

Appendix A

Fixed point theorems for first- and second-order polynomial mappings

A.1 Introduction

In this appendix we provide a deeper insight into the fixed point theorem that was presented and used in Chapter 3. In particular, we concentrate on two of the simplest but most important special cases of this theorem, one for linear or affine mappings (in Sec. A.2), and the other for second-order polynomial mappings (in Sec. A.3). The importance of these particular cases is twofold: On the one hand, these cases cover many of the layer transformations being used in the figures and in the examples throughout this book; but on the other hand, these cases may also serve as simple illustrations to the general case involving more complex mappings. A further generalization to the case of mutual fixed points between two mappings is discussed in Sec. A.4.

A.2 The fixed point theorem for linear or affine mappings

As we have seen in Sec. 3.2, the affine fixed point theorem states that all non-degenerate affine mappings $\mathbf{g}(x,y)$ from \mathbb{R}^2 onto itself have a single fixed point.

In order to more deeply understand this theorem, we start by analyzing the different possible types of affine mappings $\mathbf{g}(x,y)$. The most general form of an affine mapping is:

$$\begin{aligned}x' &= a_1x + b_1y + x_0 \\y' &= a_2x + b_2y + y_0\end{aligned}\tag{A.1}$$

Let us first consider the homogeneous mapping that is associated with $\mathbf{g}(x,y)$, i.e. the corresponding linear transformation where the shift (x_0, y_0) is zero:

$$\begin{aligned}x' &= a_1x + b_1y \\y' &= a_2x + b_2y\end{aligned}\tag{A.2}$$

Such a linear transformation may either:

- (a) map \mathbb{R}^2 onto the whole of \mathbb{R}^2 (this occurs, for example, in rotations, scalings, flipping over an axis, etc.);
- (b) map \mathbb{R}^2 onto \mathbb{R} (this occurs, for example, in a projection onto the x axis); or
- (c) map \mathbb{R}^2 onto the origin $(0,0)$ (this occurs in the zero transformation that maps all the points (x,y) onto $(0,0)$).

Cases (b) and (c) occur when the linear transformation (A.2) is singular, i.e. when its determinant equals zero:

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = a_1 b_2 - a_2 b_1 = 0 \quad (\text{A.3})$$

The same three cases occur also in the affine mapping (A.1), except that here, in cases (b) and (c) \mathbb{R}^2 is mapped, respectively, onto a shifted line $(x_0, y_0) + \mathbb{R}$ or onto a shifted point (x_0, y_0) .

The degenerate linear or affine transformations of types (b) and (c) do not interest us, of course, in our study on superpositions of transformed layers, since their application would completely destroy our 2D layers. Consequently, we are only interested in linear or affine transformations belonging to type (a), namely, cases in which determinant (A.3) is non-zero. These cases are called *regular* or *non-singular*.

Let us now proceed to the fixed points of such non-singular transformations. A non-singular linear transformation (A.2) may either:

- (1) have a single fixed point, located at the origin (this occurs, for example, in rotations, scalings, etc.);
- (2) have a full line of fixed points that passes through the origin (this occurs, for example, in transformations such as flipping over the x axis, or scaling in the y direction alone, in both of which all points of the x axis are fixed points); or
- (3) have all the points of the entire x, y plane as fixed points (this occurs in the identity transformation).

When does each of these cases occur? As we know, the fixed points of the linear transformation (A.2) are those points of the plane which satisfy $(x', y') = (x, y)$, namely:¹

$$\begin{aligned} x &= a_1 x + b_1 y \\ y &= a_2 x + b_2 y \end{aligned} \quad (\text{A.4})$$

This gives us the following linear set of equations for x and y :

$$\begin{aligned} (1 - a_1)x - b_1 y &= 0 \\ -a_2 x + (1 - b_2)y &= 0 \end{aligned} \quad (\text{A.5})$$

Clearly, cases (2) and (3) above occur when this linear set of equations is singular, i.e. when:

$$\begin{vmatrix} 1 - a_1 & -b_1 \\ -a_2 & 1 - b_2 \end{vmatrix} = 1 - a_1 - b_2 + a_1 b_2 - a_2 b_1 = 0 \quad (\text{A.6})$$

¹ Note that here $\mathbf{g}(x, y) = (x, y)$ (or equivalently, in terms of matrices, $\mathbf{A}\mathbf{x} = \mathbf{x}$) does not mean that $\lambda = 1$ is an eigenvalue of $\mathbf{g}(x, y)$ [Kreyszig93 p. 157], since in our case it may certainly happen that $\mathbf{x} = (0, 0)$ is the only solution of $\mathbf{g}(\mathbf{x}) = \mathbf{x}$ (as in the case of a scaling or rotation transformation \mathbf{g}).

In such cases the associated affine mapping (A.1), which is obtained by adding to (A.2) a shift of (x_0, y_0) , may have either infinitely many fixed points, or no fixed points at all. Only in case (1), i.e. when determinant (A.6) is non-zero, the associated affine mapping (A.1) has precisely one single fixed point; its location is given then by Eq. (3.13).

Let us first consider transformations that satisfy both conditions (a) and (1), i.e. where both of the determinants (A.3) and (A.6) are non-zero. We call such affine mappings *non-degenerate affine mappings*.

It is clear, therefore, that all non-degenerate affine mappings $\mathbf{g}(x, y)$ from \mathbb{R}^2 onto itself have a single fixed point; this is, indeed, precisely what is claimed by our affine fixed point theorem in Sec. 3.2. But this theorem does not say anything about degenerate transformations that do not satisfy condition (a) or condition (1).

As an illustration, let us mention that mappings such as rotations, scalings, etc. as well as their combinations have, indeed, a single fixed point. This is also true for all of their combinations with translation, but not for pure translations. Note that pure translations are excluded, since their determinant (A.6) is zero; this can be understood more intuitively as follows: The homogeneous transformation (A.2) associated with a pure translation is the identity transformation, that belongs to class (3) above and has all the points of the x, y plane as fixed points. But the addition of a translation to the identity transformation destroys all of its fixed points, so that a pure translation has no fixed points at all.

As a second example, let us consider the linear transformation which consists of vertical scaling. This transformation belongs to class (2) above, and has all the points of the x axis as fixed points. What happens now when we add to this linear transformation a translation? In this case, the answer depends on the direction of the translation: If the translation is horizontal, it is clear that all the fixed points on the x axis are destroyed, and the resulting affine mapping $\mathbf{g}(x, y)$ has no fixed points. But if the translation is vertical, the resulting affine mapping $\mathbf{g}(x, y)$ will still have a full line of fixed points, which is parallel to the x axis. Note, however, that such cases are not treated by our affine fixed point theorem, since their determinant (A.6) is zero: As we have seen, this theorem only considers non-degenerate affine mappings, but it does not say anything about degenerate affine mappings. Indeed, some degenerate affine mappings have a full line of fixed points, while others have no fixed points at all.

Thus, in order to cover all of the possible cases we need to introduce a more general version of our theorem, that treats all affine mappings from \mathbb{R}^2 onto itself, including degenerate cases such as vertical scalings and translations:

Generalized affine fixed point theorem: An affine mapping $\mathbf{g}(x, y)$ from \mathbb{R}^2 onto itself has a fixed point (either one or infinitely many) *iff* $\text{rank}A = \text{rank}B$, where A is the 2×2 coefficient matrix of the homogeneous system of Eqs. (A.5) and B is the 2×3 extended matrix that includes x_0 and y_0 in its third column:

$$A = \begin{pmatrix} 1 - a_1 & -b_1 \\ -a_2 & 1 - b_2 \end{pmatrix} \quad B = \begin{pmatrix} 1 - a_1 & -b_1 & x_0 \\ -a_2 & 1 - b_2 & y_0 \end{pmatrix}$$

Moreover, if the rank of both A and B is 2, the fixed point is unique; if their rank is 1, there exists a full line of fixed points; and if their rank is 0, all the points of the x, y plane are fixed points of $\mathbf{g}(x, y)$ (this occurs if $\mathbf{g}(x, y)$ is the identity mapping without translation). But if $\text{rank} A \neq \text{rank} B$, the affine mapping $\mathbf{g}(x, y)$ has no fixed points at all. ■

This generalized theorem is, in fact, an application to the particular case of Eqs. (A.5) of the algebraic theorem on the dimension of the solution space of a system of linear equations [Bronshtein97 p. 143].

