end{aligned}$$

Next we compute

$$\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}(F\circ(\alpha,\beta))\right)_{(0,0)} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}\alpha\beta\alpha^{-1}\beta^{-1}\right)_{(0,0)} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}\mu(\alpha,\mathbf{Ad}_{\beta}(\iota(\alpha)))\right)_{(0,0)}.$$

Using the chain rule, we get

$$\begin{aligned}\left(\frac{\partial}{\partial x}\mu(\alpha(x),\mathbf{Ad}_{\beta(y)}(\iota(\alpha(x))))\right)_{(0,0)} &= d\mu_{(1,1)}(u, d(\mathbf{Ad}_{\beta(y)})_1(d\iota_1(u)))\\ &= u + \mathrm{Ad}_{\beta(y)}(-u)\\ &= u - \mathrm{Ad}_{\beta(y)}(u).\end{aligned}$$

Finally we compute

$$\begin{aligned}\frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}\mu(\alpha,\mathbf{Ad}_{\beta}(\iota(\alpha)))\right)_{(0,0)} &= \frac{\partial}{\partial y}\left(u - \mathrm{Ad}_{\beta(y)}(u)\right)_{(0,0)}\\ &= -\mathrm{ad}_v(u)\\ &= -[u,v].\end{aligned}$$

Since the Hessian is bilinear symmetric, we get $[u,v] = -[v,u]$, as claimed. □

Remark: After proving that $\mathfrak{g}$ is isomorphic to the vector space of left-invariant vector fields on G, we get another proof of Proposition 18.2.

18.2 Left and Right Invariant Vector Fields, the Exponential Map

The purpose of this section is to define the exponential map for an arbitrary Lie group in a way that is consistent with our previous definition of the exponential defined for a linear Lie group, namely

$$e^X = I_n + \sum_{p\geq 1} \frac{X^p}{p!} = \sum_{p\geq 0} \frac{X^p}{p!},$$

where $X \in \mathrm{M}_n(\mathbb{R})$ or $X \in \mathrm{M}_n(\mathbb{C})$. We obtain the desired generalization by recalling Proposition 11.25 which states that for a linear Lie group, the maximal integral curve through initial point $p \in G$ with initial velocity X is given by $\gamma_p(t) = e^{tX}p$; see Sections 11.3 and 9.3. Thus the exponential may be defined in terms of maximal integral curves. Since the notion of maximal integral curve relies on vector fields, we begin our construction of the exponential map for an abstract Lie group G by defining left and right invariant vector fields.

Definition 18.7. If G is a Lie group, a vector field X on G is *left-invariant* (resp. *right-invariant*) iff

$$d(L_a)_b(X(b)) = X(L_a(b)) = X(ab), \quad \text{for all } a, b \in G.$$

(resp.

$$d(R_a)_b(X(b)) = X(R_a(b)) = X(ba), \quad \text{for all } a, b \in G.)$$

Equivalently, a vector field X is left-invariant iff the following diagram commutes (and similarly for a right-invariant vector field):

$$\begin{tikzcd} TG \arrow[r, "d(L_a)"] & TG \\ G \arrow[u, "X"] \arrow[r, "L_a"'] & G \arrow[u, "X"'] \end{tikzcd}$$

If X is a left-invariant vector field, setting $b = 1$, we see that

$$d(L_a)_1(X(1)) = X(L_a(1)) = X(a),$$

which shows that X is determined by its value $X(1) \in \mathfrak{g}$ at the identity (and similarly for right-invariant vector fields).

Conversely, given any $v \in \mathfrak{g}$, since $d(L_a)_1 \colon \mathfrak{g} \to T_aG$ is a linear isomorphism between $\mathfrak{g}$ and T_aG for every $a \in G$, we can define the vector field v^L by

$$v^L(a) = d(L_a)_1(v), \quad \text{for all } a \in G.$$

We claim that v^L is left-invariant. This follows by an easy application of the chain rule:

$$\begin{aligned} v^L(ab) &= d(L_{ab})_1(v) \\ &= d(L_a \circ L_b)_1(v) \\ &= d(L_a)_b(d(L_b)_1(v)) \\ &= d(L_a)_b(v^L(b)). \end{aligned}$$

Furthermore, $v^L(1) = v$.

In summary, we proved the following result.

Proposition 18.3. *Given a Lie group G, the map $X \mapsto X(1)$ establishes an isomorphism between the space of left-invariant vector fields on G and $\mathfrak{g}$. In fact, the map $G \times \mathfrak{g} \longrightarrow TG$ given by $(a, v) \mapsto v^L(a)$ is an isomorphism between $G \times \mathfrak{g}$ and the tangent bundle TG.*

Definition 18.8. The vector space of left-invariant vector fields on a Lie group G is denoted by $\mathfrak{g}^L$.

Because Proposition 18.14 implies that the derivative of any Lie group homomorphism is a Lie algebra homomorphism, $(dL_a)_b$ is a Lie algebra homomorphism, so if X and Y are left-invariant vector fields, then the vector field $[X, Y]$ is also left-invariant. In particular

$$\begin{aligned} (dL_a)_b[X(b), Y(b)] &= [(dL_a)_b X(b), (dL_a)_b Y(b)], && (dL_a)_b \text{ is a Lie algebra homomorphism} \\ &= [X(L_a(b)), Y(L_a(b))], && X \text{ and } Y \text{ are left-invariant vector fields} \\ &= [X(ab), Y(ab)]. \end{aligned}$$

It follows that $\mathfrak{g}^L$ is a Lie algebra.

Given any $v \in \mathfrak{g}$, since $(dR_a)_1 \colon \mathfrak{g} \to T_aG$ is a linear isomorphism between $\mathfrak{g}$ and T_aG for every $a \in G$, we can also define the vector field v^R by

$$v^R(a) = d(R_a)_1(v), \quad \text{for all } a \in G.$$

It is easily shown that v^R is right-invariant and we also have an isomorphism $G \times \mathfrak{g} \longrightarrow TG$ given by $(a, v) \mapsto v^R(a)$.

Definition 18.9. The vector space of right-invariant vector fields on a Lie group G is denoted by $\mathfrak{g}^R$.

Since $(dR_a)_b$ is a Lie algebra homomorphism, if X and Y are right-invariant vector fields, then the vector field $[X, Y]$ is also right-invariant. It follows that $\mathfrak{g}^R$ is a Lie algebra.

We will see later in this section that the Lie algebras $\mathfrak{g}$ and $\mathfrak{g}^L$ are isomorphic, and the Lie algebras $\mathfrak{g}$ and $\mathfrak{g}^R$ are anti-isomorphic.

Another reason why left-invariant (resp. right-invariant) vector fields on a Lie group are important is that they are complete; that is, they define a flow whose domain is $\mathbb{R} \times G$. To prove this we begin with the following easy proposition.

Proposition 18.4. *Given a Lie group G, if X is a left-invariant (resp. right-invariant) vector field and Φ is its flow and γ_g is the associated maximal integral curve with initial condition $g \in G$, then*

$$\gamma_g(t) = \Phi_t^X(g) = \Phi(t,g) = g\Phi(t,1) = g\Phi_t^X(1) = g\gamma_1(t)$$
$$(\textit{resp.}\quad \Phi(t,g) = \Phi(t,1)g), \quad \textit{for all } (t,g) \in \mathcal{D}(X).$$

Proof. Write

$$\gamma(t) = g\gamma_1(t) = g\Phi(t,1) = L_g(\Phi(t,1)).$$

Then $\gamma(0) = g$, and by the chain rule,

$$\dot{\gamma}(t) = d(L_g)_{\Phi(t,1)}(\dot{\Phi}(t,1)) = d(L_g)_{\Phi(t,1)}(X(\Phi(t,1))) = X(L_g(\Phi(t,1))) = X(\gamma(t)),$$

where the third equality made use of the fact that X is a left-invariant vector field. By the uniqueness of maximal integral curves, $\gamma(t) = \Phi(t,g)$ for all t, and so

$$\Phi(t,g) = g\Phi(t,1).$$

A similar argument applies to right-invariant vector fields. □

Proposition 18.5. *Given a Lie group G, for every $v \in \mathfrak{g}$, there is a unique smooth homomorphism $h_v \colon (\mathbb{R},+) \to G$ such that $\dot{h}_v(0) = (dh_v/ds)(0) = v$. Furthermore, $h_v(t) = \gamma_1(t)$ is the maximal integral curve of both v^L and v^R with initial condition 1, and the flows of v^L and v^R are defined for all $t \in \mathbb{R}$.*

Proof. Let $\Phi_t^v(g) = \gamma_g(t)$ denote the flow of v^L. As far as defined, we know that

$$\begin{aligned}\Phi_{s+t}^v(1) &= \Phi^v(s+t,1) = \Phi_s^v(\Phi_t^v(1)), && \text{by Proposition 9.10}\\ &= \Phi^v(s,\Phi_t^v(1)) = \Phi^v(s,\Phi^v(t,1)) &&\\ &= \Phi_t^v(1)\Phi_s^v(1). && \text{by Proposition 18.4}\end{aligned}$$

Now, if $\Phi_t^v(1) = \gamma_1(t)$ is defined on $(-\epsilon,\epsilon)$, setting $s = t$, we see that $\Phi_t^v(1)$ is actually defined on $(-2\epsilon, 2\epsilon)$. By induction we see that $\Phi_t^v(1)$ is defined on $(-2^n\epsilon, 2^n\epsilon)$, for all $n \geq 0$, and so $\Phi_t^v(1)$ is defined on $\mathbb{R}$, and the map $t \mapsto \Phi_t^v(1)$ is a homomorphism $h_v \colon (\mathbb{R},+) \to G$, with $\dot{h}_v(0) = v$. Since $\Phi_t^v(g) = g\Phi_t^v(1)$, the flow $\Phi_t^v(g)$ is defined for all $(t,g) \in \mathbb{R} \times G$. A similar proof applies to v^R. To show that h_v is smooth, consider the map

$$\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad \text{where} \quad (t,g,v) \mapsto (g\Phi_t^v(1), v).$$

It can be shown that the above is the flow of the vector field

$$(g,v) \mapsto (v^L(g), 0),$$

and thus it is smooth. Consequently, the restriction of this smooth map to $\mathbb{R} \times \{1\} \times \{v\}$, which is just $t \mapsto \Phi_t^v(1) = h_v(t)$, is also smooth.

Assume $h\colon (\mathbb{R}, +) \to G$ is a smooth homomorphism with $\dot{h}(0) = v$. From

$$h(t+s) = h(t)h(s) = h(s)h(t) = h(s+t),$$

we have

$$h(t+s) = L_{h(t)}h(s), \qquad h(t+s) = R_{h(t)}h(s).$$

If we differentiate these equations with respect to s at $s = 0$, we get via the chain rule

$$\frac{dh}{ds}(t) = d(L_{h(t)})_1(v) = v^L(h(t))$$

and

$$\frac{dh}{ds}(t) = d(R_{h(t)})_1(v) = v^R(h(t)).$$

Therefore, $h(t)$ is an integral curve for v^L and v^R with initial condition $h(0) = 1$, and $h(t) = \Phi^v_t(1) = \gamma_1(t)$. □

Since $h_v\colon (\mathbb{R}, +) \to G$ is a homomorphism, the following terminology is often used.

Definition 18.10. The integral curve $h_v\colon (\mathbb{R}, +) \to G$ of Proposition 18.5 is often referred to as a *one-parameter group*.

Proposition 18.5 yields the definition of the exponential map in terms of maximal integral curves.

Definition 18.11. Given a Lie group G, the *exponential map* $\exp\colon \mathfrak{g} \to G$ is given by

$$\exp(v) = h_v(1) = \Phi^v_1(1) = \gamma_1(1), \quad \text{for all } v \in \mathfrak{g}.$$

We can see that exp is smooth as follows. As in the proof of Proposition 18.5, we have the smooth map

$$\mathbb{R} \times G \times \mathfrak{g} \longrightarrow G \times \mathfrak{g}, \quad \text{where} \quad (t, g, v) \mapsto (g\Phi^v_t(1), v),$$

which is the flow of the vector field

$$(g, v) \mapsto (v^L(g), 0).$$

Consequently, the restriction of this smooth map to $\{1\} \times \{1\} \times \mathfrak{g}$, which is just
$v \mapsto \Phi^v_1(1) = \exp(v)$, is also smooth.

Observe that for any fixed $t \in \mathbb{R}$, the map

$$s \mapsto h_v(st) = \gamma_1(st)$$

is a smooth homomorphism h such that $\dot{h}(0) = tv$. By uniqueness of the maximal integral curves, we have

$$\Phi^v_{st}(1) = h_v(st) = h_{tv}(s) = \Phi^{tv}_s(1).$$

Setting $s = 1$, we find that

$$\gamma_1(t) = h_v(t) = \exp(tv), \quad \text{for all } v \in \mathfrak{g} \text{ and all } t \in \mathbb{R}.$$

If G is a linear Lie group, the preceding equation is equivalent to Proposition 11.25.

Differentiating this equation with respect to t at $t = 0$, we get

$$v = d\exp_0(v),$$

i.e., $d\exp_0 = \mathrm{id}_{\mathfrak{g}}$. By the inverse function theorem, exp is a local diffeomorphism at 0. This means that there is some open subset $U \subseteq \mathfrak{g}$ containing 0, such that the restriction of exp to U is a diffeomorphism onto $\exp(U) \subseteq G$, with $1 \in \exp(U)$. This argument is very similar to the argument used in proving Proposition 15.4.

In fact, by left-translation, the map $v \mapsto g\exp(v)$ is a local diffeomorphism between some open subset $U \subseteq \mathfrak{g}$ containing 0 and the open subset $\exp(U)$ containing g. The above facts are recorded in the following proposition.

Proposition 18.6. *Given a Lie group G, the exponential map* $\exp\colon \mathfrak{g} \to G$ *is smooth and is a local diffeomorphism at* 0*.*

Remark: Given any Lie group G, we have a notion of exponential map $\exp\colon \mathfrak{g} \to G$ given by the maximal integral curves of left-invariant vector fields on G (see Proposition 18.5 and Definition 18.11). This exponential does not require any connection or any metric in order to be defined; let us call it the *group exponential*. If G is endowed with a connection or a Riemannian metric (the Levi-Civita connection if G has a Riemannnian metric), then we also have the notion of exponential induced by geodesics (see Definition 15.6); let us call this exponential the *geodesic exponential*. To avoid ambiguities when both kinds of exponentials arise, we propose to denote the group exponential by $\exp_{\mathrm{gr}}$ and the geodesic exponential by exp, as before. Even if the geodesic exponential is defined on the whole of $\mathfrak{g}$ (which may not be the case), these two notions of exponential differ in general.

The group exponential map is natural in the following sense.

Proposition 18.7. *Given any two Lie groups G and H, for every Lie group homomorphism $f\colon G \to H$, the following diagram commutes.*

$$\begin{tikzcd} G \arrow[r, "f"] & H \\ \mathfrak{g} \arrow[u, "\exp"] \arrow[r, "df_1"'] & \mathfrak{h}. \arrow[u, "\exp"'] \end{tikzcd}$$

Proof. Observe that for every $v \in \mathfrak{g}$, the map $h\colon t \mapsto f(\exp(tv))$ is a homomorphism from $(\mathbb{R}, +)$ to G such that $\dot{h}(0) = df_1(v)$. On the other hand, Proposition 18.5 shows that the map $t \mapsto \exp(tdf_1(v))$ is the unique maximal integral curve whose tangent at 0 is $df_1(v)$, so $f(\exp(v)) = \exp(df_1(v))$. $\square$

Proposition 18.7 is the generalization of Proposition 3.13.

A useful corollary of Proposition 18.7 is

Proposition 18.8. *Let G be a connected Lie group and H be any Lie group. For any two homomorphisms $\phi_1\colon G \to H$ and $\phi_2\colon G \to H$, if $d(\phi_1)_1 = d(\phi_2)_1$, then $\phi_1 = \phi_2$.*

Proof. We know that the exponential map is a diffeomorphism on some small open subset U containing 0. By Proposition 18.7, for all $a \in \exp_G(U)$, we have

$$\phi_i(a) = \exp_H(d(\phi_i)_1(\exp_G^{-1}(a))), \quad i = 1, 2,$$

as illustrated in the following diagram:

$$\begin{tikzcd}
G \arrow[r, "\phi_i"] \arrow[d, "\exp_G^{-1}"'] & H \\
U \subseteq \mathfrak{g} \arrow[r, "d(\phi_i)_1"'] & \mathfrak{h}. \arrow[u, "\exp_H"']
\end{tikzcd}$$

Since $d(\phi_1)_1 = d(\phi_2)_1$, we conclude that $\phi_1 = \phi_2$ on $\exp_G(U)$. However, as G is connected, Proposition 4.9 implies that G is generated by $\exp_G(U)$ (we can easily find a symmetric neighborhood of 1 in $\exp_G(U)$). Therefore, $\phi_1 = \phi_2$ on G. □

Corollary 18.9. *If G is a connected Lie group, then a Lie group homomorphism $\phi\colon G \to H$ is uniquely determined by the Lie algebra homomorphism $d\phi_1\colon \mathfrak{g} \to \mathfrak{h}$.*

We obtain another useful corollary of Proposition 18.7 when we apply it to the adjoint representation of G

$$\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g}),$$

and to the conjugation map

$$\mathbf{Ad}_a\colon G \to G,$$

where $\mathbf{Ad}_a(b) = aba^{-1}$. In the first case, $d\mathrm{Ad}_1 = \mathrm{ad}$, with $\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$, and in the second case, $d(\mathbf{Ad}_a)_1 = \mathrm{Ad}_a$.

Proposition 18.10. *Given any Lie group G, the following properties hold.*

(1)

$$\mathrm{Ad}(\exp(u)) = e^{\mathrm{ad}(u)}, \qquad \textit{for all } u \in \mathfrak{g},$$

where $\exp\colon \mathfrak{g} \to G$ *is the exponential of the Lie group G, and $f \mapsto e^f$ is the exponential map given by*

$$e^f = \sum_{k=0}^{\infty} \frac{f^k}{k!},$$

for any linear map (matrix) $f \in \mathfrak{gl}(\mathfrak{g})$. Equivalently, the following diagram commutes.

$$\begin{tikzcd}
G \arrow[r, "\mathrm{Ad}"] & \mathbf{GL}(\mathfrak{g}) \\
\mathfrak{g} \arrow[u, "\exp"] \arrow[r, "\mathrm{ad}"'] & \mathfrak{gl}(\mathfrak{g}). \arrow[u, "f \mapsto e^f"']
\end{tikzcd}$$

(2)

$$\exp(t\mathrm{Ad}_g(u)) = g\exp(tu)g^{-1},$$

for all $u \in \mathfrak{g}$, all $g \in G$ and all $t \in \mathbb{R}$. Equivalently, the following diagram commutes.

$$\begin{tikzcd}
G \arrow[r, "\mathbf{Ad}_g"] & G \\
\mathfrak{g} \arrow[u, "\exp"] \arrow[r, "\mathrm{Ad}_g"'] & \mathfrak{g}. \arrow[u, "\exp"']
\end{tikzcd}$$

Since the Lie algebra $\mathfrak{g} = T_1G$, as a vector space, is isomorphic to the vector space of left-invariant vector fields on G and since the Lie bracket of vector fields makes sense (see Definition 9.5), it is natural to ask if there is any relationship between $[u, v]$, where $[u, v] = \mathrm{ad}(u)(v)$, and the Lie bracket $[u^L, v^L]$ of the left-invariant vector fields associated with $u, v \in \mathfrak{g}$. The answer is: Yes, they coincide (*via* the correspondence $u \mapsto u^L$). This fact is recorded in the proposition below whose proof involves some rather acrobatic uses of the chain rule found in Warner [114] (Chapter 3, Proposition 3.47), Bröcker and tom Dieck [24] (Chapter 1, Section 2, formula 2.11), or Marsden and Ratiu [77] (Chapter 9, Proposition 9.1.5).

Proposition 18.11. *Given a Lie group G, we have*

$$[u^L, v^L](1) = \mathrm{ad}(u)(v), \quad \textit{for all } u, v \in \mathfrak{g},$$

where $[u^L, v^L](1)$ is the element of the vector field $[u^L, v^L]$ at the identity.

Proposition 18.11 shows that the Lie algebras $\mathfrak{g}$ and $\mathfrak{g}^L$ are isomorphic (where $\mathfrak{g}^L$ is the Lie algebra of left-invariant vector fields on G). In view of this isomorphism, we make the following definition.

Definition 18.12. Let X and Y be any two left-invariant vector fields on G. We define $\mathrm{ad}(X)(Y)$ by

$$\mathrm{ad}(X)(Y) = [X, Y],$$

where the Lie bracket on the right-hand side is the Lie bracket on vector fields.

It is shown in Marsden and Ratiu [77] (Chapter 9, just after Definition 9.1.2) that if $\iota\colon G \to G$ is the inversion map $\iota(g) = g^{-1}$, then for any $u \in \mathfrak{g}$, the vector fields u^L and u^R are related by the equation

$$\iota_*(u^L) = -u^R,$$

where $\iota_*(u^L)$ is the push-forward of u^L (that is,

$$\iota_*(u^L) = d\iota_{g^{-1}}(u^L(g^{-1}))$$

for all $g \in G$.) This implies that

$$[u^L, v^L] = -[u^R, v^R],$$

and so

$$[u^R, v^R](1) = -\mathrm{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g}.$$

It follows that the Lie algebras $\mathfrak{g}$ and $\mathfrak{g}^R$ are anti-isomorphic (where $\mathfrak{g}^R$ is the Lie algebra of right-invariant vector fields on G). In summary we have the following result.

Proposition 18.12. *Given a Lie group G, the Lie algebra $\mathfrak{g}$ and $\mathfrak{g}^L$ are isomorphic, and the Lie algebra $\mathfrak{g}$ and $\mathfrak{g}^R$ are anti-isomorphic.*

We can apply Proposition 4.10 and use the exponential map to prove a useful result about Lie groups. If G is a Lie group, let G_0 be the connected component of the identity. We know G_0 is a topological normal subgroup of G and it is a submanifold in an obvious way, so it is a Lie group.

Proposition 18.13. *If G is a Lie group and G_0 is the connected component of 1, then G_0 is generated by $\exp(\mathfrak{g})$. Moreover, G_0 is countable at infinity.*

Proof. We can find a symmetric open U in $\mathfrak{g}$ in containing 0, on which exp is a diffeomorphism. Then apply Proposition 4.10 to $V = \exp(U)$. That G_0 is countable at infinity follows from Proposition 4.11. □

18.3 Homomorphisms of Lie Groups and Lie Algebras, Lie Subgroups

If G and H are two Lie groups and $\phi\colon G \to H$ is a homomorphism of Lie groups, then $d\phi_1\colon \mathfrak{g} \to \mathfrak{h}$ is a linear map between the Lie algebras $\mathfrak{g}$ and $\mathfrak{h}$ of G and H. In fact, it is a Lie algebra homomorphism, as shown below. This proposition is the generalization of Proposition 3.14.

Proposition 18.14. *If G and H are two Lie groups and $\phi\colon G \to H$ is a homomorphism of Lie groups, then*

$$d\phi_1 \circ \mathrm{Ad}_g = \mathrm{Ad}_{\phi(g)} \circ d\phi_1, \quad \textit{for all } g \in G;$$

that is, the following diagram commutes

$$\begin{tikzcd}
\mathfrak{g} \arrow[r, "d\phi_1"] \arrow[d, "\mathrm{Ad}_g"'] & \mathfrak{h} \arrow[d, "\mathrm{Ad}_{\phi(g)}"] \\
\mathfrak{g} \arrow[r, "d\phi_1"'] & \mathfrak{h}
\end{tikzcd}$$

and $d\phi_1 \colon \mathfrak{g} \to \mathfrak{h}$ *is a Lie algebra homomorphism.*

Proof. Recall that

$$\mathbf{Ad}_a(b) = R_{a^{-1}} L_a(b) = aba^{-1}, \quad \text{for all } a, b \in G$$

and that the derivative

$$d(\mathbf{Ad}_a)_1 \colon \mathfrak{g} \to \mathfrak{g}$$

of $\mathbf{Ad}_a$ at 1 is an isomorphism of Lie algebras, denoted by $\mathrm{Ad}_a \colon \mathfrak{g} \to \mathfrak{g}$. The map $a \mapsto \mathrm{Ad}_a$ is a map of Lie groups

$$\mathrm{Ad} \colon G \to \mathbf{GL}(\mathfrak{g}),$$

(where $\mathbf{GL}(\mathfrak{g})$ denotes the Lie group of all bijective linear maps on $\mathfrak{g}$) and the derivative

$$d\mathrm{Ad}_1 \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$$

of Ad at 1 is map of Lie algebras, denoted by

$$\mathrm{ad} \colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

called the adjoint representation of $\mathfrak{g}$ (where $\mathfrak{gl}(\mathfrak{g})$ denotes the Lie algebra of all linear maps on $\mathfrak{g}$). Then the Lie bracket is defined by

$$[u,\, v] = \mathrm{ad}(u)(v), \quad \text{for all } u, v \in \mathfrak{g}.$$

Now as ϕ is a homomorphism, we have $\phi(1) = 1$, and we have

$$\phi(\mathbf{Ad}_a(b)) = \phi(aba^{-1}) = \phi(a)\phi(b)\phi(a)^{-1} = R_{\phi(a)^{-1}} L_{\phi(a)}(\phi(b)) = \mathbf{Ad}_{\phi(a)}(\phi(b)).$$

By differentiating w.r.t. b at $b = 1$ in the direction, $v \in \mathfrak{g}$, we get

$$d\phi_1(\mathrm{Ad}_a(v)) = \mathrm{Ad}_{\phi(a)}(d\phi_1(v)),$$

proving the first part of the proposition. Differentiating again with respect to a at $a = 1$ in the direction, $u \in \mathfrak{g}$, (and using the chain rule, along with the fact that $d\phi_1$), we get

$$d\phi_1(\mathrm{ad}(u)(v)) = \mathrm{ad}(d\phi_1(u))(d\phi_1(v)),$$

i.e.,

$$d\phi_1[u, v] = [d\phi_1(u), d\phi_1(v)],$$

which proves that $d\phi_1$ is indeed a Lie algebra homomorphism. □

Remark: If we identify the Lie algebra $\mathfrak{g}$ of G with the space $\mathfrak{g}^L$ of left-invariant vector fields on G, then the map $d\phi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is viewed as the map such that, for every left-invariant vector field X on G, the vector field $d\phi_1(X)$ is the unique left-invariant vector field on H such that

$$d\phi_1(X)(1) = d\phi_1(X(1)),$$

i.e., $d\phi_1(X) = d\phi_1(X(1))^L$. Then we can give another proof of the fact that $d\phi_1$ is a Lie algebra homomorphism using the notion of ϕ-related vector fields.

Proposition 18.15. *If G and H are two Lie groups and if $\phi\colon G \to H$ is a homomorphism of Lie groups, if we identify $\mathfrak{g}$ (resp. $\mathfrak{h}$) with the space of left-invariant vector fields on G (resp. left-invariant vector fields on H), then*

(a) X and $d\phi_1(X)$ are ϕ-related for every left-invariant vector field X on G;

(b) $d\phi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is a Lie algebra homomorphism.

Proof. The proof uses Proposition 9.6. For details see Warner [114] (Chapter 3). □

We now consider Lie subgroups. The following proposition shows that an injective Lie group homomorphism is an immersion.

Proposition 18.16. *If $\phi\colon G \to H$ is an injective Lie group homomorphism, then the map $d\phi_g \colon T_gG \to T_{\phi(g)}H$ is injective for all $g \in G$.*

Proof. As $\mathfrak{g} = T_1G$ and T_gG are isomorphic for all $g \in G$ (and similarly for $\mathfrak{h} = T_1H$ and T_hH for all $h \in H$), it is sufficient to check that $d\phi_1 \colon \mathfrak{g} \to \mathfrak{h}$ is injective. However, by Proposition 18.7, the diagram

$$\begin{tikzcd}
G \arrow[r, "\phi"] & H \\
\mathfrak{g} \arrow[u, "\exp"] \arrow[r, "d\phi_1"'] & \mathfrak{h} \arrow[u, "\exp"']
\end{tikzcd}$$

commutes, and since the exponential map is a local diffeomorphism at 0, as ϕ is injective, then $d\phi_1$ is injective, too. □

Therefore, if $\phi\colon G \to H$ is injective, it is automatically an immersion.

Definition 18.13. Let G be a Lie group. A set H is an *immersed (Lie) subgroup* of G iff

(a) H is a Lie group;

(b) There is an injective Lie group homomorphism $\phi\colon H \to G$ (and thus, ϕ is an immersion, as noted above).

We say that H is a *Lie subgroup* (or *closed Lie subgroup*) of G iff H is a Lie group which is a subgroup of G, and also a submanifold of G.

Observe that an immersed Lie subgroup H is an immersed submanifold, since ϕ is an injective immersion (see Definition 7.27.) However, $\phi(H)$ may *not* have the subspace topology inherited from G and $\phi(H)$ may not be closed, so H is not necessarily a submanifold.

An example of this situation is provided by the 2-torus $T^2 \cong \mathbf{SO}(2) \times \mathbf{SO}(2)$, which can be identified with the group of 2×2 complex diagonal matrices of the form

$$\begin{pmatrix} e^{i\theta_1} & 0 \\ 0 & e^{i\theta_2} \end{pmatrix}$$

where $\theta_1, \theta_2 \in \mathbb{R}$. For any $c \in \mathbb{R}$, let S_c be the subgroup of T^2 consisting of all matrices of the form

$$\begin{pmatrix} e^{it} & 0 \\ 0 & e^{ict} \end{pmatrix}, \quad t \in \mathbb{R}.$$

It is easily checked that S_c is an immersed Lie subgroup of T^2 iff c is irrational. However, when c is irrational, one can show that S_c is dense in T^2 but not closed.

As we will see below, *a Lie subgroup is always closed*. We borrowed the terminology "immersed subgroup" from Fulton and Harris [46] (Chapter 7), but we warn the reader that most books call such subgroups "Lie subgroups" and refer to the second kind of subgroups (that are submanifolds) as "closed subgroups."

Theorem 18.17. *Let G be a Lie group and let (H, ϕ) be an immersed Lie subgroup of G. Then ϕ is an embedding iff $\phi(H)$ is closed in G. As as consequence, any Lie subgroup of G is closed.*

Proof. The proof can be found in Warner [114] (Chapter 1, Theorem 3.21) and Lee [76] (Chapter 20, Theorem 20.10), and uses a little more machinery than we have introduced. However, we prove that a Lie subgroup H of G is closed. The key to the argument is this. Since H is a submanifold of G, there is chart (U, φ) of G, with $1 \in U$, Definition 7.26 implies that

$$\varphi(U \cap H) = \varphi(U) \cap (R^m \times \{0_{n-m}\}).$$

By Proposition 4.4, we can find some open subset $V \subseteq U$ with $1 \in V$, so that $V = V^{-1}$ and $\overline{V} \subseteq U$. Observe that

$$\varphi(\overline{V} \cap H) = \varphi(\overline{V}) \cap (R^m \times \{0_{n-m}\})$$

and since $\overline{V}$ is closed and φ is a homeomorphism, it follows that $\overline{V} \cap H$ is closed. Thus, $\overline{V} \cap H = \overline{V} \cap \overline{H}$ (as $\overline{\overline{V} \cap H} = \overline{V} \cap \overline{H}$). Now pick any $y \in \overline{H}$. As $1 \in V^{-1}$, the open set yV^{-1} contains y and since $y \in \overline{H}$, we must have $yV^{-1} \cap H \neq \emptyset$. Let $x \in yV^{-1} \cap H$, then $x \in H$ and $y \in xV$. Then, $y \in xV \cap \overline{H}$, which implies $x^{-1}y \in V \cap \overline{H} \subseteq \overline{V} \cap \overline{H} = \overline{V} \cap H$. Therefore, $x^{-1}y \in H$ and since $x \in H$, we get $y \in H$ and H is closed. □

We also have the following important and useful theorem: If G is a Lie group, say that a subset $H \subseteq G$ is an *abstract subgroup* iff it is just a subgroup of the underlying group of G (i.e., we forget the topology and the manifold structure).

Theorem 18.18. *Let G be a Lie group. An abstract subgroup H of G is a submanifold (i.e., a Lie subgroup) of G iff H is closed (i.e, H with the induced topology is closed in G).*

Proof. We proved the easy direction of this theorem above. Conversely, we need to prove that if the subgroup H with the induced topology is closed in G, then it is a manifold. This can be done using the exponential map, but it is harder. For details, see Bröcker and tom Dieck [24] (Chapter 1, Section 3, Theorem 3.11) or Warner [114], (Chapter 3, Theorem 3.42). □

18.4 The Correspondence Lie Groups–Lie Algebras

Historically, Lie was the first to understand that a lot of the structure of a Lie group is captured by its Lie algebra, a simpler object (since it is a vector space). In this short section, we state without proof some of the "Lie theorems," although not in their original form.

Definition 18.14. If $\mathfrak{g}$ is a Lie algebra, a *subalgebra* $\mathfrak{h}$ of $\mathfrak{g}$ is a (linear) subspace of $\mathfrak{g}$ such that $[u, v] \in \mathfrak{h}$, for all $u, v \in \mathfrak{h}$. If $\mathfrak{h}$ is a (linear) subspace of $\mathfrak{g}$ such that $[u, v] \in \mathfrak{h}$ for all $u \in \mathfrak{h}$ and all $v \in \mathfrak{g}$, we say that $\mathfrak{h}$ is an *ideal* in $\mathfrak{g}$.

For a proof of the theorem below see Warner [114] (Chapter 3, Theorem 3.19), Duistermaat and Kolk [43] (Chapter 1, Section 10, Theorem 1.10.3), and Lee [76] (Chapter 20, Theorem 20.13).

Theorem 18.19. *Let G be a Lie group with Lie algebra $\mathfrak{g}$, and let (H, ϕ) be an immersed Lie subgroup of G with Lie algebra $\mathfrak{h}$; then $d\phi_1\mathfrak{h}$ is a Lie subalgebra of $\mathfrak{g}$. Conversely, for each subalgebra $\widetilde{\mathfrak{h}}$ of $\mathfrak{g}$, there is a unique connected immersed subgroup (H, ϕ) of G so that $d\phi_1\mathfrak{h} = \widetilde{\mathfrak{h}}$. In fact, as a group, $\phi(H)$ is the subgroup of G generated by $\exp(\widetilde{\mathfrak{h}})$. Furthermore, if G is connected, connected normal subgroups correspond to ideals.*

Theorem 18.19 shows that *there is a one-to-one correspondence between connected immersed subgroups of a Lie group and subalgebras of its Lie algebra.*

Theorem 18.20. *Let G and H be Lie groups with G connected and simply connected and let $\mathfrak{g}$ and $\mathfrak{h}$ be their Lie algebras. For every homomorphism $\psi\colon \mathfrak{g} \to \mathfrak{h}$, there is a unique Lie group homomorphism $\phi\colon G \to H$ so that $d\phi_1 = \psi$.*

Again a proof of the theorem above is given in Warner [114] (Chapter 3, Theorem 3.27) and in Lee [76] (Chapter 20, Theorem 20.15).

Corollary 18.21. *If G and H are connected and simply connected Lie groups, then G and H are isomorphic iff $\mathfrak{g}$ and $\mathfrak{h}$ are isomorphic.*

It can also be shown that for every finite-dimensional Lie algebra $\mathfrak{g}$, there is a connected and simply connected Lie group G such that $\mathfrak{g}$ is the Lie algebra of G. This result is known as *Lie's third theorem.*

Lie's third theorem was first prove by Élie Cartan; see Serre [105]. It is also a consequence of deep theorem known as *Ado's theorem.* Ado's theorem states that every finite-dimensional Lie algebra has a faithful representation in $\mathfrak{gl}(n, \mathbb{R}) = \mathrm{M}_n(\mathbb{R})$ for some n. The proof is quite involved; see Knapp [68] (Appendix C) Fulton and Harris [46] (Appendix E), or Bourbaki [19] (Chapter 1, Section §7).

As a corollary of Lie's third theorem, there is a one-to-one correspondence between isomorphism classes of finite-dimensional Lie algebras and isomorphism classes of simply-connected Lie groups, given by associating each simply connnected Lie group with its Lie algebra.; see Lee [76] (Theorem 20.20) and Warner [114] (Theorem 3.28).

In summary, following Fulton and Harris, we have the following two principles of the Lie group/Lie algebra correspondence:

First Principle: (*restatement of Proposition 18.8*:) If G and H are Lie groups, with G connected, then a homomorphism of Lie groups $\phi\colon G \to H$ is uniquely determined by the Lie algebra homomorphism $d\phi_1\colon \mathfrak{g} \to \mathfrak{h}$.

Second Principle: (*restatement of Theorem 18.20*:) Let G and H be Lie groups with G connected and simply connected and let $\mathfrak{g}$ and $\mathfrak{h}$ be their Lie algebras. A linear map $\psi\colon \mathfrak{g} \to \mathfrak{h}$ is a Lie algebra map iff there is a unique Lie group homomorphism $\phi\colon G \to H$ so that $d\phi_1 = \psi$.

18.5 Semidirect Products of Lie Algebras and Lie Groups

The purpose of this section is to construct an entire class of Lie algebras and Lie groups by combining two "smaller" pieces in a manner which preserves the algebraic structure. We begin with two Lie algebras and form a new vector space via the direct sum. If $\mathfrak{a}$ and $\mathfrak{b}$ are two Lie algebras, recall that the direct sum $\mathfrak{a} \oplus \mathfrak{b}$ of $\mathfrak{a}$ and $\mathfrak{b}$ is $\mathfrak{a} \times \mathfrak{b}$ with the product vector space structure where

$$(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$$

for all $a_1, a_2 \in \mathfrak{a}$ and all $b_1, b_2 \in \mathfrak{b}$, and

$$\lambda(a, b) = (\lambda a, \lambda b)$$

for all $\lambda \in \mathbb{R}$, all $a \in \mathfrak{a}$, and all $b \in \mathfrak{b}$. The map $a \mapsto (a, 0)$ is an isomorphism of $\mathfrak{a}$ with the subspace $\{(a, 0) \mid a \in \mathfrak{a}\}$ of $\mathfrak{a} \oplus \mathfrak{b}$ and the map $b \mapsto (0, b)$ is an isomorphism of $\mathfrak{b}$ with the subspace $\{(0, b) \mid b \in \mathfrak{b}\}$ of $\mathfrak{a} \oplus \mathfrak{b}$. These isomorphisms allow us to identify $\mathfrak{a}$ with the subspace $\{(a, 0) \mid a \in \mathfrak{a}\}$ and $\mathfrak{b}$ with the subspace $\{(0, b) \mid b \in \mathfrak{b}\}$.

The simplest way to make the direct sum $\mathfrak{a} \oplus \mathfrak{b}$ into a Lie algebra is by defining the Lie bracket $[-,-]$ such that $[a_1, a_2]$ agrees with the Lie bracket on $\mathfrak{a}$ for all $a_1, a_2, \in \mathfrak{a}$, $[b_1, b_2]$ agrees with the Lie bracket on $\mathfrak{b}$ for all $b_1, b_2, \in \mathfrak{b}$, and $[a, b] = [b, a] = 0$ for all $a \in \mathfrak{a}$ and all $b \in \mathfrak{b}$. In particular, if $[-,-]_{\mathfrak{a}}$ and $[-,-]_{\mathfrak{b}}$ denote the Lie bracket on $\mathfrak{a}$ and $\mathfrak{b}$ respectively, the preceding sentence says

$$\begin{aligned}
[(a_1, 0), (a_2, 0)] &= [a_1, a_2]_{\mathfrak{a}} \\
[(0, b_1), (0, b_2)] &= [b_1, b_2]_{\mathfrak{b}} \\
[(a_1, 0), (0, b_1)] &= 0 = [(0, b_1), (a_1, 0)].
\end{aligned}$$

Hence

$$[(a_1, b_1), (a_2, b_2)] = [(a_1, 0), (a_2, 0)] + [(0, b_1), (0, b_2)] = ([a_1, a_2]_{\mathfrak{a}}, [b_1, b_2]_{\mathfrak{b}}).$$

Definition 18.15. If $\mathfrak{a}$ and $\mathfrak{b}$ are two Lie algebras, the direct sum $\mathfrak{a} \oplus \mathfrak{b}$ with the bracket defined by

$$[(a_1, b_1), (a_2, b_2)] = ([a_1, a_2]_{\mathfrak{a}}, [b_1, b_2]_{\mathfrak{b}})$$

for all $a_1, a_2, \in \mathfrak{a}$ and all $b_1, b_2, \in \mathfrak{b}$ is a Lie algebra is called the *Lie algebra direct sum* of $\mathfrak{a}$ and $\mathfrak{b}$.

Observe that with this Lie algebra structure, $\mathfrak{a}$ and $\mathfrak{b}$ are ideals.

For example, let $\mathfrak{a} = \mathbb{R}^n$ with the zero bracket, and let $\mathfrak{b} = \mathfrak{so}(n)$ be the Lie algebra of $n \times n$ skew symmetric matrices with the commutator bracket. Then $\mathfrak{g} = \mathbb{R}^n \oplus \mathfrak{so}(n)$ is a Lie algebra with $[-,-]$ defined as $[u, v] = 0$ for all $u, v \in \mathbb{R}^n$, $[A, B] = AB - BA$ for all $A, B \in \mathfrak{so}(n)$, and $[u, A] = 0$ for all $u \in \mathbb{R}^n, A \in \mathfrak{so}(n)$.

The above construction is sometimes called an "external direct sum" because it does not assume that the constituent Lie algebras $\mathfrak{a}$ and $\mathfrak{b}$ are subalgebras of some given Lie algebra $\mathfrak{g}$.

Definition 18.16. If $\mathfrak{a}$ and $\mathfrak{b}$ are subalgebras of a given Lie algebra $\mathfrak{g}$ such that $\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}$ is a direct sum as a vector space and if both $\mathfrak{a}$ and $\mathfrak{b}$ are ideals, then for all $a \in \mathfrak{a}$ and all $b \in \mathfrak{b}$, we have $[a, b] \in \mathfrak{a} \cap \mathfrak{b} = (0)$, so $\mathfrak{a} \oplus \mathfrak{b}$ is the Lie algebra direct sum of $\mathfrak{a}$ and $\mathfrak{b}$. This Lie algeba is called an *internal direct sum.*

We now would like to generalize this construction to the situation where the Lie bracket $[a, b]$ of some $a \in \mathfrak{a}$ and some $b \in \mathfrak{b}$ is given in terms of a map from $\mathfrak{b}$ to $\mathrm{Hom}(\mathfrak{a}, \mathfrak{a})$. For this to work, we need to consider derivations.

Definition 18.17. Given a Lie algebra $\mathfrak{g}$, a *derivation* is a linear map $D\colon \mathfrak{g} \to \mathfrak{g}$ satisfying the following condition:

$$D([X, Y]) = [D(X), Y] + [X, D(Y)], \quad \text{for all } X, Y \in \mathfrak{g}.$$

The vector space of all derivations on $\mathfrak{g}$ is denoted by $\mathrm{Der}(\mathfrak{g})$.

Given a Lie algebra with $[-,-]$, we may use this bracket structure to define $\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$ as $\mathrm{ad}(u)(v) = [u,v]$. Then the Jacobi identity can be expressed as

$$[Z,[X,Y]] = [[Z,X],Y] + [X,[Z,Y]],$$

which holds iff

$$(\mathrm{ad}\, Z)[X,Y] = [(\mathrm{ad}\, Z)X, Y] + [X, (\mathrm{ad}\, Z)Y],$$

and the above equation means that $\mathrm{ad}(Z)$ is a derivation. In fact, it is easy to check that the Jacobi identity holds iff $\mathrm{ad}\, Z$ is a derivation for every $Z \in \mathfrak{g}$. It tuns out that the vector space of derivations $\mathrm{Der}(\mathfrak{g})$ is a Lie algebra under the commutator bracket.

Proposition 18.22. *For any Lie algebra $\mathfrak{g}$, the vector space $\mathrm{Der}(\mathfrak{g})$ is a Lie algebra under the commutator bracket. Furthermore, the map $\mathrm{ad}\colon \mathfrak{g} \to \mathrm{Der}(\mathfrak{g})$ is a Lie algebra homomorphism.*

Proof. For any $D, E \in \mathrm{Der}(\mathfrak{g})$ and any $X, Y \in \mathfrak{g}$, we have

$$\begin{aligned}
[D,E][X,Y] &= (DE - ED)[X,Y] = DE[X,Y] - ED[X,Y] \\
&= D[EX,Y] + D[X,EY] - E[DX,Y] - E[X,DY] \\
&= [DEX,Y] + [EX,DY] + [DX,EY] + [X,DEY] \\
&\quad - [EDX,Y] - [DX,EY] - [EX,DY] - [X,EDY] \\
&= [DEX,Y] - [EDX,Y] + [X,DEY] - [X,EDY] \\
&= [[D,E]X,Y] + [X,[D,E]Y],
\end{aligned}$$

which proves that $[D,E]$ is a derivation. Thus, $\mathrm{Der}(\mathfrak{g})$ is a Lie algebra. We already know that $\mathrm{ad}\, X$ is a derivation for all $X \in \mathfrak{g}$, so $\mathrm{ad}\, \mathfrak{g} \subseteq \mathrm{Der}(\mathfrak{g})$. For all $X, Y \in \mathfrak{g}$, we need to show that

$$\mathrm{ad}\, [X,Y] = (\mathrm{ad}\, X) \circ (\mathrm{ad}\, Y) - (\mathrm{ad}\, Y) \circ (\mathrm{ad}\, X).$$

If we apply both sides to any $Z \in \mathfrak{g}$, we get

$$(\mathrm{ad}\, [X,Y])(Z) = (\mathrm{ad}\, X)((\mathrm{ad}\, Y)(Z)) - (\mathrm{ad}\, Y)((\mathrm{ad}\, X)(Z)),$$

that is,

$$[[X,Y],Z] = [X,[Y,Z]] - [Y,[X,Z]],$$

which is equivalent to

$$[[X,Y],Z] + [[Y,Z],X] + [[Z,X],Y] = 0,$$

which is the Jacobi identity. Therefore, ad is a Lie algebra homomorphism. □

Proposition 18.23. *For any Lie algebra $\mathfrak{g}$ If $D \in \mathrm{Der}(\mathfrak{g})$ and $X \in \mathfrak{g}$, then*

$$[D, \mathrm{ad}\, X] = \mathrm{ad}\, (DX).$$

Proof. For all $Z \in \mathfrak{g}$, $D \in \mathrm{Der}(\mathfrak{g})$, and $X \in \mathfrak{g}$, we have

$$\begin{aligned}[D, \mathrm{ad}\, X]Z &= D(\mathrm{ad}\, X(Z)) - \mathrm{ad}\, X(D(Z)) \\ &= D[X, Z] - [X, DZ] \\ &= [DX, Z] + [X, DZ] - [X, DZ] \\ &= [DX, Z] = \mathrm{ad}\,(DX)(Z). \qquad \square\end{aligned}$$

We would like to describe another way of defining a bracket structure on $\mathfrak{a} \oplus \mathfrak{b}$ using $\mathrm{Der}(\mathfrak{a})$. To best understand this construction, let us go back to our previous example where $\mathfrak{a} = \mathbb{R}^n$ with $[-,-]_{\mathfrak{a}} = 0$, and $\mathfrak{b} = \mathfrak{so}(n)$ with $[A, B]_{\mathfrak{b}} = AB - BA$ for all $A, B \in \mathfrak{so}(n)$. The underlying vector space is $\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b} = \mathbb{R}^n \oplus \mathfrak{so}(n)$, but this time the bracket on $\mathfrak{g}$ is defined as

$$[(u, A), (v, B)] = (Av - Bu, [A, B]_{\mathfrak{b}}), \qquad u, v \in \mathbb{R}^n, \qquad A, B \in \mathfrak{so}(n).$$

By using the isomorphism between $\mathfrak{a}$ and $\{(a, 0) \mid a \in \mathfrak{a}\}$ and the isomorphism between $\mathfrak{b}$ and $\{(0, b) \mid b \in \mathfrak{b}\}$, we have

$$[u, v]_{\mathfrak{a}} = [(u, 0), (v, 0)] = (0, 0),$$

and

$$[A, B]_{\mathfrak{b}} = [(0, A), (0, B)] = (0, [A, B]_{\mathfrak{b}}).$$

Furthermore

$$[(u, 0), (0, B)] = (-Bu, [0, B]_{\mathfrak{b}}) = (-Bu, 0) \in \mathfrak{a}.$$

Hence, $\mathfrak{a}$ is an ideal in $\mathfrak{g}$. With this bracket structure, we have $\mathfrak{g} = \mathfrak{se}(n)$, the Lie algebra of $\mathbf{SE}(n)$ (see Section 1.6).

How does this bracket structure on $\mathfrak{g} = \mathfrak{se}(n)$ relate to $\mathrm{Der}(\mathfrak{a})$? Since $\mathfrak{a} = \mathbb{R}^n$ is an abelian Lie algebra, $\mathrm{Der}(\mathfrak{a}) = \mathfrak{gl}(n, \mathbb{R})$. Define $\tau \colon \mathfrak{b} \to \mathfrak{gl}(n, \mathbb{R})$ to be the inclusion map, i.e. $\tau(B) = B$ for $B \in \mathfrak{so}(n)$. Then

$$[(u, A), (v, B)] = ([u, v]_{\mathfrak{a}} + \tau(A)v - \tau(B)v, [A, B]_{\mathfrak{b}}) = (Av - Bu, [A, B]_{\mathfrak{b}}).$$

In other words

$$[(0, A), (v, 0)] = (\tau(A)v, 0),$$

and $[a, b]$ for $a \in \mathfrak{a} = \mathbb{R}^n$ and $b \in \mathfrak{b} = \mathfrak{so}(n)$ is determined by the map τ.

The construction illustrated by this example is summarized in the following proposition.

Proposition 18.24. *Let $\mathfrak{a}$ and $\mathfrak{b}$ be two Lie algebras, and suppose τ is a Lie algebra homomorphism $\tau\colon \mathfrak{b} \to \mathrm{Der}(\mathfrak{a})$. Then there is a unique Lie algebra structure on the vector space $\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}$ whose Lie bracket agrees with the Lie bracket on $\mathfrak{a}$ and the Lie bracket on $\mathfrak{b}$, and such that*

$$[(0, B), (A, 0)]_{\mathfrak{g}} = \tau(B)(A) \quad \textit{for all } A \in \mathfrak{a} \textit{ and all } B \in \mathfrak{b}. \tag{$*$}$$

The Lie bracket on $\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}$ is given by

$$[(A, B), (A', B')]_{\mathfrak{g}} = ([A, A']_{\mathfrak{a}} + \tau(B)(A') - \tau(B')(A),\ [B, B']_{\mathfrak{b}}),$$

for all $A, A' \in \mathfrak{a}$ and all $B, B' \in \mathfrak{b}$. In particular,

$$[(0, B), (A', 0)]_{\mathfrak{g}} = \tau(B)(A') \in \mathfrak{a}.$$

With this Lie algebra structure, $\mathfrak{a}$ is an ideal and $\mathfrak{b}$ is a subalgebra.

Proof. Uniqueness of the Lie algebra structure is forced by the fact that the Lie bracket is bilinear and skew symmetric. The problem is to check the Jacobi identity. Pick $X, Y, Z \in \mathfrak{g}$. If all three are in $\mathfrak{a}$ or in $\mathfrak{b}$, we are done. By skew symmetry, we are reduced to two cases:

1. X is in $\mathfrak{a}$ and Y, Z are in $\mathfrak{b}$, to simplify notation, write X for $(X, 0)$ and Y, Z for $(0, Y)$ and $(0, Z)$. Since τ is a Lie algebra homomorphism,

$$\tau([Y, Z]) = \tau(Y)\tau(Z) - \tau(Z)\tau(Y).$$

If we apply both sides to X, we get

$$\tau([Y, Z])(X) = (\tau(Y)\tau(Z))(X) - (\tau(Z)\tau(Y))(X),$$

that is, by $(*)$,

$$[[Y, Z], X] = [Y, [Z, X]] - [Z, [Y, X]],$$

or equivalently

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0,$$

which is the Jacobi identity.

2. X, Y are in $\mathfrak{a}$ and Z is in $\mathfrak{b}$, again to simplify notation, write X, Y for $(X, 0)$ and $(Y, 0)$ and Z for $(0, Z)$ Since $\tau(Z)$ is a derivation, we have

$$\tau(Z)([X, Y]) = [\tau(Z)(X), Y] + [X, \tau(Z)(Y)],$$

which, by $(*)$, is equivalent to

$$[Z, [X, Y]] = [[Z, X], Y] + [X, [Z, Y]],$$

a version of the Jacobi identity.

Since both $\mathfrak{a}$ and $\mathfrak{b}$ bracket into $\mathfrak{a}$, we conclude that $\mathfrak{a}$ is an ideal. □

Definition 18.18. The Lie algebra obtained in Proposition 18.24 is denoted by

$$\mathfrak{a} \oplus_\tau \mathfrak{b} \quad \text{or} \quad \mathfrak{a} \rtimes_\tau \mathfrak{b}$$

and is called the *semidirect product of* $\mathfrak{b}$ *by* $\mathfrak{a}$ *with respect to* $\tau\colon \mathfrak{b} \to \mathrm{Der}(\mathfrak{a})$.

When τ is the zero map, we get back the Lie algebra direct sum.

Remark: A sequence of Lie algebra maps

$$\begin{tikzcd} \mathfrak{a} \arrow[r, "\varphi"] & \mathfrak{g} \arrow[r, "\psi"] & \mathfrak{b} \end{tikzcd}$$

with φ injective, ψ surjective, and with $\mathrm{Im}\,\varphi = \mathrm{Ker}\,\psi = \mathfrak{n}$, is called an *extension of* $\mathfrak{b}$ *by* $\mathfrak{a}$ *with kernel* $\mathfrak{n}$. If there is a subalgebra $\mathfrak{p}$ of $\mathfrak{g}$ such that $\mathfrak{g}$ is a direct sum $\mathfrak{g} = \mathfrak{n} \oplus \mathfrak{p}$, then we say that this extension is *inessential*. Given a semidirect product $\mathfrak{g} = \mathfrak{a} \rtimes_\tau \mathfrak{b}$ of $\mathfrak{b}$ by $\mathfrak{a}$, if $\varphi\colon \mathfrak{a} \to \mathfrak{g}$ is the map given $\varphi(a) = (a, 0)$ and ψ is the map $\psi\colon \mathfrak{g} \to \mathfrak{b}$ given by $\psi(a, b) = b$, then $\mathfrak{g}$ is an inessential extension of $\mathfrak{b}$ by $\mathfrak{a}$. Conversely, it is easy to see that every inessential extension of $\mathfrak{b}$ by $\mathfrak{a}$ is a semidirect product of $\mathfrak{b}$ by $\mathfrak{a}$.

Proposition 18.24 is an external construction. The notion of semidirect product has a corresponding internal construction. If $\mathfrak{g}$ is a Lie algebra and if $\mathfrak{a}$ and $\mathfrak{b}$ are subspaces of $\mathfrak{g}$ such that

$$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b},$$

$\mathfrak{a}$ is an ideal in $\mathfrak{g}$ and $\mathfrak{b}$ is a subalgebra of $\mathfrak{g}$, then for every $B \in \mathfrak{b}$, because $\mathfrak{a}$ is an ideal, the restriction of $\mathrm{ad}\, B$ to $\mathfrak{a}$ leaves $\mathfrak{a}$ invariant, so by Proposition 18.22, the map $B \mapsto \mathrm{ad}\, B \restriction \mathfrak{a}$ is a Lie algebra homomorphism $\tau\colon \mathfrak{b} \to \mathrm{Der}(\mathfrak{a})$. Observe that $[B, A] = \tau(B)(A)$, for all $A \in \mathfrak{a}$ and all $B \in \mathfrak{b}$, so the Lie bracket on $\mathfrak{g}$ is completely determined by the Lie brackets on $\mathfrak{a}$ and $\mathfrak{b}$ and the homomorphism τ. We say that $\mathfrak{g}$ is the *semidirect product* of $\mathfrak{b}$ and $\mathfrak{a}$ and we write

$$\mathfrak{g} = \mathfrak{a} \oplus_\tau \mathfrak{b}.$$

Semidirect products of Lie algebras are discussed in Varadarajan [113] (Section 3.14), Bourbaki [19], (Chapter 1, Section 8), and Knapp [68] (Chapter 1, Section 4). However, beware that Knapp switches the roles of $\mathfrak{a}$ and $\mathfrak{b}$, and τ is a Lie algebra map $\tau\colon \mathfrak{a} \to \mathrm{Der}(\mathfrak{b})$.

Before turning our attention to semidirect products of Lie groups, let us consider the group $\mathrm{Aut}(\mathfrak{g})$ of Lie algebra isomorphisms of a Lie algebra $\mathfrak{g}$.

Definition 18.19. Given a Lie algebra $\mathfrak{g}$, the *group of Lie algebra automorphisms* of $\mathfrak{g}$ is denoted by $\mathrm{Aut}(\mathfrak{g})$.

The group $\mathrm{Aut}(\mathfrak{g})$ is a subgroup of the group $\mathbf{GL}(\mathfrak{g})$ of linear automorphisms φ of $\mathfrak{g}$, and since the condition

$$\varphi([u, v]) = [\varphi(u), \varphi(v)]$$

passes to the limit, it is easy to see that it is closed, so it is a Lie group. It turns out that its Lie algebra is $\mathrm{Der}(\mathfrak{g})$.

Proposition 18.25. *For any (real) Lie algebra $\mathfrak{g}$, the Lie algebra* $\mathrm{L}(\mathrm{Aut}(\mathfrak{g}))$ *of the group* $\mathrm{Aut}(\mathfrak{g})$ *is* $\mathrm{Der}(\mathfrak{g})$*, the Lie algebra of derivations of* $\mathfrak{g}$*.*

Proof. For any $f \in \mathrm{L}(\mathrm{Aut}(\mathfrak{g}))$, let $\gamma(t)$ be a smooth curve in $\mathrm{Aut}(\mathfrak{g})$ such that $\gamma(0) = I$ and $\gamma'(0) = f$. Since $\gamma(t)$ is a Lie algebra automorphism

$$\gamma(t)([X, Y]) = [\gamma(t)(X), \gamma(t)(Y)]$$

for all $X, Y \in \mathfrak{g}$, and using the product rule and taking the derivative for $t = 0$, we get

$$\begin{aligned}\gamma'(0)([X, Y]) = f([X, Y]) &= [\gamma'(0)(X), \gamma(0)(Y)] + [\gamma(0)(X), \gamma'(0)(Y)] \\ &= [f(X), Y] + [Y, f(X)],\end{aligned}$$

which shows that f is a derivation.

Conversely, pick any $f \in \mathrm{Der}(\mathfrak{g})$. We prove that $e^{tf} \in \mathrm{Aut}(\mathfrak{g})$ for all $t \in \mathbb{R}$, which shows that $\mathrm{Der}(\mathfrak{g}) \subseteq \mathrm{L}(\mathrm{Aut}(\mathfrak{g}))$. For any $X, Y \in \mathfrak{g}$, consider the two curves in $\mathfrak{g}$ given by

$$\gamma_1(t) = e^{tf}[X, Y] \quad \text{and} \quad \gamma_2(t) = [e^{tf}X, e^{tf}Y].$$

For $t = 0$, we have $\gamma_1(0) = \gamma_2(0) = [X, Y]$. We find that

$$\gamma_1'(t) = fe^{tf}[X, Y] = f\gamma_1(t),$$

and since f is a derivation

$$\begin{aligned}\gamma_2'(t) &= [fe^{tf}X, e^{tf}Y] + [e^{tf}, fe^{tf}Y] \\ &= f[e^{tf}X, e^{tf}Y] \\ &= f\gamma_2(t).\end{aligned}$$

Since γ_1 and γ_2 are maximal integral curves for the linear vector field defined by f, and with the same initial condition, by uniqueness, we have

$$e^{tf}[X, Y] = [e^{tf}X, e^{tf}Y] \quad \text{for all } t \in \mathbb{R},$$

which shows that e^{tf} is a Lie algebra automorphism. Therefore, $f \in \mathrm{L}(\mathrm{Aut}(\mathfrak{g}))$. □

Since $(d\mathbf{Ad}_a)_1 = \mathrm{Ad}_a$ is a Lie algebra isomorphism of $\mathfrak{g}$, Proposition 18.1 implies that Ad is a Lie group homomorphism

$$\mathrm{Ad}\colon G \to \mathrm{Aut}(\mathfrak{g}),$$

and Propositions 18.14 and 18.25 imply that ad is a Lie algebra homomorphism

$$\mathrm{ad}\colon \mathfrak{g} \to \mathrm{Der}(\mathfrak{g}).$$

Remark: It can be shown that if $\mathfrak{g}$ is semisimple (see Section 20.5 for the definition of a semisimple Lie algebra), then $\mathrm{ad}(\mathfrak{g}) = \mathrm{Der}(\mathfrak{g})$.

We now define semidirect products of Lie groups and show how their algebras are semidirect products of Lie algebras. We begin with the definition of the semidirect product of two groups.

Proposition 18.26. *Let H and K be two groups and let $\tau\colon K \to \mathrm{Aut}(H)$ be a homomorphism of K into the automorphism group of H, i.e. the set of isomorphisms of H with the group structure given by composition. Let $G = H \times K$ with multiplication defined as follows:*

$$(h_1, k_1)(h_2, k_2) = (h_1\tau(k_1)(h_2), k_1k_2),$$

for all $h_1, h_2 \in H$ and all $k_1, k_2 \in K$. Then the following properties hold:

(1) This multiplication makes G into a group with identity $(1, 1)$ and with inverse given by

$$(h, k)^{-1} = (\tau(k^{-1})(h^{-1}), k^{-1}).$$

(2) The maps $h \mapsto (h, 1)$ for $h \in H$ and $k \mapsto (1, k)$ for $k \in K$ are isomorphisms from H to the subgroup $\{(h, 1) \mid h \in H\}$ of G and from K to the subgroup $\{(1, k) \mid k \in K\}$ of G.

(3) Using the isomorphisms from (2), the group H is a normal subgroup of G.

(4) Using the isomorphisms from (2), $H \cap K = (1)$.

(5) For all $h \in H$ an all $k \in K$, we have

$$(1, k)(h, 1)(1, k)^{-1} = (\tau(k)(h), 1).$$

Proof. We leave the proof of these properties as an exercise, except for (5). Checking associativity takes a little bit of work.

