Solving Differential Equations on Manifolds

Ernst Hairer

Université de Genève
Section de mathématiques
2-4 rue du Lièvre, CP 64
CH-1211 Genève 4

June 2011

Acknowledgement. These notes have been distributed during the lecture "Équations différentielles sur des sous-variétés" (2 hours per week) given in the spring term 2011. At several places, the text and figures are taken from one of the monographs by the author.

Contents

Recommended Literature

There are many monographs treating manifolds and submanifolds. Many of them can be found under the numbers 53 and 57 in the mathematics library. Books specially devoted to the numerical treatment of differential equations on manifolds (differential-algebraic equations) are listed under number 65 in the library.

We give here an incomplete list for further reading: the numbers in brackets (e.g. [MA 65/403]) allow one to find the book without computer search.

R. Abraham, J.E. Marsden and T. Ratiu, *Manifolds, Tensor Analysis, and Applications, 2nd edition*, Applied Mathematical Sciences **75**, Springer-Verlag, 1988. [MA 57/266]

V. Arnold, *Equations Différentielles Ordinaires*, Editions Mir (traduction française), Moscou, 1974. [MA 34/102]

U.M. Ascher and L.R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

K.E. Brenan, S.L. Campbell and L.R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. Revised and corrected reprint of the 1989 original*, Classics in Applied Mathematics **14**, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996. [MA 65/294]

E. Eich-Soellner and C. Führer, *Numerical methods in multibody dynamics*, European Consortium for Mathematics in Industry. B.G. Teubner, Stuttgart, 1998.

E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Texte zur Mathematik **88**, Teubner Verlagsgesellschaft, Leipzig, 1986. [MA 65/256]

E. Hairer, C. Lubich and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, Lecture Notes in Mathematics **1409**, Springer Berlin, 1989. [MA 00.04/3 1409]

E. Hairer, C. Lubich and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations, 2nd edition*, Springer Series in Computational Mathematics **31**, Springer Berlin, 2006. [MA 65/448]

E. Hairer, S.P. Nørsett and G. Wanner, *Solving Ordinary Differential Equations I. Nonstiff Problems, 2nd edition*, Springer Series in Computational Mathematics **8**, Springer Berlin, 1993. [MA 65/245]

E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Agebraic Problems, 2nd edition*, Springer Series in Computational Mathematics **14**, Springer Berlin, 1996. [MA 65/245]

E. Hairer and G. Wanner, *Analysis by Its History*, Undergraduate Texts in Mathematics, Springer New York, 1995. [MA 27/256]

P. Kunkel and V. Mehrmann, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Textbooks in Mathematics. European Mathematical Society (EMS), Zürich, 2006. [MA 34/325]

S. Lang, *Introduction to Differentiable Manifolds, 2nd edition*, Universitext, Springer New York, 2002. [MA 57/15]

J.M. Lee, *Introduction to Smooth Manifolds*, Graduate Texts in Mathematics, Springer New York, 2003. [MA 53/302]

Chapter I

Introduction by Examples

Systems of ordinary differential equations in the Euclidean space $\mathbb{R}^n$ are given by

$$\dot{y} = f(y), \tag{0.1}$$

where $f : U \to \mathbb{R}^n$ with an open set $U \subset \mathbb{R}^n$. If f is sufficiently smooth and an initial value $y(0) = y_0$ is prescribed, it is known that the problem has a unique solution $y : (-\alpha, \alpha) \to \mathbb{R}^n$ for some $\alpha > 0$. This solution can be extended until it approaches the border of U.

In the present lecture we are interested in differential equations, where the solution is known to evolve on a submanifold of $\mathbb{R}^n$, and the vector field $f(y)$ is often only defined on this submanifold. We start with presenting a few typical examples, which serve as motivation for the topic. Later, we shall give a precise definition of differential equations on submanifolds, we shall discuss their numerical treatment, and we shall analyze them rigorously.

I.1 Differential equation on a sphere – the rigid body

Let I_1, I_2, I_3 be the principal moments of inertia of a rigid body. The angular momentum vector $y = (y_1, y_2, y_3)^\mathsf{T}$ then satisfies Euler's equations of motion

$$\begin{aligned} \dot{y}_1 &= (I_3^{-1} - I_2^{-1})\, y_3\, y_2 \\ \dot{y}_2 &= (I_1^{-1} - I_3^{-1})\, y_1\, y_3 \\ \dot{y}_3 &= (I_2^{-1} - I_1^{-1})\, y_2\, y_1 \end{aligned} \qquad \text{or} \qquad \begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix} \begin{pmatrix} y_1/I_1 \\ y_2/I_2 \\ y_3/I_3 \end{pmatrix}. \tag{1.1}$$

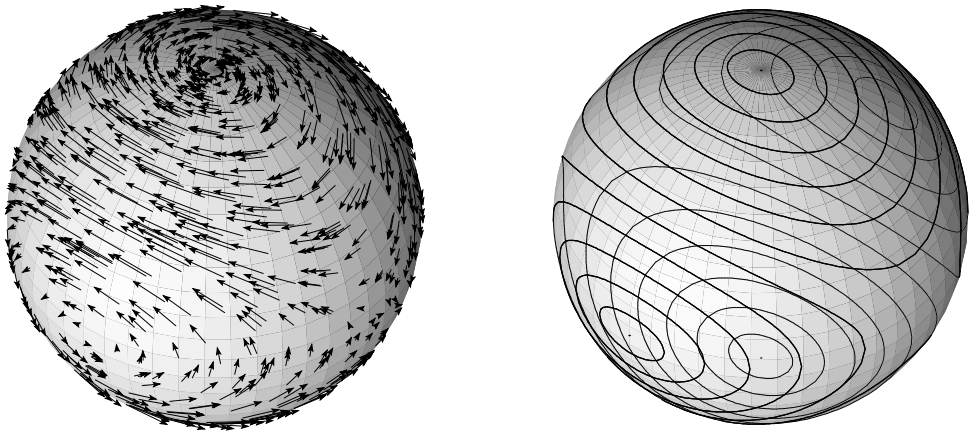


FIG. I.1: Euler's equation of motion for $I_1 = 1.6$, $I_2 = 1$, $I_3 = 2/3$; left picture: vector field on the sphere; right picture: some solution curves.

This differential equation has the property that the function

$$C(y) = \frac{1}{2}\Big(y_1^2 + y_2^2 + y_3^2\Big) \tag{1.2}$$

is exactly preserved along solutions, a property that can be checked by differentiation: $\frac{\mathrm{d}}{\mathrm{d}t} C\big(y(t)\big) = \ldots = 0$. As a consequence, the solution remains forever on the sphere with radius that is determined by the initial values. The left picture of Figure I.1 shows the vector $f(y)$ attached to selected points y of the unit sphere.

To study further properties of the solution we write the differential equation as

$$\dot{y} = B(y)\nabla H(y) \qquad \text{with} \qquad B(y) = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix}, \quad H(y) = \frac{1}{2}\Big(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3}\Big).$$

The function $H(y)$ is called Hamiltonian of the system, whereas $C(y)$ of (1.2) is called Casimir function. Exploiting the skew-symmetry of the matrix $B(y)$, we obtain $\frac{\mathrm{d}}{\mathrm{d}t} H(y(t)) = \nabla H(y(t))^{\mathsf{T}} B(y) \nabla H(y(t)) = 0$, which implies the preservation of the Hamiltonian $H(y)$ along solutions of (1.1). Consequently, solutions lie on the intersection of a sphere $C(y) = \mathit{Const}$ with an ellipsoid $H(y) = \mathit{Const}$, and give rise to the closed curves of the right picture in Figure I.1. Solutions are therefore typically periodic.

Numerical solutions are displayed in Figure I.2. The top picture shows the numerical result, when the explicit Euler method $y_{n+1} = y_n + hf(y_n)$ is applied with step size $h = 0.025$ and with the initial value $y_0 = (\cos(0.9), 0, \sin(0.9))$. The numerical solution drifts away from the manifold. The bottom left picture shows the result of the trapezoidal rule $y_{n+1} = y_n + \frac{h}{2}(f(y_{n+1}) + f(y_n))$ with $h = 1$, where the numerical solution is orthogonally

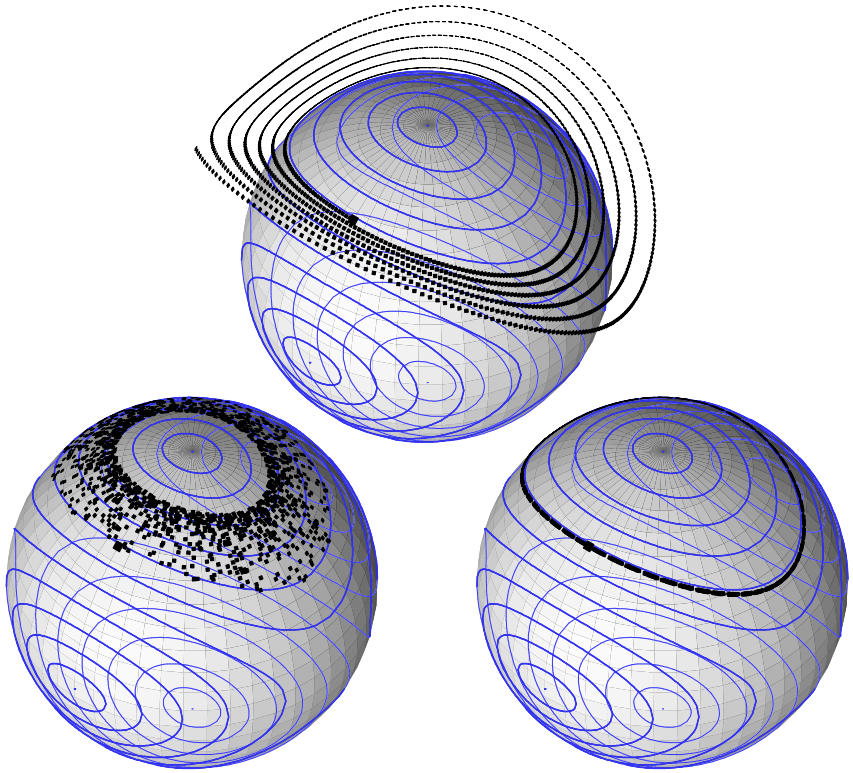


FIG. I.2: Top picture: integration with explicit Euler; bottom left picture: trapezoidal rule with projection onto the sphere; bottom right picture: implicit midpoint rule.

projected onto the sphere after every step. The bottom right picture considers the implicit mid-point rule $y_{n+1} = y_n + hf(\frac{1}{2}(y_{n+1} + y_n))$ with $h = 1$. Even without any projection, the solution agrees extremely well with the exact solution. All these behaviours will be explained in later chapters.

I.2 Problems in control theory

In control theory one often encounters problems of the form

$$\begin{aligned} \dot{y} &= f(y, u) \\ 0 &= g(y), \end{aligned} \tag{2.1}$$

where $u(t)$ is a control function that permits to steer the motion $y(t)$ of a mechanical system. Differentiating the algebraic equation $g(y(t)) = 0$ with respect to time yields $g'(y)f(y, u) = 0$. Under suitable regularity assumptions this relation permits us to express u as a function of y (using the implicit function theorem). Inserting $u = G(y)$ into (2.1) gives a differential equation for y on the manifold $\mathcal{M} = \{y\,;\, g(y) = 0\}$.

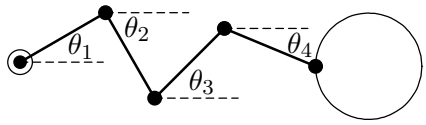


FIG. I.3: Sketch of an articulated robot arm.

Example 2.1 (articulated robot arm)**.** Consider $n > 2$ segments of fixed length 1 that are connected with joints as illustrated in Figure I.3. We assume that the starting point of the first segment is fixed at the origin. Denoting by θ_j the angle of the jth segment with respect to the horizontal axis, the endpoint of the last segment is given by $g(\theta)$, where for $\theta = (\theta_1, \ldots, \theta_n)$ we have

$$g(\theta) = \begin{pmatrix} \cos\theta_1 + \cos\theta_2 + \ldots + \cos\theta_n \\ \sin\theta_1 + \sin\theta_2 + \ldots + \sin\theta_n \end{pmatrix}. \tag{2.2}$$

The problem consists in finding the motion $\theta(t)$ of the articulated robot arm such that the endpoint of the last segment follows a given parametrized curve $\gamma(t)$ in the plane and

$$\|\dot{\theta}(t)\| \to \min \qquad \text{subject to} \qquad g(\theta(t)) = \gamma(t).$$

Differentiating the algebraic relation with respect to time yields the underdetermined linear equation $g'(\theta(t))\dot{\theta}(t) = \dot{\gamma}(t)$ for $\dot{\theta}(t)$ (two linear equations for $n > 2$ unknowns). Among all solutions of this linear system, the Euclidean norm of $\dot{\theta}(t)$ is minimized when this vector is perpendicular to $\ker g'(\theta(t))$. Because of $(\ker g'(\theta))^{\perp} = \operatorname{Im} g'(\theta)^{\mathsf{T}}$, this leads to the problem

$$\dot{\theta} = g'(\theta)^{\mathsf{T}} u, \qquad g(\theta) = \gamma(t), \tag{2.3}$$

which is of the form (2.1), if we add the trivial equation $\dot{t} = 1$ to the system, and interpret $y = (\theta, t)$. This is a differential equation on the manifold $\mathcal{M} = \{(\theta, t)\ ;\ g(\theta) - \gamma(t) = 0\}$.

The differentiated constraint yields $g'(\theta)g'(\theta)^{\mathsf{T}} u = \dot{\gamma}(t)$, which permits to express u in terms of (θ, t) as long as the Jacobian matrix $g'(\theta)$ has full rank 2. In this case we get

$$\dot{\theta} = g'(\theta)^{\mathsf{T}} \big(g'(\theta)g'(\theta)^{\mathsf{T}}\big)^{-1} \dot{\gamma}(t),$$

a differential equation that can be solved numerically by standard approaches. As in the previous example, care has to be taken to avoid a drift from the manifold $\mathcal{M}$.

I.3 Constrained mechanical systems

A rich treasure trove of differential equations on manifolds are constrained mechanical systems (or multi-body systems). Let $q = (q_1, \dots, q_n)^\mathsf{T}$ be generalized coordinates of a conservative mechanical system with kinetic energy $T(\dot q) = \frac{1}{2}\,\dot q^\mathsf{T} M\,\dot q$ (symmetric positive definite mass matrix M) and potential energy $U(q)$, which is subject to holonomic constraints $g(q) = 0$ (here, $g : \mathbb{R}^n \to \mathbb{R}^m$ with $m < n$). The equations of motion are then given by

$$\begin{aligned} \dot q &= v \\ M\,\dot v &= -\nabla U(q) - g'(q)^\mathsf{T}\lambda \\ 0 &= g(q). \end{aligned} \tag{3.1}$$

To find a differential equation on a submanifold we differentiate the algebraic constraint to obtain $g'(q)v = 0$. A second differentiation yields

$$g''(q)(v,v) - g'(q)M^{-1}(\nabla U(q) + g'(q)^\mathsf{T}\lambda) = 0,$$

which permits to express λ in terms of (q, v) provided that $g'(q)$ is of full rank m (Exercise 6). Inserted into (3.1) we obtain a differential equations for (q, v) on the submanifold

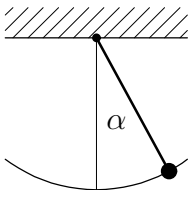
$$\mathcal{M} = \{(q,v)\,;\, g(q) = 0,\, g'(q)v = 0\}.$$

Example 3.1 (mathematical pendulum)**.** Consider a weight on the end of a massless cord suspended from a pivot, without friction. We let the pivot be at the origin and denote by $q = (q_1, q_2)^\mathsf{T}$ the Cartesian coordinates of the weight. Assuming unit mass, unit gravity constant, and unit length of the cord we have $T(\dot q) = \frac{1}{2}\dot q^\mathsf{T}\dot q$, $U(q) = q_2$, and constraint $g(q) = q^\mathsf{T}q - 1$. The equations of motion are therefore

$$\begin{aligned} \dot q_1 &= v_1, & \dot v_1 &= -\lambda q_1, \\ \dot q_2 &= v_2, & \dot v_2 &= -1 - \lambda q_2, \end{aligned} \qquad 0 = q_1^2 + q_2^2 - 1, \tag{3.2}$$

which represent a differential equation on the submanifold

$$\mathcal{M} = \{(q_1, q_2, v_1, v_2)\,;\, q_1^2 + q_2^2 = 1,\, q_1 v_1 + q_2 v_2 = 0\}.$$



We have presented this simple example to become familiar with multi-body systems. It should not be misleading, because a much simpler formulation is possible in this case by the use of polar coordinates $q_1 = \sin\alpha$, $q_2 = -\cos\alpha$. A short computation shows that the system (3.2) is indeed equivalent to the familiar equation

$$\ddot\alpha + \sin\alpha = 0.$$

Remark. In general it is not possible to determine minimal coordinates (where the number of coordinates equals the number of degrees of freedom of the mechanical system). Even if it is possible, they are usually only locally defined and the differential equations become much more complicated as the formulation (3.1). Our next example illustrates such a situation and shows the importance of considering differential equations on manifolds.

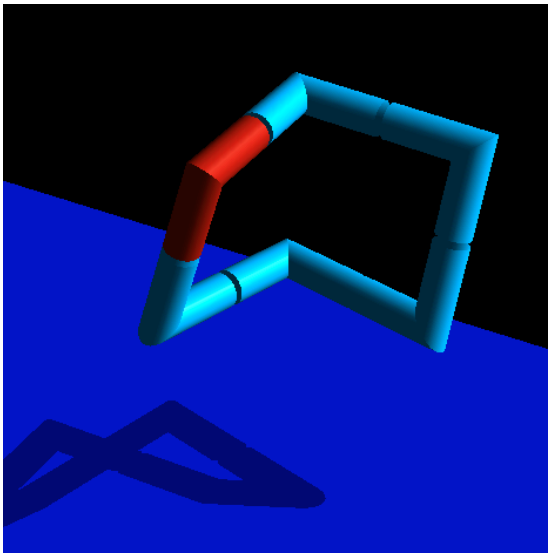


FIG. I.4: Multi-body system; graphics by J.-P. Eckmann & M. Hairer.

Example 3.2. We consider a mechanical system, where six rigid corner pieces are joined together to form a ring (see the illustration of Figure I.4). At the contact of any two pieces the only degree of freedom is rotation around their common axis. For a mathematical formulation we denote the position of the corners of the six pieces by $q_i = (q_{i1}, q_{i2}, q_{i3})^\mathsf{T} \in \mathbb{R}^3$, $i = 1, \dots, 6$, which constitute $6 \times 3 = 18$ variables. It is convenient to use in addition the notation $q_0 = q_6$. Let us start with the description of the constraints:

(1) the motion of one piece (red) is prescribed, i.e., q_0, q_1, q_2 are given functions of time that satisfy $\|q_1 - q_0\|_2 = \|q_2 - q_1\|_2 = 1$ and $(q_1 - q_0) \perp (q_2 - q_1)$ (9 conditions),

(2) distance between neighbor corners is unity (4 additional conditions),

(3) orthogonality between neighbor edges $(q_{n-1} - q_n) \perp (q_{n+1} - q_n)$ (5 additional conditions).

These conditions define the constraint $g(q) = \gamma(t)$, where $g : \mathbb{R}^{18} \to \mathbb{R}^{18}$ is given by

$$\begin{aligned} g_{3i+j}(q) &= q_{ij} && \text{for } \ i = 0, 1, 2, \ j = 1, 2, 3, \\ g_{8+j}(q) &= \|q_{j+1} - q_j\|_2^2 - 1 && \text{for } \ j = 2, 3, 4, 5, \\ g_{12+j}(q) &= (q_{j+1} - q_j)^\mathsf{T}(q_j - q_{j-1}) && \text{for } \ j = 2, 3, 4, 5, 6, \end{aligned} \tag{3.3}$$

and

$$\gamma_k(t) = \begin{cases} q_{ij}(t) & \text{for } \ k = 3(i-1) + j, \ i = 1, 2, 3, \ j = 1, 2, 3, \\ 0 & \text{else.} \end{cases}$$

The constraint condition $g(q) = \gamma(t)$ represents 18 (linear and quadratic) equations for 18 unknowns $q = (q_{11}, q_{12}, q_{13}, q_{21}, q_{22}, q_{23}, \dots, q_{63})^\mathsf{T}$. For a consistent vector $\gamma(t)$, this nonlinear equation possesses as solution a discrete point and a one-dimensional closed curve in $\mathbb{R}^{18}$ (without proof, see also Exercise 7). To get a nontrivial dynamics we assume that the initial value lies on the one-dimensional curve.

To complete the description of the problem, we assume that the mass of the pieces is unity and concentrated in their corner, and the motion is without friction. The kinetic and potential energies are then given by

$$T(\dot q) = \frac{1}{2} \sum_{i=1}^{6} \dot q_i^\mathsf{T} \dot q_i, \qquad U(q) = \sum_{i=1}^{6} q_{i3},$$

where the potential only takes gravity into account. The equations of motion are obtained by (3.1) with the constraint replaced by $g(q) - \gamma(t) = 0$.

Remark. The fact that the equation $g(q) = \gamma(t)$ admits a one-dimensional submanifold as solution shows that the 18 equations are not independent. For a numerical treatment we can remove one (carefully chosen) constraint and work with the remaining 17 constraints. This can be done with the help of a QR decomposition of $g'(q)$ which is anyway required during the integration.

I.4 Exercises

1. Compute all stationary solutions of the system (1.1) and identify them in Figure I.1. Explain the behaviour of the solutions close to these points.
2. If the principal moments of inertia satisfy $I_1 = I_2 \neq I_3$, the rigid body is called a *symmetrical top*. In this situation, solve analytically Euler's equations of motion (1.1).
3. For a vector $\theta = (\theta_1, \ldots, \theta_n)$ of angles consider the function $g(\theta)$ of (2.2). Prove that $g'(\theta)$ is of full rank 2 if and only if there exists a pair of subscripts i, j such that
$$\theta_i - \theta_j \neq 0 \mod \pi.$$
4. Consider the problem (2.3) and assume that initial values satisfy $\theta_i(t_0) = \theta_j(t_0)$. Prove that the solution then satisfies $\theta_i(t) = \theta_j(t)$ wherever it exists.
5. Find a differential equation (on a submanifold) that describes the solution of the problem
$$\dot\theta_1^2 + (\dot\theta_2 - \dot\theta_1)^2 + \ldots + (\dot\theta_n - \dot\theta_{n-1})^2 \to \min$$
subject to the constraint $g(\theta) - \gamma(t) = 0$, where $g(\theta)$ is as in (2.2).
6. Consider a $n \times n$ matrix M and a $m \times n$ matrix G (with $m < n$). Under the assumptions that M is a symmetric positive definite matrix and G is of full rank m, prove that the matrices
$$\begin{pmatrix} M & G^\mathsf{T} \\ G & 0 \end{pmatrix} \qquad \text{and} \qquad G\,M^{-1}G^\mathsf{T}$$
are invertible.
7. Consider the function $g : \mathbb{R}^{18} \to \mathbb{R}^{18}$ defined in (3.3), and compute the Jacobian matrix $g'(q)$ for the two (admissible) points
$$\begin{aligned} a &= (1,0,0;0,0,0;0,1,0;0,1,1;0,0,1;1,0,1) \\ b &= (1,0,0;0,0,0;0,1,0;0,1,1;1,1,1;1,0,1). \end{aligned}$$
Prove that $g'(a)$ is invertible, but $g'(b)$ is singular and of rank 17.