A.3 The fixed point theorem for second-order polynomial mappings

As we have seen in Secs. 3.2 and A.1, the affine fixed point theorem states that all non-degenerate affine mappings $\mathbf{g}(x, y)$ from \mathbb{R}^2 onto itself have a single fixed point.

This theorem can be generalized to the case of second-order polynomial mappings. By a second-order polynomial mapping (or simply, a mapping of order 2) we mean a mapping $\mathbf{g}(x, y)$ that is defined by a pair of algebraic equations of order 2, namely:

$$\begin{aligned} x' &= a_1x^2 + b_1xy + c_1y^2 + d_1x + e_1y + x_0 \\ y' &= a_2x^2 + b_2xy + c_2y^2 + d_2x + e_2y + y_0 \end{aligned} \quad (\text{A.7})$$

Just as in the affine case, such transformations do not necessarily map \mathbb{R}^2 onto the entire \mathbb{R}^2 : in some degenerate cases the transformation (A.7) maps \mathbb{R}^2 onto a straight or a curved line, or even into a single point. For example, the second order transformation $\mathbf{g}(x, y) = (x, x^2)$ maps \mathbb{R}^2 onto the parabola $y = x^2$. Such situations occur when the mapping $\mathbf{g}(x, y)$ defined by the set of equations (A.7) is *singular*, namely, when the two equations forming the set are not independent. Two equations $g_1(x, y) = 0$ and $g_2(x, y) = 0$ are said to be *independent* (or *functionally independent*) if there exists no function $f(u, v)$ other than $f(u, v) \equiv 0$ such that $f(g_1(x, y), g_2(x, y)) = 0$ is satisfied for all (x, y) . Equivalently, this means that the Jacobian:

$$J(x, y) = \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \frac{\partial g_1}{\partial x} \frac{\partial g_2}{\partial y} - \frac{\partial g_2}{\partial x} \frac{\partial g_1}{\partial y} \quad (\text{A.8})$$

is not identically zero (see [Bronstein90 pp. 226, 430–431] or [Courant88 pp. 154–155]). If $g_1(x, y)$ and $g_2(x, y)$ are *dependent*, for instance if $g_2(x, y) = g_1(x, y)^2$, they are a consequence of each other, and hence the 2D transformation $\mathbf{g}(x, y) = (g_1(x, y), g_2(x, y))$ they define is singular, and it maps \mathbb{R}^2 onto a 1D curve or even a single point in \mathbb{R}^2 .

In the particular case of polynomial mappings of order 2, the condition for Eq. (A.7) to be singular is, therefore, that the Jacobian (A.8) be identically zero, namely:

$$(2a_1x + b_1y + d_1)(b_2x + 2c_2y + e_2) - (2a_2x + b_2y + d_2)(b_1x + 2c_1y + e_1) \equiv 0 \quad (\text{A.9})$$

which gives:

$$2(a_1b_2 - a_2b_1)x^2 + 4(a_1c_2 - a_2c_1)xy + 2(b_1c_2 - b_2c_1)y^2 + (b_2d_1 - b_1d_2 + 2a_1e_2 - 2a_2e_1)x \\ + (2c_2d_1 - 2c_1d_2 + b_1e_2 - b_2e_1)y + (d_1e_2 - d_2e_1) \equiv 0$$

But since this expression must be identically zero for any values of x and y , this means that Eq. (A.7) is singular when all of the following conditions are simultaneously satisfied:

$$\begin{aligned} a_1b_2 &= a_2b_1 \\ a_1c_2 &= a_2c_1 \\ b_1c_2 &= b_2c_1 \\ b_2d_1 + 2a_1e_2 &= b_1d_2 + 2a_2e_1 \\ 2c_2d_1 + b_1e_2 &= 2c_1d_2 + b_2e_1 \\ d_1e_2 &= d_2e_1 \end{aligned} \quad (\text{A.10})$$

Such degenerate cases do not interest us, of course, in our study on superpositions of transformed layers, since the mappings $\mathbf{g}(x,y)$ they represent do not map \mathbb{R}^2 onto itself; we will only be interested in *non-singular* mappings of order 2, namely, cases in which Eq. (A.7) is non-singular.

We now proceed to discuss the fixed points of non-singular mappings of order 2. The fixed points of transformation (A.7) are those points of the plane for which (x',y') equals (x,y) , namely:

$$\begin{aligned} x &= a_1x^2 + b_1xy + c_1y^2 + d_1x + e_1y + x_0 \\ y &= a_2x^2 + b_2xy + c_2y^2 + d_2x + e_2y + y_0 \end{aligned} \quad (\text{A.11})$$

This gives us the following system of equations for x and y :

$$\begin{aligned} a_1x^2 + b_1xy + c_1y^2 + (d_1 - 1)x + e_1y + x_0 &= 0 \\ a_2x^2 + b_2xy + c_2y^2 + d_2x + (e_2 - 1)y + y_0 &= 0 \end{aligned} \quad (\text{A.12})$$

A general rule in algebra states that a system of two equations $p_1(x,y) = 0$, $p_2(x,y) = 0$, where $p_1(x,y)$ is an m -order polynomial in x and y and $p_2(x,y)$ is an n -order polynomial in x and y , has mn solutions (x,y) , real or complex [Bronstein90 pp. 226–227]. This means that in our case the system of equations (A.12) has up to 4 *real* solutions (x,y) . This agrees, indeed, with our geometric intuition, since each equation of order 2 represents in fact a conic curve in the plane, and the intersection of two such conics clearly gives up to 4 real solutions. However, depending on the locations and orientations of the two conic curves, they may have only 3, 2, 1 or even 0 intersection points. Moreover, just like in the affine

case, there may exist here, too, degenerate systems (A.12) with infinitely many solutions, meaning that the corresponding mapping $g(x,y)$ defined by (A.7) has infinitely many fixed points. In general, the system of equations (A.12) may either:

- (1) have between 0 and 4 solutions (which are fixed points of the mapping (A.7));
- (2) have a full straight or curved line of solutions (fixed points of (A.7)); or
- (3) have the full x,y plane as solutions (fixed points of (A.7)).

Cases (2) and (3) occur when the set of equations (A.12) is singular, or in other words, if the two equations forming the set are not independent. This happens when their Jacobian is identically zero:²

$$\begin{aligned}
 J(x,y) &= \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \frac{\partial g_1}{\partial x} \frac{\partial g_2}{\partial y} - \frac{\partial g_2}{\partial x} \frac{\partial g_1}{\partial y} = \\
 &= (2a_1x + b_1y + d_1 - 1)(b_2x + 2c_2y + e_2 - 1) \\
 &\quad - (2a_2x + b_2y + d_2)(b_1x + 2c_1y + e_1) \equiv 0
 \end{aligned} \tag{A.13}$$

which gives:

$$\begin{aligned}
 &2(a_1b_2 - a_2b_1)x^2 + 4(a_1c_2 - a_2c_1)xy + 2(b_1c_2 - b_2c_1)y^2 \\
 &+ [b_2(d_1 - 1) - b_1d_2 + 2a_1(e_2 - 1) - 2a_2e_1]x \\
 &+ [2c_2(d_1 - 1) - 2c_1d_2 + b_1(e_2 - 1) - b_2e_1]y \\
 &+ [(d_1 - 1)(e_2 - 1) - d_2e_1] \equiv 0
 \end{aligned}$$

But since this expression must be identically zero for any values of x and y , it follows that Eq. (A.12) is singular when all of the following conditions are simultaneously satisfied:

$$\begin{aligned}
 a_1b_2 &= a_2b_1 \\
 a_1c_2 &= a_2c_1 \\
 b_1c_2 &= b_2c_1 \\
 b_2(d_1 - 1) + 2a_1(e_2 - 1) &= b_1d_2 + 2a_2e_1 \\
 2c_2(d_1 - 1) + b_1(e_2 - 1) &= 2c_1d_2 + b_2e_1 \\
 (d_1 - 1)(e_2 - 1) &= d_2e_1
 \end{aligned} \tag{A.14}$$

² Note that $g_1(x,y)$ and $g_2(x,y)$ in this Jacobian are the functions in the left hand side of Eqs. (A.12), while $g_1(x,y)$ and $g_2(x,y)$ in the Jacobian (A.8) are the functions in the right hand side of Eqs. (A.7).