Using the definition of multiplication, since $\tau(k_1)$ is an automorphism of H for all $k_1 \in K$, we have $\tau(k_1)(1) = 1$, which means that

$$(1, k)^{-1} = (1, k^{-1}),$$

so we have

$$\begin{aligned}(1, k)(h, 1)(1, k)^{-1} &= ((1, k)(h, 1))(1, k^{-1})\\ &= (\tau(k)(h), k)(1, k^{-1})\\ &= (\tau(k)(h)\tau(k)(1), kk^{-1})\\ &= (\tau(k)(h), 1),\end{aligned}$$

as claimed. □

In view of Proposition 18.26, we make the following definition.

Definition 18.20. Let H and K be two groups and let $\tau\colon K \to \mathrm{Aut}(H)$ be a homomorphism of K into the automorphism group of H. The group defined in Proposition 18.26 is called the *semidirect product of K by H with respect to τ*, and it is denoted $H \rtimes_\tau K$ (or even $H \rtimes K$).

Note that $\tau\colon K \to \mathrm{Aut}(H)$ can be viewed as a left action $\cdot\colon K \times H \to H$ of K on H "acting by automorphisms," which means that for every $k \in K$, the map $h \mapsto \tau(k, h)$ is an automorphism of H.

Note that when τ is the trivial homomorphism (that is, $\tau(k) = \mathrm{id}$ for all $k \in K$), the semidirect product is just the direct product $H \times K$ of the groups H and K, and K is also a normal subgroup of G.

Semidirect products are used to construct affine groups. For example, let $H = \mathbb{R}^n$ under addition, let $K = \mathbf{SO}(n)$, and let τ be the inclusion map of $\mathbf{SO}(n)$ into $\mathrm{Aut}(\mathbb{R}^n)$. In other words, τ is the action of $\mathbf{SO}(n)$ on $\mathbb{R}^n$ given by $R \cdot u = Ru$. Then the semidirect product $\mathbb{R}^n \rtimes \mathbf{SO}(n)$ is isomorphic to the group $\mathbf{SE}(n)$ of direct affine rigid motions of $\mathbb{R}^n$ (translations and rotations), since the multiplication is given by

$$(u, R)(v, S) = (Rv + u, RS), \qquad u, v \in \mathbb{R}^2, \qquad R, S \in \mathbf{SO}(n).$$

We obtain other affine groups by letting K be $\mathbf{SL}(n)$, $\mathbf{GL}(n)$, *etc.*

Semidirect products of groups are discussed in Varadarajan [113] (Section 3.15), Bourbaki [19], (Chapter 3, Section 1.4), and Knapp [68] (Chapter 1, Section 15). Note that some authors (such as Knapp) define the semidirect product of two groups H and K by letting H act on K. In this case, in order to work, the multiplication must be defined as

$$(h_1, k_1)(h_2, k_2) = (h_1h_2, \tau(h_2^{-1})(k_1)k_2),$$

which involves the inverse h_2^{-1} of h_2. This is because h_2 acts on the element k_1 *on its left*, which makes it a right action. To work properly, we must use h_2^{-1}. In fact, $\tau\colon K \times H \to K$ is a right action of H on K, and in this case, the map from H to $\mathrm{Aut}(K)$ should send h to the map $k \mapsto \tau(h^{-1}, k)$, in order to be a homomorphism.

On the other hand, the way we have defined multiplication as

$$(h_1, k_1)(h_2, k_2) = (h_1\tau(k_1)(h_2), k_1k_2),$$

the element k_1 acts on the element h_2 *on its right*, which makes it a left action and works fine with no inversion needed. The left action seems simpler.

Definition 18.21. A sequence of groups homomorphisms

$$\begin{tikzcd} H \arrow[r, "\varphi"] & G \arrow[r, "\psi"] & K \end{tikzcd}$$

with φ injective, ψ surjective, and with $\mathrm{Im}\,\varphi = \mathrm{Ker}\,\psi = N$, is called an *extension of K by H with kernel N.*

If $H \rtimes_\tau K$ is a semidirect product, we have the homomorphisms $\varphi\colon H \to G$ and $\psi\colon G \to K$ given by

$$\varphi(h) = (h, 1), \qquad \psi(h, k) = k,$$

and it is clear that we have an extension of K by H with kernel $N = \{(h, 1) \mid h \in H\}$. Note that we have a homomorphism $\gamma\colon K \to G$ (a section of ψ) given by

$$\gamma(k) = (1, k),$$

and that

$$\psi \circ \gamma = \mathrm{id}.$$

Conversely, it can be shown that if an extension of K by H has a section $\gamma\colon K \to G$, then G is isomorphic to a semidirect product of K by H with respect to a certain homomorphism τ; find it!

Proposition 18.27. *If H and K are two Lie groups and if the map from $H \times K$ to H given by $(h, k) \mapsto \tau(k)(h)$ is smooth, then the semidirect product $H \rtimes_\tau K$ is a Lie group.*

Proof. (See Varadarajan [113] (Section 3.15), or Bourbaki [19], (Chapter 3, Section 1.4)). This is because

$$\begin{aligned}(h_1, k_1)(h_2, k_2)^{-1} &= (h_1, k_1)(\tau(k_2^{-1})(h_2^{-1}), k_2^{-1})\\ &= (h_1\tau(k_1)(\tau(k_2^{-1})(h_2^{-1})), k_1k_2^{-1})\\ &= (h_1\tau(k_1k_2^{-1})(h_2^{-1}), k_1k_2^{-1}),\end{aligned}$$

which shows that multiplication and inversion in $H \rtimes_\tau K$ are smooth. □

It it not very surprising that the Lie algebra of $H \rtimes_\tau K$ is a semidirect product of the Lie algebras $\mathfrak{h}$ of H and $\mathfrak{k}$ of K.

For every $k \in K$, the derivative of $d(\tau(k))_1$ of $\tau(k)$ at 1 is a Lie algebra isomorphism of $\mathfrak{h}$, and just like Ad, it can be shown that the map $\widetilde{\tau}\colon K \to \mathrm{Aut}(\mathfrak{h})$ given by

$$\widetilde{\tau}(k) = d(\tau(k))_1 \quad k \in K$$

is a smooth homomorphism from K into $\mathrm{Aut}(\mathfrak{h})$. It follows by Proposition 18.25 that its derivative $d\widetilde{\tau}_1\colon \mathfrak{k} \to \mathrm{Der}(\mathfrak{h})$ at 1 is a homomorphism of $\mathfrak{k}$ into $\mathrm{Der}(\mathfrak{h})$.

Proposition 18.28. *Using the notations just introduced, the Lie algebra of the semidirect product $H \rtimes_\tau K$ of K by H with respect to τ is the semidirect product $\mathfrak{h} \rtimes_{d\widetilde{\tau}_1} \mathfrak{k}$ of $\mathfrak{k}$ by $\mathfrak{h}$ with respect to $d\widetilde{\tau}_1$.*

Proof. We follow Varadarajan [113] (Section 3.15), and provide a few more details. The tangent space at the identity of $H \rtimes_\tau K$ is $\mathfrak{h} \oplus \mathfrak{k}$ as a vector space. The bracket structure on $\mathfrak{h} \times \{0\}$ is inherited by the bracket on $\mathfrak{h}$, and similarly the bracket structure on $\{0\} \times \mathfrak{k}$ is inherited by the bracket on $\mathfrak{k}$. We need to figure out the bracket between elements of $\{0\} \times \mathfrak{k}$ and elements of $\mathfrak{h} \times \{0\}$. For any $X \in \mathfrak{h}$ and any $Y \in \mathfrak{k}$, for all $t, s \in \mathbb{R}$, using Proposition

18.10(2), Property (5) of Proposition 18.26, and the fact that $\exp(X, Y) = (\exp(X), \exp(Y))$, we have

$$\begin{aligned}\exp(\mathrm{Ad}(\mathbf{exp}(\mathbf{t}(\mathbf{0}, \mathbf{Y}))(s(X, 0))) &= (\mathbf{exp}(\mathbf{t}(\mathbf{0}, \mathbf{Y})))(\exp(s(X, 0)))(\mathbf{exp}(\mathbf{t}(\mathbf{0}, \mathbf{Y})))^{-1}\\ &= (1, \exp(tY))(\exp(sX), 1)(1, \exp(tY))^{-1}\\ &= (1, \exp(tY))(\exp(sX), 1)(1, \exp(tY)^{-1})\\ &= (1, \exp(tY))(\exp(sX), 1)(1, \exp(-tY))\\ &= (\tau(\exp(tY))(\exp(sX)), 1).\end{aligned}$$

For fixed t, taking the derivative with respect to s at $s = 0$, and using the chain rule, we deduce that

$$\mathrm{Ad}(\exp(t(0, Y)))(X, 0) = (\widetilde{\tau}(\exp(tY))(X), 0).$$

Taking the derivative with respect to t at $t = 0$, and using the chain rule, we get

$$[(0, Y), (X, 0)] = (\mathrm{ad}\,(0, Y))(X, 0) = (d\widetilde{\tau}_1(Y)(X), 0),$$

which shows that the Lie bracket between elements of $\{0\} \times \mathfrak{k}$ and elements of $\mathfrak{h} \times \{0\}$ is given by $d\widetilde{\tau}_1$. The reader should fill in the details of the above computations. □

Proposition 18.28 applied to the semidirect product $\mathbb{R}^n \rtimes_\tau \mathbf{SO}(n) \cong \mathbf{SE}(n)$ where τ is the inclusion map of $\mathbf{SO}(n)$ into $\mathrm{Aut}(\mathbb{R}^n)$ confirms that $\mathbb{R}^n \rtimes_{d\widetilde{\tau}_1} \mathfrak{so}(n)$ is the Lie algebra of $\mathbf{SE}(n)$, where $d\widetilde{\tau}_1$ is inclusion map of $\mathfrak{so}(n)$ into $\mathfrak{gl}(n, \mathbb{R})$ (and $\widetilde{\tau}$ is the inclusion of $\mathbf{SO}(n)$ into $\mathrm{Aut}(\mathbb{R}^n)$).

As a special case of Proposition 18.28, when our semidirect product is just a direct product $H \times K$ (τ is the trivial homomorphism mapping every $k \in K$ to id), we see that the Lie algebra of $H \times K$ is the Lie algebra direct sum $\mathfrak{h} \oplus \mathfrak{k}$ (where the bracket between elements of $\mathfrak{h}$ and elements of $\mathfrak{k}$ is 0).

18.6 Universal Covering Groups ⊛

Every connected Lie group G is a manifold, and as such, from results in Section 10.2, it has a universal cover $\pi\colon \widetilde{G} \to G$, where $\widetilde{G}$ is simply connected. It is possible to make $\widetilde{G}$ into a group so that $\widetilde{G}$ is a Lie group and π is a Lie group homomorphism. We content ourselves with a sketch of the construction whose details can be found in Warner [114], Chapter 3.

Consider the map $\alpha\colon \widetilde{G} \times \widetilde{G} \to G$, given by

$$\alpha(\widetilde{a}, \widetilde{b}) = \pi(\widetilde{a})\pi(\widetilde{b})^{-1},$$

for all $\widetilde{a}, \widetilde{b} \in \widetilde{G}$, and pick some $\widetilde{e} \in \pi^{-1}(1)$. Since $\widetilde{G} \times \widetilde{G}$ is simply connected, it follows by Proposition 10.13 that there is a unique map $\widetilde{\alpha}\colon \widetilde{G} \times \widetilde{G} \to \widetilde{G}$ such that

$$\alpha = \pi \circ \widetilde{\alpha} \quad \text{and} \quad \widetilde{e} = \widetilde{\alpha}(\widetilde{e}, \widetilde{e}),$$

as illustrated below:

$$\begin{tikzcd}
 & \widetilde{G} \ni \widetilde{e} \arrow[d, "\pi"] \\
\widetilde{G} \times \widetilde{G} \arrow[ur, dashed, "\widetilde{\alpha}"] \arrow[r, "\alpha"'] & G \ni 1.
\end{tikzcd}$$

For all $\widetilde{a}, \widetilde{b} \in \widetilde{G}$, define

$$\widetilde{b}^{-1} = \widetilde{\alpha}(\widetilde{e}, \widetilde{b}), \qquad \widetilde{a}\widetilde{b} = \widetilde{\alpha}(\widetilde{a}, \widetilde{b}^{-1}). \tag{$*$}$$

Using Proposition 10.13, it can be shown that the above operations make $\widetilde{G}$ into a group, and as $\widetilde{\alpha}$ is smooth, into a Lie group. Moreover, π becomes a Lie group homomorphism. We summarize these facts as

Theorem 18.29. *Every connected Lie group has a simply connected covering map $\pi\colon \widetilde{G} \to G$, where $\widetilde{G}$ is a Lie group and π is a Lie group homomorphism.*

The group $\widetilde{G}$ is called the *universal covering group* of G. Consider $D = \ker \pi$. Since the fibres of π are countable, the group D is a countable closed normal subgroup of $\widetilde{G}$; that is, a discrete normal subgroup of $\widetilde{G}$. It follows that $G \cong \widetilde{G}/D$, where $\widetilde{G}$ is a simply connected Lie group and D is a discrete normal subgroup of $\widetilde{G}$.

We conclude this section by stating the following useful proposition whose proof can be found in Warner [114] (Chapter 3, Proposition 3.26).

Proposition 18.30. *Let $\phi\colon G \to H$ be a homomorphism of connected Lie groups. Then ϕ is a covering map iff $d\phi_1\colon \mathfrak{g} \to \mathfrak{h}$ is an isomorphism of Lie algebras.*

For example, we know that $\mathfrak{su}(2) = \mathfrak{so}(3)$, so the homomorphism from $\mathbf{SU}(2)$ to $\mathbf{SO}(3)$ provided by the representation of 3D rotations by the quaternions is a covering map.

18.7 The Lie Algebra of Killing Fields ⊛

In Section 17.4 we defined Killing vector fields. Recall that a Killing vector field X on a manifold M satisfies the condition

$$L_X g(Y, Z) = X(\langle Y, Z\rangle) - \langle [X, Y], Z\rangle - \langle Y, [X, Z]\rangle = 0,$$

for all $X, Y, Z \in \mathfrak{X}(M)$. By Proposition 17.9, X is a Killing vector field iff the diffeomorphisms Φ_t induced by the flow Φ of X are isometries (on their domain).

The isometries of a Riemannian manifold (M, g) form a group $\mathrm{Isom}(M, g)$, called the *isometry group of* (M, g). An important theorem of Myers and Steenrod asserts that the isometry group $\mathrm{Isom}(M, g)$ is a Lie group. It turns out that the Lie algebra $\mathfrak{i}(M)$ of the group $\mathrm{Isom}(M, g)$ is closely related to a certain Lie subalgebra of the Lie algebra of Killing fields. In this section we briefly explore this relationship.

We begin by observing that the Killing fields form a Lie algebra.

Proposition 18.31. *The Killing fields on a smooth manifold M form a Lie subalgebra $\mathcal{K}i(M)$ of the Lie algebra $\mathfrak{X}(M)$ of vector fields on M.*

Proof. The Lie derivative L_X is $\mathbb{R}$-linear in X, and since

$$L_X \circ L_Y - L_Y \circ L_X = [L_X, L_Y] = L_{[X,Y]},$$

if X and Y are Killing fields, then $L_X g = L_Y g = 0$, and we get

$$L_{[X,Y]} g = (L_X \circ L_Y - L_Y \circ L_X) g = (L_X \circ L_Y) g - (L_Y \circ L_X) g = 0,$$

proving that $[X, Y]$ is a Killing vector field. □

However, unlike $\mathfrak{X}(M)$, the Lie algebra $\mathcal{K}i(M)$ is finite-dimensional. In fact, the Lie subalgebra $c\mathcal{K}i(M)$ of complete Killing vector fields is anti-isomorphic to the Lie algebra $\mathfrak{i}(M)$ of the Lie group $\mathrm{Isom}(M)$ of isometries of M (complete vector fields are defined in Definition 9.12). The following result is proved in O'Neill [91] (Chapter 9, Lemma 28) and Sakai [100] (Chapter III, Lemma 6.4 and Proposition 6.5).

Proposition 18.32. *Let (M, g) be a connected Riemannian manifold of dimension n (equipped with the Levi–Civita connection on M induced by g). The Lie algebra $\mathcal{K}i(M)$ of Killing vector fields on M has dimension at most $n(n + 1)/2$.*

We also have the following result proved in O'Neill [91] (Chapter 9, Proposition 30) and Sakai [100] (Chapter III, Corollary 6.3).

Proposition 18.33. *Let (M, g) be a Riemannian manifold of dimension n (equipped with the Levi–Civita connection on M induced by g). If M is complete, then every Killing vector field on M is complete.*

The relationship between the Lie algebra $\mathfrak{i}(M)$ and Killing vector fields is obtained as follows. For every element X in the Lie algebra $\mathfrak{i}(M)$ of $\mathrm{Isom}(M)$ (viewed as a left-invariant vector field), define the vector field X^+ on M by

$$X^+(p) = \frac{d}{dt}(\varphi_t(p))\bigg|_{t=0}, \quad p \in M,$$

where $t \mapsto \varphi_t = \exp(tX)$ is the one-parameter group associated with X. Because φ_t is an isometry of M, the vector field X^+ is a Killing vector field, and it is also easy to show that (φ_t) is the one-parameter group of X^+. Since φ_t is defined for all t, the vector field X^+ is complete. The following result is shown in O'Neill [91] (Chapter 9, Proposition 33).

Theorem 18.34. *Let (M, g) be a Riemannian manifold (equipped with the Levi–Civita connection on M induced by g). The following properties hold:*

(1) The set $c\mathcal{K}i(M)$ of complete Killing vector fields on M is a Lie subalgebra of the Lie algebra $\mathcal{K}i(M)$ of Killing vector fields.

(2) The map $X \mapsto X^+$ is a Lie anti-isomorphism between $\mathfrak{i}(M)$ and $c\mathcal{K}i(M)$, which means that

$$[X^+, Y^+] = -[X, Y]^+, \quad X, Y \in \mathfrak{i}(M).$$

For more on Killing vector fields, see Sakai [100] (Chapter III, Section 6). In particular, complete Riemannian manifolds for which $\mathfrak{i}(M)$ has the maximum dimension $n(n+1)/2$ are characterized.

18.8 Problems

Problem 18.1. Prove that in a Lie group, the smoothness of inversion follows from the smoothness of multiplication.

Hint. Apply the inverse function theorem to the map $(g, h) \mapsto (g, gh)$, from $G \times G$ to $G \times G$.

Problem 18.2. Prove the following two facts:

1. The derivative of multiplication in a Lie group $\mu \colon G \times G \to G$ is given by

$$d\mu_{a,b}(u, v) = (dR_b)_a(u) + (dL_a)_b(v),$$

for all $u \in T_aG$ and all $v \in T_bG$. At $(1, 1)$, the above yields

$$d\mu_{1,1}(u, v) = u + v.$$

2. The derivative of the inverse map $\iota \colon G \to G$ is given by

$$d\iota_a(u) = -(dR_{a^{-1}})_1 \circ (dL_{a^{-1}})_a(u) = -(dL_{a^{-1}})_1 \circ (dR_{a^{-1}})_a(u)$$

for all $u \in T_aG$. At 1, we get

$$d\iota_1(u) = -u.$$

Problem 18.3. Prove Proposition 18.11.

Hint. See Warner [114] (Chapter 3, Proposition 3.47), Bröcker and tom Dieck [24] (Chapter 1, Section 2, formula 2.11), or Marsden and Ratiu [77] (Chapter 9, Proposition 9.1.5).

Problem 18.4. Prove Proposition 18.15.

Hint. See Warner [114] (Chapter 3).

Problem 18.5. Prove Statements (1) through (4) of Proposition 18.26.

Problem 18.6. All Lie algebras in this problem are finite-dimensional. Let $\mathfrak{g}$ be a Lie algebra (over $\mathbb{R}$ or $\mathbb{C}$). Given two subsets $\mathfrak{a}$ and $\mathfrak{b}$ of $\mathfrak{g}$, we let $[\mathfrak{a}, \mathfrak{b}]$ be the subspace of $\mathfrak{g}$ consisting of all linear combinations of elements of the form $[a, b]$ with $a \in \mathfrak{a}$ and $b \in \mathfrak{b}$.

(1) Check that if $\mathfrak{a}$ and $\mathfrak{b}$ are ideals, then $[\mathfrak{a}, \mathfrak{b}]$ is an ideal.

(2) The *lower central series* $(C^k\mathfrak{g})$ of $\mathfrak{g}$ is defined as follows:

$$\begin{aligned} C^0\mathfrak{g} &= \mathfrak{g} \\ C^{k+1}\mathfrak{g} &= [\mathfrak{g}, C^k\mathfrak{g}], \quad k \geq 0. \end{aligned}$$

We have a decreasing sequence

$$\mathfrak{g} = C^0\mathfrak{g} \supseteq C^1\mathfrak{g} \supseteq C^2\mathfrak{g} \supseteq \cdots .$$

We say that $\mathfrak{g}$ is *nilpotent* iff $C^k\mathfrak{g} = (0)$ for some $k \geq 1$.

Prove that the following statements are equivalent:

1. The algebra $\mathfrak{g}$ is nilpotent.
2. There is some $n \geq 1$ such that

   $$[x_1, [x_2, [x_3, \cdots , [x_n, x_{n+1}] \cdots]]] = 0$$

   for all $x_1, \ldots, x_{n+1} \in \mathfrak{g}$.
3. There is a chain of ideals

   $$\mathfrak{g} = \mathfrak{a}_0 \supseteq \mathfrak{a}_1 \supseteq \cdots \supseteq \mathfrak{a}_n = (0)$$

   such that $[\mathfrak{g}, \mathfrak{a}_i] \subseteq \mathfrak{a}_{i+1}$ for $i = 0, \ldots, n-1$ $(n \geq 1)$.

(3) Given a vector space E of dimension n, a *flag* in E is a sequence $F = (V_i)$ of subspaces of E such that

$$(0) = V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots \subseteq V_n = E,$$

such that $\dim(V_i) = i$. Define $\mathfrak{n}(F)$ by

$$\mathfrak{n}(F) = \{f \in \mathrm{End}(E) \mid f(V_i) \subseteq V_{i-1},\ i = 1, \ldots, n\}.$$

If we pick a basis $(e_1, \ldots, e_n)$ of E such that $e_i \in V_i$, then check that every $f \in \mathfrak{n}(F)$ is represented by a strictly upper triangular matrix (the diagonal entries are 0).

Prove that $\mathfrak{n}(F)$ is a Lie subalgebra of $\mathrm{End}(E)$ and that it is nilpotent.

If $\mathfrak{g}$ is a nilpotent Lie algebra, then prove that ad_x is nilpotent for every $x \in \mathfrak{g}$.

(4) The *derived series* (or *commutator series*) $(D^k\mathfrak{g})$ of $\mathfrak{g}$ is defined as follows:

$$\begin{aligned} D^0\mathfrak{g} &= \mathfrak{g} \\ D^{k+1}\mathfrak{g} &= [D^k\mathfrak{g}, D^k\mathfrak{g}], \quad k \geq 0. \end{aligned}$$

We have a decreasing sequence

$$\mathfrak{g} = D^0\mathfrak{g} \supseteq D^1\mathfrak{g} \supseteq D^2\mathfrak{g} \supseteq \cdots .$$

We say that $\mathfrak{g}$ is *solvable* iff $D^k\mathfrak{g} = (0)$ for some $k \geq 1$.

Recall that a Lie algebra $\mathfrak{g}$ is *abelian* if $[X, Y] = 0$ for all $X, Y \in \mathfrak{g}$. Check that If $\mathfrak{g}$ is abelian, then $\mathfrak{g}$ is solvable.

Prove that a nonzero solvable Lie algebra has a nonzero abelian ideal.

Prove that the following statements are equivalent:

1. The algebra $\mathfrak{g}$ is solvable.
2. There is a chain of ideals
$$\mathfrak{g} = \mathfrak{a}_0 \supseteq \mathfrak{a}_1 \supseteq \cdots \supseteq \mathfrak{a}_n = (0)$$
such that $[\mathfrak{a}_i, \mathfrak{a}_i] \subseteq \mathfrak{a}_{i+1}$ for $i = 0, \ldots, n-1$ $(n \geq 1)$.

Given any flag $F = (V_i)$ in E (where E is a vector space of dimension n), define $\mathfrak{b}(F)$ by

$$\mathfrak{b}(F) = \{f \in \mathrm{End}(E) \mid f(V_i) \subseteq V_i,\ i = 0, \ldots, n\}.$$

If we pick a basis $(e_1, \ldots, e_n)$ of E such that $e_i \in V_i$, then check that every $f \in \mathfrak{b}(F)$ is represented by an upper triangular matrix.

Prove that $\mathfrak{b}(F)$ is a Lie subalgebra of $\mathrm{End}(E)$ and that it is solvable (observe that $D^1(\mathfrak{b}(F)) \subseteq \mathfrak{n}(F)$).

(5) Prove that

$$D^k\mathfrak{g} \subseteq C^k\mathfrak{g} \quad k \geq 0.$$

Deduce that every nilpotent Lie algebra is solvable.

(6) If $\mathfrak{g}$ is a solvable Lie algebra, then prove that every Lie subalgebra of $\mathfrak{g}$ is solvable, and for every ideal $\mathfrak{a}$ of $\mathfrak{g}$, the quotient Lie algebra $\mathfrak{g}/\mathfrak{a}$ is solvable.

Given a Lie algebra $\mathfrak{g}$, if $\mathfrak{a}$ is a solvable ideal and if $\mathfrak{g}/\mathfrak{a}$ is also solvable, then $\mathfrak{g}$ is solvable.

Given any two ideals $\mathfrak{a}$ and $\mathfrak{b}$ of a Lie algebra $\mathfrak{g}$, prove that $(\mathfrak{a} + \mathfrak{b})/\mathfrak{a}$ and $\mathfrak{b}/(\mathfrak{a} \cap \mathfrak{b})$ are isomorphic Lie algebras.

Given any two solvable ideals $\mathfrak{a}$ and $\mathfrak{b}$ of a Lie algebra $\mathfrak{g}$, prove that $\mathfrak{a} + \mathfrak{b}$ is solvable. Conclude from this that there is a largest solvable ideal $\mathfrak{r}$ in $\mathfrak{g}$ (called the *radical* of $\mathfrak{g}$).

Problem 18.7. Refer to the notion of an extension

$$\begin{tikzcd} \mathfrak{a} \arrow[r, "\varphi"] & \mathfrak{g} \arrow[r, "\psi"] & \mathfrak{b} \end{tikzcd}$$

with φ injective, ψ surjective, and with $\mathrm{Im}\,\varphi = \mathrm{Ker}\,\psi = \mathfrak{n}$, given just after Definition 18.18. Prove that every inessential extension of $\mathfrak{b}$ by $\mathfrak{a}$ is a semidirect product of $\mathfrak{b}$ by $\mathfrak{a}$.

Problem 18.8. Prove that if an extension of K by H has a section $\gamma\colon K \to G$, then G is isomorphic to a semidirect product of K by H with respect to a certain homomorphism τ that you need to find.

Problem 18.9. Fill in the details of the computations in the proof of Proposition 18.28.

Problem 18.10. We know that the Lie algebra $\mathfrak{se}(3)$ of $\mathbf{SE}(3)$ consists of all 4×4 matrices of the form

$$\begin{pmatrix} B & u \\ 0 & 0 \end{pmatrix},$$

where $B \in \mathfrak{so}(3)$ is a skew symmetric matrix and $u \in \mathbb{R}^3$. The following 6 matrices form a basis of $\mathfrak{se}(3)$:

$$X_1 = \begin{pmatrix} E_1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad X_2 = \begin{pmatrix} E_2 & 0 \\ 0 & 0 \end{pmatrix}, \qquad X_3 = \begin{pmatrix} E_3 & 0 \\ 0 & 0 \end{pmatrix},$$
$$X_4 = \begin{pmatrix} 0 & e_1^3 \\ 0 & 0 \end{pmatrix}, \qquad X_5 = \begin{pmatrix} 0 & e_2^3 \\ 0 & 0 \end{pmatrix}, \qquad X_6 = \begin{pmatrix} 0 & e_3^3 \\ 0 & 0 \end{pmatrix},$$

with

$$E_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \qquad E_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$
$$e_1^3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad e_2^3 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad e_3^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Also recall the isomorphism between $(\mathbb{R}^3, \times)$ and $\mathfrak{so}(3)$ given by

$$u = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \mapsto u_\times = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix}.$$

We define the bijection $\psi\colon \mathbb{R}^6 \to \mathfrak{se}(3)$ by

$$\psi(e_i^6) = X_i, \quad i = 1, \ldots, 6,$$

where $(e_1^6, \ldots, e_6^6)$ is the canonical basis of $\mathbb{R}^6$. If we split a vector in $\mathbb{R}^6$ as two vectors $\omega, u \in \mathbb{R}^3$ and write

$$\begin{pmatrix} \omega \\ u \end{pmatrix}$$

for such a vector in $\mathbb{R}^6$, then ψ is given by

$$\psi \begin{pmatrix} \omega \\ u \end{pmatrix} = \begin{pmatrix} \omega_\times & u \\ 0 & 0 \end{pmatrix}.$$

We define a bracket structure on $\mathbb{R}^6$ by

$$\left[\begin{pmatrix} \omega \\ u \end{pmatrix}, \begin{pmatrix} \theta \\ v \end{pmatrix} \right] = \begin{pmatrix} \omega \times \theta \\ u \times \theta + \omega \times v \end{pmatrix}.$$

(1) Check that $\psi \colon (\mathbb{R}^6, [-,-]) \to \mathfrak{se}(3)$ is a Lie algebra isomorphism.

Hint. Prove that

$$[\omega_\times, \theta_\times] = \omega_\times \theta_\times - \theta_\times \omega_\times = (\omega \times \theta)_\times.$$

(2) For any

$$X = \begin{pmatrix} B & u \\ 0 & 0 \end{pmatrix} \in \mathfrak{se}(3)$$

and any

$$\begin{pmatrix} \theta \\ v \end{pmatrix} \in \mathbb{R}^6,$$

prove that

$$\psi^{-1} \circ \mathrm{ad}(X) \circ \psi \begin{pmatrix} \theta \\ v \end{pmatrix} = \begin{pmatrix} B & 0 \\ u_\times & B \end{pmatrix} \begin{pmatrix} \theta \\ v \end{pmatrix}.$$

(3) For any

$$g = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \in \mathbf{SE}(3),$$

where $R \in \mathbf{SO}(3)$ and $t \in \mathbb{R}^3$ and for any

$$\begin{pmatrix} \theta \\ v \end{pmatrix} \in \mathbb{R}^6,$$

prove that

$$\psi^{-1} \mathrm{Ad}(g) \circ \psi \begin{pmatrix} \theta \\ v \end{pmatrix} = \begin{pmatrix} R & 0 \\ t_\times R & R \end{pmatrix} \begin{pmatrix} \theta \\ v \end{pmatrix}.$$

Problem 18.11. We can let the group $\mathbf{SO}(3)$ act on itself by conjugation, so that

$$R \cdot S = RSR^{-1} = RSR^{\top}.$$

The orbits of this action are the *conjugacy classes* of $\mathbf{SO}(3)$.

(1) Prove that the conjugacy classes of $\mathbf{SO}(3)$ are in bijection with the following sets:

1. $\mathcal{C}_0 = \{(0,0,0)\}$, the sphere of radius 0.
2. $\mathcal{C}_\theta$, with $0 < \theta < \pi$ and
$$\mathcal{C}_\theta = \{u \in \mathbb{R}^3 \mid \|u\| = \theta\},$$
the sphere of radius θ.
3. $\mathcal{C}_\pi = \mathbb{RP}^2$, viewed as the quotient of the sphere of radius π by the equivalence relation of being antipodal.

(2) Give $\mathrm{M}_3(\mathbb{R})$ the Euclidean structure where

$$\langle A, B\rangle = \frac{1}{2}\mathrm{tr}(A^{\top}B).$$

Consider the following three curves in $\mathbf{SO}(3)$:

$$c(t) = \begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

for $0 \leq t \leq 2\pi$,

$$\alpha(\theta) = \begin{pmatrix} -\cos 2\theta & 0 & \sin 2\theta \\ 0 & -1 & 0 \\ \sin 2\theta & 0 & \cos 2\theta \end{pmatrix},$$

for $-\pi/2 \leq \theta \leq \pi/2$, and

$$\beta(\theta) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\cos 2\theta & \sin 2\theta \\ 0 & \sin 2\theta & \cos 2\theta \end{pmatrix},$$

for $-\pi/2 \leq \theta \leq \pi/2$.

Check that $c(t)$ is a rotation of angle t and axis $(0,0,1)$, that $\alpha(\theta)$ is a rotation of angle π whose axis is in the (x,z)-plane, and that $\beta(\theta)$ is a rotation of angle π whose axis is in the (y,z)-plane. Show that a log of $\alpha(\theta)$ is

$$B_\alpha = \pi \begin{pmatrix} 0 & -\cos\theta & 0 \\ \cos\theta & 0 & -\sin\theta \\ 0 & \sin\theta & 0 \end{pmatrix},$$

and that a log of $\beta(\theta)$ is

$$B_\beta = \pi \begin{pmatrix} 0 & -\cos\theta & \sin\theta \\ \cos\theta & 0 & 0 \\ -\sin\theta & 0 & 0 \end{pmatrix}.$$

(3) The curve $c(t)$ is a closed curve starting and ending at I that intersects $\mathcal{C}_\pi$ for $t = \pi$, and α, β are contained in $\mathcal{C}_\pi$ and coincide with $c(\pi)$ for $\theta = 0$. Compute the derivative $c'(\pi)$ of $c(t)$ at $t = \pi$, and the derivatives $\alpha'(0)$ and $\beta'(0)$, and prove that they are pairwise orthogonal (under the inner product $\langle -, - \rangle$).

Conclude that $c(t)$ intersects $\mathcal{C}_\pi$ transversally in $\mathbf{SO}(3)$, which means that

$$T_{c(\pi)}\, c + T_{c(\pi)}\, \mathcal{C}_\pi = T_{c(\pi)}\, \mathbf{SO}(3).$$

This fact can be used to prove that all closed curves smoothly homotopic to $c(t)$ must intersect $\mathcal{C}_\pi$ transversally, and consequently $c(t)$ is not (smoothly) homotopic to a point. This implies that $\mathbf{SO}(3)$ is not simply connected,

Problem 18.12. Let G be a Lie group with Lie algebra $\mathfrak{g}$. Prove that if $[X, Y] = 0$, then $\exp(X)\exp(Y) = \exp(Y)\exp(X)$. If G is connected, prove that $[X, Y] = 0$ for all $X, Y \in \mathfrak{g}$ iff G is abelian.

Hint. For help, see Duistermaat and Kolk [43] (Chapter 1, Section 1.9).

Problem 18.13. Let G be a connected Lie group with Lie algebra $\mathfrak{g}$, and assume that G is abelian.

(1) Prove that the exponential map $\exp\colon \mathfrak{g} \to G$ is surjective.

(2) If $\Gamma = \mathrm{Ker}\, \exp$, then show that Γ is a discrete closed subgroup of $\mathfrak{g}$, and that exp induces an Lie group isomorphism between $\mathfrak{g}/\Gamma$ and G.

Hint. For help, see Duistermaat and Kolk [43] (Chapter 1, Section 1.12).

Problem 18.14. It is a standard result of algebra that every nontrivial discrete subgroup Γ of a finite-dimensional vector space V of dimension n is of the form

$$\Gamma = \{n_1 e_1 + \cdots + n_k e_k \mid n_i \in \mathbb{Z},\ 1 \leq i \leq k\},$$

where $e_1, \ldots, e_k \in V$ are linearly independent vectors. See Duistermaat and Kolk [43] (Chapter 1, Theorem 1.12.3). Such a group is called an *integral lattice*.

Use the above result to prove that every connected abelian Lie group G is isomorphic (as a Lie group) to the additive group

$$(\mathbb{R}/\mathbb{Z})^k \times \mathbb{R}^{n-k},$$

where $n = \dim(\mathfrak{g})$. Deduce that every compact connected abelian Lie group G is isomorphic to the torus $(\mathbb{R}/\mathbb{Z})^n$, and that G is isomorphic to $\mathbb{R}^n$ iff $\mathrm{Ker}\, \exp = (0)$.

Problem 18.15. (1) Check that the set of Killing vector field on a Riemannian manifold M is a Lie algebra denoted $\mathcal{K}i(M)$.

(2) Prove that if M is and connected and has dimension n, then $\mathcal{K}i(M)$ has dimension at most $n(n+1)/2$.

Problem 18.16. The "right way" (meaning convenient and rigorous) to define the *unit quaternions* is to define them as the elements of the unitary group $\mathbf{SU}(2)$, namely the group of 2×2 complex matrices of the form

$$\begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \qquad \alpha, \beta \in \mathbb{C},\ \alpha\overline{\alpha} + \beta\overline{\beta} = 1.$$

Then, the *quaternions* are the elements of the real vector space $\mathbb{H} = \mathbb{R}\,\mathbf{SU}(2)$. Let $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ be the matrices

$$\mathbf{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{i} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \mathbf{j} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \mathbf{k} = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

then $\mathbb{H}$ is the set of all matrices of the form

$$X = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}, \quad a, b, c, d \in \mathbb{R}.$$

Indeed, every matrix in $\mathbb{H}$ is of the form

$$X = \begin{pmatrix} a + ib & c + id \\ -(c - id) & a - ib \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}.$$

(1) Prove that the quaternions $\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k}$ satisfy the famous identities discovered by Hamilton:

$$\begin{aligned} \mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{i}\mathbf{j}\mathbf{k} &= -\mathbf{1}, \\ \mathbf{i}\mathbf{j} = -\mathbf{j}\mathbf{i} &= \mathbf{k}, \\ \mathbf{j}\mathbf{k} = -\mathbf{k}\mathbf{j} &= \mathbf{i}, \\ \mathbf{k}\mathbf{i} = -\mathbf{i}\mathbf{k} &= \mathbf{j}. \end{aligned}$$

Prove that $\mathbb{H}$ is a skew field (a noncommutative field) called the *quaternions*, and a real vector space of dimension 4 with basis $(\mathbf{1}, \mathbf{i}, \mathbf{j}, \mathbf{k})$; thus as a vector space, $\mathbb{H}$ is isomorphic to $\mathbb{R}^4$.

A concise notation for the quaternion X defined by $\alpha = a + ib$ and $\beta = c + id$ is

$$X = [a, (b, c, d)].$$

We call a the *scalar part* of X and (b, c, d) the *vector part* of X. With this notation, $X^* = [a, -(b, c, d)]$, which is often denoted by $\overline{X}$. The quaternion $\overline{X}$ is called the *conjugate* of q. If q is a unit quaternion, then $\overline{q}$ is the multiplicative inverse of q. A *pure quaternion* is a quaternion whose scalar part is equal to zero.

(2) Given a unit quaternion

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \in \mathbf{SU}(2),$$

the usual way to define the rotation ρ_q (of $\mathbb{R}^3$) induced by q is to embed $\mathbb{R}^3$ into $\mathbb{H}$ as the pure quaternions, by

$$\psi(x, y, z) = \begin{pmatrix} ix & y + iz \\ -y + iz & -ix \end{pmatrix}, \quad (x, y, z) \in \mathbb{R}^3.$$

Observe that the above matrix is skew-Hermitian ($\psi(x, y, z)^* = -\psi(x, y, z)$). But, the space of skew-Hermitian matrices is the Lie algebra $\mathfrak{su}(2)$ of $\mathbf{SU}(2)$, so $\psi(x, y, z) \in \mathfrak{su}(2)$. Then, q defines the map ρ_q (on $\mathbb{R}^3$) given by

$$\rho_q(x, y, z) = \psi^{-1}(q\psi(x, y, z)q^*),$$

where q^* is the inverse of q (since $\mathbf{SU}(2)$ is a unitary group) and is given by

$$q^* = \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

Actually, the *adjoint representation* of the group $\mathbf{SU}(2)$ is the group homomorphism $\mathrm{Ad}\colon \mathbf{SU}(2) \to \mathbf{GL}(\mathfrak{su}(2))$ defined such that for every $q \in \mathbf{SU}(2)$,

$$\mathrm{Ad}_q(A) = qAq^*, \quad A \in \mathfrak{su}(2).$$

Therefore, modulo the isomorphism ψ, the linear map ρ_q is the linear isomorphism Ad_q. In fact, ρ_q is a rotation (and so is Ad_q), which you will prove shortly.

Since the matrix $\psi(x, y, z)$ is skew-Hermitian, the matrix $-i\psi(x, y, z)$ is Hermitian, and we have

$$-i\psi(x, y, z) = \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} = x\sigma_3 + y\sigma_2 + z\sigma_1,$$

where $\sigma_1, \sigma_2, \sigma_3$ are the *Pauli spin matrices*

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Check that $\mathbf{i} = i\sigma_3$, $\mathbf{j} = i\sigma_2$, $\mathbf{k} = i\sigma_1$. Prove that matrices of the form $x\sigma_3 + y\sigma_2 + z\sigma_1$ (with $x, y, x \in \mathbb{R}$) are exactly the 2×2 Hermitian matrix with zero trace.

(3) Prove that for every $q \in \mathbf{SU}(2)$, if A is any 2×2 Hermitian matrix with zero trace as above, then qAq^* is also a Hermitian matrix with zero trace.

Prove that

$$\det(x\sigma_3 + y\sigma_2 + z\sigma_1) = \det(qAq^*) = -(x^2 + y^2 + z^2).$$

We can embed $\mathbb{R}^3$ into the space of Hermitian matrices with zero trace by

$$\varphi(x, y, z) = x\sigma_3 + y\sigma_2 + z\sigma_1.$$

Check that

$$\varphi = -i\psi$$

and

$$\varphi^{-1} = i\psi^{-1}.$$

Prove that every quaternion q induces a map r_q on $\mathbb{R}^3$ by

$$r_q(x, y, z) = \varphi^{-1}(q\varphi(x, y, z)q^*) = \varphi^{-1}(q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^*)$$

which is clearly linear, and an isometry. Thus, $r_q \in \mathbf{O}(3)$.

(4) Find the fixed points of r_q, where $q = (a, (b, c, d))$. If $(b, c, d) \neq (0, 0, 0)$, then show that the fixed points (x, y, z) of r_q are solutions of the equations

$$\begin{aligned} -dy + cz &= 0 \\ cx - by &= 0 \\ dx - bz &= 0. \end{aligned}$$

This linear system has the nontrivial solution (b, c, d) and the matrix of this system is

$$\begin{pmatrix} 0 & -d & c \\ c & -b & 0 \\ d & 0 & -b \end{pmatrix}.$$

Prove that the above matrix has rank 2, so the fixed points of r_q form the one-dimensional space spanned by (b, c, d). Deduce from this that r_q must be a rotation.

Prove that $r\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$ given by $r(q) = r_q$ is a group homomorphism whose kernel is $\{I, -I\}$.

(5) Find the matrix R_q representing r_q explicitly by computing

$$q(x\sigma_3 + y\sigma_2 + z\sigma_1)q^* = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix} \begin{pmatrix} x & z - iy \\ z + iy & -x \end{pmatrix} \begin{pmatrix} \overline{\alpha} & -\beta \\ \overline{\beta} & \alpha \end{pmatrix}.$$

You should find

$$R_q = \begin{pmatrix} a^2 + b^2 - c^2 - d^2 & 2bc - 2ad & 2ac + 2bd \\ 2bc + 2ad & a^2 - b^2 + c^2 - d^2 & -2ab + 2cd \\ -2ac + 2bd & 2ab + 2cd & a^2 - b^2 - c^2 + d^2 \end{pmatrix}.$$

Since $a^2 + b^2 + c^2 + d^2 = 1$, this matrix can also be written as

$$R_q = \begin{pmatrix} 2a^2 + 2b^2 - 1 & 2bc - 2ad & 2ac + 2bd \\ 2bc + 2ad & 2a^2 + 2c^2 - 1 & -2ab + 2cd \\ -2ac + 2bd & 2ab + 2cd & 2a^2 + 2d^2 - 1 \end{pmatrix}.$$

Prove that $r_q = \rho_q$.

(6) To prove the surjectivity of r algorithmically, proceed as follows.

First, prove that $\mathrm{tr}(R_q) = 4a^2 - 1$, so

$$a^2 = \frac{\mathrm{tr}(R_q) + 1}{4}.$$

If $R \in \mathbf{SO}(3)$ is any rotation matrix and if we write

$$R = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33}, \end{pmatrix}$$

we are looking for a unit quaternion $q \in \mathbf{SU}(2)$ such that $r_q = R$. Therefore, we must have

$$a^2 = \frac{\mathrm{tr}(R) + 1}{4}.$$

We also know that

$$\mathrm{tr}(R) = 1 + 2\cos\theta,$$

where $\theta \in [0, \pi]$ is the angle of the rotation R. Deduce that

$$|a| = \cos\left(\frac{\theta}{2}\right) \quad (0 \leq \theta \leq \pi).$$

There are two cases.

Case 1. $\mathrm{tr}(R) \neq -1$, or equivalently $\theta \neq \pi$. In this case $a \neq 0$. Pick

$$a = \frac{\sqrt{\mathrm{tr}(R) + 1}}{2}.$$

Then, show that

$$b = \frac{r_{32} - r_{23}}{4a}, \quad c = \frac{r_{13} - r_{31}}{4a}, \quad d = \frac{r_{21} - r_{12}}{4a}.$$

Case 2. $\mathrm{tr}(R) = -1$, or equivalently $\theta = \pi$. In this case $a = 0$. Prove that

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31} \\ 4cd &= r_{32} + r_{23} \end{aligned}$$

and

$$\begin{aligned} b^2 &= \frac{1 + r_{11}}{2} \\ c^2 &= \frac{1 + r_{22}}{2} \\ d^2 &= \frac{1 + r_{33}}{2}. \end{aligned}$$

Since $q \neq 0$ and $a = 0$, at least one of b, c, d is nonzero.

If $b \neq 0$, let

$$b = \frac{\sqrt{1 + r_{11}}}{\sqrt{2}},$$

and determine c, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4bd &= r_{13} + r_{31}. \end{aligned}$$

If $c \neq 0$, let

$$c = \frac{\sqrt{1 + r_{22}}}{\sqrt{2}},$$

and determine b, d using

$$\begin{aligned} 4bc &= r_{21} + r_{12} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

If $d \neq 0$, let

$$d = \frac{\sqrt{1 + r_{33}}}{\sqrt{2}},$$

and determine b, c using

$$\begin{aligned} 4bd &= r_{13} + r_{31} \\ 4cd &= r_{32} + r_{23}. \end{aligned}$$

(7) Given any matrix $A \in \mathfrak{su}(2)$, with

$$A = \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix},$$

write $\theta = \sqrt{u_1^2 + u_2^2 + u_3^2}$ and prove that

$$e^A = \cos\theta I + \frac{\sin\theta}{\theta} A, \quad \theta \neq 0,$$

with $e^0 = I$. Therefore, e^A is a unit quaternion representing the rotation of angle 2θ and axis (u_1, u_2, u_3) (or I when $\theta = k\pi$, $k \in \mathbb{Z}$). The above formula shows that we may assume that $0 \leq \theta \leq \pi$.

An equivalent but often more convenient formula is obtained by assuming that $u = (u_1, u_2, u_3)$ is a unit vector, equivalently $\det(A) = -1$, in which case $A^2 = -I$, so we have

$$e^{\theta A} = \cos\theta I + \sin\theta A.$$

Using the quaternion notation, this read as

$$e^{\theta A} = [\cos\theta, \sin\theta\, u].$$

Prove that the logarithm $A \in \mathfrak{su}(2)$ of a unit quaternion

$$q = \begin{pmatrix} \alpha & \beta \\ -\overline{\beta} & \overline{\alpha} \end{pmatrix}$$

with $\alpha = a + bi$ and $\beta = c + id$ can be determined as follows:

If $q = I$ (*i.e.* $a = 1$) then $A = 0$. If $q = -I$ (*i.e.* $a = -1$), then

$$A = \pm\pi \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}.$$

Otherwise, $a \neq \pm 1$ and $(b, c, d) \neq (0, 0, 0)$, and we are seeking some $A = \theta B \in \mathfrak{su}(2)$ with $\det(B) = 1$ and $0 < \theta < \pi$, such that

$$q = e^{\theta B} = \cos\theta I + \sin\theta B.$$

Then,

$$\begin{aligned} \cos\theta &= a \qquad (0 < \theta < \pi) \\ (u_1, u_2, u_3) &= \frac{1}{\sin\theta}(b, c, d). \end{aligned}$$

Since $a^2+b^2+c^2+d^2 = 1$ and $a = \cos\theta$, the vector $(b, c, d)/\sin\theta$ is a unit vector. Furthermore if the quaternion q is of the form $q = [\cos\theta, \sin\theta u]$ where $u = (u_1, u_2, u_3)$ is a unit vector (with $0 < \theta < \pi$), then

$$A = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix}$$

is a logarithm of q.

Show that the exponential map $\exp\colon \mathfrak{su}(2) \to \mathbf{SU}(2)$ is surjective, and injective on the open ball

$$\{\theta B \in \mathfrak{su}(2) \mid \det(B) = 1, 0 \leq \theta < \pi\}.$$

(8) You are now going to derive a formula for interpolating between two quaternions. This formula is due to Ken Shoemake, once a Penn student and my TA! Since rotations in $\mathbf{SO}(3)$ can be defined by quaternions, this has applications to computer graphics, robotics, and computer vision.

First, we observe that multiplication of quaternions can be expressed in terms of the inner product and the cross-product in $\mathbb{R}^3$. Indeed, if $q_1 = [a, u_1]$ and $q_2 = [a_2, u_2]$, then check that

$$q_1 q_2 = [a_1, u_1][a_2, u_2] = [a_1 a_2 - u_1 \cdot u_2,\ a_1 u_2 + a_2 u_1 + u_1 \times u_2].$$

We will also need the identity

$$u \times (u \times v) = (u \cdot v)u - (u \cdot u)v.$$

Given a quaternion q expressed as $q = [\cos\theta, \sin\theta\, u]$, where u is a unit vector, we can interpolate between I and q by finding the logs of I and q, interpolating in $\mathfrak{su}(2)$, and then exponentiating. We have

$$A = \log(I) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \log(q) = \theta \begin{pmatrix} iu_1 & u_2 + iu_3 \\ -u_2 + iu_3 & -iu_1 \end{pmatrix}.$$

Since $\mathbf{SU}(2)$ is a compact Lie group and since the inner product on $\mathfrak{su}(2)$ given by

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$$

is $\mathrm{Ad}(\mathbf{SU}(2))$-invariant, it induces a biinvariant Riemannian metric on $\mathbf{SU}(2)$, and the curve

$$\lambda \mapsto e^{\lambda B}, \quad \lambda \in [0, 1]$$

is a geodesic from I to q in $\mathbf{SU}(2)$. We write $q^\lambda = e^{\lambda B}$. Given two quaternions q_1 and q_2, because the metric is left invariant, the curve

$$\lambda \mapsto Z(\lambda) = q_1(q_1^{-1} q_2)^\lambda, \quad \lambda \in [0, 1]$$

is a geodesic from q_1 to q_2. Remarkably, there is a closed-form formula for the interpolant $Z(\lambda)$. Say $q_1 = [\cos\theta, \sin\theta\, u]$ and $q_2 = [\cos\varphi, \sin\varphi\, v]$, and assume that $q_1 \neq q_2$ and $q_1 \neq -q_2$.

Define Ω by

$$\cos\Omega = \cos\theta\cos\varphi + \sin\theta\sin\varphi(u \cdot v).$$

Since $q_1 \neq q_2$ and $q_1 \neq -q_2$, we have $0 < \Omega < \pi$. Prove that

$$Z(\lambda) = q_1(q_1^{-1} q_2)^\lambda = \frac{\sin(1-\lambda)\Omega}{\sin\Omega} q_1 + \frac{\sin\lambda\Omega}{\sin\Omega} q_2.$$

(9) We conclude by discussing the problem of a consistent choice of sign for the quaternion q representing a rotation $R = \rho_q \in \mathbf{SO}(3)$. We are looking for a "nice" section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$, that is, a function s satisfying the condition

$$\rho \circ s = \mathrm{id},$$

where ρ is the surjective homomorphism $\rho\colon \mathbf{SU}(2) \to \mathbf{SO}(3)$.

I claim that any section $s\colon \mathbf{SO}(3) \to \mathbf{SU}(2)$ of ρ is neither a homomorphism nor continuous. Intuitively, this means that there is no “nice and simple ” way to pick the sign of the quaternion representing a rotation.

To prove the above claims, let Γ be the subgroup of $\mathbf{SU}(2)$ consisting of all quaternions of the form $q = [a, (b, 0, 0)]$. Then, using the formula for the rotation matrix R_q corresponding to q (and the fact that $a^2 + b^2 = 1$), show that

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2a^2 - 1 & -2ab \\ 0 & 2ab & 2a^2 - 1 \end{pmatrix}.$$

Since $a^2 + b^2 = 1$, we may write $a = \cos\theta, b = \sin\theta$, and we see that

$$R_q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos 2\theta & -\sin 2\theta \\ 0 & \sin 2\theta & \cos 2\theta \end{pmatrix},$$

a rotation of angle 2θ around the x-axis. Thus, both Γ and its image are isomorphic to $\mathbf{SO}(2)$, which is also isomorphic to $\mathbf{U}(1) = \{w \in \mathbb{C} \mid |w| = 1\}$. By identifying $\mathbf{i}$ and i and identifying Γ and its image to $\mathbf{U}(1)$, if we write $w = \cos\theta + i\sin\theta \in \Gamma$, show that the restriction of the map ρ to Γ is given by $\rho(w) = w^2$.

Prove that any section s of ρ is not a homomorphism. (Consider the restriction of s to the image $\rho(\Gamma)$).

Prove that any section s of ρ is not continuous.

Chapter 19

The Derivative of exp and Dynkin's Formula ⊛

19.1 The Derivative of the Exponential Map

By Proposition 1.5, we know that if $[X, Y] = 0$, then $\exp(X+Y) = \exp(X)\exp(Y)$, whenever $X, Y \in \mathfrak{g}$, but this generally false if X and Y do not commute. For X and Y in a small enough open subset $U \subseteq \mathfrak{g}$ containing 0, we know by Proposition 18.6 that exp is a diffeomorphism from U to its image, so the function $\mu\colon U \times U \to U$ given by

$$\mu(X, Y) = \log(\exp(X)\exp(Y))$$

is well-defined and it turns out that for U small enough, it is analytic. Thus, it is natural to seek a formula for the Taylor expansion of μ near the origin.

This problem was investigated by Campbell (1897/98), Baker (1905) and in a more rigorous fashion by Hausdorff (1906). These authors gave recursive identities expressing the Taylor expansion of μ at the origin and the corresponding result is often referred to as the *Campbell-Baker-Hausdorff formula*. F. Schur (1891) and Poincaré (1899) also investigated the exponential map, in particular formulae for its derivative and the problem of expressing the function μ. However, it was Dynkin who finally gave an explicit formula (see Section 19.3) in 1947.

The proof that μ is analytic in a suitable domain can be proved using a formula for the derivative of the exponential map, a formula that was obtained by F. Schur and Poincaré. Thus, we begin by presenting such a formula.

First we introduce a convenient notation. If A is any real (or complex) $n \times n$ matrix, the following formula is clear:

$$\int_0^1 e^{tA} dt = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}.$$

If A is invertible, then the right-hand side can be written explicitly as

$$\sum_{k=0}^{\infty} \frac{A^k}{(k+1)!} = A^{-1}(e^A - I),$$

and we also write the latter as

$$\frac{e^A - I}{A} = \sum_{k=0}^{\infty} \frac{A^k}{(k+1)!}. \tag{$*$}$$

Even if A is not invertible, we use $(*)$ as the definition of $\frac{e^A - I}{A}$.

We can use the following trick to figure out what $(d\exp_X)(Y)$ is:

$$(d\exp_X)(Y) = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \exp(X + \epsilon Y) = \left.\frac{d}{d\epsilon}\right|_{\epsilon=0} d(R_{\exp(X+\epsilon Y)})_1,$$

since by Proposition 18.4, the map $s \mapsto R_{\exp s(X+\epsilon Y)}$ is the flow of the left-invariant vector field $(X + \epsilon Y)^L$ on G. Now, $(X + \epsilon Y)^L$ is an ϵ-dependent vector field which depends on ϵ in a C^1 fashion. From the theory of ODE's, if $p \mapsto v_\epsilon(p)$ is a smooth vector field depending in a C^1 fashion on a real parameter ϵ and if $\Phi_t^{v_\epsilon}$ denotes its flow (after time t), then we have the variational formula

$$\frac{\partial \Phi_t^{v_\epsilon}}{\partial \epsilon}(x) = \int_0^t d(\Phi_{t-s}^{v_\epsilon})_{\Phi_t^{v_\epsilon}(x)} \frac{\partial v_\epsilon}{\partial \epsilon}(\Phi_s^{v_\epsilon}(x)) ds.$$

See Duistermaat and Kolk [43], Appendix B, Formula (B.10). Using this, the following is proved in Duistermaat and Kolk [43] (Chapter 1, Section 1.5):

Proposition 19.1. *Given any Lie group G, for any $X \in \mathfrak{g}$, the linear map $d\exp_X \colon \mathfrak{g} \to T_{\exp(X)}G$ is given by*

$$\begin{aligned} d\exp_X &= d(R_{\exp(X)})_1 \circ \int_0^1 e^{s\,\mathrm{ad}\,X} ds \;=\; d(R_{\exp(X)})_1 \circ \frac{e^{\mathrm{ad}\,X} - I}{\mathrm{ad}\,X} \\ &= d(L_{\exp(X)})_1 \circ \int_0^1 e^{-s\,\mathrm{ad}\,X} ds \;=\; d(L_{\exp(X)})_1 \circ \frac{I - e^{-\mathrm{ad}\,X}}{\mathrm{ad}\,X}. \end{aligned}$$

Remark: If G is a matrix group of $n \times n$ matrices, we see immediately that the derivative of left multiplication ($X \mapsto L_A X = AX$) is given by

$$d(L_A)_X Y = AY,$$

for all $n \times n$ matrices X, Y. Consequently, for a matrix group, we get

$$d\exp_X = e^X \left(\frac{I - e^{-\mathrm{ad}\,X}}{\mathrm{ad}\,X} \right).$$

An alternative proof sketch of this fact is provided in Section 2.1.

Now, if A is an $n\times n$ matrix, the argument provided at the end of Section 2.1 is applicable, and hence it is clear that the (complex) eigenvalues of $\int_0^1 e^{sA}ds$ are of the form

$$\frac{e^{\lambda}-1}{\lambda} \quad (=1 \quad \text{if } \lambda = 0),$$

where λ ranges over the (complex) eigenvalues of A. Consequently, we get

Proposition 19.2. *The singular points of the exponential map* $\exp\colon \mathfrak{g} \to G$*, that is, the set of* $X \in \mathfrak{g}$ *such that* $d\exp_X$ *is singular (not invertible), are the* $X \in \mathfrak{g}$ *such that the linear map* $\mathrm{ad}\, X\colon \mathfrak{g} \to \mathfrak{g}$ *has an eigenvalue of the form* $k2\pi i$*, with* $k \in \mathbb{Z}$ *and* $k \neq 0$.

Another way to describe the *singular locus* Σ of the exponential map is to say that it is the disjoint union

$$\Sigma = \bigcup_{k\in\mathbb{Z}-\{0\}} k\Sigma_1,$$

where Σ_1 is the algebraic variety in $\mathfrak{g}$ given by

$$\Sigma_1 = \{X \in \mathfrak{g} \mid \det(\mathrm{ad}\, X - 2\pi i\, I) = 0\}.$$

For example, for $\mathrm{SL}(2,\mathbb{R})$,

$$\Sigma_1 = \left\{ \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2) \mid a^2 + bc = -\pi^2 \right\},$$

a two-sheeted hyperboloid mapped to $-I$ by exp.

Definition 19.1. Let $\mathfrak{g}_e = \mathfrak{g} - \Sigma$ be the set of $X \in \mathfrak{g}$ such that $\frac{e^{\mathrm{ad}\, X}-I}{\mathrm{ad}\, X}$ is invertible.

The set $\mathfrak{g}_e$ is an open subset of $\mathfrak{g}$ containing 0.

19.2 The Product in Logarithmic Coordinates

Since the map

$$X \mapsto \frac{e^{\mathrm{ad}\, X}-I}{\mathrm{ad}\, X}$$

is invertible for all $X \in \mathfrak{g}_e = \mathfrak{g} - \Sigma$, in view of the chain rule, the reciprocal (multiplicative inverse) of the above map

$$X \mapsto \frac{\mathrm{ad}\, X}{e^{\mathrm{ad}\, X}-I},$$

is an analytic function from $\mathfrak{g}_e$ to $\mathfrak{gl}(\mathfrak{g},\mathfrak{g})$.

Definition 19.2. Let $\mathfrak{g}_e^2$ be the subset of $\mathfrak{g} \times \mathfrak{g}_e$ consisting of all (X, Y) such that the solution $t \mapsto Z(t)$ of the differential equation

$$\frac{dZ(t)}{dt} = \frac{\mathrm{ad}\, Z(t)}{e^{\mathrm{ad}\, Z(t)} - I}(X)$$

with initial condition $Z(0) = Y (\in \mathfrak{g}_e)$ is defined for all $t \in [0, 1]$.

Set

$$\mu(X, Y) = Z(1), \quad (X, Y) \in \mathfrak{g}_e^2.$$

The following theorem is proved in Duistermaat and Kolk [43] (Chapter 1, Section 1.6, Theorem 1.6.1):

Theorem 19.3. *Given any Lie group G with Lie algebra $\mathfrak{g}$, the set $\mathfrak{g}_e^2$ is an open subset of $\mathfrak{g} \times \mathfrak{g}$ containing $(0, 0)$, and the map $\mu\colon \mathfrak{g}_e^2 \to \mathfrak{g}$ is real-analytic. Furthermore, we have*

$$\exp(X)\exp(Y) = \exp(\mu(X, Y)), \qquad (X, Y) \in \mathfrak{g}_e^2,$$

where $\exp\colon \mathfrak{g} \to G$. *If $\mathfrak{g}$ is a complex Lie algebra, then μ is complex-analytic.*

We may think of μ as the product in logarithmic coordinates. It is explained in Duistermaat and Kolk [43] (Chapter 1, Section 1.6) how Theorem 19.3 implies that a Lie group can be provided with the structure of a real-analytic Lie group. Rather than going into this, we will state a remarkable formula due to Dynkin expressing the Taylor expansion of μ at the origin.