Chapter II

Submanifolds of $\mathbb{R}^n$

The Euclidean space $\mathbb{R}^n$ is a differentiable manifold. In this chapter we give a short introduction to submanifolds of $\mathbb{R}^n$. Our emphasis is on characterizations that are suitable for numerical computations. We further discuss the tangent space, differentiable mappings, and differential equations on submanifolds.

II.1 Definition and characterization of submanifolds

Submanifolds of $\mathbb{R}^n$ are nonlinear analogues of linear subspaces. They extend the notion of curves and surfaces. In the following a diffeomorphism $\varphi : U \to V$ between open sets is a continuously differentiable mapping having a continuously differentiable inverse.

Definition 1.1 (submanifold)**.** *A set $\mathcal{M} \subset \mathbb{R}^n$ is a submanifold of $\mathbb{R}^n$ if for every $a \in \mathcal{M}$ there exist open sets $U, V \subset \mathbb{R}^n$ with $a \in U$ and a diffeomorphism $\varphi : U \to V$ such that*

$$\varphi(U \cap \mathcal{M}) = \varphi(U) \cap (\mathbb{R}^k \times \{\mathbf{0}\}).$$

The number k is called dimension of $\mathcal{M}$ and $n-k$ is its codimension. A pair (U, φ) is called chart on $\mathcal{M}$, and the union of all charts is called (maximal) atlas.

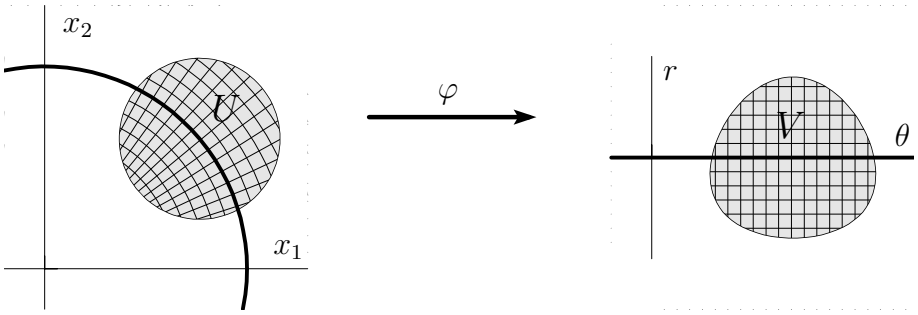


FIG. II.1: Definition of a submanifold of $\mathbb{R}^n$.

Figure II.1 illustrates the circle $\{(x_1, x_2)\,;\, x_1^2 + x_2^2 = 1\}$ as a submanifold of $\mathbb{R}^2$. A possible choice for the diffeomorphism $\varphi(x_1, x_2) = (\theta, r)$ is the mapping defined by polar coordinates $x_1 = (1 + r)\cos\theta$, $x_2 = (1 + r)\sin\theta$.

Submanifolds of dimension $k = 0$ are discrete points in $\mathbb{R}^n$. Submanifolds of maximal dimension $k = n$ are open sets in $\mathbb{R}^n$. Every linear or affine subspace of $\mathbb{R}^n$ is a submanifold. However, the set $\{(x, y)\,;\, xy = 0\}$ is not a submanifold of $\mathbb{R}^2$ because, close to the origin, it is not diffeomorph to a straight line.

Lemma 1.2 (level set representation)**.** *A set $\mathcal{M} \subset \mathbb{R}^n$ is a submanifold of $\mathbb{R}^n$ if and only if for every $a \in \mathcal{M}$ there exist an open set $U \subset \mathbb{R}^n$ with $a \in U$, and a differentiable mapping $g : U \to \mathbb{R}^{n-k}$ with $g(a) = \mathbf{0}$ and $g'(a)$ of maximal rank $n-k$, such that*

$$U \cap \mathcal{M} = g^{-1}(\mathbf{0}).$$

Proof. For the "only if" part it is sufficient to take for $g(x)$ the last $n-k$ components of $\varphi(x)$. For the proof of the "if" part we assume (after a possible permutation of the components of $x \in \mathbb{R}^n$) that the submatrix of $g'(a)$ consisting of the last $n-k$ columns is invertible. The function

$$\varphi(x) = \Big(x_1 \ldots, x_k, g_1(x), \ldots, g_{n-k}(x)\Big)^{\mathsf{T}}$$

is then a local diffeomorphism close to a and satisfies the condition of Definition 1.1. □

Lemma 1.3 (local parametrization)**.** *A set $\mathcal{M} \subset \mathbb{R}^n$ is a submanifold of $\mathbb{R}^n$ if and only if for every $a \in \mathcal{M}$ there exist open sets $a \in U \subset \mathbb{R}^n$ and $W \subset \mathbb{R}^k$, and a continuously differentiable mapping $\eta : W \to U$ with $\eta(\mathbf{0}) = a$ and $\eta'(\mathbf{0})$ of maximal rank k, such that*

$$U \cap \mathcal{M} = \eta(W),$$

and $\eta : W \to U \cap \mathcal{M}$ is a homeomorphism.

Proof. For the "only if" part we put $W = \{z \in \mathbb{R}^k\,;\, (z, \mathbf{0}) \in \varphi(U)\}$ and $\eta(z) = \varphi^{-1}(z, \mathbf{0})$. Here, $(z, \mathbf{0})$ denotes the vector, where z is completed with zeros to a vector in $\mathbb{R}^n$.

To prove the "if" part, we consider $\eta : W \to U$ and we assume (after a possible permutation of the coordinates in the image space) that the submatrix of $\eta'(\mathbf{0})$ consisting of the first k rows is invertible. We then define for $y \in W \times \mathbb{R}^{n-k} \subset \mathbb{R}^n$

$$\psi(y) = \Big(\eta_1(\widehat{y}), \ldots, \eta_k(\widehat{y}), \eta_{k+1}(\widehat{y}) - y_{k+1}, \ldots, \eta_n(\widehat{y}) - y_n\Big)^{\mathsf{T}}$$

where $\widehat{y}$ denotes the vector consisting of the first k components of y. The Jacobian matrix $\psi'(\mathbf{0})$ is invertible, so that ψ is a local diffeomorphism close to $\psi(\mathbf{0}) = a$, i.e., there exist open neighborhoods $U_1 \subset U$ of a and $V \subset W \times \mathbb{R}^{n-k}$ of $\mathbf{0}$, such that $\psi : V \to U_1$ is a diffeomorphism. We now put $\varphi = \psi^{-1} : U_1 \to V$.

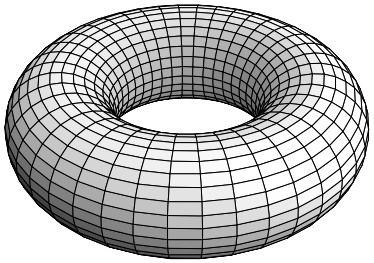
The property $\varphi(U_1 \cap \mathcal{M}) \supset \varphi(U_1) \cap (\mathbb{R}^k \times \{\mathbf{0}\})$ follows immediately from the fact that, for $y \in \varphi(U_1)$ with $y_{k+1} = \ldots = y_n = 0$, we have $\psi(y) = \eta(\widehat{y}) \in U_1 \cap \eta(W) = U_1 \cap \mathcal{M}$. To prove the inverse inclusion, we take $y \in \varphi(U_1 \cap \mathcal{M}) = \varphi(U_1 \cap \eta(W))$ so that $y = \varphi(\eta(z))$ for some $z \in W$ and hence also $\psi(y) = \eta(z)$. If U_1 is chosen as a sufficiently small neighborhood of a, the vectors z and $\widehat{y} = (y_1, \ldots, y_k)^{\mathsf{T}}$ are both close to $\mathbf{0}$ (this follows from the fact that $\eta : W \to U \cap \mathcal{M}$ is a homeomorphism). If we denote by $\widehat{\eta}$ the first k components of the function η, it follows from $\psi(y) = \eta(z)$ that $\widehat{\eta}(\widehat{y}) = \widehat{\eta}(z)$. However, since $\widehat{\eta}'(\mathbf{0})$ is nonsingular, $\widehat{\eta}$ is a local diffeomorphism close to $\mathbf{0}$, and we obtain $\widehat{y} = z$. The relation $\psi(y) = \eta(z) = \eta(\widehat{y})$ thus implies $y_{k+1} = \ldots = y_n = 0$, which completes the proof. □

FIG. II.2: Curves in $\mathbb{R}^2$ which are not submanifolds.

Remark. One cannot remove the assumption "$\eta : W \to U \cap \mathcal{M}$ is a homeomorphism" from the characterization in Lemma 1.3. As counter-example serves the curve $\eta(t) = ((1 + 0.1t^2)\cos t, (1+0.1t^2)\sin t)$ which satisfies all other assumptions of Lemma 1.3, but the image $\eta(\mathbb{R})$ is not a submanifold of $\mathbb{R}^2$ (left picture of Figure II.2). The injectivity of $\eta(t)$ is even not sufficient as shown in the right picture of Figure II.2.

Example 1.4 (torus)**.** Consider the circle $(x, z) = (d + \rho\cos\alpha, \rho\sin\alpha)$ (with $0 < \rho < d$) and rotate it around the z-axis. This gives the parametrization

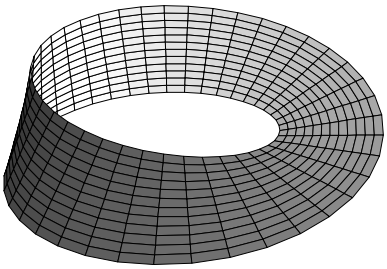
$$\eta(\alpha, \beta) = \begin{pmatrix} (d + \rho\cos\alpha)\cos\beta \\ (d + \rho\cos\alpha)\sin\beta \\ \rho\sin\alpha \end{pmatrix}$$



of a torus. One can check that $\eta'(\alpha, \beta)$ is of maximal rank 2 and that η is locally a homeomorphism.

Example 1.5 (Möbius strip)**.** Consider a segment of lenght 2 (parametrized by $-1 < t < 1$), rotate it around its centre and, at the same time, move this centre twice as fast along a circle of radius d. This gives the parametrization

$$\eta(t, \alpha) = \begin{pmatrix} (d + t\cos\alpha)\cos 2\alpha \\ (d + t\cos\alpha)\sin 2\alpha \\ t\sin\alpha \end{pmatrix}.$$



Example 1.6 (orthogonal group)**.** The set

$$\mathrm{O}(n) = \{X \ ; \ X^T X = I\}$$

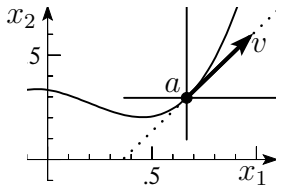
is a submanifold of dimension $n(n-1)/2$ of the space $\mathbb{M}_n(\mathbb{R}) = \mathbb{R}^{n\cdot n}$ of all n-dimensional matrices. For the proof of this statement we consider the mapping

$$g : \mathbb{M}_n(\mathbb{R}) \to Sym_n(\mathbb{R}) \approx \mathbb{R}^{n(n+1)/2}$$

defined by $g(X) = X^T X - I$ (the symbol $Sym_n(\mathbb{R})$ denotes the space of all symmetric matrices of dimension n). This mapping is differentiable and we have $g^{-1}(\mathbf{0}) = \mathrm{O}(n)$. It therefore suffices to prove that $g'(A)$ is of maximal rank for every matrix $A \in \mathrm{O}(n)$. The derivative of $g(X)$ is $g'(A)H = A^T H + H^T A$. For an arbitrary symmetric matrix B, the choice $H = AB/2$ shows that $g'(A)H = B$. Therefore, $g'(A) : \mathbb{M}_n(\mathbb{R}) \to Sym_n(\mathbb{R})$ is surjective (i.e., of maximal rank), and $\mathrm{O}(n)$ is a submanifold of codimension $n(n+1)/2$.

II.2 Tangent space

Curves. For a regular parametric curve $\gamma : I \to \mathbb{R}^n$, the tangent at $a = \gamma(0)$ is the straight line given by $\tau(t) = a + tv$, where $v = \dot\gamma(0) \neq \mathbf{0}$. Shifting the origin to the point a, the tangent in a at the curve $\mathcal{M} = \gamma(I)$ becomes the linear space



$$T_a\mathcal{M} = \{tv \,|\, t \in \mathbb{R}\}.$$

In the original variables, the tangent is the affine space $a + T_a\mathcal{M} \subset \mathbb{R}^n$.

Surfaces in $\mathbb{R}^3$. As an example, consider the ellipsoid

$$\mathcal{M} = \left\{(x,y,z) \;\middle|\; \frac{x^2}{a^2}+\frac{y^2}{b^2}+\frac{z^2}{c^2}=1\right\} \quad \text{with parametrization} \quad \varphi(\alpha,\theta) = \begin{pmatrix} a\cos\alpha\,\sin\theta \\ b\sin\alpha\,\sin\theta \\ c\cos\theta \end{pmatrix}.$$

To determine the tangent plane at $a = (x_0, y_0, z_0) = \varphi(\alpha_0, \theta_0) \in \mathcal{M}$, we consider the parametric curves $\gamma(t) = \varphi(t, \theta_0)$ and $\delta(t) = \varphi(\alpha_0, t)$; see Figure II.3. The left picture shows also the tangents (in grey) $\tau(t) = a + tv_1$ with $v_1 = \dot\gamma(\alpha_0)$ and $\sigma(t) = a + tv_2$ with $v_2 = \dot\delta(\theta_0)$. The vectors v_1 and v_2 span the tangent space. It is given by $a + T_a\mathcal{M}$, where

$$T_a\mathcal{M} = \{t_1v_1 + t_2v_2 \mid t_1, t_2 \in \mathbb{R}\} \qquad \text{with} \qquad v_1 = \frac{\partial\varphi}{\partial\alpha}(\alpha_0,\theta_0), \quad v_2 = \frac{\partial\varphi}{\partial\theta}(\alpha_0,\theta_0).$$

The tangent of other curves lying in $\mathcal{M}$ and passing through a is also in $a + T_a\mathcal{M}$ (see the right picture of Figure II.3).

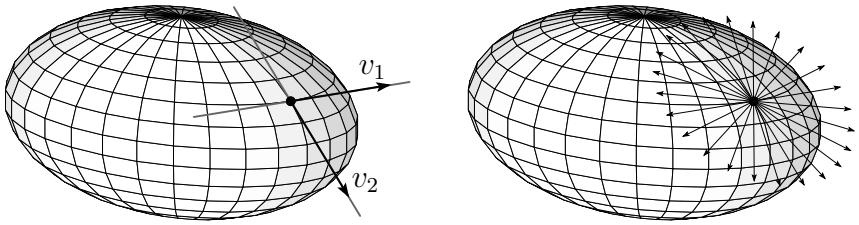


FIG. II.3: Illustration of the definition of the tangent space.

Definition 2.1 (tangent space)**.** *Let $\mathcal{M} \subset \mathbb{R}^n$ be a submanifold of $\mathbb{R}^n$ and let $a \in \mathcal{M}$. The tangent space of $\mathcal{M}$ at a is the linear space given by*

$$T_a\mathcal{M} = \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{array}{c} \text{there exists a continuously differentiable } \gamma : (-\varepsilon,\varepsilon) \to \mathbb{R}^n \text{ such that} \\ \gamma(t) \in \mathcal{M} \text{ for } t \in (-\varepsilon,\varepsilon) \text{ and } \gamma(0) = a \text{ and } \dot\gamma(0) = v \end{array} \right\}.$$

This definition gives a nice geometric interpretation of the tangent space. Algebraic characterizations with explicit formulas are given in the following theorem.

Theorem 2.2. *Consider a submanifold $\mathcal{M} \subset \mathbb{R}^n$ of dimension k and let $a \in \mathcal{M}$.*

- *If, close to a, $\mathcal{M}$ is given by a local parametrization $\eta : W \to \mathbb{R}^n$, i.e., we have $U \cap \mathcal{M} = \{\eta(z) \mid z \in W\}$, where $\eta(z_0) = a$ with $z_0 \in W \subset \mathbb{R}^k$, then*

$$T_a\mathcal{M} = \operatorname{Im}\eta'(z_0) = \{\eta'(z_0)\,t \mid t \in \mathbb{R}^k\}. \tag{2.1}$$

- *If $\mathcal{M}$ is locally given by $U \cap \mathcal{M} = \{x \in U \mid g(x) = \mathbf{0}\}$, then*

$$T_a\mathcal{M} = \ker g'(a) = \{v \in \mathbb{R}^n \mid g'(a)v = \mathbf{0}\}. \tag{2.2}$$

Proof. Let $\delta(s)$ be a curve in $\mathbb{R}^k$ satisfying $\delta(0) = z_0$ and $\dot\delta(0) = t$ (for example the straight line $\delta(s) = z_0 + st$). The curve $\gamma(s) := \eta(\delta(s))$ is then a curve lying in $\mathcal{M}$, and satisfying $\gamma(0) = \eta(z_0) = a$ and $\dot\gamma(0) = \eta'(z_0)\,\dot\delta(0) = \eta'(z_0)\,t$. This implies $\operatorname{Im}\eta'(z_0) \subset T_a\mathcal{M}$.

If $\gamma(t)$ is a curve lying in $\mathcal{M}$ and satisfying $\gamma(0) = a$ and $\dot\gamma(0) = v$, then we have $g(\gamma(t)) = \mathbf{0}$ and hence also $g'(a)\dot\gamma(0) = \mathbf{0}$. This implies $T_a\mathcal{M} \subset \ker g'(a)$.

By definition of the submanifold $\mathcal{M}$, the two linear spaces $\operatorname{Im}\eta'(z_0)$ and $\ker g'(a)$ have the same dimension k. From the inclusions $\operatorname{Im}\eta'(z_0) \subset T_a\mathcal{M} \subset \ker g'(a)$ we therefore deduce the identities $\operatorname{Im}\eta'(z_0) = T_a\mathcal{M} = \ker g'(a)$. □

The definition of differentiability of a function $f : A \to \mathbb{R}^m$ at a point a requires that f is defined in an open neighborhood of $a \in A$. Our aim is to extend the notion of differentiability to functions that are defined on a manifold $\mathcal{M}$ with positive codimension, but not on a neighborhood of it.

Consider a function $f : \mathcal{M} \to \mathcal{N}$ between two submanifolds, and let (U, φ) and (V, ψ) be charts of $\mathcal{M}$ and $\mathcal{N}$, respectively. If $f(U \cap \mathcal{M}) \subset V \cap \mathcal{N}$, we can consider the mapping $f_{\varphi\psi}$ defined by

$$\Big(\psi \circ f \circ \varphi^{-1}\Big)(z, \mathbf{0}) = \Big(f_{\varphi\psi}(z), \mathbf{0}\Big). \tag{2.3}$$

Here, z is a vector of the dimension of $\mathcal{M}$, such that $(z, \mathbf{0}) \in \varphi(U)$. By the definition of a submanifold, $\varphi^{-1}(z, \mathbf{0}) \in U \cap \mathcal{M}$, so that f can be applied to $\varphi^{-1}(z, \mathbf{0})$ and has an image in $V \cap \mathcal{N}$. A final application of ψ yields $\Big(f_{\varphi\psi}(z), \mathbf{0}\Big)$, where $f_{\varphi\psi}(z)$ is a vector with dimension of that of the submanifold $\mathcal{N}$.

Definition 2.3. *A function $f : \mathcal{M} \to \mathcal{N}$ is differentiable at $a \in \mathcal{M}$, if there exist a chart (U, φ) of $\mathcal{M}$ with $a \in U$ and a chart (V, ψ) of $\mathcal{N}$ with $f(a) \in V$, such that the function $f_{\varphi\psi}(z)$ of (2.3) is differentiable at z_0 given by $\varphi(a) = (z_0, \mathbf{0})$.*

If this property is satisfied for all $a \in \mathcal{M}$ and if $f_{\varphi\psi}(z)$ is continuously differentiable, then the function f is called continuously differentiable (or of class $\mathcal{C}^1$).

This definition is meaningful, because $f_{\varphi\psi}(z)$ is well-defined in a neighborhood of z_0. Moreover, it is independent of the choice of charts, because for two charts (U_i, φ_i) the function $\varphi_1^{-1} \circ \varphi_2$ is differentiable, where it is defined. We remark that due to the fact that $\mathcal{N}$ is a submanifold of $\mathbb{R}^n$, an equivalent definition would be to require that $(f \circ \varphi^{-1})(z, \mathbf{0})$ is differentiable at z_0. In the following we denote $f_\varphi(z) := (f \circ \varphi^{-1})(z, \mathbf{0})$.

Next we give a meaning to the derivative of a $\mathcal{C}^1$-function $f : \mathcal{M} \to \mathcal{N}$. We consider a continuously differentiable curve $\gamma : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ with $\gamma(t) \in \mathcal{M}$, $\gamma(0) = a$, and $\dot\gamma(0) = v$, so that $v \in T_a\mathcal{M}$. The curve $\delta(t) := f(\gamma(t))$ then satisfies $\delta(t) \in \mathcal{N}$, $\delta(0) = f(a)$, and it is continuously differentiable, because it can be written as $\delta = (f \circ \varphi^{-1}) \circ (\varphi \circ \gamma) = f_\varphi \circ (\varphi \circ \gamma)$. Its derivative at the origin is $\dot\delta(0) = f'_\varphi(\varphi(a))\varphi'(a)\dot\gamma(0) \in T_{f(a)}\mathcal{N}$. This formula may give the impression that $\dot\delta(0)$ depends on the diffeomorphism φ. This is not the case, because $\delta(t)$ is independent of φ. Moreover, we see that $\dot\delta(0)$ only depends on $v = \dot\gamma(0)$ and not on the complete curve $\gamma(t)$. This justifies the following definition.

Definition 2.4. *For a mapping $f : \mathcal{M} \to \mathcal{N}$ of class $\mathcal{C}^1$ we define*

$$T_a f : T_a\mathcal{M} \to T_{f(a)}\mathcal{N} \qquad \text{by} \qquad (T_a f)(v) = w,$$

where for a tangent vector $v = \dot\gamma(0) \in T_a\mathcal{M}$ we have $w = \dot\delta(0) \in T_{f(a)}\mathcal{N}$ with $\delta(t) = f(\gamma(t))$. The linear mapping $T_a f$ is called tangent map (or derivative) of f at a.

It is straight-forward to define mappings $f : \mathcal{M} \to \mathcal{N}$ of class $\mathcal{C}^k$ (k times continuously differentiable mappings). In this case one has to require that the diffeomorphisms φ of the charts (U, φ) of the manifold are also mappings of class $\mathcal{C}^k$.

II.3 Differential equations on submanifolds

We are interested in differential equations whose solutions evolve on a submanifold. If this is the case, so that $y(t)$ is a differentiable curve with values in $\mathcal{M} \subset \mathbb{R}^n$, then (by definition of the tangent space) its derivative $\dot y(t)$ satisfies $\dot y(t) \in T_{y(t)}\mathcal{M}$ for all t. This motivates the following definition.

Definition 3.1. *Let $\mathcal{M} \subset \mathbb{R}^n$ be a submanifold. A vector field on $\mathcal{M}$ is a $\mathcal{C}^1$-mapping $f : \mathcal{M} \to \mathbb{R}^n$ such that*

$$f(y) \in T_y\mathcal{M} \qquad \text{for all} \quad y \in \mathcal{M}. \tag{3.1}$$

For such a vector field,

$$\dot{y} = f(y)$$

is called differential equation on the submanifold $\mathcal{M}$, and a function $y : I \to \mathcal{M}$ satisfying $\dot{y}(t) = f(y(t))$ for all $t \in I$ is called integral curve or simply solution of the equation.