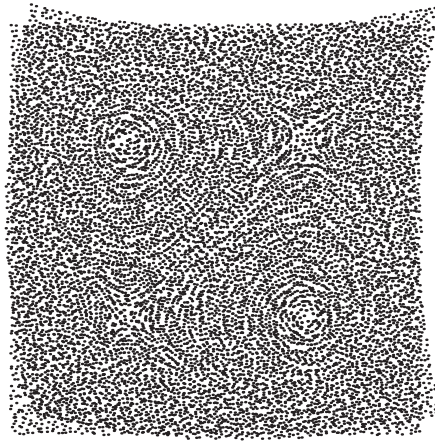


Figure A.1: An example with 4 fixed points: The superposition of two originally identical aperiodic dot screens, one of which has undergone the second-order transformation $\mathbf{g}(x,y) = (x - ay^2 + x_0, y - ax^2 + y_0)$. Fig. 3.15 shows a slightly different variant in which both of the layers have undergone second-order transformations.

We call second-order polynomial mappings $\mathbf{g}(x,y)$ for which both the Jacobians (A.9) and (A.13) are not identically zero *non-degenerate* second order mappings. We obtain, therefore, the following result:

The fixed point theorem for second-order polynomial mappings: A non-degenerate second-order polynomial mapping $\mathbf{g}(x,y)$ from \mathbb{R}^2 onto itself may have up to 4 fixed points. ■

An example of a layer superposition in which one of the two layers has undergone a second-order mapping $\mathbf{g}(x,y)$ having 4 fixed points is shown in Fig. A.1. In this case we have $a_1 = 0, b_1 = 0, c_1 = -a, d_1 = 1, e_1 = 0, a_2 = -a, b_2 = 0, c_2 = 0, d_2 = 0$ and $e_2 = 1$, so that $a_1c_2 \neq a_2c_1$ in conditions (A.10) and (A.14), meaning that both of the Jacobians (A.9) and (A.13) are not identically zero. Note, however, that if $\mathbf{g}(x,y)$ is only *non-singular* (meaning that only the Jacobian (A.9) is not identically zero), it will have infinitely many fixed points, for example one or two full lines of fixed points. Such cases are illustrated in Figs. 3.5(c),(d) and 3.6; the mappings in these cases are clearly non-singular (they map \mathbb{R}^2 onto the whole of \mathbb{R}^2), and yet they have infinitely many fixed points. (Explanation: in these mappings, given by Eq. (3.19), we have $a_1 = 0, b_1 = 0, c_1 = -a, d_1 = 1, e_1 = 0, a_2 = 0, b_2 = 0, c_2 = 0, d_2 = 0$ and $e_2 = 1$. Therefore we have in conditions (A.10) $d_1e_2 \neq d_2e_1$, while in conditions (A.14) we have instead $(d_1 - 1)(e_2 - 1) = d_2e_1$ and all the equalities are satisfied.)

As we have seen above, a detailed analysis of all the different possible cases for mappings of order 2 amounts to an analysis of the intersection points between two conics

in the plane. A full discussion on the intersection of conics can be found, for example, in [Barrett97].

A.4 Mutual fixed points between two mappings; application to the moiré theory

As we have seen in Chapter 3, when we superpose two originally identical aperiodic layers (such as random or pseudo-random screens) that have undergone transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$, respectively, we may obtain in the superposition a visible Glass pattern about each mutual fixed point of the transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ (see Sec. 3.5). The fixed point theorems described in Secs. A.2 and A.3 above correspond, in fact, to the case where one of the two layer transformations, say, $\mathbf{g}_2(x,y)$, is the identity transformation, meaning that only one of the two layers has been transformed. In such cases the mutual fixed points of the two layers are obtained at the points (x,y) where $\mathbf{g}_1(x,y) = (x,y)$, which is precisely the situation described in Eqs. (A.4) and (A.11). In the more general case where both of the superposed layers have been transformed, the mutual fixed points of $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ are the points that satisfy $\mathbf{g}_1(x,y) = \mathbf{g}_2(x,y)$, namely:

$$\mathbf{g}_m(x,y) = \mathbf{g}_1(x,y) - \mathbf{g}_2(x,y) = (0,0) \quad (\text{A.15})$$

In componentwise notation, these points are the solutions of the system of equations:

$$\begin{aligned} g_{m_1}(x,y) &= g_{1,1}(x,y) - g_{2,1}(x,y) = 0 \\ g_{m_2}(x,y) &= g_{1,2}(x,y) - g_{2,2}(x,y) = 0 \end{aligned} \quad (\text{A.16})$$

where $\mathbf{g}_1(x,y) = (g_{1,1}(x,y), g_{1,2}(x,y))$, $\mathbf{g}_2(x,y) = (g_{2,1}(x,y), g_{2,2}(x,y))$ and $\mathbf{g}_m(x,y) = (g_{m_1}(x,y), g_{m_2}(x,y))$.

Therefore, in cases where both of the layers have been transformed, conditions (1)–(3) in Secs. A.2 and A.3 apply, in fact, to the solutions of equations (A.15) or (A.16). For example, the layer superposition may have a *linear* Glass pattern when $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ have a full line of mutual fixed points, i.e. when $\mathbf{g}_m(x,y) = (0,0)$ has a full continuous line (or curve) of solutions within the x,y plane. Note that these points are *not* fixed points of the transformation $\mathbf{g}_m(x,y)$ itself (the fixed points of $\mathbf{g}_m(x,y)$ are given by the solutions of $\mathbf{g}_m(x,y) - (x,y) = (0,0)$, not by the solutions of $\mathbf{g}_m(x,y) = (0,0)$).

This generalization to the case of two transformed layers is valid for *any* transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$, and not only for first- or second-order polynomial mappings.

Appendix B

The various interpretations of a 2D transformation

B.1 Introduction

Consider a system of two equations in two independent variables x and y :

$$\begin{aligned} u &= g_1(x,y) \\ v &= g_2(x,y) \end{aligned} \tag{B.1}$$

or in vector notation:

$$\mathbf{u} = \mathbf{g}(\mathbf{x}) \tag{B.2}$$

where $\mathbf{x} = (x,y)$, $\mathbf{u} = (u,v)$, and $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$. Clearly, both $g_1(x,y)$ and $g_2(x,y)$ are *scalar functions*, i.e. functions that return for each point $(x,y) \in \mathbb{R}^2$ a single real value:

$$g_1: \mathbb{R}^2 \rightarrow \mathbb{R}, \quad g_2: \mathbb{R}^2 \rightarrow \mathbb{R}$$

whereas $\mathbf{g}(x,y)$ is a *mapping* (or, equivalently, a *transformation*), i.e. a *vector function* that returns for each point $(x,y) \in \mathbb{R}^2$ a new point $(u,v) \in \mathbb{R}^2$:

$$\mathbf{g}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$$

We denote this function by a boldface letter \mathbf{g} since the value it returns, $\mathbf{g}(x,y)$, is a vector.

The mathematical relationship defined by (B.1) (or alternatively by (B.2)) can be interpreted in several different yet completely equivalent ways, as explained in the following sections. Because all of these interpretations are mathematically equivalent, we are free in each application to choose any of them according to our convenience. It is important, however, to be aware of the different interpretations, and to know which of them is being used in each case, in order to avoid any possible confusions.