19.3 Dynkin's Formula

As we said in Section 19.1, the problem of finding the Taylor expansion of μ near the origin was investigated by Campbell (1897/98), Baker (1905) and Hausdorff (1906). However, it was Dynkin who finally gave an explicit formula in 1947. There are actually slightly different versions of Dynkin's formula. One version is given (and proved convergent) in Duistermaat and Kolk [43] (Chapter 1, Section 1.7). Another slightly more explicit version (because it gives a formula for the homogeneous components of $\mu(X, Y)$) is given (and proved convergent) in Bourbaki [19] (Chapter II, §6, Section 4) and Serre [105] (Part I, Chapter IV, Section 8). We present the version in Bourbaki and Serre without proof. The proof uses formal power series and free Lie algebras.

Given $X, Y \in \mathfrak{g}_e^2$, we can write

$$\mu(X, Y) = \sum_{n=1}^{\infty} z_n(X, Y),$$

where $z_n(X, Y)$ is a homogeneous polynomial of degree n in the non-commuting variables X, Y.

Theorem 19.4. *(Dynkin's Formula) If we write $\mu(X,Y) = \sum_{n=1}^{\infty} z_n(X,Y)$, then we have*

$$z_n(X,Y) = \frac{1}{n} \sum_{p+q=n} (z'_{p,q}(X,Y) + z''_{p,q}(X,Y)),$$

with

$$z'_{p,q}(X,Y) = \sum_{\substack{p_1+\cdots+p_m=p \\ q_1+\cdots+q_{m-1}=q-1 \\ p_i+q_i\geq 1,\, p_m\geq 1,\, m\geq 1}} \frac{(-1)^{m+1}}{m} \left(\left(\prod_{i=1}^{m-1} \frac{(\mathrm{ad}\, X)^{p_i}}{p_i!} \frac{(\mathrm{ad}\, Y)^{q_i}}{q_i!} \right) \frac{(\mathrm{ad}\, X)^{p_m}}{p_m!} \right) (Y)$$

and

$$z''_{p,q}(X,Y) = \sum_{\substack{p_1+\cdots+p_{m-1}=p-1 \\ q_1+\cdots+q_{m-1}=q \\ p_i+q_i\geq 1,\, m\geq 1}} \frac{(-1)^{m+1}}{m} \left(\prod_{i=1}^{m-1} \frac{(\mathrm{ad}\, X)^{p_i}}{p_i!} \frac{(\mathrm{ad}\, Y)^{q_i}}{q_i!} \right) (X).$$

As a concrete illustration of Dynkin's formula, after some labor, the following Taylor expansion up to order 4 is obtained:

$$\mu(X,Y) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}[X,[X,Y]] + \frac{1}{12}[Y,[Y,X]] - \frac{1}{24}[X,[Y,[X,Y]]] \\ + \text{ higher order terms.}$$

Observe that due to the lack of associativity of the Lie bracket quite different looking expressions can be obtained using the Jacobi identity. For example,

$$-[X,[Y,[X,Y]]] = [Y,[X,[Y,X]]].$$

There is also an integral version of the Campbell-Baker-Hausdorff formula; see Hall [56] (Chapter 3).

19.4 Problems

Problem 19.1. Let X and Y be two $n \times n$ matrices with complex entries. Prove that if $[X,[X,Y]] = [Y,[X,Y]] = 0$, then

$$e^X e^Y = e^{X+Y+\frac{1}{2}[X,Y]}.$$

Hint. For help, see Hall [56], Chapter 3.

Problem 19.2. The (complex) function g given by

$$g(z) = \frac{z \log z}{z-1}$$

has a series expansion that converges for $|z-1|<1$.

(1) Prove that

$$g(z) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k(k+1)}\,(z-1)^k.$$

(2) It can be shown that if X and Y are two $n\times n$ matrices with complex entries and if $\|X\|$ and $\|Y\|$ are small enough, then

$$\log(e^X e^Y) = X + \int_0^1 g(e^{\mathrm{ad}_X}\ e^{t\mathrm{ad}_Y})(Y)dt.$$

The above formula is the integral form of the Campbell-Baker-Hausdorff formula; see Hall [56] (Chapter 3, Theorem 3.3).

Use the above fornula to prove that

$$\log(e^X e^Y) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}[X,[X,Y]] + \frac{1}{12}[Y,[Y,X]]$$
$$+\ \text{higher order terms}.$$

Chapter 20

Metrics, Connections, and Curvature on Lie Groups

Since a Lie group G is a smooth manifold, we can endow G with a Riemannian metric. Among all the Riemannian metrics on a Lie groups, those for which the left translations (or the right translations) are isometries are of particular interest because they take the group structure of G into account. As a consequence, it is possible to find explicit formulae for the Levi-Civita connection and the various curvatures, especially in the case of metrics which are both left and right-invariant.

In Section 20.1 we define left-invariant and right-invariant metrics on a Lie group. We show that left-invariant metrics are obtained by picking some inner product on $\mathfrak{g}$ and moving it around the group to the other tangent spaces T_gG using the maps $(dL_{g^{-1}})_g$ (with $g \in G$). Right-invariant metrics are obtained by using the maps $(dR_{g^{-1}})_g$.

In Section 20.2 we give four characterizations of bi-invariant metrics. The first one refines the criterion of the existence of a left-invariant metric and states that every bi-invariant metric on a Lie group G arises from some Ad-invariant inner product on the Lie algebra $\mathfrak{g}$.

In Section 20.3 we show that if G is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of G, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae are be obtained. In particular the geodesics coincide with the one-parameter groups induced by left-invariant vector fields.

Section 20.5 introduces simple and semisimple Lie algebras. They play a major role in the structure theory of Lie groups

Section 20.6 is devoted to the Killing form. It is an important concept, and we establish some of its main properties. Remarkably, the Killing form yields a simple criterion due to Élie Cartan for testing whether a Lie algebra is semisimple. Indeed, a Lie algebra $\mathfrak{g}$ is semisimple iff its Killing form B is non-degenerate. We also show that a connected Lie group is compact and semisimple iff its Killing form is negative definite.

We conclude this chapter with a section on Cartan connections (Section 20.7). Unfortunately, if a Lie group G does not admit a bi-invariant metric, under the Levi-Civita connection, geodesics are generally not given by the exponential map $\exp\colon \mathfrak{g} \to G$. If we are willing to consider connections not induced by a metric, then it turns out that there is a fairly natural connection for which the geodesics coincide with integral curves of left-invariant vector fields. We are led to consider left-invariant connections. It turns out that there is a one-to-one correspondence between left-invariant connections and bilinear maps $\alpha\colon \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. Connections for which the geodesics are given by the exponential map are those for which α is skew-symmetric. These connections are called *Cartan connections*.

This chapter makes extensive use of results from a beautiful paper of Milnor [84].

20.1 Left (resp. Right) Invariant Metrics

In a Lie group G, since the operations dL_a and dR_a are diffeomorphisms for all $a \in G$, it is natural to consider the metrics for which these maps are isometries.

Definition 20.1. A metric $\langle -, - \rangle$ on a Lie group G is called *left-invariant* (resp. *right-invariant*) iff

$$\langle u, v\rangle_b = \langle (dL_a)_b u, (dL_a)_b v\rangle_{ab} \quad (\text{resp. } \langle u, v\rangle_b = \langle (dR_a)_b u, (dR_a)_b v\rangle_{ba}),$$

for all $a, b \in G$ and all $u, v \in T_bG$. A Riemannian metric that is both left and right-invariant is called a *bi-invariant metric*.

As shown in the next proposition, left-invariant (resp. right-invariant) metrics on G are induced by inner products on the Lie algebra $\mathfrak{g}$ of G. In what follows the identity element of the Lie group G will be denoted by e or 1.

Proposition 20.1. *There is a bijective correspondence between left-invariant (resp. right invariant) metrics on a Lie group G, and inner products on the Lie algebra $\mathfrak{g}$ of G.*

Proof. If the metric on G is left-invariant, then for all $a \in G$ and all $u, v \in T_aG$, we have

$$\langle u, v\rangle_a = \langle (dL_{a^{-1}})_a u, (dL_{a^{-1}})_a v\rangle_e,$$

which shows that our metric is completely determined by its restriction to $\mathfrak{g} = T_eG$. Conversely, let $\langle -, - \rangle$ be an inner product on $\mathfrak{g}$ and set

$$\langle u, v\rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v\rangle,$$

for all $u, v \in T_gG$ and all $g \in G$. Obviously, the family of inner products, $\langle -, - \rangle_g$, yields a Riemannian metric on G. To prove that it is left-invariant, we use the chain rule and the fact that left translations are group isomorphisms. For all $a, b \in G$ and all $u, v \in T_bG$, we have

$$\begin{aligned}
\langle (\mathbf{dL_a})_\mathbf{b}\mathbf{u}, (\mathbf{dL_a})_\mathbf{b}\mathbf{v}\rangle_{ab} &= \langle (dL_{(ab)^{-1}})_{ab}((\mathbf{dL_a})_\mathbf{b}\mathbf{u})), (dL_{(ab)^{-1}})_{ab}((\mathbf{dL_a})_\mathbf{b}\mathbf{v})\rangle \\
&= \langle d(L_{(ab)^{-1}} \circ L_a)_b u, d(L_{(ab)^{-1}} \circ L_a)_b v\rangle \\
&= \langle d(L_{b^{-1}a^{-1}} \circ L_a)_b u, d(L_{b^{-1}a^{-1}} \circ L_a)_b v\rangle \\
&= \langle (dL_{b^{-1}})_b u, (dL_{b^{-1}})_b v\rangle \\
&= \langle u, v\rangle_b,
\end{aligned}$$

as desired.

To get a right-invariant metric on G, set

$$\langle u, v\rangle_g = \langle (dR_{g^{-1}})_g u, (dR_{g^{-1}})_g v\rangle,$$

for all $u, v \in T_gG$ and all $g \in G$. The verification that this metric is right-invariant is analogous. □

If G has dimension n, then since inner products on $\mathfrak{g}$ are in one-to-one correspondence with $n \times n$ positive definite matrices, we see that G possesses a family of left-invariant metrics of dimension $\frac{1}{2}\,n(n+1)$.

If G has a left-invariant (resp. right-invariant) metric, since left-invariant (resp. right-invariant) translations are isometries and act transitively on G, the space G is called a *homogeneous Riemannian manifold*.

Proposition 20.2. *Every Lie group G equipped with a left-invariant (resp. right-invariant) metric is complete.*

Proof. As G is locally compact, we can pick some $\epsilon > 0$ small enough so that the closed ϵ-ball about the identity is compact. By translation, every ϵ-ball is compact, hence every Cauchy sequence eventually lies within a compact set, and thus, converges. □

We now give four characterizations of bi-invariant metrics.

20.2 Bi-Invariant Metrics

Recall that the adjoint representation $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ of the Lie group G is the map defined such that $\mathrm{Ad}_a\colon \mathfrak{g} \to \mathfrak{g}$ is the linear isomorphism given by

$$\mathrm{Ad}_a = d(\mathbf{Ad}_a)_e = d(R_{a^{-1}} \circ L_a)_e, \quad \text{for every } a \in G.$$

Clearly,

$$\mathrm{Ad}_a = (dR_{a^{-1}})_a \circ (dL_a)_e.$$

Here is the *first* of four criteria for the existence of a bi-invariant metric on a Lie group.

Definition 20.2. Given a Lie group G with Lie algebra $\mathfrak{g}$, we say that an inner product $\langle -, -\rangle$ on $\mathfrak{g}$ is Ad-*invariant* if

$$\langle \mathrm{Ad}_a u, \mathrm{Ad}_a v\rangle = \langle u, v\rangle,$$

for all $a \in G$ and all $u, v \in \mathfrak{g}$.

Proposition 20.3. *There is a bijective correspondence between bi-invariant metrics on a Lie group G and* Ad*-invariant inner products on the Lie algebra $\mathfrak{g}$ of G, namely inner products $\langle -, -\rangle$ on $\mathfrak{g}$ such that Ad_a is an isometry of $\mathfrak{g}$ for all $a \in G$.*

Proof. If $\langle -, -\rangle$ is a bi-invariant metric on G, as

$$\mathrm{Ad}_a = (dR_{a^{-1}})_a \circ (dL_a)_e,$$

we claim that

$$\langle \mathrm{Ad}_a u, \mathrm{Ad}_a v\rangle = \langle u, v\rangle,$$

which means that Ad_a is an isometry on $\mathfrak{g}$. To prove this claim, first observe that the left-invariance of the metric gives

$$\langle (dL_a)_e u, (dL_a)_e v\rangle_a = \langle u, v\rangle.$$

Define $U = (dL_a)_e u \in T_aG$ and $V = (dL_a)_e v \in T_aG$. This time, the right-invariance of the metric implies

$$\langle (dR_{a^{-1}})_a U, (dR_{a^{-1}})_a V\rangle = \langle U, V\rangle_a = \langle (dL_a)_e u, (dL_a)_e v\rangle_a.$$

Since $\langle (dR_{a^{-1}})_a U, (dR_{a^{-1}})_a V\rangle = \langle \mathrm{Ad}_a u, \mathrm{Ad}_a v\rangle$, the previous equation verifies the claim.

Conversely, if $\langle -, -\rangle$ is any inner product on $\mathfrak{g}$ such that Ad_a is an isometry of $\mathfrak{g}$ for all $b \in G$, we need to prove that the metric on G given by

$$\langle u, v\rangle_b = \langle (dL_{b^{-1}})_b u, (dL_{b^{-1}})_b v\rangle, \tag{$\dagger_1$}$$

where $u, v \in T_bG$, is also right-invariant. We have

$$\begin{aligned}
\langle (dR_a)_b u, (dR_a)_b v\rangle_{ba} &= \langle (dL_{(ba)^{-1}})_{ba}((dR_a)_b u), (dL_{(ba)^{-1}})_{ba}((dR_a)_b v)\rangle\\
&= \langle d(L_{a^{-1}} \circ L_{b^{-1}} \circ R_a)_b u, d(L_{a^{-1}} \circ L_{b^{-1}} \circ R_a)_b v\rangle\\
&= \langle d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}} \circ L_{b^{-1}})_b v\rangle\\
&= \langle d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b u, d(R_a \circ L_{a^{-1}})_e \circ d(L_{b^{-1}})_b v\rangle\\
&= \langle \mathrm{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b u, \mathrm{Ad}_{a^{-1}} \circ d(L_{b^{-1}})_b v\rangle\\
&= \langle d(L_{b^{-1}})_b u, d(L_{b^{-1}})_b v\rangle\\
&= \langle u, v\rangle_b,
\end{aligned}$$

as $\langle -, -\rangle$ is left-invariant, (as defined at ($\dagger_1$)), and Ad_g-invariant for all $g \in G$. □

Proposition 20.3 shows that if a Lie group G possesses a bi-invariant metric, then every linear map Ad_a is an orthogonal transformation of $\mathfrak{g}$. It follows that $\mathrm{Ad}(G)$ is a subgroup of the orthogonal group of $\mathfrak{g}$, and so its closure $\overline{\mathrm{Ad}(G)}$ is compact. It turns out that this condition is also sufficient!

To prove the above fact, we make use of an "averaging trick" used in representation theory. But first we need the following definition.

Definition 20.3. A *representation* of a Lie group G is a (smooth) homomorphism $\rho\colon G \to \mathbf{GL}(V)$, where V is some finite-dimensional vector space. For any $g \in G$ and any $u \in V$, we often write $g \cdot u$ for $\rho(g)(u)$. We say that an inner-product $\langle -, - \rangle$ on V is *G-invariant* iff

$$\langle g \cdot u, g \cdot v \rangle = \langle u, v \rangle, \quad \text{for all } g \in G \text{ and all } u, v \in V.$$

If G is compact, then the "averaging trick," also called "Weyl's unitarian trick," yields the following important result.

Theorem 20.4. *If G is a compact Lie group, then for every representation $\rho\colon G \to \mathbf{GL}(V)$, there is a G-invariant inner product on V.*

Proof. This proof uses the fact that a notion of integral invariant with respect to left and right multiplication can be defined on any compact Lie group.

In Section 4.11 of Warner [114], it is shown that a Lie group is orientable, has a left-invariant volume form ω, and for every continuous function f with compact support, we can define the integral $\int_G f = \int_G f\omega$. Furthermore, when G is compact, we may assume that our integral is normalized so that $\int_G \omega = 1$ and in this case, our integral is both left and right invariant. Given any inner product $\langle -, - \rangle$ on V, set

$$\langle\langle u, v \rangle\rangle = \int_G \langle g \cdot u, g \cdot v \rangle, \quad \text{for all } u, v \in V,$$

where $\langle g \cdot u, g \cdot v \rangle$ denotes the function $g \mapsto \langle g \cdot u, g \cdot v \rangle$. It is easily checked that $\langle\langle -, - \rangle\rangle$ is an inner product on V. Furthermore, using the right-invariance of our integral (that is, $\int_G f = \int_G (f \circ R_h)$, for all $h \in G$), we have

$$\begin{aligned}
\langle\langle h \cdot u, h \cdot v \rangle\rangle &= \int_G \langle g \cdot (h \cdot u), g \cdot (h \cdot v) \rangle, && \text{definition of } \langle\langle -, - \rangle\rangle \\
&= \int_G \langle (gh) \cdot u, (gh) \cdot v \rangle, && \text{definition of representation} \\
&= \int_G \langle g \cdot u, g \cdot v \rangle, && \text{right invariance of integral} \\
&= \langle\langle u, v \rangle\rangle,
\end{aligned}$$

which shows that $\langle\langle -, - \rangle\rangle$ is G-invariant. □

Using Theorem 20.4, we can prove the following result giving a criterion for the existence of a G-invariant inner product for any representation of a Lie group G (see Sternberg [110], Chapter 5, Theorem 5.2).

Theorem 20.5. *Let $\rho\colon G \to \mathbf{GL}(V)$ be a (finite-dimensional) representation of a Lie group G. There is a G-invariant inner product on V iff $\overline{\rho(G)}$ is compact. In particular, if G is compact, then there is a G-invariant inner product on V.*

Proof. If V has a G-invariant inner product on V, then each linear map, $\rho(g)$, is an isometry, so $\rho(G)$ is a subgroup of the orthogonal group $\mathbf{O}(V)$ of V. As $\mathbf{O}(V)$ is compact, $\overline{\rho(G)}$ is also compact.

Conversely, assume that $\overline{\rho(G)}$ is compact. In this case, $H = \overline{\rho(G)}$ is a closed subgroup of the Lie group $\mathbf{GL}(V)$, so by Theorem 18.18, H is a compact Lie subgroup of $\mathbf{GL}(V)$. The inclusion homomorphism $H \hookrightarrow \mathbf{GL}(V)$ is a representation of H ($f \cdot u = f(u)$, for all $f \in H$ and all $u \in V$), so by Theorem 20.4, there is an inner product on V which is H-invariant. However, for any $g \in G$, if we write $f = \rho(g) \in H$, then we have

$$\langle g \cdot u, g \cdot v\rangle = \langle f(u), f(v)\rangle = \langle u, v\rangle,$$

proving that $\langle -, -\rangle$ is G-invariant as well. □

Applying Theorem 20.5 to the adjoint representation $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$, we get our *second* criterion for the existence of a bi-invariant metric on a Lie group.

Proposition 20.6. *Given any Lie group G, an inner product $\langle -, -\rangle$ on $\mathfrak{g}$ induces a bi-invariant metric on G iff $\overline{\mathrm{Ad}(G)}$ is compact. In particular, every compact Lie group has a bi-invariant metric.*

Proof. Proposition 20.3 is equivalent to the fact that G possesses a bi-invariant metric iff there is some Ad-invariant inner product on $\mathfrak{g}$. By Theorem 20.5, there is some Ad-invariant inner product on $\mathfrak{g}$ iff $\overline{\mathrm{Ad}(G)}$ is compact, which is the statement of our theorem. □

Proposition 20.6 can be used to prove that certain Lie groups do not have a bi-invariant metric. For example, Arsigny, Pennec and Ayache use Proposition 20.6 to give a short and elegant proof of the fact that $\mathbf{SE}(n)$ does not have any bi-invariant metric for all $n \geq 2$. As noted by these authors, other proofs found in the literature are a lot more complicated and only cover the case $n = 3$.

Recall the adjoint representation of the Lie algebra $\mathfrak{g}$,

$$\mathrm{ad}\colon \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g}),$$

given by $\mathrm{ad} = d\mathrm{Ad}_1$. Here is our *third* criterion for the existence of a bi-invariant metric on a connected Lie group.

Proposition 20.7. *If G is a connected Lie group, an inner product $\langle -,-\rangle$ on $\mathfrak{g}$ induces a bi-invariant metric on G iff the linear map* $\mathrm{ad}(u)\colon \mathfrak{g} \to \mathfrak{g}$ *is skew-adjoint for all $u \in \mathfrak{g}$, which means that*

$$\langle \mathrm{ad}(u)(v), w\rangle = -\langle v, \mathrm{ad}(u)(w)\rangle, \quad \textit{for all } u, v, w \in \mathfrak{g},$$

or equivalently that

$$\langle [v,u], w\rangle = \langle v, [u,w]\rangle, \quad \textit{for all } u, v, w \in \mathfrak{g}.$$

Proof. We follow Milnor [84], Lemma 7.2. By Proposition 20.3 an inner product on $\mathfrak{g}$ induces a bi-invariant metric on G iff Ad_g is an isometry for all $g \in G$. Recall the notion of adjoint of a linear map. Given a linear map $f\colon V \to V$ on a vector space V equipped with an inner product $\langle -,-\rangle$, we define $f^*\colon V \to V$ to be the unique linear map such that

$$\langle f(u), v\rangle = \langle u, f^*(v)\rangle, \qquad \text{for all } u, v \in V,$$

and call f^* the *adjoint* of f. It is a standard fact of linear algebra that f is an isometry iff $f^{-1} = f^*$. Thus $\mathrm{Ad}(g)$ is an isometry iff $\mathrm{Ad}(g)^{-1} = \mathrm{Ad}(g)^*$. By definition, a linear map f is skew-adjoint iff

$$\langle f(u), v\rangle = -\langle u, f(v)\rangle, \qquad \text{for all } u, v \in V,$$

and it is immediately verified that this is equivalent to $f^* = -f$.

First assume that $\mathrm{ad}(u)$ is skew-adjoint for all $u \in \mathfrak{g}$. Proposition 18.6 shows that we can choose a small enough open subset U of $\mathfrak{g}$ containing 0 so that $\exp\colon \mathfrak{g} \to G$ is a diffeomorphism from U to $\exp(U)$. For any $g \in \exp(U)$, there is a unique $u \in \mathfrak{g}$ so that $g = \exp(u)$. By Proposition 18.10,

$$\mathrm{Ad}(g) = \mathrm{Ad}(\exp(u)) = e^{\mathrm{ad}(u)}.$$

Since $\mathrm{Ad}(g)^{-1} = \mathrm{Ad}(g)^*$, the preceding equation implies that

$$\mathrm{Ad}(g)^{-1} = e^{-\mathrm{ad}(u)} \quad \text{and} \quad \mathrm{Ad}(g)^* = e^{\mathrm{ad}(u)^*}.$$

But since $\mathrm{ad}(u)$ is skew-adjoint, we have

$$\mathrm{ad}(u)^* = -\mathrm{ad}(u),$$

which implies that $\mathrm{Ad}(g)^{-1} = \mathrm{Ad}(g)^*$, namely, $\mathrm{Ad}(g)$ is an isometry. Since a connected Lie group is generated by any open subset containing the identity, every $g \in G$ can be written as $g = g_1 \cdots g_m$ with $g_1, \ldots, g_m \in \exp(U)$. Since $\mathrm{Ad}_g = \mathrm{Ad}_{g_1 \cdots g_m} = \mathrm{Ad}_{g_1} \circ \cdots \circ \mathrm{Ad}_{g_m}$, and since by the previous reasoning each $\mathrm{Ad}(g_i)$ is an isometry, we deduce that $\mathrm{Ad}(g)$ is an isometry.

Conversely, we prove that if every $\mathrm{Ad}(g)$ is an isometry, then $\mathrm{ad}(u)$ is skew-adjoint for all $u \in \mathfrak{g}$. It is enough to prove that for any basis $(u_1, \ldots, u_n)$ of $\mathfrak{g}$, that $\mathrm{ad}(u_i)$ is skew-adjoint. By the remark before Proposition 3.10, the matrix exponential is also a diffeomorphism on an open subset V containing $0 \in \mathrm{End}(\mathfrak{g})$. We can pick the basis $(u_1, \ldots, u_n)$ in such a way that $\mathrm{ad}(u_i) \in V$ and $\mathrm{ad}(u_i)^* \in V$ for $i = 1, \ldots, n$, and we let $g_i = \exp(u_i)$. Then as in the previous part

$$\mathrm{Ad}(g_i)^{-1} = e^{-\mathrm{ad}(u_i)} \quad \text{and} \quad \mathrm{Ad}(g_i)^* = e^{\mathrm{ad}(u_i)^*}.$$

Since each $\mathrm{Ad}(g_i)$ is an isometry, we have $\mathrm{Ad}(g_i)^{-1} = \mathrm{Ad}(g_i)^*$, which implies that

$$e^{-\mathrm{ad}(u_i)} = e^{\mathrm{ad}(u_i)^*}.$$

Since $\mathrm{ad}(u_i) \in V$ and $\mathrm{ad}(u_i)^* \in V$, and since the matrix exponential is bijective in V, we conclude that

$$\mathrm{ad}(u_i)^* = -\mathrm{ad}(u_i),$$

which means that $\mathrm{ad}(u_i)$ is skew-adjoint.

The skew-adjointness of $\mathrm{ad}(u)$ means that

$$\langle \mathrm{ad}(u)(v), w \rangle = -\langle v, \mathrm{ad}(u)(w) \rangle \quad \text{for all } u, v, w \in \mathfrak{g},$$

and since $\mathrm{ad}(u)(v) = [u, v]$ and $[u, v] = -[v, u]$, we get

$$\langle [v, u], w \rangle = \langle v, [u, w] \rangle$$

which is the last claim of the proposition. □

Remark: It will be convenient to say that an inner product on $\mathfrak{g}$ is *bi-invariant* iff every $\mathrm{ad}(u)$ is skew-adjoint.

The following variant of Proposition 20.7 will also be needed. This is a special case of Lemma 3 in O'Neill [91] (Chapter 11).

Proposition 20.8. *If G is Lie group equipped with an inner product $\langle -, - \rangle$ on $\mathfrak{g}$ that induces a bi-invariant metric on G, then $\mathrm{ad}(X)\colon \mathfrak{g}^L \to \mathfrak{g}^L$ is skew-adjoint for all left-invariant vector fields $X \in \mathfrak{g}^L$, which means that*

$$\langle \mathrm{ad}(X)(Y), Z \rangle = -\langle Y, \mathrm{ad}(X)(Z) \rangle, \quad \textit{for all } X, Y, Z \in \mathfrak{g}^L,$$

or equivalently that

$$\langle [Y, X], Z \rangle = \langle Y, [X, Z] \rangle, \quad \textit{for all } X, Y, Z \in \mathfrak{g}^L.$$

Proof. By the bi-invariance of the metric, Proposition 20.3 implies that the inner product $\langle -, - \rangle$ on $\mathfrak{g}$ is Ad-invariant. For any two left-invariant vector fields $X, Y \in \mathfrak{g}^L$, we have

$$\begin{aligned} \langle \mathrm{Ad}_a X, \mathrm{Ad}_a Y \rangle_e &:= \langle \mathrm{Ad}_a X(e), \mathrm{Ad}_a Y(e) \rangle \\ &= \langle X(e), Y(e) \rangle, \end{aligned}$$

which shows that the function $a \mapsto \langle \mathrm{Ad}_a X, \mathrm{Ad}_a Y\rangle_e$ is constant. For any left-invariant vector field Z, by taking the derivative of this function with $a = \exp(tZ(e))$ at $t = 0$, we get

$$\langle [Z(e), X(e)], Y(e)\rangle + \langle X(e), [Z(e), Y(e)]\rangle = 0.$$

Since dL_g is a diffeomorphism for every $g \in G$, the metric on G is assumed to be bi-invariant, and $X(g) = (dL_g)_e(X(e))$ for any left-invariant vector field X, we have

$$\begin{aligned}\langle [Z(g), X(g)], Y(g)\rangle_g + \langle X(g), [Z(g), Y(g)]\rangle_g = \\ &\langle [(dL_g)_e(Z(e)), (dL_g)_e(X(e))], (dL_g)_e(Y(e))\rangle_g \\ &+ \langle (dL_g)_e(X(e)), [(dL_g)_e(Z(e)), (dL_g)_e(Y(e))]\rangle_g \\ &= \langle [Z(e), X(e)], Y(e)\rangle + \langle X(e), [Z(e), Y(e)]\rangle = 0.\end{aligned}$$

Therefore,

$$\langle [Z, X], Y\rangle + \langle X, [Z, Y]\rangle = 0,$$

which is equivalent to

$$\langle [X, Z], Y\rangle = \langle X, [Z, Y]\rangle,$$

and to

$$\langle \mathrm{ad}(Z)(X), Y\rangle = -\langle X, \mathrm{ad}(Z)(Y)\rangle.$$

If we apply the permutation $(X, Y, Z) \mapsto Y, Z, X$, we obtain our proposition. □

We now turn to our *fourth* criterion. If G is a connected Lie group, then the existence of a bi-invariant metric on G places a heavy restriction on its group structure, as shown by the following result from Milnor's paper [84] (Lemma 7.5).

Theorem 20.9. *A connected Lie group G admits a bi-invariant metric iff it is isomorphic to the Cartesian product of a compact Lie group and a vector space ($\mathbb{R}^m$, for some $m \geq 0$).*

A proof of Theorem 20.9 can be found in Milnor [84] (Lemma 7.4 and Lemma 7.5). The proof uses the universal covering group and it is a bit involved. Because it is really quite beautiful,we will outline the structure of the proof.

First, recall from Definition 18.14 that a subset $\mathfrak{h}$ of a Lie algebra $\mathfrak{g}$ is a *Lie subalgebra* iff it is a subspace of $\mathfrak{g}$ (as a vector space) and if it is closed under the bracket operation on $\mathfrak{g}$. A subalgebra $\mathfrak{h}$ of $\mathfrak{g}$ is *abelian* iff $[x, y] = 0$ for all $x, y \in \mathfrak{h}$. An *ideal* in $\mathfrak{g}$ is a Lie subalgebra $\mathfrak{h}$ such that

$$[h, g] \in \mathfrak{h}, \qquad \text{for all } h \in \mathfrak{h} \text{ and all } g \in \mathfrak{g}.$$

Definition 20.4. A Lie algebra $\mathfrak{g}$ is *simple* iff it is non-abelian and if it has no ideal other than (0) and $\mathfrak{g}$. A Lie group is *simple* iff its Lie algebra is simple.

In a first step for the proof of Theorem 20.9, it is shown that if G has a bi-invariant metric, then its Lie algebra $\mathfrak{g}$ can be written as an orthogonal direct sum

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is either a simple ideal or a one-dimensional abelian ideal; that is, $\mathfrak{g}_i \cong \mathbb{R}$.

The next step is to lift the ideals $\mathfrak{g}_i$ to simply connected normal subgroups G_i of the universal covering group $\widetilde{G}$ of G. For every simple ideal $\mathfrak{g}_i$ in the decomposition, it is proved that there is some constant $c_i > 0$, so that all Ricci curvatures are strictly positive and bounded from below by c_i. Therefore, by Myers' theorem (Theorem 16.28), G_i is compact. It follows that $\widetilde{G}$ is isomorphic to a product of compact simple Lie groups and some vector space $\mathbb{R}^m$. Finally, we know that G is isomorphic to the quotient of $\widetilde{G}$ by a discrete normal subgroup of $\widetilde{G}$, which yields our theorem.

Because it is a fun proof, we prove the statement about the structure of a Lie algebra for which each $\mathrm{ad}(u)$ is skew-adjoint.

Proposition 20.10. *Let $\mathfrak{g}$ be a Lie algebra with an inner product such that the linear map $\mathrm{ad}(u)$ is skew-adjoint for every $u \in \mathfrak{g}$. Then the orthogonal complement $\mathfrak{a}^\perp$ of any ideal $\mathfrak{a}$ is itself an ideal. Consequently, $\mathfrak{g}$ can be expressed as an orthogonal direct sum*

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is either a simple ideal or a one-dimensional abelian ideal ($\mathfrak{g}_i \cong \mathbb{R}$).

Proof. Assume $u \in \mathfrak{g}$ is orthogonal to $\mathfrak{a}$, i.e. $u \in \mathfrak{a}^\perp$. We need to prove that $[u, v] = -[v, u]$ is orthogonal to $\mathfrak{a}$ for all $v \in \mathfrak{g}$. But, as $\mathrm{ad}(v)$ is skew-adjoint, $\mathrm{ad}(v)(u) = [v, u]$, and $\mathfrak{a}$ is an ideal with $[v, a] \in \mathfrak{a}$ for all $v \in \mathfrak{g}$ and $a \in \mathfrak{a}$, we have

$$\langle [u, v], a\rangle = -\langle [v, u], a\rangle = \langle u, [v, a]\rangle = 0, \qquad \text{for all } a \in \mathfrak{a},$$

which shows that $\mathfrak{a}^\perp$ is an ideal.

For the second statement we use induction on the dimension of $\mathfrak{g}$, *but for this proof, we redefine a simple Lie algebra to be an algebra with no nontrivial proper ideals*. The case where $\dim \mathfrak{g} = 1$ is clear.

For the induction step, if $\mathfrak{g}$ is simple, we are done. Else, $\mathfrak{g}$ has some nontrivial proper ideal $\mathfrak{h}$, and if we pick $\mathfrak{h}$ of minimal dimension p, with $1 \leq p < n = \dim \mathfrak{g}$, then $\mathfrak{h}$ is simple. Now, $\mathfrak{h}^\perp$ is also an ideal and $\dim \mathfrak{h}^\perp < n$, so the induction hypothesis applies. Therefore, we have an orthogonal direct sum

$$\mathfrak{g} = \mathfrak{g}_1 \oplus \cdots \oplus \mathfrak{g}_k,$$

where each $\mathfrak{g}_i$ is simple *in our relaxed sense*. However, if $\mathfrak{g}_i$ is not abelian, then it is simple in the usual sense, and if $\mathfrak{g}_i$ is abelian, having no proper nontrivial ideal, it must be one-dimensional and we get our decomposition. ☐

We now investigate connections and curvature on Lie groups with a left-invariant metric.

20.3 Connections and Curvature of Left-Invariant Metrics on Lie Groups

If G is a Lie group equipped with a left-invariant metric, then it is possible to express the Levi-Civita connection and the sectional curvature in terms of quantities defined over the Lie algebra of G, at least for left-invariant vector fields. When the metric is bi-invariant, much nicer formulae are be obtained. *In this section we always assume that our Lie groups are equipped with the Levi-Civita connection.*

If $\langle -, - \rangle$ is a left-invariant metric on G, then for any two left-invariant vector fields X, Y, we have

$$\langle X, Y\rangle_g = \langle X(g), Y(g)\rangle_g = \langle (dL_g)_e X(e), (dL_g)_e Y(e)\rangle_g = \langle X_e, Y_e\rangle_e = \langle X, Y\rangle_e,$$

which shows that the function $g \mapsto \langle X, Y\rangle_g$ is constant. Therefore, for any vector field Z,

$$Z(\langle X, Y\rangle) = 0.$$

If we go back to the Koszul formula (Proposition 14.9)

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= X(\langle Y, Z\rangle) + Y(\langle X, Z\rangle) - Z(\langle X, Y\rangle) \\ &\quad -\ \langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle, \end{aligned}$$

we deduce that for all left-invariant vector fields $X, Y, Z \in \mathfrak{g}^L$, we have

$$2\langle \nabla_X Y, Z\rangle = -\langle Y, [X, Z]\rangle - \langle X, [Y, Z]\rangle - \langle Z, [Y, X]\rangle,$$

which can be rewritten as

$$2\langle \nabla_X Y, Z\rangle = \langle [X, Y], Z\rangle - \langle [Y, Z], X\rangle + \langle [Z, X], Y\rangle. \tag{$\dagger$}$$

Note that ($\dagger$) is equivalent to

$$\begin{aligned} 2\langle \nabla_X Y, Z\rangle &= \langle [X, Y], Z\rangle - \langle [Y, Z], X\rangle - \langle [X, Z], Y\rangle \\ &= \langle [X, Y], Z\rangle - \langle \mathrm{ad}(Y)(Z), X\rangle - \langle \mathrm{ad}(X)(Z), Y\rangle \\ &= \langle [X, Y], Z\rangle - \langle Z, \mathrm{ad}(Y)^*(X)\rangle - \langle Z, \mathrm{ad}(X)^*(Y)\rangle. \end{aligned}$$

The above yields the formula

$$\nabla_X Y = \frac{1}{2}\left([X, Y] - \mathrm{ad}(X)^* Y - \mathrm{ad}(Y)^* X\right), \qquad X, Y \in \mathfrak{g}^L,$$

where $\mathrm{ad}(X)^*$ denotes the adjoint of $\mathrm{ad}(X)$ as defined in Definition 18.12.

Remark: Given any two vector $u, v \in \mathfrak{g}$, it is common practice (even though this is quite confusing) to denote by $\nabla_u v$ the result of evaluating the vector field $\nabla_{u^L} v^L$ at e (so, $\nabla_u v = (\nabla_{u^L} v^L)(e)$).

Following Milnor, if we pick an orthonormal basis $(e_1, \ldots, e_n)$ *w.r.t.* our inner product on $\mathfrak{g}$, and if we define the structure constants α_{ijk} by

$$\alpha_{ijk} = \langle [e_i, e_j], e_k \rangle,$$

we see by (†) that

$$\nabla_{e_i} e_j = \frac{1}{2} \sum_k (\alpha_{ijk} - \alpha_{jki} + \alpha_{kij}) e_k. \tag{$*$}$$

For example, let $G = \mathbf{SO}(3)$, the group of 3×3 rotation matrices. Then $\mathfrak{g} = \mathfrak{so}(3)$ is the vector space of skew symmetric 3×3 matrices with orthonormal basis

$$e_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \qquad e_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \qquad e_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

since the left invariant, indeed bi-invariant, metric on $\mathfrak{g}$ is

$$\langle B_1, B_2 \rangle = \mathrm{tr}(B_1^\top B_2) = -\mathrm{tr}(B_1 B_2).$$

Matrix multiplication shows that

$$\begin{aligned}
[e_1, e_2] &= e_1 e_2 - e_2 e_1 = \frac{1}{\sqrt{2}} e_3 \\
[e_2, e_3] &= e_2 e_3 - e_3 e_2 = \frac{1}{\sqrt{2}} e_1 \\
[e_3, e_1] &= e_3 e_1 - e_1 e_3 = \frac{1}{\sqrt{2}} e_2 \\
[e_1, e_1] &= [e_2, e_2] = [e_3, e_3] = 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
-\alpha_{213} = \alpha_{123} &= \langle [e_1, e_2], e_3 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_3 \rangle = \frac{1}{\sqrt{2}} \\
-\alpha_{211} = \alpha_{121} &= \langle [e_1, e_2], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_1 \rangle = 0 \\
-\alpha_{212} = \alpha_{122} &= \langle [e_1, e_2], e_2 \rangle = \frac{1}{\sqrt{2}} \langle e_3, e_2 \rangle = 0 \\
\alpha_{112} &= \langle [e_1, e_1], e_2 \rangle = 0, \qquad \alpha_{221} = \langle [e_2, e_2], e_1 \rangle = 0 \\
-\alpha_{321} = \alpha_{231} &= \langle [e_2, e_3], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_1, e_1 \rangle = \frac{1}{\sqrt{2}} \\
-\alpha_{131} = \alpha_{311} &= \langle [e_3, e_1], e_1 \rangle = \frac{1}{\sqrt{2}} \langle e_2, e_1 \rangle = 0
\end{aligned}$$

$$\begin{aligned}
-\alpha_{133} &= \alpha_{313} = \langle [e_3, e_1], e_3\rangle = \frac{1}{\sqrt{2}}\langle e_2, e_3\rangle = 0\\
\alpha_{113} &= \langle [e_1, e_1], e_3\rangle = 0, \qquad \alpha_{331} = \langle [e_3, e_3], e_1\rangle = 0\\
-\alpha_{132} &= \alpha_{312} = \langle [e_3, e_1], e_2\rangle = \frac{1}{\sqrt{2}}\langle e_2, e_2\rangle = \frac{1}{\sqrt{2}}\\
-\alpha_{322} &= \alpha_{232} = \langle [e_2, e_3], e_2\rangle = \frac{1}{\sqrt{2}}\langle e_1, e_2\rangle = 0\\
-\alpha_{323} &= \alpha_{233} = \langle [e_2, e_3], e_3\rangle = \frac{1}{\sqrt{2}}\langle e_1, e_3\rangle = 0\\
\alpha_{223} &= \langle [e_2, e_2], e_3\rangle = 0, \qquad \alpha_{332} = \langle [e_3, e_3], e_2\rangle = 0,
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{e_1} e_2 &= -\nabla_{e_2} e_1 = \frac{1}{2}\sum_{k=1}^{3}(\alpha_{12k} - \alpha_{2k1} + \alpha_{k12}) e_k\\
&= \frac{1}{2}(\alpha_{123} - \alpha_{231} + \alpha_{312}) e_3 = \frac{1}{2\sqrt{2}} e_3 = \frac{1}{2}[e_1, e_2]\\
\nabla_{e_1} e_3 &= -\nabla_{e_3} e_1 = \frac{1}{2}\sum_{k=1}^{3}(\alpha_{13k} - \alpha_{3k1} + \alpha_{k13}) e_k\\
&= \frac{1}{2}(\alpha_{132} - \alpha_{321} + \alpha_{213}) e_2 = -\frac{1}{2\sqrt{2}} e_2 = \frac{1}{2}[e_1, e_3]\\
\nabla_{e_2} e_3 &= -\nabla_{e_3} e_2 = \frac{1}{2}\sum_{k=1}^{3}(\alpha_{23k} - \alpha_{3k2} + \alpha_{k23}) e_k\\
&= \frac{1}{2}(\alpha_{213} - \alpha_{312} + \alpha_{123}) e_1 = -\frac{1}{2\sqrt{2}} e_1 = \frac{1}{2}[e_2, e_3]
\end{aligned}$$

$$\begin{aligned}
\nabla_{e_1} e_1 &= \frac{1}{2}\sum_{k=1}^{3}(\alpha_{11k} - \alpha_{1k1} + \alpha_{k11}) e_k = 0\\
\nabla_{e_2} e_2 &= \frac{1}{2}\sum_{k=1}^{3}(\alpha_{22k} - \alpha_{2k2} + \alpha_{k22}) e_k = 0\\
\nabla_{e_3} e_3 &= \frac{1}{2}\sum_{k=1}^{3}(\alpha_{33k} - \alpha_{3k3} + \alpha_{k33}) e_k = 0.
\end{aligned}$$

Now for orthonormal vectors u, v, the sectional curvature is given by

$$K(u, v) = \langle R(u, v)u, v\rangle,$$

with

$$R(u, v) = \nabla_{[u,v]} - \nabla_u \nabla_v + \nabla_v \nabla_u.$$

If we plug the expressions from Equation $(*)$ into the definitions, we obtain the following proposition from Milnor [84] (Lemma 1.1).

Proposition 20.11. *Given a Lie group G equipped with a left-invariant metric, for any orthonormal basis $(e_1, \ldots, e_n)$ of $\mathfrak{g}$, and with the structure constants $\alpha_{ijk} = \langle [e_i, e_j], e_k \rangle$, the sectional curvature $K(e_1, e_2)$ is given by*

$$\begin{aligned} K(e_i, e_j) = \sum_k \biggl(& \frac{1}{2}\alpha_{ijk}(-\alpha_{ijk} + \alpha_{jki} + \alpha_{kij}) \\ & - \frac{1}{4}(\alpha_{ijk} - \alpha_{jki} + \alpha_{kij})(\alpha_{ijk} + \alpha_{jki} - \alpha_{kij}) - \alpha_{kii}\alpha_{kjj} \biggr). \end{aligned}$$

For $\mathbf{SO}(3)$, the formula of Proposition 20.11, when evaluated with the previously computed structure constants, gives

$$\begin{aligned} K(e_1, e_2) &= \frac{1}{8} = \frac{1}{8}\langle e_3, e_3 \rangle = \frac{1}{4}\langle [e_1, e_2], [e_1, e_2] \rangle \\ K(e_1, e_3) &= \frac{1}{8} = \frac{1}{8}\langle e_2, e_2 \rangle = \frac{1}{4}\langle [e_1, e_3], [e_1, e_3] \rangle \\ K(e_2, e_3) &= \frac{1}{8} = \frac{1}{8}\langle e_1, e_1 \rangle = \frac{1}{4}\langle [e_2, e_3], [e_3, e_3] \rangle \\ K(e_1, e_1) &= K(e_2, e_2) = K(e_3, e_3) = 0. \end{aligned}$$

Although the above formula is not too useful in general, in some cases of interest, a great deal of cancellation takes place so that a more useful formula can be obtained. An example of this situation is provided by the next proposition (Milnor [84], Lemma 1.2).

Proposition 20.12. *Given a Lie group G equipped with a left-invariant metric, for any $u \in \mathfrak{g}$, if the linear map $\mathrm{ad}(u)$ is skew-adjoint, then*

$$K(u, v) \geq 0 \quad \textit{for all } v \in \mathfrak{g},$$

where equality holds iff u is orthogonal to $[v, \mathfrak{g}] = \{[v, x] \mid x \in \mathfrak{g}\}$.

Proof. We may assume that u and v are orthonormal. If we pick an orthonormal basis such that $e_1 = u$ and $e_2 = v$, the fact that $\mathrm{ad}(e_1)$ is skew-adjoint means that the array (α_{1jk}) is skew-symmetric (in the indices j and k). It follows that the formula of Proposition 20.11 reduces to

$$K(e_1, e_2) = \frac{1}{4}\sum_k \alpha_{2k1}^2,$$

so $K(e_1, e_2) \geq 0$, as claimed. Furthermore, $K(e_1, e_2) = 0$ iff $\alpha_{2k1} = 0$ for $k = 1, \ldots, n$; that is, $\langle [e_2, e_k], e_1 \rangle = 0$ for $k = 1, \ldots, n$, which means that e_1 is orthogonal to $[e_2, \mathfrak{g}]$. □

For the next proposition we need the following definition.

Definition 20.5. The *center* $Z(\mathfrak{g})$ of a Lie algebra $\mathfrak{g}$ is the set of all elements $u \in \mathfrak{g}$ such that $[u, v] = 0$ for all $v \in \mathfrak{g}$, or equivalently, such that $\mathrm{ad}(u) = 0$.

Proposition 20.13. *Given a Lie group G equipped with a left-invariant metric, for any u in the center $Z(\mathfrak{g})$ of $\mathfrak{g}$,*

$$K(u, v) \geq 0 \quad \text{for all } v \in \mathfrak{g}.$$

Proof. For any element u in the center of $\mathfrak{g}$, we have $\mathrm{ad}(u) = 0$, and the zero map is obviously skew-adjoint. □

Recall that the Ricci curvature $\mathrm{Ric}(u, v)$ is the trace of the linear map $y \mapsto R(u, y)v$. With respect to any orthonormal basis $(e_1, \ldots, e_n)$ of $\mathfrak{g}$, we have

$$\mathrm{Ric}(u, v) = \sum_{j=1}^{n} \langle R(u, e_j)v, e_j\rangle = \sum_{j=1}^{n} R(u, e_j, v, e_j).$$

The Ricci curvature is a symmetric form, so it is completely determined by the quadratic form

$$r(u) = \mathrm{Ric}(u, u) = \sum_{j=1}^{n} R(u, e_j, u, e_j).$$

Definition 20.6. If u is a unit vector, $r(u) = \mathrm{Ric}(u, u)$ is called the *Ricci curvature in the direction u.* If we pick an orthonormal basis such that $e_1 = u$, then

$$r(e_1) = \sum_{i=2}^{n} K(e_1, e_i).$$

For computational purposes it may be more convenient to introduce the *Ricci transformation* $\mathrm{Ric}^{\#}$, defined by

$$\mathrm{Ric}^{\#}(x) = \sum_{i=1}^{n} R(e_i, x)e_i.$$

Observe that

$$\begin{aligned}
\langle \mathrm{Ric}^{\#}(x), y\rangle &= \langle \sum_{i=1}^{n} R(e_i, x)e_i, y\rangle = \sum_{i=1}^{n} \langle R(e_i, x)e_i, y\rangle = \sum_{i=1}^{n} R(e_i, x, e_i, y) \\
&= \sum_{i=1}^{n} R(e_i, y, e_i, x), \qquad \text{by Proposition16.3 (4)} \\
&= \sum_{i=1}^{n} \langle R(e_i, y)e_i, x\rangle = \langle \sum_{i=1}^{n} R(e_i, y)e_i, x\rangle = \langle x, \mathrm{Ric}^{\#}(y)\rangle.
\end{aligned}$$

Hence, we showed the following proposition.

Proposition 20.14. *The Ricci transformation defined by*

$$\mathrm{Ric}^{\#}(x) = \sum_{i=1}^{n} R(e_i, x)e_i$$

is self-adjoint, and it is also the unique map so that

$$r(x) = \mathrm{Ric}(x, x) = \langle \mathrm{Ric}^{\#}(x), x\rangle, \qquad \textit{for all } x \in \mathfrak{g}.$$

Definition 20.7. The eigenvalues of $\mathrm{Ric}^{\#}$ are called the *principal Ricci curvatures*.

Proposition 20.15. *Given a Lie group G equipped with a left-invariant metric, if the linear map* $\mathrm{ad}(u)$ *is skew-adjoint, then $r(u) \geq 0$, where equality holds iff u is orthogonal to the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$.*

Proof. This follows from Proposition 20.12 since

$$r(u) = \mathrm{Ric}(u, u) = \sum_{j=1}^{n} R(u, e_j, u, e_j) = \sum_{j=1}^{n} \langle R(u, e_j)u, e_j\rangle = \sum_{j=1}^{n} K(u, e_j). \qquad \square$$

In particular, if u is in the center of $\mathfrak{g}$, then $r(u) \geq 0$.

As a corollary of Proposition 20.15, we have the following result which is used in the proof of Theorem 20.9.

Proposition 20.16. *If G is a connected Lie group equipped with a bi-invariant metric and if the Lie algebra of G is simple, then there is a constant $c > 0$ so that $r(u) \geq c$ for all unit vector $u \in T_gG$ and for all $g \in G$.*

Proof. First of all, by Proposition 20.7, the linear maps $\mathrm{ad}(u)$ are skew-adjoint for all $u \in \mathfrak{g}$, which implies that $r(u) \geq 0$. As $\mathfrak{g}$ is simple, the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$ is either (0) or $\mathfrak{g}$. But, if $[\mathfrak{g}, \mathfrak{g}] = (0)$, then then $\mathfrak{g}$ is abelian, which is impossible since $\mathfrak{g}$ is simple. Therefore $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which implies $r(u) > 0$ for all $u \neq 0$ (otherwise, u would be orthogonal to $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$, which is impossible). As the set of unit vectors in $\mathfrak{g}$ is compact, the function $u \mapsto r(u)$ achieves it minimum c, and $c > 0$ as $r(u) > 0$ for all $u \neq 0$. But, $dL_g \colon \mathfrak{g} \to T_gG$ is an isometry for all $g \in G$, so $r(u) \geq c$ for all unit vectors $u \in T_gG$, and for all $g \in G$. $\square$

By Myers' theorem (Theorem 16.28), if the Lie group G satisfies the conditions of Proposition 20.16, it is compact and has a finite fundamental group.

The following interesting theorem is proved in Milnor (Milnor [84], Theorem 2.2).

Theorem 20.17. *A connected Lie group G admits a left-invariant metric with $r(u) > 0$ for all unit vectors $u \in \mathfrak{g}$ (all Ricci curvatures are strictly positive) iff G is compact and has a finite fundamental group.*

The following criterion for obtaining a direction of negative curvature is also proved in Milnor (Milnor [84], Lemma 2.3).

Proposition 20.18. *Given a Lie group G equipped with a left-invariant metric, if u is orthogonal to the commutator ideal $[\mathfrak{g}, \mathfrak{g}]$, then $r(u) \leq 0$, where equality holds iff $\mathrm{ad}(u)$ is self-adjoint.*

20.4 Connections and Curvature of Bi-Invariant Metrics on Lie Groups

When G possesses a bi-invariant metric and G is equipped with the Levi-Civita connection, the group exponential coincides with the exponential defined in terms of geodesics. Much nicer formulae are also obtained for the Levi-Civita connection and the curvatures.

First of all, since by Proposition 20.8,

$$\langle [Y, Z], X\rangle = \langle Y, [Z, X]\rangle,$$

the last two terms in equation (†), namely

$$2\langle \nabla_X Y, Z\rangle = \langle [X, Y], Z\rangle - \langle [Y, Z], X\rangle + \langle [Z, X], Y\rangle,$$

cancel out, and we get

$$\nabla_X Y = \frac{1}{2}\,[X, Y], \quad \text{for all } X, Y \in \mathfrak{g}^L.$$

This is equivalent to

$$\nabla_X = \frac{1}{2}\mathrm{ad}(X), \quad \text{for all } X \in \mathfrak{g}^L.$$

Then since

$$R(u, v) = \nabla_{[u,v]} - \nabla_u\nabla_v + \nabla_v\nabla_u,$$

we get

$$R(u, v) = \frac{1}{2}\mathrm{ad}([u, v]) - \frac{1}{4}\mathrm{ad}(u)\mathrm{ad}(v) + \frac{1}{4}\mathrm{ad}(v)\mathrm{ad}(u).$$

Using the Jacobi identity,

$$\mathrm{ad}([u, v]) = \mathrm{ad}(u)\mathrm{ad}(v) - \mathrm{ad}(v)\mathrm{ad}(u),$$

we get

$$R(u, v) = \frac{1}{4}\,\mathrm{ad}[u, v],$$

so

$$R(u, v)w = \frac{1}{4}\,[[u, v], w].$$

Hence, for unit orthogonal vectors u, v, the sectional curvature $K(u, v) = \langle R(u, v)u, v\rangle$ is given by

$$K(u, v) = \frac{1}{4}\,\langle [[u, v], u], v\rangle,$$

which (by the Proposition 20.7 equality $\langle [x, y], z\rangle = \langle x, [y, z]\rangle$) is rewritten as

$$K(u, v) = \frac{1}{4}\,\langle [u, v], [u, v]\rangle.$$

To compute the Ricci curvature $\mathrm{Ric}(u, v)$, we observe that $\mathrm{Ric}(u, v)$ is the trace of the linear map

$$y \mapsto R(u, y)v = \frac{1}{4}\,[[u, y], v] = -\frac{1}{4}\,[v, [u, y]] = -\frac{1}{4}\,\mathrm{ad}(v) \circ \mathrm{ad}(u)(y).$$

However, the bilinear form B on $\mathfrak{g}$ given by

$$B(u, v) = \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v))$$

is a famous object known as the *Killing form* of the Lie algebra $\mathfrak{g}$. We will take a closer look at the Killing form shortly. For the time being, we observe that as $\mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)) = \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}(u))$, we get

$$\mathrm{Ric}(u, v) = -\frac{1}{4}\,B(u, v), \quad \text{for all } u, v \in \mathfrak{g}.$$

We summarize all this in

Proposition 20.19. *For any Lie group G equipped with a bi-invariant metric, the following properties hold:*

(a) The Levi-Civita connection $\nabla_X Y$ is given by

$$\nabla_X Y = \frac{1}{2}\,[X, Y], \qquad \textit{for all } X, Y \in \mathfrak{g}^L.$$

(b) The curvature tensor $R(u, v)$ is given by

$$R(u, v) = \frac{1}{4}\,\mathrm{ad}[u, v], \qquad \textit{for all } u, v \in \mathfrak{g},$$

or equivalently,

$$R(u, v)w = \frac{1}{4}\,[[u, v], w], \qquad \textit{for all } u, v, w \in \mathfrak{g}.$$

(c) The sectional curvature $K(u, v)$ is given by

$$K(u, v) = \frac{1}{4}\,\langle [u, v], [u, v]\rangle,$$

for all pairs of orthonormal vectors $u, v \in \mathfrak{g}$.

(d) *The Ricci curvature* $\mathrm{Ric}(u, v)$ *is given by*

$$\mathrm{Ric}(u, v) = -\frac{1}{4}\, B(u, v), \qquad \textit{for all } u, v \in \mathfrak{g},$$

where B is the Killing form, with

$$B(u, v) = \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)), \qquad \textit{for all } u, v \in \mathfrak{g}.$$

Consequently, $K(u, v) \geq 0$, with equality iff $[u, v] = 0$ and $r(u) = \mathrm{Ric}(u, u) \geq 0$, with equality iff u belongs to the center of $\mathfrak{g}$.

Remark: Proposition 20.19 shows that if a Lie group admits a bi-invariant metric, then its Killing form is negative semi-definite.

What are the geodesics in a Lie group equipped with a bi-invariant metric and the Levi-Civita connection? *The answer is simple: they are the integral curves of left-invariant vector fields*.

Proposition 20.20. *For any Lie group G equipped with a bi-invariant metric, we have:*

(1) *The inversion map $\iota\colon g \mapsto g^{-1}$ is an isometry.*

(2) *For every $a \in G$, if I_a denotes the map given by*

$$I_a(b) = ab^{-1}a, \quad \textit{for all } a, b \in G,$$

then I_a is an isometry fixing a which reverses geodesics; that is, for every geodesic γ through a, we have

$$I_a(\gamma)(t) = \gamma(-t).$$

(3) *The geodesics through e are the integral curves $t \mapsto \exp_{\mathrm{gr}}(tu)$, where $u \in \mathfrak{g}$; that is, the one-parameter groups. Consequently, the Lie group exponential map $\exp_{\mathrm{gr}}\colon \mathfrak{g} \to G$ coincides with the Riemannian exponential map (at e) from T_eG to G, where G is viewed as a Riemannian manifold.*

Proof. (1) Since

$$\iota(g) = g^{-1} = g^{-1}h^{-1}h = (hg)^{-1}h = (R_h \circ \iota \circ L_h)(g),$$

we have

$$\iota = R_h \circ \iota \circ L_h, \qquad \text{for all } h \in G.$$

In particular, for $h = g^{-1}$, we get

$$d\iota_g = (dR_{g^{-1}})_e \circ d\iota_e \circ (dL_{g^{-1}})_g.$$

As $(dR_{g^{-1}})_e$ and $d(L_{g^{-1}})_g$ are isometries (since G has a bi-invariant metric), $d\iota_g$ is an isometry iff $d\iota_e$ is. Thus, it remains to show that $d\iota_e$ is an isometry. However, if we can prove that $d\iota_e = -\mathrm{id}$, then $d\iota_g$ will be an isometry for all $g \in G$.

It remains to prove that $d\iota_e = -\mathrm{id}$. This can be done in several ways. If we denote the multiplication of the group by $\mu\colon G \times G \to G$, then $T_e(G \times G) = T_eG \oplus T_eG = \mathfrak{g} \oplus \mathfrak{g}$, and it is easy to see that

$$d\mu_{(e,e)}(u, v) = u + v, \qquad \text{for all } u, v \in \mathfrak{g}.$$

See the proof of Proposition 18.2. This is because $d\mu_{(e,e)}$ is a homomorphism, and because $g \mapsto \mu(e, g)$ and $g \mapsto \mu(g, e)$ are the identity maps. As the map $g \mapsto \mu(g, \iota(g))$ is the constant map with value e, by differentiating and using the chain rule, we get

$$d\iota_e(u) = -u,$$

as desired. (Another proof makes use of the fact that for every $u \in \mathfrak{g}$, the integral curve γ through e with $\gamma'(0) = u$ is a group homomorphism. Therefore,

$$\iota(\gamma(t)) = \gamma(t)^{-1} = \gamma(-t),$$

and by differentiating, we get $d\iota_e(u) = -u$.)

(2) We follow Milnor [81] (Lemma 21.1). From (1), the map ι is an isometry, so by Proposition 17.3 (3), it preserves geodesics through e. Since $d\iota_e$ reverses $T_eG = \mathfrak{g}$, it reverses geodesics through e. Observe that

$$I_a = R_a \circ \iota \circ R_{a^{-1}},$$

so by (1), I_a is an isometry, and obviously $I_a(a) = a$. Again, by Proposition 17.3 (3), the isometry I_a preserve geodesics, and since R_a and $R_{a^{-1}}$ translate geodesics but ι reverses geodesics, it follows that I_a reverses geodesics.

(3) We follow Milnor [81] (Lemma 21.2). Assume γ is the unique geodesic through e such that $\gamma'(0) = u$, and let $X = u^L$ be the left invariant vector field such that $X(e) = u$. The first step is to prove that γ has domain $\mathbb{R}$ and that it is a group homomorphism; that is,

$$\gamma(s + t) = \gamma(s)\gamma(t).$$

Details of this argument are given in Milnor [81] (Lemma 20.1 and Lemma 21.2) and in Gallot, Hulin and Lafontaine [49] (Appendix B, Solution of Exercise 2.90). We present Milnor's proof.

Claim. The isometries I_a have the following property: For every geodesic ω through $p = \omega(0)$, if we let $q = \omega(r)$, then

$$I_q \circ I_p(\omega(t)) = \omega(t + 2r),$$

whenever $\omega(t)$ and $\omega(t + 2r)$ are defined.

Let $\alpha(t) = \omega(t+r)$. Then α is a geodesic with $\alpha(0) = q$. As I_p reverses geodesics through p (and similarly for I_q), we get

$$\begin{aligned} I_q \circ I_p(\omega(t)) &= I_q(\omega(-t)) \\ &= I_q(\alpha(-t-r)) \\ &= \alpha(t+r) = \omega(t+2r). \end{aligned}$$

It follows from the claim that ω can be indefinitely extended; that is, the domain of ω is $\mathbb{R}$.