All examples of Chapter I are differential equations on submanifolds. Euler's equations of motion (I.1.1) are a differential equation on the sphere $\mathcal{M} = \{y \in \mathbb{R}^3 \,;\, \|y\|_2^2 - 1 = 0\}$. Since $y^{\mathsf{T}} f(y) = 0$ in this case, it follows from the second characterization of Theorem 2.2 that (3.1) is satisfied.

We next study the existence and uniqueness of solutions for differential equations on a submanifold. If the vector field $f(y)$ is defined on an open neighborhood of $\mathcal{M}$, this is a consequence from classical theory. If it is only defined on $\mathcal{M}$, additional considerations are necessary.

Theorem 3.2 (existence and uniqueness)**.** *Consider a differential equation $\dot{y} = f(y)$ on a submanifold $\mathcal{M} \subset \mathbb{R}^n$ with a $\mathcal{C}^1$ vector field $f : \mathcal{M} \to \mathbb{R}^n$. For every $y_0 \in \mathcal{M}$ there then exist a maximal open interval $I = I(y_0)$ and a $\mathcal{C}^2$ function $y : I \to \mathcal{M}$ satisfying*
(1) $y(t)$ is a solution of $\dot{y} = f(y)$ on I satisfying $y(0) = y_0$,
(2) if $\widehat{y} : J \to \mathcal{M}$ is a solution of $\dot{y} = f(y)$, $y(0) = y_0$ on the interval J, then $J \subset I$ and $\widehat{y}(t) = y(t)$ for $t \in J$.

Proof. Local existence and uniqueness. Consider an arbitrary $y_1 \in \mathcal{M}$ and a chart (U, φ) with $y_1 \in U$. We use the local parameterization $y = \eta(z) = \varphi^{-1}(z, \mathbf{0})$ of the manifold (see Lemma 1.3) and we write the differential equation in terms of z. For functions $y(t)$ and $z(t)$ related by $y(t) = \eta(z(t))$ we have $\dot{y}(t) = \eta'(z(t))\dot{z}(t)$, so that the initial value problem $\dot{y} = f(y)$, $y(t_1) = y_1$ becomes

$$\eta'(z)\,\dot{z} = f(\eta(z)), \qquad z(t_1) = z_1$$

with z_1 given by $y_1 = \eta(z_1)$. Premultiplication with $\eta'(z)^{\mathsf{T}}$ yields the following differential equation for z:

$$\dot{z} = \widetilde{f}(z), \qquad \widetilde{f}(z) = \Big(\eta'(z)^{\mathsf{T}}\eta'(z)\Big)^{-1}\eta'(z)^{\mathsf{T}} f(\eta(z)). \tag{3.2}$$

The matrix $\eta'(z)^{\mathsf{T}}\eta'(z)$ is invertible in z_1 (and hence also in a neighborhood), because $\eta'(z_1)$ is known to be of maximal rank. For a sufficiently smooth manifold $\mathcal{M}$ the function $\widetilde{f}$ is of class $\mathcal{C}^1$. Since (3.2) is a differential equation in an Euclidean space, we can apply the classical theory which yields the local existence and uniqueness of a solution $z(t)$. Because of $f(\eta(z)) \in T_{\eta(z)}\mathcal{M} = \operatorname{Im}\eta'(z)$, the function $y(t) = \eta(z(t))$ is seen to be a solution of $\dot{y} = f(y)$.

Global uniqueness. Let I, J be open intervals, and let $y : I \to \mathcal{M}$ and $\widehat{y} : J \to \mathcal{M}$ be two solutions of $\dot{y} = f(y)$ satisfying $y(0) = \widehat{y}(0) = y_0$. To prove that both functions coincide on the interval $I \cap J$, we consider the set

$$K = \{t \in I \cap J \,;\, y(t) = \widehat{y}(t)\}.$$

This set in nonempty ($0 \in K$) and closed in $I \cap J$ ($y(t)$ and $\widehat{y}(t)$ are continuous). Since $I \cap J$ is a connected set (an interval), it is sufficient to prove that K is also open. In fact, for $t_1 \in K$, we can choose a chart of $\mathcal{M}$ containing $y(t_1) = \widehat{y}(t_1)$. The above local existence and uniqueness result shows that we have $y(t) = \widehat{y}(t)$ for t in a neighborhood of t_1. This proves that K is open and, consequently, $K = I \cap J$.

Maximality of the interval $I = I(y_0)$. We consider all open intervals J such that the problem $\dot{y} = f(y)$, $y(0) = y_0$ admits a solution on J. We then let $I(y_0)$ be the union of all these intervals. For $t \in I(y_0)$ there exists J with $t \in J$, and we can define $y(t)$ as the value of the function $y : J \to \mathcal{M}$. By the uniqueness result, this is well defined and provides a solution on the maximal interval $I(y_0)$. □

The solution of a differential equation depends on the initial data. We adopt the notation $\phi_t(y_0) = y(t)$ for the solution of $\dot{y} = f(y)$ at time t corresponding to the initial condition $y(0) = y_0$. It is called the flow (exact flow in contrast to a discrete flow) of the differential equation. We also consider

$$D = \{(t, y_0)\,;\, y_0 \in \mathcal{M},\ t \in I(y_0)\} \qquad \text{and} \qquad \phi : D \to \mathcal{M}, \quad \phi(t, y_0) := \phi_t(y_0). \tag{3.3}$$

Theorem 3.3 (dependence on initial values)**.** *Consider a differential equation* $\dot{y} = f(y)$ *on a submanifold* $\mathcal{M} \subset \mathbb{R}^n$ *with a* $\mathcal{C}^1$ *vector field* $f : \mathcal{M} \to \mathbb{R}^n$*. Then, the set* D *of (3.3) is open in* $\mathbb{R} \times \mathcal{M}$*, and the flow mapping* $\phi : D \to \mathcal{M}$ *is of class* $\mathcal{C}^1$*.*

Proof. We first study differentiability for small t. We fix $y_0 \in \mathcal{M}$, we consider a chart (U, φ) with $y_0 \in U$, and we let z_0 be defined by $\varphi(y_0) = (z_0, \mathbf{0})$. As in the first part of the proof of Theorem 3.2 we consider the differential equation (3.2) in local coordinates, which allows us to apply classical results. In fact, the flow $\widehat{\phi}(t, z) := \widehat{\phi}_t(z)$ of (3.2) is well defined in an open neighborhood of $(0, z_0)$, and it is of class $\mathcal{C}^1$ as function of (t, z). Since the flow of the original differential equation can be expressed via $y = \eta(z)$ as $\phi_t(y) = (\phi_t \circ \varphi^{-1})(z, \mathbf{0}) = \eta(\widehat{\phi}_t(z))$, it is well defined in an open neighborhood of $(0, y_0)$ (as long as $\phi(t, y)$ remains in the same chart). It follows from Definition 2.3 that $\phi(t, y)$ is of class $\mathcal{C}^1$ in this neighborhood.

We next consider an initial value $y_0 \in \mathcal{M}$ and a finite interval $[0, \widehat{t}\,]$, on which the solution $y(t) = \phi_t(y_0)$ exists, i.e., $(\widehat{t}, y_0) \in D$. We shall prove below that it is possible to partition this interval into subintervals $0 = t_0 < t_1 < t_2 < \ldots < t_N = \widehat{t}$, such that for every $i \in \{0, \ldots, N-1\}$ there exists a chart (U_i, φ_i) such that $y(s) \in U_i$ for all $s \in [t_i, t_{i+1}]$ (see Figure II.4). The statement then follows from the fact that

$$\phi(t, y) = \phi_t(y) = \Big(\phi_{t-t_{N-1}} \circ \phi_{t_{N-1}-t_{N-2}} \circ \ldots \circ \phi_{t_2-t_1} \circ \phi_{t_1}\Big)(y). \tag{3.4}$$

By the local argument above each of the mappings $\phi_{t_{i+1}-t_i}$ (for $i \in \{0, \ldots, N-2\}$) is of class $\mathcal{C}^1$ in a neighborhood of $\phi_{t_i}(y_0)$, and the mapping $(t, y) \mapsto \phi_{t-t_{N-1}}(y)$ is defined and of class $\mathcal{C}^1$ for (t, y) in a neighborhood of $(t_N, \phi_{t_{N-1}}(y_0))$. This proves that D is open and that the composition (3.4) is of class $\mathcal{C}^1$.

The existence of such a partitioning follows from a compactness argument. For a fixed $\tau \in [0, \widehat{t}\,]$ there exists an open interval I_τ (with $\tau \in I_\tau$) and a chart (U_τ, φ_τ), such that

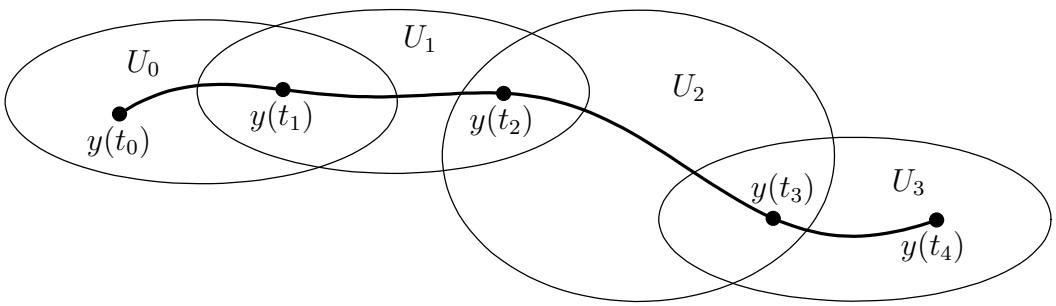


FIG. II.4: Patching together of the solution defined on charts.

$y(s) \in U_\tau$ for all $s \in I_\tau$. The family $\{I_\tau\}_{\tau\in[0,\hat{t}\,]}$ is an open covering of the compact interval $[0,\widehat{t}\,]$. By the Heine–Borel theorem we know that already finitely many intervals I_τ cover the whole interval. This completes the proof of the theorem. □

The following result on the propagation of perturbations in initial values will be an essential ingredient of the convergence analysis of numerical integrators on submanifolds.

Corollary 3.4 (propagation of perturbations)**.** *Consider a differential equation $\dot{y} = f(y)$ on a submanifold $\mathcal{M} \subset \mathbb{R}^n$ with a $\mathcal{C}^1$ vector field $f : \mathcal{M} \to \mathbb{R}^n$. Suppose that the solution $y(t) = \phi_t(y_0)$ exists for $0 \le t \le \widehat{t}$. Then there exist $\delta > 0$ and a constant C, such that*

$$\|\phi_{t-\tau}(y_1) - \phi_{t-\tau}(y_2)\| \le C\,\|y_1 - y_2\| \qquad \text{for} \qquad 0 \le \tau \le t \le \widehat{t}$$

for all $y_1, y_2 \in K_\tau(\delta)$, where the compact neighborhood of the solution is given by

$$K_\tau(\delta) = \Big\{ y \in \mathcal{M}\,;\, \|y - \phi_\tau(y_0)\| \le \delta \Big\}. \tag{3.5}$$

Proof. As in the proof of Theorem 3.3 we cover the solution $\phi_t(y_0)$ for $0 \le t \le \widehat{t}$ by finitely many charts (U_i, φ_i). Since the sets U_i are open, there exists $\delta_0 > 0$, such that $K_\tau(\delta_0) \subset U_i$ for all $\tau \in [t_i, t_{i+1}]$ and all $i \in \{0, \dots, N-1\}$. By the smoothness of the flow mapping (Theorem 3.3) and a compactness argument there exists $0 < \delta \le \delta_0$ such that for $\tau \in [t_i, t_{i+1}]$, for $y \in K_\tau(\delta)$, and for $\tau \le t \le \widehat{t}$, the solution $\phi_{t-\tau}(y)$ remains in $K_t(\delta_0)$.

We now consider $\tau \in [t_i, t_{i+1}]$ and $y_1, y_2 \in K_\tau(\delta)$ and we let local coordinates z_1, z_2 be given by $\varphi_i(y_1) = (z_1, \mathbf{0})$ and $\varphi_i(y_2) = (z_2, \mathbf{0})$. The mean value theorem, applied to the $\mathcal{C}^1$ mapping $\nu(z) = (\phi_{t-\tau} \circ \varphi_i^{-1})(z, \mathbf{0})$, yields the existence of a constant C_i such that

$$\|\phi_{t-\tau}(y_1) - \phi_{t-\tau}(y_2)\| = \|(\phi_{t-\tau} \circ \varphi_i^{-1})(z_1, \mathbf{0}) - (\phi_{t-\tau} \circ \varphi_i^{-1})(z_2, \mathbf{0})\| \le C_i\,\|z_1 - z_2\|$$

for all $y_1, y_2 \in K_\tau(\delta)$. A compactness argument implies that the constant C_i can be chosen independent of $\tau \in [t_i, t_{i+1}]$ and of $t \in [\tau, \widehat{t}\,]$. A further application of the mean value theorem yields

$$\|z_1 - z_2\| = \|\varphi_i(y_1) - \varphi_i(y_2)\| \le D_i \|y_1 - y_2\|,$$

which proves the statement of the corollary with $C = \max_{i=0,\dots,N-1} C_i D_i$. □

II.4 Differential equations on Lie groups

A *Lie group* is a group G which is a differentiable manifold, and for which the product is a differentiable mapping $G \times G \to G$. We restrict our considerations to *matrix Lie groups*, that is, Lie groups which are subgroups of $GL(n)$, the group of invertible $n \times n$ matrices with the usual matrix product as the group operation.[1]

An important example is the set

$$\mathrm{O}(n) = \Big\{ X \in \mathrm{GL}(n)\,;\, X^{\mathsf{T}} X = I \Big\}$$

of all orthogonal matrices, which is a submanifold of dimension $n(n-1)/2$ (see Example 1.6). With the usual product of matrices the set $\mathrm{O}(n)$ is a group with unit element I (the identity). Since the matrix multiplication is a differentiable mapping, $\mathrm{O}(n)$ is a Lie group.

[1]Section II.4 is nearly identical to Section IV.6 of the monograph *Geometric Numerical Integration* by Hairer, Lubich, and Wanner. For further reading on Lie groups we refer to the monographs *Applications of Lie Groups to Differential Equations* by Olver (1986) and to *Lie Groups, Lie Algebras and Their Representations* by Varadarajan (1974).

TAB. II.1: Some matrix Lie groups and their corresponding Lie algebras.

Lie group	Lie algebra
$\mathrm{GL}(n) = \{X\,;\,\det X \neq 0\}$ general linear group	$\mathfrak{gl}(n) = \{Z\,;\,\text{arbitrary matrix}\}$ Lie algebra of $n \times n$ matrices
$\mathrm{SL}(n) = \{X\,;\,\det X = 1\}$ special linear group	$\mathfrak{sl}(n) = \{Z\,;\,\mathrm{trace}(Z) = 0\}$ special linear Lie algebra
$\mathrm{O}(n) = \{X\,;\,X^TX = I\}$ orthogonal group	$\mathfrak{so}(n) = \{Z\,;\,Z^T + Z = 0\}$ skew-symmetric matrices
$\mathrm{SO}(n) = \{X \in \mathrm{O}(n)\,;\,\det X = 1\}$ special orthogonal group	$\mathfrak{so}(n) = \{Z\,;\,Z^T + Z = 0\}$ skew-symmetric matrices
$\mathrm{Sp}(n) = \{X\,;\,X^TJX = J\}$ symplectic group	$\mathfrak{sp}(n) = \{Z\,;\,JZ + Z^TJ = 0\}$

Table II.1 lists further prominent examples. The symplectic group is only defined for even n, and the matrix J given by

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

determines the symplectic structure on $\mathbb{R}^{2n}$.

As the following lemma shows, the tangent space $\mathfrak{g} = T_IG$ at the identity I of a matrix Lie group G is closed under forming commutators of its elements. This makes $\mathfrak{g}$ an algebra, the *Lie algebra* of the Lie group G.

Lemma 4.1 (Lie Bracket and Lie Algebra)**.** *Let G be a matrix Lie group and let $\mathfrak{g} = T_IG$ be the tangent space at the identity. The Lie bracket (or commutator)*

$$[A, B] = AB - BA \tag{4.1}$$

defines an operation $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ which is bilinear, skew-symmetric ($[A, B] = -[B, A]$), and satisfies the Jacobi identity

$$\Big[A, [B, C]\Big] + \Big[C, [A, B]\Big] + \Big[B, [C, A]\Big] = 0. \tag{4.2}$$

Proof. By definition of the tangent space, for $A, B \in \mathfrak{g}$, there exist differentiable paths $\alpha(t), \beta(t)$ $(|t| < \varepsilon)$ in G such that $\alpha(t) = I + tA + o(t)$ and $\beta(t) = I + tB + o(t)$. Consider now the path $\gamma(t)$ in G defined by

$$\gamma(t) = \alpha(\sqrt{t})\beta(\sqrt{t})\alpha(\sqrt{t})^{-1}\beta(\sqrt{t})^{-1} \qquad \text{for} \qquad t \geq 0.$$

An elementary computation then yields

$$\gamma(t) = I + t[A, B] + o(t).$$

With the extension $\gamma(t) = \gamma(-t)^{-1}$ for negative t, this is a differentiable path in G satisfying $\gamma(0) = I$ and $\dot\gamma(0) = [A, B]$. Hence $[A, B] \in \mathfrak{g}$ by definition of the tangent space. The properties of the Lie bracket can be verified in a straightforward way. □

Example 4.2. Consider again the orthogonal group $\mathrm{O}(n)$, see Example 1.6. Since the derivative of $g(X) = X^\mathsf{T}X - I$ at the identity is $g'(I)H = I^\mathsf{T}H + H^\mathsf{T}I = H + H^\mathsf{T}$, it follows from the second part of Theorem 2.2 that the Lie algebra corresponding to $\mathrm{O}(n)$ consists of all skew-symmetric matrices. The right column of Table II.1 gives the Lie algebras of the other Lie groups listed there.

The following basic lemma shows that the exponential map

$$\exp(A) = \sum_{k\geq 0} \frac{1}{k!}\, A^k$$

yields a local parametrization of the Lie group near the identity, with the Lie algebra (a linear space) as the parameter space. We recall that the mapping $Y(t) = \exp(tA)Y_0$ is the solution of the matrix differential equation $\dot{Y} = AY$, $Y(0) = Y_0$.

Lemma 4.3 (Exponential map)**.** *Consider a matrix Lie group G and its Lie algebra $\mathfrak{g}$. The matrix exponential maps the Lie algebra into the Lie group,*

$$\exp : \mathfrak{g} \to G,$$

i.e., for $A \in \mathfrak{g}$ we have $\exp(A) \in G$*. Moreover,* exp *is a local diffeomorphism in a neighbourhood of $A = 0$.*

Proof. For $A \in \mathfrak{g}$, it follows from the definition of the tangent space $\mathfrak{g} = T_IG$ that there exists a differentiable path $\alpha(t)$ in G satisfying $\alpha(0) = I$ and $\dot{\alpha}(0) = A$. For a fixed $Y \in G$, the path $\gamma(t) := \alpha(t)Y$ is in G and satisfies $\gamma(0) = Y$ and $\dot{\gamma}(0) = AY$. Consequently, $AY \in T_YG$ and $\dot{Y} = AY$ defines a differential equation on the manifold G. The solution $Y(t) = \exp(tA)$ is therefore in G for all t.

Since $\exp(H) - \exp(\mathbf{0}) = H + \mathcal{O}(H^2)$, the derivative of the exponential map at $A = \mathbf{0}$ is the identity, and it follows from the inverse function theorem that exp is a local diffeomorphism close to $A = \mathbf{0}$. □

The proof of Lemma 4.3 shows that for a matrix Lie group G the tangent space at $Y \in G$ has the form

$$T_YG = \{AY\,;\, A \in \mathfrak{g}\}. \tag{4.3}$$

By Definition 3.1, differential equations on a matrix Lie group (considered as a manifold) can therefore be written as

$$\dot{Y} = A(Y)Y \tag{4.4}$$

where $A(Y) \in \mathfrak{g}$ for all $Y \in G$. The following theorem summarizes this discussion.

Theorem 4.4. *Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. If $A(Y) \in \mathfrak{g}$ for all $Y \in G$ and if $Y_0 \in G$, then the solution of (4.4) satisfies $Y(t) \in G$ for all t.* □

II.5 Exercises

1. Consider the 2-dimensional torus of Example 1.4. Find a function $g : \mathbb{R}^3 \to \mathbb{R}$ such that the manifold is given by $\mathcal{M} = \{x\,;\, g(x) = 0\}$. Prove that $g'(x) \neq \mathbf{0}$ for all $x \in \mathcal{M}$.
   Result. $g(x) = \left(x_1^2 + x_2^2 + x_3^2 + d^2 - \rho^2\right)^2 - 4\,d^2(x_1^2 + x_2^2)$.

2. Which of the following sets are submanifolds? Draw pictures if possible.

$$\begin{array}{ll}
\{(t,t^2)\in\mathbb{R}^2\ ;\ t\in\mathbb{R}\} & \{(t,t^2)\in\mathbb{R}^2\ ;\ t\geq 0\}\\
\{(t^2,t^3)\in\mathbb{R}^2\ ;\ t\in\mathbb{R}\} & \{(t^2,t^3)\in\mathbb{R}^2\ ;\ t\neq 0\}\\
\{(x,y)\in\mathbb{R}^2\ ;\ x>0, y>0\} & \{(x,y,z)\in\mathbb{R}^3\ ;\ x=y=z=0\}\\
\{(x,y,z)\in\mathbb{R}^3\ ;\ x^2+y^2-z^2=1\} & \{(x,y,z)\in\mathbb{R}^3\ ;\ x^2+y^2-z^2=0\}
\end{array}$$

3. The circle $\mathbb{S}^1=\{x\in\mathbb{R}^2\,;\,\|x\|=1\}$ is a submanifold of $\mathbb{R}^2$. Prove that it cannot be covered by only one chart (U,φ).

4. Prove that the cylinder $\mathcal{M}=\mathbb{S}^1\times\mathbb{R}$ is a submanifold of $\mathbb{R}^3$.
   a) Find a function $g:\mathbb{R}^3\to\mathbb{R}$ such that $\mathcal{M}$ is a level set of g.
   b) Find local parametrizations of the cylinder.
   c) Find an atlas of the cylinder (i.e., a union of charts that cover $\mathcal{M}$).

5. Let $\mathcal{M}\subset\mathbb{R}^n$ and $\mathcal{N}\subset\mathbb{R}^m$ be two submanifolds. Prove that the product

$$\mathcal{M}\times\mathcal{N}=\{(x,y)\in\mathbb{R}^n\times\mathbb{R}^m\,;\,x\in\mathcal{M},\ y\in\mathcal{N}\}$$

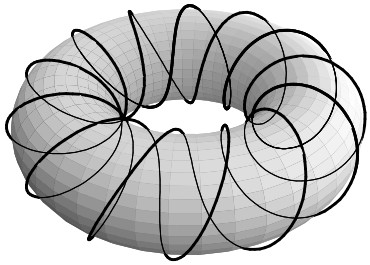
   is a submanifold of $\mathbb{R}^n\times\mathbb{R}^m$.