B.2 Interpretation as two surfaces over the plane or as two sets of level lines

Each of the two real valued functions of the system (B.1) defines a surface (manifold) $z = g(x,y)$ over the x,y plane, where z represents the altitude of the surface at the point (x,y) in terms of the vertical axis, perpendicularly to the x,y plane. Therefore, the 2D mapping $\mathbf{g}(x,y)$ can be interpreted geometrically as a pair of surfaces (see Fig. B.1).

The level lines (or level curves) of each of these surfaces are given by:

$$g_i(x,y) = \text{const.}$$

In particular, the set of points (x,y) satisfying $g_i(x,y) = 0$ (i.e. the solution of the equation $g_i(x,y) = 0$) can be interpreted as the zero level line of the surface $g_i(x,y)$, namely, the intersection of the surface with the x,y plane. Therefore, the set of points (x,y) satisfying the two equations:

$$\begin{aligned} g_1(x,y) &= 0 \\ g_2(x,y) &= 0 \end{aligned} \tag{B.3}$$

(i.e. the simultaneous solution of both equations) consists of the points of intersection between the zero level curves of $g_1(x,y)$ and the zero level curves of $g_2(x,y)$. Depending on the case, there may exist zero, one, several, or even infinitely many such intersection points. For example, if the zero level curves of the two functions are two intersecting straight lines, then the system (B.3) has a single solution (intersection point). As a further example, if the zero level lines of the two functions are second-order curves such as parabolas or ellipses, they may have up to 4 intersection points (see Sec. A.3 in Appendix A). Finally, in the case where the zero level curves of $g_1(x,y)$ and $g_2(x,y)$ coincide there exist infinitely many intersection points; and in the case where the two zero level curves are parallel to each other (or when at least one of the two surfaces does not intersect the x,y plane at all) the system (B.3) has no solutions.

In the example shown in Fig. B.1, where $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$, the zero level lines of the surface $g_1(x,y)$ are two perpendicular lines that coincide with the x and y axes, and the zero level lines of the surface $g_2(x,y)$ are two perpendicular lines that coincide with the main diagonals. Their intersection consists, therefore, of a single point at the origin. But if we consider, instead, the two functions defined by $\mathbf{g}(x,y) = (2xy - 1, y^2 - x^2 - 1)$, the zero level lines of the two surfaces become hyperbolic (see in Fig. B.1 the level lines corresponding to the altitude $z = 1$), and their intersection consists of two points.

It should be noted that the system of equations (B.3) is not equivalent to (B.1) because it only takes into consideration the subsets of the surfaces $u = g_1(x,y)$ and $v = g_2(x,y)$ where the surface altitude is zero. The generalization to any other altitudes c,k is straightforward (by considering the equations $g_1(x,y) = c$ and $g_2(x,y) = k$); but none of these equation pairs is equivalent to (B.1), either.

B.3 Interpretation as a mapping from the plane into itself

So far, we considered (B.1) as a system consisting of two scalar functions, $g_1(x,y)$ and $g_2(x,y)$. However, using an alternative interpretation, we may consider the vector function $\mathbf{g}(x,y)$ of (B.2) as a mapping from the x,y plane onto the u,v plane (or a subset thereof). Thus, the transformation \mathbf{g} maps each point (x,y) of the domain of \mathbf{g} in the x,y plane into its image point $(u,v) = \mathbf{g}(x,y)$ in the u,v plane. This is concisely expressed by the notation: $(x,y) \mapsto \mathbf{g}(x,y)$. The interpretation of \mathbf{g} as a mapping is illustrated in Fig. B.2 for the same transformation as in Fig. B.1.

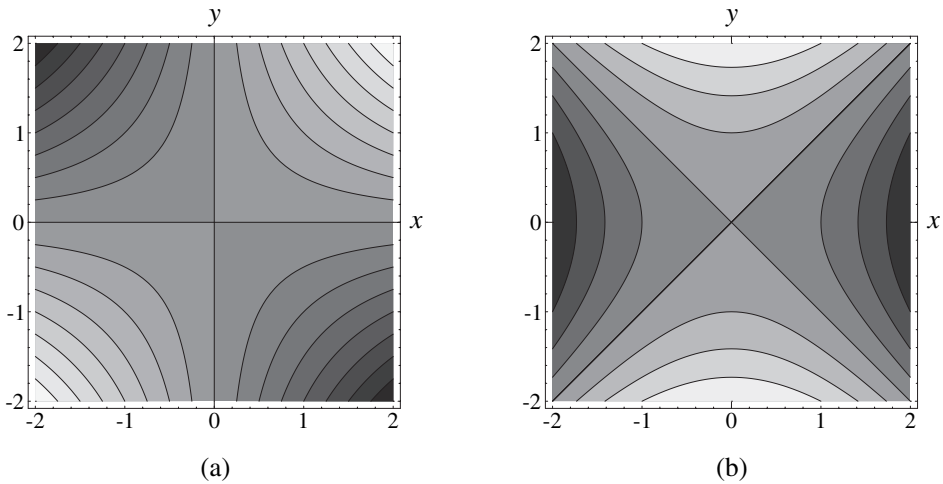


Figure B.1: Illustration of the transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a pair of surfaces. (a) The surface $g_1(x,y) = 2xy$. (b) The surface $g_2(x,y) = y^2 - x^2$. Gray levels indicate the surface altitude: brighter shades represent higher values and darker shades represent lower values. The curves plotted on each of the surfaces are its level lines.

Note, however, that a mapping $\mathbf{g}(x,y)$ does not necessarily map \mathbb{R}^2 onto the entire \mathbb{R}^2 , or even onto a 2D subregion thereof: In some degenerate cases $\mathbf{g}(x,y)$ may map \mathbb{R}^2 into a straight or a curved line, into a single point, or even into an empty set. For example, the second order transformation $\mathbf{g}(x,y) = (x, x^2)$ (namely, $u = x, v = x^2$) maps \mathbb{R}^2 onto the parabola $v = u^2$ which is only a 1D curve within the u,v plane. As another example, the transformation $u = \sqrt{x}, v = \sqrt{-x}$ maps \mathbb{R}^2 onto the single point $(0,0)$, while the transformation $u = \sqrt{x}, v = 1/\sqrt{-x}$ maps \mathbb{R}^2 onto an empty set. Such situations occur when the mapping $\mathbf{g}(x,y)$ (or equivalently, the system of equations (B.1)) is *singular*, namely, when the two equations forming the system are not *independent*.

Definition B.1: Two functions (or equations) $u = g_1(x,y)$ and $v = g_2(x,y)$ are said to be *independent* (or *functionally independent*) if there exists no function $f(u,v)$ other than $f(u,v) \equiv 0$ such that $f(g_1(x,y), g_2(x,y)) = 0$ is satisfied for all (x,y) . Equivalently, this means that the Jacobian:

$$J(x,y) = \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \frac{\partial g_1}{\partial x} \frac{\partial g_2}{\partial y} - \frac{\partial g_2}{\partial x} \frac{\partial g_1}{\partial y} \quad (\text{B.4})$$

is not identically zero (see, for example, [Bronstein90 pp. 226, 430–431] or [Courant88 pp. 154–155]). ■

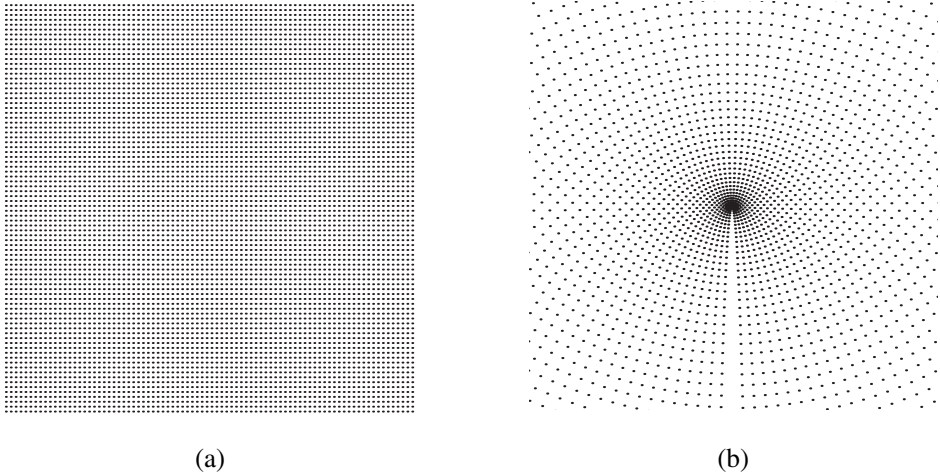


Figure B.2: Illustration of the transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a direct mapping $(x,y) \mapsto \mathbf{g}(x,y)$. (a) A periodic dot screen in the original x,y plane. (b) The transformed dot screen after each of its dots (x,y) has been moved by the mapping \mathbf{g} to its new location $\mathbf{g}(x,y)$. The “seam” along the negative part of the y axis in (b) was left intentionally, to clearly illustrate how the upper half plane of (a) is deformed around the origin in (b), and covers the entire destination plane. The lower half plane of (a) is deformed upwards in a similar way, and it covers once again the entire destination plane. See also Fig. B.4(a),(b).