Next we prove that γ is a homomorphism. By the claim, $I_{\gamma(t)} \circ I_e$ takes $\gamma(u)$ into $\gamma(u+2t)$. Now by definition of I_a and I_e,

$$I_{\gamma(t)} \circ I_e(a) = \gamma(t)a\gamma(t),$$

so, with $a = \gamma(u)$, we get

$$\gamma(t)\gamma(u)\gamma(t) = \gamma(u+2t).$$

By induction, it follows that

$$\gamma(nt) = \gamma(t)^n, \qquad \text{for all } n \in \mathbb{Z}.$$

We now use the (usual) trick of approximating every real by a rational number. For all $r, s \in \mathbb{R}$ with $s \neq 0$, if r/s is rational, say $r/s = m/n$ where m, n are integers, then $r = mt$ and $s = nt$ with $t = r/m = s/n$ and we get

$$\gamma(r+s) = \gamma(t)^{m+n} = \gamma(t)^m\gamma(t)^n = \gamma(r)\gamma(s).$$

Given any $t_1, t_2 \in \mathbb{R}$ with $t_2 \neq 0$, since t_1 and t_2 can be approximated by rationals r and s, as r/s is rational, $\gamma(r+s) = \gamma(r)\gamma(s)$, and by continuity, we get

$$\gamma(t_1 + t_2) = \gamma(t_1)\gamma(t_2),$$

as desired (the case $t_2 = 0$ is trivial as $\gamma(0) = e$).

As γ is a homomorphism, by differentiating the equation $\gamma(s+t) = \gamma(s)\gamma(t) = L_{\gamma(s)}\gamma(t)$, we get

$$\frac{d}{dt}(\gamma(s+t))|_{t=0} = (dL_{\gamma(s)})_e \left(\frac{d}{dt}(\gamma(t))|_{t=0} \right),$$

that is

$$\gamma'(s) = (dL_{\gamma(s)})_e(\gamma'(0)) = X(\gamma(s)),$$

which means that γ is the integral curve of the left-invariant vector field X, a one-parameter group.

Conversely, let c be the one-parameter group determined by a left-invariant vector field $X = u^L$, with $X(e) = u$ and let γ be the unique geodesic through e such that $\gamma'(0) = u$. Since we have just shown that γ is a homomorphism with $\gamma'(0) = u$, by uniqueness of one-parameter groups, $c = \gamma$; that is, c is a geodesic. $\square$

Remarks:

(1) As $R_g = \iota \circ L_{g^{-1}} \circ \iota$, we deduce that if G has a left-invariant metric, then this metric is also right-invariant iff ι is an isometry.

(2) Property (2) of Proposition 20.20 says that a Lie group with a bi-invariant metric is a *symmetric space*, an important class of Riemannian spaces invented and studied extensively by Élie Cartan. Symmetric spaces are briefly discussed in Section 22.8.

(3) The proof of 20.20 (3) given in O'Neill [91] (Chapter 11, equivalence of (5) and (6) in Proposition 9) appears to be missing the "hard direction," namely, that a geodesic is a one-parameter group. Also, since left and right translations are isometries and since isometries map geodesics to geodesics, the geodesics through any point $a \in G$ are the left (or right) translates of the geodesics through e, and thus are expressed in terms of the group exponential. Therefore, the geodesics through $a \in G$ are of the form

$$\gamma(t) = L_a(\exp_{\mathrm{gr}}(tu)),$$

where $u \in \mathfrak{g}$. Observe that $\gamma'(0) = (dL_a)_e(u)$.

(4) Some of the other facts stated in Proposition 20.19 and Proposition 20.20 are equivalent to the fact that a left-invariant metric is also bi-invariant; see O'Neill [91] (Chapter 11, Proposition 9).

Many more interesting results about left-invariant metrics on Lie groups can be found in Milnor's paper [84]. For example, flat left-invariant metrics on Lie a group are characterized (Theorem 1.5). We conclude this section by stating the following proposition (Milnor [84], Lemma 7.6).

Proposition 20.21. *If G is any compact, simple, Lie group, then the bi-invariant metric is unique up to a constant. Such a metric necessarily has constant Ricci curvature.*

20.5 Simple and Semisimple Lie Algebras and Lie Groups

In this section we introduce semisimple Lie algebras. They play a major role in the structure theory of Lie groups, but we only scratch the surface.

Definition 20.8. A Lie algebra $\mathfrak{g}$ is *simple* iff it is non-abelian and if it has **no** ideal other than (0) and $\mathfrak{g}$. A Lie algebra $\mathfrak{g}$ is *semisimple* iff it has **no abelian ideal other than** (0). A Lie group is *simple* (resp. *semisimple*) iff its Lie algebra is simple (resp. semisimple).

Clearly, the trivial subalgebras (0) and $\mathfrak{g}$ itself are ideals, and the center of a Lie algebra is an abelian ideal. It follows that the center $Z(\mathfrak{g})$ of a semisimple Lie algebra must be the trivial ideal (0).

Definition 20.9. Given two subsets $\mathfrak{a}$ and $\mathfrak{b}$ of a Lie algebra $\mathfrak{g}$, we let $[\mathfrak{a}, \mathfrak{b}]$ be the subspace of $\mathfrak{g}$ consisting of all linear combinations $[a, b]$, with $a \in \mathfrak{a}$ and $b \in \mathfrak{b}$.

If $\mathfrak{a}$ and $\mathfrak{b}$ are ideals in $\mathfrak{g}$, then $\mathfrak{a} + \mathfrak{b}$, $\mathfrak{a} \cap \mathfrak{b}$, and $[\mathfrak{a}, \mathfrak{b}]$, are also ideals (for $[\mathfrak{a}, \mathfrak{b}]$, use the Jacobi identity). The last fact allows us to make the following definition.

Definition 20.10. Let $\mathfrak{g}$ be a Lie algebra. The ideal $[\mathfrak{g}, \mathfrak{g}]$ is called the *commutator ideal* of $\mathfrak{g}$. The commutator ideal $[\mathfrak{g}, \mathfrak{g}]$ is also denoted by $D^1\mathfrak{g}$ (or $D\mathfrak{g}$).

If $\mathfrak{g}$ is a simple Lie agebra, then $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$ (because $[\mathfrak{g}, \mathfrak{g}]$ is an ideal, so the simplicity of $\mathfrak{g}$ implies that either $[\mathfrak{g}, \mathfrak{g}] = (0)$ or $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$. However, if $[\mathfrak{g}, \mathfrak{g}] = (0)$, then $\mathfrak{g}$ is abelian, a contradiction).

Definition 20.11. The *derived series* (or *commutator series*) $(D^k\mathfrak{g})$ of a Lie algebra (or ideal) $\mathfrak{g}$ is defined as follows:

$$\begin{aligned} D^0\mathfrak{g} &= \mathfrak{g} \\ D^{k+1}\mathfrak{g} &= [D^k\mathfrak{g}, D^k\mathfrak{g}], \quad k \geq 0. \end{aligned}$$

The first three $D^k\mathfrak{g}$ are

$$\begin{aligned} D^0\mathfrak{g} &= \mathfrak{g} \\ D^1\mathfrak{g} &= [\mathfrak{g}, \mathfrak{g}] \\ D^2\mathfrak{g} &= [D^1\mathfrak{g}, D^1\mathfrak{g}]. \end{aligned}$$

We have a decreasing sequence

$$\mathfrak{g} = D^0\mathfrak{g} \supseteq D^1\mathfrak{g} \supseteq D^2\mathfrak{g} \supseteq \cdots .$$

If $\mathfrak{g}$ is an ideal, by induction we see that each $D^k\mathfrak{g}$ is an ideal.

Definition 20.12. We say that a Lie algebra $\mathfrak{g}$ is *solvable* iff $D^k\mathfrak{g} = (0)$ for some $k \geq 0$.

If $\mathfrak{g}$ is abelian, then $[\mathfrak{g}, \mathfrak{g}] = 0$, so $\mathfrak{g}$ is solvable. Observe that a nonzero solvable Lie algebra has a nonzero abelian ideal, namely, the last nonzero $D^j\mathfrak{g}$. As a consequence, a Lie algebra is semisimple iff it has no nonzero solvable ideal.

It can be shown that every Lie algebra $\mathfrak{g}$ has a largest solvable ideal $\mathfrak{r}$, called the radical of $\mathfrak{g}$ (see Knapp [68], Chapter I, Proposition 1.12).

Definition 20.13. The *radical* of a Lie algebra $\mathfrak{g}$ is its largest solvable ideal, and it is denoted $\operatorname{rad}\mathfrak{g}$.

Then a Lie algebra is semisimple iff $\operatorname{rad}\mathfrak{g} = (0)$.

It can also be shown that for every (finite-dimensional) Lie algebra $\mathfrak{g}$, there is some semisimple Lie algebra $\mathfrak{s}$ such that $\mathfrak{g}$ is a semidirect product

$$\mathfrak{g} = \operatorname{rad}\mathfrak{g} \oplus_\tau \mathfrak{s}.$$

The above is called a *Levi decomposition*; see Knapp [68] (Appendix B), Serre [105] (Chapter VI, Theorem 4.1 and Corollary 1), and Fulton and Harris [46] (Appendix E). The Levi decomposition shows the importance of semisimple and solvable Lie algebras: the structure of these algebras determines the structure of arbitrary Lie algebras.

Definition 20.14. The *lower central series* $(C^k\mathfrak{g})$ of a Lie algebra (or ideal) $\mathfrak{g}$ is defined as follows:

$$\begin{aligned} C^0\mathfrak{g} &= \mathfrak{g} \\ C^{k+1}\mathfrak{g} &= [\mathfrak{g}, C^k\mathfrak{g}], \quad k \geq 0. \end{aligned}$$

We have a decreasing sequence

$$\mathfrak{g} = C^0\mathfrak{g} \supseteq C^1\mathfrak{g} \supseteq C^2\mathfrak{g} \supseteq \cdots .$$

Since $\mathfrak{g}$ is an ideal, by induction, each $C^k\mathfrak{g}$ is an ideal.

Definition 20.15. We say that an ideal $\mathfrak{g}$ is *nilpotent* iff $C^k\mathfrak{g} = (0)$ for some $k \geq 0$.

Proposition 20.22. *Every nilpotent Lie algebra is solvable.*

Proof. By induction, it is easy to show that

$$D^k\mathfrak{g} \subseteq C^k\mathfrak{g} \quad k \geq 0,$$

which immediately implies the proposition. □

Note that, by definition, simple and semisimple Lie algebras are non-abelian, and a simple algebra is a semisimple algebra. It turns out that a Lie algebra $\mathfrak{g}$ is semisimple iff it can be expressed as a direct sum of ideals $\mathfrak{g}_i$, with each $\mathfrak{g}_i$ a simple algebra (see Knapp [68], Chapter I, Theorem 1.54).

As a consequence we have the following result.

Proposition 20.23. *If $\mathfrak{g}$ is a semisimple Lie algebra, then $[\mathfrak{g}, \mathfrak{g}] = \mathfrak{g}$.*

Proof. If

$$\mathfrak{g} = \bigoplus_{i=1}^{m} \mathfrak{g}_i$$

where each $\mathfrak{g}_i$ is a simple ideal, then

$$[\mathfrak{g}, \mathfrak{g}] = \left[\bigoplus_{i=1}^{m} \mathfrak{g}_i, \bigoplus_{j=1}^{m} \mathfrak{g}_j\right] = \bigoplus_{i,j=1}^{m} [\mathfrak{g}_i, \mathfrak{g}_j] = \bigoplus_{i=1}^{m} [\mathfrak{g}_i, \mathfrak{g}_i] = \bigoplus_{i=1}^{m} \mathfrak{g}_i = \mathfrak{g},$$

since the $\mathfrak{g}_i$ being simple and forming a direct sum, $[\mathfrak{g}_i, \mathfrak{g}_j] = (0)$ whenever $i \neq j$ and $[\mathfrak{g}_i, \mathfrak{g}_i] = \mathfrak{g}_i$. □

If we drop the requirement that a simple Lie algebra be non-abelian, thereby allowing one dimensional Lie algebras to be simple, we run into the trouble that a simple Lie algebra is no longer semisimple, and the above theorem fails for this stupid reason. Thus, it seems technically advantageous to require that simple Lie algebras be non-abelian.

Nevertheless, in certain situations, it is desirable to drop the requirement that a simple Lie algebra be non-abelian and this is what Milnor does in his paper because it is more convenient for one of his proofs. This is a minor point but it could be confusing for uninitiated readers.

20.6 The Killing Form

The Killing form showed the tip of its nose in Proposition 20.19. It is an important concept, and in this section we establish some of its main properties. First we recall its definition.

Definition 20.16. For any Lie algebra $\mathfrak{g}$ over the field K (where $K = \mathbb{R}$ or $K = \mathbb{C}$), the *Killing form B of $\mathfrak{g}$* is the symmetric K-bilinear form $B\colon \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ given by

$$B(u, v) = \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)), \qquad \text{for all } u, v \in \mathfrak{g}.$$

If $\mathfrak{g}$ is the Lie algebra of a Lie group G, we also refer to B as the *Killing form of G*.

Remark: The *Killing form* as defined above is not due to Killing, and is closer to a variant due to Élie Cartan, as explained in Knapp [68] (page 754) and Armand Borel [17] (Chapter 1, §2), who claims to be responsible for this misnomer. On the other hand, the notion of "Cartan matrix" is due to Wilhelm Killing!

Example 20.1. For example, consider the group $\mathbf{SU}(2)$. Its Lie algebra $\mathfrak{su}(2)$ is the three-dimensional Lie algebra consisting of all skew-Hermitian 2×2 matrices with zero trace; that is, matrices of the form

$$X = \begin{pmatrix} ai & b + ic \\ -b + ic & -ai \end{pmatrix}, \qquad a, b, c \in \mathbb{R}.$$

Let

$$Y = \begin{pmatrix} di & e + if \\ -e + if & -di \end{pmatrix}, \qquad d, e, f \in \mathbb{R}.$$

By picking a suitable basis of $\mathfrak{su}(2)$, namely

$$e_1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad e_2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \qquad e_3 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

it can be shown that

$$\begin{aligned}\mathrm{ad}_X(e_1) &= L_X(e_1) - R_X(e_1)\\ &= \begin{pmatrix} ai & b+ic\\ -b+ic & -ai\end{pmatrix}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} - \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix} ai & b+ic\\ -b+ic & -ai\end{pmatrix}\\ &= \begin{pmatrix}-2ic & 2ia\\ 2ia & 2ic\end{pmatrix} = -2ce_3 + 2ae_2\\ \mathrm{ad}_X(e_2) &= L_X(e_2) - R_X(e_2) = \begin{pmatrix}2ib & -2a\\ 2a & -2ib\end{pmatrix} = -2ae_1 + 2be_3\\ \mathrm{ad}_X(e_3) &= L_X(e_3) - R_X(e_3) = \begin{pmatrix}0 & 2c-2ib\\ -2c-2ib & 0\end{pmatrix} = 2ce_1 - 2be_2,\end{aligned}$$

which in turn implies that

$$\mathrm{ad}_X = \begin{pmatrix}0 & -2a & 2c\\ 2a & 0 & -2b\\ -2c & 2b & 0\end{pmatrix}.$$

Similarly

$$\mathrm{ad}_Y = \begin{pmatrix}0 & -2d & 2f\\ 2d & 0 & -2e\\ -2f & 2e & 0\end{pmatrix}.$$

Thus

$$\begin{aligned}B(X,Y) = \mathrm{tr}(\mathrm{ad}_X \circ \mathrm{ad}_Y) &= \mathrm{tr}\begin{pmatrix}-4ad-4cf & 4ce & 4ae\\ 4bf & -4ad-4be & 4af\\ 4bd & 4cd & -4be-4cf\end{pmatrix}\\ &= -8ad - 8be - 8cf.\end{aligned}$$

However

$$\begin{aligned}\mathrm{tr}(XY) &= \mathrm{tr}\begin{pmatrix}-ad-cf-be+i(bf-ce) & -af+cd+i(ae-bd)\\ af-cd+i(ae-bd) & -ad-cf-be+i(-bf+ce)\end{pmatrix}\\ &= -2ad-2be-2cf.\end{aligned}$$

Hence

$$B(X,Y) = 4\mathrm{tr}(XY).$$

Example 20.2. Now if we consider the group $\mathbf{U}(2)$, its Lie algebra $\mathfrak{u}(2)$ is the four-dimensional Lie algebra consisting of all skew-Hermitian 2×2 matrices; that is, matrices of the form

$$\begin{pmatrix} ai & b+ic\\ -b+ic & id\end{pmatrix}, \qquad a,b,c,d \in \mathbb{R},$$

By using the basis

$$e_1 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad e_2 = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \qquad e_3 = \begin{pmatrix} i & 0 \\ 0 & 0 \end{pmatrix} \qquad e_4 = \begin{pmatrix} 0 & 0 \\ 0 & i \end{pmatrix},$$

it can be shown that

$$B(X,Y) = 4\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y).$$

Example 20.3. For $\mathbf{SO}(3)$, we know that $\mathfrak{so}(3) \cong \mathfrak{su}(2)$, and we get

$$B(X,Y) = \mathrm{tr}(XY).$$

Actually, the following proposition can be shown.

Proposition 20.24. *The following identities hold:*

$$\begin{aligned} \mathbf{GL}(n,\mathbb{R}), \mathbf{U}(n)\colon & \qquad B(X,Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y) \\ \mathbf{SL}(n,\mathbb{R}), \mathbf{SU}(n)\colon & \qquad B(X,Y) = 2n\mathrm{tr}(XY) \\ \mathbf{SO}(n)\colon & \qquad B(X,Y) = (n-2)\mathrm{tr}(XY). \end{aligned}$$

To prove Proposition 20.24, it suffices to compute the quadratic form $B(X,X)$, because $B(X,Y)$ is symmetric bilinear so it can be recovered using the polarization identity

$$B(X,Y) = \frac{1}{2}(B(X+Y,X+Y) - B(X,X) - B(Y,Y)).$$

Furthermore, if $\mathfrak{g}$ is the Lie algebra of a matrix group, since $\mathrm{ad}_X = L_X - R_X$ and L_X and R_X commute, for all $X, Z \in \mathfrak{g}$, we have

$$(\mathrm{ad}_X \circ \mathrm{ad}_X)(Z) = (L_X^2 - 2L_X \circ R_X + R_X^2)(Z) = X^2Z - 2XZX + ZX^2.$$

Therefore, to compute $B(X,X) = \mathrm{tr}(\mathrm{ad}_X \circ \mathrm{ad}_X)$, we can pick a convenient basis of $\mathfrak{g}$ and compute the diagonal entries of the matrix representing the linear map

$$Z \mapsto X^2Z - 2XZX + ZX^2.$$

Unfortunately, this is usually quite laborious. Some of the computations can be found in Jost [64] (Chapter 5, Section 5.5) and in Helgason [58] (Chapter III, §8).

Recall that a homomorphism of Lie algebras $\varphi\colon \mathfrak{g} \to \mathfrak{h}$ is a linear map that preserves brackets; that is,

$$\varphi([u,v]) = [\varphi(u), \varphi(v)].$$

Proposition 20.25. *The Killing form B of a Lie algebra $\mathfrak{g}$ has the following properties.*

(1) It is a symmetric bilinear form invariant under all automorphisms of $\mathfrak{g}$. In particular, if $\mathfrak{g}$ is the Lie algebra of a Lie group G, then B is Ad_g-invariant, for all $g \in G$; that is

$$B(\mathrm{Ad}_g(u), \mathrm{Ad}_g(v)) = B(u,v), \quad \textit{for all } u, v \in \mathfrak{g} \textit{ and all } g \in G.$$

(2) *The linear map* $\mathrm{ad}(u)$ *is skew-adjoint w.r.t* B *for all* $u \in \mathfrak{g}$*; that is,*

$$B(\mathrm{ad}(u)(v), w) = -B(v, \mathrm{ad}(u)(w)), \quad \textit{for all } u, v, w \in \mathfrak{g},$$

or equivalently,

$$B([u, v], w) = B(u, [v, w]), \quad \textit{for all } u, v, w \in \mathfrak{g}.$$

Proof. (1) The form B is clearly bilinear, and as $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, it is symmetric. If φ is an automorphism of $\mathfrak{g}$, the preservation of the bracket implies that

$$\mathrm{ad}(\varphi(u)) \circ \varphi = \varphi \circ \mathrm{ad}(u),$$

so

$$\mathrm{ad}(\varphi(u)) = \varphi \circ \mathrm{ad}(u) \circ \varphi^{-1}.$$

From $\mathrm{tr}(XY) = \mathrm{tr}(YX)$, we get $\mathrm{tr}(A) = \mathrm{tr}(BAB^{-1})$, so we get

$$\begin{aligned} B(\varphi(u), \varphi(v)) &= \mathrm{tr}(\mathrm{ad}(\varphi(u)) \circ \mathrm{ad}(\varphi(v)) \\ &= \mathrm{tr}(\varphi \circ \mathrm{ad}(u) \circ \varphi^{-1} \circ \varphi \circ \mathrm{ad}(v) \circ \varphi^{-1}) \\ &= \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v)) = B(u, v). \end{aligned}$$

Since Ad_g is an automorphism of $\mathfrak{g}$ for all $g \in G$, B is Ad_g-invariant.

(2) We have

$$B(\mathrm{ad}(u)(v), w) = B([u, v], w) = \mathrm{tr}(\mathrm{ad}([u, v]) \circ \mathrm{ad}(w))$$

and

$$B(v, \mathrm{ad}(u)(w)) = B(v, [u, w]) = \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}([u, w])).$$

However, the Jacobi identity is equivalent to

$$\mathrm{ad}([u, v]) = \mathrm{ad}(u) \circ \mathrm{ad}(v) - \mathrm{ad}(v) \circ \mathrm{ad}(u).$$

Consequently,

$$\begin{aligned} \mathrm{tr}(\mathrm{ad}([u, v]) \circ \mathrm{ad}(w)) &= \mathrm{tr}((\mathrm{ad}(u) \circ \mathrm{ad}(v) - \mathrm{ad}(v) \circ \mathrm{ad}(u)) \circ \mathrm{ad}(w)) \\ &= \mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v) \circ \mathrm{ad}(w)) - \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}(u) \circ \mathrm{ad}(w)) \end{aligned}$$

and

$$\begin{aligned} \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}([u, w])) &= \mathrm{tr}(\mathrm{ad}(v) \circ (\mathrm{ad}(u) \circ \mathrm{ad}(w) - \mathrm{ad}(w) \circ \mathrm{ad}(u))) \\ &= \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}(u) \circ \mathrm{ad}(w)) - \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}(w) \circ \mathrm{ad}(u)). \end{aligned}$$

As

$$\mathrm{tr}(\mathrm{ad}(u) \circ \mathrm{ad}(v) \circ \mathrm{ad}(w)) = \mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}(w) \circ \mathrm{ad}(u)),$$

we deduce that

$$B(\mathrm{ad}(u)(v), w) = \mathrm{tr}(\mathrm{ad}([u, v]) \circ \mathrm{ad}(w)) = -\mathrm{tr}(\mathrm{ad}(v) \circ \mathrm{ad}([u, w])) = -B(v, \mathrm{ad}(u)(w)),$$

as claimed. □

Remarkably, the Killing form yields a simple criterion due to Élie Cartan for testing whether a Lie algebra is semisimple. Recall that a symmetric bilinear form $\varphi\colon \mathfrak{g}\times\mathfrak{g}\to\mathbb{C}$ is nondegenerate if and only if for every $u\in\mathfrak{g}$, if $\varphi(u,v)=0$ for all $v\in\mathfrak{g}$, then $u=0$.

Theorem 20.26. *(Cartan's Criterion for Semisimplicity) A Lie algebra $\mathfrak{g}$ is semisimple iff its Killing form B is non-degenerate.*

As far as we know, all the known proofs of Cartan's criterion are quite involved. A fairly easy going proof can be found in Knapp [68] (Chapter 1, Theorem 1.45). A more concise proof is given in Serre [105] (Chapter VI, Theorem 2.1).

Since a Lie group with trivial Lie algebra is discrete, this implies that the center of a simple Lie group is discrete (because the Lie algebra of the center of a Lie group is the center of its Lie algebra. Prove it!).

We can also characterize which Lie groups have a Killing form which is negative definite.

Theorem 20.27. *A connected Lie group is compact and semisimple iff its Killing form is negative definite.*

Proof. First, assume that G is compact and semisimple. Then by Proposition 20.6, there is an inner product on $\mathfrak{g}$ inducing a bi-invariant metric on G, and by Proposition 20.7, every linear map $\mathrm{ad}(u)$ is skew-adjoint. Therefore, if we pick an orthonormal basis of $\mathfrak{g}$, the matrix X representing $\mathrm{ad}(u)$ is skew-symmetric, and

$$B(u,u)=\mathrm{tr}(\mathrm{ad}(u)\circ\mathrm{ad}(u))=\mathrm{tr}(XX)=\sum_{i,j=1}^{n}a_{ij}a_{ji}=-\sum_{i,j=1}^{n}a_{ij}^2\leq 0.$$

Since G is semisimple, Cartan's criterion implies that B is nondegenerate, and so it is negative definite.

Now assume that B is negative definite. If so, $-B$ is an inner product on $\mathfrak{g}$, and by Proposition 20.25, it is Ad-invariant. By Proposition 20.3, the inner product $-B$ induces a bi-invariant metric on G, and by Proposition 20.19 (d), the Ricci curvature is given by

$$\mathrm{Ric}(u,v)=-\frac{1}{4}\,B(u,v),$$

which shows that $r(u)>0$ for all units vectors $u\in\mathfrak{g}$. As in the proof of Proposition 20.16, there is some constant $c>0$, which is a lower bound on all Ricci curvatures $r(u)$, and by Myers' theorem (Theorem 16.28), G is compact (with finite fundamental group). By Cartan's criterion, as B is non-degenerate, G is also semisimple. □

Remark: A compact semisimple Lie group equipped with $-B$ as a metric is an Einstein manifold, since Ric is proportional to the metric (see Definition 16.8).

By using Theorems 20.26 and 20.27, since the Killing forms for $\mathbf{U}(n)$, $\mathbf{SU}(n)$ and $\mathbf{SO}(n)$ are given by

$$\begin{aligned} \mathbf{GL}(n,\mathbb{R}), \mathbf{U}(n)\colon &\qquad B(X,Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y) \\ \mathbf{SL}(n,\mathbb{R}), \mathbf{SU}(n)\colon &\qquad B(X,Y) = 2n\mathrm{tr}(XY) \\ \mathbf{SO}(n)\colon &\qquad B(X,Y) = (n-2)\mathrm{tr}(XY), \end{aligned}$$

we obtain the following result:

Proposition 20.28. *The Lie group $\mathbf{SU}(n)$ is compact and semisimple for $n \geq 2$, $\mathbf{SO}(n)$ is compact and semisimple for $n \geq 3$, and $\mathbf{SL}(n,\mathbb{R})$ is noncompact and semisimple for $n \geq 2$. However, $\mathbf{U}(n)$, even though it is compact, is not semisimple.*

Another way to determine whether a Lie algebra is semisimple is to consider reductive Lie algebras. We give a quick exposition without proofs. Details can be found in Knapp [68] (Chapter I, Sections, 7, 8).

Definition 20.17. A Lie algebra $\mathfrak{g}$ is *reductive* iff for every ideal $\mathfrak{a}$ in $\mathfrak{g}$, there is some ideal $\mathfrak{b}$ in $\mathfrak{g}$ such that $\mathfrak{g}$ is the direct sum

$$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{b}.$$

If $\mathfrak{g}$ is semisimple, we can pick $\mathfrak{b} = \mathfrak{a}^{\perp}$, the orthogonal complement of $\mathfrak{a}$ with respect to the Killing form of $\mathfrak{g}$. Therefore, every semisimple Lie algebra is reductive. More generally, if $\mathfrak{g}$ is the direct sum of a semisimple Lie algebra and an abelian Lie algebra, then $\mathfrak{g}$ is reductive. In fact, there are no other reductive Lie algebra. The following result is proved in Knapp [68] (Chapter I, Corollary 1.56).

Proposition 20.29. *If $\mathfrak{g}$ is a reductive Lie algebra, then*

$$\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}] \oplus Z(\mathfrak{g}),$$

with $[\mathfrak{g}, \mathfrak{g}]$ semisimple and $Z(\mathfrak{g})$ abelian.

Consequently, if $\mathfrak{g}$ is reductive, then it is semisimple iff its center $Z(\mathfrak{g})$ is trivial. For Lie algebras of matrices, a simple condition implies that a Lie algera is reductive. The following result is proved in Knapp [68] (Chapter I, Proposition 1.59).

Proposition 20.30. *If $\mathfrak{g}$ is a real Lie algebra of matrices over $\mathbb{R}$ or $\mathbb{C}$, and if $\mathfrak{g}$ is closed under conjugate transpose (that is, if $A \in \mathfrak{g}$, then $A^* \in \mathfrak{g}$), then $\mathfrak{g}$ is reductive.*

The familiar Lie algebras $\mathfrak{gl}(n,\mathbb{R})$, $\mathfrak{sl}(n,\mathbb{R})$, $\mathfrak{gl}(n,\mathbb{C})$, $\mathfrak{sl}(n,\mathbb{C})$, $\mathfrak{so}(n)$, $\mathfrak{so}(n,\mathbb{C})$, $\mathfrak{u}(n)$, $\mathfrak{su}(n)$, $\mathfrak{so}(p,q)$, $\mathfrak{u}(p,q)$, $\mathfrak{su}(p,q)$ are all closed under conjugate transpose. Among those, by computing their center, we obtain the following result:

Proposition 20.31. *The Lie algebra $\mathfrak{sl}(n, \mathbb{R})$ and $\mathfrak{sl}(n, \mathbb{C})$ are semisimple for $n \geq 2$, $\mathfrak{so}(n)$, $\mathfrak{so}(n, \mathbb{C})$ is semisimple for $n \geq 3$, $\mathfrak{su}(n)$ is semisimple for $n \geq 2$, $\mathfrak{so}(p, q)$ is semisimple for $p + q \geq 3$, and $\mathfrak{su}(p, q)$ is semisimple for $p + q \geq 2$.*

Semisimple Lie algebras and semisimple Lie groups have been investigated extensively, starting with the complete classification of the complex semisimple Lie algebras by Killing (1888) and corrected by Élie Cartan in his thesis (1894). One should read the Notes, especially on Chapter II, at the end of Knapp's book [68] for a fascinating account of the history of the theory of semisimple Lie algebras.

The theories and the body of results that emerged from these classification investigations play a very important role, not only in mathematics, but also in physics, and constitute one of the most beautiful chapters of mathematics. A quick introduction to these theories can be found in Arvanitoyeorgos [11] and in Carter, Segal, Macdonald [29]. A more comprehensive but yet still introductory presentation is given in Hall [56]. The most comprehensive treatment and yet accessible is probably Knapp [68]. The most encyclopedic but very abstract treatment is given in Bourbaki's nine volumes [19, 22, 23]. An older is classic is Helgason [58], which also discusses differential geometric aspects of Lie groups. Other "advanced" presentations can be found in Adams [3], Bröcker and tom Dieck [24], Serre [106, 105], Samelson [101], Humphreys [63], Fulton and Harris [46], and Kirillov [66]. A fascinating account of the history of Lie groups and Lie algebras is found in Armand Borel [17].

20.7 Left-Invariant Connections and Cartan Connections

Unfortunately, if a Lie group G does not admit a bi-invariant metric, under the Levi-Civita connection, geodesics are generally not given by the Lie group exponential map $\exp_{\mathrm{gr}} \colon \mathfrak{g} \to G$. If we are willing to consider connections not induced by a metric, then it turns out that there is a fairly natural connection for which the geodesics coincide with integral curves of left-invariant vector fields. These connections are called *Cartan connections.* Such connections are torsion-free (symmetric), but *the price that we pay is that in general they are not compatible with the chosen metric.* As a consequence, even though geodesics exist for all $t \in \mathbb{R}$, Hopf–Rinow's theorem fails; worse, it is generally false that any two points can be connected by a geodesic. This has to do with the failure of the exponential to be surjective. This section is heavily inspired by Postnikov [96] (Chapter 6, Sections 3–6); see also Kobayashi and Nomizu [69] (Chapter X, Section 2).

Recall that a vector field X on a Lie group G is left-invariant if the following diagram commutes for all $a \in G$:

$$\begin{tikzcd}
TG \arrow[r, "d(L_a)"] & TG \\
G \arrow[u, "X"] \arrow[r, "L_a"'] & G \arrow[u, "X"']
\end{tikzcd}$$

In this section we use freely the fact that there is an isomorphism between the Lie algebra $\mathfrak{g}$ and the Lie algebra $\mathfrak{g}^L$ of left-invariant vector fields on G. For every $X \in \mathfrak{g}$, we denote by $X^L \in \mathfrak{g}^L$ the unique left-invariant vector field such that $X^L_1 = X$.

Definition 20.18. A connection ∇ on a Lie group G is *left-invariant* if for any two left-invariant vector fields X^L, Y^L with $X, Y \in \mathfrak{g}$, the vector field $\nabla_{X^L} Y^L$ is also left-invariant.

By analogy with left-invariant metrics, there is a version of Proposition 20.1 stating that there is a one-to-one correspondence between left-invariant connections and bilinear maps $\alpha\colon \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. This is shown as follows.

Given a left-invariant connection ∇ on G, we get the map $\alpha\colon \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ given by

$$\alpha(X, Y) = (\nabla_{X^L} Y^L)_1, \quad X, Y \in \mathfrak{g}.$$

To define a map in the opposite direction, pick any basis $X_1, \ldots, X_n$ of $\mathfrak{g}$. Then every vector field X on G can be written as

$$X = f_1 X_1^L + \cdots + f_n X_n^L,$$

for some smooth functions $f_1, \ldots, f_n$ on G. If ∇ is a left-invariant connection on G, for any left-invariant vector fields $X = \sum_{i=1}^n f_i X_i^L$ and $Y = \sum_{j=1}^n g_j X_j^L$, we have

$$\begin{aligned}
\nabla_X Y &= \nabla_{\sum_{i=1}^n f_i X_i^L} Y = \sum_{i=1}^n f_i \nabla_{X_i^L} Y \\
&= \sum_{i=1}^n f_i \nabla_{X_i^L} \sum_{i=1}^n g_j X_j^L \\
&= \sum_{i,j=1}^n f_i \big((X_i^L g_j) X_j^L + g_j \nabla_{X_i^L} X_j^L \big).
\end{aligned}$$

This shows that ∇ is completely determined by the matrix with entries

$$\alpha_{ij} = \alpha(X_i, X_j) = (\nabla_{X_i^L} X_j^L)_1.$$

Conversely, any bilinear map α on $\mathfrak{g}$ is determined by the matrix (α_{ij}) with $\alpha_{ij} = \alpha(X_i, X_j) \in \mathfrak{g}$, and it is immediately checked that Formula (†) shown below

$$\nabla_X Y = \sum_{i,j=1}^n f_i \big((X_i^L g_j) X_j^L + g_j \alpha_{ij} \big), \tag{†}$$

defines a left-invariant connection such that $(\nabla_{X_i^L} X_j^L)_1 = \alpha_{ij}$ for $i, j = 1, \ldots, n$. In summary, we proved the following result.

Proposition 20.32. *There is a one-to-one correspondence between left-invariant connections on G and bilinear maps $\alpha\colon \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$.*

Let us now investigate the conditions under which the geodesic curves coincide with the integral curves of left-invariant vector fields. Let X^L be any left-invariant vector field, and let γ be the integral curve such that $\gamma(0) = 1$ and $\gamma'(0) = X$ (in other words, $\gamma(t) = \exp_{\mathrm{gr}}(tX) = e^{tX}$). Since the vector field $t \mapsto \gamma'(t)$ along γ is the restriction of the vector field X^L, we have

$$\frac{D}{dt}(\gamma'(t)) = (\nabla_{X^L} X^L)_{\gamma(t)} = \alpha(X, X)^L_{\gamma(t)}, \quad \text{for all } t \in \mathbb{R}.$$

Since a left-invariant vector field is determined by its value at 1, and γ is a geodesic iff $\frac{D\gamma'}{dt} = 0$, we have $(\nabla_{X^L} X^L)_{\gamma(t)} = 0$ for all $t \in \mathbb{R}$ iff

$$\alpha(X, X) = 0.$$

Every bilinear map α can be written as the sum of a symmetric bilinear map

$$\alpha_H(X, Y) = \frac{\alpha(X, Y) + \alpha(Y, X)}{2}$$

and a skew-symmetric bilinear map

$$\alpha_S(X, Y) = \frac{\alpha(X, Y) - \alpha(Y, X)}{2},$$

Clearly $\alpha_S(X, X) = 0$. Thus $\alpha(X, X) = 0$ implies that $\alpha_H(X, X) = 0$. Hence we conclude that for every $X \in \mathfrak{g}$, the curve $t \mapsto e^{tX}$ is a geodesic iff α is skew-symmetric.

Proposition 20.33. *The left-invariant connection ∇ induced by a bilinear map α on $\mathfrak{g}$ has the property that, for every $X \in \mathfrak{g}$, the curve $t \mapsto \exp_{\mathrm{gr}}(tX) = e^{tX}$ is a geodesic iff α is skew-symmetric.*

Definition 20.19. A left-invariant connection satisfying the property that for every $X \in \mathfrak{g}$, the curve $t \mapsto e^{tX}$ is a geodesic, is called a *Cartan connection.*

It is easy to find out when the Cartan connection ∇ associated with a bilinear map α on $\mathfrak{g}$ is torsion-free (symmetric).

Proposition 20.34. *The Cartan connection ∇ associated with a bilinear map α on $\mathfrak{g}$ is torsion-free (symmetric) iff*

$$\alpha_S(X, Y) = \frac{1}{2}[X, Y], \quad \textit{for all } X, Y \in \mathfrak{g},$$

Proof. In order for the connection ∇ to be torsion-free we must have

$$\nabla_{X^L} Y^L - \nabla_{Y^L} X^L = [X, Y]^L, \quad \text{for all } X, Y \in \mathfrak{g}.$$

that is,

$$\alpha(X, Y) - \alpha(Y, X) = [X, Y], \quad \text{for all } X, Y \in \mathfrak{g}.$$

so we deduce that the Cartan connection induced by α is torsion-free iff

$$\alpha_S(X, Y) = \frac{1}{2}[X, Y], \quad \text{for all } X, Y \in \mathfrak{g}. \qquad \square$$

In view of Proposition 20.34, we have the following fact.

Proposition 20.35. *Given any Lie group G, there is a unique torsion-free (symmetric) Cartan connection ∇ given by*

$$\nabla_{X^L} Y^L = \frac{1}{2}[X, Y]^L, \quad \textit{for all } X, Y \in \mathfrak{g}.$$

Then the same calculation that we used in the case of a bi-invariant metric on a Lie group shows that the curvature tensor is given by

$$R(X, Y)Z = \frac{1}{4}[[X, Y], Z], \quad \text{for all } X, Y, Z \in \mathfrak{g}.$$

The following fact is easy to check.

Proposition 20.36. *For any $X \in \mathfrak{g}$ and any point $a \in G$, the unique geodesic $\gamma_{a,X}$ such that $\gamma_{a,X}(0) = a$ and $\gamma'_{a,X}(0) = X$, is given by*

$$\gamma_{a,X}(t) = e^{td(R_{a^{-1}})_a X} a;$$

that is,

$$\gamma_{a,X} = R_a \circ \gamma_{d(R_{a^{-1}})_a X},$$

where $\gamma_{d(R_{a^{-1}})_a X}(t) = e^{td(R_{a^{-1}})_a X}$.

Remark: Observe that the bilinear maps given by

$$\alpha(X, Y) = \lambda[X, Y] \quad \text{for some } \lambda \in \mathbb{R}$$

are skew-symmetric, and thus induce Cartan connections. Let us show that the torsion is given by

$$T(X, Y) = (2\lambda - 1)[X, Y],$$

and the curvature by

$$R(X, Y)Z = \lambda(1 - \lambda)[[X, Y], Z].$$

For the torsion, we have

$$\begin{aligned} T(X,Y) &= \nabla_X Y - \nabla_Y X - [X,Y] \\ &= \alpha(X,Y) - \alpha(Y,X) - [X,Y] \\ &= \lambda[X,Y] - \lambda[Y,X] - [X,Y] \\ &= (2\lambda - 1)[X,Y]. \end{aligned}$$

For the curvature, we get

$$\begin{aligned} R(X,Y)Z &= \nabla_{[X,Y]}Z + \nabla_Y\nabla_X Z - \nabla_X\nabla_Y Z \\ &= \alpha([X,Y],Z) + \nabla_Y\alpha(X,Z) - \nabla_X\alpha(Y,Z) \\ &= \lambda[[X,Y],Z] + \lambda\nabla_Y[X,Z] - \lambda\nabla_X[Y,Z] \\ &= \lambda[[X,Y],Z] + \lambda\alpha(Y,[X,Z]) - \lambda\alpha(X,[Y,Z]) \\ &= \lambda[[X,Y],Z] + \lambda^2([Y,[X,Z]] - [X,[Y,Z]]). \end{aligned}$$

The Jacobi identity

$$[X,[Y,Z]] + [Z,[X,Y]] + [Y,[Z,X]] = 0$$

yields

$$[X,[Y,Z]] - [Y,[X,Z]] = -[Z,[X,Y]] = [[X,Y],Z],$$

so we get

$$[Y,[X,Z]] - [X,[Y,Z]] = -[[X,Y],Z],$$

and thus

$$R(X,Y)Z = \lambda[[X,Y],Z] + \lambda^2([Y,[X,Z]] - [X,[Y,Z]]) = \lambda(1-\lambda)[[X,Y],Z].$$

It follows that for $\lambda = 0$ and $\lambda = 1$, we get connections where the curvature vanishes. However, these connections have torsion. Again, we see that $\lambda = 1/2$ is the only value for which the Cartan connection is symmetric.

In the case of a bi-invariant metric, the *Levi-Civita connection coincides with the Cartan connection*.

20.8 Problems

Problem 20.1. Prove Proposition 20.11.
Hint. See Milnor [84] (Lemma 1.1).

Problem 20.2. Consider the Lie group $\mathbf{SO}(n)$ with the bi-invariant metric induced by the inner product on $\mathfrak{so}(n)$ given by

$$\langle B_1, B_2\rangle = \frac{1}{2}\mathrm{tr}(B_1^\top B_2).$$

For any two matrices $B_1, B_2 \in \mathfrak{so}(n)$, let γ be the curve given by

$$\gamma(t) = e^{(1-t)B_1 + tB_2}, \quad 0 \leq t \leq 1.$$

This is a curve "interpolating" between the two rotations $R_1 = e^{B_1}$ and $R_2 = e^{B_2}$.

(1) Prove that the length $L(\gamma)$ of the curve γ is given by

$$L(\gamma) = \left(-\frac{1}{2}\mathrm{tr}((B_2 - B_1)^2)\right)^{\frac{1}{2}}.$$

(2) We know that the geodesic from R_1 to R_2 is given by

$$\gamma_g(t) = R_1 e^{tB}, \quad 0 \leq t \leq 1,$$

where $B \in \mathfrak{so}(n)$ is the principal log of $R_1^\top R_2$ (if we assume that $R_1^\top R_2$ is not a rotation by π, i.e, does not admit -1 as an eigenvalue).

Conduct numerical experiments to verify that in general, $\gamma(1/2) \neq \gamma_g(1/2)$.

Problem 20.3. Consider the set of affine maps ρ of $\mathbb{R}^3$ defined such that

$$\rho(X) = \alpha RX + W,$$

where R is a rotation matrix (an orthogonal matrix of determinant $+1$), W is some vector in $\mathbb{R}^3$, and $\alpha \in \mathbb{R}$ with $\alpha > 0$. Every such a map can be represented by the 4×4 matrix

$$\begin{pmatrix} \alpha R & W \\ 0 & 1 \end{pmatrix}$$

in the sense that

$$\begin{pmatrix} \rho(X) \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha R & W \\ 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix}$$

iff

$$\rho(X) = \alpha RX + W.$$

(a) Prove that these maps form a group, denoted by $\mathbf{SIM}(3)$ (the *direct affine similitudes* of $\mathbb{R}^3$).

(b) Let us now consider the set of 4×4 real matrices of the form

$$B = \begin{pmatrix} \Gamma & W \\ 0 & 0 \end{pmatrix},$$

where Γ is a matrix of the form

$$\Gamma = \lambda I_3 + \Omega,$$

with

$$\Omega = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

so that

$$\Gamma = \begin{pmatrix} \lambda & -c & b \\ c & \lambda & -a \\ -b & a & \lambda \end{pmatrix},$$

and W is a vector in $\mathbb{R}^3$.

Verify that this set of matrices is a vector space isomorphic to $(\mathbb{R}^7, +)$. This vector space is denoted by $\mathfrak{sim}(3)$.

(c) Given a matrix

$$B = \begin{pmatrix} \Gamma & W \\ 0 & 0 \end{pmatrix}$$

as in (b), prove that

$$B^n = \begin{pmatrix} \Gamma^n & \Gamma^{n-1} W \\ 0 & 0 \end{pmatrix}$$

where $\Gamma^0 = I_3$. Prove that

$$e^B = \begin{pmatrix} e^\Gamma & VW \\ 0 & 1 \end{pmatrix},$$

where

$$V = I_3 + \sum_{k \geq 1} \frac{\Gamma^k}{(k+1)!}.$$

(d) Prove that if $\Gamma = \lambda I_3 + \Omega$ as in (b), then

$$V = I_3 + \sum_{k \geq 1} \frac{\Gamma^k}{(k+1)!} = \int_0^1 e^{\Gamma t} dt.$$

(e) For any matrix $\Gamma = \lambda I_3 + \Omega$, with

$$\Omega = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

if we let $\theta = \sqrt{a^2 + b^2 + c^2}$, then prove that

$$e^\Gamma = e^\lambda e^\Omega = e^\lambda \left(I_3 + \frac{\sin\theta}{\theta} \Omega + \frac{(1 - \cos\theta)}{\theta^2} \Omega^2 \right), \quad \text{if } \theta \neq 0,$$

and $e^\Gamma = e^\lambda I_3$ if $\theta = 0$.

Hint. You may use the fact that if $AB = BA$, then $e^{A+B} = e^A e^B$. In general, $e^{A+B} \neq e^A e^B$!

(f) Prove that

1. If $\theta = 0$ and $\lambda = 0$, then
$$V = I_3.$$

2. If $\theta = 0$ and $\lambda \neq 0$, then
$$V = \frac{(e^{\lambda} - 1)}{\lambda} I_3;$$

3. If $\theta \neq 0$ and $\lambda = 0$, then
$$V = I_3 + \frac{(1 - \cos\theta)}{\theta^2}\Omega + \frac{(\theta - \sin\theta)}{\theta^3}\Omega^2.$$

4. If $\theta \neq 0$ and $\lambda \neq 0$, then
$$\begin{aligned} V = {} & \frac{(e^{\lambda} - 1)}{\lambda} I_3 + \frac{(\theta(1 - e^{\lambda}\cos\theta) + e^{\lambda}\lambda\sin\theta)}{\theta(\lambda^2 + \theta^2)}\Omega \\ & + \left(\frac{(e^{\lambda} - 1)}{\lambda\theta^2} - \frac{e^{\lambda}\sin\theta}{\theta(\lambda^2 + \theta^2)} - \frac{\lambda(e^{\lambda}\cos\theta - 1)}{\theta^2(\lambda^2 + \theta^2)}\right)\Omega^2. \end{aligned}$$

Hint. You will need to compute $\int_0^1 e^{\lambda t} \sin\theta t \, dt$ and $\int_0^1 e^{\lambda t} \cos\theta t \, dt$.

(g) Prove that V is invertible iff $\lambda \neq 0$ or $\theta \neq k2\pi$, with $k \in \mathbb{Z} - \{0\}$.

Hint. Express the eigenvalues of V in terms of the eigenvalues of Γ.

In the special case where $\lambda = 0$, show that
$$V^{-1} = I - \frac{1}{2}\Omega + \frac{1}{\theta^2}\left(1 - \frac{\theta\sin\theta}{2(1 - \cos\theta)}\right)\Omega^2, \quad \text{if } \theta \neq 0.$$

Hint. Assume that the inverse of V is of the form
$$Z = I_3 + a\Omega + b\Omega^2,$$
and show that a, b, are given by a system of linear equations that always has a unique solution.

(h) Prove that the exponential map $\exp\colon \mathfrak{sim}(3) \to \mathbf{SIM}(3)$, given by $\exp(B) = e^B$, is surjective. You may use the fact that $\exp\colon \mathfrak{so}(3) \to \mathbf{SO}(3)$ is surjective, proved in Problem 1.4.

Problem 20.4. Refer to Problem 20.3. Similitudes can be used to describe certain deformations (or flows) of a deformable body $\mathcal{B}_t$ in 3D. Given some initial shape $\mathcal{B}$ in $\mathbb{R}^3$ (for example, a sphere, a cube, etc.), a deformation of $\mathcal{B}$ is given by a piecewise differentiable curve
$$\mathcal{D}\colon [0, T] \to \mathbf{SIM}(3),$$

where each $\mathcal{D}(t)$ is a similitude (for some $T > 0$). The deformed body $\mathcal{B}_t$ at time t is given by

$$\mathcal{B}_t = \mathcal{D}(t)(\mathcal{B}),$$

where $\mathcal{D}(t) \in \mathbf{SIM}(3)$ is a similitude.

The surjectivity of the exponential map $\exp\colon \mathfrak{sim}(3) \to \mathbf{SIM}(3)$ implies that there is a map $\log\colon \mathbf{SIM}(3) \to \mathfrak{sim}(3)$, although it is multivalued. The exponential map and the log "function" allows us to work in the simpler (noncurved) Euclidean space $\mathfrak{sim}(3)$ (which has dimension 7).

For instance, given two similitudes $A_1, A_2 \in \mathbf{SIM}(3)$ specifying the shape of $\mathcal{B}$ at two different times, we can compute $\log(A_1)$ and $\log(A_2)$, which are just elements of the Euclidean space $\mathfrak{sim}(3)$, form the linear interpolant $(1-t)\log(A_1) + t\log(A_2)$, and then apply the exponential map to get an interpolating deformation

$$t \mapsto e^{(1-t)\log(A_1)+t\log(A_2)}, \quad t \in [0,1].$$

Also, given a sequence of "snapshots" of the deformable body $\mathcal{B}$, say $A_0, A_1, \ldots, A_m$, where each is A_i is a similitude, we can try to find an interpolating deformation (a curve in $\mathbf{SIM}(3)$) by finding a simpler curve $t \mapsto C(t)$ in $\mathfrak{sim}(3)$ (say, a B-spline) interpolating $\log A_1, \log A_1, \ldots, \log A_m$. Then, the curve $t \mapsto e^{C(t)}$ yields a deformation in $\mathbf{SIM}(3)$ interpolating $A_0, A_1, \ldots, A_m$.

(1) Write a program interpolating between two deformations, using the formulae found in Problem 20.3. (not the built-in `Matlab` functions!).

(2) Write a program using cubic spline interpolation program to interpolate a sequence of deformations given by similitudes $A_0, A_1, \ldots, A_m$ in $\mathbf{SIM}(3)$. Use the formulae found in Problem 20.3 (not the built-in `Matlab` functions!).

Problem 20.5. Prove that if $\mathfrak{g}$ is an ideal, then each $D^k\mathfrak{g}$ is an ideal.

Problem 20.6. Given a finite dimensional Lie algebra $\mathfrak{g}$ (as a vector space over $\mathbb{R}$), we define the function $B\colon \mathfrak{g} \times \mathfrak{g} \to \mathbb{C}$ (Killing form) by

$$B(X,Y) = \mathrm{tr}(\mathrm{ad}(X) \circ \mathrm{ad}(Y)), \quad X, Y \in \mathfrak{g}.$$

(1) Check that B is $\mathbb{R}$-bilinear and symmetric.

(2) Let $\mathfrak{g} = \mathfrak{gl}(2,\mathbb{R}) = \mathrm{M}_2(\mathbb{R})$. Given any matrix $A \in \mathrm{M}_2(\mathbb{R})$ with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

show that in the basis $(E_{12}, E_{11}, E_{22}, E_{21})$, the matrix of $\mathrm{ad}(A)$ is given by

$$\begin{pmatrix} a-d & -b & b & 0 \\ -c & 0 & 0 & b \\ c & 0 & 0 & -b \\ 0 & c & -c & d-a \end{pmatrix}.$$

Show that

$$\det(xI - \mathrm{ad}(A)) = x^2(x^2 - ((a-d)^2 + 4bc)).$$

(3) Given $A, A' \in \mathrm{M}_2(\mathbb{R})$ with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad A' = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix},$$

prove that

$$B(A, A') = 2(d-a)(d'-a') + 4bc' + 4cb' = 4\mathrm{tr}(AA') - 2\mathrm{tr}(A)\mathrm{tr}(A').$$

(4) Next, let $\mathfrak{g} = \mathfrak{sl}(2, \mathbb{R})$. Check that the following three matrices form a basis of $\mathfrak{sl}(2, \mathbb{R})$:

$$H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Prove that in the basis (H, X, Y), for any

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \in \mathfrak{sl}(2, \mathbb{R}),$$

the matrix of $\mathrm{ad}(A)$ is

$$\begin{pmatrix} 0 & -c & b \\ -2b & 2a & 0 \\ 2c & 0 & -2a \end{pmatrix}.$$

Prove that

$$\det(xI - \mathrm{ad}(A)) = x(x^2 - 4(a^2 + bc)).$$

(5) Given $A, A' \in \mathfrak{sl}(2, \mathbb{R})$ with

$$A = \begin{pmatrix} a & b \\ c & -a \end{pmatrix}, \quad A' = \begin{pmatrix} a' & b' \\ c' & -a'' \end{pmatrix},$$

prove that

$$B(A, A') = 8aa' + 4bc' + 4cb' = 4\mathrm{tr}(AA').$$

(6) Let $\mathfrak{g} = \mathfrak{so}(3)$. For any $A \in \mathfrak{so}(3)$, with

$$A = \begin{pmatrix} 0 & -c & b \\ c & 0 & -a \\ -b & a & 0 \end{pmatrix},$$

we know from Proposition 2.39 that in the basis (E_1, E_2, E_3), the matrix of $\mathrm{ad}(A)$ is A itself. Prove that

$$B(A, A') = -2(aa' + bb' + cc') = \mathrm{tr}(AA').$$

(7) Recall that a symmetric bilinear form B is *nondegenerate* if for every X, if $B(X, Y) = 0$ for all Y, then $X = 0$.

Prove that B on $\mathfrak{gl}(2, \mathbb{R}) = \mathrm{M}_2(\mathbb{R})$ is degenerate; B on $\mathfrak{sl}(2, \mathbb{R})$ is nondegenerate but neither positive definite nor negative definite; B on $\mathfrak{so}(3)$ is nondegenerate negative definite.

(8) Recall that a subspace $\mathfrak{h}$ of a Lie algebra $\mathfrak{g}$ is a *subalgebra* of $\mathfrak{g}$ if $[x, y] \in \mathfrak{h}$ for all $x, y \in \mathfrak{h}$, and an *ideal* if $[h, x] \in \mathfrak{h}$ for all $h \in \mathfrak{h}$ and all $x \in \mathfrak{g}$. Check that $\mathfrak{sl}(n, \mathbb{R})$ is an ideal in $\mathfrak{gl}(n, \mathbb{R})$, and that $\mathfrak{so}(n)$ is a subalgebra of $\mathfrak{sl}(n, \mathbb{R})$, but not an ideal. Prove that if $\mathfrak{h}$ is an ideal in $\mathfrak{g}$, then the bilinear form B on $\mathfrak{h}$ is equal to the restriction of the bilinear form B on $\mathfrak{g}$ to $\mathfrak{h}$.

Prove the following facts: for all $n \geq 2$:

$$\begin{aligned}
\mathfrak{gl}(n, \mathbb{R})\colon &\qquad B(X, Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y) \\
\mathfrak{sl}(n, \mathbb{R})\colon &\qquad B(X, Y) = 2n\mathrm{tr}(XY) \\
\mathfrak{so}(n)\colon &\qquad B(X, Y) = (n-2)\mathrm{tr}(XY).
\end{aligned}$$

Problem 20.7. Given a group G, recall that its *center* is the subset

$$Z(G) = \{a \in G,\ ag = ga \quad \text{for all } g \in G\}.$$

(1) Check that $Z(G)$ is a commutative normal subgroup of G.

(2) Prove that a matrix $A \in \mathrm{M}_n(\mathbb{R})$ commutes with all matrices $B \in \mathbf{GL}(n, \mathbb{R})$ iff $A = \lambda I$ for some $\lambda \in \mathbb{R}$.

Hint. Remember the elementary matrices.

Prove that

$$Z(\mathbf{GL}(n, \mathbb{R})) = \{\lambda I \mid \lambda \in \mathbb{R}, \lambda \neq 0\}.$$

(3) Prove that for any $m \geq 1$,

$$\begin{aligned}
Z(\mathbf{SO}(2(m+1))) &= \{I, -I\} \\
Z(\mathbf{SO}(2m-1)) &= \{I\} \\
Z(\mathbf{SL}(m, \mathbb{R})) &= \{\lambda I \mid \lambda \in \mathbb{R}, \lambda^m = 1\}.
\end{aligned}$$

(4) Prove that a matrix $A \in \mathrm{M}_n(\mathbb{C})$ commutes with all matrices $B \in \mathbf{GL}(n, \mathbb{C})$ iff $A = \lambda I$ for some $\lambda \in \mathbb{C}$.

(5) Prove that for any $n \geq 1$,

$$\begin{aligned}
Z(\mathbf{GL}(n, \mathbb{C})) &= \{\lambda I \mid \lambda \in \mathbb{C}, \lambda \neq 0\} \\
Z(\mathbf{SL}(n, \mathbb{C})) &= \{e^{\frac{k2\pi}{n}i} I \mid k = 0, 1, \ldots, n-1\} \\
Z(\mathbf{U}(n)) &= \{e^{i\theta} I \mid 0 \leq \theta < 2\pi\} \\
Z(\mathbf{SU}(n)) &= \{e^{\frac{k2\pi}{n}i} I \mid k = 0, 1, \ldots, n-1\}.
\end{aligned}$$

(6) Prove that the groups $\mathbf{SO}(3)$ and $\mathbf{SU}(2)$ are not isomorphic (although their Lie algebras *are* isomorphic).

Problem 20.8. Consider a finite dimensional Lie algebra $\mathfrak{g}$, but this time a vector space over $\mathbb{C}$, and define the function $B\colon \mathfrak{g}\times\mathfrak{g}\to\mathbb{C}$ by

$$B(x,y) = \mathrm{tr}(\mathrm{ad}(x)\circ\mathrm{ad}(y)), \quad x,y\in\mathfrak{g}.$$

The bilinear form B is called the *Killing form* of $\mathfrak{g}$. Recall that a *homomorphism* $\varphi\colon\mathfrak{g}\to\mathfrak{g}$ is a linear map such that $\varphi([x,y]) = [\varphi(x),\varphi(y)]$ for all $x,y\in\mathfrak{g}$, or equivalently such that

$$\varphi\circ\mathrm{ad}(x) = \mathrm{ad}(\varphi(x))\circ\varphi, \quad \text{for all } x\in\mathfrak{g},$$

and that an *automorphism* of $\mathfrak{g}$ is a homomorphism of $\mathfrak{g}$ that has an inverse which is also a homomorphism of $\mathfrak{g}$.

(1) Prove that for every automorphism $\varphi\colon\mathfrak{g}\to\mathfrak{g}$, we have

$$B(\varphi(x),\varphi(y)) = B(x,y), \quad \text{for all } x,y\in\mathfrak{g}.$$

Prove that for all $x,y,z\in\mathfrak{g}$, we have

$$B(\mathrm{ad}(x)(y),z) = -B(y,\mathrm{ad}(x)(z)),$$

or equivalently

$$B([y,x],z) = B(y,[x,z]).$$

(2) Review the primary decomposition theorem. For any $x\in\mathfrak{g}$, we can apply the primary decomposition theorem to the linear map $\mathrm{ad}(x)$. Write

$$m(X) = (X-\lambda_1)^{r_1}\cdots(X-\lambda_k)^{r_k}$$

for the minimal polynomial of $\mathrm{ad}(x)$, where $\lambda_1,\ldots,\lambda_k$ are the distinct eigenvalues of $\mathrm{ad}(x)$, and let

$$\mathfrak{g}_x^{\lambda_i} = \mathrm{Ker}\,(\lambda_i I - \mathrm{ad}(x))^{r_i}, \quad i=1,\ldots,k.$$

We know that 0 is an eigenvalue of $\mathrm{ad}(x)$, and we agree that $\lambda_0 = 0$. Then, we have a direct sum

$$\mathfrak{g} = \bigoplus_{\lambda_i}\mathfrak{g}_x^{\lambda_i}.$$

It is convenient to define $\mathfrak{g}_x^{\lambda}$ when λ is not an eigenvalue of $\mathrm{ad}(x)$ as

$$\mathfrak{g}_x^{\lambda} = (0).$$

Prove that

$$[\mathfrak{g}_x^{\lambda},\mathfrak{g}_x^{\mu}] \subseteq \mathfrak{g}_x^{\lambda+\mu}, \quad \text{for all } \lambda,\mu\in\mathbb{C}.$$

Hint. First, show that

$$((\lambda+\mu)I - \mathrm{ad}(x))[y,z] = [(\lambda I - \mathrm{ad}(x))(y), z] + [y, (\mu I - \mathrm{ad}(x))(z)],$$

for all $x, y, z \in \mathfrak{g}$, and then that

$$((\lambda+\mu)I - \mathrm{ad}(x))^n[y,z] = \sum_{p=0}^{n} \binom{n}{p} [(\lambda I - \mathrm{ad}(x))^p(y), (\mu I - \mathrm{ad}(x))^{n-p}(z)],$$

by induction on n.

Prove that $\mathfrak{g}_x^0$ is a Lie subalgebra of $\mathfrak{g}$.

(3) Prove that if $\lambda + \mu \neq 0$, then $\mathfrak{g}_x^\lambda$ and $\mathfrak{g}_x^\mu$ are orthogonal with respect to B (which means that $B(X,Y) = 0$ for all $X \in \mathfrak{g}_x^\lambda$ and all $Y \in \mathfrak{g}_x^\mu$).