6. Prove that the set

$$\{(\cos t+2)\cos\lambda t,(\cos t+2)\sin\lambda t,\sin t)\in\mathbb{R}^3\ ;\ t\in\mathbb{R}\} \tag{5.1}$$

   is a submanifold of $\mathbb{R}^3$ for $\lambda=2/13$ (see the figure). For $\lambda=\sqrt{2}$ the set (5.1) is not a submanifold and it is everywhere dense in the torus

$$\{(\cos u+2)\cos v,(\cos u+2)\sin v,\sin u)\ ;\ u,v\in\mathbb{R}\}.$$



   Hint. Using continued fractions (see Section I.6 of the textbook "Analysis by Its History" by Hairer & Wanner), prove that the set $\{\ell+k\sqrt{2}\ ;\ \ell,k\in\mathbb{Z}\}$ is dense in $\mathbb{R}$.

7. Consider the n-dimensional sphere $\mathbb{S}^n\subset\mathbb{R}^{n+1}$, and let $N=(0,\dots,0,1)$ be its 'north pole'. Define the *stereographic projection* $\sigma:\mathbb{S}^n\setminus\{N\}\to\mathbb{R}^n$ by

$$\sigma(x_1,\dots,x_n,x_{n+1})=\frac{1}{1-x_{n+1}}\big(x_1,\dots,x_n\big)^{\mathsf{T}}.$$

   a) For any $x\in\mathbb{S}^n\setminus\{N\}$, prove that $\sigma(x)$ is the point where the line trough N and x intersects the hyperplane $x_{n+1}=0$ (which is identified with $\mathbb{R}^n$).

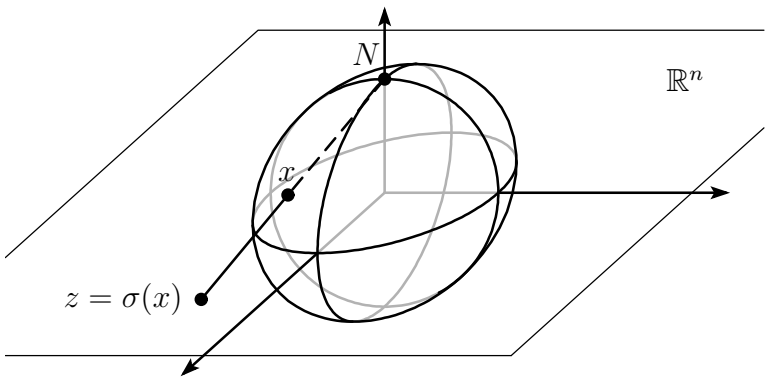


FIG. II.5: Stereographic projection.

b) Prove that σ is bijective, and that its inverse $\eta = \sigma^{-1}$ is given by

$$\eta(z_1, \dots, z_n) = \frac{1}{\|z\|^2 + 1}\Big(2z_1, \dots, 2z_n, \|z\|^2 - 1\Big)^{\mathsf{T}}.$$

c) For any $z \in \mathbb{R}^n$, prove that the matrix $\eta'(z)$ is of full rank n. Determine for which $z \in \mathbb{R}^n$ the first n lines of $\eta'(z)$ are not linearly independent.
d) For a fixed $x \in \mathbb{S}^n \setminus \{N\}$ with $x_{n+1} \neq 0$, find a chart (U, φ) with $x \in U$ by following the proof of Lemma 1.3.

8. Let $\mathcal{M}$, $\mathcal{N}$, $\mathcal{P}$ be submanifolds, and let $g : \mathcal{M} \to \mathcal{N}$, $f : \mathcal{N} \to \mathcal{P}$ be $\mathcal{C}^1$-mappings. Prove that the composition $f \circ g$ is a $\mathcal{C}^1$-mapping, and that its tangent map satisfies

$$T_a(f \circ g) = T_{g(a)}f \circ T_a g\,.$$

9. Consider a compact submanifold $\mathcal{M}$ (e.g., the sphere or the torus) and a $\mathcal{C}^1$ vector field $f(y)$ on $\mathcal{M}$. Prove that for every $y_0 \in \mathcal{M}$ the solution $y(t)$ of the initial value problem $\dot{y} = f(y)$, $y(0) = y_0$ exists for all $t \in (-\infty, +\infty)$.

10. Prove that $\mathrm{SL}(n)$ is a Lie group of dimension $n^2 - 1$, and that $\mathfrak{sl}(n)$ is its Lie algebra (see Table II.1 for the definitions of $\mathrm{SL}(n)$ and $\mathfrak{sl}(n)$).

11. Let G be a matrix Lie group and $\mathfrak{g}$ its Lie algebra. Prove that for $X \in G$ and $A \in \mathfrak{g}$ we have $XAX^{-1} \in \mathfrak{g}$.
Hint. Consider the path $\gamma(t) = X\alpha(t)X^{-1}$.

Chapter III

Integrators on Manifolds

We consider ordinary differential equations

$$\dot{y} = f(y), \qquad y(0) = y_0 \tag{0.1}$$

on a submanifold $\mathcal{M}$, i.e., we assume that $f(y) \in T_y\mathcal{M}$ for all $y \in \mathcal{M}$. This chapter is devoted to the numerical solution of such problems. We discuss projection methods, integrators based on local coordinates, and Magnus series methods for linear differential equations on Lie groups. We also show how the global error can be estimated (global convergence).

III.1 Projection methods

We start by assuming that the vector field $f(y)$ is well defined in an open neighborhood of the manifold $\mathcal{M}$. In principle it is then possible to apply any numerical integrator (Runge–Kutta, multistep, etc.) to the differential equation (0.1) without taking care of the manifold. However, as we have seen in Chapter I (for example in Figure I.2), the numerical solution will usually drift away from the manifold and often looses a physical interpretation. A natural approach for avoiding such unphysical approximations is by projection[1].

Algorithm 1.1 (Standard projection method)**.** *Assume that $y_n \in \mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. III.1):*

- *Compute $\widetilde{y}_{n+1} = \Phi_h(y_n)$, where Φ_h is an arbitrary one-step method applied to $\dot{y} = f(y)$;*
- *project the value $\widetilde{y}_{n+1}$ onto the manifold $\mathcal{M}$ to obtain $y_{n+1} \in \mathcal{M}$.*

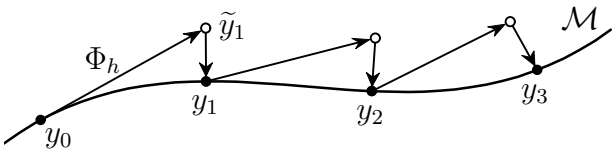


FIG. III.1: Illustration of the standard projection method.

For $y_n \in \mathcal{M}$ the distance of $\widetilde{y}_{n+1}$ to the manifold $\mathcal{M}$ is of the size of the local error, i.e., $\mathcal{O}(h^{p+1})$ for a method of order p. Therefore, we do not expect that this projection algorithm destroys the convergence order of the method.

[1]For more details consult the following monographs: Sections IV.4 and V.4.1 of *Geometric Numerical Integration* by Hairer, Lubich and Wanner (2006), Section VII.2 of *Solving Ordinary Differential Equations II* by Hairer and Wanner (1996), and Section 5.3.3 of *Numerical Methods in Multibody Dynamics* by Eich-Soellner and Führer (1998).

In some situations the projection step is straight-forward. If $\mathcal{M}$ is the unit sphere (e.g., for Euler's equation of motion for a rigid body, Section I.1), we simply divide the approximation $\widetilde{y}_{n+1}$ by its Euclidean norm to get a vector of length one.

If the manifold is given by a local parametrization $y = \eta(z)$, we compute z_{n+1} by minimizing $\|\eta(z_{n+1}) - \widetilde{y}_{n+1}\|$ in a suitable norm, and then we put $y_{n+1} = \eta(z_{n+1})$. But this situation is not important in practice, because we can treat directly the differential equation (II.3.2) for z, if explicit formulas for the parametrization are known. This yields approximations z_n and $y_n := \eta(z_n)$, which lie on the manifold by definition.

Projection step, if the manifold is given as a level set. For all examples of Chapter I the manifold $\mathcal{M}$ is given as the level set of a smooth function $g(y) = (g_1(y), \dots, g_m(y))^\mathsf{T}$. This is by far the most important situation. For the computation of y_{n+1} (projection step) we have to solve the constrained minimization problem

$$\|y_{n+1} - \widetilde{y}_{n+1}\| \to \min \qquad \text{subject to} \qquad g(y_{n+1}) = 0. \tag{1.1}$$

In the case of the Euclidean norm, a standard approach is to introduce Lagrange multipliers $\lambda = (\lambda_1, \dots, \lambda_m)^\mathsf{T}$, and to consider the Lagrange function

$$\mathcal{L}(y_{n+1}, \lambda) = \frac{1}{2}\,\|y_{n+1} - \widetilde{y}_{n+1}\|^2 - g(y_{n+1})^\mathsf{T}\lambda.$$

The necessary condition $\partial\mathcal{L}/\partial y_{n+1} = 0$ then leads to the system

$$\begin{aligned} y_{n+1} &= \widetilde{y}_{n+1} + g'(\widetilde{y}_{n+1})^\mathsf{T}\lambda \\ 0 &= g(y_{n+1}). \end{aligned} \tag{1.2}$$

We have replaced y_{n+1} with $\widetilde{y}_{n+1}$ in the argument of $g'(y)$ in order to save some evaluations of $g'(y)$. Inserting the first relation of (1.2) into the second gives a nonlinear equation for λ, which can be efficiently solved by simplified Newton iterations:

$$\Delta\lambda_i = -\Big(g'(\widetilde{y}_{n+1})g'(\widetilde{y}_{n+1})^\mathsf{T}\Big)^{-1} g\Big(\widetilde{y}_{n+1} + g'(\widetilde{y}_{n+1})^\mathsf{T}\lambda_i\Big), \qquad \lambda_{i+1} = \lambda_i + \Delta\lambda_i.$$

For the choice $\lambda_0 = 0$ the first increment $\Delta\lambda_0$ is of size $\mathcal{O}(h^{p+1})$, so that the convergence is usually extremely fast. Often, one simplified Newton iteration is sufficient to achieve the desired precision.

Internal projection. We assume here that the vector field $f(y)$ is only defined on the manifold $\mathcal{M}$, and not on a whole neighborhood. It may also happen that the differential equation has a different (stability) behavior outside the manifold. In this case we are interested in numerical methods that evaluate the vector field only on the manifold.

The idea is the following. We denote by $\pi(y)$ a smooth projection of a vector y onto the manifold. Since $\pi(y) = y$ for $y \in \mathcal{M}$, the solution of the differential equation

$$\dot{y} = f(\pi(y)), \qquad y(0) = y_0 \in \mathcal{M} \tag{1.3}$$

is identical to that of (0.1). We then apply our integrator to (1.3) instead of (0.1). For a Runge–Kutta method, e.g.,

$$\begin{aligned} k_1 &= f(\pi(y_n)) \\ k_2 &= f(\pi(y_n + a_{21}hk_1)) \\ y_{n+1} &= y_n + h(b_1k_1 + b_2k_2), \end{aligned}$$

this means that we do not only project y_{n+1} onto the manifold, but also the vector $y_n + a_{21}hk_1$ before computing k_2.

Example 1.2 (volume preservation)**.** Consider a matrix differential equation $\dot{Y} = A(Y)Y$, where $\operatorname{trace} A(Y) = 0$ for all Y. We know from Theorem II.4.4 that the solution stays on the manifold $\mathcal{M} = \{Y\,;\, \det Y = \mathit{Const}\}$. Let $\widetilde{Y}_{n+1}$ be the numerical approximation obtained with an arbitrary one-step method. We consider the Frobenius norm $\|Y\|_F = \sqrt{\sum_{i,j} |y_{ij}|^2}$ for measuring the distance to the manifold $\mathcal{M}$. Using $g'(Y)(HY) = \operatorname{trace} H \det Y$ for the function $g(Y) = \det Y$ with H chosen such that the product HY contains only one non-zero element, the projection step (1.2) is seen to become (see Exercises 1 and 2)

$$Y_{n+1} = \widetilde{Y}_{n+1} + \mu \widetilde{Y}_{n+1}^{-\mathsf{T}}, \tag{1.4}$$

where the scalar μ is given by $\mu = \lambda \det \widetilde{Y}_{n+1}$. This leads to the scalar nonlinear equation $\det(\widetilde{Y}_{n+1} + \mu \widetilde{Y}_{n+1}^{-\mathsf{T}}) = \det Y_n$, for which simplified Newton iterations become

$$\det\Big(\widetilde{Y}_{n+1} + \mu_i \widetilde{Y}_{n+1}^{-\mathsf{T}}\Big)\Big(1 + (\mu_{i+1} - \mu_i) \operatorname{trace}\Big((\widetilde{Y}_{n+1}^{\mathsf{T}} \widetilde{Y}_{n+1})^{-1}\Big)\Big) = \det Y_n.$$

If the QR-decomposition of $\widetilde{Y}_{n+1}$ is available from the computation of $\det \widetilde{Y}_{n+1}$, the value of $\operatorname{trace}((\widetilde{Y}_{n+1}^{\mathsf{T}} \widetilde{Y}_{n+1})^{-1})$ can be computed efficiently with $\mathcal{O}(n^3/3)$ flops.

The above projection is preferable to $Y_{n+1} = c\,\widetilde{Y}_{n+1}$, where $c \in \mathbb{R}$ is chosen such that $\det Y_{n+1} = \det Y_n$. This latter projection is already ill-conditioned for diagonal matrices with entries that differ by several magnitudes.

Example 1.3 (orthogonal matrices)**.** As a second example let us consider $\dot{Y} = F(Y)$, where the solution $Y(t)$ is known to be an orthogonal matrix or, more generally, an $n \times k$ matrix satisfying $Y^T Y = I$ (Stiefel manifold). The projection step (1.1) requires the solution of the problem

$$\|Y - \widetilde{Y}\|_F \to \min \qquad \text{subject to} \qquad Y^{\mathsf{T}} Y = I, \tag{1.5}$$

where $\widetilde{Y}$ is a given matrix. This projection can be computed as follows: compute the singular value decomposition $\widetilde{Y} = U^{\mathsf{T}} \Sigma V$, where U^{T} and V are $n \times k$ and $k \times k$ matrices with orthonormal columns, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$, and the singular values $\sigma_1 \geq \ldots \geq \sigma_k$ are all close to 1. Then the solution of (1.5) is given by the product $Y = U^{\mathsf{T}} V$ (see Exercise 3 for some hints).

This procedure has a different interpretation: the orthogonal projection is the first factor of the *polar decomposition* $\widetilde{Y} = YR$ (where Y has orthonormal columns and R is symmetric positive definite). The equivalence is seen from the polar decomposition $\widetilde{Y} = (U^{\mathsf{T}} V)(V^{\mathsf{T}} \Sigma V)$.

III.2 Numerical methods based on local coordinates

Let $y = \eta(z)$ be a local parametrization of the manifold $\mathcal{M}$. As discussed in Section II.3, the differential equation (0.1) on the manifold $\mathcal{M}$ is then equivalent to

$$\dot{z} = \eta'(z)^+ f(\eta(z)), \tag{2.1}$$

where $A^+ = (A^T A)^{-1} A^T$ denotes the pseudo-inverse of a matrix with full column rank. The solutions of (0.1) and (2.1) are related via $y(t) = \eta(z(t))$, so that any approximation z_n of $z(t_n)$ also provides an approximation $y_n = \eta(z_n) \approx y(t_n)$. The idea is to apply the numerical integrator in the parameter space rather than in the space where $\mathcal{M}$ is embedded. In contrast to projection methods (Section III.1), the numerical integrators of this section evaluate $f(y)$ only on the manifold $\mathcal{M}$.

Algorithm 2.1 (Local Coordinates Approach)**.** *Assume that $y_n \in \mathcal{M}$ and that $y = \eta(z)$ is a local parametrization of $\mathcal{M}$. One step $y_n \mapsto y_{n+1}$ is defined as follows (see Fig. III.2):*

- *determine z_n in the parameter space, such that $\eta(z_n) = y_n$;*
- *compute $\widetilde{z}_{n+1} = \Phi_h(z_n)$, the result of the numerical method Φ_h applied to (2.1);*
- *define the numerical solution by $y_{n+1} = \eta(\widetilde{z}_{n+1})$.*

It is important to remark that the parametrization $y = \eta(z)$ can be changed at every step.

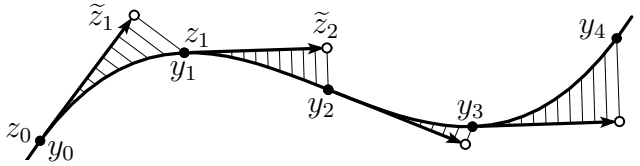
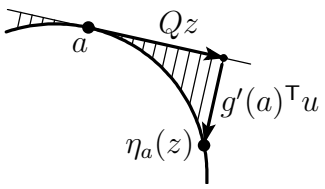


FIG. III.2: The numerical solution of differential equations on manifolds via local coordinates.

There are many possible choices of local coordinates. For the mathematical pendulum of Example I.3.1, where $\mathcal{M} = \big\{(q_1, q_2, v_1, v_2) \,|\, q_1^2 + q_2^2 = 1,\ q_1 v_1 + q_2 v_2 = 0\big\}$, a standard parametrization is $q_1 = \sin\alpha$, $q_2 = -\cos\alpha$, $v_1 = \omega\cos\alpha$, and $v_2 = \omega\sin\alpha$. In the new coordinates (α, ω) the problem becomes simply $\dot\alpha = \omega$, $\dot\omega = -\sin\alpha$. Another typical choice is the exponential map $\eta(Z) = \exp(Z)$ for differential equations on Lie groups. In this section we are mainly interested in the situation where the manifold is given as the level set of a smooth function $g(y)$, and we discuss two commonly used choices which do not use any special structure of the manifold.

Generalized Coordinate Partitioning. We assume that the manifold is given by $\mathcal{M} = \{y \in \mathbb{R}^n\,;\, g(y) = \mathbf{0}\}$, where $g : \mathbb{R}^n \to \mathbb{R}^m$ has a Jacobian with full rank $m < n$ at $y = a$. We can then find a partitioning $y = (y_1, y_2)$, such that $\partial g/\partial y_2(a)$ is invertible. In this case we can choose the components of y_1 as local coordinates. The function $y = \eta(z)$ is then given by $y_1 = z$ and $y_2 = \eta_2(z)$, where $\eta_2(z)$ is implicitly defined by $g(z, \eta_2(z)) = 0$, and (2.1) reduces to $\dot z = f_1(\eta(z))$, where $f_1(y)$ denotes the first $n - m$ components of $f(y)$. This approach has been promoted by Wehage and Haug[2] in the context of constrained mechanical systems, and the partitioning is found by Gaussian elimination with full pivoting applied to the matrix $g'(a)$. Another way of finding the partitioning is by the use of the QR decomposition with column change.

Tangent Space Parametrization. Let the manifold $\mathcal{M}$ be given as the level set of a smooth function $g : \mathbb{R}^n \to \mathbb{R}^m$. We compute an orthonormal basis of the tangent space $T_a\mathcal{M} = \ker g'(a)$ at $a = y_n$, and we collect the basis vectors as columns of the matrix Q, which is of dimension $n \times (n - m)$. This matrix satisfies $Q^\mathsf{T} Q = I$ and $g'(a)Q = 0$. We then consider the parametrization



$$\eta_a(z) = a + Qz + g'(a)^\mathsf{T} u(z), \tag{2.2}$$

where $u(z)$ is defined by $g(\eta_a(z)) = 0$. The existence and local uniqueness of $u(z)$ with $u(\mathbf{0}) = \mathbf{0}$ follows for small z from the implicit function theorem. In fact, the function $F(z, u) := g(a + Qz + g'(a)^\mathsf{T} u)$ satisfies $F(\mathbf{0}, \mathbf{0}) = \mathbf{0}$ and its derivative with respect to u is

[2] *Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems*, Mechanical Design 104 (1982) 247–255.

for $(z,u) = (\mathbf{0},\mathbf{0})$ the matrix $g'(a)g'(a)^\mathsf{T}$, which is invertible because $g'(a)$ is assumed to be of full rank. Differentiating $y(t) = \eta_a(z(t))$ with respect to time yields

$$\Big(Q + g'(a)^\mathsf{T} u'(z)\Big)\dot z = \dot y = f(y) = f(\eta_a(z)).$$

Because of $Q^TQ = I$ and $g'(a)Q = 0$, a premultiplication with Q^T leads to the equation

$$\dot z = Q^\mathsf{T} f(\eta_a(z)), \tag{2.3}$$

which corresponds to (2.1). If we apply a numerical method to (2.3), every function evaluation requires the projection of an element of the tangent space onto the manifold. This procedure is illustrated in Fig. III.2, and was originally proposed by Potra and Rheinboldt[3] for the solution of the Euler–Lagrange equations of constrained multibody systems.

III.3 Derivative of the exponential and its inverse

The exponential function exp plays an important role as local parametrization of Lie groups (Section II.4). In view of the differential equation (2.1) we need the derivative of exp and its inverse. Elegant formulas are obtained by the use of matrix commutators $[\Omega, A] = \Omega A - A\Omega$. If we suppose Ω fixed, this expression defines a linear operator $A \mapsto [\Omega, A]$

$$\mathrm{ad}_\Omega(A) = [\Omega, A], \tag{3.1}$$

which is called the *adjoint operator*. Let us start by computing the derivatives of Ω^k. The product rule for differentiation yields

$$\Big(\frac{d}{d\Omega}\,\Omega^k\Big)H = H\Omega^{k-1} + \Omega H\Omega^{k-2} + \ldots + \Omega^{k-1}H, \tag{3.2}$$

and this equals $kH\Omega^{k-1}$ if Ω and H commute. Therefore, it is natural to write (3.2) as $kH\Omega^{k-1}$ to which are added correction terms involving commutators and iterated commutators. In the cases $k = 2$ and $k = 3$ we have

$$\begin{aligned} H\Omega + \Omega H &= 2H\Omega + \mathrm{ad}_\Omega(H) \\ H\Omega^2 + \Omega H\Omega + \Omega^2 H &= 3H\Omega^2 + 3\Big(\mathrm{ad}_\Omega(H)\Big)\Omega + \mathrm{ad}_\Omega^2(H), \end{aligned}$$

where ad_Ω^i denotes the iterated application of the linear operator ad_Ω. With the convention $\mathrm{ad}_\Omega^0(H) = H$ we obtain by induction on k that

$$\Big(\frac{d}{d\Omega}\,\Omega^k\Big)H = \sum_{i=0}^{k-1}\binom{k}{i+1}\Big(\mathrm{ad}_\Omega^i(H)\Big)\Omega^{k-i-1}. \tag{3.3}$$

This is seen by applying Leibniz' rule to $\Omega^{k+1} = \Omega\cdot\Omega^k$ and by using the identity $\Omega(\mathrm{ad}_\Omega^i(H)) = (\mathrm{ad}_\Omega^i(H))\Omega + \mathrm{ad}_\Omega^{i+1}(H)$.

Lemma 3.1. *The derivative of* $\exp\Omega = \sum_{k\ge0}\frac{1}{k!}\,\Omega^k$ *is given by*

$$\Big(\frac{d}{d\Omega}\,\exp\Omega\Big)H = \Big(d\exp_\Omega(H)\Big)\exp\Omega,$$

where

$$d\exp_\Omega(H) = \sum_{k\ge0}\frac{1}{(k+1)!}\,\mathrm{ad}_\Omega^k(H). \tag{3.4}$$

The series (3.4) converges for all matrices Ω.