If $g_1(x,y)$ and $g_2(x,y)$ are *dependent*, for instance if $g_2(x,y) = g_1(x,y)^2$, they are a consequence of each other, and hence the 2D transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ they define is singular, and it maps \mathbb{R}^2 onto a 1D curve or even into a single point or an empty set in \mathbb{R}^2 . This can be easily seen from Definition B.1: If there exists a function $f(u,v)$ other than $f(u,v) \equiv 0$ such that $f(g_1(x,y), g_2(x,y)) = 0$ for all x,y , then the image of our transformation $(u,v) = \mathbf{g}(x,y)$ satisfies $f(u,v) = 0$, which means, indeed, that the image of \mathbf{g} is a 1D curve (or a point, or even an empty set) within the 2D u,v space.

Thus, in order to avoid such degenerate cases, we must request that the functions $g_1(x,y)$ and $g_2(x,y)$ be independent.

Example B.1: Consider the system consisting of the two functions:

$$u = f(x) \cos y$$

$$v = f(x) \sin y$$

This transformation is, in fact, a generalization of the polar to Cartesian coordinate transformation $u = r \cos \theta$, $v = r \sin \theta$ where the radius length r is replaced by its modulated version $f(r)$. It is easy to see that the Jacobian of this transformation is given by:

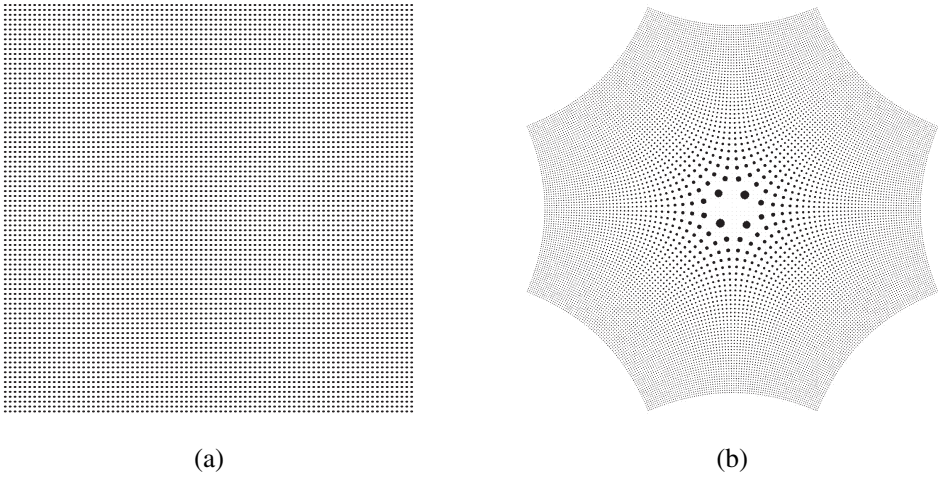


Figure B.3: Illustration of $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a domain transformation. It operates on the original image $r(x,y)$ shown in (a), and gives the transformed image $r(\mathbf{g}(x,y))$ shown in (b). The original image $r(x,y)$ in (a) is the same periodic dot screen as in Fig. B.2(a). Note that the left half plane of (a) is mapped twice into the two left quadrants of (b); similarly, the right half plane of (a) is also mapped twice into the two right quadrants of (b). See also Fig. B.4(c),(d).

$$J(x,y) = f(x) \frac{d}{dx}f(x) \cos^2 y + f(x) \frac{d}{dx}f(x) \sin^2 y = f(x) \frac{d}{dx}f(x)$$

If we take, for example, $f(x) = \sin x$ we obtain the system:

$$\begin{aligned} u &= \sin x \cos y \\ v &= \sin x \sin y \end{aligned} \tag{B.5}$$

These two functions are clearly independent, since their Jacobian is not identically zero; and indeed, they map \mathbb{R}^2 onto a 2D subregion of \mathbb{R}^2 , the entire unit disk. On the other hand, if we take $f(x) = 1$ then the Jacobian becomes identically zero, meaning that the two functions $u = \cos x$, $v = \sin x$ are dependent; and indeed, this system maps \mathbb{R}^2 onto a 1D curve within \mathbb{R}^2 , the *perimeter* of the unit circle.

It may be also instructive to see what happens to the independent system $u = r \cos \theta$, $v = r \sin \theta$ (that maps \mathbb{R}^2 onto \mathbb{R}^2) as it gradually approaches its dependent counterpart $u = \cos \theta$, $v = \sin \theta$ (that maps \mathbb{R}^2 onto a 1D curve). This can be done, for example, by observing the system $u = (1+\varepsilon r) \cos \theta$, $v = (1+\varepsilon r) \sin \theta$ while ε gradually tends to zero. ■

Remark B.1: When $(u,v) = \mathbf{g}(x,y)$ is a *linear* transformation, the two surfaces defined over the x,y plane by the functions $u = g_1(x,y)$, $v = g_2(x,y)$ (see Sec. B.2) are planes that pass through the origin. Hence, the equations $g_1(x,y) = 0$, $g_2(x,y) = 0$ may have either a

single common solution at the origin (if the zero level lines of the two planes have a single intersection point), a full straight line of solutions passing through the origin (if the two planes have a common zero level line), or a full plane of solutions (in the degenerate case where both planes fully coincide with the x,y plane).¹ Now, because \mathbf{g} is a linear transformation, it must satisfy the relationship $\dim \text{Ker } \mathbf{g} + \dim \text{Im } \mathbf{g} = 2$ (see Sec. 5.4.1 in *Vol. I*). It follows, therefore, that $g_1(x,y)$ and $g_2(x,y)$ are independent (i.e. $\dim \text{Im } \mathbf{g} = 2$) iff they have a single solution point (i.e. $\dim \text{Ker } \mathbf{g} = 0$). Thus, if $g_1(x,y)$ and $g_2(x,y)$ have a full continuous line of solutions in common, they must be dependent and the transformation $\mathbf{g}(x,y)$ necessarily maps \mathbb{R}^2 onto a 1D line. For example, in the case of the linear transformation $\mathbf{g}(x,y) = (x, 2x)$ the planes $u = g_1(x,y) = x$ and $u = g_2(x,y) = 2x$ have a full continuous line of solutions in common along the y axis; and indeed, $\mathbf{g}(x,y)$ maps the entire plane \mathbb{R}^2 into the line $y = 2x$.²

It is interesting to note, however, that the situation in non-linear transformations is more flexible than in the linear case: The two functions $u = g_1(x,y)$, $v = g_2(x,y)$ may have a full continuous curve of solutions in common (or even several or infinitely many such curves) without necessarily being dependent (and hence, without implying that the system (B.1) is singular and maps its entire 2D domain into a 1D curve). For example, both of the functions $u = x$, $v = xe^y$ have a full line of solutions along the y axis; but they are still independent (their Jacobian is not identically zero), and $\mathbf{g}(x,y) = (x, xe^y)$ still maps \mathbb{R}^2 onto a 2D subrange of \mathbb{R}^2 (the first and third quadrants of the plane). As a second example, consider the two functions of (B.5). These functions are clearly independent, and indeed, they map \mathbb{R}^2 onto a 2D subrange of \mathbb{R}^2 (the unit disk). And yet, they have infinitely many continuous lines of solutions in common (all the vertical lines $x = n\pi$, $n \in \mathbb{Z}$). Incidentally, in this case each of the two functions has also infinitely many additional zeros that are not shared with the other: The first function has the horizontal lines $y = (m + \frac{1}{2})\pi$, $m \in \mathbb{Z}$ as zeros, while the second function has the horizontal lines $y = m\pi$, $m \in \mathbb{Z}$ as zeros.