Hint. For any $X \in \mathfrak{g}_x^\lambda$ and any $Y \in \mathfrak{g}_x^\mu$, prove that $\mathrm{ad}(X) \circ \mathrm{ad}(Y)$ is nilpotent. Note that for any ν and any $Z \in \mathfrak{g}_x^\nu$,

$$(\mathrm{ad}(X) \circ \mathrm{ad}(Y))(Z) = [X, [Y, Z]],$$

so by (2),

$$[\mathfrak{g}_x^\lambda, [\mathfrak{g}_x^\mu, \mathfrak{g}_x^\nu]] \subseteq \mathfrak{g}_x^{\lambda+\mu+\nu}.$$

Conclude that we have an orthogonal direct sum decomposition

$$\mathfrak{g} = \mathfrak{g}_x^0 \oplus \bigoplus_{\lambda \neq 0} (\mathfrak{g}_x^\lambda \oplus \mathfrak{g}_x^{-\lambda}).$$

Prove that if B is nondegenerate, then B is nondegenerate on each of the summands.

Chapter 21

The Log-Euclidean Framework Applied to SPD Matrices

21.1 Introduction

In this chapter we present an application of Lie groups and Riemannian geometry. We describe an approach due to Arsigny, Fillard, Pennec and Ayache, to define a Lie group structure and a class of metrics on symmetric, positive-definite matrices (SPD matrices) which yield a new notion of mean on SPD matrices generalizing the standard notion of geometric mean.

SPD matrices are used in diffusion tensor magnetic resonance imaging (for short, DTI), and they are also a basic tool in numerical analysis, for example, in the generation of meshes to solve partial differential equations more efficiently.

As a consequence, there is a growing need to interpolate or to perform statistics on SPD matrices, such as computing the mean of a finite number of SPD matrices.

Recall that the set of $n \times n$ SPD matrices is not a vector space (because if $A \in \mathbf{SPD}(n)$, then $\lambda A \notin \mathbf{SPD}(n)$ if $\lambda < 0$), but it is a convex cone. Thus, the *arithmetic mean* of n SPD matrices $S_1, \ldots, S_n$ can be defined as $(S_1 + \cdots + S_n)/n$, which is SPD. However, there are many situations, especially in DTI, where this mean is not adequate. There are essentially two problems.

(1) The arithmetic mean is not invariant under inversion, which means that if $S = (S_1 + \cdots + S_n)/n$, then in general $S^{-1} \neq (S_1^{-1} + \cdots + S_n^{-1})/n$.

(2) The swelling effect: the determinant $\det(S)$ of the mean S may be strictly larger than the original determinants $\det(S_i)$. This effect is undesirable in DTI because it amounts to introducing more diffusion, which is physically unacceptable.

To circumvent these difficulties, various metrics on SPD matrices have been proposed. One class of metrics is the *affine-invariant metrics* (see Arsigny, Pennec and Ayache [9]).

The swelling effect disappears and the new mean is invariant under inversion, but computing this new mean has a high computational cost, and in general, there is no closed-form formula for this new kind of mean.

Arsigny, Fillard, Pennec and Ayache [8] have defined a new family of metrics on $\mathbf{SPD}(n)$ named *Log-Euclidean metrics*, and have also defined a novel structure of Lie group on $\mathbf{SPD}(n)$ which yields a notion of mean that has the same advantages as the affine mean but is a lot cheaper to compute. Furthermore, this new mean, called *Log-Euclidean mean*, is given by a simple closed-form formula. We will refer to this approach as the *Log-Euclidean framework*.

The key point behind the Log-Euclidean framework is the fact that the exponential map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a bijection, where $\mathbf{S}(n)$ is the space of $n \times n$ symmetric matrices; see Proposition 1.8. Consequently, the exponential map has a well-defined inverse, the *logarithm* $\log\colon \mathbf{SPD}(n) \to \mathbf{S}(n)$.

But more is true. It turns out that $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a diffeomorphism, a fact stated as Theorem 2.8 in Arsigny, Fillard, Pennec and Ayache [8].

Since exp is a bijection, the above result follows from the fact that exp is a local diffeomorphism on $\mathbf{S}(n)$, because $d\exp_S$ is non-singular for all $S \in \mathbf{S}(n)$. In Arsigny, Fillard, Pennec and Ayache [8], it is proved that the non-singularity of $d\exp_I$ near 0, which is well-known, "propagates" to the whole of $\mathbf{S}(n)$.

Actually, the non-singularity of $d\exp$ on $\mathbf{S}(n)$ is a consequence of a more general result stated in Theorem 2.2.

With this preparation, we are ready to present the natural Lie group structure on $\mathbf{SPD}(n)$ introduced by Arsigny, Fillard, Pennec and Ayache [8] (see also Arsigny's thesis [6]).

21.2 A Lie Group Structure on $\mathbf{SPD}(n)$

Using the diffeomorphism $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ and its inverse $\log\colon \mathbf{SPD}(n) \to \mathbf{S}(n)$, an abelian group structure can be defined on $\mathbf{SPD}(n)$ as follows.

Definition 21.1. For any two matrices $S_1, S_2 \in \mathbf{SPD}(n)$, define the *logarithmic product* $S_1 \odot S_2$ by

$$S_1 \odot S_2 = \exp(\log(S_1) + \log(S_2)).$$

Obviously, the multiplication operation $\odot$ is commutative. The following proposition is shown in Arsigny, Fillard, Pennec and Ayache [8] (Proposition 3.2).

Proposition 21.1. *The set* $\mathbf{SPD}(n)$ *with the binary operation* $\odot$ *is an abelian group with identity* I*, and with inverse operation the usual inverse of matrices. Whenever* S_1 *and* S_2 *commute, then* $S_1 \odot S_2 = S_1 S_2$ *(the usual multiplication of matrices).*

For the last statement, we need to show that if $S_1, S_2 \in \mathbf{SPD}(n)$ commute, then S_1S_2 is also in $\mathbf{SPD}(n)$, and that $\log(S_1)$ and $\log(S_2)$ commute, which follows from the fact that if two diagonalizable matrices commute, then they can be diagonalized over the same basis of eigenvectors.

Actually, $(\mathbf{SPD}(n), \odot, I)$ is an abelian Lie group isomorphic to the vector space (also an abelian Lie group!) $\mathbf{S}(n)$, as shown in Arsigny, Fillard, Pennec and Ayache [8] (Theorem 3.3 and Proposition 3.4).

Theorem 21.2. *The abelian group $(\mathbf{SPD}(n), \odot, I)$ is a Lie group isomorphic to its abelian Lie algebra $\mathfrak{spd}(n) = \mathbf{S}(n)$. In particular, the Lie group exponential in $\mathbf{SPD}(n)$ is identical to the usual (matrix) exponential on $\mathbf{S}(n)$.*

We now investigate bi-invariant metrics on the Lie group, $\mathbf{SPD}(n)$.

21.3 Log-Euclidean Metrics on $\mathbf{SPD}(n)$

In general a Lie group does not admit a bi-invariant metric, but an abelian Lie group always does because $\mathrm{Ad}_g = \mathrm{id} \in \mathbf{GL}(\mathfrak{g})$ for all $g \in G$, and so the adjoint representation $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ is trivial (that is, $\mathrm{Ad}(G) = \{\mathrm{id}\}$), and then the existence of bi-invariant metrics is a consequence of Proposition 20.3.

Then given any inner product $\langle -, - \rangle$ on $\mathfrak{g}$, the induced bi-invariant metric on G is given by

$$\langle u, v \rangle_g = \langle (dL_{g^{-1}})_g u, (dL_{g^{-1}})_g v \rangle,$$

where $u, v \in T_gG$.

The geodesics on a Lie group equipped with a bi-invariant metric are the left (or right) translates of the geodesics through e, and the geodesics through e are given by the group exponential, as stated in Proposition 20.20 (3).

Let us apply Proposition 20.20 to the abelian Lie group $\mathbf{SPD}(n)$ and its Lie algebra $\mathfrak{spd}(n) = \mathbf{S}(n)$. Let $\langle -, - \rangle$ be any inner product on $\mathbf{S}(n)$ and let $\langle -, - \rangle_S$ be the induced bi-invariant metric on $\mathbf{SPD}(n)$. We find that the geodesics through $S \in \mathbf{SPD}(n)$ are of the form

$$\gamma(t) = S \odot e^{tV},$$

where $V \in \mathbf{S}(n)$. But $S = e^{\log S}$, so

$$S \odot e^{tV} = e^{\log S} \odot e^{tV} = e^{\log S + tV},$$

so every geodesic through S is of the form

$$\gamma(t) = e^{\log S + tV} = \exp_{\mathrm{gr}}(\log S + tV). \tag{$*$}$$

Remark: To avoid confusion between the exponential and the logarithm as Lie group maps and as Riemannian manifold maps, we will denote the former by exp (instead of $\exp_{\mathrm{gr}}$ and log (instead of $\log_{\mathrm{gr}}$), and their Riemannian counterparts by Exp and Log.

We are going to show that Exp, Log, the bi-invariant metric on $\mathbf{SPD}(n)$, and the distance $d(S,T)$ between two matrices $S,T \in \mathbf{SPD}(n)$ can be expressed in terms of exp and log.

We begin with Exp. Note by an application of the chain rule to $(*)$, we obtain

$$\gamma'(0) = d\exp_{\log S}(V),$$

and since the exponential map of $\mathbf{SPD}(n)$, as a Riemannian manifold, is given by

$$\mathrm{Exp}_S(\widehat{U}) = \gamma_{\widehat{U}}(1), \quad \widehat{U} \in T_S\mathbf{SPD}(n),$$

where $\gamma_{\widehat{U}}$ is the unique geodesic such that $\gamma_{\widehat{U}}(0) = S$ and $\gamma'_{\widehat{U}}(0) = \widehat{U}$.

Remark: Since $\mathbf{SPD}(n)$ is an abelian Lie group, $\mathrm{ad} = 0$ and Proposition 19.1 implies that $T_S\mathbf{SPD}(n) = d(L_S)_I(\mathbf{S}(n))$, so $T_S\mathbf{SPD}(n)$ is isomorphic to $\mathbf{S}(n)$. To compute $d(L_S)_I$, it suffices to take a curve through I with tangent vector $U \in \mathbf{S}(n)$, namely $c(t) = e^{tU}$, and calculate

$$(L_S \circ c)'(0) = (S \odot e^{tU})'(0) = \frac{d}{dt}\left(e^{\log S + tU}\right)_{|t=0} = \sum_{n=0}^{\infty} \frac{d}{dt}\left(\frac{(\log S + tU)^k}{k!}\right)_{|t=0}.$$

The answer is given by the formula for $d\exp_{\log S}(U)$ for the derivative of the matrix exponential; see Section 2.1 just after Proposition 2.1. This calculation yields some complicated linear matrix expression for U unless S and U commute, in which case we get

$$(L_S \circ c)'(0) = \frac{d}{dt}\left(e^{\log S + tU}\right)_{|t=0} = \frac{d}{dt}\left(Se^{tU}\right)_{|t=0} = SU.$$

Since Remark (3) of Proposition 20.20 implies that $\gamma(t) = \gamma_{\widehat{U}}(t)$, we must have $d\exp_{\log S}(V) = \widehat{U}$, so $V = (d\exp_{\log S})^{-1}(\widehat{U})$ and

$$\mathrm{Exp}_S(\widehat{U}) = e^{\log S + V} = e^{\log S + (d\exp_{\log S})^{-1}(\widehat{U})}.$$

However, $\exp \circ \log = \mathrm{id}$, so by differentiation, we get

$$(d\exp_{\log S})^{-1}(\widehat{U}) = d\log_S(\widehat{U}),$$

which yields

$$\mathrm{Exp}_S(\widehat{U}) = e^{\log S + d\log_S(\widehat{U})}, \qquad \widehat{U} \in T_S\mathbf{SPD}(n).$$

To get a formula for $\mathrm{Log}_S T = \widehat{U}$ with $T \in \mathbf{SPD}(n)$ and $\widehat{U} \in T_S\mathbf{SPD}(n)$, we solve the equation $T = \mathrm{Exp}_S(\widehat{U})$ with respect to $\widehat{U}$, that is

$$e^{\log S + (d\exp_{\log S})^{-1}(\widehat{U})} = T,$$

which yields

$$\log S + (d\exp_{\log S})^{-1}(\widehat{U}) = \log T,$$

so $\widehat{U} = d\exp_{\log S}(\log T - \log S)$. Therefore,

$$\mathrm{Log}_S\, T = d\exp_{\log S}(\log T - \log S).$$

Finally, we can find an explicit formula for the Riemannian metric. Let $\widehat{U}, \widehat{V} \in T_S\mathbf{SPD}(n)$. Then

$$\langle \widehat{U}, \widehat{V}\rangle_S = \langle d(L_{S^{-1}})_S(\widehat{U}), d(L_{S^{-1}})_S(\widehat{V})\rangle,$$

We claim that $d(L_{S^{-1}})_S = d\log_S$, which can be shown as follows. Observe that

$$(\log \circ L_{S^{-1}})(T) = \log(S^{-1} \odot T) = \log(\exp(\log(S^{-1}) + \log(T)) = \log S^{-1} + \log T,$$

so $d(\log \circ L_{S^{-1}})_T = d\log_T$ (because S is held fixed), that is

$$d\log_{S^{-1}\odot T} \circ d(L_{S^{-1}})_T = d\log_T,$$

which, for $T = S$, yields $(dL_{S^{-1}})_S = d\log_S$ since $d\log_I = I$. Therefore,

$$\langle \widehat{U}, \widehat{V}\rangle_S = \langle d\log_S(\widehat{U}), d\log_S(\widehat{V})\rangle.$$

Now the proof of Part (3) in Proposition 20.20 shows that a Lie group with a bi-invariant metric is complete; so given any two matrices $S, T \in \mathbf{SPD}(n)$, their distance is the length of the geodesic segment $\gamma_{\widehat{V}}$ such that $\gamma_{\widehat{V}}(0) = S$ and $\gamma_{\widehat{V}}(1) = T$, namely $\left\|\widehat{V}\right\|_S = \sqrt{\langle \widehat{V}, \widehat{V}\rangle_S}$, where $\widehat{V} \in T_S\mathbf{SPD}(n)$ and the norm is given by the Riemannian metric. But since $\mathrm{Exp}_S(\widehat{V}) = \gamma_{\widehat{V}}(1) = T$, we observe that $\widehat{V} = \mathrm{Log}_S T$. Hence

$$d(S, T) = \|\mathrm{Log}_S T\|_S\,.$$

Using the equation

$$\mathrm{Log}_S\, T = d\exp_{\log S}(\log T - \log S)$$

and the fact that $d(\log \circ \exp)_{\log S} = d\log_S \circ d\exp_{\log S} = \mathrm{id}$, we deduce that $d\exp_{\log S} = (d\log_S)^{-1}$. But since $d(L_{S^{-1}})_S = d\log_S$, we may rewrite the previous equality as

$$d\exp_{\log S} = (d\log_S)^{-1} = (d(L_{S^{-1}})_S)^{-1}\,.$$

To simplify $(d(L_{S^{-1}})_S)^{-1}$, we apply the chain rule to the identity $L_{S^{-1}} \circ L_S = L_I = \mathrm{id}$ and deduce that $(d(L_{S^{-1}})_S)^{-1} = d(L_S)_I$. Hence we find that

$$d\exp_{\log S} = d(L_S)_I,$$

which in turn implies that

$$\begin{aligned}\langle \mathrm{Log}_S T, \mathrm{Log}_S T\rangle_S &= \langle d\exp_{\log S}(\log T - \log S), d\exp_{\log S}(\log T - \log S)\rangle_S \\ &= \langle d(L_S)_I(\log T - \log S), \langle d(L_S)_I(\log T - \log S)\rangle \\ &= \langle \log T - \log S, \log T - \log S\rangle,\end{aligned}$$

where the last equality used the bi-invariance of the metric on $S(n)$. Thus we get

$$d(S, T) = \|\log T - \log S\|,$$

where $\| \;\|$ is the norm corresponding to the inner product on $\mathfrak{spd}(n) = \mathbf{S}(n)$. Since $\langle -, -\rangle$ is a bi-invariant metric on $\mathbf{SPD}(n)$, and since

$$\langle \widehat{U}, \widehat{V}\rangle_S = \langle d\log_S(\widehat{U}), d\log_S(\widehat{V})\rangle,$$

we see that the map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is an isometry (since $d\exp \circ d\log = \mathrm{id}$).

In summary, we have proved Corollary 3.9 of Arsigny, Fillard, Pennec and Ayache [8].

Theorem 21.3. *For any inner product $\langle -, -\rangle$ on $\mathbf{S}(n)$, if we give the Lie group $\mathbf{SPD}(n)$ the bi-invariant metric induced by $\langle -, -\rangle$, then the following properties hold:*

(1) For any $S \in \mathbf{SPD}(n)$, the geodesics through S are of the form

$$\gamma(t) = e^{\log S + tV}, \qquad V \in \mathbf{S}(n).$$

(2) The exponential and logarithm associated with the bi-invariant metric on $\mathbf{SPD}(n)$ are given by

$$\begin{aligned}\mathrm{Exp}_S(\widehat{U}) &= e^{\log S + d\log_S(\widehat{U})} \\ \mathrm{Log}_S(T) &= d\exp_{\log S}(\log T - \log S),\end{aligned}$$

for all $S, T \in \mathbf{SPD}(n)$ and all $\widehat{U} \in T_S\mathbf{SPD}(n)$.

(3) The bi-invariant metric on $\mathbf{SPD}(n)$ is given by

$$\langle \widehat{U}, \widehat{V}\rangle_S = \langle d\log_S(\widehat{U}), d\log_S(\widehat{V})\rangle,$$

for all $\widehat{U}, \widehat{V} \in T_S\mathbf{SPD}(n)$ and all $S \in \mathbf{SPD}(n)$, and the distance $d(S, T)$ between any two matrices $S, T \in \mathbf{SPD}(n)$ is given by

$$d(S, T) = \|\log T - \log S\|,$$

where $\| \;\|$ is the norm corresponding to the inner product on $\mathfrak{spd}(n) = \mathbf{S}(n)$.

(4) The map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is an isometry.

In view of Theorem 21.3 Part (3), bi-invariant metrics on the Lie group $\mathbf{SPD}(n)$ are called *Log-Euclidean metrics*. Since $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is an isometry and $\mathbf{S}(n)$ is a vector space, the Riemannian Lie group $\mathbf{SPD}(n)$ is a complete, simply-connected, and flat manifold (the sectional curvature is zero at every point); that is, a flat *Hadamard manifold* (see Sakai [100], Chapter V, Section 4).

Although, in general, Log-Euclidean metrics are not invariant under the action of arbitrary invertible matrices, they are invariant under similarity transformations (an isometry composed with a scaling). Recall that $\mathbf{GL}(n)$ acts on $\mathbf{SPD}(n)$ *via*

$$A \cdot S = ASA^{\top},$$

for all $A \in \mathbf{GL}(n)$ and all $S \in \mathbf{SPD}(n)$. We say that a Log-Euclidean metric is *invariant under* $A \in \mathbf{GL}(n)$ iff

$$d(A \cdot S, A \cdot T) = d(S, T),$$

for all $S, T \in \mathbf{SPD}(n)$. The following result is proved in Arsigny, Fillard, Pennec and Ayache [8] (Proposition 3.11).

Proposition 21.4. *There exist metrics on* $\mathbf{S}(n)$ *that are invariant under all similarity transformations, for example the metric* $\langle S, T\rangle = \mathrm{tr}(ST)$*.*

21.4 A Vector Space Structure on SPD(n)

The vector space structure on $\mathbf{S}(n)$ can also be transferred onto $\mathbf{SPD}(n)$.

Definition 21.2. For any matrix $S \in \mathbf{SPD}(n)$, for any scalar $\lambda \in \mathbb{R}$, define the scalar multiplication $\lambda \circledast S$ by

$$\lambda \circledast S = \exp(\lambda \log(S)).$$

It is easy to check that $(\mathbf{SPD}(n), \odot, \circledast)$ is a vector space with addition $\odot$ and scalar multiplication $\circledast$. By construction, the map $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a linear isomorphism. What happens is that the vector space structure on $\mathbf{S}(n)$ is transfered onto $\mathbf{SPD}(n)$ *via* the log and exp maps.

21.5 Log-Euclidean Means

One of the major advantages of Log-Euclidean metrics is that they yield a computationally inexpensive notion of mean with many desirable properties. If $(x_1, \ldots, x_n)$ is a list of n data points in $\mathbb{R}^m$, then it is a simple exercise to see that the mean $\overline{x} = (x_1 + \cdots + x_n)/n$ is the unique minimum of the map

$$x \mapsto \sum_{i=1}^{n} d_2(x, x_i)^2,$$

where d_2 is the Euclidean distance on $\mathbb{R}^m$. We can think of the quantity

$$\sum_{i=1}^{n} d_2(x, x_i)^2$$

as the *dispersion* of the data.

More generally, if (X, d) is a metric space, for any $\alpha > 0$ and any positive weights $w_1, \ldots, w_n$, with $\sum_{i=1}^{n} w_i = 1$, we can consider the problem of minimizing the function

$$x \mapsto \sum_{i=1}^{n} w_i d(x, x_i)^{\alpha}.$$

The case $\alpha = 2$ corresponds to a generalization of the notion of mean in a vector space and was investigated by Fréchet. In this case, any minimizer of the above function is known as a *Fréchet mean*. Fréchet means are not unique, but if X is a complete Riemannian manifold, certain sufficient conditions on the dispersion of the data are known that ensure the existence and uniqueness of the Fréchet mean (see Pennec [92]). The case $\alpha = 1$ corresponds to a generalization of the notion of *median*. When the weights are all equal, the points that minimize the map

$$x \mapsto \sum_{i=1}^{n} d(x, x_i)$$

are called *Steiner points*. On a Hadamard manifold, Steiner points can be characterized (see Sakai [100], Chapter V, Section 4, Proposition 4.9).

In the case where $X = \mathbf{SPD}(n)$ and d is a Log-Euclidean metric, it turns out that the Fréchet mean is unique and is given by a simple closed-form formula. We have the following theorem from Arsigny, Fillard, Pennec and Ayache [8] (Theorem 3.13), in the case where $w_i = 1/N$ for $i = 1, \ldots, N$:

Theorem 21.5. *Given N matrices $S_1, \ldots, S_N \in \mathbf{SPD}(n)$, their Log-Euclidean Fréchet mean exists and is uniquely determined by the formula*

$$\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N) = \exp\left(\frac{1}{N}\sum_{i=1}^{N} \log(S_i)\right).$$

Furthermore, the Log-Euclidean mean is similarity-invariant, invariant by group multiplication, inversion, and exponential-invariant.

Similarity-invariance means that for any similarity A,

$$\mathbb{E}_{\mathrm{LE}}(AS_1A^{\top}, \ldots, AS_NA^{\top}) = A\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N)A^{\top},$$

and similarly for the other types of invariance.

Observe that the Log-Euclidean mean is a generalization of the notion of geometric mean. Indeed, if $x_1, \ldots, x_n$ are n positive numbers, then their *geometric mean* is given by

$$\mathbb{E}_{\mathrm{geom}}(x_1, \ldots, x_n) = (x_1 \cdots x_n)^{\frac{1}{n}} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log(x_i)\right).$$

The Log-Euclidean mean also has a good behavior with respect to determinants. The following theorem is proved in Arsigny, Fillard, Pennec and Ayache [8] (Theorem 4.2):

Theorem 21.6. *Given N matrices $S_1, \ldots, S_N \in \mathbf{SPD}(n)$, we have*

$$\det(\mathbb{E}_{\mathrm{LE}}(S_1, \ldots, S_N)) = \mathbb{E}_{\mathrm{geom}}(\det(S_1), \ldots, \det(S_N)).$$

Remark: The last line of the proof in Arsigny, Fillard, Pennec and Ayache [8] seems incorrect.

Arsigny, Fillard, Pennec and Ayache [8] also compare the Log-Euclidean mean with the affine mean. We highly recommend the above paper as well as Arsigny's thesis [6] for further details.

21.6 Problems

Problem 21.1. Read Arsigny, Fillard, Pennec and Ayache [8], especially Section 4.6, and implement linear interpolation of two SPD matrices.

Chapter 22

Manifolds Arising from Group Actions

This chapter provides the culmination of the theory presented in the previous nineteen chapters, the concept of a homogeneous naturally reductive space.

We saw in Chapter 4 that many topological spaces arise from a group action. The scenario is that we have a smooth action $\varphi\colon G \times M \to M$ of a Lie group G acting on a manifold M. If G acts transitively on M, then for any point $x \in M$, if G_x is the stabilizer of x, Theorem 4.14 ensures that M is homeomorphic to G/G_x. For simplicity of notation, write $H = G_x$. What we would really like is that G/H actually be a manifold. This is indeed the case, because the transitive action of G on G/H is equivalent to a *right action* of H on G which is no longer transitive, but which has some special properties (to be proper and free).

We are thus led to considering left (and right) actions $\varphi\colon G \times M \to M$ of a Lie group G on a manifold M that are not necessarily transitive. If the action is not transitive, then we consider the *orbit space* M/G of orbits $G \cdot x$ ($x \in M$). However, in general, M/G is not even Hausdorff. It is thus desirable to look for sufficient conditions that ensure that M/G is Hausdorff. A sufficient condition can be given using the notion of a *proper map*. If our action is also *free*, then the orbit space M/G is indeed a smooth manifold. These results are presented in Sections 22.1 and 22.2; see Theorem 22.11 and its corollary Theorem 22.12.

Sharper results hold if we consider Riemannian manifolds. Given a Riemannian manifold N and a Lie group G acting on N, Theorem 22.14 gives us a method for obtaining a Riemannian manifold N/G such that $\pi\colon N \to N/G$ is a Riemannian submersion (when $\cdot\colon G \times N \to N$ is a free and proper action and G acts by isometries). Theorem 22.18 gives us a method for obtaining a Riemannian manifold N/G such that $\pi\colon N \to N/G$ is a Riemannian covering (when $\cdot\colon G \times N \to N$ is a free and proper action of a discrete group G acting by isometries).

In the rest of this chapter, we consider the situation where our Lie group G acts transitively on a manifold M. In this case, we know that M is diffeomorphic to G/H, where H is the stabilizer of any given point in M. Our goal is to endow G/H with Riemannian metrics that arise from inner products on the Lie algebra $\mathfrak{g}$, in a way that is reminiscent of the way in which left-invariant metrics on a Lie group are in one-to-one correspondence with inner products on $\mathfrak{g}$ (see Proposition 20.1). Our goal is realized by the class of reductive homogeneous spaces, which is the object of much of the following sections.

The first step is to consider *G-invariant metrics* on G/H. For any $g \in G$, let $\tau_g \colon G/H \to G/H$ be the diffeomorphism given by

$$\tau_g(g_2H) = gg_2H.$$

The τ_g are left-multiplications on cosets. A metric on G/H is said to be G-invariant iff the τ_g are isometries of G/H. The existence of G-invariant metrics on G/H depends on properties of a certain representation of H called the isotropy representation (see Proposition 22.21). We will also need to express the derivative $d\pi_1 \colon \mathfrak{g} \to T_o(G/H)$ of the natural projection $\pi \colon G \to G/H$ (where o is the point of G/H corresponding to the coset H). This can be done in terms of the Lie group exponential $\exp_{\mathrm{gr}} \colon \mathfrak{g} \to G$ (see Definition 18.11). Then it turns out that $\mathrm{Ker}\,(d\pi_1) = \mathfrak{h}$, the Lie algebra of $\mathfrak{h}$, and $d\pi_1$ factors through $\mathfrak{g}/\mathfrak{h}$ and yields an isomorphism between $\mathfrak{g}/\mathfrak{h}$ and $T_o(G/H)$.

In general, it is difficult to deal with the quotient $\mathfrak{g}/\mathfrak{h}$, and this suggests considering the situation where $\mathfrak{g}$ splits as a direct sum

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}.$$

In this case, $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, and $d\pi_1$ restricts to an isomorphism between $\mathfrak{m}$ and $T_o(G/H)$. This isomorphism can be used to transport an inner product on $\mathfrak{m}$ to an inner product on $T_o(G/H)$. It is remarkable that a simple condition on $\mathfrak{m}$, namely $\mathrm{Ad}(H)$ invariance, yields a one-to-one correspondence between G-invariant metrics on G/H and $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$ (see Proposition 22.22). This is a generalization of the situation of Proposition 20.3 characterizing the existence of bi-invariant metrics on Lie groups. All this is built into the definition of a *reductive homogeneous space* given by Definition 22.8.

It is possible to express the Levi-Civita connection on a reductive homogeneous space in terms of the Lie bracket on $\mathfrak{g}$, but in general this formula is not very useful. A simplification of this formula is obtained if a certain condition holds. The corresponding spaces are said to be *naturally reductive*; see Definition 22.9. A naturally reductive space has the "nice" property that its geodesics at o are given by applying the coset exponential map to $\mathfrak{m}$; see Proposition 22.27. As we will see from the explicit examples provided in Section 22.7, naturally reductive spaces "behave" just as nicely as their Lie group counterpart G, and the coset exponential of $\mathfrak{m}$ will provide *all* the necessary geometric information.

A large supply of naturally reductive homogeneous spaces are the *symmetric spaces*. Such spaces arise from a Lie group G equipped with an involutive automorphism $\sigma \colon G \to G$ (with $\sigma \neq \mathrm{id}$ and $\sigma^2 = \mathrm{id}$). Let G^σ be the set of fixed points of σ, the subgroup of G given by

$$G^\sigma = \{g \in G \mid \sigma(g) = g\},$$

and let G_0^σ be the identity component of G^σ (the connected component of G^σ containing 1). Consider the $+1$ and -1 eigenspaces of the derivative $d\sigma_1\colon \mathfrak{g} \to \mathfrak{g}$ of σ, given by

$$\begin{aligned}\mathfrak{k} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}\\ \mathfrak{m} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = -X\}.\end{aligned}$$

Pick a closed subgroup K of G such that $G_0^\sigma \subseteq K \subseteq G^\sigma$. Then it can be shown that G/K is a reductive homogenous space and that $\mathfrak{g}$ factors as a direct sum $\mathfrak{k} \oplus \mathfrak{m}$, which makes G/K a reductive space. Furthermore, if G is connected and if both G_0^σ and K are compact, then G/K is naturally reductive.

There is an extensive theory of symmetric spaces and our goal is simply to show that the additional structure afforded by an involutive automorphism of G yields spaces that are naturally reductive. The theory of symmetric spaces was entirely created by one person, Élie Cartan, who accomplished the tour de force of giving a complete classification of these spaces using the classification of semisimple Lie algebras that he had obtained earlier. In Sections 22.8, 22.9, and 22.10, we provide an introduction to symmetric spaces.

22.1 Proper Maps

We saw in Chapter 4 that many manifolds arise from a group action. The scenario is that we have a smooth action $\varphi\colon G \times M \to M$ of a Lie group G acting on a manifold M (recall that an action φ is smooth if it is a smooth map). If G acts transitively on M, then for any point $x \in M$, if G_x is the stabilizer of x, then Proposition 22.13 will show that G/G_x is diffeomorphic to M and that the projection $\pi\colon G \to G/G_x$ is a submersion.

If the action is not transitive, then we consider the *orbit space* M/G of orbits $G \cdot x$. However, in general, M/G is not even Hausdorff. It is thus desirable to look for sufficient conditions that ensure that M/G is Hausdorff. A sufficient condition can be given using the notion of a *proper map*.

Before we go any further, let us observe that the case where our action is transitive is subsumed by the more general situation of an orbit space. Indeed, if our action $\cdot\colon G \times M \to M$ is transitive, for any $x \in M$, we know that M is homeomorphic to G/H, where $H = G_x$ is the stabilizer of x. Furthermore, for any continuous action (not necessarily transitive), the subgroup H is a closed subgroup of G. Then we can consider the *right* action $G \times H \to G$ of H on G given by

$$g \cdot h = gh, \quad g \in G, h \in H.$$

The orbits of this (right) action are precisely the left cosets gH of H. Therefore, the set of left cosets G/H (the homogeneous space induced by the action $\cdot\colon G \times M \to M$) is the set of orbits of the right action $G \times H \to G$.

Observe that we have a transitive left action of G on the space G/H of left cosets, given by

$$g_1 \cdot g_2H = g_1g_2H.$$

The stabilizer of $1H$ is obviously H itself. Thus we recover the original transitive left action of G on $M = G/H$.

Now it turns out that a right action of the form $G \times H \to G$, where H is a closed subgroup of a Lie group G, is a special case of a free and proper right action $M \times G \to M$, in which case the orbit space M/G is a manifold, and the projection $\pi\colon G \to M/G$ is a submersion.

Let us now define proper maps.

Definition 22.1. If X and Y are two Hausdorff topological spaces,[1] a function a $\varphi\colon X \to Y$ is *proper* iff it is continuous and for every topological space Z, the map $\varphi \times \mathrm{id}\colon X \times Z \to Y \times Z$ is a *closed map* (recall that f is a closed map iff the image of any closed set by f is a closed set).

If we let Z be a one-point space, we see that *a proper map is closed*.

At first glance, it is not obvious how to check that a map is proper just from Definition 22.1. Proposition 22.2 gives a more palatable criterion.

The following proposition is easy to prove (see Bourbaki, General Topology [20], Chapter 1, Section 10).

Proposition 22.1. *If $\varphi\colon X \to Y$ is any proper map, then for any closed subset F of X, the restriction of φ to F is proper.*

The following result providing a "good" criterion for checking that a map is proper can be shown (see Bourbaki, General Topology [20], Chapter 1, Section 10).

Proposition 22.2. *A continuous map $\varphi\colon X \to Y$ is proper iff φ is closed and if $\varphi^{-1}(y)$ is compact for every $y \in Y$.*

Proposition 22.2 shows that a homeomorphism (or a diffeomorphism) is proper.

If φ is proper, it is easy to show that $\varphi^{-1}(K)$ is compact in X whenever K is compact in Y. Moreover, if Y is also locally compact, then we have the following result (see Bourbaki, General Topology [20], Chapter 1, Section 10).

Proposition 22.3. *If Y is locally compact, a continuous map $\varphi\colon X \to Y$ is a proper map iff $\varphi^{-1}(K)$ is compact in X whenever K is compact in Y*

In particular, this is true if Y is a manifold since manifolds are locally compact. This explains why Lee [76] (Chapter 9) takes the property stated in Proposition 22.3 as the definition of a proper map (because he only deals with manifolds).[2]

[1] It is not necessary to assume that X and Y are Hausdorff but, if X and/or Y are not Hausdorff, we have to replace "compact" by "quasi-compact." We have no need for this extra generality.

[2] However, Duistermaat and Kolk [43] seem to have overlooked the fact that a condition on Y (such as local compactness) is needed in their remark on lines 5-6, page 53, just before Lemma 1.11.3.

Finally we can define proper actions.

Remark: It is remarkable that a great deal of material discussed in this chapter, especially in Sections 22.4–22.9, can be found in Volume IV of Dieudonné's classical treatise on Analysis [36]. However, it is spread over 400 pages, which does not make it easy to read.

22.2 Proper and Free Actions

Definition 22.2. Given a Hausdorff topological group G and a topological space M, a left action $\cdot\colon G \times M \to M$ is *proper* if it is continuous and if the map

$$\theta\colon G \times M \longrightarrow M \times M, \quad (g, x) \mapsto (g \cdot x, x)$$

is proper.

The right actions associated with the transitive actions presented in Section 4.2 are examples of proper actions.

Proposition 22.4. *The action $\cdot\colon H \times G \to G$ of a closed subgroup H of a group G on G (given by $(h, g) \mapsto hg$) is proper. The same is true for the right action of H on G.*

Proof. If H is a closed subgroup of G and if $\cdot\colon G \times M \to M$ is a proper action, then the restriction of this action to H is also proper (by Proposition 22.1, because $H \times M$ is closed in $G \times M$). If we let $M = G$, then G acts on itself by left translation, and the map $\theta\colon G \times G \to G \times G$ given by $\theta(g, x) = (gx, x)$ is a homeomorphism, so it is proper. □

As desired, proper actions yield Hausdorff orbit spaces.

Proposition 22.5. *If the action $\cdot\colon G \times M \to M$ is proper (where G is Hausdorff), then the orbit space M/G is Hausdorff. Furthermore, M is also Hausdorff.*

Proof. If the action is proper, then the map $\theta\colon G \times M \to M \times M$ as defined in Definition 22.2 is closed. Hence the orbit equivalence relation is closed since it is the image of $G \times M$ in $M \times M$. Furthermore, $\pi\colon M \to M/G$ is an open map and so by the paragraph following Proposition 12.32, M/G is Hausdorff. The second part is left as an exercise. □

We also have the following properties (see Bourbaki, General Topology [20], Chapter 3, Section 4).

Proposition 22.6. *Let $\cdot\colon G \times M \to M$ be a proper action, with G Hausdorff. For any $x \in M$, let $G \cdot x$ be the orbit of x and let G_x be the stabilizer of x. Then*

(a) The map $g \mapsto g \cdot x$ is a proper map from G to M.

(b) G_x is compact.

(c) *The canonical map from G/G_x to $G \cdot x$ is a homeomorphism.*

(d) *The orbit $G \cdot x$ is closed in M.*

If G is locally compact, then we have the following necessary and sufficient conditions for an action $\cdot\colon G \times M \to M$ to be proper (see Bourbaki, General Topology [20], Chapter 3, Section 4).

Proposition 22.7. *If G and M are Hausdorff and if G is locally compact, then the action $\cdot\colon G \times M \to M$ is proper iff for all $x, y \in M$, there exist some open sets, V_x and V_y in M, with $x \in V_x$ and $y \in V_y$, so that the closure $\overline{K}$ of the set $K = \{g \in G \mid (g \cdot V_x) \cap V_y \neq \emptyset\}$, is compact in G.*

In particular, if G has the discrete topology, the above condition holds iff the sets $\{g \in G \mid (g \cdot V_x) \cap V_y \neq \emptyset\}$ are finite. Also, if G is compact, then $\overline{K}$ is automatically compact, so every compact group acts properly.

Corollary 22.8. *If G and M are Hausdorff and if G is compact, then the action $\cdot\colon G \times M \to M$ is proper.*

If M is locally compact, we have the following characterization of being proper (see Bourbaki, General Topology [20], Chapter 3, Section 4).

Proposition 22.9. *Let $\cdot\colon G \times M \to M$ be a continuous action, with G and M Hausdorff. For any compact subset K of M we have*

(a) *The set $G_K = \{g \in G \mid (g \cdot K) \cap K \neq \emptyset\}$ is closed.*

(b) *If M is locally compact, then the action is proper iff G_K is compact for every compact subset K of M.*

In the special case where G is discrete (and M is locally compact), Condition (b) says that the action is proper iff G_K is finite. We use this criterion to show that the action $\cdot\colon \mathbb{Z} \times \mathbb{R} \to \mathbb{R}$ given by $n \cdot x = 2^n x$ is not proper. Note that $\mathbb{R}$ is locally compact. Take $K = \{0, 1\}$, a set which is clearly compact in $\mathbb{R}$. Then $n \cdot K = \{0, 2^n\}$ and $\{0\} \subseteq (n \cdot K) \cap K$ Thus, by definition, $G_K = \mathbb{Z}$, which is not compact or finite in $\mathbb{R}$. Intuitively, proper actions on manifolds involve translations, rotations, and constrained expansions. The action $n \cdot x = 2^n x$ provides too much dilation on $\mathbb{R}$ to be a proper action.

Remark: If G is a Hausdorff topological group and if H is a subgroup of G, then it can be shown that the action of G on G/H ($(g_1, g_2H) \mapsto g_1g_2H$) is proper iff H is compact in G.

Definition 22.3. An action $\cdot\colon G \times M \to M$ is *free* if for all $g \in G$ and all $x \in M$, if $g \neq 1$ then $g \cdot x \neq x$.

An equivalent way to state that an action $\cdot\colon G \times M \to M$ is free is as follows. For every $g \in G$, let $\tau_g\colon M \to M$ be the diffeomorphism of M given by

$$\tau_g(x) = g \cdot x, \quad x \in M.$$

Then the action $\cdot\colon G \times M \to M$ is free iff for all $g \in G$, if $g \neq 1$ then τ_g has no fixed point.

Consequently, an action $\cdot\colon G \times M \to M$ is free iff for every $x \in M$, the stabilizer G_x of x is reduced to the trivial group $\{1\}$.

For example, the action of $\mathbf{SO}(3)$ on S^2 given by Example 4.2 of Section 4.2 is not free since any rotation of S^2 fixes the two points of the rotation axis.

If H is a subgroup of G, obviously H acts freely on G (by multiplication on the left or on the right). This fact together with Proposition 22.4 yields the following corollary which provides a large supply of free and proper actions.

Corollary 22.10. *The action $\cdot\colon H \times G \to G$ of a closed subgroup H of a group G on G (given by $(h, g) \mapsto hg$) is free and proper. The same is true for the right action of H on G.*

There is a stronger version of the results that we are going to state next that involves the notion of principal bundle. Since this notion is not discussed in this book, we state weaker versions not dealing with principal bundles. The weaker version that does not mention principal bundles is usually stated for left actions; for instance, in Lee [76] (Chapter 9, Theorem 9.16). We formulate both a left and a right version.

Theorem 22.11. *Let M be a smooth manifold, G be a Lie group, and let $\cdot\colon G \times M \to M$ be a left smooth action (resp. right smooth action $\cdot\colon M \times G \to M$) which is proper and free. Then the canonical projection $\pi\colon G \to M/G$ is a submersion (which means that $d\pi_g$ is surjective for all $g \in G$), and there is a unique manifold structure on M/G with this property.*

Theorem 22.11 has some interesting corollaries. Because a closed subgroup H of a Lie group G is a Lie group, and because the action of a closed subgroup is free and proper, if we apply Theorem 22.11 to the right action $\cdot\colon G \times H \to G$ (here $M = G$ and $G = H$), we get the following result (proofs can also be found in Bröcker and tom Dieck [24] (Chapter I, Section 4) and in Duistermaat and Kolk [43] (Chapter 1, Section 11)). *This is the result we use to verify reductive homogeneous spaces are indeed manifolds.*

Theorem 22.12. *If G is a Lie group and H is a closed subgroup of G, then the canonical projection $\pi\colon G \to G/H$ is a submersion (which means that $d\pi_g$ is surjective for all $g \in G$), and there is a unique manifold structure on G/H with this property.*

In the special case where G acts transitively on M, for any $x \in M$, if G_x is the stabilizer of x, then with $H = G_x$, Theorem 22.12 shows that there is a manifold structure on G/H such $\pi\colon G \to G/H$ is a submersion.

Actually, G/H is diffeomorphic to M, as shown by the following theorem whose proof can be found in Lee [76] (Chapter 9, Theorem 9.24).

Theorem 22.13. *Let* $\cdot\colon G \times M \to M$ *be a smooth transitive action of a Lie group* G *on a smooth manifold* M *(so that* M *is a homogeneous space). For any* $x \in M$*, if* G_x *is the stabilizer of* x *and if we write* $H = G_x$*, then the map* $\overline{\pi}_x\colon G/H \to M$ *given by*

$$\overline{\pi}_x(gH) = g \cdot x$$

is a diffeomorphism and an equivariant map (with respect to the action of G *on* G/H *and the action of* G *on* M*).*

The proof of Theorem 22.13 is not particularly difficult. It relies on technical properties of equivariant maps that we have not discussed. We refer the reader to the excellent account in Lee [76] (Chapter 9).

By Theorem 22.12 and Theorem 22.13, every homogeneous space M (with a smooth G-action) is equivalent to a manifold G/H as above. This is an important and very useful result that reduces the study of homogeneous spaces to the study of coset manifolds of the form G/H where G is a Lie group and H is a closed subgroup of G.

Here is a simple example of Theorem 22.12. Let $G = \mathbf{SO}(3)$ and

$$H = \left\{ M \in \mathbf{SO}(3) \mid M = \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix},\ S \in \mathbf{SO}(2) \right\}.$$

The right action $\cdot\colon \mathbf{SO}(3) \times H \to \mathbf{SO}(3)$ given by the matrix multiplication

$$g \cdot h = gh, \qquad g \in \mathbf{SO}(3),\ \ h \in H,$$

yields the left cosets gH, and the orbit space $\mathbf{SO}(3)/\mathbf{SO}(2)$, which by Theorem 22.12 and Theorem 22.13 is diffeomorphic to S^2.

22.3 Riemannian Submersions and Coverings Induced by Group Actions ⊛

The purpose of this section is to equip the orbit space M/G of Theorem 22.11 with the inner product structure of a Riemannian manifold. Because we provide a different proof for the reason why reductive homogeneous manifolds are Riemannian manifolds, namely Proposition 22.23, this section is not necessary for understanding the material in Section 22.4 and may be skipped on the first reading.

Definition 22.4. Given a Riemannian manifold (N, h), we say that a Lie group G *acts by isometries on* N if for every $g \in G$, the diffeomorphism $\tau_g\colon N \to N$ given by

$$\tau_g(p) = g \cdot p, \quad p \in N,$$

is an isometry ($(d\tau_g)_p\colon T_pN \to T_{\tau_g(p)}N$ is an isometry for all $p \in N$).

If (N, h) is a Riemannian manifold and if G is a Lie group, then $\pi\colon N \to N/G$ can be made into a Riemannian submersion.

Theorem 22.14. *Let (N, h) be a Riemannian manifold and let $\cdot\colon G \times N \to N$ be a smooth, free and proper action, with G a Lie group acting by isometries of N. Then there is a unique Riemannian metric g on $M = N/G$ such that $\pi\colon N \to M$ is a Riemannian submersion.*

Sketch of proof. We follow Gallot, Hulin, Lafontaine [49] (Chapter 2, Proposition 2.28). Pick any $x \in M = N/G$, and any $u, v \in T_xM$. For any $p \in \pi^{-1}(x)$, there exist unique lifts $\overline{u}, \overline{v} \in \mathcal{H}_p$ such that

$$d\pi_p(\overline{u}) = u \quad \text{and} \quad d\pi_p(\overline{v}) = v.$$

See Definitions 17.2 and 17.4. Set

$$g_x(u, v) = h_p(\overline{u}, \overline{v}),$$

which makes (T_xM, g_x) isometric to $(\mathcal{H}_p, h_p)$. See Figure 22.1. We need to check that g_x does not depend on the choice of p in the fibre $\pi^{-1}(x)$, and that (g_x) is a smooth family. We check the first property (for the second property, see Gallot, Hulin, Lafontaine [49]). If $\pi(q) = \pi(p)$, then there is some $g \in G$ such that $\tau_g(p) = q$, and $(d\tau_g)_p$ induces an isometry between $\mathcal{H}_p$ and $\mathcal{H}_q$ which commutes with π. Therefore, g_x does not depend on the choice of $p \in \pi^{-1}(x)$. □

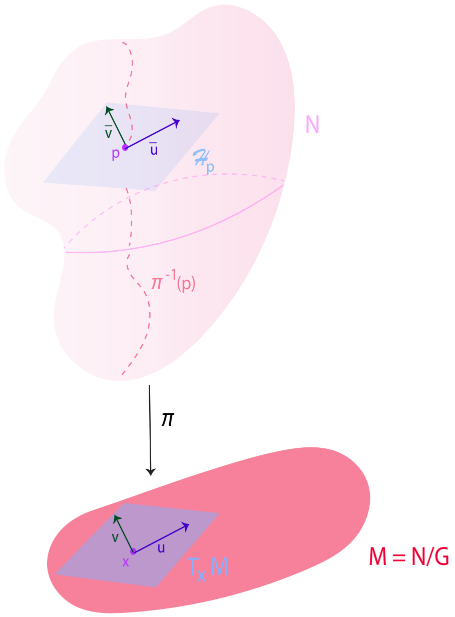


Figure 22.1: A schematic illustration of the metric on N inducing the metric on $M = N/G$ via a lift to horizontal tangent vectors.

As an example, take $N = S^{2n+1}$, where N is isomorphic to the subspace of $\mathbb{C}^{n+1}$ given by

$$\Sigma^n = \left\{ (z_1, z_2, \cdots, z_{n+1}) \in \mathbb{C}^{n+1} \mid \sum_{i=1}^{n+1} z_i \overline{z_i} = 1 \right\}.$$

The group $G = S^1 = \mathbf{SU}(1)$ acts by isometries on S^{2n+1} by complex multiplication. In other words, given $p \in \Sigma^n$ and $e^{i\theta} \in \mathbf{SU}(1)$,

$$e^{i\theta} \cdot p = (e^{i\theta} z_1, e^{i\theta} z_2, \cdots, e^{i\theta} z_{n+1}) \in \Sigma^n.$$

Since the action of G on N is free and proper, Theorem 22.14 and Example 4.8 imply that we obtain the Riemann submersion $\pi\colon S^{2n+1} \to \mathbb{CP}^n$. If we pick the canonical metric on S^{2n+1}, by Theorem 22.14, we obtain a Riemannian metric on $\mathbb{CP}^n$ known as the *Fubini–Study metric*. Using Proposition 17.8, it is possible to describe the geodesics of $\mathbb{CP}^n$; see Gallot, Hulin, Lafontaine [49] (Chapter 2).

Another situation where a group action yields a Riemannian submersion is the case where a transitive action is reductive, considered in the next section.

We now consider the case of a smooth action $\cdot\colon G \times M \to M$, where G is a discrete group (and M is a manifold). In this case, we will see that $\pi\colon M \to M/G$ is a Riemannian covering map.

Assume G is a discrete group. By Proposition 22.7, the action $\cdot\colon G \times M \to M$ is proper iff for all $x, y \in M$, there exist some open sets, V_x and V_y in M, with $x \in V_x$ and $y \in V_y$, so that the set $K = \{g \in G \mid (g \cdot V_x) \cap V_y \neq \emptyset\}$ is finite. By Proposition 22.9, the action $\cdot\colon G \times M \to M$ is proper iff $G_K = \{g \in G \mid g \cdot K \cap K \neq \emptyset\}$ is finite for every compact subset K of M.

It is shown in Lee [76] (Chapter 9) that the above conditions are equivalent to the conditions below.

Proposition 22.15. *If $\cdot\colon G \times M \to M$ is a smooth action of a discrete group G on a manifold M, then this action is proper iff*

(i) For every $x \in M$, there is some open subset V with $x \in V$ such that $gV \cap V \neq \emptyset$ for only finitely many $g \in G$.

(ii) For all $x, y \in M$, if $y \notin G \cdot x$ (y is not in the orbit of x), then there exist some open sets V, W with $x \in V$ and $y \in W$ such that $gV \cap W = \emptyset$ for all $g \in G$.

The following proposition gives necessary and sufficient conditions for a discrete group to act freely and properly often found in the literature (for instance, O'Neill [91], Berger and Gostiaux [15], and do Carmo [39], but beware that in this last reference Hausdorff separation is not required!).

Proposition 22.16. *If X is a locally compact space and G is a discrete group, then a smooth action of G on X is free and proper iff the following conditions hold.*

(i) For every $x \in X$, there is some open subset V with $x \in V$ such that $gV \cap V = \emptyset$ for all $g \in G$ such that $g \neq 1$.

(ii) For all $x, y \in X$, if $y \notin G \cdot x$ (y is not in the orbit of x), then there exist some open sets V, W with $x \in V$ and $y \in W$ such that $gV \cap W = \emptyset$ for all $g \in G$.

Proof. Condition (i) of Proposition 22.16 implies Condition (i) of Proposition 22.15, and Condition (ii) is the same in Proposition 22.16 and Proposition 22.15. If Condition (i) holds, then the action must be free since if $g \cdot x = x$, then $gV \cap V \neq \emptyset$, which implies that $g = 1$.

Conversely, we just have to prove that the conditions of Proposition 22.15 imply Condition (i) of Proposition 22.16. By Condition (i) of Proposition 22.15, there is some open subset W containing x and a finite number of elements of G, say $g_1, \ldots, g_m$, with $g_i \neq 1$, such that

$$g_i W \cap W \neq \emptyset, \quad i = 1, \ldots, m.$$

Since our action is free and $g_i \neq 1$, we have $g_i \cdot x \neq x$, so by Hausdorff separation, there exist some open subsets W_i, W_i', with $x \in W_i$ and $g_i \cdot x \in W_i'$, such that $W_i \cap W_i' = \emptyset$, $i = 1, \ldots, m$. Then if we let

$$V = W \cap \left(\bigcap_{i=1}^{m} (W_i \cap g_i^{-1} W_i') \right),$$

we see that $V \cap g_i V = \emptyset$, and since $V \subseteq W$, we also have $V \cap gV = \emptyset$ for all other $g \in G$. □

Remark: The action of a discrete group satisfying the properties of Proposition 22.16 is often called “properly discontinuous.” However, as pointed out by Lee ([76], just before Proposition 9.18), this term is self-contradictory since such actions are smooth, and thus continuous!

Then we have the following useful result.

Theorem 22.17. *Let N be a smooth manifold and let G be discrete group acting smoothly, freely and properly on N. Then there is a unique structure of smooth manifold on N/G such that the projection map $\pi \colon N \to N/G$ is a covering map.*

For a proof, see Gallot, Hulin, Lafontaine [49] (Theorem 1.88) or Lee [76] (Theorem 9.19).

Real projective spaces are illustrations of Theorem 22.17. Indeed, if N is the unit n-sphere $S^n \subseteq \mathbb{R}^{n+1}$ and $G = \{I, -I\}$, where $-I$ is the antipodal map, then the conditions of Proposition 22.16 are easily checked (since S^n is compact), and consequently the quotient

$$\mathbb{RP}^n = S^n/G$$

is a smooth manifold and the projection map $\pi \colon S^n \to \mathbb{RP}^n$ is a covering map. The fiber $\pi^{-1}([x])$ of every point $[x] \in \mathbb{RP}^n$ consists of two antipodal points: $x, -x \in S^n$.

The next step is to see how a Riemannian metric on N induces a Riemannian metric on the quotient manifold N/G. The following theorem is the Riemannian version of Theorem 22.17.

Theorem 22.18. *Let (N, h) be a Riemannian manifold and let G be discrete group acting smoothly, freely and properly on N, and such that the map $x \mapsto \sigma \cdot x$ is an isometry for all $\sigma \in G$. Then there is a unique structure of Riemannian manifold on $M = N/G$ such that the projection map $\pi \colon N \to M$ is a Riemannian covering map.*

Proof sketch. For a complete proof see Gallot, Hulin, Lafontaine [49] (Proposition 2.20). To define a Riemannian metric g on $M = N/G$ we need to define an inner product g_p on the tangent space T_pM for every $p \in M$. Pick any $q_1 \in \pi^{-1}(p)$ in the fibre of p. Because π is a covering map, it is a local diffeomorphism, and thus $d\pi_{q_1} \colon T_{q_1}N \to T_pM$ can be made into an isometry as follows. Given any two tangent vectors $u, v \in T_pM$, we define their inner product $g_p(u, v)$ by

$$g_p(u, v) = h_{q_1}(d\pi_{q_1}^{-1}(u), d\pi_{q_1}^{-1}(v)).$$

See Figure 22.2. We need to show that g_p does not depend on the choice of $q_1 \in \pi^{-1}(p)$. Let $q_2 \in \pi^{-1}(p)$ be any other point in the fibre of p. By definition of $M = N/G$, we have $q_2 = g \cdot q_1$ for some $g \in G$, and we know that the map $f \colon q \mapsto g \cdot q$ is an isometry of N. Since $\pi = \pi \circ f$, we have

$$d\pi_{q_1} = d\pi_{q_2} \circ df_{q_1},$$

and since $d\pi_{q_1} \colon T_{q_1}N \to T_pM$ and $d\pi_{q_2} \colon T_{q_2}N \to T_pM$ are isometries, we get

$$d\pi_{q_2}^{-1} = df_{q_1} \circ d\pi_{q_1}^{-1}.$$

But $df_{q_1} \colon T_{q_1}N \to T_{q_2}N$ is also an isometry, so

$$h_{q_2}(d\pi_{q_2}^{-1}(u), d\pi_{q_2}^{-1}(v)) = h_{q_2}(df_{q_1}(d\pi_{q_1}^{-1}(u)), df_{q_1}(d\pi_{q_2}^{-1}(v))) = h_{q_1}(d\pi_{q_1}^{-1}(u), d\pi_{q_1}^{-1}(v)).$$

Therefore, the inner product g_p is well defined on T_pM. It remains to prove that (g_p) is a smooth family; see Gallot, Hulin, Lafontaine [49] (Proposition 2.20). □

Theorem 22.18 implies that every Riemannian metric g on the sphere S^n induces a Riemannian metric $\widehat{g}$ on the projective space $\mathbb{RP}^n$, in such a way that the projection $\pi \colon S^n \to \mathbb{RP}^n$ is a Riemannian covering. In particular, if U is an open hemisphere obtained by removing its boundary S^{n-1} from a closed hemisphere, then π is an isometry between U and its image $\mathbb{RP}^n - \pi(S^{n-1}) \cong \mathbb{RP}^n - \mathbb{RP}^{n-1}$.

We also observe that for any two points $p = [x]$ and $q = [y]$ in $\mathbb{RP}^n$, where $x, y \in S^n$, if $x \cdot y = \cos\theta$, with $0 \leq \theta \leq \pi$, then there are two possibilities:

1. $x \cdot y \geq 0$, which means that $0 \leq \theta \leq \pi/2$, or
2. $x \cdot y < 0$, which means that $\pi/2 < \theta \leq \pi$.

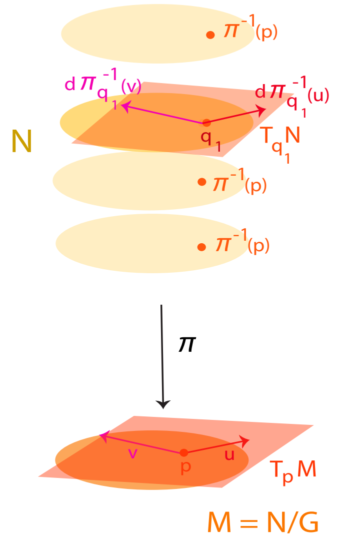


Figure 22.2: A schematic illustration of the metric on the covering space N inducing the metric on $M = N/G$.

In the second case, since $[-y] = [y]$ and $x \cdot (-y) = -x \cdot y$, we can replace the representative y of q by $-y$, and we have $x \cdot (-y) = \cos(\pi - \theta)$, with $0 \leq \pi - \theta < \pi/2$. Therefore, in all cases, for any two points $p, q \in \mathbb{RP}^n$, we can find an open hemisphere U such that $p = [x], q = [y]$, $x, y \in U$, and $x \cdot y \geq 0$; that is, the angle $\theta \geq 0$ between x and y is at most $\pi/2$.

Applying Theorem 22.18 to $\mathbb{RP}^n$ and the canonical Euclidean metric induced by $\mathbb{R}^{n+1}$, since geodesics of S^n are great circles (see Section 22.7), by the discussion above, for any two points $p = [x]$ and $q = [y]$ in $\mathbb{RP}^n$, with $x, y \in S^n$, the distance between them is given by

$$d(p, q) = d([x], [y]) = \begin{cases} \cos^{-1}(x \cdot y) & \text{if } x \cdot y \geq 0 \\ \cos^{-1}(-x \cdot y) & \text{if } x \cdot y < 0. \end{cases}$$

Here $\cos^{-1}(z) = \arccos(z)$ is the unique angle $\theta \in [0, \pi]$ such that $\cos(\theta) = z$. Equivalently,

$$d([x], [y]) = \cos^{-1}(|x \cdot y|).$$

If the representatives $x, y \in \mathbb{R}^{n+1}$ of $p = [x]$ and $q = [q]$ are not unit vectors, then

$$d([x], [y]) = \cos^{-1}\left(\frac{|x \cdot y|}{\|x\| \, \|y\|}\right).$$

Note that $0 \leq d(p, q) \leq \pi/2$.

In summary, given a Riemannian manifold N and a group G acting on N, Theorem 22.14 gives us a method for obtaining a Riemannian manifold N/G such that $\pi\colon N \to N/G$ is a Riemannian submersion ($\cdot\colon G \times N \to N$ is a free and proper action and G acts by isometries). Theorem 22.18 gives us a method for obtaining a Riemannian manifold N/G such that $\pi\colon N \to N/G$ is a Riemannian covering ($\cdot\colon G \times N \to N$ is a free and proper action of a discrete group G acting by isometries).

In the next section we show that Riemannian submersions arise from a reductive homogeneous space.

22.4 Reductive Homogeneous Spaces

If $\cdot\colon G \times M \to M$ is a smooth action of a Lie group G on a manifold M, then a certain class of Riemannian metrics on M is particularly interesting. Recall that for every $g \in G$, $\tau_g\colon M \to M$ is the diffeomorphism of M given by

$$\tau_g(p) = g \cdot p, \quad \text{for all } p \in M.$$

If $M = G$ and G acts on itself (on the left) by left multiplication, then $\tau_g = L_g$ for all $g \in G$, as defined earlier in Section 18.1. Thus the left multiplications τ_g generalize left multiplications in a group.

Definition 22.5. Given a smooth action $\cdot\colon G \times M \to M$, a metric $\langle -, - \rangle$ on M is G-*invariant* if τ_g is an isometry for all $g \in G$; that is, for all $p \in M$, we have

$$\langle d(\tau_g)_p(u), d(\tau_g)_p(v)\rangle_{\tau_g(p)} = \langle u, v\rangle_p \quad \text{for all } u, v \in T_pM.$$

If the action is transitive, then for any fixed $p_0 \in M$ and for every $p \in M$, there is some $g \in G$ such that $p = g \cdot p_0$, so it is sufficient to require that $d(\tau_g)_{p_0}$ be an isometry for every $g \in G$.

From now on we are dealing with a *smooth transitive action* $\cdot\colon G \times M \to M$, and for any given $p_0 \in M$, if $H = G_{p_0}$ is the stabilizer of p_0, then by Theorem 22.13, M is diffeomorphic to G/H.

Recall the notion of representation given in Definition 20.3. The existence of G-invariant metrics on G/H depends on properties of a certain representation of H called the isotropy representation (see Proposition 22.21). The isotropy representation is equivalent to another representation $\mathrm{Ad}^{G/H}\colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ of H involving the quotient algebra $\mathfrak{g}/\mathfrak{h}$.

This representation is too complicated to deal with, so we consider the more tractable situation where the Lie algebra $\mathfrak{g}$ of G factors as a direct sum

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m},$$

for some subspace $\mathfrak{m}$ of $\mathfrak{g}$ such that $\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $h \in H$, where $\mathfrak{h}$ is the Lie algebra of H. Then $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, and the representation $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ becomes the representation $\mathrm{Ad} \colon H \to \mathbf{GL}(\mathfrak{m})$, where Ad_h is the restriction of Ad_h to $\mathfrak{m}$ for every $h \in H$. In this situation there is an isomorphism between $T_{p_0}M \cong T_o(G/H)$ and $\mathfrak{m}$ (where o denotes the point in G/H corresponding to the coset H). It is also the case that if H is "nice" (for example, compact), then $M = G/H$ will carry G-invariant metrics, and that under such metrics, the projection $\pi \colon G \to G/H$ is a Riemannian submersion.