[3] *On the numerical solution of Euler–Lagrange equations*, Mech. Struct. & Mech. 19 (1991) 1–18; see also page 476 of the monograph *Solving Ordinary Differential Equations II* by Hairer and Wanner (1996).

Proof. Multiplying (3.3) by $(k!)^{-1}$ and summing, then exchanging the sums and putting $j = k - i - 1$ yields

$$\Big(\frac{d}{d\Omega}\,\exp\Omega\Big)H = \sum_{k\ge 0}\frac{1}{k!}\sum_{i=0}^{k-1}\binom{k}{i+1}\Big(\mathrm{ad}_\Omega^i(H)\Big)\Omega^{k-i-1} = \sum_{i\ge 0}\sum_{j\ge 0}\frac{1}{(i+1)!\,j!}\Big(\mathrm{ad}_\Omega^i(H)\Big)\Omega^j.$$

The convergence of the series follows from the boundedness of the linear operator ad_Ω (we have $\|\mathrm{ad}_\Omega\| \le 2\|\Omega\|$). □

Lemma 3.2. *If the eigenvalues of the linear operator* ad_Ω *are different from* $2\ell\pi i$ *with* $\ell \in \{\pm 1, \pm 2, \ldots\}$*, then* $d\exp_\Omega$ *is invertible. Furthermore, we have for* $\|\Omega\| < \pi$ *that*

$$d\exp_\Omega^{-1}(H) = \sum_{k\ge 0}\frac{B_k}{k!}\,\mathrm{ad}_\Omega^k(H), \tag{3.5}$$

where B_k *are the Bernoulli numbers, defined by* $\sum_{k\ge 0}(B_k/k!)x^k = x/(e^x-1)$.

Proof. The eigenvalues of $d\exp_\Omega$ are $\mu = \sum_{k\ge 0}\lambda^k/(k+1)! = (e^\lambda - 1)/\lambda$, where λ is an eigenvalue of ad_Ω. By our assumption, the values μ are non-zero, so that $d\exp_\Omega$ is invertible. By definition of the Bernoulli numbers, the composition of (3.5) with (3.4) gives the identity. Convergence for $\|\Omega\| < \pi$ follows from $\|\mathrm{ad}_\Omega\| \le 2\|\Omega\|$ and from the fact that the radius of convergence of the series for $x/(e^x-1)$ is 2π. □

III.4 Methods based on the Magnus series expansion

Our next aim is the numerical solution of differential equations (II.4.4) on Lie groups. For this purpose we consider linear matrix differential equations of the form

$$\dot Y = A(t)\,Y. \tag{4.1}$$

No assumption on the matrix $A(t)$ is made for the moment (apart from continuous dependence on t). For the scalar case, the solution of (4.1) with $Y(0) = Y_0$ is given by

$$Y(t) = \exp\Big(\int_0^t A(\tau)\,d\tau\Big)\,Y_0. \tag{4.2}$$

Also in the case where the matrices $A(t)$ and $\int_0^t A(\tau)\,d\tau$ commute, (4.2) is the solution of (4.1). In the general non-commutative case we search for a matrix function $\Omega(t)$ such that

$$Y(t) = \exp(\Omega(t))\,Y_0$$

solves (4.1). The main ingredient for the solution will be the inverse of the derivative of the matrix exponential. It has been studied in Section III.3.

Theorem 4.1 (Magnus 1954)**.** *The solution of the differential equation (4.1) can be written as* $Y(t) = \exp(\Omega(t))Y_0$ *with* $\Omega(t)$ *given by*

$$\dot\Omega = d\exp_\Omega^{-1}(A(t)), \qquad \Omega(0) = 0. \tag{4.3}$$

As long as $\|\Omega(t)\| < \pi$*, the convergence of the series expansion (3.5) of* $d\exp_\Omega^{-1}$ *is assured.*

Proof. Comparing the derivative of $Y(t) = \exp(\Omega(t))Y_0$,

$$\dot Y(t) = \Big(\frac{d}{d\Omega}\exp(\Omega(t))\Big)\,\dot\Omega(t)\,Y_0 = \Big(d\exp_{\Omega(t)}(\dot\Omega(t))\Big)\,\exp(\Omega(t))\,Y_0,$$

with (4.1) we obtain $A(t) = d\exp_{\Omega(t)}(\dot\Omega(t))$. Applying the inverse operator $d\exp_\Omega^{-1}$ to this relation yields the differential equation (4.3) for $\Omega(t)$. The statement on the convergence is a consequence of Lemma 3.2. □

The first few Bernoulli numbers are $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_3 = 0$. The differential equation (4.3) therefore becomes

$$\dot\Omega = A(t) - \frac{1}{2}\big[\Omega, A(t)\big] + \frac{1}{12}\big[\Omega, \big[\Omega, A(t)\big]\,\big] + \dots\ ,$$

which is nonlinear in Ω. Applying Picard fixed point iteration after integration yields

$$\begin{aligned}\Omega(t) \;=\; \int_0^t A(\tau)\,d\tau \;&-\; \frac{1}{2}\int_0^t \Big[\int_0^\tau A(\sigma)\,d\sigma, A(\tau)\Big]\,d\tau \\ &+\; \frac{1}{4}\int_0^t \Big[\int_0^\tau \Big[\int_0^\sigma A(\mu)\,d\mu, A(\sigma)\Big]\,d\sigma, A(\tau)\Big]\,d\tau \\ &+\; \frac{1}{12}\int_0^t \Big[\int_0^\tau A(\sigma)\,d\sigma, \Big[\int_0^\tau A(\mu)\,d\mu, A(\tau)\Big]\Big]\,d\tau + \dots\ ,\end{aligned} \tag{4.4}$$

which is the so-called *Magnus expansion*. For smooth matrices $A(t)$ the remainder in (4.4) is of size $\mathcal{O}(t^5)$ so that the truncated series inserted into $Y(t) = \exp(\Omega(t))Y_0$ gives an excellent approximation to the solution of (4.1) for small t.

Numerical Methods Based on the Magnus Expansion. The matrix Ω can be considered as local coordinates for $Y = \exp(\Omega)Y_n$. The differential equation (4.3) corresponds to equation (2.1) in the general situation. Following the steps in Algorithm 2.1 we let $\Omega_n = \mathbf{0}$, we compute an approximation Ω_{n+1} of $\Omega(h)$ given by (4.4) with $A(t_n + \tau)$ instead of $A(\tau)$, and we finally put $Y_{n+1} = \exp(\Omega_{n+1})\,Y_n$. For Ω_{n+1} it is natural to take a suitable truncation of the Magnus expansion with the integrals approximated by numerical quadrature.[4] A related approach is to replace $A(t)$ locally by an interpolation polynomial

$$\widehat{A}(t) = \sum_{i=1}^{s} \ell_i(t)\, A(t_n + c_i h),$$

and to solve $\dot Y = \widehat{A}(t)\,Y$ on $[t_n, t_n + h]$ by the use of the truncated series (4.4).

Theorem 4.2. *Consider a quadrature formula $(b_i, c_i)_{i=1}^s$ of order $p \ge s$, and let $Y(t)$ and $Z(t)$ be solutions of $\dot Y = A(t)\,Y$ and $\dot Z = \widehat{A}(t)\,Z$, respectively, satisfying $Y(t_n) = Z(t_n)$. Then, $Z(t_n + h) - Y(t_n + h) = \mathcal{O}(h^{p+1})$.*

Proof. We write the differential equation for Z as $\dot Z = A(t)Z + (\widehat{A}(t) - A(t))Z$ and use the variation of constants formula to get

$$Z(t_n + h) - Y(t_n + h) = \int_{t_n}^{t_n+h} R(t_n + h, \tau)\Big(\widehat{A}(\tau) - A(\tau)\Big)Z(\tau)\,d\tau.$$

Applying our quadrature formula to this integral gives zero as result, and the remainder is of size $\mathcal{O}(h^{p+1})$. Details of the proof are omitted. □

[4] Iserles and Nørsett, *On the solution of linear differential equations in Lie groups* (1999); Zanna, *Collocation and relaxed collocation for the Fer and the Magnus expansions* (1999).

Example 4.3. As a first example, we use the midpoint rule ($c_1 = 1/2$, $b_1 = 1$). In this case the interpolation polynomial is constant, and the method becomes

$$Y_{n+1} = \exp\Big(hA(t_n + h/2)\Big)\, Y_n, \tag{4.5}$$

which is of order 2.

Example 4.4. The two-stage Gauss quadrature is given by $c_{1,2} = 1/2 \pm \sqrt{3}/6$, $b_{1,2} = 1/2$. The interpolation polynomial is of degree one and we have to apply (4.4) to get an approximation Y_{n+1}. Since we are interested in a fourth order approximation, we can neglect the remainder term (indicated by ... in (4.4)). Computing analytically the iterated integrals over products of $\ell_i(t)$ we obtain

$$Y_{n+1} = \exp\Big(\frac{h}{2}(A_1 + A_2) + \frac{\sqrt{3}\,h^2}{12}[A_2, A_1]\Big)\, Y_n, \tag{4.6}$$

where $A_1 = A(t_n + c_1 h)$ and $A_2 = A(t_n + c_2 h)$. This is a method of order four. The terms of (4.4) with triple integrals give $\mathcal{O}(h^4)$ expressions, whose leading term vanishes by the symmetry of the method (Exercise 6). Therefore, they need not be considered.

Remark. All numerical methods of this section are of the form $Y_{n+1} = \exp(h\Omega_n)\, Y_n$, where Ω_n is a linear combination of $A(t_n + c_i h)$ and of their commutators. If $A(t) \in \mathfrak{g}$ for all t, then also $h\Omega_n$ lies in the Lie algebra $\mathfrak{g}$, so that the numerical solution stays in the Lie group G if $Y_0 \in G$ (this is a consequence of Lemma II.4.3).

III.5 Convergence of methods on submanifolds

Consider a differential equation $\dot y = f(y)$ on a submanifold $\mathcal{M} \subset \mathbb{R}^m$, and a numerical integrator $y_{n+1} = \Phi_h(y_n)$ with a discrete flow map[5] $\Phi_h : \mathcal{M} \to \mathcal{M}$. If the vector field $f(y)$ and the integrator $\Phi_h(y)$ are defined and sufficiently smooth in a neighborhood of $\mathcal{M}$, then we can apply the well-established convergence results in the linear space $\mathbb{R}^m$. In particular, for a method of order p we have that the global error can be estimated as $\|y_n - y(nh)\| \le ch^p$ for $nh \le \widehat t$ with c depending on $\widehat t$, if the step size h is sufficiently small. Since the exact solution $y(t)$ stays on the submanifold $\mathcal{M}$, this implies that the numerical approximation stays $\mathcal{O}(h^p)$-close to the submanifold on compact time intervals $[0, \widehat t\,]$.

In the rest of this section we focus on the situation, where $f(y)$ and $\Phi_h(y)$ are only defined (and smooth) on the submanifold $\mathcal{M}$, and there is no natural smooth extension to a neighborhood of $\mathcal{M}$.

Definition 5.1 (local error and order of consistency)**.** *Let a differential equation $\dot y = f(y)$ with sufficiently smooth vector field be given on a submanifold $\mathcal{M} \subset \mathbb{R}^m$. A numerical integrator $\Phi_h : \mathcal{M} \to \mathcal{M}$ is of order p, if for every compact set $K \subset \mathcal{M}$ there exists $h_0 > 0$ such that for all h satisfying $0 < h \le h_0$ and for all $y_0 \in K$ the local truncation error can be estimated as*

$$\|\Phi_h(y_0) - y(h)\| \le C_0\, h^{p+1}.$$

Here, $y(t)$ denotes the solution of $\dot y = f(y)$ that satisfies $y(0) = y_0$, the norm is that of $\mathbb{R}^m$, and the constant C_0 depends on K but is independent of h.

[5]Typically, $\Phi_h(y)$ is defined implicitly by algebraic equations, and it is well defined only for sufficiently small $h \le h_0$ with h_0 depending on y. It may happen that there is no uniform $h_0 > 0$ such that $\Phi_h(y)$ exists for all $y \in \mathcal{M}$ and for $h \le h_0$. By abuse of notation, we nevertheless write $\Phi_h : \mathcal{M} \to \mathcal{M}$ in this situation.

Notice that the local error has to be estimated only for y_0 in the submanifold $\mathcal{M}$. This is usually much easier than estimating a suitable extension on an open neighborhood of $\mathcal{M}$. However, this makes sense only if $\Phi_h : \mathcal{M} \to \mathcal{M}$, which implies that the numerical solution stays for ever on the submanifold.

Theorem 5.2 (convergence)**.** *Consider a sufficiently smooth differential equation $\dot{y} = f(y)$ on a submanifold $\mathcal{M} \subset \mathbb{R}^m$, and an initial value $y_0 \in \mathcal{M}$ such that the solution $y(t) = \phi_t(y_0)$ exists for $0 \le t \le \widehat{t}$. If the numerical integrator $y_{n+1} = \Phi_h(y_n)$ is of order p and yields approximations satisfying $y_n \in \mathcal{M}$ for $nh \le \widehat{t}$, then there exists $h_0 > 0$ such that for $0 < h \le h_0$ the global error can be estimated as*

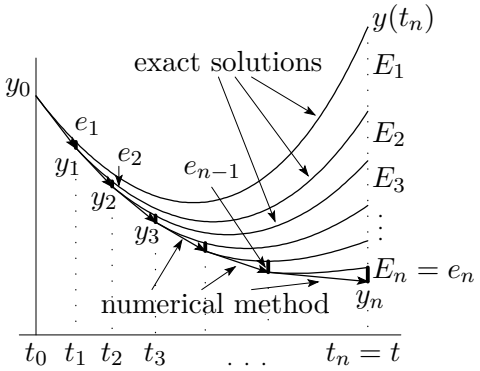
$$\|y_n - y(nh)\| \le c\, h^p \qquad \text{for} \quad nh \le \widehat{t}.$$

The constant c is independent on h, but depends on the length $\widehat{t}$ of the considered interval.

Proof. We consider the compact neighborhood

$$K = \{y \in \mathcal{M}\,;\, \exists\, \tau \in [0, \widehat{t}\,] \text{ with } \|y - y(\tau)\| \le \delta\}$$

of the solution, where $\delta > 0$ is given by Corollary II.3.4. As long as $y_n \in K$, it follows from Definition 5.1 that $\|y_{n+1} - \phi_h(y_n)\| \le C_0\, h^{p+1}$.



Assume for the moment that $y_n \in K_{nh}(\delta)$ and $\phi_h(y_{n-1}) \in K_{nh}(\delta)$ for $nh = t_n \le \widehat{t}$, where $K_\tau(\delta) = \{y \in \mathcal{M}\,;\, \|y - y(\tau)\| \le \delta\}$. Using $\phi_{t-t_n}(y_n) = \phi_{t-t_{n+1}}(\phi_h(y_n))$, Corollary II.3.4 then yields

$$\|\phi_{t-t_{n+1}}(y_{n+1}) - \phi_{t-t_n}(y_n)\| \le C\, \|y_{n+1} - \phi_h(y_n)\| \le C\, C_0\, h^{p+1}$$

for $t_{n+1} \le t \le \widehat{t}$. Summing up, we thus obtain for $nh = t_n \le \widehat{t}$

$$\|y_n - y(t_n)\| \le \sum_{j=0}^{n-1} \|\phi_{t_n - t_{j+1}}(y_{j+1}) - \phi_{t_n - t_j}(y_j)\| \le C\, C_0\, (nh) h^p,$$

which proves the statement with $c = C\, C_0\, \widehat{t}$. Our assumptions $y_n \in K_{nh}(\delta)$ and $\phi_h(y_{n-1}) \in K_{nh}(\delta)$ are justified a posteriori, if h is assumed to be sufficiently small.

In the figure, the local errors are $e_{j+1} = y_{j+1} - \phi_h(y_j)$, and the transported local errors are $E_{j+1} = \phi_{t_n - t_{j+1}}(y_{j+1}) - \phi_{t_n - t_j}(y_j)$. □

III.6 Exercises

1. For n-dimensional square matrices Y consider the function $g(Y) = \det Y$. Prove that

$$g'(Y)(HY) = \text{trace}H \det Y.$$

   Hint. Expand $\det(Y + \varepsilon HY)$ in powers of ε.

2. Elaborate Example 1.2 for the special case where Y is a matrix of dimension 2. In particular, show that (1.4) is the same as (1.2), and check the formulas for the simplified Newton iterations.

3. Show that for given $\widetilde{Y}$ the solution of the problem (1.5) is $Y = U^{\mathsf{T}}V$, where $\widetilde{Y} = U^{\mathsf{T}}\Sigma V$ is the singular value decomposition of $\widetilde{Y}$.
Hint. Since $\|U^{\mathsf{T}}SV\|_F = \|S\|_F$ holds for all orthogonal matrices U and V, it is sufficient to consider the case $\widetilde{Y} = (\Sigma, 0)^{\mathsf{T}}$ with $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$. Prove that

$$\|(\Sigma, 0)^{\mathsf{T}} - Y\|_F^2 \geq \sum_{i=1}^{k} (\sigma_i - 1)^2$$

for all matrices Y satisfying $Y^{\mathsf{T}}Y = I$.

4. *Rodrigues formula.* Prove that

$$\exp(\Omega) = I + \frac{\sin\alpha}{\alpha}\,\Omega + \frac{1}{2}\Big(\frac{\sin(\alpha/2)}{\alpha/2}\Big)^2 \Omega^2 \qquad \text{for} \qquad \Omega = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix}$$

where $\alpha = \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2}$. This formula allows for an efficient computation of the matrix exponential, if we are concerned with problems in the Lie proup O(3).

5. The solution of $\dot{Y} = A(Y)Y$, $Y(0) = Y_0$, is given by $Y(t) = \exp\big(\Omega(t)\big)Y_0$, where $\Omega(t)$ solves the differential equation

$$\dot{\Omega} = d\exp_\Omega^{-1}(A(Y(t))).$$

Prove that the first terms of the t-expansion of $\Omega(t)$ are given by

$$\begin{aligned} \Omega(t) &= tA(Y_0) + \frac{t^2}{2}A'(Y_0)A(Y_0)Y_0 + \frac{t^3}{6}\Big(A'(Y_0)^2A(Y_0)Y_0^2 + A'(Y_0)A(Y_0)^2Y_0 \\ &+ A''(Y_0)(A(Y_0)Y_0, A(Y_0)Y_0) - \frac{1}{2}[A(Y_0), A'(Y_0)A(Y_0)Y_0]\Big) + \ldots \end{aligned}$$

6. For the numerical solution of $\dot{Y} = A(t)\,Y$ consider the method $Y_n \mapsto Y_{n+1}$ defined by $Y_{n+1} = Z(t_n + h)$, where $Z(t)$ is the solution of

$$\dot{Z} = \widehat{A}(t)\,Z, \qquad Z(t_n) = Y_n,$$

and $\widehat{A}(t)$ is the interpolation polynomial based on symmetric nodes $c_1, \ldots, c_s$, i.e., we have $c_{s+1-i} + c_i = 1$ for all i.
a) Prove that this method is symmetric.
b) Show that $Y_{n+1} = \exp(\Omega(h))\,Y_n$ holds, where $\Omega(h)$ has an expansion in odd powers of h. This justifies the omission of the terms involving triple integrals in Example 4.4.

7. Consider the projection method of Algorithm 1.1, where Φ_h represents an explicit Runge-Kutta method of order p (e.g., the explicit Euler method) and the numerical approximation is obtained by orthogonal projection onto the submanifold. Prove that, for sufficiently small h, the projection method is of order p according to Definition 5.1.

Chapter IV

Differential-Algebraic Equations

The most general form of a differential-algebraic system is that of an implicit differential equation

$$F(\dot{u}, u) = 0 \tag{0.1}$$

where F and u have the same dimension. We always assume F to be sufficiently differentiable. A non-autonomous system is brought to the form (0.1) by appending t to the vector u, and by adding the equation $\dot{t} = 1$. If $\partial F/\partial \dot{u}$ is invertible we can locally solve (0.1) for $\dot{u}$ to obtain an ordinary differential equation. In this chapter we are interested in problems (0.1) where $\partial F/\partial \dot{u}$ is singular.[1]

IV.1 Linear equations with constant coefficients

The simplest and best understood problems of the form (0.1) are linear differential equations with constant coefficients

$$B\dot{u} + Au = d(t). \tag{1.1}$$

In looking for solutions of the form $u(t) = e^{\lambda t}u_0$ (if $d(t) \equiv 0$) we are led to consider the "matrix pencil" $A + \lambda B$. When $A + \lambda B$ is singular for all values of λ, then (1.1) has either no solution or infinitely many solutions for a given initial value (Exercise 1). We shall therefore deal only with *regular matrix pencils*, i.e., with problems where the polynomial $\det(A + \lambda B)$ does not vanish identically. The key to the solution of (1.1) is the following simultaneous transformation of A and B to canonical form.

Theorem 1.1 (Weierstrass 1868, Kronecker 1890)**.** *Let $A + \lambda B$ be a regular matrix pencil. Then there exist nonsingular matrices P and Q such that*

$$PAQ = \begin{pmatrix} C & 0 \\ 0 & I \end{pmatrix}, \qquad PBQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix} \tag{1.2}$$

where $N = \mathit{blockdiag}\,(N_1, \ldots, N_k)$, each N_i is of the form

$$N_i = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & 0 & 1 \\ 0 & & & 0 \end{pmatrix} \qquad \textit{of dimension } m_i\ , \tag{1.3}$$

and C can be assumed to be in Jordan canonical form.

[1] The text of this chapter is taken from Section VII.1 of the monograph *Solving Ordinary Differential Equations II* by Hairer and Wanner (1996).

Proof. (Gantmacher 1954 (Chap. XII), see also Exercises 3 and 4). We fix some c such that $A + cB$ is invertible. If we multiply

$$A + \lambda B = A + cB + (\lambda - c)B$$

by the inverse of $A + cB$ and then transform $(A + cB)^{-1}B$ to Jordan canonical form we obtain

$$\begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + (\lambda - c)\begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix}. \tag{1.4}$$

Here, J_1 contains the Jordan blocks with non-zero eigenvalues, J_2 those with zero eigenvalues (the dimension of J_1 is just the degree of the polynomial $\det(A+\lambda B)$). Consequently, J_1 and $I - cJ_2$ are both invertible and multiplying (1.4) from the left by $\operatorname{blockdiag}(J_1^{-1}, (I - cJ_2)^{-1})$ gives

$$\begin{pmatrix} J_1^{-1}(I - cJ_1) & 0 \\ 0 & I \end{pmatrix} + \lambda \begin{pmatrix} I & 0 \\ 0 & (I - cJ_2)^{-1}J_2 \end{pmatrix}.$$

The matrices $J_1^{-1}(I - cJ_1)$ and $(I - cJ_2)^{-1}J_2$ can then be brought to Jordan canonical form. Since all eigenvalues of $(I - cJ_2)^{-1}J_2$ are zero, we obtain the desired decomposition (1.2). □

Theorem 1.1 allows us to solve the differential-algebraic system (1.1) as follows: we premultiply (1.1) by P and use the transformation

$$u = Q\begin{pmatrix} y \\ z \end{pmatrix}, \qquad P\,d(t) = \begin{pmatrix} \eta(t) \\ \delta(t) \end{pmatrix}.$$

This decouples the differential-algebraic system (1.1) into

$$\dot{y} + Cy = \eta(t), \qquad N\dot{z} + z = \delta(t). \tag{1.5}$$

The equation for y is just an ordinary differential equation. The relation for z decouples again into k subsystems, each of the form (with $m = m_i$)

$$\begin{aligned} \dot{z}_2 + z_1 &= \delta_1(t) \\ &\vdots \\ \dot{z}_m + z_{m-1} &= \delta_{m-1}(t) \\ z_m &= \delta_m(t). \end{aligned} \tag{1.6}$$

Here z_m is determined by the last equation, and the other components are computed recursively by repeated differentiation. Exactly $m - 1$ differentiations are necessary to obtain

$$z_1(t) = \delta_1(t) - \dot{\delta}_2(t) + \ddot{\delta}_3(t) \mp \ldots + (-1)^{m-1}\delta_m^{(m-1)}(t). \tag{1.7}$$

The integer $(\max m_i)$ is called the *index of nilpotency* of the matrix pencil $A + \lambda B$. It does not depend on the particular transformation used to get (1.2) (see Exercise 5).