It turns out that if the Jacobian is non-zero at a solution point (x_0, y_0) of the system $(u,v) = \mathbf{g}(x,y)$ then that solution point is isolated [Howse95 p. 14]. Note, however, that although the converse is true for linear transformations, it is not necessarily true in the general case. For example, the point $(0,0)$ is clearly an isolated solution of the system $(u,v) = (2xy, y^2 - x^2)$, and yet the Jacobian $J(x,y) = 4x^2 + 4y^2$ vanishes at this point. ■

B.4 Interpretation as a domain transformation $r(\mathbf{g}(x,y))$

Suppose we are given a scalar function (i.e. a surface) $r(u,v)$, and that we apply to it the transformation $(u,v) = \mathbf{g}(x,y)$. The resulting distorted function (or surface) is given,

¹ Note that the case with no solutions at all (where the two planes, and hence their zero level lines, are parallel to each other) is excluded when \mathbf{g} is linear, since both planes of a linear transformation must pass through the origin. This case may occur, however, if \mathbf{g} is an affine transformation.

² More generally, this situation occurs in all linear transformations having the form $u = ax + by$, $v = s(ax + by)$. The common zeros of these two planes (i.e. $\text{Ker } \mathbf{g}$) consist of the entire line $ax + by = 0$, and the image of the transformation ($\text{Im } \mathbf{g}$) consists of the line $v = su$.

therefore, by $r(\mathbf{g}(x,y))$. This is illustrated in Fig. B.3, where the original function $r(u,v)$ is shown in (a), and the resulting distorted function $r(\mathbf{g}(x,y))$ is shown in (b). This figure uses, again, the same transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as in Fig. B.2, and yet, the effect of the transformation seems to be completely different: While in Fig. B.2 the distortion generated by $\mathbf{g}(x,y)$ seems to be parabolic, in Fig. B.3 the distortion seems to be hyperbolic. How can we explain this fact?

As pointed out in Sec. D.6 of Appendix D, each transformation $\mathbf{g}(x,y)$ can be used in two different ways: either as a direct transformation, or as a domain, inverse transformation. Consider, for example, the transformation $\mathbf{g}(x,y) = (2x, 2y)$. Clearly, this transformation maps each point (x,y) to the new location $(2x, 2y)$, and thus it expands the original layer by two. This is, indeed, the interpretation of $\mathbf{g}(x,y)$ as a *direct* transformation. However, when the same transformation $\mathbf{g}(x,y)$ is used as a *domain* transformation, for example, when it acts on the original layer $r(u,v)$ to give $r(2x, 2y)$, its effect is inverted: $r(2x, 2y)$ is a two-fold shrunk version of $r(u,v)$, while the two-fold expansion of $r(u,v)$ is expressed by $r(x/2, y/2)$, i.e. by using the *inverse* transformation $\mathbf{g}^{-1}(x,y) = (x/2, y/2)$. This inversion effect of domain transformations is explained in detail in Sections D.6 and D.10 of Appendix D; see also Remark 4.1 in Sec. 4.4.

Returning now to our case, we see that the transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ is used in Fig. B.2 as a direct transformation $(x,y) \mapsto (2xy, y^2 - x^2)$, while in Fig. B.3 it is used as a domain (and hence, inverse) transformation that distorts $r(u,v)$ into $r(2xy, y^2 - x^2)$. This explains, indeed, the different geometric shapes that are obtained by the same transformation in Figs. B.2 and B.3. The effects of the same transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a direct transformation and as an inverse transformation are also illustrated in Fig. B.4.

The lesson is, therefore, that in situations where $\mathbf{g}(x,y)$ can be used in both ways, it is important to clearly indicate which of the interpretations of $\mathbf{g}(x,y)$ is intended, in order to avoid any possible confusion.

B.5 Interpretation as a coordinate change

If we consider the level lines of the function $u = g_1(x,y)$ and the level lines of the function $v = g_2(x,y)$ as two sets of curvilinear coordinates, we may interpret system (B.1) or its vector representation (B.2) as a transformation from Cartesian to curvilinear coordinates in the plane. In other words, we can regard (B.1) or (B.2) as a mapping from \mathbb{R}^2 onto itself that defines a new curvilinear coordinate system u,v in the plane, instead of the original Cartesian coordinate system x,y (see Fig. B.1 and Fig. B.4(d)).

According to this interpretation of $\mathbf{g}(x,y)$, the zero level curves $g_1(x,y) = 0$ and $g_2(x,y) = 0$ are simply the curvilinear axes $u = 0$ and $v = 0$ of the new curvilinear coordinate system defined by the transformation $\mathbf{g}(x,y)$. The other coordinate curves $u = m$ and $v = n$ for all

$m, n \in \mathbb{Z}$ are given by the two curve families defined by $g_1(x, y) = m$ and $g_2(x, y) = n$, which are, respectively, integer level lines of $g_1(x, y)$ and integer level lines of $g_2(x, y)$.

It is interesting to note, however, that the curvilinear coordinate system obtained from the level lines of $u = g_1(x, y)$ and $v = g_2(x, y)$ corresponds to the effect of $\mathbf{g}(x, y)$ as an *inverse* transformation (for example, in the case shown in Fig. B.1 we obtain a hyperbolic coordinate net rather than a parabolic one). This phenomenon is explained in detail in Sec. D.4 of Appendix D.

Obviously, in order to provide a useful coordinate system the transformation $(u, v) = \mathbf{g}(x, y)$ must satisfy the condition $J(x, y) \neq 0$ set by definition B.1. We have seen in Sec. B.3 that this condition eliminates the risk that $\mathbf{g}(x, y)$ maps \mathbb{R}^2 into a degenerate subset of \mathbb{R}^2 such as a 1D curve, a single point, or an empty set. In terms of coordinate lines, the condition $J(x, y) \neq 0$ guarantees that the two components of $(u, v) = \mathbf{g}(x, y)$, namely, $u = g_1(x, y)$ and $v = g_2(x, y)$, do not have exactly the same level curves [Kaplan03 p. 162]. But while this condition is obviously *necessary*, it is not yet *sufficient* to guarantee that the resulting coordinate system is useful, and it does not exclude other pathologic situations that can make our new coordinate system useless. For instance, each of the surfaces $u = g_1(x, y)$, $v = g_2(x, y)$ still may have several disjoint zero level curves, meaning that each of the resulting curvilinear coordinate axes $g_1(x, y) = 0$ and $g_2(x, y) = 0$ consists of several disjoint branches. This occurs, for example, in the case of $\mathbf{g}(x, y) = (2xy, y^2 - x^2)$, where each of the surfaces has two perpendicular zero level lines (see Fig. B.1), or in its variant $\mathbf{g}(x, y) = (2xy - 1, y^2 - x^2 - 1)$, where each of the surfaces has a hyperbolic zero level line composed of two disjoint branches. Furthermore, the curvilinear axes $g_1(x, y) = 0$ and $g_2(x, y) = 0$ may have several intersection points even if the Jacobian is not identically zero; this may happen, for example, if the axes are second order curves such as parabolas, hyperbolas or ellipses, since such curves may have up to 4 intersection points. Even worse, as we have already seen in Remark B.1, the two functions $u = g_1(x, y)$, $v = g_2(x, y)$ may have coinciding zero level curves (and hence generate a useless, degenerate coordinate system) even if their Jacobian is not identically zero: Indeed, the condition $J(x, y) \neq 0$ excludes the possibility that *all* the level curves of $g_1(x, y)$ and $g_2(x, y)$ be identical, but $g_1(x, y)$ and $g_2(x, y)$ still may have *some* (and even infinitely many) coinciding level lines. For example, as we have seen in Remark B.1, they may have coinciding zero level lines, and thus give coinciding coordinate axes. It is clear, therefore, that in order to obtain a useful coordinate system we need a stronger condition than simply having a non-identically zero Jacobian $J(x, y)$. For example, we may require that $g_1(x, y)$ and $g_2(x, y)$ satisfy also the following identities, which are known as the Cauchy-Riemann conditions:

$$(a) \quad \frac{\partial g_1}{\partial x} = \frac{\partial g_2}{\partial y}, \quad \frac{\partial g_1}{\partial y} = -\frac{\partial g_2}{\partial x} \quad \text{or} \quad (b) \quad \frac{\partial g_1}{\partial x} = -\frac{\partial g_2}{\partial y}, \quad \frac{\partial g_1}{\partial y} = \frac{\partial g_2}{\partial x} \quad (B.6)$$