In order to proceed it is necessary to express the derivative $d\pi_1 \colon \mathfrak{g} \to T_o(G/H)$ of the projection map $\pi \colon G \to G/H$ in terms of certain vector fields. This is a special case of a process in which an action $\cdot \colon G \times M \to M$ associates a vector field X^* on M to every vector $X \in \mathfrak{g}$ in the Lie algebra of G.

Definition 22.6. Given a smooth action $\varphi \colon G \times M \to M$ of a Lie group on a manifold M, for every $X \in \mathfrak{g}$, we define the vector field X^* (or X_M) on M called an *action field* or *infinitesimal generator* of the action corresponding to X, by

$$X^*(p) = \frac{d}{dt}(\exp(tX) \cdot p)\bigg|_{t=0}, \quad p \in M.$$

For a fixed $X \in \mathfrak{g}$, the map $t \mapsto \exp(tX)$ is a curve through 1 in G, so the map $t \mapsto \exp(tX) \cdot p$ is a curve through p in M, and $X^*(p)$ is the tangent vector to this curve at p.

For example, in the case of the adjoint action $\mathrm{Ad} \colon G \times \mathfrak{g} \to \mathfrak{g}$, for every $X \in \mathfrak{g}$, Proposition 18.10 implies that

$$X^*(Y) = \frac{d}{dt}(\mathrm{Ad}(\exp(tX))Y)\bigg|_{t=0} = \frac{d}{dt}(e^{\mathrm{ad}(tX)}Y)\bigg|_{t=0} = \mathrm{ad}(X)(Y) = [X, Y],$$

so $X^* = \mathrm{ad}(X)$.

For any $p_0 \in M$, there is a diffeomorphism $G/G_{p_0} \to G \cdot p_0$ onto the orbit $G \cdot p_0$ of p_0 viewed as a manifold, and it is not hard to show that for any $p \in G \cdot p_0$, we have an isomorphism

$$T_p(G \cdot p_0) = \{X^*(p) \mid X \in \mathfrak{g}\};$$

see Marsden and Ratiu [77] (Chapter 9, Section 9.3). It can also be shown that the Lie algebra $\mathfrak{g}_p$ of the stabilizer G_p of p is given by

$$\mathfrak{g}_p = \{X \in \mathfrak{g} \mid X^*(p) = 0\}.$$

The following technical proposition will be needed. It is shown in Marsden and Ratiu [77] (Chapter 9, Proposition 9.3.6 and Lemma 9.3.7).

Proposition 22.19. *Given a smooth action $\varphi \colon G \times M \to M$ of a Lie group on a manifold M, the following properties hold.*

(1) For every $X \in \mathfrak{g}$, we have

$$(\mathrm{Ad}_g X)^* = \tau_{g^{-1}}^* X^* = (\tau_g)_* X^*, \quad \textit{for every } g \in G,$$

where $\tau_{g^{-1}}^*$ is the pullback associated with $\tau_{g^{-1}}$, and $(\tau_g)_*$ is the push-forward associated with τ_g. This is equivalent to

$$(\mathrm{Ad}_g X)^*(p) = (d\tau_g)_{g^{-1}\cdot p} X^*(g^{-1} \cdot p), \qquad p \in M.$$

(2) The map $X \mapsto X^*$ from $\mathfrak{g}$ to $\mathfrak{X}(M)$ is a Lie algebra anti-homomorphism, which means that

$$[X^*, Y^*] = -[X, Y]^* \quad \textit{for all } X, Y \in \mathfrak{g}.$$

Remark: If the metric on M is G-invariant (that is, every τ_g is an isometry of M), then the vector field X^* is a Killing vector field on M for every $X \in \mathfrak{g}$.

Given a pair (G, H), where G is a Lie group and H is a closed subgroup of G, it turns out that there is a criterion for the existence of some G-invariant metric on the homogeneous space G/H in terms of a certain representation of H called the isotropy representation. Let us explain what this representation is.

Recall that G acts on the left on G/H *via*

$$g_1 \cdot (g_2 H) = g_1 g_2 H, \quad g_1, g_2 \in G.$$

For any $g_1 \in G$, the diffeomorphism $\tau_{g_1} \colon G/H \to G/H$ is left coset multiplication, given by

$$\tau_{g_1}(g_2 H) = g_1 \cdot (g_2 H) = g_1 g_2 H.$$

In this situation, Part (1) of Proposition 22.19 is easily proved as follows.

Proposition 22.20. *For any $X \in \mathfrak{g}$ and any $g \in G$, we have*

$$(\tau_g)_* X^* = (\mathrm{Ad}_g(X))^*.$$

Proof. By definition, for any $p = bH$, we have $\tau_g(bH) = gbH$, and

$$\begin{aligned}
((\tau_g)_* X^*)_{\tau_g(p)} &= (d\tau_g)_p (X^*(p)) \\
&= \frac{d}{dt}(g \exp(tX) bH)\Big|_{t=0} \\
&= \frac{d}{dt}(g \exp(tX) g^{-1} gbH)\Big|_{t=0} \\
&= \frac{d}{dt}(\exp(t\mathrm{Ad}_g(X)) gbH)\Big|_{t=0} \\
&= (\mathrm{Ad}_g(X))^*_{\tau_g(p)},
\end{aligned}$$

which shows that $(\tau_g)_* X^* = (\mathrm{Ad}_g(X))^*$. □

Denote the point in G/H corresponding to the coset $1H = H$ by o. Then we have the map

$$\chi^{G/H} \colon H \to \mathbf{GL}(T_o(G/H)),$$

given by

$$\chi^{G/H}(h) = (d\tau_h)_o, \quad \text{for all } h \in H.$$

Using the same kind of technique that we used in proving that $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ is a homomorphism (just before Proposition 18.1), we can prove that $\chi^{G/H} \colon H \to \mathbf{GL}(T_o(G/H))$ is a homomorphism.

Definition 22.7. The homomorphism $\chi^{G/H}$ is called the *isotropy representation* of the homogeneous space G/H.

The homomorphism $\chi^{G/H}$ is a representation of the group H, and since we can view H as the isotropy group (the stabilizer) of the element $o \in G/H$ corresponding to the coset H, it makes sense to call it the isotropy representation. It is not easy to deal with the isotropy representation directly. Fortunately, the isotropy representation is *equivalent* to another representation $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ obtained from the representation $\mathrm{Ad}\colon G \to \mathbf{GL}(\mathfrak{g})$ by a quotient process that we now describe.

Recall that $\mathbf{Ad}_{g_1}(g_2) = g_1 g_2 g_1^{-1}$ for all $g_1, g_2 \in G$, and that the canonical projection $\pi\colon G \to G/H$ is given by $\pi(g) = gH$. Then following O'Neill [91] (see Proposition 22, Chapter 11), observe that

$$\tau_h \circ \pi = \pi \circ \mathbf{Ad}_h \quad \text{for all } h \in H,$$

since $h \in H$ implies that $h^{-1}H = H$, so for all $g \in G$,

$$(\tau_h \circ \pi)(g) = hgH = hgh^{-1}H = (\pi \circ \mathbf{Ad}_h)(g).$$

By taking derivatives at 1, we get

$$(d\tau_h)_o \circ d\pi_1 = d\pi_1 \circ \mathrm{Ad}_h,$$

which is equivalent to the commutativity of the diagram

$$\begin{tikzcd}
\mathfrak{g} \arrow[r, "\mathrm{Ad}_h"] \arrow[d, "d\pi_1"'] & \mathfrak{g} \arrow[d, "d\pi_1"] \\
T_o(G/H) \arrow[r, "(d\tau_h)_o"'] & T_o(G/H).
\end{tikzcd}$$

For any $X \in \mathfrak{g}$, we can express $d\pi_1(X)$ in terms of the vector field X^* introduced in Definition 22.6. Indeed, to compute $d\pi_1(X)$, we can use the curve $t \mapsto \exp(tX)$, and we have

$$d\pi_1(X) = \frac{d}{dt}(\pi(\exp(tX)))\bigg|_{t=0} = \frac{d}{dt}(\exp(tX)H)\bigg|_{t=0} = X^*_o.$$

For every $X \in \mathfrak{h}$, since the curve $t \mapsto \exp(tX)H$ in G/H has the *constant value* o, we see that

$$\mathrm{Ker}\, d\pi_1 = \mathfrak{h},$$

and thus, $d\pi_1$ factors through $\mathfrak{g}/\mathfrak{h}$ as $d\pi_1 = \varphi \circ \pi_{\mathfrak{g}/\mathfrak{h}}$, where $\pi_{\mathfrak{g}/\mathfrak{h}}\colon \mathfrak{g} \to \mathfrak{g}/\mathfrak{h}$ is the quotient map and $\varphi\colon \mathfrak{g}/\mathfrak{h} \to T_o(G/H)$ is the isomorphism given by the First Isomorphism theorem. Explicitly, the map φ is given by $\varphi(X + \mathfrak{h}) = d\pi_1(X)$, for all $X \in \mathfrak{g}$. Since Ad_h is an isomorphism, the kernel of the map $\pi_{\mathfrak{g}/\mathfrak{h}} \circ \mathrm{Ad}_h$ is $\mathfrak{h}$, and by the First Isomophism theorem there is a unique map $\mathrm{Ad}_h^{G/H}\colon \mathfrak{g}/\mathfrak{h} \to \mathfrak{g}/\mathfrak{h}$ such that

$$\pi_{\mathfrak{g}/\mathfrak{h}} \circ \mathrm{Ad}_h = \mathrm{Ad}_h^{G/H} \circ \pi_{\mathfrak{g}/\mathfrak{h}}$$

making the following diagram commute:

$$\begin{tikzcd}
\mathfrak{g} \arrow[r, "\mathrm{Ad}_h"] \arrow[d, "\pi_{\mathfrak{g}/\mathfrak{h}}"'] & \mathfrak{g} \arrow[d, "\pi_{\mathfrak{g}/\mathfrak{h}}"] \\
\mathfrak{g}/\mathfrak{h} \arrow[r, "\mathrm{Ad}_h^{G/H}"'] & \mathfrak{g}/\mathfrak{h}.
\end{tikzcd}$$

Explicitly, the map $\mathrm{Ad}_h^{G/H}$ is given by $\mathrm{Ad}_h^{G/H}(X + \mathfrak{h}) = (\pi_{\mathfrak{g}/\mathfrak{h}} \circ \mathrm{Ad}_h)(X)$ for all $X \in \mathfrak{g}$. Then we have the following diagram in which the outermost rectangle commutes and the upper rectangle commutes:

$$\begin{tikzcd}
\mathfrak{g} \arrow[r, "\mathrm{Ad}_h"] \arrow[d, "\pi_{\mathfrak{g}/\mathfrak{h}}"'] & \mathfrak{g} \arrow[d, "\pi_{\mathfrak{g}/\mathfrak{h}}"] \\
\mathfrak{g}/\mathfrak{h} \arrow[r, "\mathrm{Ad}_h^{G/H}"] \arrow[d, "\varphi"'] & \mathfrak{g}/\mathfrak{h} \arrow[d, "\varphi"'] \\
T_o(G/H) \arrow[r, "(d\tau_h)_o"'] & T_o(G/H).
\end{tikzcd}$$

Since $\pi_{\mathfrak{g}/\mathfrak{h}}$ is surjective, it follows that the lower rectangle commutes; that is

$$\begin{tikzcd}
\mathfrak{g}/\mathfrak{h} \arrow[r, "\mathrm{Ad}_h^{G/H}"] \arrow[d, "\varphi"'] & \mathfrak{g}/\mathfrak{h} \arrow[d, "\varphi"] \\
T_o(G/H) \arrow[r, "(d\tau_h)_o"'] & T_o(G/H)
\end{tikzcd}$$

commutes. Observe that $\mathrm{Ad}_h^{G/H}$ is a linear isomorphism of $\mathfrak{g}/\mathfrak{h}$ for every $h \in H$, so that the map $\mathrm{Ad}^{G/H}\colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ is a representation of H. This proves the first part of the following proposition.

Proposition 22.21. *Let (G, H) be a pair where G is a Lie group and H is a closed subgroup of G. The following properties hold:*

(1) The representations $\chi^{G/H} \colon H \to \mathbf{GL}(T_o(G/H))$ and $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ are equivalent; this means that for every $h \in H$, we have the commutative diagram

$$\begin{tikzcd}
\mathfrak{g}/\mathfrak{h} \arrow[r, "\mathrm{Ad}_h^{G/H}"] \arrow[d, "\varphi"'] & \mathfrak{g}/\mathfrak{h} \arrow[d, "\varphi"] \\
T_o(G/H) \arrow[r, "(d\tau_h)_o"'] & T_o(G/H),
\end{tikzcd}$$

where the isomorphism $\varphi \colon \mathfrak{g}/\mathfrak{h} \to T_o(G/H)$ and the quotient map $\mathrm{Ad}_h^{G/H} \colon \mathfrak{g}/\mathfrak{h} \to \mathfrak{g}/\mathfrak{h}$ are defined as above.

(2) The homogeneous space G/H has some G-invariant metric iff the closure of $\mathrm{Ad}^{G/H}(H)$ is compact in $\mathbf{GL}(\mathfrak{g}/\mathfrak{h})$. Furthermore, this metric is unique up to a scalar if the isotropy representation is irreducible.

We just proved the first part, which is Proposition 2.40 of Gallot, Hulin, Lafontaine [49] (Chapter 2, Section A). The proof of the second part is very similar to the proof of Theorem 20.5; see Gallot, Hulin, Lafontaine [49] (Chapter 2, Theorem 2.42).

The representation $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ which involves the quotient algebra $\mathfrak{g}/\mathfrak{h}$ is hard to deal with. To make things more tractable, it is natural to assume that $\mathfrak{g}$ splits as a direct sum $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ for some *well-behaved* subspace $\mathfrak{m}$ of $\mathfrak{g}$, so that $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$.

Definition 22.8. Let (G, H) be a pair where G is a Lie group and H is a closed subgroup of G. We say that the homogeneous space G/H is *reductive* if there is some subspace $\mathfrak{m}$ of $\mathfrak{g}$ such that

$$\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m},$$

and

$$\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m} \quad \text{for all } h \in H.$$

See Figure 22.3.

Observe that unlike $\mathfrak{h}$, which is a Lie subalgebra of $\mathfrak{g}$, the subspace $\mathfrak{m}$ is *not necessarily closed* under the Lie bracket, so in general it *is not* a Lie algebra. Also, since $\mathfrak{m}$ is finite-dimensional and since Ad_h is an isomorphism, we actually have $\mathrm{Ad}_h(\mathfrak{m}) = \mathfrak{m}$.

Definition 22.8 allows us to deal with $\mathfrak{g}/\mathfrak{h}$ in a tractable manner, but does not provide any means of defining a metric on G/H. We would like to define G-invariant metrics on G/H and a key property of a reductive spaces is that there is a criterion for the existence of G-invariant metrics on G/H in terms of $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$.

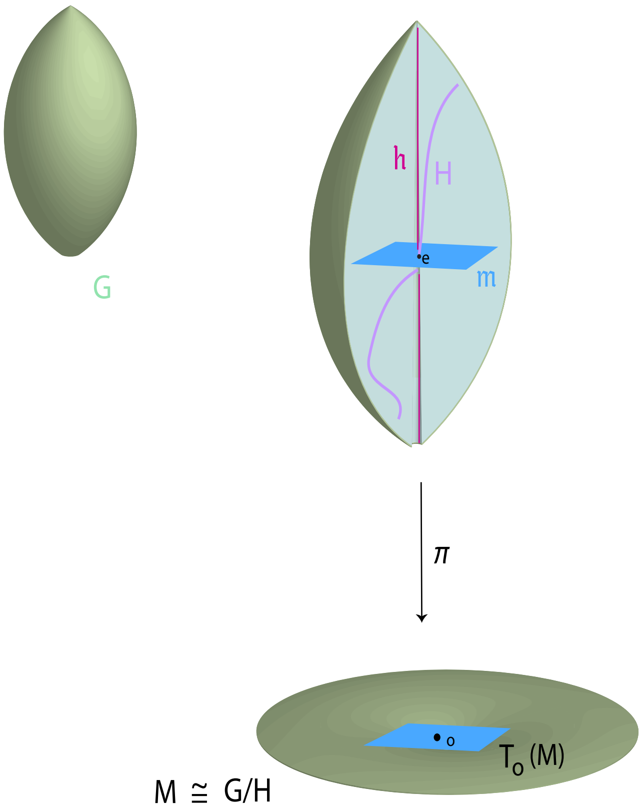


Figure 22.3: A schematic illustration of a reductive homogeneous manifold. Note that $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ and that $T_o(M) \cong \mathfrak{m}$ via $d\pi_1$.

Since $\mathfrak{g}/\mathfrak{h}$ is isomorphic to $\mathfrak{m}$, by the reasoning just before Proposition 22.21, the map $d\pi_1 \colon \mathfrak{g} \to T_o(G/H)$ restricts to an isomorphism between $\mathfrak{m}$ and $T_o(G/H)$ (where o denotes the point in G/H corresponding to the coset H). The representation $\mathrm{Ad}^{G/H} \colon H \to \mathbf{GL}(\mathfrak{g}/\mathfrak{h})$ becomes the representation $\mathrm{Ad} \colon H \to \mathbf{GL}(\mathfrak{m})$, where Ad_h is the restriction of Ad_h to $\mathfrak{m}$ for every $h \in H$.

We also know that for any $X \in \mathfrak{g}$, we can express $d\pi_1(X)$ in terms of the vector field X^* introduced in Definition 22.6 by

$$d\pi_1(X) = X^*_o,$$

and that

$$\mathrm{Ker}\, d\pi_1 = \mathfrak{h}.$$

Thus, the *restriction* of $d\pi_1$ to $\mathfrak{m}$ is an isomorphism onto $T_o(G/H)$, given by $X \mapsto X^*_o$. Also, for every $X \in \mathfrak{g}$, since $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, we can write $X = X_{\mathfrak{h}} + X_{\mathfrak{m}}$, for some unique $X_{\mathfrak{h}} \in \mathfrak{h}$ and some unique $X_{\mathfrak{m}} \in \mathfrak{m}$, and

$$d\pi_1(X) = d\pi_1(X_{\mathfrak{m}}) = X^*_o.$$

We use the isomorphism $d\pi_1$ to transfer any inner product $\langle -, -\rangle_{\mathfrak{m}}$ on $\mathfrak{m}$ to an inner product $\langle -, -\rangle$ on $T_o(G/H)$, and vice-versa, by stating that

$$\langle X, Y\rangle_{\mathfrak{m}} = \langle X^*_o, Y^*_o\rangle, \quad \text{for all } X, Y \in \mathfrak{m};$$

that is, by declaring $d\pi_1$ to be an isometry between $\mathfrak{m}$ and $T_o(G/H)$. See Figure 22.3.

If the metric on G/H is G-invariant, then the map $p \mapsto \exp(tX) \cdot p = \exp(tX)aH$ (with $p = aH \in G/H$, $a \in G$) is an isometry of G/H for every $t \in \mathbb{R}$, so by Proposition 17.9, X^* is a Killing vector field. This fact is needed in Section 22.6.

Proposition 22.22. *Let (G, H) be a pair of Lie groups defining a reductive homogeneous space $M = G/H$, with reductive decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$. The following properties hold:*

(1) The isotropy representation $\chi^{G/H} \colon H \to \mathbf{GL}(T_o(G/H))$ is equivalent to the representation $\mathrm{Ad}\colon H \to \mathbf{GL}(\mathfrak{m})$ (where Ad_h is restricted to $\mathfrak{m}$ for every $h \in H$); this means that for every $h \in H$, we have the commutative diagram

$$\begin{tikzcd}
\mathfrak{m} \arrow[r, "\mathrm{Ad}_h"] \arrow[d, "d\pi_1"'] & \mathfrak{m} \arrow[d, "d\pi_1"] \\
T_o(G/H) \arrow[r, "(d\tau_h)_o"'] & T_o(G/H),
\end{tikzcd}$$

where $d\pi_1 \colon \mathfrak{m} \to T_o(G/H)$ is the isomorphism induced by the canonical projection $\pi \colon G \to G/H$.

(2) By making $d\pi_1$ an isometry between $\mathfrak{m}$ and $T_o(G/H)$ (as explained above), there is a one-to-one correspondence between G-invariant metrics on G/H and $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$ (inner products $\langle -, -\rangle_{\mathfrak{m}}$ such that

$$\langle u, v\rangle_{\mathfrak{m}} = \langle \mathrm{Ad}_h(u), \mathrm{Ad}_h(v)\rangle_{\mathfrak{m}}, \quad \textit{for all } h \in H \textit{ and all } u, v \in \mathfrak{m}).$$

(3) The homogeneous space G/H has some G-invariant metric iff the closure of $\mathrm{Ad}(H)$ is compact in $\mathbf{GL}(\mathfrak{m})$. In particular, if H is compact, then a G-invariant metric on G/H always exists. Furthermore, if the representation $\mathrm{Ad}\colon H \to \mathbf{GL}(\mathfrak{m})$ is irreducible, then such a metric is unique up to a scalar.

Proof. Part (1) follows immediately from the fact that $\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $h \in H$ and from the identity

$$(d\tau_h)_o \circ d\pi_1 = d\pi_1 \circ \mathrm{Ad}_h,$$

which was proved just before Proposition 22.21. Part (2) is proved in O'Neill [91] (Chapter 11, Proposition 22), Arvanitoyeorgos [11] (Chapter 5, Proposition 5.1), and Ziller [119] (Chapter

6, Lemma 6.22). Since the proof is quite informative, we provide it. First assume that the metric on G/H is G-invariant. By restricting this G-invariant metric to the tangent space at o, we will show the existence of a metric on $\mathfrak{m}$ obeys the property of $\mathrm{Ad}(H)$ invariance. For every $h \in H$, the map τ_h is an isometry of G/H, so in particular we have

$$\langle (d\tau_h)_o(X^*_o), (d\tau_h)_o(Y^*_o)\rangle = \langle X^*_o, Y^*_o\rangle, \quad \text{for all } X, Y \in \mathfrak{m}.$$

However, the commutativity of the diagram in (1) can be expressed as

$$(d\tau_h)_o(X^*_o) = (\mathrm{Ad}_h(X))^*_o,$$

so we get

$$\langle (\mathrm{Ad}_h(X))^*_o, (\mathrm{Ad}_h(Y))^*_o\rangle = \langle X^*_o, Y^*_o\rangle,$$

which is equivalent to

$$\langle \mathrm{Ad}_h(X), \mathrm{Ad}_h(Y)\rangle_{\mathfrak{m}} = \langle X, Y\rangle_{\mathfrak{m}}, \quad \text{for all } X, Y \in \mathfrak{m}.$$

Conversely, assume we have an inner product $\langle -, -\rangle_{\mathfrak{m}}$ on $\mathfrak{m}$ which is $\mathrm{Ad}(H)$-invariant. The proof strategy is as follows: place the metric on $T_o(G/H)$ and then use the maps $\tau_g : G/H \to G/H$ to transfer this metric around G/H in a fashion that is consistent with the notion of G-invariance. The condition of $\mathrm{Ad}(H)$-invariance ensures that this construction of the metric on G/H is well defined.

First we transfer this metric on $T_o(G/H)$ using the isomorphism $d\pi_1$ between $\mathfrak{m}$ and $T_o(G/H)$. Since $(d\tau_a)_o\colon T_o(G/H) \to T_p(G/H)$ is a linear isomorphism with inverse $(d\tau_{a^{-1}})_p$, for any $p = aH$, we define a metric on G/H as follows: for every $p \in G/H$, for any coset representative aH of p, set

$$\langle u, v\rangle_p = \langle (d\tau_{a^{-1}})_p(u), (d\tau_{a^{-1}})_p(v)\rangle_o, \quad \text{for all } u, v \in T_p(G/H).$$

We need to show that the above does not depend on the representative aH chosen for p. This is where we make use of the $\mathrm{Ad}(H)$-invariant condition. By reversing the computation that we just made, each map $(d\tau_h)_o$ is an isometry of $T_o(G/H)$. If bH is another representative for p, so that $aH = bH$, then $b^{-1}a = h$ for some $h \in H$, so $b^{-1} = ha^{-1}$, and we have

$$\begin{aligned}\langle (d\tau_{b^{-1}})_p(u), (d\tau_{b^{-1}})_p(v)\rangle_o &= \langle (d\tau_h)_o((d\tau_{a^{-1}})_p(u)), (d\tau_h)_o((d\tau_{a^{-1}})_p(v))\rangle_o \\ &= \langle (d\tau_{a^{-1}})_p(u), (d\tau_{a^{-1}})_p(v)\rangle_o,\end{aligned}$$

since $(d\tau_h)_o$ is an isometry.

To prove that the metric that we defined is smooth, we use a result that will be proved later, so this part of the proof can be skipped during a first reading. Since G is a principal H-bundle over G/H (see Theorem 22.12), for every $p \in G/H$, there is a local trivialization $\varphi_\alpha\colon \pi^{-1}(U_\alpha) \to U_\alpha \times H$, where U_α is some open subset in G/H containing p, so smooth local sections over U_α exist (for example, pick some $h \in H$ and define $s\colon U_\alpha \to \pi^{-1}(U_\alpha)$ by

$s(q) = \varphi_\alpha^{-1}(q, h)$, for all $q \in U_\alpha$). Given any smooth local section s over U_α (as $s(q) \in G$ and $q = \pi(s(q)) = s(q)H$), we have

$$\langle u, v\rangle_q = \langle (d\tau_{s(q)^{-1}})_q(u), d\tau_{s(q)^{-1}})_q(u)\rangle_o, \quad \text{for all } q \in U_\alpha \text{ and all } u, v \in T_q(G/H),$$

which shows that the resulting metric on G/H is smooth. By definition, the metric that we just defined is G-invariant.

Part (3) is shown in Gallot, Hulin, Lafontaine [49] (Chapter 2, Theorem 2.42). □

At this stage we have a mechanism to equip G/H with a Riemannian metric from an inner product $\mathfrak{m}$ which has the special property of being $\mathrm{Ad}(H)$-invariant, but this mechanism does *not* provide a Riemannian metric on G. The construction of a Riemannian metric on G can be done by extending the $\mathrm{Ad}(H)$-invariant metric on $\mathfrak{m}$ to all of $\mathfrak{g}$, and using the bijective correspondence between left-invariant metrics on a Lie group G, and inner products on its Lie algebra $\mathfrak{g}$ given by Proposition 20.1.

Proposition 22.23. *Let (G, H) be a pair of Lie groups defining a reductive homogeneous space $M = G/H$, with reductive decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$. If $\mathfrak{m}$ has some $\mathrm{Ad}(H)$-invariant inner product $\langle -, -\rangle_{\mathfrak{m}}$, for any inner product $\langle -, -\rangle_{\mathfrak{g}}$ on $\mathfrak{g}$ extending $\langle -, -\rangle_{\mathfrak{m}}$ such that $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal, if we give G the left-invariant metric induced by $\langle -, -\rangle_{\mathfrak{g}}$, then the map $\pi\colon G \to G/H$ is a Riemannian submersion.*

Proof. (After O'Neill [91] (Chapter 11, Lemma 24). The map $\pi\colon G \to G/H$ is clearly a smooth submersion. For Condition (2) of Definition 17.3, for all $a, b \in G$, since

$$\tau_a(\pi(b)) = \tau_a(bH) = abH = L_a(b)H = \pi(L_a(b)),$$

we have

$$\tau_a \circ \pi = \pi \circ L_a,$$

and by taking derivatives at 1, we get

$$d(\tau_a)_o \circ d\pi_1 = d\pi_a \circ (dL_a)_1.$$

The horizontal subspace at $a \in G$ is $\mathcal{H}_a = (dL_a)_1(\mathfrak{m})$, and since the metric on G is left-invariant, $(dL_a)_1$ is an isometry; the map $d(\tau_a)_o$ is an isometry because the metric on G/H is G-invariant, and $d\pi_1$ is an isometry between $\mathfrak{m}$ and $T_o(G/H)$ by construction, so

$$d\pi_a = (d\tau_a)_o \circ d\pi_1 \circ (dL_a^{-1})_1$$

is an isometry between $\mathcal{H}_a$ and $T_p(G/H)$, where $p = aH$. □

By Proposition 17.8, a Riemannian submersion carries horizontal geodesics to geodesics.

22.5 Examples of Reductive Homogeneous Spaces

We now apply the theory of Propositions 22.22 and 22.23 to construct a family of reductive homogeneous spaces, the Stiefel manifolds $S(k,n)$. We first encountered the Stiefel manifolds in Section 4.4. For any $n \geq 1$ and any k with $1 \leq k \leq n$, let $S(k,n)$ be the set of all orthonormal k-frames, where an orthonormal k-frame is a k-tuples of orthonormal vectors $(u_1, \ldots, u_k)$ with $u_i \in \mathbb{R}^n$. Recall that $\mathbf{SO}(n)$ acts transitively on $S(k,n)$ via the action $\cdot \colon \mathbf{SO}(n) \times S(k,n) \to S(k,n)$

$$R \cdot (u_1, \ldots, u_k) = (Ru_1, \ldots, Ru_k).$$

and that the stabilizer of this action is

$$H = \left\{ \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \;\middle|\; R \in \mathbf{SO}(n-k) \right\}.$$

Theorem 22.13 implies that $S(k,n) \cong G/H$, with $G = \mathbf{SO}(n)$ and $H \cong \mathbf{SO}(n-k)$. Observe that the points of $G/H \cong S(k,n)$ are the cosets QH, with $Q \in \mathbf{SO}(n)$; that is, the equivalence classes $[Q]$, with the equivalence relation on $\mathbf{SO}(n)$ given by

$$Q_1 \equiv Q_2 \quad \text{iff} \quad Q_2 = Q_1\widetilde{R}, \text{ for some } \widetilde{R} \in H.$$

If we write $Q = [Y\, Y_\perp]$, where Y consists of the first k columns of Q and $Y_\perp$ consists of the last $n-k$ columns of Q, it is clear that $[Q]$ is uniquely determined by Y. In fact, if $P_{n,k}$ denotes the projection matrix consisting of the first k columns of the identity matrix I_n,

$$P_{n,k} = \begin{pmatrix} I_k \\ 0_{n-k,k} \end{pmatrix},$$

for any $Q = [Y\, Y_\perp]$, the unique representative Y of the equivalence class $[Q]$ is given by

$$Y = QP_{n,k}.$$

Furthermore $Y_\perp$ is characterized by the fact that $Q = [Y\, Y_\perp]$ is orthogonal, namely, $YY^\top + Y_\perp Y_\perp^\top = I$.

Define

$$\mathfrak{h} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(n-k) \right\}, \qquad \mathfrak{m} = \left\{ \begin{pmatrix} T & -A^\top \\ A & 0 \end{pmatrix} \;\middle|\; T \in \mathfrak{so}(k),\ A \in \mathrm{M}_{n-k,k}(\mathbb{R}) \right\}.$$

Clearly $\mathfrak{g} = \mathfrak{so}(n) = \mathfrak{h} \oplus \mathfrak{m}$. For $h \in H$ with $h = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix}$, note that $h^{-1} = \begin{pmatrix} I & 0 \\ 0 & R^\top \end{pmatrix}$. Given any $X \in \mathfrak{m}$ with $X = \begin{pmatrix} T & -A^\top \\ A & 0 \end{pmatrix}$, we see that

$$hXh^{-1} = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} T & -A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & R^\top \end{pmatrix} = \begin{pmatrix} T & -A^\top R^\top \\ RA & 0 \end{pmatrix} \in \mathfrak{m},$$

which implies that $\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$. Therefore Definition 22.8 shows that $S(k,n) \cong G/H$ is a reductive homogeneous manifold with $\mathfrak{g}/\mathfrak{h} \cong \mathfrak{m}$.

Since $H \cong \mathbf{SO}(n-k)$ is compact, Proposition 22.22 guarantees the existence of a G-invariant metric on G/H, which in turn ensures the existence of an $\mathrm{Ad}(H)$-invariant metric on $\mathfrak{m}$. Theorem 20.27 implies that we may construct such a metric by using the Killing form on $\mathfrak{so}(n)$. We know that the Killing form on $\mathfrak{so}(n)$ is given by $B(X,Y) = (n-2)\mathrm{tr}(XY)$. Now observe that if take $\begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \in \mathfrak{h}$ and $\begin{pmatrix} T & -A^\top \\ A & 0 \end{pmatrix} \in \mathfrak{m}$, then

$$\mathrm{tr}\left(\begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix}\begin{pmatrix} T & -A^\top \\ A & 0 \end{pmatrix}\right) = \mathrm{tr}\begin{pmatrix} 0 & 0 \\ SA & 0 \end{pmatrix} = 0.$$

Furthermore, it is clear that $\dim(\mathfrak{m}) = \dim(\mathfrak{g}) - \dim(\mathfrak{h})$, so $\mathfrak{m}$ is the orthogonal complement of $\mathfrak{h}$ with respect to the Killing form. If $X, Y \in \mathfrak{m}$, with

$$X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} T & -B^\top \\ B & 0 \end{pmatrix},$$

observe that

$$\mathrm{tr}\left(\begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}\begin{pmatrix} T & -B^\top \\ B & 0 \end{pmatrix}\right) = \mathrm{tr}\begin{pmatrix} ST - A^\top B & -SB^\top \\ AT & -AB^\top \end{pmatrix} = \mathrm{tr}(ST) - 2\mathrm{tr}(A^\top B),$$

and since $S^\top = -S$, we have

$$\mathrm{tr}(ST) - 2\mathrm{tr}(A^\top B)) = -\mathrm{tr}(S^\top T) - 2\mathrm{tr}(A^\top B),$$

so we define an inner product on $\mathfrak{m}$ by

$$\langle X, Y\rangle = -\frac{1}{2}\mathrm{tr}(XY) = \frac{1}{2}\mathrm{tr}(X^\top Y) = \frac{1}{2}\mathrm{tr}(S^\top T) + \mathrm{tr}(A^\top B).$$

We give $\mathfrak{h}$ the same inner product. For $X, Y \in \mathfrak{m}$ as defined above, and $h = \begin{pmatrix} I & 0 \\ 0 & R \end{pmatrix} \in H$, we have

$$\begin{aligned} \mathrm{Ad}_h(X) = hXh^{-1} &= \begin{pmatrix} S & -A^\top R^\top \\ RA & 0 \end{pmatrix} \\ \mathrm{Ad}_h(Y) = hYh^{-1} &= \begin{pmatrix} T & -B^\top R^\top \\ RB & 0 \end{pmatrix}. \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{tr}(\mathrm{Ad}_h(X)\mathrm{Ad}_h(Y)) &= \mathrm{tr}\begin{pmatrix} ST - A^\top B & -SB^\top R^\top \\ RAT & -RAB^\top R^\top \end{pmatrix} \\ &= \mathrm{tr}(ST) - \mathrm{tr}(A^\top B) - \mathrm{tr}(RAB^\top R^\top) \\ &= \mathrm{tr}(ST) - \mathrm{tr}(A^\top B) - \mathrm{tr}(AB^\top R^\top R) \\ &= \mathrm{tr}(ST) - \mathrm{tr}(A^\top B) - \mathrm{tr}(AB^\top) \\ &= \mathrm{tr}(ST) - 2\mathrm{tr}(A^\top B) = \mathrm{tr}(XY), \end{aligned}$$

and this shows that the inner product defined on $\mathfrak{m}$ is $\mathrm{Ad}(H)$-invariant.

We summarize all this in the following proposition.

Proposition 22.24. *If $X, Y \in \mathfrak{m}$, with*

$$X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} T & -B^\top \\ B & 0 \end{pmatrix},$$

then the fomula

$$\langle X, Y\rangle = -\frac{1}{2}\mathrm{tr}(XY) = \frac{1}{2}\mathrm{tr}(S^\top T) + \mathrm{tr}(A^\top B)$$

defines an $\mathrm{Ad}(H)$-invariant inner product on $\mathfrak{m}$. If we give $\mathfrak{h}$ the same inner product so that $\mathfrak{g}$ also has the inner product $\langle X, Y\rangle = -\frac{1}{2}\mathrm{tr}(XY)$, then $\mathfrak{m}$ and $\mathfrak{h}$ are orthogonal.

Observe that there is a bijection between the space $\mathfrak{m}$ of $n \times n$ matrices of the form

$$X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}$$

and the set of $n \times k$ matrices of the form

$$\Delta_1 = \begin{pmatrix} S \\ A \end{pmatrix},$$

but the inner product given by

$$\langle \Delta_1, \Delta_2\rangle = \mathrm{tr}(\Delta_1^\top \Delta_2),$$

where

$$\Delta_2 = \begin{pmatrix} T \\ B \end{pmatrix},$$

yields

$$\langle \Delta_1, \Delta_2\rangle = \mathrm{tr}(S^\top T) + \mathrm{tr}(A^\top B),$$

without the factor 1/2 in front of $S^\top T$. These metrics are different.

The vector space $\mathfrak{m}$ is the tangent space $T_o S(k, n)$ to $S(k, n)$ at $o = [H]$, the coset of the point corresponding to H. For any other point $[Q] \in G/H \cong S(k, n)$, the tangent space $T_{[Q]}S(k, n)$ is given by

$$T_{[Q]}S(k, n) = \left\{ Q \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(k),\ A \in \mathrm{M}_{n-k,k}(\mathbb{R}) \right\}.$$

Using the decomposition $Q = [Y\, Y_\perp]$, where Y consists of the first k columns of Q and $Y_\perp$ consists of the last $n - k$ columns of Q, we have

$$[Y\, Y_\perp] \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} = [YS + Y_\perp A \;\; -YA^\top].$$

If we write $\widetilde{X} = YS + Y_\perp A$, since Y and $Y_\perp$ are parts of an orthogonal matrix, we have $Y_\perp^\top Y = 0_{n-k,k}$, $Y^\top Y = I_k$, and $Y_\perp^\top Y_\perp = I_{n-k}$, so we can recover A from $\widetilde{X}$ and $Y_\perp$ and S from $\widetilde{X}$ and Y, by

$$Y_\perp^\top \widetilde{X} = Y_\perp^\top (YS + Y_\perp A) = A,$$

and

$$Y^\top \widetilde{X} = Y^\top (YS + Y_\perp A) = S.$$

Since $A = Y_\perp^\top \widetilde{X}$, we also have $A^\top A = \widetilde{X}^\top Y_\perp Y_\perp^\top \widetilde{X} = \widetilde{X}^\top (I - YY^\top)\widetilde{X}$.

Therefore, given $Q = [Y\, Y_\perp]$, the $n \times n$ matrices

$$\widehat{X} = [Y\, Y_\perp] \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \tag{$*$}$$

are in one-to-one correspondence with the $n \times k$ matrices of the form $\widetilde{X} = YS + Y_\perp A$.

Since Y describes an element of $S(k,n)$, we can say that the tangent vectors to $S(k,n)$ at Y are of the form

$$\widetilde{X} = YS + Y_\perp A, \quad S \in \mathfrak{so}(k),\ A \in \mathrm{M}_{n-k,k}(\mathbb{R}).$$

Since $[Y\ Y_\perp]$ is an orthogonal matrix, we get $Y^\top \widetilde{X} = S$, which shows that $Y^\top \widetilde{X}$ is skew-symmetric. Conversely, since the columns of $[Y\ Y_\perp]$ form an orthonormal basis of $\mathbb{R}^n$, every $n \times k$ matrix $\widetilde{X}$ can be written as

$$\widetilde{X} = \begin{pmatrix} Y & Y_\perp \end{pmatrix} \begin{pmatrix} S \\ A \end{pmatrix} = YS + Y_\perp A,$$

where $S \in \mathrm{M}_{k,k}(\mathbb{R})$ and $A \in \mathrm{M}_{n-k,k}(\mathbb{R})$, and if $Y^\top \widetilde{X}$ is skew-symmetric, then $S = Y^\top \widetilde{X}$ is also skew-symmetric. Therefore, the tangent vectors to $S(k,n)$ at Y are the vectors $\widetilde{X} \in \mathrm{M}_{n,k}(\mathbb{R})$ such that $Y^\top \widetilde{X}$ is skew-symmetric. This is the description given in Edelman, Arias and Smith [44].

Another useful observation is that if $\widetilde{X} = YS + Y_\perp A$ is a tangent vector to $S(k,n)$ at Y, then the square norm $\langle \widehat{X}, \widehat{X} \rangle$ (in the canonical metric) is given by

$$\langle \widehat{X}, \widehat{X} \rangle = \mathrm{tr}\Big(\widetilde{X}^\top \Big(I - \frac{1}{2} YY^\top\Big) \widetilde{X}\Big),$$

where $\widehat{X}$ is the matrix defined in $(*)$. Indeed, we have

$$\begin{aligned}\widetilde{X}^\top\Big(I-\frac{1}{2}YY^\top\Big)\widetilde{X} &= (S^\top Y^\top + A^\top Y_\perp^\top)\Big(I-\frac{1}{2}YY^\top\Big)(YS+Y_\perp A)\\ &= \Big(S^\top Y^\top + A^\top Y_\perp^\top - \frac{1}{2}S^\top Y^\top YY^\top - \frac{1}{2}A^\top Y_\perp^\top YY^\top\Big)(YS+Y_\perp A)\\ &= \Big(\frac{1}{2}S^\top Y^\top + A^\top Y_\perp^\top\Big)(YS+Y_\perp A)\\ &= \frac{1}{2}S^\top Y^\top YS + A^\top Y_\perp^\top Y_\perp A + \frac{1}{2}S^\top Y^\top Y_\perp A + A^\top Y_\perp^\top YS\\ &= \frac{1}{2}S^\top S + A^\top A.\end{aligned}$$

But then

$$\mathrm{tr}\Big(\widetilde{X}^\top\Big(I-\frac{1}{2}YY^\top\Big)\widetilde{X}\Big) = \frac{1}{2}\mathrm{tr}(S^\top S) + \mathrm{tr}(A^\top A) = \langle \widehat{X}, \widehat{X}\rangle,$$

because

$$\widehat{X} = [Y\, Y_\perp]\begin{pmatrix} S & -A^\top \\ A & 0\end{pmatrix}$$

and the matrix $[Y\, Y_\perp]$ is orthogonal, as claimed. By polarization we find that the canonical metric is given by

$$\langle X_1, X_2\rangle = \mathrm{tr}\Big(X_1^\top\Big(I-\frac{1}{2}YY^\top\Big)X_2\Big).$$

In that paper it is also observed that because $Y_\perp$ has rank $n-k$ (since $Y_\perp^\top Y_\perp = I$), for every $(n-k)\times k$ matrix A, there is some $n\times k$ matrix C such that $A = Y_\perp^\top C$ (every column of A must be a linear combination of the $n-k$ columns of $Y_\perp$, which are linearly independent). Thus, we have

$$YS + Y_\perp A = YS + Y_\perp Y_\perp^\top C = YS + (I - YY^\top)C.$$

In order to describe the geodesics of $S(k,n)\cong G/H$, we will need the additional requirement of naturally reductiveness which is defined in the next section.

22.6 Naturally Reductive Homogeneous Spaces

When $M = G/H$ is a reductive homogeneous space that has a G-invariant metric, it is possible to give an expression for $(\nabla_{X^*}Y^*)_o$ (where X^* and Y^* are the vector fields corresponding to $X, Y\in\mathfrak{m}$, and $\nabla_{X^*}Y^*$ is the Levi-Civita connection).

If X^*, Y^*, Z^* are the Killing vector fields associated with $X, Y, Z\in\mathfrak{m}$, then by Proposition 17.10 we have

$$\begin{aligned}X^*\langle Y^*, Z^*\rangle &= \langle [X^*, Y^*], Z^*\rangle + \langle Y^*, [X^*, Z^*]\rangle\\ Y^*\langle X^*, Z^*\rangle &= \langle [Y^*, X^*], Z^*\rangle + \langle X^*, [Y^*, Z^*]\rangle\\ Z^*\langle X^*, Y^*\rangle &= \langle [Z^*, X^*], Y^*\rangle + \langle X^*, [Z^*, Y^*]\rangle.\end{aligned}$$

Using the Koszul formula (see Proposition 14.9),

$$\begin{aligned}2\langle \nabla_{X^*}Y^*, Z^*\rangle &= X^*(\langle Y^*, Z^*\rangle) + Y^*(\langle X^*, Z^*\rangle) - Z^*(\langle X^*, Y^*\rangle)\\ &\quad - \langle Y^*, [X^*, Z^*]\rangle - \langle X^*, [Y^*, Z^*]\rangle - \langle Z^*, [Y^*, X^*]\rangle,\end{aligned}$$

we obtain

$$2\langle \nabla_{X^*}Y^*, Z^*\rangle = \langle [X^*, Y^*], Z^*\rangle + \langle [X^*, Z^*], Y^*\rangle + \langle [Y^*, Z^*], X^*\rangle.$$

Since $[X^*, Y^*] = -[X, Y]^*$ (see Proposition 22.19), we obtain

$$2\langle \nabla_{X^*}Y^*, Z^*\rangle = -\langle [X, Y]^*, Z^*\rangle - \langle [X, Z]^*, Y^*\rangle - \langle [Y, Z]^*, X^*\rangle.$$

The problem is that the vector field $\nabla_{X^*}Y^*$ is not necessarily of the form W^* for some $W \in \mathfrak{g}$. However, we can find its value at o. By evaluating at o and using the fact that $X^*_o = (X^*_{\mathfrak{m}})_o$ for any $X \in \mathfrak{g}$, we obtain

$$2\langle (\nabla_{X^*}Y^*)_o, Z^*_o\rangle = -\langle ([X, Y]^*_{\mathfrak{m}})_o, Z^*_o\rangle - \langle ([X, Z]^*_{\mathfrak{m}})_o, Y^*_o\rangle - \langle ([Y, Z]^*_{\mathfrak{m}})_o, X^*_o\rangle.$$

Hence

$$2\langle (\nabla_{X^*}Y^*)_o, Z^*_o\rangle + \langle ([X, Y]^*_{\mathfrak{m}})_o, Z^*_o\rangle = \langle ([Z, X]^*_{\mathfrak{m}})_o, Y^*_o\rangle + \langle ([Z, Y]^*_{\mathfrak{m}})_o, X^*_o\rangle,$$

and consequently,

$$(\nabla_{X^*}Y^*)_o = -\frac{1}{2}([X, Y]^*_{\mathfrak{m}})_o + U(X, Y)^*_o,$$

where $[X, Y]_{\mathfrak{m}}$ is the component of $[X, Y]$ on $\mathfrak{m}$ and $U(X, Y)$ is determined by

$$2\langle U(X, Y), Z\rangle = \langle [Z, X]_{\mathfrak{m}}, Y\rangle + \langle X, [Z, Y]_{\mathfrak{m}}\rangle,$$

for all $Z \in \mathfrak{m}$. Here we are using the isomorphism $X \mapsto X^*_o$ between $\mathfrak{m}$ and $T_o(G/H)$ and the fact that the inner product on $\mathfrak{m}$ is chosen so that $\mathfrak{m}$ and $T_o(G/H)$ are isometric.

Since the term $U(X, Y)^*_o$ clearly complicates matters, it is natural to make the following definition, which is equivalent to requiring that $U(X, Y) = 0$ for all $X, Y \in \mathfrak{m}$.

Definition 22.9. A homogeneous space G/H is *naturally reductive* if it is reductive with some reductive decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, if it has a G-invariant metric, and if

$$\langle [X, Z]_{\mathfrak{m}}, Y\rangle = \langle X, [Z, Y]_{\mathfrak{m}}\rangle, \quad \text{for all } X, Y, Z \in \mathfrak{m}.$$

Note that one of the requirements of Definition 22.9 is that G/H must have a G-invariant metric.

The above computation yield the following result.

Proposition 22.25. *If G/H is naturally reductive, then the Levi-Civita connection associated with the G-invariant metric on G/H is given by*

$$(\nabla_{X^*}Y^*)_o = -\frac{1}{2}([X,Y]^*_{\mathfrak{m}})_o = -\frac{1}{2}[X,Y]_{\mathfrak{m}},$$

for all $X, Y \in \mathfrak{m}$.

We can now find the geodesics on a naturally reductive homogeneous space. Indeed, if $M = (G,H)$ is a reductive homogeneous space and if M has a G-invariant metric, then there is an $\mathrm{Ad}(H)$-invariant inner product $\langle -, - \rangle_{\mathfrak{m}}$ on $\mathfrak{m}$. Pick any inner product $\langle -, - \rangle_{\mathfrak{h}}$ on $\mathfrak{h}$, and define an inner product on $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ by setting $\mathfrak{h}$ and $\mathfrak{m}$ to be orthogonal. Then Proposition 22.23 provides a left-invariant metric on G for which the elements of $\mathfrak{h}$ are vertical vectors and the elements of $\mathfrak{m}$ are horizontal vectors.

Observe that in this situation, the condition for being naturally reductive extends to left-invariant vector fields on G induced by vectors in $\mathfrak{m}$. Since $(dL_g)_1 \colon \mathfrak{g} \to T_gG$ is a linear isomorphism for all $g \in G$, the direct sum decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ yields a direct sum decomposition $T_gG = (dL_g)_1(\mathfrak{h}) \oplus (dL_g)_1(\mathfrak{m})$. Given a left-invariant vector field X^L induced by a vector $X \in \mathfrak{g}$, if $X = X_{\mathfrak{h}} + X_{\mathfrak{m}}$ is the decomposition of X onto $\mathfrak{h} \oplus \mathfrak{m}$, we obtain a decomposition

$$X^L = X^L_{\mathfrak{h}} + X^L_{\mathfrak{m}},$$

into a left-invariant vector field $X^L_{\mathfrak{h}} \in \mathfrak{h}^L$ and a left-invariant vector field $X^L_{\mathfrak{m}} \in \mathfrak{m}^L$, with

$$X^L_{\mathfrak{h}}(g) = (dL_g)_1(X_{\mathfrak{h}}), \quad X^L_{\mathfrak{m}} = (dL_g)_1(X_{\mathfrak{m}}).$$

Since the $(dL_g)_1$ are isometries, if $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal, so are $(dL_g)_1(\mathfrak{h})$ and $(dL_g)_1(\mathfrak{m})$, and so $X^L_{\mathfrak{h}}$ and $X^L_{\mathfrak{m}}$ are orthogonal vector fields.

Since $[X^L, Y^L] = [X,Y]^L$ (see the calculation made just after Definition 18.8), we have $[X^L, Y^L]_{\mathfrak{m}}(g) = [X,Y]^L_{\mathfrak{m}}(g) = (dL_g)_1([X,Y]_{\mathfrak{m}})$, so if X^L, Y^L, Z^L are the left-invariant vector fields induced by $X, Y, Z \in \mathfrak{m}$, since the metric on G is left-invariant, for any $g \in G$, we have

$$\begin{aligned}\langle [X^L, Z^L]_{\mathfrak{m}}(g), Y^L(g) \rangle &= \langle (dL_g)_1([X,Z]_{\mathfrak{m}}), (dL_g)_1(Y) \rangle \\ &= \langle [X,Z]_{\mathfrak{m}}, Y \rangle.\end{aligned}$$

Similarly, we have

$$\langle X^L(g), [Z^L, Y^L]_{\mathfrak{m}}(g) \rangle = \langle X, [Z,Y]_{\mathfrak{m}} \rangle.$$

In summary, we showed the following result.

Proposition 22.26. *If the condition for being naturally reductive holds, namely*

$$\langle [X,Z]_{\mathfrak{m}}, Y \rangle = \langle X, [Z,Y]_{\mathfrak{m}} \rangle, \quad \textit{for all } X, Y, Z \in \mathfrak{m},$$

then a similar condition holds for left-invariant vector fields:

$$\langle [X^L, Z^L]_{\mathfrak{m}}, Y^L \rangle = \langle X^L, [Z^L, Y^L]_{\mathfrak{m}} \rangle, \quad \textit{for all } X^L, Y^L, Z^L \in \mathfrak{m}^L.$$

Recall that the left action of G on G/H is given by $g_1 \cdot g_2H = g_1g_2H$, and that o denotes the coset $1H$.

Proposition 22.27. *If $M = G/H$ is a naturally reductive homogeneous space, for every $X \in \mathfrak{m}$, the geodesic $\gamma_{d\pi_1(X)}$ through o is given by*

$$\gamma_{d\pi_1(X)}(t) = \pi \circ \exp(tX) = \exp(tX) \cdot o, \quad \textit{for all } t \in \mathbb{R}.$$

Proof. As explained earlier, since there is a G-invariant metric on G/H, we can construct a left-invariant metric $\langle -, - \rangle$ on G such that its restriction to $\mathfrak{m}$ is $\mathrm{Ad}(H)$-invariant, and such that $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal. The curve $\alpha(t) = \exp(tX)$ is horizontal in G, since it is an integral curve of the horizontal vector field $X^L \in \mathfrak{m}^L$. By Proposition 17.8, the Riemannian submersion π carries horizontal geodesics in G to geodesics in G/H. Thus it suffices to show that α is a geodesic in G. Following O'Neill (O'Neill [91], Chapter 11, Proposition 25), we prove that

$$\nabla_{X^L} Y^L = \frac{1}{2}[X^L, Y^L], \quad X, Y \in \mathfrak{m}.$$

As noted in Section 20.3, since the metric on G is left-invariant, the Koszul formula reduces to

$$2\langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle - \langle [Y^L, Z^L], X^L \rangle + \langle [Z^L, X^L], Y^L \rangle;$$

that is

$$2\langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle + \langle [Z^L, Y^L], X^L \rangle - \langle [X^L, Z^L], Y^L \rangle, \text{ for all } X, Y, Z \in \mathfrak{g}.$$

Since $\langle -, - \rangle$ and $\mathfrak{m}$ are $\mathrm{Ad}(H)$-invariant, as in the proof of Proposition 20.8, for all $a \in H$,

$$\langle \mathrm{Ad}_a(X), \mathrm{Ad}_a(Y) \rangle = \langle X, Y \rangle, \quad \text{for all } X, Y \in \mathfrak{m},$$

so the function $a \mapsto \langle \mathrm{Ad}_a(X), \mathrm{Ad}_a(Y) \rangle$ is constant, and by taking the derivative with $a = \exp(tZ)$ at $t = 0$, we get

$$\langle [X, Z], Y \rangle = \langle X, [Z, Y] \rangle, \quad X, Y \in \mathfrak{m}, \ Z \in \mathfrak{h}.$$

Since the metric on G is left-invariant, as in the proof of Proposition 20.8, by applying $(dL_g)_e$ to X, Y, Z, we obtain

$$\langle [X^L, Z^L], Y^L \rangle = \langle X^L, [Z^L, Y^L] \rangle, \quad X, Y \in \mathfrak{m}, \ Z \in \mathfrak{h}. \tag{h}$$

The natural reductivity condition is

$$\langle [X^L, Z^L]_{\mathfrak{m}}, Y^L \rangle = \langle X^L, [Z^L, Y^L]_{\mathfrak{m}} \rangle \quad \text{for all } X, Y, Z \in \mathfrak{m}. \tag{m}$$

Also recall that $\mathfrak{h}$ and $\mathfrak{m}$ are orthogonal. Let us now consider the Koszul formula for $X, Y \in \mathfrak{m}$ and $Z \in \mathfrak{g}$. If $Z \in \mathfrak{m}$, then by (m), the last two terms cancel out. Similarly, if $Z \in \mathfrak{h}$, then by (h), the last two terms cancel out. Therefore,

$$2\langle \nabla_{X^L} Y^L, Z^L \rangle = \langle [X^L, Y^L], Z^L \rangle \quad \text{for all } X \in \mathfrak{g},$$

which shows that

$$\nabla_{X^L} Y^L = \frac{1}{2}[X^L, Y^L], \quad X, Y \in \mathfrak{g}.$$

To finish the proof, the above formula implies that

$$\nabla_{X^L} X^L = 0,$$

but since α is a one-parameter group, $\alpha' = X^L$, which shows that α is indeed a geodesic.

If γ is any geodesic through o with initial condition $X^*_o = d\pi_1(X)$ $(X \in \mathfrak{m})$, then the curve $t \mapsto \exp(tX) \cdot o$ is also a geodesic through o with the same initial condition, so γ must coincide with this curve. □

Proposition 22.27 shows that the geodesics in G/H are given by the orbits of the one-parameter groups $(t \mapsto \exp tX)$ generated by the members of $\mathfrak{m}$.

We can also obtain a formula for the geodesic through every point $p = gH \in G/H$. Recall from Definition 22.6 that the vector field X^* associated with a vector $X \in \mathfrak{m}$ is given by

$$X^*(p) = \frac{d}{dt}(\exp(tX) \cdot p)\bigg|_{t=0}, \quad p \in G/H.$$

We have an isomorphism between $\mathfrak{m}$ and $T_o(G/H)$ given by $X \mapsto X^*_o$. Furthermore, $(\tau_g)_*$ induces an isomorphism between $T_o(G/H)$ and $T_p(G/H)$. By Proposition 22.19 (1), we have

$$(\mathrm{Ad}_g X)^* = (\tau_g)_* X^*,$$

so the isomorphism from $\mathfrak{m}$ to $T_p(G/H)$ is given by

$$X \mapsto (\mathrm{Ad}_g X)^*_p.$$

It follows that the geodesic through p with initial velocity $(\mathrm{Ad}_g X)^*_p$ is given by

$$t \mapsto \exp(t\mathrm{Ad}_g X) \cdot p.$$

Since Proposition 18.10 implies that $\exp(t\mathrm{Ad}_g X) = g \exp(tX) g^{-1}$ and $g^{-1} \cdot p = o$, the geodesic through $p = gH$ with initial velocity $(\mathrm{Ad}_g X)^*_p = (\tau_g)_* X^*_p$ is given by

$$t \mapsto g \exp(tX) \cdot o.$$

We record this fact as the following proposition.

Proposition 22.28. *If $M = G/H$ is a naturally reductive homogeneous space, for every $X \in \mathfrak{m}$, the geodesic through $p = gH$ with initial velocity $(\mathrm{Ad}_g X)^*_p = (\tau_g)_* X^*_p$ is given*

$$t \mapsto g \exp(tX) \cdot o.$$

An important corollary of Proposition 22.27 is that naturally reductive homogeneous spaces are complete. Indeed, the one-parameter group $t \mapsto \exp(tX)$ is defined for all $t \in \mathbb{R}$.

One can also figure out a formula for the sectional curvature (see (O'Neill [91], Chapter 11, Proposition 26). Under the identification of $\mathfrak{m}$ and $T_o(G/H)$ given by the restriction of $d\pi_1$ to $\mathfrak{m}$, we have

$$\langle R(X,Y)X,Y\rangle = \frac{1}{4}\langle [X,Y]_{\mathfrak{m}}, [X,Y]_{\mathfrak{m}}\rangle + \langle [[X,Y]_{\mathfrak{h}}, X]_{\mathfrak{m}}, Y\rangle, \quad \text{for all } X, Y \in \mathfrak{m}.$$

Conditions on a homogeneous space that ensure that such a space is naturally reductive are obviously of interest. Here is such a condition.

Proposition 22.29. *Let $M = G/H$ be a homogeneous space with G a connected Lie group, assume that $\mathfrak{g}$ admits an $\mathrm{Ad}(G)$-invariant inner product $\langle -,-\rangle$, and let $\mathfrak{m} = \mathfrak{h}^{\perp}$ be the orthogonal complement of $\mathfrak{h}$ with respect to $\langle -,-\rangle$. Then the following properties hold.*

(1) The space G/H is reductive with respect to the decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$.

(2) Under the G-invariant metric induced by $\langle -,-\rangle$, the homogeneous space G/H is naturally reductive.

(3) The sectional curvature is determined by

$$\langle R(X,Y)X,Y\rangle = \frac{1}{4}\langle [X,Y]_{\mathfrak{m}}, [X,Y]_{\mathfrak{m}}\rangle + \langle [X,Y]_{\mathfrak{h}}, [X,Y]_{\mathfrak{h}}\rangle.$$

Sketch of proof. Since H is closed under $\mathbf{Ad}_h$ for every $h \in H$, by taking the derivative at 1 we see that $\mathfrak{h}$ is closed under Ad_h for all $h \in H$. In fact, since Ad_h is an isomorphism, we have $\mathrm{Ad}_h(\mathfrak{h}) = \mathfrak{h}$. Since $\mathfrak{m} = \mathfrak{h}^{\perp}$, we can show that $\mathfrak{m}$ is also closed under Ad_h. If $u \in \mathfrak{m} = \mathfrak{h}^{\perp}$, then

$$\langle u, v\rangle = 0 \quad \text{for all } v \in \mathfrak{h}.$$

Since the inner product $\langle -,-\rangle$ is $\mathrm{Ad}(G)$-invariant, for any $h \in H$ we get

$$\langle \mathrm{Ad}_h(u), \mathrm{Ad}_h(v)\rangle = 0 \quad \text{for all } v \in \mathfrak{h}.$$

Since $\mathrm{Ad}_h(\mathfrak{h}) = \mathfrak{h}$, the above means that

$$\langle \mathrm{Ad}_h(u), w\rangle = 0 \quad \text{for all } w \in \mathfrak{h},$$

proving that $\mathrm{Ad}_h(u) \in \mathfrak{h}^{\perp} = \mathfrak{m}$. Therefore $\mathrm{Ad}_h(\mathfrak{m}) \subseteq \mathfrak{m}$ for all $a \in H$.

To prove (2), since $\langle -,-\rangle$ is $\mathrm{Ad}(G)$-invariant, for all $a \in G$, we have

$$\langle \mathrm{Ad}_a(X), \mathrm{Ad}_a(Y)\rangle = \langle X, Y\rangle, \quad \text{for all } X, Y \in \mathfrak{m},$$

so for $a = \exp(tZ)$ with $Z \in \mathfrak{m}$, by taking derivatives at $t = 0$, we get

$$\langle [X, Z], Y \rangle = \langle X, [Z, Y] \rangle, \quad X, Y, Z \in \mathfrak{m}.$$

However, since $\mathfrak{m}$ and $\mathfrak{h}$ are orthogonal, the above implies that

$$\langle [X, Z]_{\mathfrak{m}}, Y \rangle = \langle X, [Z, Y]_{\mathfrak{m}} \rangle, \quad X, Y, Z \in \mathfrak{m},$$

which is the natural reductivity condition.