IV.2 Differentiation index

The previous example shows that certain equations of the differential-algebraic system (1.6) have to be differentiated $m-1$ times to get an explicit expression of all solution components. One more differentiation gives ordinary differential equations for all components. This motivates the following index definition for general nonlinear problems (Gear and Petzold 1983, 1984; Gear, Gupta, and Leimkuhler 1985, Gear 1990, Campbell and Gear 1995).

Definition 2.1. *Equation (0.1) has differentiation index m, if m is the minimal number of analytical differentiations*

$$F(\dot u, u) = 0, \quad \frac{\mathrm{d}F(\dot u, u)}{\mathrm{d}t} = 0\ , \quad \dots\ , \quad \frac{\mathrm{d}^m F(\dot u, u)}{\mathrm{d}t^m} = 0 \tag{2.1}$$

such that equations (2.1) allow us to extract by algebraic manipulations an explicit ordinary differential system $\dot u = a(u)$ (which is called the "underlying ODE").

Note that for linear equations with constant coefficients the differentiation index and the index of nilpotency are the same. Let us discuss the (differentiation) index for some important special cases.

Systems of index 1. Differential-algebraic systems of the form

$$\begin{aligned} \dot y &= f(y,z) \\ 0 &= g(y,z) \end{aligned} \tag{2.2}$$

have no occurrence of $\dot z$. We therefore differentiate the second equation of (2.2) to obtain

$$\dot z = -g_z^{-1}(y,z) g_y(y,z) f(y,z)$$

which is possible if g_z is invertible in a neighbourhood of the solution. The problem (2.2), for invertible g_z, is thus of differentiation index 1.

In practice, it is not necessary to know the differential equation for z. If initial values satisfy $g(y_0, z_0) = 0$ (we call them consistent) and if the matrix $g_z(y_0, z_0)$ is invertible, then the implicit function theorem guarantees the existence of a unique function $z = \zeta(y)$ (defined close to (y_0, z_0)) such that $g(y, \zeta(y)) = 0$. The problem then reduces locally to the ordinary differential equation $\dot y = f(y, \zeta(y))$, which can be solved by any numerical integrator.

Systems of index 2. In the system

$$\begin{aligned} \dot y &= f(y,z) \\ 0 &= g(y) \end{aligned} \tag{2.3}$$

where the variable z is absent in the algebraic constraint, we obtain by differentiation of the second relation of (2.3) the "hidden constraint"

$$0 = g_y(y) f(y,z). \tag{2.4}$$

If $g_y(y) f_z(y,z)$ is invertible in a neighbourhood of the solution, then the first equation of (2.3) together with (2.4) constitute an index 1 problem. Differentiation of (2.4) yields the missing differential equation for z, so that the problem (2.3) is of differentiation index 2.

If the initial values satisfy $0 = g(y_0)$ and $0 = g_y(y_0) f(y_0, z_0)$, we call them consistent. If in addition the matrix $g_y(y_0) f_z(y_0, z_0)$ is invertible, the implicit function theorem implies the local existence of a function $z = \zeta(y)$ satisfying $g_y(y) f(y, \zeta(y)) = 0$ in a neighborhood of y_0. We thus obtain the differential equation

$$\dot y = f(y, \zeta(y)) \qquad \text{on the manifold} \qquad \mathcal{M} = \{y\,;\, g(y) = 0\}.$$

The property $f(y, \zeta(y)) \in T_y\mathcal{M}$ follows from $g_y(y) f(y, \zeta(y)) = 0$. All numerical approaches of Chapter III can be applied to solve such problems.

System (2.3) is a representative of the larger class of problems of type (2.2) with *singular* g_z. If we assume that g_z has constant rank in a neighbourhood of the solution, we can eliminate certain algebraic variables from $0 = g(y, z)$ until the system is of the form (2.3). This can be done as follows: if there exists a pair (i, j) such that $\partial g_i/\partial z_j \neq 0$ at the initial value then, by the implicit function theorem, the relation $g_i(y, z) = 0$ permits us to express z_j in terms of y and the other components of z. We can thus eliminate the variable z_j from the system. Repeating this procedure we arrive at the situation, where g_z vanishes at the initial value. From the constant rank assumption it follows that g_z vanishes in a whole neighborhood of the initial value, so that g is already independent of z.

Systems of index 3. Problems of the form

$$\begin{aligned} \dot{y} &= f(y, z) \\ \dot{z} &= k(y, z, u) \\ 0 &= g(y) \end{aligned} \tag{2.5}$$

are of differentiation index 3, if

$$g_y(y) f_z(y, z) k_u(y, z, u) \qquad \text{is invertible} \tag{2.6}$$

in a neighborhood of the solution. To see this, we differentiate twice the algebraic relation of (2.5), which yields

$$0 = (g_y f)(y, z), \qquad 0 = (g_{yy}(f, f))(y, z) + (g_y f_y f)(y, z) + (g_y f_z k)(y, z, u). \tag{2.7}$$

A third differentiation permits to express $\dot{u}$ in terms of (y, z, u) provided that (2.6) is satisfied. This proves index 3 of the system (2.5).

Consistent inital values (y_0, z_0, u_0) must satisfy $g(y_0) = 0$ and the two conditions (2.7). Under the condition (2.6) an application of the implicit function theorem permits to express u in terms of (y, z) from the second relation of (2.7), i.e., $u = \nu(y, z)$. Inserting this relation into the differential-algebraic system (2.5) yields an ordinary differential equation for (y, z) on the manifold

$$\mathcal{M} = \{(y, z)\,;\, g(y) = 0,\ g_y(y) f(y, z) = 0\}.$$

The assumption (2.6) implies that $g_y(y)$ and $g_y(y) f_z(y, z)$ have full rank, so that $\mathcal{M}$ is a manifold. It follows from (2.7) that the vector field lies in the tangent space $T_{(y,z)}\mathcal{M}$ for all $(y, z) \in \mathcal{M}$.

IV.3 Control problems

Many problems of control theory lead to ordinary differential equations of the form

$$\dot{y} = f(y, u),$$

where u represents a set of controls. These controls must be applied so that the solution satisfies some constraints $0 = g(y)$ (or $0 = g(y, u)$). They often lead to a differential-algebraic system of index 2, as it is the case for the example of Section I.2.

Optimal control problems are differential equations $\dot{y} = f(y, u)$ formulated in such a way that the control $u(t)$ has to minimize some cost functional. The Euler–Lagrange equation then often becomes a differential-algebraic system (Pontryagin, Boltyanskij, Gamkrelidze

& Mishchenko 1961, Athans & Falb 1966, Campbell 1982). We demonstrate this on the problem

$$\dot{y} = f(y,u), \qquad y(0) = y_0 \tag{3.1}$$

with cost functional

$$J(u) = \int_0^1 \varphi\Big(y(t),u(t)\Big)\,\mathrm{d}t. \tag{3.2}$$

For a given function $u(t)$ the solution $y(t)$ is determined by (3.1). In order to find conditions for $u(t)$ that minimize $J(u)$ of (3.2), we consider the perturbed control $u(t)+\varepsilon\delta u(t)$ where $\delta u(t)$ is an arbitrary function and ε a small parameter. To this control there corresponds a solution $y(t)+\varepsilon\delta y(t)+\mathcal{O}(\varepsilon^2)$ of (3.1); we have (by comparing powers of ε)

$$\delta\dot{y}(t) = f_y(t)\delta y(t) + f_u(t)\delta u(t), \qquad \delta y(0) = 0,$$

where, as usual, $f_y(t) = f_y(y(t),u(t))$, etc. Linearization of (3.2) shows that

$$J(u+\varepsilon\delta u) - J(u) = \varepsilon\int_0^1 \Big(\varphi_y(t)\delta y(t) + \varphi_u(t)\delta u(t)\Big)\,\mathrm{d}t + \mathcal{O}(\varepsilon^2)$$

so that

$$\int_0^1 \Big(\varphi_y(t)\delta y(t) + \varphi_u(t)\delta u(t)\Big)\,\mathrm{d}t = 0 \tag{3.3}$$

is a necessary condition for $u(t)$ to be an optimal solution of our problem. In order to express δy in terms of δu in (3.3), we introduce the adjoint differential equation

$$\dot{v} = -f_y(t)^{\mathsf{T}} v - \varphi_y(t)^{\mathsf{T}}, \qquad v(1) = 0$$

with inhomogeneity $\varphi_y(t)^{\mathsf{T}}$. Hence we have (see Exercise 6)

$$\int_0^1 \varphi_y(t)\delta y(t)\,\mathrm{d}t = \int_0^1 v^{\mathsf{T}}(t) f_u(t)\delta u(t)\,\mathrm{d}t.$$

Inserted into (3.3) this gives the necessary condition

$$\int_0^1 \Big(v^{\mathsf{T}}(t) f_u(t) + \varphi_u(t)\Big)\delta u(t)\,\mathrm{d}t = 0.$$

Since this relation has to be satisfied for all δu we obtain the necessary relation

$$v^{\mathsf{T}}(t) f_u(t) + \varphi_u(t) = 0$$

by the so-called "fundamental lemma of variational calculus".

In summary, we have proved that a solution of the above optimal control problem has to satisfy the system

$$\begin{array}{rcll} \dot{y} &=& f(y,u), & y(0) = y_0 \\ \dot{v} &=& -f_y(y,u)^{\mathsf{T}} v - \varphi_y(y,u)^{\mathsf{T}}, & v(1) = 0 \\ 0 &=& v^{\mathsf{T}} f_u(y,u) + \varphi_u(y,u). & \end{array} \tag{3.4}$$

This is a boundary value differential-algebraic problem. It can also be obtained directly from the Pontryagin minimum principle (see Pontryagin et al. 1961, Athans and Falb 1966).

Differentiation of the algebraic relation in (3.4) shows that the system (3.4) has index 1 if the matrix

$$\sum_{i=1}^{n} v_i \frac{\partial^2 f_i}{\partial u^2}(y,u) + \frac{\partial^2 \varphi}{\partial u^2}(y,u)$$

is invertible along the solution. A situation where the system (3.4) has index 3 is presented in Exercise 7.

IV.4 Mechanical systems

An interesting class of differential-algebraic systems appears in mechanical modeling of constrained systems. A choice method for deriving the equations of motion of mechanical systems is the Lagrange-Hamilton principle, whose long history goes back to merely theological ideas of Leibniz and Maupertuis.

Mechanical systems in minimal coordinates. Let $q = (q_1, \ldots, q_n)^\mathsf{T}$ be minimal[2] generalized coordinates of a system and $v_i = \dot{q}_i$ the velocities. Suppose a function $L(q, \dot{q})$ is given; then the Euler equations of the variational problem

$$\int_{t_1}^{t_2} L(q, \dot{q})\, \mathrm{d}t = \min!$$

are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0, \qquad k = 1, \ldots, n, \tag{4.1}$$

which represent a second order differential equations for the coordinates q_k. The great discovery of Lagrange (1788) is that for $L = T - U$, where $T(q, \dot{q}) = \frac{1}{2}\dot{q}^\mathsf{T} M(q)\, \dot{q}$ (with a symmetric positive matrix $M(q)$) is the *kinetic energy* and $U(q)$ the *potential energy*, the differential equation (4.1) describes the movement of the corresponding "conservative system". Written as a first order differential equation, it is given by

$$\begin{aligned} \dot{q} &= v \\ M(q)\, \dot{v} &= f(q, v), \end{aligned} \tag{4.2}$$

where $f(q, v) = -\frac{\partial}{\partial q}(M(q)v)v + \nabla_q T(q, v) - \nabla_q U(q)$. For the important special case where M is constant, we simply have $f(q, v) = -\nabla_q U(q)$.

Example 4.1. The mathematical pendulum of length ℓ has one degree of freedom. We choose as generalized coordinate the angle $\alpha = q_1$ such that $T = m\,\ell^2\dot{\alpha}^2/2$ and $U = -\ell m g \cos\alpha$. Then (4.1) becomes $\ell\,\ddot{\alpha} = -g \sin\alpha$, the well-known pendulum equation.

Constrained mechanical systems. Suppose now that the generalized coordinates $q = (q_1, \ldots, q_n)^\mathsf{T}$ are constrained by the relations $g_1(q) = 0, \ldots, g_m(q) = 0$ (or shortly $g(q) = 0$) on their movement. If these relations are independent (we assume that $g'(q)$ has full rank m) the number of degrees of freedom is $n - m$. An example is the mathematical pendulum considered in Cartesian coordinates. We again assume that the kinetic energy is given by $T(q, \dot{q}) = \frac{1}{2}\dot{q}^\mathsf{T} M(q)\, \dot{q}$ with a symmetric positive matrix $M(q)$, and the potential energy is $U(q)$. To obtain the equations of motion we proceed in three steps:

- we introduce minimal coordinates of the system, i.e., a parametrization $q = \eta(z)$ of the submanifold $\mathcal{N} = \{q\,;\, g(q) = 0\}$,
- we write down the equations of motion in minimal coordinates z, and
- we rewrite these equations in the original variables q.

Using our parametrization $q = \eta(z)$ and its time derivative $\dot{q} = \eta'(z)\dot{z}$, the kinetic and potential energies become

$$\widehat{T}(z, \dot{z}) = T(\eta(z), \eta'(z)\dot{z}) = \frac{1}{2}\, \dot{z}^\mathsf{T} \widehat{M}(z)\, \dot{z} \qquad \text{with} \qquad \widehat{M}(z) = \eta'(z)^\mathsf{T} M(\eta(z))\, \eta'(z)$$

[2] Minimal means that the dimension of q equals the number of degrees of freedom in the system.

and $\widehat{U}(z) = U(\eta(z))$. With the Lagrangian $\widehat{L}(z,\dot z) = L(\eta(z), \eta'(z)\dot z) = \widehat{T}(z,\dot z) - \widehat{U}(z)$ the equations of motion, written in minimal coordinates z, are therefore

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial \widehat{L}}{\partial \dot z}(z,\dot z)\Big) - \frac{\partial \widehat{L}}{\partial z}(z,\dot z) = 0. \tag{4.3}$$

We have to rewrite these equations in the original variables q. Using the relations

$$\begin{aligned}
\frac{\partial \widehat{L}}{\partial \dot z}(z,\dot z) &= \frac{\partial L}{\partial \dot q}(q,\dot q)\,\eta'(z) \\
\frac{\partial \widehat{L}}{\partial z}(z,\dot z) &= \frac{\partial L}{\partial q}(q,\dot q)\,\eta'(z) + \frac{\partial L}{\partial \dot q}(q,\dot q)\,\eta''(z)(\dot z,\cdot) \\
\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial \widehat{L}}{\partial \dot z}(z,\dot z)\Big) &= \frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial L}{\partial \dot q}(q,\dot q)\Big)\eta'(z) + \frac{\partial L}{\partial \dot q}(q,\dot q)\,\eta''(z)(\dot z,\cdot)
\end{aligned}$$

the equations (4.3) become

$$\Big(\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial L}{\partial \dot q}(q,\dot q)\Big) - \frac{\partial L}{\partial q}(q,\dot q)\Big)\eta'(z) = 0. \tag{4.4}$$

Any vector w satisfying $w^{\mathsf T}\eta'(z) = 0$ is orthogonal to the image $\operatorname{Im}\eta'(z)$. However, from the characterization of the tangent space (Theorem II.2.2) we know that $\operatorname{Im}\eta'(z) = T_q\mathcal{N} = \ker g'(q)$. Using the identity $(\ker g'(q))^\perp = \operatorname{Im} g'(q)^{\mathsf T}$, we obtain that the equation (4.4) is equivalent to

$$\Big(\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial L}{\partial \dot q}(q,\dot q)\Big) - \frac{\partial L}{\partial q}(q,\dot q)\Big)^{\mathsf T} = -g'(q)^{\mathsf T}\lambda$$

which can also be written as

$$\begin{aligned}
\dot q &= v \\
M(q)\,\dot v &= f(q,v) - G(q)^{\mathsf T}\lambda \\
0 &= g(q),
\end{aligned} \tag{4.5}$$

where we denote $G(q) = g'(q)$, and $f(q,v)$ is as in (4.2). For the mathematical pendulum, written in Cartesian coordinates, these equations have been considered in Example I.3.1. Various formulations are possible for such a problem, each of which leads to a different numerical approach.

Index 3 Formulation (position level, descriptor form). If we formally multiply the second equation of (4.5) by $M(q)^{-1}$, the system (4.5) becomes of the form (2.5) with (q, v, λ) in the roles of (y, z, u). The condition (2.6), written out for (4.5), is

$$G(q)M(q)^{-1}G(q)^{\mathsf T} \qquad \text{is invertible .} \tag{4.6}$$

This is satisfied, if the rows of the matrix $G(q)$ are linearly independent, i.e., the constraints $g(q) = 0$ are independent. Under this assumption, the system (4.5) is an index 3 problem.

Index 2 Formulation (velocity level). Differentiation of the algebraic relation in (4.5) gives

$$0 = G(q)v. \tag{4.7}$$

If we replace the algebraic relation in (4.5) by (4.7), we obtain a system of the form (2.3) with (q, u) in the role of y and λ in that of z. One verifies that because of (4.6) the first two equations of (4.5) together with (4.7) represent a problem of index 2.

Index 1 Formulation (acceleration level). If we differentiate twice the constraint in (4.5), the resulting equation together with the second equation of (4.5) yield

$$\begin{pmatrix} M(q) & G^T(q) \\ G(q) & 0 \end{pmatrix} \begin{pmatrix} \dot v \\ \lambda \end{pmatrix} = \begin{pmatrix} f(q,v) \\ -g_{qq}(q)(v,v) \end{pmatrix}. \tag{4.8}$$

This allows us to express $\dot v$ and λ as functions of q, v, provided that the matrix in (4.8) is invertible (see Exercise I.6). Hence, the first equation of (4.5) together with (4.8) consitute an index 1 problem.

All these formulations are mathematically equivalent, if the initial values are consistent, i.e., if (q_0, v_0) satisfy $g(q_0) = 0$ and $g'(q_0)v_0 = 0$, and if $\lambda_0 = \lambda(q_0, v_0)$ where the function $\lambda(q, v)$ is defined by (4.8). However, if for example the index 1 or the index 2 system is integrated numerically, the constraints of the original problem will no longer be exactly satisfied. It is recommended to consider the problem as a differential equation on the manifold, and to force the solution to remain on the manifold.

Constrained mechanical system as differential equation on a manifold. Inserting the function $\lambda(q, v)$ obtained from (4.8) into the system (4.5), the first two equations of (4.5) represent an ordinary differential equation on the submanifold

$$\mathcal{M} = \{(q, v)\,;\, g(q) = 0,\ g'(q)v = 0\}.$$

This is equivalent to the index 1 formulation. Applying the numerical techniques of Chapter III (projection methods and local state space form approaches) to the problem, one has to be careful that the numerical solution not only satisfies the given constraint $g(q) = 0$, but also the hidden constraint $g'(q)v = 0$.

IV.5 Exercises

1. Compute the general solution of the linear differential-algebraic equation

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} \dot u_1 \\ \dot u_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} t^2 \\ 3\,t \end{pmatrix}.$$

   Is there a solution for every pair of initial values $(u_1(0), u_2(0))$?

2. Prove that the initial value problem

$$B\dot u + Au = 0, \qquad u(0) = 0$$

   has a unique solution if and only if the matrix pencil $A + \lambda B$ is regular.

   Hint for the "only if" part. If n is the dimension of u, choose arbitrarily $n + 1$ distinct λ_i and vectors $v_i \neq 0$ satisfying $(A + \lambda_i B)v_i = 0$. Then take a linear combination, such that $\sum \alpha_i v_i = 0$, but $\sum \alpha_i e^{\lambda_i x} v_i \not\equiv 0$.

3. (Stewart 1972). Let $A + \lambda B$ be a regular matrix pencil. Show that there exist unitary matrices Q and Z such that

$$QAZ = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \qquad QBZ = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix} \tag{5.1}$$

   are both triangular. Furthermore, the submatrices A_{22} and B_{11} are invertible, and the diagonal elements of B_{22} are all 0.

Hint (compare with the Schur decomposition of a matrix). Let λ_1 be a zero of $\det(A+\lambda B)$ and $v_1 \neq 0$ be such that $(A+\lambda_1 B)v_1 = 0$. Verify that $Bv_1 \neq 0$ and that

$$AZ_1 = Q_1 \begin{pmatrix} -\lambda_1\beta & * \\ 0 & \widetilde{A} \end{pmatrix}, \qquad BZ_1 = Q_1 \begin{pmatrix} \beta & * \\ 0 & \widetilde{B} \end{pmatrix}$$

where $\beta = \|Bv_1\|/\|v_1\|$, and Q_1, Z_1 are unitary matrices (orthogonal if λ_1 is real) whose first columns are scalar multiples of Bv_1 and v_1, respectively. The matrix pencil $\widetilde{A}+\lambda\widetilde{B}$ is again regular and this procedure can be continued until $\det(\widetilde{A}+\lambda\widetilde{B}) = \mathit{Const}$ which implies that $\det \widetilde{B} = 0$. In this case we take a vector $v_2 \neq 0$ such that $\widetilde{B}v_2 = 0$ and transform $\widetilde{A}+\lambda\widetilde{B}$ with unitary matrices Q_2, Z_2, whose first columns are $\widetilde{A}v_2$ and v_2, respectively. For a practical computation of the decomposition (5.1) see the monograph of Golub and Van Loan (1989), Section 7.7.

4. Under the assumptions of Exercise 3 show that there exist matrices S and T such that

$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix},$$
$$\begin{pmatrix} I & S \\ 0 & I \end{pmatrix}\begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}\begin{pmatrix} I & T \\ 0 & I \end{pmatrix} = \begin{pmatrix} B_{11} & 0 \\ 0 & B_{22} \end{pmatrix}.$$

Hint. These matrices have to satisfy

$$A_{11}T + A_{12} + SA_{22} = 0 \tag{5.2}$$
$$B_{11}T + B_{12} + SB_{22} = 0 \tag{5.3}$$

and can be computed as follows: the first column of T is obtained from (5.3) because B_{11} is invertible and the first column of SB_{22} vanishes; then the first column of S is given by (5.2) because A_{22} is invertible; the second column of SB_{22} is then known and we can compute the second column of T from (5.3), etc.

5. Prove that the index of nilpotency of a regular matrix pencil $A+\lambda B$ does not depend on the choice of P and Q in (1.2).

   Hint. Consider two different decompositions of the form (1.2) and denote the matrices which appear by C_1, N_1 and C_2, N_2, respectively. Show the existence of a regular matrix T such that $N_2 = T^{-1}N_1T$.