In this case, the transformation $\mathbf{g}(x, y)$ is called *conformal* [Courant88 pp. 166–167], and it maps the straight lines $x = \text{const.}$, $y = \text{const.}$ into curve families $u = \text{const.}$ and $v = \text{const.}$ which intersect *at right angles*. This orthogonality is clearly stronger than the mere independence of $g_1(x, y)$ and $g_2(x, y)$; and indeed, condition (a) implies $J(x, y) > 0$, and

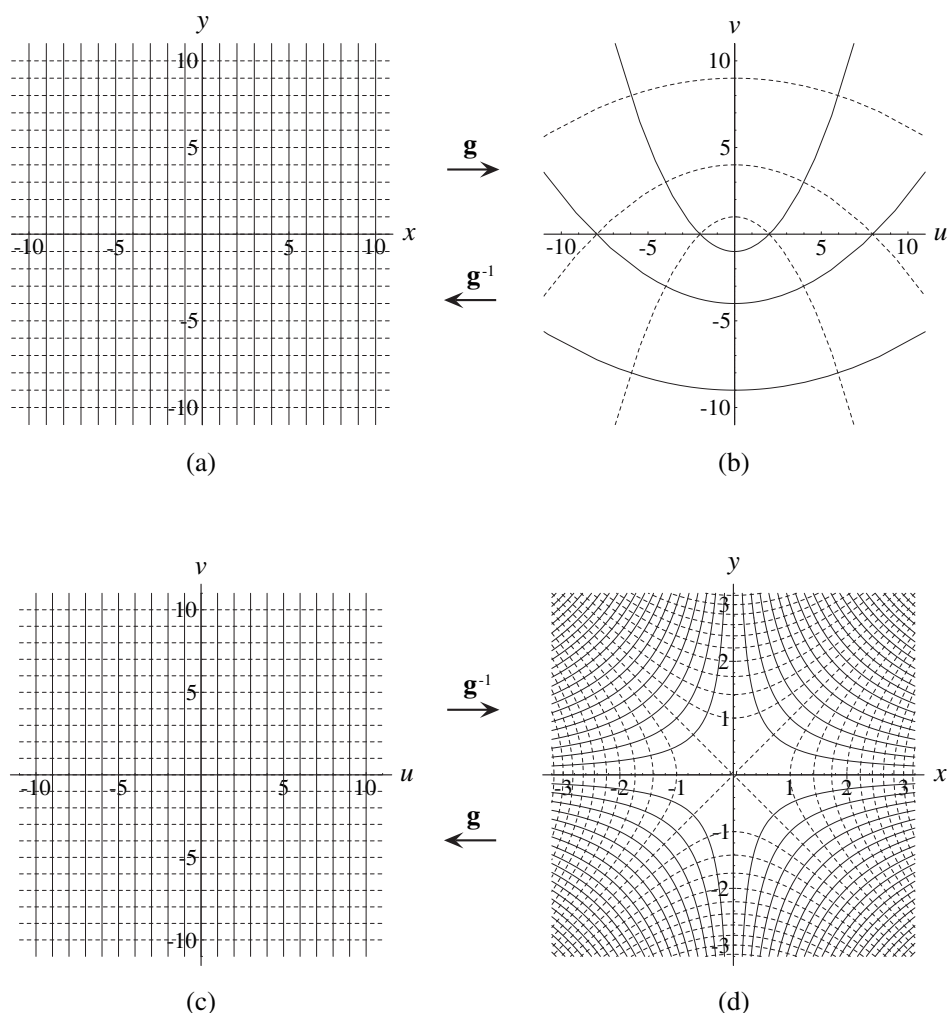


Figure B.4: (a),(b) Representation of the same transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a coordinate change in \mathbb{R}^2 . (c),(d) Representation of the inverse transformation \mathbf{g}^{-1} as a coordinate change in \mathbb{R}^2 (see Example D.5 in Appendix D).

condition (b) implies $J(x,y) < 0$. Such an orthogonality is not *required* for having a useful coordinate system (see for instance Fig. 10.2(b) in *Vol. I*), but it is certainly advantageous. But on the other hand, this condition is not yet sufficient for excluding cases with multiple-branch coordinate axes or with coordinate axes having several intersection points. For example, although the transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ shown in Fig. B.1 is clearly conformal, each of its new coordinate axes $g_1(x,y) = 0$ and $g_2(x,y) = 0$ consists of

two perpendicular lines. Usually such pathologies can be resolved, however, by considering our transformation on a suitable subrange of \mathbb{R}^2 .

B.6 Interpretation as a 2D vector field

Another useful interpretation is obtained by considering the vector function $\mathbf{g}(x,y)$ of Eq. (B.2) as a *vector field*. A vector field in \mathbb{R}^2 is a function $\mathbf{g}(x,y)$ that assigns to each point (x,y) in the x,y plane a vector $(u,v) = \mathbf{g}(x,y)$. Well known examples in physics include electric or magnetic fields as well as the gravitation field of the earth, all of which are vector fields that are defined in the 3D space \mathbb{R}^3 .

A vector field in \mathbb{R}^2 can be illustrated graphically by drawing an arrow emanating from each point (x,y) of the x,y plane (or, more practically, from some representative points on a given grid within the x,y plane), where the length and the orientation of each arrow indicate the length and the orientation of the vector $\mathbf{g}(x,y)$ that has been assigned by \mathbf{g} to the point (x,y) (see Fig. B.5(a)).³ It is important to note, however, that each such arrow does not connect the point (x,y) to its destination $\mathbf{g}(x,y)$ under the transformation \mathbf{g} , but rather to the point $(x,y) + \mathbf{g}(x,y)$. Note also that the null vector $(0,0)$ is assigned to a point (x,y) iff (x,y) is a solution of the system (B.3). As we have seen, depending on the case there may exist one such point, several such points, infinitely many, or even none at all. It is important to stress, however, that these points are *not* fixed points of $\mathbf{g}(x,y)$, since they do not satisfy $\mathbf{g}(x,y) = (x,y)$ (they *are*, however, fixed points of the transformation $\mathbf{g}(x,y) + (x,y)$, since they do satisfy, of course, $\mathbf{g}(x,y) + (x,y) = (x,y)$).

The vector field interpretation of $\mathbf{g}(x,y)$ is closely related to the interpretation of $\mathbf{g}(x,y)$ as a direct mapping. For example, by comparing Fig. B.5 with Fig. B.4 it can be seen that the vector $\mathbf{g}(x,y)$ attached to the point $(x,y) = (1,0)$ of the vector field is precisely the destination of the point $(1,0)$ under the *direct* transformation, namely $(0,-1)$ (Fig. B.4(b)), and not the destination of $(1,0)$ under the *inverse* transformation, which is $(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$ (Fig. B.4(d)). And yet, as we can clearly see in the figures, the geometric shapes of the direct transformation and of its corresponding vector field may look significantly different. In fact, all the various representations of the same transformation $\mathbf{g}(x,y)$ — as a vector field, as a direct mapping and as an inverse mapping — can be completely different from each other. We will return to this point in more detail in Sec. B.7.