Part (3) is proved in Kobayashi and Nomizu [69] (Chapter X, Theorem 3.5). □

By Proposition 20.3, the condition that $\mathfrak{g}$ admits a $\mathrm{Ad}(G)$-invariant inner product is equivalent to the fact that G has a bi-invariant metric. By Proposition 20.6, this is equivalent to requiring $\overline{\mathrm{Ad}(G)}$ to be compact. *In practice, this means that G is compact.*

Recall a Lie group G is said to be semisimple if its Lie algebra $\mathfrak{g}$ is semisimple. From Theorem 20.26, a Lie algebra $\mathfrak{g}$ is semisimple iff its Killing form B is nondegenerate, and from Theorem 20.27, a connected Lie group G is compact and semisimple iff its Killing form B is negative definite. By Proposition 20.25, the Killing form is $\mathrm{Ad}(G)$-invariant. Thus, for any connected compact semisimple Lie group G, for any constant $c > 0$, the bilinear form $-cB$ is an $\mathrm{Ad}(G)$-invariant inner product on $\mathfrak{g}$. Then as a corollary of Proposition 22.29, we obtain the following result, which is that we use in practice.

Proposition 22.30. *Let $M = G/H$ be a homogeneous space such that G is a connected compact semisimple group. Then under any inner product $\langle -, - \rangle$ on $\mathfrak{g}$ given by $-cB$, where B is the Killing form of $\mathfrak{g}$ and $c > 0$ is any positive real number, the space G/H is naturally reductive with respect to the decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, where $\mathfrak{m} = \mathfrak{h}^{\perp}$ be the orthogonal complement of $\mathfrak{h}$ with respect to $\langle -, - \rangle$. The sectional curvature is non-negative.*

A homogeneous space as in Proposition 22.30 is called a *normal homogeneous space*.

22.7 Examples of Naturally Reductive Homogeneous Spaces

Since $\mathbf{SO}(n)$ is connected, semisimple, and compact for $n \geq 3$, the Stiefel manifolds $S(k, n) \cong \mathbf{SO}(n)/\mathbf{SO}(n-k)$ described in Section 22.5 are reductive spaces which satisfy the assumptions of Proposition 22.30 (with an inner product induced by a scalar factor of $-1/2$ of the Killing form on $\mathbf{SO}(n)$). Therefore, Stiefel manifolds $S(k, n)$ are naturally reductive homogeneous spaces for $n \geq 3$ (under the reduction $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$ induced by the Killing form).

Another class of naturally reductive homogeneous spaces is the Grassmannian manifolds $G(k, n)$ which may obtained via a refinement of the Stiefel manifold $S(k, n)$. Given any $n \geq 1$, for any k, with $0 \leq k \leq n$, let $G(k, n)$ be the set of all linear k-dimensional subspaces of

$\mathbb{R}^n$, where the k-dimensional subspace U of $\mathbb{R}$ is spanned by k linearly independent vectors $u_1, \ldots, u_k$ in $\mathbb{R}^n$; write $U = \mathrm{span}(u_1, \ldots, u_k)$. In Section 4.4 we showed that the action $\cdot\colon \mathbf{SO}(n) \times G(k,n) \to G(k,n)$

$$R \cdot U = \mathrm{span}(Ru_1, \ldots, Ru_k).$$

is well-defined, transitive, and has the property that stabilizer of U is the set of matrices in $\mathbf{SO}(n)$ with the form

$$R = \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix},$$

where $S \in \mathbf{O}(k)$, $T \in \mathbf{O}(n-k)$ and $\det(S)\det(T) = 1$. We denote this group by $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$. Since $\mathbf{SO}(n)$ is a connected, compact semisimple Lie group whenever $n \geq 3$, Proposition 22.30 implies that

$$G(k,n) \cong \mathbf{SO}(n)/S(\mathbf{O}(k) \times \mathbf{O}(n-k))$$

is a naturally reductive homogeneous manifold whenever $n \geq 3$.

If $n = 2$, then $\mathbf{SO}(2)$ is an abelian group, and thus not semisimple. However, in this case, $G(1,2) = \mathbb{RP}(1) \cong \mathbf{SO}(2)/S(\mathbf{O}(1) \times \mathbf{O}(1)) \cong \mathbf{SO}(2)/\mathbf{O}(1)$, and $S(1,2) = S^1 \cong \mathbf{SO}(2)/\mathbf{SO}(1) \cong \mathbf{SO}(2)$. These are special cases of symmetric spaces discussed in Section 22.9. In the first case, $H = S(\mathbf{O}(1) \times \mathbf{O}(1))$, and in the second case, $H = \mathbf{SO}(1)$. In both cases,

$$\mathfrak{h} = (0),$$

and we can pick

$$\mathfrak{m} = \mathfrak{so}(2),$$

which is trivially $\mathrm{Ad}(H)$-invariant. In Section 22.9 we show that the inner product on $\mathfrak{so}(2)$ given by

$$\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$$

is $\mathrm{Ad}(H)$-invariant, and with the induced metric, $\mathbb{RP}(1)$ and $S^1 \cong \mathbf{SO}(2)$ are are examples of naturally reductive homogeneous spaces which are *also* symmetric spaces.

For $n \geq 3$, we have $S(1,n) = S^{n-1}$ and $S(n-1,n) = \mathbf{SO}(n)$, which are symmetric spaces. On the other hand, $S(k,n)$ it is not a symmetric space if $2 \leq k \leq n-2$. A justification is given in Section 22.10.

To construct yet another class of naturally reductive homogeneous spaces known as the *oriented Grassmannian* $G^0(k,n)$, we consider the set of k-dimensional *oriented subspaces* of $\mathbb{R}^n$. An oriented k-subspace is a k-dimensional subspace W together with the choice of a basis $(u_1, \ldots, u_k)$ determining the orientation of W. Another basis $(v_1, \ldots, v_k)$ of W is *positively oriented* if $\det(f) > 0$, where f is the unique linear map f such that $f(u_i) = v_i$, $i = 1, \ldots, k$.

Definition 22.10. The set of k-dimensional oriented subspaces of $\mathbb{R}^n$ is called the *oriented Grassmannian*, and it is denoted by $G^0(k,n)$.

The action of $\mathbf{SO}(n)$ on $G(k,n)$ is readily adjusted to become a transitive action $G^0(k,n)$. By a reasoning similar to the one used in the case where $\mathbf{SO}(n)$ acts on $G(k,n)$, we find that the stabilizer of the oriented subspace $(e_1,\ldots,e_k)$ is the set of orthogonal matrices of the form

$$\begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix},$$

where $Q \in \mathbf{SO}(k)$ and $R \in \mathbf{SO}(n-k)$, because this time, Q has to preserve the orientation of the subspace spanned by $(e_1,\ldots,e_k)$. Thus the isotropy group is isomorphic to

$$\mathbf{SO}(k) \times \mathbf{SO}(n-k).$$

It follows from Proposition 22.30 that

$$G^0(k,n) \cong \mathbf{SO}(n)/\mathbf{SO}(k) \times \mathbf{SO}(n-k)$$

is a naturally reductive homogeneous space whenever $n \geq 3$. Furthermore, since $G^0(1,2) \cong \mathbf{SO}(2)/\mathbf{SO}(1) \times \mathbf{SO}(1) \cong \mathbf{SO}(2)/\mathbf{SO}(1) \cong S(1,2)$, the same reasoning that shows why $S(1,2)$ is a symmetric space explains why $G^0(1,2) \cong S^1$ is also a symmetric space

Since the Grassmann manifolds $G(k,n)$ and the oriented Grassmann manifolds $G^0(k,n)$ have more structure (they are symmetric spaces), in this section we restrict our attention to the Stiefel manifolds $S(k,n)$. The Grassmannian manifolds $G(k,n)$ and $G^0(k,n)$ are discussed in Section 22.9.

Stiefel manifolds have been presented as reductive homogeneous spaces in Section 22.5, but since they are also naturally reductive, we can describe their geodesics.

By Proposition 22.27, the geodesic through o with initial velocity

$$X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \in \mathfrak{m}$$

is given by

$$\gamma(t) = \exp\left(t\begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix}\right) P_{n,k}.$$

Recall that $S \in \mathfrak{so}(k)$ and $A \in \mathrm{M}_{n-k,k}(\mathbb{R})$. This is not a very explicit formula. It is possible to do better, see later in this section for details.

Let us consider the case where $k = n-1$, which is simpler.

If $k = n-1$, then $n-k = 1$, so $S(n-1,n) = \mathbf{SO}(n)$, $H \cong \mathbf{SO}(1) = \{1\}$, $\mathfrak{h} = (0)$ and $\mathfrak{m} = \mathfrak{so}(n)$. The inner product on $\mathfrak{so}(n)$ is given by

$$\langle X, Y\rangle = -\frac{1}{2}\mathrm{tr}(XY) = \frac{1}{2}\mathrm{tr}(X^\top Y), \qquad X, Y \in \mathfrak{so}(n).$$

Every matrix $X \in \mathfrak{so}(n)$ is a skew-symmetric matrix, and we know that every such matrix can be written as $X = P^\top DP$, where P is orthogonal and where D is a block diagonal matrix whose blocks are either a 1-dimensional block consisting of a zero, of a 2×2 matrix of the form

$$D_j = \begin{pmatrix} 0 & -\theta_j \\ \theta_j & 0 \end{pmatrix},$$

with $\theta_j > 0$. Then, $e^X = P^\top e^D P = P^\top \Sigma P$, where Σ is a block diagonal matrix whose blocks are either a 1-dimensional block consisting of a 1, of a 2×2 matrix of the form

$$D_j = \begin{pmatrix} \cos\theta_j & -\sin\theta_j \\ \sin\theta_j & \cos\theta_j \end{pmatrix}.$$

We also know that every matrix $R \in \mathbf{SO}(n)$ can be written as

$$R = e^X,$$

for some matrix $X \in \mathfrak{so}(n)$ as above, with $0 < \theta_j \leq \pi$. Then we can give a formula for the distance $d(I, Q)$ between the identity matrix and any matrix $Q \in \mathrm{SO}(n)$. Since the geodesics from I through Q are of the form

$$\gamma(t) = e^{tX} \quad \text{with} \quad e^X = Q,$$

and since the length $L(\gamma)$ of the geodesic from I to e^X is

$$L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t) \rangle^{\frac{1}{2}} dt,$$

we have

$$\begin{aligned}
d(I, Q) &= \min_{X | e^X = Q} \int_0^1 \langle (e^{tX})', (e^{tX})' \rangle^{\frac{1}{2}} dt \\
&= \min_{X | e^X = Q} \int_0^1 \langle X e^{tX}, X e^{tX} \rangle^{\frac{1}{2}} dt \\
&= \min_{X | e^X = Q} \int_0^1 \left(\frac{1}{2} \mathrm{tr}((e^{tX})^\top X^\top X e^{tX}) \right)^{\frac{1}{2}} dt \\
&= \min_{X | e^X = Q} \int_0^1 \left(\frac{1}{2} \mathrm{tr}(X^\top X e^{tX} e^{tX^\top}) \right)^{\frac{1}{2}} dt \\
&= \min_{X | e^X = Q} \int_0^1 \left(\frac{1}{2} \mathrm{tr}(X^\top X e^{tX} e^{-tX}) \right)^{\frac{1}{2}} dt \\
&= \min_{X | e^X = Q} \left(\frac{1}{2} \mathrm{tr}(X^\top X) \right)^{\frac{1}{2}} \\
&= (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},
\end{aligned}$$

where $\theta_1, \ldots, \theta_m$ are the angles associated with the eigenvalues $e^{\pm i\theta_1}, \ldots, e^{\pm i\theta_m}$ of Q distinct from 1, and with $0 < \theta_j \leq \pi$. Therefore,

$$d(I, Q) = (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},$$

and if $Q, R \in \mathbf{SO}(n)$, then

$$d(Q, R) = (\theta_1^2 + \cdots + \theta_m^2)^{\frac{1}{2}},$$

where $\theta_1, \ldots, \theta_m$ are the angles associated with the eigenvalues $e^{\pm i\theta_1}, \ldots, e^{\pm i\theta_m}$ of $Q^{-1}R = Q^\top R$ distinct from 1, and with $0 < \theta_j \leq \pi$.

Remark: Since $X^\top = -X$, the square distance $d(I, Q)^2$ can also be expressed as

$$d(I, Q)^2 = -\frac{1}{2} \min_{X \mid e^X = Q} \operatorname{tr}(X^2),$$

or even (with some abuse of notation, since log is multi-valued) as

$$d(I, Q)^2 = -\frac{1}{2} \min \operatorname{tr}((\log Q)^2).$$

In the other special case where $k = 1$, we have $S(1, n) = S^{n-1}$, $H \cong \mathbf{SO}(n-1)$,

$$\mathfrak{h} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(n-1) \right\},$$

and

$$\mathfrak{m} = \left\{ \begin{pmatrix} 0 & -u^\top \\ u & 0 \end{pmatrix} \;\middle|\; u \in \mathbb{R}^{n-1} \right\}.$$

Therefore, there is a one-to-one correspondence between $\mathfrak{m}$ and $\mathbb{R}^{n-1}$. Given any $Q \in \mathbf{SO}(n)$, the equivalence class $[Q]$ of Q is uniquely determined *by the first column* of Q, and we view it as a *point on* S^{n-1}.

If we let $\|u\| = \sqrt{u^\top u}$, we leave it as an exercise to prove that for any

$$X = \begin{pmatrix} 0 & -u^\top \\ u & 0 \end{pmatrix},$$

we have

$$e^{tX} = \begin{pmatrix} \cos(\|u\|\, t) & -\sin(\|u\|\, t) \frac{u^\top}{\|u\|} \\ \sin(\|u\|\, t) \frac{u}{\|u\|} & I + (\cos(\|u\|\, t) - 1) \frac{uu^\top}{\|u\|^2} \end{pmatrix}.$$

Consequently (under the identification of S^{n-1} with the first column of matrices $Q \in \mathbf{SO}(n)$), the geodesic γ through e_1 (the column vector corresponding to the point $o \in S^{n-1}$) with initial tangent vector u is given by

$$\gamma(t) = \begin{pmatrix} \cos(\|u\|\, t) \\ \sin(\|u\|\, t) \frac{u}{\|u\|} \end{pmatrix} = \cos(\|u\|\, t) e_1 + \sin(\|u\|\, t) \frac{u}{\|u\|},$$

where $u \in \mathbb{R}^{n-1}$ is viewed as the vector in $\mathbb{R}^n$ whose first component is 0. Then we have

$$\gamma'(t) = \|u\| \left(-\sin(\|u\|\, t)e_1 + \cos(\|u\|\, t)\frac{u}{\|u\|} \right),$$

and we find the that the length $L(\gamma)(\theta)$ of the geodesic from e_1 to the point

$$p(\theta) = \gamma(\theta) = \cos(\|u\|\, \theta)e_1 + \sin(\|u\|\, \theta)\frac{u}{\|u\|}$$

is given by

$$L(\gamma)(\theta) = \int_0^\theta \langle \gamma'(t), \gamma'(t)\rangle^{\frac{1}{2}}\, dt = \theta\, \|u\|\, .$$

Since

$$\langle e_1, p(\theta)\rangle = \cos(\theta\, \|u\|),$$

we see that for a unit vector u and for any angle θ such that $0 \leq \theta \leq \pi$, the length of the geodesic from e_1 to $p(\theta)$ can be expressed as

$$L(\gamma)(\theta) = \theta = \arccos(\langle e_1, p\rangle);$$

that is, the angle between the unit vectors e_1 and p. This is a generalization of the distance between two points on a circle.

Geodesics can also be determined in the general case where $2 \leq k \leq n-2$; we follow Edelman, Arias and Smith [44], with one change because some point in that paper requires some justification which is not provided.

Given a point $Q = [Y\, Y_\perp] \in S(k, n)$, and given and any tangent vector $\widetilde{X} = YS + Y_\perp A$, Proposition 22.28 implies that we need to compute

$$\gamma(t) = [Y\, Y_\perp] \exp\left(t \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \right) P_{n,k}.$$

We can compute this exponential if we replace the matrix by a more "regular matrix," and for this, we use a QR-decomposition of A. Let

$$A = U \begin{pmatrix} R \\ 0 \end{pmatrix}$$

be a QR-decomposition of A, with U an orthogonal $(n-k) \times (n-k)$ matrix and R an upper triangular $k \times k$ matrix. We can write $U = [U_1\, U_2]$, where U_1 consists of the first k columns on U and U_2 of the last $n - 2k$ columns of U (if $2k \leq n$). We have

$$A = U_1 R,$$

and we can write

$$\begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & U_1 \end{pmatrix} \begin{pmatrix} S & -R^\top \\ R & 0 \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & U_1^\top \end{pmatrix}.$$

Then we have

$$\begin{aligned}\gamma(t) &= [Y\,Y_\perp]\begin{pmatrix} I & 0\\ 0 & U_1\end{pmatrix}\exp\left(t\begin{pmatrix} S & -R^\top\\ R & 0\end{pmatrix}\right)\begin{pmatrix} I & 0\\ 0 & U_1^\top\end{pmatrix}P_{n,k}\\ &= [Y\,Y_\perp U_1]\exp\left(t\begin{pmatrix} S & -R^\top\\ R & 0\end{pmatrix}\right)\begin{pmatrix} I & 0\\ 0 & U^\top\end{pmatrix}P_{n,k}\\ &= [Y\,Y_\perp U_1]\exp\left(t\begin{pmatrix} S & -R^\top\\ R & 0\end{pmatrix}\right)\begin{pmatrix} I_k\\ 0\end{pmatrix}.\end{aligned}$$

This is essentially the formula given by Corollary 2.2, Section 2.4.2 of Edelman, Arias and Smith [44], except for the term $Y_\perp U_1$. To explain the difference, observe that Edelman, Arias and Smith [44] derived their formula by taking a QR decomposition of $(I - YY^\top)\widetilde{X}$ and implicitly assume that a QR decomposition of $(I-YY^\top)\widetilde{X}$ yields a QR decomposition of A. But unfortunately, this assumption does not appear to be true. What is true is that a QR decomposition of A yields a QR decomposition of $(I - YY^\top)\widetilde{X}$. To justify this statement, observe that since $A = U_1R$, we have $Y_\perp A = Y_\perp U_1 R$, but $A = Y_\perp^\top \widetilde{X}$ so $Y_\perp A = (I-YY^\top)\widetilde{X}$, and thus

$$(I - YY^\top)\widetilde{X} = Y_\perp U_1 R.$$

If we write $Q = Y_\perp U_1$, then we have

$$Q^\top Q = U_1^\top Y_\perp^\top Y_\perp U_1 = I,$$

since $Y_\perp^\top Y_\perp = I$ and $U_1^\top U_1 = I$. Therefore,

$$(I - YY^\top)\widetilde{X} = QR$$

is a compact QR-decomposition of $(I - YY^\top)\widetilde{X}$.

Furthermore, given a QR-decomposition of $(I - YY^\top)\widetilde{X}$,

$$(I - YY^\top)\widetilde{X} = QR,$$

since $(I - YY^\top)\widetilde{X} = Y_\perp A$, we get

$$A = Y_\perp^\top QR.$$

But,

$$(Y_\perp^\top Q)^\top Y_\perp^\top Q = Q^\top Y_\perp Y_\perp^\top Q,$$

and there is no reason why this term should be equal to I.

Thus, it seems to us that one has to use a QR-decomposition of A. In any case, there are efficient algorithms to compute the exponential of the $2k \times 2k$ matrix

$$t\begin{pmatrix} S & -R^\top\\ R & 0\end{pmatrix}.$$

Since by Proposition 17.8(1), the length of the geodesic γ from o to $p = e^{sX} \cdot o$ is the same as the the length of the geodesic $\overline{\gamma}$ in G from 1 to e^{sX}, for any $X \in \mathfrak{m}$, we can easily compute the length $L(\gamma)(s)$ of the geodesic γ from o to $p = e^{sX} \cdot o$.

Indeed, for any

$$X = \begin{pmatrix} S & -A^\top \\ A & 0 \end{pmatrix} \in \mathfrak{m},$$

we know that the geodesic (in G) from 1 with initial velocity X is $\overline{\gamma}(t) = e^{tX}$, so we have

$$L(\gamma)(s) = L(\overline{\gamma})(s) = \int_0^s \langle (e^{tX})', (e^{tX})' \rangle^{\frac{1}{2}} dt,$$

but we already did this computation and found that

$$(L(\gamma)(s))^2 = s^2 \left(\frac{1}{2} \mathrm{tr}(X^\top X) \right) = s^2 \left(\frac{1}{2} \mathrm{tr}(S^\top S) + \mathrm{tr}(A^\top A) \right).$$

We can compute these traces using the eigenvalues of S and the singular values of A. If $\pm i\theta_1, \ldots, \pm i\theta_m$ are the nonzero eigenvalues of S and $\sigma_1, \ldots, \sigma_k$ are the singular values of A, then

$$L(\gamma)(s) = s(\theta_1^2 + \cdots + \theta_m^2 + \sigma_1^2 + \cdots + \sigma_k^2)^{\frac{1}{2}}.$$

We conclude this section with a proposition that shows that under certain conditions, G is determined by $\mathfrak{m}$ and H. A point $p \in M = G/H$ is called a *pole* if the exponential map at p is a diffeomorphism. The following proposition is proved in O'Neill [91] (Chapter 11, Lemma 27).

Proposition 22.31. *If $M = G/H$ is a naturally reductive homogeneous space, then for any pole $o \in M$, there is a diffeomorphism $\mathfrak{m} \times H \cong G$ given by the map $(X, h) \mapsto (\exp(X))h$.*

Next we will see that there exists a large supply of naturally reductive homogeneous spaces: symmetric spaces.

22.8 A Glimpse at Symmetric Spaces

There is an extensive theory of symmetric spaces and our goal is simply to show that the additional structure afforded by an involutive automorphism of G yields spaces that are naturally reductive. The theory of symmetric spaces was entirely created by one person, Élie Cartan, who accomplished the tour de force of giving a complete classification of these spaces using the classification of semisimple Lie algebras that he had obtained earlier. One of the most complete exposition is given in Helgason [58]. O'Neill [91], Petersen [93], Sakai [100] and Jost [64] have nice and more concise presentations. Ziller [119] is also an excellent introduction, and Borel [17] contains a fascinating historical account.

Until now, we have denoted a homogeneous space by G/H, but when dealing with symmetric spaces, it is customary to denote the closed subgroup of G by K rather than H.

Given a homogeneous space G/K, the new ingredient is that we have an involutive automorphism σ of G.

Definition 22.11. Given a Lie group G, an automorphism σ of G such that $\sigma \neq \mathrm{id}$ and $\sigma^2 = \mathrm{id}$ called an *involutive automorphism* of G. Let G^σ be the set of fixed points of σ, the subgroup of G given by

$$G^\sigma = \{g \in G \mid \sigma(g) = g\},$$

and let G_0^σ be the identity component of G^σ (the connected component of G^σ containing 1).

If we have an involutive automorphism $\sigma\colon G \to G$, then we can consider the $+1$ and -1 eigenspaces of $d\sigma_1\colon \mathfrak{g} \to \mathfrak{g}$, given by

$$\begin{aligned}\mathfrak{k} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}\\ \mathfrak{m} &= \{X \in \mathfrak{g} \mid d\sigma_1(X) = -X\}.\end{aligned}$$

Definition 22.12. An involutive automorphism of G satisfying $G_0^\sigma \subseteq K \subseteq G^\sigma$ is called a *Cartan involution*. The map $d\sigma_1$ is often denoted by θ.

The following proposition will be needed later.

Proposition 22.32. *Let σ be an involutive automorphism of G and let $\mathfrak{k}$ and $\mathfrak{m}$ be the $+1$ and -1 eigenspaces of $d\sigma_1\colon \mathfrak{g} \to \mathfrak{g}$. Then for all $X \in \mathfrak{m}$ and all $Y \in \mathfrak{k}$, we have*

$$B(X, Y) = 0,$$

where B is the Killing form of $\mathfrak{g}$.

Proof. By Proposition 20.25, B is invariant under automorphisms of $\mathfrak{g}$. Since $\theta = d\sigma_1\colon \mathfrak{g} \to \mathfrak{g}$ is an automorphism and since $\mathfrak{m}$ and $\mathfrak{k}$ are eigenspaces of θ for the eigenvalues -1 and $+1$ respectively, we have

$$B(X, Y) = B(\theta(X), \theta(Y)) = B(-X, Y) = -B(X, Y),$$

so $B(X, Y) = 0$. □

As before, K can be viewed as the stabilizer of the left action of G on G/K, but remarkably, the fact that $\mathfrak{k}$ and $\mathfrak{m}$ are the eigenspaces of $d\sigma_1$ implies that they yield a reductive decomposition of G/K.

Proposition 22.33. *Given a homogeneous space G/K with a Cartan involution σ ($G_0^\sigma \subseteq K \subseteq G^\sigma$), if $\mathfrak{k}$ and $\mathfrak{m}$ are defined as above, then*

(1) $\mathfrak{k}$ is indeed the Lie algebra of K.

(2) We have a direct sum

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}.$$

(3) $\mathrm{Ad}(K)(\mathfrak{m}) \subseteq \mathfrak{m}$*; in particular,* $[\mathfrak{k}, \mathfrak{m}] \subseteq \mathfrak{m}$.

(4) We have

$$[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k} \quad \textit{and} \quad [\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k}.$$

In particular, the pair (G, K) *is a reductive homogeneous space (as in Definition 22.8), with reductive decomposition* $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$.

Proof. We follow the proof given in O'Neill [91] (Chapter 11, Lemma 30). Another proof is given in Ziller [119] (Chapter 6).

(1) Let us rename $\{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}$ as $\mathcal{K}$, and let us denote the Lie algebra of K as $\mathfrak{k}$. Since $K \subseteq G^{\sigma}$, the restriction of the map σ to K is the identity, so if $X \in \mathfrak{k}$, then $d\sigma_1(X) = X$. This shows that $\mathfrak{k} \subseteq \mathcal{K}$. Conversely, assume that $X \in \mathcal{K}$, that is, $d\sigma_1(X) = X$. If $h_X(t) = \exp(tX)$ is the one-parameter subgroup of X, then h_X and $\sigma \circ h_X$ have the same initial velocity since $(\sigma \circ h_X)'(0) = d\sigma_1(h_X'(0)) = d\sigma_1(X) = X$. But $\sigma \circ h_X$ is also an integral curve through 1, so by uniqueness of integral curves, $\sigma \circ h_X = h_X$. Therefore h_X lies in G^{σ}, in fact in G_0^{σ} (as the image of the connected set $\mathbb{R}$ under a continuous map), and since $G_0^{\sigma} \subseteq K$, we deduce that $X \in \mathfrak{k}$. This proves that $\mathcal{K} \subseteq \mathfrak{k}$, and thus $\mathfrak{k} = \mathcal{K} = \{X \in \mathfrak{g} \mid d\sigma_1(X) = X\}$.

(2) This is purely a matter of linear algebra. Since $\sigma^2 = \mathrm{id}$, by taking the derivative at 1 we get $d\sigma_1^2 = \mathrm{id}$. Let E be any vector space and let $f\colon E \to E$ be a linear map such that $f^2 = \mathrm{id}$. Let E_1 and E_{-1} be the eigenspaces of E associated with $+1$ and -1,

$$\begin{aligned} E_1 &= \{u \in E \mid f(u) = u\} \\ E_{-1} &= \{u \in E \mid f(u) = -u\}. \end{aligned}$$

Then we have a direct sum

$$E = E_1 \oplus E_{-1}.$$

Pick any $u \in E$ and write $u = u_1 + u_{-1}$, with

$$u_1 = \frac{u + f(u)}{2}, \quad u_{-1} = \frac{u - f(u)}{2}.$$

Since $f^2 = \mathrm{id}$, we have

$$f(u_1) = f\left(\frac{u + f(u)}{2}\right) = \frac{f(u) + f^2(u)}{2} = \frac{f(u) + u}{2} = u_1,$$

and

$$f(u_{-1}) = f\left(\frac{u - f(u)}{2}\right) = \frac{f(u) - f^2(u)}{2} = \frac{f(u) - u}{2} = -u_{-1}.$$

Therefore, $u_1 \in E_1$ and $u_{-1} \in E_{-1}$, and since $u = u_1 + u_{-1}$, we have

$$E = E_1 + E_{-1}.$$

If $u \in E_1 \cap E_{-1}$, then $f(u) = u$ and $f(u) = -u$, so $u = -u$, which means that $u = 0$. Therefore, $E_1 \cap E_{-1} = (0)$ and we have the direct sum

$$E = E_1 \oplus E_{-1}.$$

Applying the above to $f = d\sigma_1$, we get

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}.$$

(3) If $X \in \mathfrak{m} = \{X \in \mathfrak{g} \mid d\sigma_1(X) = -X\}$ we must show that $d\sigma_1(\mathrm{Ad}_k(X)) = -\mathrm{Ad}_k(X)$ for all $k \in K$. As usual, let $\mathbf{Ad}_k(g) = kgk^{-1}$. Since σ is the identity on K, observe that σ and $\mathbf{Ad}_k$ commute, since

$$\sigma(\mathbf{Ad}_k(g)) = \sigma(kgk^{-1}) = \sigma(k)\sigma(g)\sigma(k^{-1}) = k\sigma(g)k^{-1} = \mathbf{Ad}_k(\sigma(g)).$$

By taking derivatives, for any $X \in \mathfrak{m}$ and any $k \in K$, we get

$$\begin{aligned} d\sigma_1(\mathrm{Ad}_k(X)) &= d(\sigma \circ \mathbf{Ad}_k)_1(X) \\ &= d(\mathbf{Ad}_k \circ \sigma)_1(X) \\ &= \mathrm{Ad}_k(d\sigma_1(X)) \\ &= \mathrm{Ad}_k(-X) && \text{since } X \in \mathfrak{m} \\ &= -\mathrm{Ad}_k(X). \end{aligned}$$

The Ad_K-invariance of $\mathfrak{m}$ implies that $[\mathfrak{k}, \mathfrak{m}] \subseteq \mathfrak{m}$. This is also shown directly using the fact $\mathfrak{k}$ is the $+1$ eigenspace and $\mathfrak{m}$ is the -1-eigenspace of $d\sigma_1$. For all $X \in \mathfrak{k}$ and all $Y \in \mathfrak{m}$, we have

$$\begin{aligned} d\sigma_1([X, Y]) &= [d\sigma_1(X), d\sigma_1(Y)] \\ &= [X, -Y] && \text{since } X \in \mathfrak{k} \text{ and } Y \in \mathfrak{m} \\ &= -[X, Y], \end{aligned}$$

which means that $[X, Y] \in \mathfrak{m}$.

(4) Since $\mathfrak{k}$ is the Lie algebra of the Lie group of K, we have $[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}$. Since $\mathfrak{k}$ is the $+1$ eigenspace and $\mathfrak{m}$ is the -1-eigenspace of $d\sigma_1$, if $X, Y \in \mathfrak{m}$, we have

$$\begin{aligned} d\sigma_1([X, Y]) &= [d\sigma_1(X), d\sigma_1(Y)] \\ &= [-X, -Y] && \text{since } X, Y \in \mathfrak{m} \\ &= [X, Y], \end{aligned}$$

which means that $[X, Y] \in \mathfrak{k}$. We also obtain a direct proof of the inclusion $[\mathfrak{k}, \mathfrak{k}] \subseteq \mathfrak{k}$. For all $X, Y \in \mathfrak{k}$, we have

$$\begin{aligned} d\sigma_1([X, Y]) &= [d\sigma_1(X), d\sigma_1(Y)] \\ &= [X, Y] && \text{since } X, Y \in \mathfrak{k} \end{aligned}$$

which means that $[X, Y] \in \mathfrak{k}$. □

Observe that since σ is a Cartan involution, by Proposition 22.33, we have

$$[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k},$$

so $[X, Z], [Z, Y] \in \mathfrak{k}$ for all $X, Y, Z \in \mathfrak{m}$, and since $\mathfrak{k} \cap \mathfrak{m} = (0)$, we have $[X, Z]_{\mathfrak{m}} = [Z, Y]_{\mathfrak{m}} = 0$, which implies the natural reductivity condition

$$\langle [X, Z]_{\mathfrak{m}}, Y \rangle = \langle X, [Z, Y]_{\mathfrak{m}} \rangle, \quad \text{for all } X, Y, Z \in \mathfrak{m}.$$

Note that Proposition 22.33 holds without any assumption on K besides the fact that it is a closed subgroup of G. If we also assume that G is connected and that G_0^σ is compact, we then obtain the following remarkable result.

Theorem 22.34. *Let G be a connected Lie group and let $\sigma\colon G \to G$ be an automorphism such that $\sigma^2 = \mathrm{id}$, $\sigma \neq \mathrm{id}$ (an involutive automorphism), and G_0^σ is compact. For every compact subgroup K of G, if $G_0^\sigma \subseteq K \subseteq G^\sigma$, then G/K has G-invariant metrics, and for every such metric G/K is a naturally reductive space with reductive decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$ given by the $+1$ and -1 eigenspaces of $d\sigma_1$. For every $p \in G/K$, there is an isometry $s_p\colon G/K \to G/K$ such that $s_p(p) = p$, $d(s_p)_p = -\mathrm{id}$, and*

$$s_p \circ \pi = \pi \circ \sigma,$$

as illustrated in the diagram below:

$$\begin{tikzcd} G \arrow[r, "\sigma"] \arrow[d, "\pi"'] & G \arrow[d, "\pi"] \\ G/K \arrow[r, "s_p"'] & G/K. \end{tikzcd}$$

Proof. Since K is assumed to be compact, and since by Proposition 22.33, we know that G is a reductive homogeneous space, by Proposition 22.22 (3), there is a G-invariant metric on G/K. We observed just before stating the theorem that the natural reductivity condition holds. Therefore, under any G-invariant metric, G/K is indeed a naturally reductive space with reductive decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$ as described in terms of the eigenspaces of $d\sigma_1$. The existence of the isometry $s_p\colon G/K \to G/K$ is proved in O'Neill [91] (Chapter 11) and Ziller [119] (Chapter 6). □

Definition 22.13. A triple (G, K, σ) satisfying the assumptions of Theorem 22.34 is called a *symmetric pair*.[3]

[3] Once again we fall victims of tradition. A symmetric pair is actually a triple!

A triple (G, K, σ) as above defines a special kind of naturally homogeneous space G/K known as a symmetric space.

Definition 22.14. If M is a connected Riemannian manifold, for any $p \in M$, an isometry s_p such that $s_p(p) = p$ and $d(s_p)_p = -\mathrm{id}$ is a called a *global symmetry at* p. A connected Riemannian manifold M for which there is a global symmetry for every point of M is called a *symmetric space*.

Theorem 22.34 implies that the naturally reductive homogeneous space G/K defined by a symmetric pair (G, K, σ) is a symmetric space.

It can be shown that a global symmetry s_p reverses geodesics at p and that $s_p^2 = \mathrm{id}$, so s_p is an involution. It should be noted that although $s_p \in \mathrm{Isom}(M)$, the isometry s_p does not necessarily lie in $\mathrm{Isom}(M)_0$. (For the definition of $\mathrm{Isom}(M)$, see the beginning of Section 18.7.)

The following facts are proved in O'Neill [91] (Chapters 9 and 11), Ziller [119] (Chapter 6), and Sakai [100] (Chapter IV).

1. Every symmetric space M is complete, and $\mathrm{Isom}(M)$ acts transitively on M. In fact the identity component $\mathrm{Isom}(M)_0$ acts transitively on M.

2. Thus, every symmetric space M is a homogeneous space of the form $\mathrm{Isom}(M)_0/K$, where K is the isotropy group of any chosen point $p \in M$ (it turns out that K is compact).

3. The symmetry s_p gives rise to a Cartan involution σ of $G = \mathrm{Isom}(M)_0$ defined so that
$$\sigma(g) = s_p \circ g \circ s_p \quad g \in G.$$
Then we have
$$G_0^\sigma \subseteq K \subseteq G^\sigma.$$

4. Thus, every symmetric space M is presented by a symmetric pair $(\mathrm{Isom}(M)_0, K, \sigma)$.

However, beware that in the presentation of the symmetric space $M = G/K$ given by a symmetric pair (G, K, σ), *the group G is not necessarily equal to* $\mathrm{Isom}(M)_0$. Thus, we do not have a one-to-one correspondence between symmetric spaces and symmetric pairs; there are more presentations of symmetric pairs than symmetric spaces. From our point of view, this does not matter since we are more interested in getting symmetric spaces from the data (G, K, σ). By abuse of terminology (and notation), we refer to the homogeneous space G/K defined by a symmetric pair (G, K, σ) as the *symmetric space* (G, K, σ).

Since the homogeneous space G/K defined by a symmetric pair (G, K, σ) is naturally reductive and has a G-invariant metric, by Proposition 22.27, its geodesics coincide with the one-parameter groups (they are given by the Lie group exponential).

The Levi-Civita connection on a symmetric space depends only on the Lie bracket on $\mathfrak{g}$. Indeed, we have the following formula proved in Ziller [119] (Chapter 6).

Proposition 22.35. *Given any symmetric space M defined by the triple (G, K, σ), for any $X \in \mathfrak{m}$ and and vector field Y on $M \cong G/K$, we have*

$$(\nabla_{X^*} Y)_o = [X^*, Y]_o.$$

Proof. If X^*, Z^* are the Killing vector fields induced by any $X, Z \in \mathfrak{m}$, by the Koszul formula,

$$\begin{aligned} 2\langle \nabla_{X^*} Y, Z^* \rangle &= X^*(\langle Y, Z^* \rangle) + Y(\langle X^*, Z^* \rangle) - Z^*(\langle X^*, Y \rangle) \\ &\quad - \langle Y, [X^*, Z^*] \rangle - \langle X^*, [Y, Z^*] \rangle - \langle Z^*, [Y, X^*] \rangle. \end{aligned}$$

Since X^* and Z^* are Killing vector fields, by Proposition 17.10 we have

$$\begin{aligned} X^*\langle Y, Z^* \rangle &= \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle \\ Z^*\langle X^*, Y \rangle &= \langle [Z^*, X^*], Y \rangle + \langle X^*, [Z^*, Y] \rangle, \end{aligned}$$

and because the Levi-Civita connection is symmetric and torsion-free,

$$\begin{aligned} Y\langle X^*, Z^* \rangle &= \langle \nabla_Y X^*, Z^* \rangle + \langle X^*, \nabla_Y Z^* \rangle \\ \langle Z^*, \nabla_Y X^* \rangle &= \langle Z^*, \nabla_{X^*} Y \rangle - \langle Z^*, [X^*, Y] \rangle, \end{aligned}$$

so we get

$$Y\langle X^*, Z^* \rangle = \langle \nabla_Y Z^*, X^* \rangle + \langle \nabla_{X^*} Y, Z^* \rangle - \langle [X^*, Y], Z^* \rangle.$$

Plugging these expressions in the Koszul formula, we get

$$\begin{aligned} 2\langle \nabla_{X^*} Y, Z^* \rangle = {} & \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle + \langle \nabla_Y Z^*, X^* \rangle \\ & + \langle \nabla_{X^*} Y, Z^* \rangle - \langle [X^*, Y], Z^* \rangle - \langle [Z^*, X^*], Y \rangle - \langle X^*, [Z^*, Y] \rangle \\ & - \langle Y, [X^*, Z^*] \rangle - \langle X^*, [Y, Z^*] \rangle - \langle Z^*, [Y, X^*] \rangle \\ = {} & \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle + \langle \nabla_{X^*} Y, Z^* \rangle + \langle \nabla_Y Z^*, X^* \rangle, \end{aligned}$$

and thus,

$$\begin{aligned} \langle \nabla_{X^*} Y, Z^* \rangle &= \langle [X^*, Y], Z^* \rangle + \langle Y, [X^*, Z^*] \rangle + \langle \nabla_Y Z^*, X^* \rangle \\ &= \langle [X^*, Y], Z^* \rangle - \langle Y, [X, Z]^* \rangle + \langle \nabla_Y Z^*, X^* \rangle, \end{aligned}$$

where the second equality follows from Proposition 22.19 (2). Therefore, evaluating at o and using the fact that $[X, Z]^*_o = ([X, Z]^*_{\mathfrak{m}})_o$, we have

$$\langle (\nabla_{X^*} Y)_o, Z^*_o \rangle = \langle [X^*, Y]_o, Z^*_o \rangle - \langle Y_o, ([X, Z]^*_{\mathfrak{m}})_o \rangle + \langle (\nabla_Y Z^*)_o, X^*_o \rangle.$$

Since $[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k}$ and $\mathfrak{m} \cap \mathfrak{k} = (0)$, we have $[X, Z]_{\mathfrak{m}} = 0$, so $\langle Y_o, ([X, Z]^*_{\mathfrak{m}})_o \rangle = 0$.

Since $Y_o \in T_o(G/H)$, there is some $W \in \mathfrak{m}$ such that $Y_o = W^*_o$, so

$$(\nabla_Y Z^*)_o = (\nabla_{Y_o} Z^*)_o = (\nabla_{W^*_o} Z^*)_o = (\nabla_{W^*} Z^*)_o.$$

Furthermore, since a symmetric space is naturally reductive, we showed in Proposition 22.25 that

$$(\nabla_{W^*} Z^*)_o = -\frac{1}{2}([W, Z]^*_{\mathfrak{m}})_o,$$

and since $[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k}$, and $\mathfrak{m} \cap \mathfrak{k} = (0)$, we have $[W, Z]_{\mathfrak{m}} = 0$, which implies that

$$(\nabla_{W^*} Z^*)_o = 0.$$

Therefore, $(\nabla_Y Z^*)_o = 0$, so $\langle (\nabla_{X^*} Y)_o, Z^*_o \rangle = \langle [X^*, Y]_o, Z^*_o \rangle$ for all $Z \in \mathfrak{m}$, and we conclude that

$$(\nabla_{X^*} Y)_o = [X^*, Y]_o,$$

as claimed. □

Another nice property of symmetric space is that the curvature formulae are quite simple. If we use the isomorphism between $\mathfrak{m}$ and $T_o(G/K)$ induced by the restriction of $d\pi_1$ to $\mathfrak{m}$, then for all $X, Y, Z \in \mathfrak{m}$ we have

1. The curvature at o is given by

$$R(X, Y)Z = [[X, Y], Z],$$

or more precisely by

$$R(d\pi_1(X), d\pi_1(Y))d\pi_1(Z) = d\pi_1([[X, Y], Z]).$$

In terms of the vector fields X^*, Y^*, Z^*, we have

$$R(X^*, Y^*)Z^* = [[X, Y], Z]^* = [[X^*, Y^*], Z^*].$$

2. The sectional curvature $K(X^*, Y^*)$ at o is determined by

$$\langle R(X^*, Y^*)X^*, Y^* \rangle = \langle [[X, Y], X], Y \rangle.$$

3. The Ricci curvature at o is given by

$$\mathrm{Ric}(X^*, X^*) = -\frac{1}{2}B(X, X),$$

where B is the Killing form associated with $\mathfrak{g}$.

Proof of the above formulae can be found in O'Neill [91] (Chapter 11), Ziller [119] (Chapter 6), Sakai [100] (Chapter IV) and Helgason [58] (Chapter IV, Section 4). However, beware that Ziller, Sakai and Helgason use the opposite of the sign convention that we are using for the curvature tensor (which is the convention used by O'Neill [91], Gallot, Hulin, Lafontaine [49], Milnor [81], and Arvanitoyeorgos [11]). Recall that we define the Riemann tensor by

$$R(X,Y) = \nabla_{[X,Y]} + \nabla_Y \circ \nabla_X - \nabla_X \circ \nabla_Y,$$

whereas Ziller, Sakai and Helgason use

$$R(X,Y) = -\nabla_{[X,Y]} - \nabla_Y \circ \nabla_X + \nabla_X \circ \nabla_Y.$$

With our convention, the sectional curvature $K(x,y)$ is determined by $\langle R(x,y)x, y\rangle$, and the Ricci curvature $\mathrm{Ric}(x,y)$ as the trace of the map $v \mapsto R(x,v)y$. With the opposite sign convention, the sectional curvature $K(x,y)$ is determined by $\langle R(x,y)y, x\rangle$, and the Ricci curvature $\mathrm{Ric}(x,y)$ as the trace of the map $v \mapsto R(v,x)y$. Therefore, the sectional curvature and the Ricci curvature are identical under both conventions (as they should!). In Ziller, Sakai and Helgason, the curvature formula is

$$R(X^*, Y^*)Z^* = -[[X,Y], Z]^*.$$

We are now going to see that basically all of the familiar spaces are symmetric spaces.

22.9 Examples of Symmetric Spaces

We now apply Theorem 22.34 and construct five families of symmetric spaces. In the first four cases, the Cartan involution is either a conjugation, or the map $\sigma(A) = (A^\top)^{-1}$. We begin by explaining why the Grassmannian manifolds $G(k,n) \cong \mathbf{SO}(n)/S(\mathbf{O}(k) \times \mathbf{O}(n-k))$ and the oriented Grassmannian manifolds $G^0(k,n) \cong \mathbf{SO}(n)/\mathbf{SO}(k) \times \mathbf{SO}(n-k)$ are symmetric spaces. Readers may find material from Absil, Mahony and Sepulchre [2], especially Chapters 1 and 2, a good complement to our presentation.

1. Grassmannians as Symmetric Spaces

Let $G = \mathbf{SO}(n)$ (with $n \geq 2$), let

$$I_{k,n-k} = \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix},$$

where I_k is the $k \times k$-identity matrix, and let σ be given by

$$\sigma(P) = I_{k,n-k} P I_{k,n-k}, \quad P \in \mathbf{SO}(n).$$

It is clear that σ is an involutive automorphism of G. Let us find the set $F = G^\sigma$ of fixed points of σ. If we write

$$P = \begin{pmatrix} Q & U \\ V & R \end{pmatrix}, \qquad Q \in M_{k,k}(\mathbb{R}),\ U \in M_{k,n-k}(\mathbb{R}),\ V \in M_{n-k,k}(\mathbb{R}),\ R \in M_{n-k,n-k}(\mathbb{R}),$$

then $P = I_{k,n-k}PI_{k,n-k}$ iff

$$\begin{pmatrix} Q & U \\ V & R \end{pmatrix} = \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix} \begin{pmatrix} Q & U \\ V & R \end{pmatrix} \begin{pmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{pmatrix}$$

iff

$$\begin{pmatrix} Q & U \\ V & R \end{pmatrix} = \begin{pmatrix} Q & -U \\ -V & R \end{pmatrix},$$

so $U = 0$, $V = 0$, $Q \in \mathbf{O}(k)$ and $R \in \mathbf{O}(n-k)$. Since $P \in \mathbf{SO}(n)$, we conclude that $\det(Q)\det(R) = 1$, so

$$G^{\sigma} = \left\{ \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \;\middle|\; Q \in \mathbf{O}(k),\; R \in \mathbf{O}(n-k),\; \det(R)\det(S) = 1 \right\};$$

that is,

$$F = G^{\sigma} = S(\mathbf{O}(k) \times \mathbf{O}(n-k)),$$

and

$$G_0^{\sigma} = \mathbf{SO}(k) \times \mathbf{SO}(n-k).$$

Therefore, there are two choices for K.

1. $K = \mathbf{SO}(k) \times \mathbf{SO}(n-k)$, in which case we get the Grassmannian $G^0(k,n)$ of oriented k-subspaces.

2. $K = S(\mathbf{O}(k) \times \mathbf{O}(n-k))$, in which case we get the Grassmannian $G(k,n)$ of k-subspaces.

As in the case of Stiefel manifolds, given any $Q \in \mathbf{SO}(n)$, the first k columns Y of Q constitute a representative of the equivalence class $[Q]$, but these representatives are not unique; there is a further equivalence relation given by

$$Y_1 \equiv Y_2 \quad \text{iff} \quad Y_2 = Y_1 R \quad \text{for some } R \in \mathbf{O}(k).$$

Nevertheless, it is useful to consider the first k columns of Q, given by $QP_{n,k}$, as representative of $[Q] \in G(k,n)$.

Because σ is a linear map, its derivative $d\sigma$ is equal to σ, and since $\mathfrak{so}(n)$ consists of all skew-symmetric $n \times n$ matrices, the $+1$-eigenspace is given by

$$\mathfrak{k} = \left\{ \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(k),\; T \in \mathfrak{so}(n-k) \right\},$$

and the -1-eigenspace by

$$\mathfrak{m} = \left\{ \begin{pmatrix} 0 & -A^{\top} \\ A & 0 \end{pmatrix} \;\middle|\; A \in \mathrm{M}_{n-k,k}(\mathbb{R}) \right\}.$$

Thus, $\mathfrak{m}$ is isomorphic to $M_{n-k,k}(\mathbb{R}) \cong \mathbb{R}^{(n-k)k}$. By using the equivalence provided by Proposition 22.22 (1), we can show that the isotropy representation is given by

$$\mathrm{Ad}((Q,R))A = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & R^\top \end{pmatrix} = \begin{pmatrix} 0 & -QA^\top R^\top \\ RAQ^\top & 0 \end{pmatrix} = RAQ^\top,$$

where (Q,R) represents an element of $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$, and A represents an element of $\mathfrak{m}$.

It can be shown that this representation is irreducible iff $(k,n) \neq (2,4)$. It can also be shown that if $n \geq 3$, then $G^0(k,n)$ is simply connected, $\pi_1(G(k,n)) = \mathbb{Z}_2$, and $G^0(k,n)$ is a double cover of $G(k,n)$.

An $\mathrm{Ad}(K)$-invariant inner product on $\mathfrak{m}$ is given by

$$\left\langle \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}, \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix} \right\rangle = -\frac{1}{2}\mathrm{tr}\left(\begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix} \right) = \mathrm{tr}(AB^\top) = \mathrm{tr}(A^\top B).$$

We also give $\mathfrak{g}$ the same inner product. Then we immediately check that $\mathfrak{k}$ and $\mathfrak{m}$ are orthogonal.

In the special case where $k = 1$, we have $G^0(1,n) = S^{n-1}$ and $G(1,n) = \mathbb{RP}^{n-1}$, and then the $\mathbf{SO}(n)$-invariant metric on S^{n-1} (resp. $\mathbb{RP}^{n-1}$) is the canonical one.

For any point $[Q] \in G(k,n)$ with $Q \in \mathbf{SO}(n)$, if we write $Q = [Y\ Y_\perp]$, where Y denotes the first k columns of Q and $Y_\perp$ denotes the last $n-k$ columns of Q, the tangent vectors $X \in T_{[Q]}G(k,n)$ are of the form

$$X = [Y\ Y_\perp] \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} = [Y_\perp A\ \ -YA^\top], \quad A \in M_{n-k,k}(\mathbb{R}).$$

Consequently, there is a one-to-one correspondence between matrices X as above and $n \times k$ matrices of the form $X' = Y_\perp A$, for any matrix $A \in M_{n-k,k}(\mathbb{R})$. As noted in Edelman, Arias and Smith [44], because the spaces spanned by Y and $Y_\perp$ form an orthogonal direct sum in $\mathbb{R}^n$, there is a one-to-one correspondence between $n \times k$ matrices of the form $Y_\perp A$ for any matrix $A \in M_{n-k,k}(\mathbb{R})$, and matrices $X' \in M_{n,k}(\mathbb{R})$ such that

$$Y^\top X' = 0.$$

This second description of tangent vectors to $G(k,n)$ at $[Y]$ is sometimes more convenient. The tangent vectors $X' \in M_{n,k}(\mathbb{R})$ to the Stiefel manifold $S(k,n)$ at Y satisfy the weaker condition that $Y^\top X'$ is *skew-symmetric*.

Indeed, the tangent vectors at Y to the Stiefel manifold $S(k,n)$ are of the form

$$YS + Y_\perp A,$$

with S skew-symmetric, and since the Grassmanian $G(k,n)$ is obtained from the Stiefel manifold $S(k,n)$ by forming the quotient under the equivalence $Y_1 \equiv Y_2$ iff $Y_2 = Y_1 R$, for some $R \in \mathbf{O}(k)$, the contribution YS is a vertical tangent vector at Y in $S(k, n)$, and thus the horizontal tangent vector is $Y_\perp A$; these vectors can be viewed as tangent vectors at $[Y]$ to $G(k, n)$.

Given any $X \in \mathfrak{m}$ of the form

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix},$$

the geodesic starting at o is given by

$$\gamma(t) = \exp(tX) \cdot o.$$

Thus we need to compute

$$\exp(tX) = \exp \begin{pmatrix} 0 & -tA^\top \\ tA & 0 \end{pmatrix}.$$

This can be done using SVD.

Since $G(k, n)$ and $G(n - k, n)$ are isomorphic, without loss of generality, assume that $2k \leq n$. Then let

$$A = U \begin{pmatrix} \Sigma \\ 0_{n-2k,k} \end{pmatrix} V^\top$$

be an SVD for A, with U a $(n - k) \times (n - k)$ orthogonal matrix, Σ a diagonal $k \times k$ matrix, and V a $k \times k$ orthogonal matrix. Since we assumed that $k \leq n - k$, we can write

$$U = [U_1 \, U_2],$$

with U_1 is a $(n - k) \times k$ matrix and U_2 an $(n - k) \times (n - 2k)$ matrix. Then from

$$A = [U_1 \, U_2] \begin{pmatrix} \Sigma \\ 0_{n-2k,k} \end{pmatrix} V^\top = U_1 \Sigma V^\top,$$

we get

$$\begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} = \begin{pmatrix} V & 0 & 0 \\ 0 & U_1 & U_2 \end{pmatrix} \begin{pmatrix} 0 & -\Sigma & 0 \\ \Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} V^\top & 0 \\ 0 & U_1^\top \\ 0 & U_2^\top \end{pmatrix}.$$

(where the middle matrix is $n \times n$). Since

$$\begin{pmatrix} V^\top & 0 \\ 0 & U^\top \end{pmatrix} \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} = \begin{pmatrix} V^\top V & 0 \\ 0 & U^\top U \end{pmatrix} = I_n,$$

the $n \times n$ matrix

$$R = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} = \begin{pmatrix} V & 0 & 0 \\ 0 & U_1 & U_2 \end{pmatrix}$$

is orthogonal, so we have

$$\exp(tX) = \exp\left(t\begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}\right) = R\exp\begin{pmatrix} 0 & -t\Sigma & 0 \\ t\Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} R^\top.$$

Then the computation of the middle exponential proceeds just as in the case where Σ is a scalar, so we get

$$\exp\begin{pmatrix} 0 & -t\Sigma & 0 \\ t\Sigma & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \cos t\Sigma & -\sin t\Sigma & 0 \\ \sin t\Sigma & \cos t\Sigma & 0 \\ 0 & 0 & I \end{pmatrix},$$

so

$$\exp(tX) = \exp\left(t\begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}\right) = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix}\begin{pmatrix} \cos t\Sigma & -\sin t\Sigma & 0 \\ \sin t\Sigma & \cos t\Sigma & 0 \\ 0 & 0 & I \end{pmatrix}\begin{pmatrix} V^\top & 0 \\ 0 & U^\top \end{pmatrix}.$$

Now, $\exp(tX)P_{n,k}$ is certainly a representative of the equivalence class of $[\exp(tX)]$, so as a $n \times k$ matrix, the geodesic through o with initial velocity

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}$$

(with A any $(n-k)\times k$ matrix with $n-k \geq k$) is given by

$$\gamma(t) = \begin{pmatrix} V & 0 \\ 0 & U_1 \end{pmatrix}\begin{pmatrix} \cos t\Sigma \\ \sin t\Sigma \end{pmatrix} V^\top,$$

where $A = U_1\Sigma V^\top$, a compact SVD of A.

Remark: Because symmetric spaces are geodesically complete, we get an interesting corollary. Indeed, every equivalence class $[Q] \in G(k,n)$ possesses some representative of the form e^X for some $X \in \mathfrak{m}$, so we conclude that for every orthogonal matrix $Q \in \mathbf{SO}(n)$, there exist some orthogonal matrices $V, \widetilde{V} \in \mathbf{O}(k)$ and $U, \widetilde{U} \in \mathbf{O}(n-k)$, and some diagonal matrix Σ with nonnegatives entries, so that

$$Q = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix}\begin{pmatrix} \cos \Sigma & -\sin \Sigma & 0 \\ \sin \Sigma & \cos \Sigma & 0 \\ 0 & 0 & I \end{pmatrix}\begin{pmatrix} (\widetilde{V})^\top & 0 \\ 0 & (\widetilde{U})^\top \end{pmatrix},$$

because a matrix in the coset of Q is obtained by multiplying on the right by a matrix in the stabilizer, and these matrices are of the form

$$\begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \quad \text{with } Q_1 \in \mathbf{O}(k) \text{ and } Q_2 \in \mathbf{O}(n-k).$$

The above is an instance of the CS-decomposition; see Golub and Van Loan [51]. The matrices $\cos\Sigma$ and $\sin\Sigma$ are actually diagonal matrices of the form

$$\cos\Sigma = \mathrm{diag}(\cos\theta_1, \ldots, \cos\theta_k) \quad \text{and} \quad \sin\Sigma = \mathrm{diag}(\sin\theta_1, \ldots, \sin\theta_k),$$

so we may assume that $0 \leq \theta_i \leq \pi/2$, because if $\cos\theta_i$ or $\sin\theta_i$ is negative, we can change the sign of the ith row of V (resp. the sign of the i-th row of U) and still obtain orthogonal matrices U' and V' that do the job. One should also observe that the first k columns of Q are

$$Y = \begin{pmatrix} V & 0 \\ 0 & U \end{pmatrix} \begin{pmatrix} \cos\Sigma \\ \sin\Sigma \\ 0 \end{pmatrix} (\widetilde{V})^\top,$$

and that the matrix

$$V(\cos\Sigma)(\widetilde{V})^\top$$

is an SVD for the matrix $P_{n,k}^\top Y$, which consists of the first k rows of Y. Now it is known that $(\theta_1, \ldots, \theta_k)$ are the *principal angles* (or *Jordan angles*) between the subspaces spanned the first k columns of I_n and the subspace spanned by the columns of Y (see Golub and van Loan [51]).

Recall that given two k-dimensional subspaces $\mathcal{U}$ and $\mathcal{V}$ determined by two $n \times k$ matrices Y_1 and Y_2 of rank k, the *principal angles* $\theta_1, \ldots, \theta_k$ between $\mathcal{U}$ and $\mathcal{V}$ are defined recursively as follows. Let $\mathcal{U}_1 = \mathcal{U}$, $\mathcal{V}_1 = \mathcal{V}$, let

$$\cos\theta_1 = \max_{\substack{u \in \mathcal{U}, v \in \mathcal{V} \\ \|u\|_2 = 1, \|v\|_2 = 1}} \langle u, v\rangle,$$

let $u_1 \in \mathcal{U}$ and $v_1 \in \mathcal{V}$ be any two unit vectors such that $\cos\theta_1 = \langle u_1, v_1\rangle$, and for $i = 2, \ldots, k$, if $\mathcal{U}_i = \mathcal{U}_{i-1} \cap \{u_{i-1}\}^\perp$ and $\mathcal{V}_i = \mathcal{V}_{i-1} \cap \{v_{i-1}\}^\perp$, let

$$\cos\theta_i = \max_{\substack{u \in \mathcal{U}_i, v \in \mathcal{V}_i \\ \|u\|_2 = 1, \|v\|_2 = 1}} \langle u, v\rangle,$$

and let $u_i \in \mathcal{U}_i$ and $v_i \in \mathcal{V}_i$ be any two unit vectors such that $\cos\theta_i = \langle u_i, v_i\rangle$.

The vectors u_i and v_i are not unique, but it is shown in Golub and van Loan [51] that $(\cos\theta_1, \ldots, \cos\theta_k)$ are the singular values of $Y_1^\top Y_2$ (with $0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_k \leq \pi/2$).

We can also determine the length $L(\gamma)(s)$ of the geodesic $\gamma(t)$ from o to $p = e^{sX} \cdot o$, for any $X \in \mathfrak{m}$, with

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}.$$

Since by Proposition 17.8(1), the length of the geodesic γ from o to $p = e^{sX} \cdot o$ is the same as the the length of the geodesic $\overline{\gamma}$ in G from 1 to e^{sX}, for any $X \in \mathfrak{m}$, the computation from Section 22.7 remains valid, and we obtain

$$(L(\gamma)(s))^2 = (L(\overline{\gamma})(s))^2 = s^2\left(\frac{1}{2}\mathrm{tr}(X^\top X)\right) = s^2\mathrm{tr}(A^\top A).$$

Then if $\theta_1, \ldots, \theta_k$ are the singular values of A, we get

$$L(\gamma)(s) = s(\theta_1^2 + \cdots + \theta_k^2)^{\frac{1}{2}}.$$

In view of the above discussion regarding principal angles, we conclude that if Y_1 consists of the first k columns of an orthogonal matrix Q_1 and if Y_2 consists of the first k columns of an orthogonal matrix Q_2 then the distance between the subspaces $[Q_1]$ and $[Q_2]$ is given by

$$d([Q_1], [Q_2]) = (\theta_1^2 + \cdots + \theta_k^2)^{\frac{1}{2}},$$

where $(\cos\theta_1, \ldots, \cos\theta_k)$ are the singular values of $Y_1^\top Y_2$ (with $0 \leq \theta_i \leq \pi/2$); the angles $(\theta_1, \ldots, \theta_k)$ are the principal angles between the spaces $[Q_1]$ and $[Q_2]$.

In Golub and van Loan, a different distance between subspaces is defined, namely

$$d_{p2}([Q_1], [Q_2]) = \left\| Y_1 Y_1^\top - Y_2 Y_2^\top \right\|_2.$$

If we write $\Theta = \mathrm{diag}(\theta_1 \ldots, \theta_k)$, then it is shown that

$$d_{p2}([Q_1], [Q_2]) = \|\sin\Theta\|_\infty = \max_{1 \leq i \leq k} \sin\theta_i.$$

This metric is derived by embedding the Grassmannian in the set of $n \times n$ projection matrices of rank k, and then using the 2-norm. Other metrics are proposed in Edelman, Arias and Smith [44].

We leave it to the brave readers to compute $\langle [[X, Y], X], Y \rangle$, where

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix},$$

and check that

$$\langle [[X, Y], X], Y \rangle = \langle BA^\top - AB^\top, BA^\top - AB^\top \rangle + \langle A^\top B - B^\top A, A^\top B - B^\top A \rangle,$$

which shows that the sectional curvature is nonnegative. When $k = 1$ (or $k = n - 1$), which corresponds to $\mathbb{RP}^{n-1}$ (or S^{n-1}), we get a metric of constant positive curvature.