6. For the linear initial value problem

$$\dot{y} = A(t)y + f(t), \qquad y(0) = 0$$

consider the *adjoint* problem

$$\dot{v} = -A(t)^{\mathsf{T}}v - g(t), \qquad v(1) = 0.$$

Prove that

$$\int_0^1 g(t)^{\mathsf{T}}y(t)\,\mathrm{d}t = \int_0^1 v(t)^{\mathsf{T}}f(t)\,\mathrm{d}t.$$

7. Consider a linear optimal control problem with quadratic cost functional

$$\begin{aligned} \dot{y} &= Ay + Bu + f(t), & y(0) &= y_0 \\ J(u) &= \frac{1}{2}\int_0^1 \Big(y(t)^{\mathsf{T}}Cy(t) + u(t)^{\mathsf{T}}Du(t)\Big)\mathrm{d}t, \end{aligned}$$

where C and D are symmetric, positive semi-definite matrices.

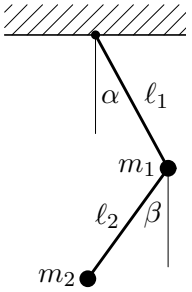
a) Prove that $J(u)$ is minimal if and only if

$$\begin{aligned} \dot{y} &= Ay + Bu + f(t), & y(0) &= y_0 \\ \dot{v} &= -A^{\mathsf{T}} v - Cy, & v(1) &= 0 \\ 0 &= B^{\mathsf{T}} v + Du. \end{aligned} \tag{5.4}$$

b) If D is positive definite, then (5.4) has index 1.

c) If $D = 0$ and $B^{\mathsf{T}}CB$ is positive definite, then (5.4) has index 3.

8. Consider the double pendulum in the configuration of the small figure to the right. The pendulum is fixed at the origin, the two mass points have coordinates (x_1, y_1) and (x_2, y_2). The kinetic and potential energies are given by



$$\begin{aligned} T &= \frac{m_1}{2}(\dot{x}_1^2 + \dot{y}_1^2) + \frac{m_2}{2}(\dot{x}_2^2 + \dot{y}_2^2) \\ U &= m_1 g\, y_1 + m_2 g\, y_2. \end{aligned}$$

a) Determine the constraints and give the descriptor form (differential-algebraic equation of index 3) of the equations of motion for the mechanical system in Cartesian coordinates.

b) Let α and β be the generalized coordinates of the double pendulum. Write the equations of motion in terms of these minimal coordinates.

Chapter V

Numerical Methods for DAEs

We have seen in Chapter IV how differential-algebraic equations (DAEs) can be interpreted as differential equations on manifolds. Therefore, all numerical approaches (projection methods and integrators based on local coordinates) discussed in Chapter III can be applied to solve these problems. Here, we consider direct numerical methods for problems of the form

$$M\,\dot u = F(u), \quad u(0) = u_0, \tag{0.1}$$

where M is a constant, but possibly singular matrix, and the initial value is such that the problem possesses a unique solution. For this it is necessary that $F(u_0)$ lies in the range of the matrix M. All problems of Chapter IV can be written in this form. For the problems of the form (IV.2.2) or (IV.2.3) the matrix M is diagonal, with entries 1 in the first part, and entries 0 in the rest. For an implicit differential equation $F_0(\dot v, v) = 0$, we can introduce a new variable for the derivative and thus obtain the system $\dot v = w$, $F_0(w, v) = 0$, which is of the form (0.1) for the vector $u = (v, w)$.

In the first sections of the present chapter, we consider a numerical approach which requires only the knowledge of the data M and $F(u)$ of the problem, and not that of the underlying manifold of the DAE. It can be sketched as follows:

- apply formally any numerical method to the differential equation $\dot u = M^{-1}F(u)$,
- rewrite the formulas in such a way that the inverse of M is no longer present,
- investigate whether the resulting numerical scheme makes sense for singular M.

Whereas the definition of the numerical schemes, following this approach, is extremely simple, their analysis (local accuracy, stability, and convergence) needs more effort.

V.1 Runge–Kutta and multistep methods

Let us start with applying the above approach to the explicit and implicit Euler methods. For the explicit Euler method we obtain

$$u_{n+1} = u_n + h\,M^{-1}F(u_n) \qquad \text{or} \qquad M(u_{n+1} - u_n) = h\,F(u_n).$$

If the matrix M is singular, this relation does not permit us to compute u_{n+1} for a given u_n, and the above approach does not lead to a numerical approximation. The implicit Euler method yields

$$M(u_{n+1} - u_n) = h\,F(u_{n+1}), \tag{1.1}$$

which represents a nonlinear system for u_{n+1}. Application of simplified Newton iterations requires the solution of linear equations with the matrix

$$M - h\,F'(u_n). \tag{1.2}$$

If the matrix pencil, formed by the matrices M and $F'(u_n)$ is regular, then the matrix (1.2) is invertible for sufficiently small step size h, and simplified Newton iterations are feasible. We shall study in the next sections, when the solution u_{n+1} of (1.1) exists, so that the implicit Euler method is well defined for small h.

Linear multistep methods. Applying a multistep formula to the system $\dot u = M^{-1}F(u)$ and multiplying the relation with M yields (notice that $\alpha_k \neq 0$)

$$M \sum_{j=0}^{k} \alpha_j\, u_{n+j} = h \sum_{j=0}^{k} \beta_j\, F(u_{n+j}). \tag{1.3}$$

If the method is explicit, i.e., $\beta_k = 0$, this relation does not permit the computation of u_{n+k} when M is singular. Therefore, only implicit methods make sense in this context. As for the implicit Euler method, an application of simplified Newton iterations leads to linear systems with the matrix

$$\alpha_k\, M - h\,\beta_k\, F'(u_{n+k-1}).$$

This again requires the matrix pencil formed by M and $F'(u_{n+k-1})$ to be regular.

Runge–Kutta methods. Using this approach with Runge–Kutta methods as numerical integrator leads to the system

$$\begin{aligned} M(U_{ni} - u_n) &= h \sum_{j=1}^{s} a_{ij}\, F(U_{nj}), \qquad i = 1, \dots, s \\ M(u_{n+1} - u_n) &= h \sum_{i=1}^{s} b_i\, F(U_{ni}). \end{aligned} \tag{1.4}$$

Consider first the upper relation of (1.4), which is supposed to define the internal stages U_{ni} for $i = 1, \dots, s$. Applying simplified Newton iterations yields linear systems with the matrix[1]

$$I \otimes M - h\,A \otimes F'(u_n). \tag{1.5}$$

Suppose that the invertible matrix T is such that $T^{-1}AT$ is upper triangular with the eigenvalues λ_i of A on the diagonal. The matrix $T \otimes I$ then transforms (1.5) to block upper triangular form with diagonal blocks of the form $M - h\lambda_i F'(u_n)$. If the matrix pencil formed by M and $F'(u_n)$ is regular, and if $\lambda_i \neq 0$ for all i (which means that A is non-singular) then the matrix (1.5) is invertible for sufficiently small h, and simplified Newton iterations can be performed.

Assume for the moment that the system (1.4) has a (locally) unique solution $U_{n1}, \dots, U_{ns}$. The right-hand side of the lower relation of (1.4) is then determined, and it seems hopeless to get a unique approximation u_{n+1} when M is singular. However, if the Runge–Kutta matrix $A = (a_{ij})_{i,j=1}^{s}$ is invertible, we can compute the vector $(F(U_{n1}), \dots, F(U_{ns}))$ from the upper part of (1.4) and insert it into the lower part. This gives

$$M(u_{n+1} - u_n) = \sum_{i=1}^{s} b_i \sum_{j=1}^{s} w_{ij}\, M(U_{nj} - u_n),$$

[1] For two matrices A and B, the tensor product is defined as $A \otimes B = (a_{ij}B)_{i,j=1}^{s}$.

where w_{ij} are the entries of the inverse A^{-1} of the Runge–Kutta matrix. As long as M is invertible, we can simplify this relation by M and thus obtain

$$u_{n+1} - u_n = \sum_{j=1}^{s} \Big(\sum_{i=1}^{s} b_i w_{ij} \Big) (U_{nj} - u_n). \tag{1.6}$$

For invertible M, the complete system (1.4) is therefore equivalent to the system, where the lower relation of (1.4) is replaced with (1.6). This formulation is perfectly adapted to the solution of problems (0.1) with singular M.[2]

Invariance with respect to linear transformations. In many situations (either for theoretical investigations or for practical issues like step size selection) it is convenient to have a very simple form of the matrix M in (0.1). We can always decompose the matrix M (e.g., by Gaussian elimination with total pivoting) as

$$M = S \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} T, \tag{1.7}$$

where S and T are invertible matrices and the dimension of I represents the rank of M. Inserting this into (0.1), multiplying by S^{-1}, and using the transformed quantities

$$Tu = \begin{pmatrix} y \\ z \end{pmatrix}, \qquad S^{-1} F(u) = S^{-1} F\Big(T^{-1} \begin{pmatrix} y \\ z \end{pmatrix} \Big) = \begin{pmatrix} f(y,z) \\ g(y,z) \end{pmatrix}, \tag{1.8}$$

gives

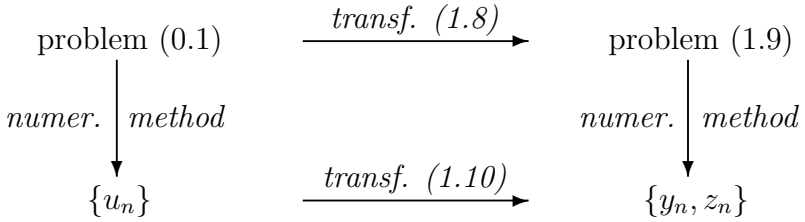
$$\begin{aligned} \dot{y} &= f(y,z) \\ 0 &= g(y,z), \end{aligned} \tag{1.9}$$

a problem that has been extensively studied in Chapter IV. At the moment we do not make any assumption on the index of the problem.

It is interesting to note that all numerical methods considered in this section are invariant with respect to this transformation. If we consider transformed variables

$$Tu_n = \begin{pmatrix} y_n \\ z_n \end{pmatrix}, \qquad TU_{ni} = \begin{pmatrix} Y_{ni} \\ Z_{ni} \end{pmatrix} \tag{1.10}$$

also for the numerical solution, this means that the diagram



commutes. An important consequence of this commutativity is that all results (existence of a numerical solution, convergence, asymptotic expansions, ...) for semi-explicit systems (1.9) and the approach of this section also apply to differential-algebraic equations (0.1) with singular M.

[2]By the way, the use of (1.6) is recommended for an implementation of implicit Runge–Kutta methods.

V.2 Index 1 problems

We consider the differential-algebraic equation

$$\begin{aligned} \dot y &= f(y,z), \qquad y(0)=y_0 \\ 0 &= g(y,z), \qquad z(0)=z_0 \end{aligned} \tag{2.1}$$

with initial values satisfying $g(y_0,z_0)=0$. In this section we assume that g_z is invertible along the solution, so that the problem is of differentiation index 1. As discussed in Section IV.2, the algebraic equation of (2.1) can then be solved for z and yields and equivalent relation $z=\zeta(y)$. In this section we study the accuracy and convergence of multistep methods as well as Runge–Kutta methods.

Linear multistep methods. For the problem (2.1), a linear multistep method applied in the form (1.3) reads

$$\sum_{j=0}^{k} \alpha_j\, y_{n+j} = h \sum_{j=0}^{k} \beta_j\, f(y_{n+j},z_{n+j}), \qquad 0 = \sum_{j=0}^{k} \beta_j\, g(y_{n+j},z_{n+j}). \tag{2.2}$$

If the starting approximations are consistent, i.e., $g(y_j,z_j)=0$ for $j=0,1,\dots,k-1$, and if $\beta_k\neq 0$, then the second relation of (2.2) is equivalent to $g(y_{n+k},z_{n+k})=0$ for all n. This means that $z_{n+k}=\zeta(y_{n+k})$ for all n, and the numerical approximation for the y component is the result of the multistep method applied to the ordinary differential equation $\dot y = f(y,\zeta(y))$. Classical convergence theory thus tells us that the global error in the y component is of the size $\mathcal{O}(h^p)$, if p is the classical order of the method. It follows from $z_n=\zeta(y_n)$ and $z(nh)=\zeta(y(nh))$ that the same bounds hold also for the z component.

In practice, for example when the relation $g(y_{n+k},z_{n+k})=0$ is not explicitly used (e.g., when the formulation with a general matrix M is used), then (due to errors in the starting approximations or due to round-off errors) the difference equation (2.2) for $g_{n+j}=g(y_{n+j},z_{n+j})$ has to be stable. This means that all zeros of the characteristic polynomial $\sigma(\zeta)=\sum_{j=0}^{k}\beta_j\zeta^j$ have to lie in the unit circle, and those with modulus one have to be simple.

Runge–Kutta methods. Applying an implicit Runge–Kutta method (1.4) with invertible Runge–Kutta matrix A to the system (2.1) yields

$$\begin{aligned} Y_{ni}-y_n &= h\sum_{j=1}^{s} a_{ij}\, f(Y_{nj},Z_{nj}), \qquad i=1,\dots,s \\ 0 &= g(Y_{ni},Z_{ni}), \qquad i=1,\dots,s \\ y_{n+1}-y_n &= h\sum_{i=1}^{s} b_i\, f(Y_{ni},Z_{ni}) \\ z_{n+1}-z_n &= \sum_{j=1}^{s}\Big(\sum_{i=1}^{s} b_i w_{ij}\Big)(Z_{nj}-z_n). \end{aligned} \tag{2.3}$$

The second relation shows that the internal stages satisfy $Z_{ni}=\zeta(Y_{ni})$ for $i=1,\dots,s$. Consequently, the y component is precisely the same as if we apply the Runge–Kutta method to the ordinary differential equation $\dot y = f(y,\zeta(y))$. Classical convergence results therefore yield $y_n - y(nh) = \mathcal{O}(h^p)$ on compact intervals $0\le nh\le T$, where p denotes the order of the method.

If the method is *stiffly accurate*, i.e., the Runge–Kutta coefficients satisfy $a_{sj}=b_j$ for all j, then we have $y_{n+1}=Y_{ns}$. Moreover, the Runge–Kutta coefficients satisfy $\sum_{i=1}^{s} b_i w_{ij}=0$ for

$j = 1, \dots, s-1$, and $\sum_{i=1}^{s} b_i w_{is} = 1$. Consequently, we have $z_{n+1} = Z_{ns}$ and thus also $z_{n+1} = \zeta(y_{n+1})$. The convergence estimate for the y component therefore implies $z_n - z(nh) = \mathcal{O}(h^p)$ on compact intervals $0 \le nh \le T$.

For methods that are not stiffly accurate, the so-called *stage order* plays an important role. One says that a Runge–Kutta method has stage order q, if the coefficients satisfy the simplifying condition

$$\mathrm{C}(q): \qquad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \dots, s, \quad k = 1, \dots q. \tag{2.4}$$

This is equivalent to $\sum_{j=1}^{s} a_{ij}\, p(c_j) = \int_0^{c_i} p(\tau)\, \mathrm{d}\tau$ for polynomials $p(\tau)$ of degree $\le q-1$, and means that the quadrature rules for the internal stages have an order at least q. For the z component we have the following convergence result.

Theorem 2.1 (order reduction[3])**.** *Consider the system (2.1) with initial values satisfying $g(y_0, z_0) = 0$, and assume that g_z is invertible in a neighborhood of the exact solution $(y(t), z(t))$. Let the Runge-Kutta method be of order p, of stage order q, with invertible matrix A, and denote[4] $\rho = 1 - \sum_{j=1}^{s} \sum_{i=1}^{s} b_i w_{ij}$. Then the numerical solution of (2.3) has global error satisfying*

$$z_n - z(nh) = \mathcal{O}(h^r) \quad \text{for} \quad t_n = nh \le T,$$

where

a) $r = p$ *for stiffly accurate methods,*

b) $r = \min(p, q+1)$ *if the stability function satisfies* $-1 \le \rho < 1$,

c) $r = \min(p-1, q)$ *if* $\rho = +1$.

d) If $|\rho| > 1$, *the numerical solution diverges.*

Proof. Part (a) has already been discussed. For the remaining cases we proceed as follows: we first observe that condition $C(q)$ and order p imply

$$z(t_n + c_i h) \;=\; z(t_n) + h \sum_{j=1}^{s} a_{ij}\, \dot z(t_n + c_j h) + \mathcal{O}(h^{q+1}) \tag{2.5}$$

$$z(t_{n+1}) \;=\; z(t_n) + h \sum_{i=1}^{s} b_i\, \dot z(t_n + c_i h) + \mathcal{O}(h^{p+1}). \tag{2.6}$$

Since A is invertible we can compute $\dot z(t_n + c_j h)$ from (2.5) and insert it into (2.6). This gives

$$z(t_{n+1}) = \rho\, z(t_n) + b^T A^{-1} \widehat{Z}_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}), \tag{2.7}$$

where $\widehat{Z}_n = (z(t_n + c_1 h), \dots, z(t_n + c_s h))^{\mathsf{T}}$. We then denote the global error by $\Delta z_n = z_n - z(t_n)$, and $\Delta Z_n = Z_n - \widehat{Z}_n$, where $Z_n = (Z_{n1}, \dots, Z_{ns})^{\mathsf{T}}$. Subtracting (2.7) from the last relation of (2.3) yields

$$\Delta z_{n+1} = \rho\, \Delta z_n + b^T A^{-1} \Delta Z_n + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}). \tag{2.8}$$

Our next aim is to estimate ΔZ_n. For this we have to consider the y component of the system. By definition of the method, the values y_n, Y_{ni} are those of the Runge-Kutta method

[3]This order reduction in the z component was first studied in a more general context by L.R. Petzold, *Order results for implicit Runge–Kutta methods applied to differential/algebraic systems.* SIAM J. Numer. Anal. 23 (1986) 837–852.

[4]The expression ρ equals the value at infinity of the stability function.

applied to $\dot y = f(y, \zeta(y))$. It thus follows from the classical convergence theory for ordinary differential equations that $y_n - y(t_n) = e_p(t_n)h^p + \mathcal{O}(h^{p+1})$. Since equation (2.5) also holds with $z(t)$ replaced by $y(t)$, we can subtract this formula from the first relation of (2.3) and so obtain

$$\begin{aligned} Y_{ni} - y(t_n + c_i h) &= y_n - y(t_n) \\ &+ h\sum_{j=1}^{s} a_{ij}\Big(f\Big(Y_{nj}, \zeta(Y_{nj})\Big) - f\Big(y(t_n + c_j h), \zeta(y(t_n + c_j h))\Big)\Big) + \mathcal{O}(h^{q+1}). \end{aligned}$$

Since $y_n - y(t_n) = \mathcal{O}(h^p)$, this implies that

$$Y_{ni} - y(t_n + c_i h) = \mathcal{O}(h^\nu) \qquad \text{with} \quad \nu = \min(p, q+1).$$

By the second relation of (2.3) we have $Z_{ni} - z(t_n + c_i h) = \zeta(Y_{ni}) - \zeta(y(t_n + c_i h)) = \mathcal{O}(h^\nu)$, and equation (2.8) becomes

$$\Delta z_{n+1} = \rho\, \Delta z_n + \delta_{n+1}, \qquad \text{where} \quad \delta_{n+1} = \mathcal{O}(h^\nu).$$

Repeated insertion of this formula gives

$$\Delta z_n = \sum_{i=1}^{n} \rho^{n-i}\, \delta_i,$$

because $\Delta z_0 = 0$. This proves the statement for $\rho \neq -1$. For the case $\rho = -1$ the error Δz_n is a sum of differences $\delta_{j+1} - \delta_j$. Since δ_{n+1} is actually of the form $\delta_{n+1} = d(t_n)h^\nu + \mathcal{O}(h^{\nu+1})$ we have $\delta_{j+1} - \delta_j = \mathcal{O}(h^{\nu+1})$ and the statement also follows in this situation. □

Example 2.2 (Radau IIA methods)**.** One of the most important integrators for the numerical solution of differential-algebraic equations are the so-called Radau IIA methods. The nodes $c_1, \ldots, c_s$ are the zeros of

$$\frac{\mathrm{d}^{s-1}}{\mathrm{d}x^{s-1}}\Big(x^{s-1}(x-1)^s\Big),$$

and the weights $b_1, \ldots, b_s$ are chosen such that the quadrature formula is interpolatory, which implies that it is of order $p = 2s - 1$. Ehle (1969) and Axelsson (1969) independently proposed to consider coefficients a_{ij} by imposing condition $C(s)$ of (2.4). The special case for $s = 1$ is nothing other than the implicit Euler method. The coefficients (matrix a_{ij} together with the c_i in the left column and the b_j in the bottom row) are given in Table V.1 for the cases $s = 2$ and $s = 3$.

The methods have classical order $p = 2s-1$, stage order $q = s$, the Runge–Kutta matrix is invertible, and the weights satisfy $b_j = a_{sj}$ for all j. For more details we refer to Section IV.5 of the monograph *Solving Ordinary Differential Equations II* by Hairer and Wanner.

Tab. V.1: Radau IIA methods of order 3 and 5

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\sqrt{6}}{360}$	$\frac{296-169\sqrt{6}}{1800}$	$\frac{-2+3\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\sqrt{6}}{1800}$	$\frac{88+7\sqrt{6}}{360}$	$\frac{-2-3\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

V.3 Index 2 problems

We next consider semi-explicit problems

$$\begin{aligned} \dot y &= f(y,z), \qquad && y(0)=y_0 \\ 0 &= g(y), && z(0)=z_0 \end{aligned} \tag{3.1}$$

where the initial values satisfy $g(y_0)=0$ and $g_y(y_0)f(y_0,z_0)=0$. We assume that f and g are sufficiently differentiable and that

$$g_y(y)f_z(y,z) \quad \text{is invertible} \tag{3.2}$$

in a neighbourhood of the solution, so that the problem has index 2. Recall that this problem can be considered as a differential equation on the manifold $\mathcal{M}=\{y\,;\,g(y)=0\}$.

In this section we restrict our considerations to implicit Runge–Kutta methods with invertible matrix (a_{ij}), and coefficients satisfying $b_j=a_{sj}$ for all j (stiffly accurate methods). For the problem (3.1) they are defined by

$$Y_{ni}-y_n = h\sum_{j=1}^{s} a_{ij}\, f(Y_{nj},Z_{nj}), \qquad 0=g(Y_{ni}), \qquad i=1,\dots,s \tag{3.3}$$

with a numerical approximation after one step given by $y_{n+1}=Y_{ns}, z_{n+1}=Z_{ns}$. Notice that the internal stages and the numerical solution do not depend on z_n. The value of z_n only specifies the solution branch of $g_y(y)f(y,z)=0$ to which the expressions Z_{nj} remain close. Moreover, the numerical solution y_n stays on the manifold $\mathcal{M}$ for all n.

The convergence results of this section are also valid for index 2 systems of the form $\dot y = f(y,z)$, $0=g(y,z)$, if they can be transformed to (3.1) without any differentiation (see the discussion of index 2 systems in Section IV.2). This is because the method (3.3) is invariant with respect to these transformations.

Theorem 3.1 (existence and uniqueness of numerical solution)**.** *Consider $y_n\in\mathcal{M}$, let ζ be a value satisfying $g_y(y_n)f(y_n,\zeta)=0$, and assume that (3.2) holds in a neighborhood of (y_n,ζ). If the Runge–Kutta matrix (a_{ij}) is invertible, then there exists $h_0>0$ such that the nonlinear system (3.3) possesses for $|h|\le h_0$ a locally unique solution which satisfies*

$$Y_{ni}-y_n=\mathcal{O}(h), \qquad Z_{ni}-\zeta=\mathcal{O}(h). \tag{3.4}$$

Proof. We shall prove that the solution (Y_{ni},Z_{ni}) of (3.3) can be expressed as a smooth function of h (for sufficiently small h). A direct application of the implicit function theorem is not possible due to the presence of the factor h in front of the Z_{nj} dependence.

The idea is to use the fundamental theorem of calculus

$$g(Y_{ni})-g(y_n)=\int_0^1 g_y\Big(y_n+\tau(Y_{ni}-y_n)\Big)(Y_{ni}-y_n)\,\mathrm{d}\tau,$$

so that the second relation of (3.3), after division by h, can be written as

$$\int_0^1 g_y\Big(y_n+\tau(Y_{ni}-y_n)\Big)\,\mathrm{d}\tau\cdot\sum_{j=1}^{s} a_{ij}f(Y_{nj},Z_{nj})=0, \qquad i=1,\dots,s, \tag{3.5}$$

which is the discrete analogue of the hidden constraint $g_y(y)f(y,z)=0$. We now apply the implicit function theorem to the system formed by (3.5) and the first relation of (3.3). For

$h = 0$, our assumptions imply that the values $Y_{ni} = y_n$ and $Z_{ni} = \zeta$ satisfy the system. Furthermore, the derivative with repect to (Y_{ni}, Z_{ni}) at $h = 0$ and $(Y_{ni}, Z_{ni}) = (y_n, \zeta)$ is of the form

$$\begin{pmatrix} I \otimes I & 0 \\ \mathcal{O}(1) & A \otimes (g_y f_z)(y_n, \zeta) \end{pmatrix},$$

which is invertible because of (3.2). Therefore the implicit function theorem yields the existence of a locally unique solution of (3.3). □

The method (3.3) represents a numerical one-step method on the manifold $\mathcal{M}$. In view of an application of the convergence theorem of Section III.5 we have to study the local error. Recall that the local error is the difference $(y_{n+1} - y_n(t_{n+1}), z_{n+1} - z_n(t_{n+1}))$, where $(y_n(t), z_n(t))$ is the solution of (3.1) with consistent initial values $y_n(t_n) = y_n, z_n(t_n) = z_n$.

Theorem 3.2 (local error estimate)**.** *Consider a differential-algebraic equation (3.1) satisfying (3.2), and apply an implicit Runge–Kutta method (3.3) with invertible matrix (a_{ij}) and coefficients satisfying $b_j = a_{sj}$ for all j. If the quadrature formula formed by $(b_i, c_i)_{i=1}^s$ is of order p, and the method has stage order q, then we have the estimate*

$$y_{n+1} - y_n(t_n + h) = \mathcal{O}(h^{\min(p+1, q+2)}), \qquad z_{n+1} - z_n(t_n + h) = \mathcal{O}(h^q).$$

Proof. Inspired by the proof of Theorem 3.1, we consider the nonlinear system for (Y_i, Z_i), $i = 1, \dots, s$,

$$\begin{aligned} Y_i - y_n &= h \sum_{j=1}^s a_{ij}\, f(Y_j, Z_j) + h\delta_i \\ \int_0^1 g_y\Big(y_n + \tau(Y_i - y_n)\Big)\, \mathrm{d}\tau \cdot \Big(\sum_{j=1}^s a_{ij} f(Y_j, Z_j) + \delta_i \Big) &= 0, \end{aligned} \tag{3.6}$$

where the second equation is known to be equivalent to $g(Y_i) = 0$. For $\delta_i = 0$ we obtain the numerical solution $(Y_i, Z_i) = (Y_{ni}, Z_{ni})$ of (3.3). The exact solution at the quadrature points $(Y_i, Z_i) = (y_n(t_n + c_i h), z_n(t_n + c_i h))$ satisfies (3.6) with $\delta_i = \mathcal{O}(h^q)$ for $i = 1, \dots, s-1$, and $\delta_s = \mathcal{O}(h^p)$. We are interested in the dependence of the solution (Y_i, Z_i) on the parameters δ_i, when the step size $h \neq 0$ is fixed. We see that the derivative of the system (3.6) with respect to (Y_i, Z_i) at the numerical approximation (Y_{ni}, Z_{ni}) is of the form

$$\begin{pmatrix} I \otimes I + \mathcal{O}(h) & \mathcal{O}(h) \\ \mathcal{O}(1) & A \otimes (g_y f_z)(y_n, z_n) + \mathcal{O}(h) \end{pmatrix},$$

which is invertible for sufficiently small h. The implicit function theorem therefore implies that

$$Y_{ni} - y_n(t_n + c_i h) = \mathcal{O}(\delta), \qquad Z_{ni} - z_n(t_n + c_i h) = \mathcal{O}(\delta),$$

where $\delta = \max_{i=1,\dots,s} \delta_i = \mathcal{O}(h^q)$. This proves the estimate for the local error of the z component. Some further considerations are necessary for the local error of the y component.

First, we notice that due to the factor h in the right-hand side of the upper equation of (3.6), we have the improved estimate $Y_{ni} - y_n(t_n + c_i h) = \mathcal{O}(h^{q+1})$ for all i. For the local error $\Delta y_{n+1} = y_{n+1} - y_n(t_n + h)$ of the y component we thus obtain

$$\Delta y_{n+1} = h\, f_z(y_n, z_n) \sum_{j=1}^s b_j \Big(Z_{nj} - z_n(t_n + c_j h) \Big) + \mathcal{O}(h^{q+2}) + \mathcal{O}(h^{p+1}). \tag{3.7}$$

On the other hand, we have

$$0 = g(y_{n+1}) - g(y_n(t_n + h)) = g_y(y_n)\Delta y_{n+1} + \mathcal{O}(h\Delta y_{n+1}) = g_y(y_n)\Delta y_{n+1} + \mathcal{O}(h^{q+2}).$$

Multiplying equation (3.7) with $g_y(y_n)$ yields

$$0 = h\, g_y(y_n) f_z(y_n, z_n) \sum_{j=1}^{s} b_j \Big(Z_{nj} - z_n(t_n + c_j h) \Big) + \mathcal{O}(h^{q+2}) + \mathcal{O}(h^{p+1}).$$

Since $g_y(y_n)f_z(y_n, z_n)$ is invertible by (3.2), the expression $h\sum_{j=1}^{s} b_j(Z_{nj} - z_n(t_n + c_j h))$ is of size $\mathcal{O}(h^{q+2}) + \mathcal{O}(h^{p+1})$. Inserted into (3.7) we finally get the stated estimate for the y component. □

Remark. Whereas the estimate for the local error of the z component is in general optimal, that for the y component can be improved in some interesting situations. For example, for the Radau IIA methods of Example 2.2, we have for the y component $y_{n+1} - y_n(t_n + h) = \mathcal{O}(h^{p+1})$. This property is known as *superconvergence*.

Convergence for the y component. The numerical method (3.3) can be considered as a mapping $y_n \mapsto y_{n+1}$ on the submanifold $\mathcal{M}$. The approximations z_n only influence the choice of the solution, when the equation $0 = g_y(y)f(y, z)$ has more than one solutions z for a given y. Theorem III.5.2 can therefore be applied and yields the estimate for the global error

$$y_n - y(t_n) = \mathcal{O}(h^{\min(p,q+1)}) \qquad \text{for} \quad t_n = nh \le T.$$

Convergence for the z component. The numerical solution z_n is defined locally and there is no propagation of errors. The error is therefore a superposition of the local error for the z component and the global error of the y component. Since we have $p \ge q$ for stiffly accurate methods, this implies

$$z_n - z(t_n) = \mathcal{O}(h^q) \qquad \text{for} \quad t_n = nh \le T.$$

V.4 Constrained mechanical systems

We consider constrained mechanical systems with kinetic energy $T(\dot q) = \frac{1}{2}\,\dot q^{\mathsf T} M\,\dot q$ and potential energy $U(q)$. Here and in the next section we assume the matrix M symmetric, positive definite, and independent of q. The equations of motion become more structured if we use the momentum coordinates $p = \frac{\partial L}{\partial \dot q} = M\dot q$ in place of the velocity coordinates $v = \dot q$. As explained in Section IV.4 the equations of motion are given by

$$\begin{aligned} \dot q &= M^{-1} p \\ \dot p &= -\nabla U(q) - G(q)^{\mathsf T}\lambda \\ 0 &= g(q), \end{aligned} \tag{4.1}$$

where $G(q) = g'(q)$. This system has many remarkable properties: it exactly preserves the total energy

$$H(p, q) = \frac{1}{2}\, p^T M^{-1} p + U(q), \tag{4.2}$$

and the flow is a symplectic and volume preserving transformation. It is not the aim of this lecture to discuss these topics[5], we concentrate on the fact that the system (4.1) is a differential-algebraic equation of index 3 and can be considered as a differential equation on the manifold

$$\mathcal{M} = \Big\{(p,q)\,;\, g(q) = 0,\ G(q)M^{-1}p = 0\Big\}. \tag{4.3}$$

Symplectic Euler method for constrained mechanical systems. We integrate the p and λ variables by the implicit Euler and the q variable by the explicit Euler method. This leads to the discretization

$$\begin{aligned} \widehat{p}_{n+1} &= p_n - h\,(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_{n+1}) \\ q_{n+1} &= q_n + h\,M^{-1}\,\widehat{p}_{n+1} \\ 0 &= g(q_{n+1}). \end{aligned} \tag{4.4}$$

The numerical approximation $(\widehat{p}_{n+1}, q_{n+1})$ satisfies the constraint $g(q) = 0$, but not the hidden constraint $G(q)M^{-1}p = 0$. To get an approximation $(p_{n+1}, q_{n+1}) \in \mathcal{M}$, we append the projection

$$\begin{aligned} p_{n+1} &= \widehat{p}_{n+1} - h\,G(q_{n+1})^{\mathsf{T}}\mu_{n+1} \\ 0 &= G(q_{n+1})\,M^{-1}\,p_{n+1}. \end{aligned} \tag{4.5}$$

Let us discuss some basic properties of this method.

Existence and Uniqueness of the Numerical Solution. Inserting the definition of q_{n+1} from the second line of (4.4) into $0 = g(q_{n+1})$ gives a nonlinear system for $\widehat{p}_{n+1}$ and $h\lambda_{n+1}$. Due to the factor h in front of $M^{-1}\widehat{p}_{n+1}$, the implicit function theorem cannot be directly applied to prove existence and uniqueness of the numerical solution. We therefore write this equation as

$$0 = g(q_{n+1}) = g(q_n) + \int_0^1 G\Big(q_n + \tau(q_{n+1} - q_n)\Big)(q_{n+1} - q_n)\,\mathrm{d}\tau.$$

We now use $g(q_n) = 0$, insert the definition of q_{n+1} from the second line of (4.4) and divide by h. Together with the first line of (4.4) this yields the system $F\Big(\widehat{p}_{n+1}, h\lambda_{n+1}, h\Big) = 0$ with

$$F\Big(p, \nu, h\Big) = \begin{pmatrix} p - p_n + h\nabla U(q_n) + G(q_n)^{\mathsf{T}}\nu \\ \displaystyle\int_0^1 G\Big(q_n + \tau h M^{-1}p\Big)\,M^{-1}p\,\mathrm{d}\tau \end{pmatrix}.$$

Since $(p_n, q_n) \in \mathcal{M}$ with $\mathcal{M}$ from (4.3), we have $F(p_n, 0, 0) = 0$. Furthermore,

$$\frac{\partial F}{\partial(p,\nu)}\Big(p_n, 0, 0\Big) = \begin{pmatrix} I & G(q_n)^{\mathsf{T}} \\ G(q_n)M^{-1} & 0 \end{pmatrix},$$

and this matrix is invertible, because we always assume the matrix $G(q)$ to be of full rank. Consequently, an application of the implicit function theorem proves that the numerical solution $(\widehat{p}_{n+1}, h\lambda_{n+1})$ (and hence also q_{n+1}) exists and is locally unique for sufficiently small h. The projection step (4.5) represents a linear system for p_{n+1} and $h\mu_{n+1}$ with invertible matrix.

[5]They are treated in the monograph *Geometric Numerical Integration* by Hairer, Lubich, and Wanner. This and the next section are taken from this monograph.

Convergence of Order 1. The above use of the implicit function theorem yields the rough estimates

$$\widehat{p}_{n+1} = p_n + \mathcal{O}(h), \quad h\lambda_{n+1} = \mathcal{O}(h), \quad h\mu_{n+1} = \mathcal{O}(h),$$

which, together with the equations (4.4) and (4.5), give

$$q_{n+1} = q_n(t_{n+1}) + \mathcal{O}(h^2), \quad p_{n+1} = p_n(t_{n+1}) - G(q_n(t_{n+1}))^{\mathsf{T}}\nu + \mathcal{O}(h^2),$$

where $(p_n(t), q_n(t))$ is the solution of (4.1) passing through $(p_n, q_n) \in \mathcal{M}$ at $t = t_n$. Inserting these relations into the second equation of (4.5) we get

$$0 = G(q_n(t))M^{-1}p_n(t) - G(q_n(t))M^{-1}G(q_n(t))^{\mathsf{T}}\nu + \mathcal{O}(h^2)$$

at $t = t_{n+1}$. Since $G(q_n(t))M^{-1}p_n(t) = 0$, and $G(q_n(t))M^{-1}G(q_n(t))^{\mathsf{T}}$ is invertible, we have $\nu = O(h^2)$. The local error is therefore of size $\mathcal{O}(h^2)$ in both components.

The convergence proof is now a direct application of Theorem III.5.2, because the method is a mapping $\Phi_h : \mathcal{M} \to \mathcal{M}$ on the solution manifold. This proves that the global error satisfies $p_n - p(t_n) = \mathcal{O}(h)$ and $q_n - q(t_n) = \mathcal{O}(h)$ as long as $t_n = nh \le Const.$

Numerical Experiment (spherical pendulum). We denote by q_1, q_2, q_3 the Cartesian coordinates of a point with mass $m = 1$ that is connected with a massless rod of length $\ell = 1$ to the origin. The kinetic and potential energies are $T = \frac{1}{2}(\dot{q}_1^2 + \dot{q}_2^2 + \dot{q}_3^2)$ and $U = q_3$, respectively, and the constraint is the fixed length of the rod. We thus get the system

$$\begin{aligned} \dot{q}_1 &= p_1 & \dot{p}_1 &= -q_1\lambda \\ \dot{q}_2 &= p_2 & \dot{p}_2 &= -q_2\lambda \\ \dot{q}_3 &= p_3 & \dot{p}_3 &= -1 - q_3\lambda \\ 0 &= \tfrac{1}{2}(q_1^2 + q_2^2 + q_3^2 - 1). \end{aligned} \tag{4.6}$$

Figure V.1 (upper picture) shows the numerical solution (vertical coordinate q_3) over many periods obtained by method (4.4)-(4.5). We observe a regular qualitatively correct behavior. For the implicit Euler method (i.e., the argument q_n is replaced with q_{n+1} in (4.4)) the numerical solution, obtained with the same step size and the same initial values, is less satisfactory. Already after one period the solution deteriorates and the system loses energy.

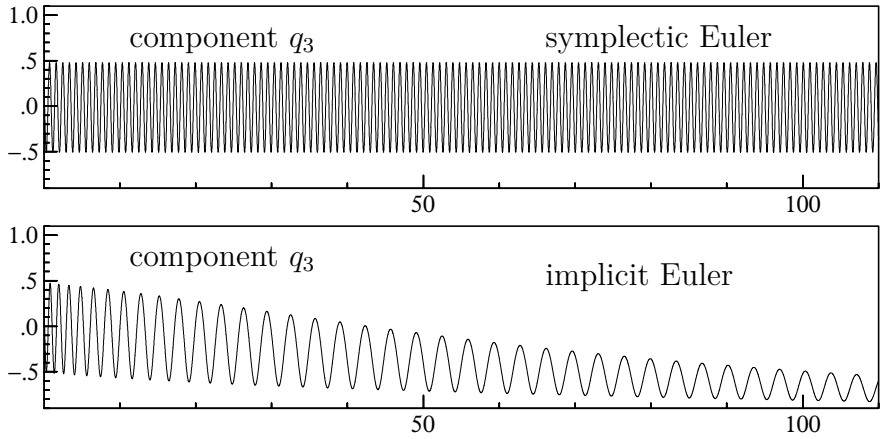


FIG. V.1: Spherical pendulum problem solved with the symplectic Euler method (4.4)-(4.5) and with the implicit Euler method; initial value $q_0 = (\sin(1.3), 0, \cos(1.3))$, $p_0 = (3\cos(1.3), 6.5, -3\sin(1.3))$, step size $h = 0.01$.

V.5 Shake and Rattle

The numerical method (4.4)-(4.5) is only of order 1 and it is not symmetric. Here we present an algorithm that is of order 2, symmetric and symplectic. The original derivation is based on the fact that the system (4.1) is equivalent to the second order differential equation $M\ddot{q} = -\nabla U(q) - G(q)^{\mathsf{T}}\lambda$ with constraint $g(q) = 0$.

SHAKE. Ryckaert, Ciccotti, and Berendsen (1977) propose the method

$$\begin{aligned} q_{n+1} - 2q_n + q_{n-1} &= -h^2 M^{-1}\big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_n\big) \\ 0 &= g(q_{n+1}) \end{aligned} \tag{5.1}$$

for computations in molecular dynamics. The p component, not used in the recursion, is approximated by the symmetric finite difference $p_n = M(q_{n+1} - q_{n-1})/2h$.

RATTLE. The three-term recursion (5.1) may lead to an accumulation of round-off errors, and a reformulation as a one-step method is desirable. Introducing a new variable via $q_{n+1} - q_n = hM^{-1}p_{n+1/2}$, the method (5.1) becomes $p_{n+1/2} - p_{n-1/2} = -h(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_n)$ and the momentum approximation leads to $p_{n+1/2} + p_{n-1/2} = 2p_n$. Elimination of either $p_{n+1/2}$ or $p_{n-1/2}$ leads to the formulae

$$\begin{aligned} p_{n+1/2} &= p_n - \frac{h}{2}\big(\nabla U(q_n) + G(q_n)^{\mathsf{T}}\lambda_n\big) \\ q_{n+1} &= q_n + hM^{-1}p_{n+1/2}, \qquad\qquad 0 = g(q_{n+1}) \\ p_{n+1} &= p_{n+1/2} - \frac{h}{2}\big(\nabla U(q_{n+1}) + G(q_{n+1})^{\mathsf{T}}\lambda_{n+1}\big). \end{aligned} \tag{5.2}$$

The difficulty with this formulation is that λ_{n+1} is not yet available at this step (it is computed together with q_{n+2}). As a remedy, Andersen (1983) suggests replacing the last line in (5.2) with a projection step similar to (4.5)

$$\begin{aligned} p_{n+1} &= p_{n+1/2} - \frac{h}{2}\big(\nabla U(q_{n+1}) + G(q_{n+1})^{\mathsf{T}}\mu_n\big) \\ 0 &= G(q_{n+1})M^{-1}p_{n+1}. \end{aligned} \tag{5.3}$$

This modification, called RATTLE, has the further advantage that the numerical approximation (p_{n+1}, q_{n+1}) lies on the solution manifold $\mathcal{M}$.

Theorem 5.1. *The RATTLE method is symmetric, symplectic, and convergent of order* 2.

Proof. If we add the consistency conditions $g(q_n) = 0$, $G(q_n)M^{-1}p_n = 0$ of the initial values to the RATTLE algorithm, the symmetry of the method follows at once by exchanging $h \leftrightarrow -h$, $p_{n+1} \leftrightarrow p_n$, $q_{n+1} \leftrightarrow q_n$, and $\lambda_n \leftrightarrow \mu_n$. We do not discuss the symplecticity in this lecture, and refer to the monograph *Geometric Numerical Integration*.

The implicit function theorem applied to the two systems (5.2) and (5.3) shows that

$$p_{n+1/2} = p_n + \mathcal{O}(h), \qquad h\lambda_n = \mathcal{O}(h), \qquad p_{n+1} = p_{n+1/2} + \mathcal{O}(h), \qquad h\mu_n = \mathcal{O}(h),$$

and, inserted into (5.2), yields

$$q_{n+1} = q(t_{n+1}) + \mathcal{O}(h^2), \qquad p_{n+1} = p(t_{n+1}) - G(q(t_{n+1}))^{\mathsf{T}}\nu + \mathcal{O}(h^2).$$

Convergence of order one follows therefore in the same way as for method (4.4)-(4.5) by applying the convergence Theorem III.5.2. Since the order of a symmetric method is always even, this implies convergence of order two. □

V.6 Exercises

1. (Gear, Hsu, and Petzold 1981, Gear and Petzold 1984). Consider the problem

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta\, t \end{pmatrix} \begin{pmatrix} \dot{y} \\ \dot{z} \end{pmatrix} + \begin{pmatrix} 1 & \eta\, t \\ 0 & 1+\eta \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix}. \tag{6.1}$$

   Prove that the system (6.1) has differentiation index 2 for all values of η, and that the z-component of the exact solution is given by $z(t) = g(t) - \frac{\mathrm{d}}{\mathrm{d}t} f(t)$.

2. A straight-forward application of the implicit Euler method to the differential-algebraic equation (6.1) would be

$$\begin{pmatrix} 0 & 0 \\ 1 & \eta\, t_{n+1} \end{pmatrix} \begin{pmatrix} y_{n+1} - y_n \\ z_{n+1} - z_n \end{pmatrix} + h \begin{pmatrix} 1 & \eta\, t_{n+1} \\ 0 & 1+\eta \end{pmatrix} \begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = h \begin{pmatrix} f(t_{n+1}) \\ g(t_{n+1}) \end{pmatrix}. \tag{6.2}$$

   Prove that this method yields the recursion

$$z_{n+1} = \frac{\eta}{1+\eta}\, z_n + \frac{1}{1+\eta} \Big(g(t_{n+1}) - \frac{f(t_{n+1}) - f(t_n)}{h} \Big).$$

   Hence, the method is convergent for $\eta > -1/2$, but unstable for $\eta < -1/2$. For $\eta = -1$ the numerical solution does not exist.

3. Introducing the new variable $u = \dot{z}$, the system (6.1) becomes equivalent to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{y} \\ \dot{z} \\ \dot{u} \end{pmatrix} + \begin{pmatrix} 0 & 1+\eta & \eta\, t \\ 0 & 0 & -1 \\ 1 & \eta\, t & 0 \end{pmatrix} \begin{pmatrix} y \\ z \\ u \end{pmatrix} = \begin{pmatrix} g(t) \\ 0 \\ f(t) \end{pmatrix}, \tag{6.3}$$

   which is of the form (0.1). Prove that this system has differentiation index 3.

4. Using the approach of the present chapter, apply the implicit Euler method to the system (6.3). Is the resulting discretization equivalent to (6.2)?

5. Consider the differential-algebraic equation

$$\dot{y} = (\mathrm{e}^{z-1} + 1)/2, \qquad 0 = y - t$$

   with consistent initial values $y(0) = 0$ and $z(0) = 1$. Prove that we are concerned with a problem of index 2, and the corresponding manifold is $\mathcal{M} = \{(t, y)\,;\, y - t = 0\}$.
   Prove that the implicit Euler method, applied to this problem with starting approximation $y_0 = h$ and $z_0 = 1$ does not have a solution.

   Remark. This exercise shows that a numerical method for index 2 problems may fail if the initial value is not $\mathcal{O}(h^2)$ close to the manifold.