2. Symmetric Positive Definite Matrices

Recall from Example 4.7 that the space $\mathbf{SPD}(n)$ of symmetric positive definite matrices ($n \geq 2$) appears as the homogeneous space $\mathbf{GL}^+(n, \mathbb{R})/\mathbf{SO}(n)$, under the action of $\mathbf{GL}^+(n, \mathbb{R})$ on $\mathbf{SPD}(n)$ given by

$$A \cdot S = ASA^\top, \qquad A \in \mathbf{GL}^+(n, \mathbb{R}),\ S \in \mathbf{SPD}(n).$$

Write $G = \mathbf{GL}^+(n, \mathbb{R})$, $K = \mathbf{SO}(n)$, and choose the Cartan involution σ given by

$$\sigma(S) = (S^\top)^{-1}, \qquad S \in \mathbf{GL}^+(n, \mathbb{R}).$$

It is immediately verified that

$$G^\sigma = \mathbf{SO}(n),$$

and that the derivative $\theta = d\sigma_1$ of σ is given by

$$\theta(S) = (\sigma(e^{tS}))'(0) = (e^{-tS^\top})'(0) = -S^\top, \quad S \in \mathrm{M}_n(\mathbb{R}),$$

since $\mathfrak{gl}^+(n) = \mathfrak{gl}(n) = \mathrm{M}_n(\mathbb{R})$. It follows that $\mathfrak{k} = \mathfrak{so}(n)$, and $\mathfrak{m} = \mathbf{S}(n)$, the vector space of symmetric matrices. We define an $\mathrm{Ad}(\mathbf{SO}(n))$-invariant inner product on $\mathfrak{gl}^+(n)$ by

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y).$$

If $X \in \mathfrak{m}$ and $Y \in \mathfrak{k} = \mathfrak{so}(n)$, then

$$\langle X, Y\rangle = \mathrm{tr}(X^\top Y) = \mathrm{tr}((X^\top Y)^\top) = \mathrm{tr}(Y^\top X) = -\mathrm{tr}(Y X^\top) = -\mathrm{tr}(X^\top Y) = -\langle X, Y\rangle,$$

so $\langle X, Y\rangle = 0$. Thus we have

$$\langle X, Y\rangle = \begin{cases} -\mathrm{tr}(XY) & \text{if } X, Y \in \mathfrak{k} \\ \mathrm{tr}(XY) & \text{if } X, Y \in \mathfrak{m} \\ 0 & \text{if } X \in \mathfrak{m},\ Y \in \mathfrak{k}. \end{cases}$$

We leave it as an exercise (see Petersen [93], Chapter 8, Section 2.5) to show that

$$\langle [[X, Y], X], Y\rangle = -\mathrm{tr}([X, Y]^\top [X, Y]), \quad \text{for all } X, Y \in \mathfrak{m}.$$

This shows that the sectional curvature is nonpositive. It can also be shown that the isotropy representation is given by

$$\chi_A(X) = AXA^{-1} = AXA^\top,$$

for all $A \in \mathbf{SO}(n)$ and all $X \in \mathfrak{m}$.

Recall that the exponential $\exp\colon \mathbf{S}(n) \to \mathbf{SPD}(n)$ is a bijection. Then given any $S \in \mathbf{SPD}(n)$, there is a unique $X \in \mathfrak{m}$ such that $S = e^X$, and the unique geodesic from I to S is given by

$$\gamma(t) = e^{tX}.$$

Let us try to find the length $L(\gamma) = d(I, S)$ of this geodesic. As in Section 22.7, we have

$$L(\gamma) = \int_0^1 \langle \gamma'(t), \gamma'(t)\rangle^{\frac{1}{2}} dt,$$

but this time, $X \in \mathfrak{m}$ is symmetric and the geodesic is unique, so we have

$$\begin{aligned} L(\gamma) &= \int_0^1 \langle (e^{tX})', (e^{tX})'\rangle^{\frac{1}{2}} dt \\ &= \int_0^1 \langle Xe^{tX}, Xe^{tX}\rangle^{\frac{1}{2}} dt \\ &= \int_0^1 (\mathrm{tr}((e^{tX})^\top X^\top X e^{tX}))^{\frac{1}{2}} dt \\ &= \int_0^1 (\mathrm{tr}(X^2 e^{2tX}))^{\frac{1}{2}} dt. \end{aligned}$$

Since X is a symmetric matrix, we can write

$$X = P^{\top}\Lambda P,$$

with P orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, a real diagonal matrix, and we have

$$\begin{aligned}\mathrm{tr}(X^2 e^{2tX}) &= \mathrm{tr}(P^{\top}\Lambda^2 P P^{\top} e^{2t\Lambda} P)\\ &= \mathrm{tr}(\Lambda^2 e^{2t\Lambda})\\ &= \lambda_1^2 e^{2t\lambda_1} + \cdots + \lambda_n^2 e^{2t\lambda_n}.\end{aligned}$$

Therefore,

$$d(I,S) = L(\gamma) = \int_0^1 (\lambda_1^2 e^{2\lambda_1 t} + \cdots + \lambda_n^2 e^{2\lambda_n t})^{\frac{1}{2}} dt.$$

Actually, since $S = e^X$ and S is SPD, $\lambda_1, \ldots, \lambda_n$ are the logarithms of the eigenvalues $\sigma_1, \ldots, \sigma_n$ of X, so we have

$$d(I,S) = L(\gamma) = \int_0^1 ((\log\sigma_1)^2 e^{2\log\sigma_1 t} + \cdots + (\log\sigma_n)^2 e^{2\log\sigma_n t})^{\frac{1}{2}} dt.$$

Unfortunately, there doesn't appear to be a closed form formula for this integral.

The symmetric space $\mathbf{SPD}(n)$ contains an interesting submanifold, namely the space of matrices S in $\mathbf{SPD}(n)$ such that $\det(S) = 1$. This the symmetric space $\mathbf{SL}(n, \mathbb{R})/\mathbf{SO}(n)$, which we suggest denoting by $\mathbf{SSPD}(n)$. For this space, $\mathfrak{g} = \mathfrak{sl}(n)$, and the reductive decomposition is given by

$$\mathfrak{k} = \mathfrak{so}(n), \quad \mathfrak{m} = \mathbf{S}(n) \cap \mathfrak{sl}(n).$$

Recall that the Killing form on $\mathfrak{gl}(n)$ is given by

$$B(X,Y) = 2n\mathrm{tr}(XY) - 2\mathrm{tr}(X)\mathrm{tr}(Y).$$

On $\mathfrak{sl}(n)$, the Killing form is $B(X,Y) = 2n\mathrm{tr}(XY)$, and restricted to $\mathbf{S}(n)$ it is proportional to the inner product

$$\langle X, Y\rangle = \mathrm{tr}(XY).$$

Therefore, we see that the restriction of the Killing form of $\mathfrak{sl}(n)$ to $\mathfrak{m} = \mathbf{S}(n) \cap \mathfrak{sl}(n)$ is positive definite, whereas it is negative definite on $\mathfrak{k} = \mathfrak{so}(n)$. The symmetric space $\mathbf{SSPD}(n) \cong \mathbf{SL}(n, \mathbb{R})/\mathbf{SO}(n)$ is an example of a symmetric space of noncompact type. On the other hand, the Grassmannians are examples of symmetric spaces of compact type (for $n \geq 3$). In the next section, we take a quick look at these special types of symmetric spaces.

3. The Hyperbolic Space $\mathcal{H}_n^+(1)$ ⊛

In Section 5.1 we defined the Lorentz group $\mathbf{SO}_0(n, 1)$ as follows: if

$$J = \begin{pmatrix} I_n & 0 \\ 0 & -1 \end{pmatrix},$$

then a matrix $A \in \mathrm{M}_{n+1}(\mathbb{R})$ belongs to $\mathbf{SO}_0(n,1)$ iff

$$A^\top JA = J, \quad \det(A) = +1, \quad a_{n+1n+1} > 0.$$

In that same section we also defined the hyperbolic space $\mathcal{H}_n^+(1)$ as the sheet of $\mathcal{H}_n(1)$ which contains $(0,\ldots,0,1)$ where

$$\mathcal{H}_n(1) = \{u = (\mathbf{u}, t) \in \mathbb{R}^{n+1} \mid \|\mathbf{u}\|^2 - t^2 = -1\}.$$

We also showed that the action $\cdot\colon \mathbf{SO}_0(n,1) \times \mathcal{H}_n^+(1) \longrightarrow \mathcal{H}_n^+(1)$ with

$$A \cdot u = Au$$

is a transitive with stabilizer $\mathbf{SO}(n)$ (see Proposition 5.10). Thus, $\mathcal{H}_n^+(1)$ arises as the homogeneous space $\mathbf{SO}_0(n,1)/\mathbf{SO}(n)$.

Since the inverse of $A \in \mathbf{SO}_0(n,1)$ is $JA^\top J$, the map $\sigma\colon \mathbf{SO}_0(n,1) \to \mathbf{SO}_0(n,1)$ given by

$$\sigma(A) = JAJ = (A^\top)^{-1}$$

is an involutive automorphism of $\mathbf{SO}_0(n,1)$. Write $G = \mathbf{SO}_0(n,1)$, $K = \mathbf{SO}(n)$. It is immediately verified that

$$G^\sigma = \left\{ \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix} \mid Q \in \mathbf{SO}(n) \right\},$$

so $G^\sigma \cong \mathbf{SO}(n)$. We have

$$\mathfrak{so}(n,1) = \left\{ \begin{pmatrix} B & u \\ u^\top & 0 \end{pmatrix} \mid B \in \mathfrak{so}(n), u \in \mathbb{R}^n \right\},$$

and the derivative $\theta\colon \mathfrak{so}(n,1) \to \mathfrak{so}(n,1)$ of σ at I is given by

$$\theta(X) = JXJ = -X^\top.$$

From this we deduce that the $+1$-eigenspace is given by

$$\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} \mid B \in \mathfrak{so}(n) \right\},$$

and the -1-eigenspace is given by

$$\mathfrak{m} = \left\{ \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \mid u \in \mathbb{R}^n \right\},$$

with

$$\mathfrak{so}(n,1) = \mathfrak{k} \oplus \mathfrak{m},$$

a reductive decomposition. We define an $\mathrm{Ad}(K)$-invariant inner product on $\mathfrak{so}(n,1)$ by

$$\langle X, Y\rangle = \frac{1}{2}\mathrm{tr}(X^\top Y).$$

In fact, on $\mathfrak{m} \cong \mathbb{R}^n$, we have

$$\left\langle \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix} \right\rangle = \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix} \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix} \right) = \frac{1}{2}\mathrm{tr}(uv^\top + u^\top v) = u^\top v,$$

the Euclidean product of u and v.

As an exercise, the reader should compute $\langle [[X,Y],X],Y\rangle$, where

$$X = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix},$$

and check that

$$\langle [[X,Y],X],Y\rangle = -\langle uv^\top - vu^\top, uv^\top - vu^\top\rangle,$$

which shows that the sectional curvature is nonpositive. In fact, $\mathcal{H}_n^+(1)$ has constant negative sectional curvature.

We leave it as an exercise to prove that for $n \geq 2$, the Killing form B on $\mathfrak{so}(n,1)$ is given by

$$B(X,Y) = (n-1)\mathrm{tr}(XY),$$

for all $X, Y \in \mathfrak{so}(n,1)$. If we write

$$X = \begin{pmatrix} B_1 & u \\ u^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} B_2 & v \\ v^\top & 0 \end{pmatrix},$$

then

$$B(X,Y) = (n-1)\mathrm{tr}(B_1B_2) + 2(n-1)u^\top v.$$

This shows that B is negative definite on $\mathfrak{k}$ and positive definite on $\mathfrak{m}$. This means that the space $\mathcal{H}_n^+(1)$ is a symmetric space of noncompact type.

The symmetric space $\mathcal{H}_n^+(1) = \mathbf{SO}_0(n,1)/\mathbf{SO}(n)$ turns out to be dual, as a symmetric space, to $S^n = \mathbf{SO}(n+1)/\mathbf{SO}(n)$. For the precise notion of duality in symmetric spaces, we refer the reader to O'Neill [91].

4. The Hyperbolic Grassmannian $G^*(q, p+q)$⊛

This is the generalization of the hyperbolic space $\mathcal{H}_n^+(1)$ in Example (3). Recall from Section 5.1 that we define $I_{p,q}$, for $p, q \geq 1$, by

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}.$$

If $n = p + q$, the matrix $I_{p,q}$ is associated with the nondegenerate symmetric bilinear form

$$\varphi_{p,q}((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{p} x_i y_i - \sum_{j=p+1}^{n} x_j y_j$$

with associated quadratic form

$$\Phi_{p,q}((x_1, \ldots, x_n)) = \sum_{i=1}^{p} x_i^2 - \sum_{j=p+1}^{n} x_j^2.$$

The group $\mathbf{SO}(p, q)$ is the set of all $n \times n$-matrices (with $n = p + q$)

$$\mathbf{SO}(p, q) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q},\ \det(A) = 1\}.$$

If we write

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}, \qquad P \in M_p(\mathbb{R}),\ \ Q \in M_q(\mathbb{R})$$

then it is shown in O'Neill [91] (Chapter 9, Lemma 6) that the connected component $\mathbf{SO}_0(p, q)$ of $\mathbf{SO}(p, q)$ containing I is given by

$$\mathbf{SO}_0(p, q) = \{A \in \mathbf{GL}(n, \mathbb{R}) \mid A^\top I_{p,q} A = I_{p,q},\ \det(P) > 0,\ \det(S) > 0\}.$$

For both $\mathbf{SO}(p, q)$ and $\mathbf{SO}_0(p, q)$, the inverse is given by

$$A^{-1} = I_{p,q} A^\top I_{p,q}.$$

This implies that the map $\sigma\colon \mathbf{SO}_0(p, q) \to \mathbf{SO}_0(p, q)$ given by

$$\sigma(A) = I_{p,q} A I_{p,q} = (A^\top)^{-1}$$

is an involution, and its fixed subgroup G^σ is given by

$$G^\sigma = \left\{ \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \mid Q \in \mathbf{SO}(p), R \in \mathbf{SO}(q) \right\}.$$

Thus G^σ is isomorphic to $\mathbf{SO}(p) \times \mathbf{SO}(q)$.

For $p, q \geq 1$, the Lie algebra $\mathfrak{so}(p, q)$ of $\mathbf{SO}_0(p, q)$ (and $\mathbf{SO}(p, q)$ as well) is given by

$$\mathfrak{so}(p, q) = \left\{ \begin{pmatrix} B & A \\ A^\top & C \end{pmatrix} \mid B \in \mathfrak{so}(p), C \in \mathfrak{so}(q), A \in \mathrm{M}_{p,q}(\mathbb{R}) \right\}.$$

Since $\theta = d\sigma_I$ is also given by $\theta(X) = I_{p,q} X I_{p,q} = -X^\top$, we find that the $+1$-eigenspace $\mathfrak{k}$ of θ is given by

$$\mathfrak{k} = \left\{ \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix} \mid B \in \mathfrak{so}(p), C \in \mathfrak{so}(q) \right\},$$

and the -1-eigenspace $\mathfrak{m}$ of θ is is given by

$$\mathfrak{m} = \left\{ \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix} \mid A \in \mathrm{M}_{p,q}(\mathbb{R}) \right\}.$$

Note that $\mathfrak{k}$ is a subalgebra of $\mathfrak{so}(p,q)$ and $\mathfrak{so}(p,q) = \mathfrak{k} \oplus \mathfrak{m}$.

Write $G = \mathbf{SO}_o(p,q)$ and $K = \mathbf{SO}(p) \times \mathbf{SO}(q)$. We define an $\mathrm{Ad}(K)$-invariant inner product on $\mathfrak{so}(p,q)$ by

$$\langle X, Y \rangle = \frac{1}{2} \mathrm{tr}(X^\top Y).$$

Therefore, for $p, q \geq 1$, the coset space $\mathbf{SO}_0(p,q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$ is a symmetric space. Observe that on $\mathfrak{m}$, the above inner product is given by

$$\langle X, Y \rangle = \frac{1}{2} \mathrm{tr}(XY).$$

On the other hand, in the case of $\mathbf{SO}(p+q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$, on $\mathfrak{m}$, the inner product is given by

$$\langle X, Y \rangle = -\frac{1}{2} \mathrm{tr}(XY).$$

This space can be described explicitly. Indeed, let $G^*(q, p+q)$ be the set of q-dimensional subspaces W of $R^n = R^{p+q}$ such that $\Phi_{p,q}$ is negative definite on W. Then we have an obvious matrix multiplication action of $\mathbf{SO}_0(p,q)$ on $G^*(q, p+q)$, and it is easy to check that this action is transitive. It is not hard to show that the stabilizer of the subspace spanned by the last q columns of the $(p+q) \times (p+q)$ identity matrix is $\mathbf{SO}(p) \times \mathbf{SO}(q)$, so the space $G^*(q, p+q)$ is isomorphic to the homogeneous (symmetric) space $\mathbf{SO}_0(p,q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$.

Definition 22.15. The symmetric space $G^*(q, p+q) \cong \mathbf{SO}_0(p,q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$ is called the *hyperbolic Grassmannian*.

Assume that $p + q \geq 3, p, q \geq 1$. Then it can be shown that the Killing form on $\mathfrak{so}(p,q)$ is given by

$$B(X, Y) = (p + q - 2)\mathrm{tr}(XY),$$

so $\mathfrak{so}(p,q)$ is semisimple. If we write

$$X = \begin{pmatrix} B_1 & A_1 \\ A_1^\top & C_1 \end{pmatrix}, \quad Y = \begin{pmatrix} B_2 & A_2 \\ A_2^\top & C_2 \end{pmatrix},$$

then

$$B(X, Y) = (p + q - 2)(\mathrm{tr}(B_1 B_2) + \mathrm{tr}(C_1 C_2)) + 2(p + q - 2) A_1^\top A_2.$$

Consequently, B is negative definite on $\mathfrak{k}$ and positive definite on $\mathfrak{m}$, so $G^*(q, p+q) = \mathbf{SO}_0(p,q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$ is another example of a symmetric space of noncompact type.

We leave it to the reader to compute $\langle [[X, Y], X], Y\rangle$, where

$$X = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & B \\ B^\top & 0 \end{pmatrix},$$

and check that

$$\langle [[X, Y], X], Y\rangle = -\langle BA^\top - AB^\top, BA^\top - AB^\top\rangle - \langle A^\top B - B^\top A, A^\top B - B^\top A\rangle,$$

which shows that the sectional curvature is nonpositive. In fact, the above expression is the negative of the expression that we found for the sectional curvature of $G^0(p, p+q)$. When $p = 1$ or $q = 1$, we get a space of constant negative curvature.

The above property is one of the consequences of the fact that the space $G^*(q, p+q) = \mathbf{SO}_0(p, q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$ is the symmetric space dual to
$G^0(p, p+q) = \mathbf{SO}(p+q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$, the Grassmannian of oriented p-planes; see O'Neill [91] (Chapter 11, Definition 37) or Helgason [58] (Chapter V, Section 2).

5. Compact Lie Groups

If H be a compact Lie group, then $G = H \times H$ is the group with multiplication given by $(h_1, h_2) \cdot (h_1', h_2') = (h_1h_1', h_2h_2')$. The group $G = H \times H$ acts on H *via*

$$(h_1, h_2) \cdot h = h_1 h h_2^{-1}.$$

The stabilizer of 1 is clearly $K = \Delta H = \{(h, h) \mid h \in H\}$. It is easy to see that the map

$$(h_1, h_2)K \mapsto h_1 h_2^{-1}$$

is a diffeomorphism between the coset space G/K and H (see Helgason [58], Chapter IV, Section 6). A Cartan involution σ on G is given by

$$\sigma(h_1, h_2) = (h_2, h_1),$$

and obviously $G^\sigma = K = \Delta H$. Therefore, H appears as the symmetric space G/K, with $G = H \times H$, $K = \Delta H$, and

$$\mathfrak{k} = \{(X, X) \mid X \in \mathfrak{h}\}, \quad \mathfrak{m} = \{(X, -X) \mid X \in \mathfrak{h}\}.$$

For every $(h_1, h_2) \in \mathfrak{g}$, we have

$$(h_1, h_2) = \left(\frac{h_1 + h_2}{2}, \frac{h_1 + h_2}{2}\right) + \left(\frac{h_1 - h_2}{2}, -\frac{h_1 - h_2}{2}\right)$$

which gives the direct sum decomposition

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}.$$

The natural projection $\pi\colon H \times H \to H$ is given by

$$\pi(h_1, h_2) = h_1 h_2^{-1},$$

which yields $d\pi_{(1,1)}(X, Y) = X - Y$ (see Helgason [58], Chapter IV, Section 6). It follows that the natural isomorphism $\mathfrak{m} \to \mathfrak{h}$ is given by

$$(X, -X) \mapsto 2X.$$

Given any bi-invariant metric $\langle -, - \rangle$ on H, define a metric on $\mathfrak{m}$ by

$$\langle (X, -X), (Y, -Y) \rangle = 4\langle X, Y \rangle.$$

The reader should check that the resulting symmetric space is isometric to H (see Sakai [100], Chapter IV, Exercise 4).

More examples of symmetric spaces are presented in Ziller [119] and Helgason [58]. For example, the complex Grassmannian

$$\mathbf{SU}(n)/S(\mathbf{U}(k) \times \mathbf{U}(n-k)) \cong G_{\mathbb{C}}(k, n)$$

is a symmetric space. The Cartan involution is also given by $\sigma(U) = I_{k,n-k} U I_{k,n-k}$, with $U \in \mathbf{SU}(n)$.

To close our brief tour of symmetric spaces, we conclude with a short discussion about the type of symmetric spaces.

22.10 Types of Symmetric Spaces

Suppose (G, K, σ) (G connected and K compact) presents a symmetric space with Cartan involution σ, and with

$$\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m},$$

where $\mathfrak{k}$ (the Lie algebra of K) is the eigenspace of $d\sigma_1$ associated with the eigenvalue $+1$ and $\mathfrak{m}$ is is the eigenspace associated with the eigenvalue -1. If B is the Killing form of $\mathfrak{g}$, it turns out that the restriction of B to $\mathfrak{k}$ is always negative semidefinite. This will be shown as the first part of the proof of Proposition 22.37. However, to guarantee that B is negative definite (that is, $B(Z, Z) = 0$ implies that $Z = 0$) some additional condition is needed.

This condition has to do with the subgroup N of G defined by

$$N = \{g \in G \mid \tau_g = \mathrm{id}\} = \{g \in G \mid gaK = aK \text{ for all } a \in G\}.$$

By setting $a = e$, we see that $N \subseteq K$. Furthermore, since $n \in N$ implies $na^{-1}bK = a^{-1}bK$ for all $a, b \in G$, we can readily show that N is a normal subgroup of both K and G. It is not hard to show that N is the largest normal subgroup that K and G have in common (see Ziller [119] (Chapter 6, Section 6.2).

We can also describe the subgroup N in a more explicit fashion. We have

$$\begin{aligned} N &= \{g \in G \mid gaK = aK \ \text{ for all } a \in G\} \\ &= \{g \in G \mid a^{-1}gaK = K \ \text{ for all } a \in G\} \\ &= \{g \in G \mid a^{-1}ga \in K \ \text{ for all } a \in G\}. \end{aligned}$$

Definition 22.16. For any Lie group G and any closed subgroup K of G, the subgroup N of G given by

$$N = \{g \in G \mid a^{-1}ga \in K \ \text{ for all } a \in G\}$$

is called the *ineffective kernel* of the left action of G on G/K. The left action of G on G/K is said to be *effective* (or *faithful*) if $N = \{1\}$, *almost effective* if N is a discrete subgroup.

If K is compact, which will be assumed from now on, since a discrete subgroup of a compact group is finite, *the action of G on G/K is almost effective if N is finite.*

For example, the action $\cdot\colon \mathbf{SU}(n+1) \times \mathbb{CP}^n \to \mathbb{CP}^n$ of $\mathbf{SU}(n+1)$ on the (complex) projective space $\mathbb{CP}^n$ discussed in Example (e) of Section 4.3 is almost effective but not effective. It presents $\mathbb{CP}^n$ as the homogeneous manifold

$$\mathbf{SU}(n+1)/S(\mathbf{U}(1) \times \mathbf{U}(n)) \cong \mathbb{CP}^n.$$

We leave it as an exercise to the reader to prove that the ineffective kernel of the above action is the finite group

$$N = \{\lambda I_{n+1} \mid \lambda^{n+1} = 1,\ \lambda \in \mathbb{C}\}.$$

It turns out that the additional requirement needed for the Killing form to be negative definite is that the action of G on G/K is almost effective.

The following technical proposition gives a criterion for the left action of G on G/K to be almost effective in terms of the Lie algebras $\mathfrak{g}$ and $\mathfrak{k}$. This is Proposition 6.27 from Ziller [119].

Proposition 22.36. *The left action of G on G/K (with K compact) is almost effective iff $\mathfrak{g}$ and $\mathfrak{k}$ have no nontrivial ideal in common.*

Proof. By a previous remark, the effective kernel N of the left action of G on G/K is the largest normal subgroup that K and G have in common. To say that N is finite is equivalent to saying that N is discrete (since K is compact), which is equivalent to the fact that its Lie algebra $\mathfrak{n} = (0)$. Since by Theorem 18.19 normal subgroups correspond to ideals, the condition that the largest normal subgroup that K and G have in common is finite is equivalent to the condition that $\mathfrak{g}$ and $\mathfrak{k}$ have no nontrivial ideal in common. □

Proposition 22.37. *Let (G, K, σ) be a symmetric space (K compact) with Cartan involution σ, and assume that the left action of G on G/K is almost effective. If B is the Killing form of $\mathfrak{g}$ and $\mathfrak{k} \neq (0)$, then the restriction of B to $\mathfrak{k}$ is negative definite.*

Proof. (After Ziller [119], Proposition 6.38). The restriction of the Ad-representation of G to K yields a representation $\mathrm{Ad}\colon K \to \mathbf{GL}(\mathfrak{g})$. Since K is compact, by Theorem 20.4 there is an $\mathrm{Ad}(K)$-invariant inner product on $\mathfrak{g}$. Then for $k \in K$, we have

$$\langle \mathrm{Ad}_k(X), \mathrm{Ad}_k(Y)\rangle = \langle X, Y\rangle, \quad \text{for all } X, Y \in \mathfrak{g},$$

so for $k = \exp(tZ)$ with $Z \in \mathfrak{k}$, by taking derivatives at $t = 0$, we get

$$\langle [X, Z], Y\rangle = \langle X, [Z, Y]\rangle, \quad X, Y \in \mathfrak{g}, Z \in \mathfrak{k},$$

which can be written as

$$-\langle [Z, X], Y\rangle = \langle [Z, Y], X\rangle, \quad X, Y, \in \mathfrak{g}, Z \in \mathfrak{k}.$$

Consequently $\mathrm{ad}(Z)$ is a skew-symmetric linear map on $\mathfrak{g}$ for all $Z \in \mathfrak{k}$. But then, $\mathrm{ad}(Z)$ is represented by a skew symmetric matrix (a_{ij}) in any orthonormal basis of $\mathfrak{g}$, and so

$$B(Z, Z) = \mathrm{tr}(\mathrm{ad}(Z) \circ \mathrm{ad}(Z)) = -\sum_{i,j=1}^{n} a_{ij}^2 \leq 0.$$

Next, we need to prove that if $B(Z, Z) = 0$, then $Z = 0$. This is equivalent to proving that if $\mathrm{ad}(Z) = 0$ then $Z = 0$. However, $\mathrm{ad}(Z) = 0$ means that $[Z, X] = 0$ for all $X \in \mathfrak{g}$, so Z belongs to the center of $\mathfrak{g}$,

$$\mathfrak{z}(\mathfrak{g}) = \{Z \in \mathfrak{g} \mid [Z, X] = 0 \text{ for all } X \in \mathfrak{g}\}.$$

It is immediately verified that $\mathfrak{z}(\mathfrak{g})$ is an ideal of $\mathfrak{g}$. But now, $Z \in \mathfrak{z}(\mathfrak{g}) \cap \mathfrak{k}$, which is an ideal of both $\mathfrak{g}$ and $\mathfrak{k}$ by definition of $\mathfrak{z}(\mathfrak{g})$, and since the left action of G on G/K is almost effective, by Proposition 22.36, the Lie algebras $\mathfrak{g}$ and $\mathfrak{k}$ have no nontrivial ideal in common, so $\mathfrak{z}(\mathfrak{g}) \cap \mathfrak{k} = (0)$, and $Z = 0$. □

In view of Proposition 22.37, it is natural to classify symmetric spaces depending on the behavior of B on $\mathfrak{m}$.

Definition 22.17. Let $M = (G, K, \sigma)$ be a symmetric space (K compact) with Cartan involution σ and Killing form B, and assume that the left action of G on G/K is almost effective. The space M is said to be of

(1) *Euclidean type* if $B = 0$ on $\mathfrak{m}$.

(2) *Compact type* if B is negative definite on $\mathfrak{m}$.

(3) *Noncompact type* if B is positive definite on $\mathfrak{m}$.

Proposition 22.38. *Let $M = (G, K, \sigma)$ be a symmetric space (K compact) with Cartan involution σ and Killing form B on $\mathfrak{g}$, and assume that the left action of G on G/K is almost effective. The following properties hold.*

(1) M is of Euclidean type iff $[\mathfrak{m}, \mathfrak{m}] = (0)$. In this case, M has zero sectional curvature.

(2) If M is of compact type, then $\mathfrak{g}$ is semisimple and both G and M are compact.

(3) If M is of noncompact type, then $\mathfrak{g}$ is semisimple and both G and M are noncompact.

Proof. (1) If B is zero on $\mathfrak{m}$, since $B(\mathfrak{m}, \mathfrak{k}) = 0$ by Proposition 22.32, we conclude that $\mathrm{rad}(B) = \mathfrak{m}$ (recall that $\mathrm{rad}(B) = \{X \in \mathfrak{g} \mid B(X, Y) = 0 \text{ for all } Y \in \mathfrak{g}\}$). However, $\mathrm{rad}(B)$ is an ideal in $\mathfrak{g}$, so $[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{m}$, and since $[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{k}$, we deduce that

$$[\mathfrak{m}, \mathfrak{m}] \subseteq \mathfrak{m} \cap \mathfrak{k} = (0).$$

Conversely, assume that $[\mathfrak{m}, \mathfrak{m}] = (0)$. Since B is determined by the quadratic form $Z \mapsto B(Z, Z)$, it suffices to prove that $B(Z, Z) = 0$ for all $Z \in \mathfrak{m}$. Recall that

$$B(Z, Z) = \mathrm{tr}(\mathrm{ad}(Z) \circ \mathrm{ad}(Z)).$$

We have

$$(\mathrm{ad}(Z) \circ \mathrm{ad}(Z))(X) = [Z, [Z, X]]$$

for all $X \in \mathfrak{g}$. If $X \in \mathfrak{m}$, then $[Z, X] = 0$ since $Z, X \in \mathfrak{m}$ and $[\mathfrak{m}, \mathfrak{m}] = (0)$, and if $X \in \mathfrak{k}$, then $[Z, X] \in [\mathfrak{m}, \mathfrak{k}] \subseteq \mathfrak{m}$, so $[Z, [Z, X]] = 0$, since $[Z, [Z, X]] \in [\mathfrak{m}, \mathfrak{m}] = (0)$. Since $\mathfrak{g} = \mathfrak{m} \oplus \mathfrak{k}$, we proved that

$$(\mathrm{ad}(Z) \circ \mathrm{ad}(Z))(X) = 0 \quad \text{for all } X \in \mathfrak{g},$$

and thus $B(Z, Z) = 0$ on $\mathfrak{m}$, as claimed.

For (2) and (3), we use the fact that B is negative definite on $\mathfrak{k}$, by Proposition 22.37.

(2) Since B is negative definite on $\mathfrak{m}$, it is negative definite on $\mathfrak{g}$, and then by Theorem 20.27 we know that G is semisimple and compact. As K is also compact, M is compact.

(3) Since B is positive definite on $\mathfrak{m}$, it is nondegenerate on $\mathfrak{g}$, and then by Theorem 20.26, G is semisimple. In this case, G is not compact since by Theorem 20.27, G is compact iff B is negative definite. As G is noncompact and K is compact, M is noncompact. □

Symmetric spaces of Euclidean type are not that interesting, since they have zero sectional curvature. The Grassmannians $G(k, n)$ and $G^0(k, n)$ are symmetric spaces of compact type, and $\mathbf{SL}(n, \mathbb{R})/\mathbf{SO}(n)$, $\mathcal{H}_n^+(1) = \mathbf{SO}_0(n, 1)/\mathbf{SO}(n)$, and the hyperbolic Grassmannian $G^*(q, p+q) = \mathbf{SO}_0(p, q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$ are of noncompact type.

Since $\mathbf{GL}^+(n, \mathbb{R})$ is not semisimple, $\mathbf{SPD}(n) \cong \mathbf{GL}^+(n, \mathbb{R})/\mathbf{SO}(n)$ is not a symmetric space of noncompact type, but it has many similar properties. For example, it has nonpositive sectional curvature and because it is diffeomorphic to $\mathbf{S}(n) \cong \mathbb{R}^{n(n-1)/2}$, it is simply connected.

Here is a quick summary of the main properties of symmetric spaces of compact and noncompact types. Proofs can be found in O'Neill [91] (Chapter 11) and Ziller [119] (Chapter 6).

Proposition 22.39. *Let $M = (G, K, \sigma)$ be a symmetric space (K compact) with Cartan involution σ and Killing form B on $\mathfrak{g}$, and assume that the left action of G on G/K is almost effective. The following properties hold.*

(1) If M is of compact type, then M has nonnegative sectional curvature and positive Ricci curvature. The fundamental group $\pi_1(M)$ of M is a finite abelian group.

(2) If M is of noncompact type, then M is simply connected, and M has nonpositive sectional curvature and negative Ricci curvature. Furthermore, M is diffeomorphic to $\mathbb{R}^n$ (with $n = \dim(M)$) and G is diffeomorphic to $K \times \mathbb{R}^n$.

There is also an interesting duality between symmetric spaces of compact type and noncompact type, but we will not discuss it here. We refer the reader to O'Neill [91] (Chapter 11), Ziller [119] (Chapter 6), and Helgason [58] (Chapter V, Section 2).

We conclude this section by explaining why the Stiefel manifolds $S(k, n)$ are not symmetric spaces for $2 \leq k \leq n-2$. This has to do with the nature of the involutions of $\mathfrak{so}(n)$. Recall that the matrices $I_{p,q}$ and J_n are defined by

$$I_{p,q} = \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix}, \quad J_n = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix},$$

with $2 \leq p + q$ and $n \geq 1$. Observe that $I_{p,q}^2 = I_{p+q}$ and $J_n^2 = -I_{2n}$. It is shown in Helgason [58] (Chapter X, Section 2 and Section 5) that, up to conjugation, the only involutive automorphisms of $\mathfrak{so}(n)$ are given by

1. $\theta(X) = I_{p,q} X I_{p,q}$, in which case the eigenspace $\mathfrak{k}$ of θ associated with the eigenvalue $+1$ is
$$\mathfrak{k}_1 = \left\{ \begin{pmatrix} S & 0 \\ 0 & T \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(k),\, T \in \mathfrak{so}(n-k) \right\}.$$

2. $\theta(X) = -J_n X J_n$, in which case the eigenspace $\mathfrak{k}$ of θ associated with the eigenvalue $+1$ is
$$\mathfrak{k}_2 = \left\{ \begin{pmatrix} S & -T \\ T & S \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(n),\, T \in \mathbf{S}(n) \right\}.$$

However, in the case of the Stiefel manifold $S(k, n)$, the Lie subalgebra $\mathfrak{k}$ of $\mathfrak{so}(n)$ associated with $\mathbf{SO}(n-k)$ is

$$\mathfrak{k} = \left\{ \begin{pmatrix} 0 & 0 \\ 0 & S \end{pmatrix} \;\middle|\; S \in \mathfrak{so}(n-k) \right\},$$

and if $2 \leq k \leq n-2$, then $\mathfrak{k} \neq \mathfrak{k}_1$ and $\mathfrak{k} \neq \mathfrak{k}_2$. Therefore, the Stiefel manifold $S(k, n)$ is not a symmetric space if $2 \leq k \leq n-2$. This also has to do with the fact that in this case, $\mathbf{SO}(n-k)$ is not a maximal subgroup of $\mathbf{SO}(n)$.

22.11 Problems

Problem 22.1. Assume is G a topological group and M is a topological space. Prove the following claim. If the action $\cdot\colon G \times M \to M$ is proper (where G is Hausdorff), then M is Hausdorff.

Remark: This is the second of clause of Proposition 22.5.

Problem 22.2. In the proof of Theorem 22.14, we claim that (g_x) is a smooth family of inner products on M. Prove this fact.

Hint. See Gallot, Hulin, Lafontaine [49] (Chapter 2, Proposition 2.28).

Problem 22.3. Prove that for any matrix

$$X = \begin{pmatrix} 0 & -u^\top \\ u & 0 \end{pmatrix},$$

where $u \in \mathbb{R}^n$ (a column vector), we have

$$e^{tX} = \begin{pmatrix} \cos(\|u\|\, t) & -\sin(\|u\|\, t)\frac{u^\top}{\|u\|} \\ \sin(\|u\|\, t)\frac{u}{\|u\|} & I + (\cos(\|u\|\, t) - 1)\frac{uu^\top}{\|u\|^2} \end{pmatrix}.$$

Problem 22.4. Let E be a real vector space of dimension $n \geq 1$, and let $\langle -, - \rangle_1$ and $\langle -, - \rangle_2$ be two inner products on E. Let $\varphi_k\colon E \to E^*$ be the linear map given by

$$\varphi_k(u)(v) = \langle u, v\rangle_k, \quad u, v \in E, k = 1, 2.$$

(1) Prove that if $(u_1, \ldots, u_n)$ is an orthonormal basis for $(E, \langle -, - \rangle_1)$, then

$$\varphi_1(u_i) = u_i^*, \quad i = 1, \ldots, n,$$

where $(u_1^*, \ldots, u_n^*)$ is the dual basis in E^* of $(u_1, \ldots, u_n)$ (recall that $u_i^*(u_j) = \delta_{ij}$).

Prove that for any basis $(u_1, \ldots, u_n)$ in E and its dual basis $(u_1^*, \ldots, u_n^*)$ in E^*, the matrix A_k representing φ_k ($k = 1, 2$) is given by

$$(A_k)_{ij} = \varphi_k(u_j)(u_i) = \langle u_j, u_i\rangle_k, \quad 1 \leq i, j, \leq n$$

Conclude that A_k is symmetric positive definite ($k = 1, 2$).

(2) Consider the linear map $f\colon E \to E$ defined by

$$f = \varphi_1^{-1} \circ \varphi_2.$$

Check that

$$\langle u, v\rangle_2 = \langle f(u), v\rangle_1, \quad \text{for all } u, v \in E,$$

and deduce from the above that f is self-adjoint with respect to $\langle -, -\rangle_1$.

(3) Prove that there is some orthonormal basis $(u_1, \ldots, u_n)$ for $(E, \langle -, -\rangle_1)$ which is also an orthogonal basis for $(E, \langle -, -\rangle_2)$. Prove that this result still holds if $\langle -, -\rangle_1$ is an inner product and $\langle -, -\rangle_2$ is any symmetric bilinear form. We say that $\langle -, -\rangle_2$ is *diagonalized* by $\langle -, -\rangle_1$.

Hint. Use the spectral theorem for symmetric matrices.

Assume that $\langle -, -\rangle_1$ is a symmetric, nondegenerate, bilinear form and that $\langle -, -\rangle_2$ is any symmetric bilinear form. Prove that for any basis $(e_1, \ldots, e_n)$ of E, if $(e_1, \ldots, e_n)$ is orthogonal for $\langle -, -\rangle_1$ implies that it is also orthogonal for $\langle -, -\rangle_2$, which means that

$$\text{if} \quad \langle e_j, e_j\rangle_1 = 0 \quad \text{then} \quad \langle e_j, e_j\rangle_2 = 0, \quad \text{for all } i \neq j,$$

then $f = \varphi_1^{-1} \circ \varphi_2$ has $(e_1, \ldots, e_n)$ as a basis of eigenvectors.

Find an example of two symmetric, nondegenerate bilinear forms that do not admit a common orthogonal basis.

(4) Given a group G and a real finite dimensional vector space E, a *representation* of G is any homomorphism $\rho\colon G \to \mathbf{GL}(E)$. A subspace $U \subseteq E$ is *invariant* under ρ if for every $g \in G$, we have $\rho(g)(u) \in U$ for all $u \in U$. A representation is said to be *irreducible* if its only invariant subspaces are (0) and E.

For any two inner products $\langle -, -\rangle_1$ and $\langle -, -\rangle_2$ on E, if $\rho(g)$ is an isometry for both $\langle -, -\rangle_1$ and $\langle -, -\rangle_2$ for all $g \in G$ (which means that $\langle \rho(g)(u), \rho(g)(v)\rangle_k = \langle u, v\rangle_k$ for all $u, v \in E$, $k = 1, 2$) and if ρ is irreducible, then prove that $\langle -, -\rangle_2 = \lambda\langle -, -\rangle_1$, for some nonzero $\lambda \in \mathbb{R}$.

Hint. Compare $\rho(g) \circ f$ and $f \circ \rho(g)$ and show that the eigenspaces of f (as defined in (2)) are invariant under each $\rho(g)$.

In the situation of Proposition 22.21 where we have a homogeneous reductive space G/H with reductive decomposition $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{m}$, prove that if the representation $\mathrm{Ad}^{\mathrm{G}}\colon H \to \mathbf{GL}(\mathfrak{m})$ is irreducible (where Ad_h is restricted to $\mathfrak{m}$ for all $h \in H$), then any two $\mathrm{Ad}(H)$-invariant inner products on $\mathfrak{m}$ are proportional to each other.

Problem 22.5. Consider the grassmannian $G(k, n)$ viewed as a symmetric space as in Section 22.9 (1).

(1) For any X and Y in $\mathfrak{m}$ given by

$$X = \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & -B^\top \\ B & 0 \end{pmatrix},$$

prove that

$$\langle [[X, Y], X], Y\rangle = \langle BA^\top - AB^\top, BA^\top - AB^\top\rangle + \langle A^\top B - B^\top A, A^\top B - B^\top A\rangle,$$

Show that the isotropy representation is given by

$$\mathrm{Ad}((Q,R))A = \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix} \begin{pmatrix} 0 & -A^\top \\ A & 0 \end{pmatrix} \begin{pmatrix} Q^\top & 0 \\ 0 & R^\top \end{pmatrix} = \begin{pmatrix} 0 & -QA^\top R^\top \\ RAQ^\top & 0 \end{pmatrix} = RAQ^\top,$$

where (Q,R) represents an element of $S(\mathbf{O}(k) \times \mathbf{O}(n-k))$, and A represents an element of $\mathfrak{m}$.

Problem 22.6. Consider the space $\mathbf{SPD}(n)$ viewed as a symmetric space as in Section 22.9 (2).

(1) Prove that for all $X, Y \in \mathfrak{m} = \mathbf{S}(n)$ we have

$$\langle [[X,Y],X],Y \rangle = -\mathrm{tr}([X,Y]^\top [X,Y]), \quad \text{for all } X, Y \in \mathfrak{m}.$$

(2) Prove that the isotropy representation is given by

$$\chi_A(X) = AXA^{-1} = AXA^\top,$$

for all $A \in \mathbf{SO}(n)$ and all $X \in \mathfrak{m}$.

Problem 22.7. Consider the hyperbolic space $\mathcal{H}_n^+(1)$ viewed as a symmetric space as in Section 22.9 (3).

(1) Given

$$X = \begin{pmatrix} 0 & u \\ u^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & v \\ v^\top & 0 \end{pmatrix},$$

prove that

$$\langle [[X,Y],X],Y \rangle = -\langle uv^\top - vu^\top, uv^\top - vu^\top \rangle,$$

(2) Prove that for $n \geq 2$, the Killing form B on $\mathfrak{so}(n,1)$ is given by

$$B(X,Y) = (n-1)\mathrm{tr}(XY),$$

for all $X, Y \in \mathfrak{so}(n,1)$.

Problem 22.8. Let $G^*(q, p+q)$ be the set of q-dimensional subspaces W of $R^n = R^{p+q}$ such that $\Phi_{p,q}$ is negative definite on W. Then we have an obvious matrix multiplication action of $\mathbf{SO}_0(p,q)$ on $G^*(q,p+q)$. Check that this action is transitive.

Show that the stabilizer of the subspace spanned by the last q columns of the $(p+q) \times (p+q)$ identity matrix is $\mathbf{SO}(p) \times \mathbf{SO}(q)$, and deduce that the space $G^*(q,p+q)$ is isomorphic to the homogeneous (symmetric) space $\mathbf{SO}_0(p,q)/(\mathbf{SO}(p) \times \mathbf{SO}(q))$.

Problem 22.9. Consider the hyperbolic Grassmannian $G^*(q, p+q)$ viewed as a symmetric space as in Section 22.9 (4).

For $X, Y \in \mathfrak{m}$ given by

$$X = \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 0 & B \\ B^\top & 0 \end{pmatrix},$$

prove that

$$\langle [[X,Y],X],Y\rangle = -\langle BA^\top - AB^\top, BA^\top - AB^\top\rangle - \langle A^\top B - B^\top A, A^\top B - B^\top A\rangle.$$

Problem 22.10. If $G = H \times H$ and $K = \{(h,h) \mid h \in H\} \subseteq G$, prove that the map

$$(h_1, h_2)K \mapsto h_1 h_2^{-1}$$

is a diffeomorphism between the coset space G/K and H.

Problem 22.11. Let (G, K, σ) (G connected and K compact) present a symmetric space. Prove that the ineffective kernel N is the largest normal subgroup of both K and G.

Problem 22.12. Consider the action $\cdot\colon \mathbf{SU}(n+1) \times \mathbb{CP}^n \to \mathbb{CP}^n$ of $\mathbf{SU}(n+1)$ on the (complex) projective space $\mathbb{CP}^n$ discussed in Example (e) of Section 4.3. Prove that the ineffective kernel of the above action is the finite group

$$N = \{\lambda I_{n+1} \mid \lambda^{n+1} = 1,\ \lambda \in \mathbb{C}\}.$$

Bibliography

[1] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. Addison Wesley, second edition, 1978.

[2] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, first edition, 2008.

[3] J. Frank Adams. *Lectures on Lie Groups*. The University of Chicago Press, first edition, 1969.

[4] Tom Apostol. *Analysis*. Addison Wesley, second edition, 1974.

[5] M. A. Armstrong. *Basic Topology*. Undergraduate Texts in Mathematics. Springer-Verlag, first edition, 1983.

[6] Vincent Arsigny. *Processing Data in Lie Groups: An Algebraic Approach. Application to Non-Linear Registration and Diffusion Tensor MRI*. PhD thesis, École Polytechnique, Palaiseau, France, 2006. Thèse de Sciences.

[7] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A fast and log-euclidean polyaffine framework for locally affine registration. Technical report, INRIA, 2004, route des Lucioles, 06902 Sophia Antipolis Cedex, France, 2006. Report No. 5865.

[8] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. on Matrix Analysis and Applications*, 29(1):328–347, 2007.

[9] Vincent Arsigny, Xavier Pennec, and Nicholas Ayache. Polyrigid and polyaffine transformations: a novel geometrical tool to deal with non-rigid deformations–application to the registration of histological slices. *Medical Image Analysis*, 9(6):507–523, 2005.

[10] Michael Artin. *Algebra*. Prentice Hall, first edition, 1991.

[11] Andreas Arvanitoyeorgos. *An Introduction to Lie Groups and the Geometry of Homogeneous Spaces*. SML, Vol. 22. AMS, first edition, 2003.

[12] Andrew Baker. *Matrix Groups. An Introduction to Lie Group Theory.* SUMS. Springer, 2002.

[13] Marcel Berger. *Géométrie 1.* Nathan, 1990. English edition: Geometry 1, Universitext, Springer Verlag.

[14] Marcel Berger. *A Panoramic View of Riemannian Geometry.* Springer, 2003.

[15] Marcel Berger and Bernard Gostiaux. *Géométrie différentielle: variétés, courbes et surfaces.* Collection Mathématiques. Puf, second edition, 1992. English edition: Differential geometry, manifolds, curves, and surfaces, GTM No. 115, Springer Verlag.

[16] William M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, second edition, 1986.

[17] Armand Borel. *Essays in the History of Lie Groups and Algebraic Groups*, volume 21 of *History of Mathematics.* AMS, first edition, 2001.

[18] Raoul Bott and Loring W. Tu. *Differential Forms in Algebraic Topology.* GTM No. 82. Springer Verlag, first edition, 1986.

[19] Nicolas Bourbaki. *Elements of Mathematics. Lie Groups and Lie Algebras, Chapters 1–3.* Springer, first edition, 1989.

[20] Nicolas Bourbaki. *Topologie Générale, Chapitres 1-4.* Eléments de Mathématiques. Masson, 1990.

[21] Nicolas Bourbaki. *Topologie Générale, Chapitres 5-10.* Eléments de Mathématiques. CCLS, 1990.

[22] Nicolas Bourbaki. *Elements of Mathematics. Lie Groups and Lie Algebras, Chapters 4–6.* Springer, first edition, 2002.

[23] Nicolas Bourbaki. *Elements of Mathematics. Lie Groups and Lie Algebras, Chapters 7–9.* Springer, first edition, 2005.

[24] T. Bröcker and T. tom Dieck. *Representations of Compact Lie Groups.* GTM, Vol. 98. Springer Verlag, first edition, 1985.

[25] R.L. Bryant. An introduction to Lie groups and symplectic geometry. In D.S. Freed and K.K. Uhlenbeck, editors, *Geometry and Quantum Field Theory*, pages 5–181. AMS, Providence, Rhode Island, 1995.

[26] N. Burgoyne and R. Cushman. Conjugacy classes in linear groups. *Journal of Algebra*, 44:339–362, 1977.

[27] Henri Cartan. *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes.* Hermann, 1961.

[28] Henri Cartan. *Cours de Calcul Différentiel.* Collection Méthodes. Hermann, 1990.

[29] Roger Carter, Graeme Segal, and Ian Macdonald. *Lectures on Lie Groups and Lie Algebras.* Cambridge University Press, first edition, 1995.

[30] Sheung H. Cheng, Nicholas J. Higham, Charles Kenney, and Alan J. Laub. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22:1112–1125, 2001.

[31] Claude Chevalley. *Theory of Lie Groups I.* Princeton Mathematical Series, No. 8. Princeton University Press, first edition, 1946. Eighth printing.

[32] Yvonne Choquet-Bruhat, Cécile DeWitt-Morette, and Margaret Dillard-Bleick. *Analysis, Manifolds, and Physics, Part I: Basics.* North-Holland, first edition, 1982.

[33] Lawrence Conlon. *Differentiable Manifolds.* Birkhäuser, second edition, 2001.

[34] Morton L. Curtis. *Matrix Groups.* Universitext. Springer Verlag, second edition, 1984.

[35] C. R. DePrima and C. R. Johnson. The range of $A^{-1}A^*$ in $\mathbf{GL}(n, \mathbf{C})$. *Linear Algebra and Its Applications*, 9:209–222, 1974.

[36] Jean Dieudonné. *Éléments d'Analyse, Tome IV. Chapitres XVIII à XX.* Edition Jacques Gabay, first edition, 2007.

[37] Dragomir Djokovic. On the exponential map in classical lie groups. *Journal of Algebra*, 64:76–88, 1980.

[38] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces.* Prentice Hall, 1976.

[39] Manfredo P. do Carmo. *Riemannian Geometry.* Birkhäuser, second edition, 1992.

[40] Norbert Dragon. *The Geometry of Special Relativity: A Concise Course.* SpringerBriefs in Physics. Springer, first edition, 2012.

[41] B.A. Dubrovin, A.T. Fomenko, and S.P. Novikov. *Modern Geometry–Methods and Applications. Part I.* GTM No. 93. Springer Verlag, second edition, 1985.

[42] B.A. Dubrovin, A.T. Fomenko, and S.P. Novikov. *Modern Geometry–Methods and Applications. Part II.* GTM No. 104. Springer Verlag, first edition, 1985.

[43] J.J. Duistermaat and J.A.C. Kolk. *Lie Groups.* Universitext. Springer Verlag, first edition, 2000.

[44] Alan Edelman, Thomas A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[45] William Fulton. *Algebraic Topology, A first course.* GTM No. 153. Springer Verlag, first edition, 1995.

[46] William Fulton and Joe Harris. *Representation Theory, A first course.* GTM No. 129. Springer Verlag, first edition, 1991.

[47] Jean H. Gallier. Logarithms and square roots of real matrices. Technical report, University of Pennsylvania, Levine Hall, 3330 Walnut Street, Philadelphia, PA 19104, 2008. Report No. MS-CIS-08-12.

[48] Jean H. Gallier. *Geometric Methods and Applications, For Computer Science and Engineering.* TAM, Vol. 38. Springer, second edition, 2011.

[49] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian Geometry.* Universitext. Springer Verlag, second edition, 1993.

[50] Christopher Michael Geyer. *Catadioptric Projective Geometry: Theory and Applications.* PhD thesis, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104, 2002. Dissertation.

[51] H. Golub, Gene and F. Van Loan, Charles. *Matrix Computations.* The Johns Hopkins University Press, third edition, 1996.

[52] Marvin J. Greenberg and John R. Harper. *Algebraic Topology: A First Course.* Addison Wesley, first edition, 1981.

[53] Cindy M. Grimm. *Modeling Surfaces of Arbitrary Topology Using Manifolds.* PhD thesis, Department of Computer Science, Brown University, Providence, Rhode Island, USA, 1996. Dissertation.

[54] Cindy M. Grimm and John F. Hughes. Modeling surfaces of arbitrary topology using manifolds. In *Proceedings of the 22nd ACM Annual Conference on Computer Graphics and Interactive Techniques (SIGRAPH'95)*, pages 359–368. ACM, August 6-11 1995.

[55] Victor Guillemin and Alan Pollack. *Differential Topology.* Prentice Hall, first edition, 1974.

[56] Brian Hall. *Lie Groups, Lie Algebras, and Representations. An Elementary Introduction.* GTM No. 222. Springer Verlag, first edition, 2003.

[57] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, first edition, 2002.

[58] Sigurdur Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces.* GSM, Vol. 34. AMS, first edition, 2001.

[59] Nicholas J. Higham. The scaling and squaring method of the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26:1179–1193, 2005.

[60] D. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination.* Chelsea Publishing Co., 1952.

[61] Morris W. Hirsch. *Differential Topology.* GTM No. 33. Springer Verlag, first edition, 1976.

[62] Roger Howe. Very basic Lie theory. *American Mathematical Monthly*, 90:600–623, 1983.

[63] James E. Humphreys. *Introduction to Lie Algebras and Representation Theory.* GTM No. 9. Springer Verlag, first edition, 1972.

[64] Jürgen Jost. *Riemannian Geometry and Geometric Analysis.* Universitext. Springer Verlag, fourth edition, 2005.

[65] Charles S. Kenney and Alan J. Laub. Condition estimates for matrix functions. *SIAM Journal on Matrix Analysis and Applications*, 10:191–209, 1989.

[66] A.A. Kirillov. *Lectures on the Orbit Method.* GSM, Vol. 64. AMS, first edition, 2004.

[67] Wilhelm Klingenberg. *Riemannian Geometry.* de Gruyter & Co, second edition, 1995.

[68] Anthony W. Knapp. *Lie Groups Beyond an Introduction.* Progress in Mathematics, Vol. 140. Birkhäuser, second edition, 2002.

[69] Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry, II.* Wiley Classics. Wiley-Interscience, first edition, 1996.

[70] Yvette Kosmann-Schwarzbach. *Groups and Symmetries. From Finite Groups to Lie Groups.* Universitext. Springer Verlag, first edition, 2010.

[71] Wolfgang Kühnel. *Differential Geometry. Curves–Surfaces–Manifolds.* Student Mathematical Library, Vol. 16. AMS, first edition, 2002.

[72] Jacques Lafontaine. *Introduction Aux Variétés Différentielles.* PUG, first edition, 1996.

[73] Serge Lang. *Real and Functional Analysis.* GTM 142. Springer Verlag, third edition, 1996.

[74] Serge Lang. *Undergraduate Analysis.* UTM. Springer Verlag, second edition, 1997.

[75] Serge Lang. *Fundamentals of Differential Geometry.* GTM No. 191. Springer Verlag, first edition, 1999.

[76] John M. Lee. *Introduction to Smooth Manifolds.* GTM No. 218. Springer Verlag, first edition, 2006.

[77] Jerrold E. Marsden and T.S. Ratiu. *Introduction to Mechanics and Symmetry.* TAM, Vol. 17. Springer Verlag, first edition, 1994.

[78] William S. Massey. *Algebraic Topology: An Introduction.* GTM No. 56. Springer Verlag, second edition, 1987.

[79] William S. Massey. *A Basic Course in Algebraic Topology.* GTM No. 127. Springer Verlag, first edition, 1991.

[80] Yukio Matsumoto. *An Introduction to Morse Theory.* Translations of Mathematical Monographs No 208. AMS, first edition, 2002.

[81] John W. Milnor. *Morse Theory.* Annals of Math. Series, No. 51. Princeton University Press, third edition, 1969.

[82] John W. Milnor. On isometries of inner product spaces. *Inventiones Mathematicae*, 8:83–97, 1969.

[83] John W. Milnor. *Topology from the Differentiable Viewpoint.* The University Press of Virginia, second edition, 1969.

[84] John W. Milnor. Curvatures of left invariant metrics on lie groups. *Advances in Mathematics*, 21:293–329, 1976.

[85] John W. Milnor and James D. Stasheff. *Characteristic Classes.* Annals of Math. Series, No. 76. Princeton University Press, first edition, 1974.

[86] R. Mneimné and F. Testard. *Introduction à la Théorie des Groupes de Lie Classiques.* Hermann, first edition, 1997.

[87] Shigeyuki Morita. *Geometry of Differential Forms.* Translations of Mathematical Monographs No 201. AMS, first edition, 2001.

[88] James R. Munkres. *Analysis on Manifolds.* Addison Wesley, 1991.

[89] James R. Munkres. *Topology.* Prentice Hall, second edition, 2000.

[90] Mitsuru Nishikawa. On the exponential map of the group $\mathbf{O}(p, q)_0$. *Memoirs of the Faculty of Science, Kyushu University, Ser. A*, 37:63–69, 1983.

[91] Barrett O'Neill. *Semi-Riemannian Geometry With Applications to Relativity.* Pure and Applies Math., Vol 103. Academic Press, first edition, 1983.

[92] Xavier Pennec. Intrinsic statistics on Riemannian Manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[93] Peter Petersen. *Riemannian Geometry.* GTM No. 171. Springer Verlag, second edition, 2006.

[94] L. Pontryagin. *Topological Groups.* Princeton University Press, first edition, 1939.

[95] L. Pontryagin. *Topological Groups.* Gordon and Breach, second edition, 1960.

[96] M.M. Postnikov. *Geometry VI. Riemannian Geometry.* Encyclopaedia of Mathematical Sciences, Vol. 91. Springer Verlag, first edition, 2001.

[97] Marcel Riesz. *Clifford Numbers and Spinors.* Kluwer Academic Press, first edition, 1993. Edited by E. Folke Bolinder and Pertti Lounesto.

[98] Wulf Rossmann. *Lie Groups. An Introduction Through Linear Groups.* Graduate Texts in Mathematics. Oxford University Press, first edition, 2002.

[99] Arthur A. Sagle and Ralph E. Walde. *Introduction to Lie Groups and Lie Algebras.* Academic Press, first edition, 1973.

[100] Takashi Sakai. *Riemannian Geometry.* Mathematical Monographs No 149. AMS, first edition, 1996.

[101] Hans Samelson. *Notes on Lie Algebras.* Universitext. Springer, second edition, 1990.

[102] D.H. Sattinger and O.L. Weaver. *Lie Groups and Algebras with Applications to Physics, Geometry, and Mechanics.* Applied Math. Science, Vol. 61. Springer Verlag, first edition, 1986.

[103] Laurent Schwartz. *Analyse I. Théorie des Ensembles et Topologie.* Collection Enseignement des Sciences. Hermann, 1991.

[104] Laurent Schwartz. *Analyse II. Calcul Différentiel et Equations Différentielles.* Collection Enseignement des Sciences. Hermann, 1992.

[105] Jean-Pierre Serre. *Lie Algebras and Lie Groups.* Lecture Notes in Mathematics, No. 1500. Springer, second edition, 1992.

[106] Jean-Pierre Serre. *Complex Semisimple Lie Algebras.* Springer Monographs in Mathematics. Springer, first edition, 2000.

[107] Richard W. Sharpe. *Differential Geometry. Cartan's Generalization of Klein's Erlangen Program.* GTM No. 166. Springer Verlag, first edition, 1997.

[108] Marcelo Siqueira, Dianna Xu, and Jean Gallier. Parametric pseudo-manifolds. *Differential Geometry and its Applications*, 30:702–736, 2012.

[109] Norman Steenrod. *The Topology of Fibre Bundles.* Princeton Math. Series, No. 1. Princeton University Press, 1956.

[110] S. Sternberg. *Lectures On Differential Geometry.* AMS Chelsea, second edition, 1983.

[111] Kristopher Tapp. *Matrix Groups for Undergraduates*, volume 29 of *Student Mathematical Library.* AMS, first edition, 2005.

[112] Loring W. Tu. *An Introduction to Manifolds.* Universitext. Springer Verlag, first edition, 2008.

[113] V.S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations.* GTM No. 102. Springer Verlag, first edition, 1984.

[114] Frank Warner. *Foundations of Differentiable Manifolds and Lie Groups.* GTM No. 94. Springer Verlag, first edition, 1983.

[115] André Weil. *Foundations of Algebraic Geometry.* Colloquium Publications, Vol. XXIX. AMS, second edition, 1946.

[116] André Weil. *L'Intégration dans les Groupes Topologiques et ses Applications.* Hermann, second edition, 1979.

[117] R.O. Wells. *Differential Analysis on Complex Manifolds.* GTM No. 65. Springer Verlag, second edition, 1980.

[118] Hermann Weyl. *The Classical Groups. Their Invariants and Representations.* Princeton Mathematical Series, No. 1. Princeton University Press, second edition, 1946.

[119] Wolfgang Ziller. Lie Groups. Representation Theory and Symmetric Spaces. Technical report, University of Pennsylvania, Math. Department, Philadelphia, PA 19104, 2010. Book in Preparation.

Symbol Index

Index