As an alternative way of visualizing a vector field one may also draw its *trajectories* (also known as *field lines*). Loosely speaking, these are the curves obtained by following the arrows in Fig. B.5(a) and joining them into continuous curves in the x,y plane (see Fig. B.5(b)). More precisely, trajectories (or field lines) are curves for which the tangent vector to the curve at each point (x,y) is exactly $\mathbf{g}(x,y)$. Thus, at every point (x,y) , the direction in

³ For practical reasons it is customary to scale the arrow length in the drawing by a constant factor, in order to avoid drawings with too short, hard-to-see arrows, or drawings with too long, overlapping arrows.

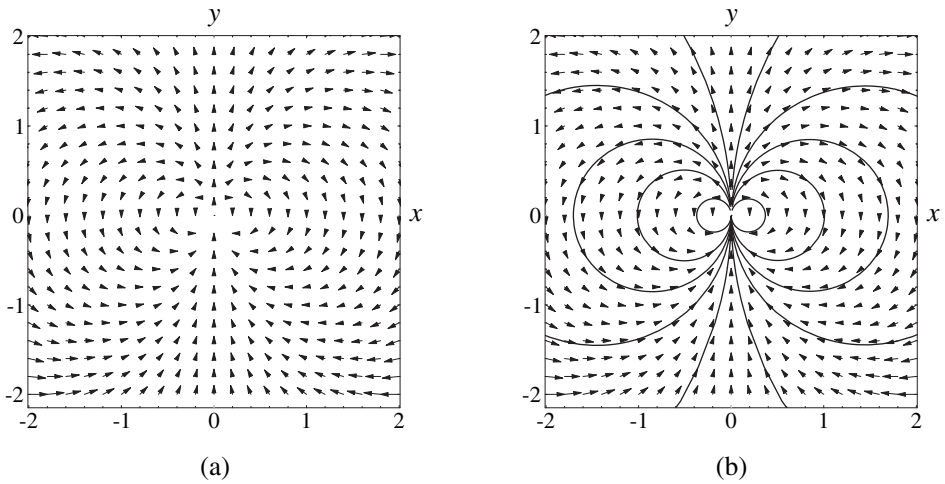


Figure B.5: (a) Illustration of the same transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ as a vector field in \mathbb{R}^2 . (b) Some trajectories of this vector field. Note that these trajectories are given by the family of the vertically tangent circles $(x - c)^2 + y^2 = c^2$ (i.e. $y = \sqrt{2cx - x^2}$) for all possible values of the constant $c \in \mathbb{R}$; this can be easily shown by calculating dx/dy and verifying that it satisfies Eq. (B.9), namely: $dx/dy = 2xy/(y^2 - x^2)$.

which the trajectory runs is determined by the vector $\mathbf{g}(x,y)$. Note that except for points where $\mathbf{g}(x,y)$ is not defined or where $\mathbf{g}(x,y) = \mathbf{0}$, every point in the plane belongs to one and only one trajectory. This also means, up to the same exceptions, that trajectories do not intersect. The advantage of drawing the trajectories of a vector field is that they are easier to picture visually. However, although the trajectories clearly show the directions of a vector field, they do not convey the information on the length of its vectors, and hence they are not a full representation of $\mathbf{g}(x,y)$. We will return to this point in Remark B.2 below.

The trajectories of a vector field $\mathbf{g}(x,y)$ are given in the parametric form by a family of curves $(x(t), y(t))$ whose members differ from each other by a constant c ; these curves are the solutions of the system of differential equations (see [Bronstein97 p. 526]):

$$\begin{aligned} \frac{d}{dt}x(t) &= g_1(x(t), y(t)) \\ \frac{d}{dt}y(t) &= g_2(x(t), y(t)) \end{aligned} \tag{B.7}$$

where t is the parameter of each of the curves.⁴ Using the vector notation $\mathbf{x}(t) = (x(t), y(t))$ this can be written more compactly as:

⁴ Remember that a curve in the x,y plane can be generally defined in Cartesian coordinates in three equivalent forms: *explicitly* by $y = h(x)$, *implicitly* by $f(x,y) = 0$, or *parametrically* by $x = f_1(t)$, $y = f_2(t)$, where the parameter t varies continuously throughout an interval such as $-\infty < t < \infty$ [Bronstein97 pp. 75–76]. Conversions between these forms can be done as explained in [Bronstein97 p. 551], but they are not always possible [Harris98 p. 121].

$$\mathbf{x}'(t) = \mathbf{g}(\mathbf{x}(t)) \quad (\text{B.8})$$

Note the close similarity between the equations of the vector field $\mathbf{g}(x,y)$ (Eqs. (B.1) or (B.2)) and the parametric equations defining its trajectories (Eqs. (B.7) or (B.8), respectively). But if we prefer the explicit form $y = f(x)$ of the trajectories instead of their parametric form given by $(x(t), y(t))$, then the corresponding differential equation that defines them is:⁵

$$\frac{dy}{dx} = \frac{g_2(x,y)}{g_1(x,y)} \quad (\text{B.9})$$

Ways of solving systems of differential equations such as (B.7) can be found, along with many illustrative examples and figures showing their trajectories, in Chapter 4 of [Kreyszig93]. Equivalent ways for the solution of the corresponding differential equation (B.9) can be found, for example, in [Bronshtein97 pp. 395–398]. A complete classification of the different trajectory shapes for *linear* differential equation systems, including nodes, saddle points, center points, spirals, etc. can be found in [Kreyszig93 pp. 176–178], [Gray97 pp. 551–567] or in Sec. 1.4 of [Tabor89]. A similar classification for *non-linear* differential equation systems can be found in [Gray97 pp. 586–602].

Remark B.2: It should be remembered that because the trajectories do not contain all the information of the vector field (they only convey the vector directions but not their lengths), a vector field $\mathbf{g}(x,y)$ cannot be uniquely determined by its trajectories $y = f(x)$. For example, the two vector fields $\mathbf{g}(x,y) = (2y, -2x)$ and $\mathbf{g}(x,y) = (-y/(x^2+y^2), x/(x^2+y^2))$ have the same trajectories (a family of circles $x^2 + y^2 = c^2$ for any constant c), the difference being only in the vector lengths along these circles. More generally, if $\mathbf{g}(x,y)$ has circular or radial trajectories, it is clear that all the vector fields $h(r)\mathbf{g}(x,y)$ with $r = \sqrt{x^2 + y^2}$ have the same trajectories $y = f(x)$ as $\mathbf{g}(x,y)$. But if we use the parametric form of the trajectories, $(x(t), y(t))$, then the trajectories may be expressed differently for each of these vector fields, the difference being not in the shape of the curves but in their tracing speed in terms of the parameter t . For example, $(\cos t, \sin t)$ and $(\cos 2t, \sin 2t)$ represent the same circle, the difference being only in the parametrization (the speed of drawing the curves as t advances). Thus, if we consider the tracing speed of the trajectories as an indication to the vector lengths in $\mathbf{g}(x,y)$, the parametric form of the trajectories may convey both the vector directions and the vector lengths of the vector field (up to a constant).

An alternative way for sorting out this problem consists of drawing the trajectories so that the strength of the vector field (i.e. the vector lengths) is represented by the density of the trajectories [Needham97 pp. 453 and 494]: the closer together the trajectories, the

⁵ Note that the explicit form of the trajectories, which is obtained by solving the differential equation (B.8), may have singular points wherever $g_1(x,y) = 0$ (i.e. at vertical tangencies of the solution). An advantage of the parametric form of the trajectories, which is obtained by solving the system of differential equation (B.7), is that such points are no longer singular points [Birkhoff89 p. 134]. Another advantage of the parametric form is that it explicitly indicates the *direction* of each trajectory: The positive sense of a trajectory is defined as the sense in which the curve is traced out for increasing values of t [Kreyszig93 p. 457].

stronger the vector field (just as the density of the level lines in a topographic map indicates the steepness of the ground). ■

Remark B.3: Note that in physics the trajectories of a vector field are often interpreted as curves which trace the motion of particles under the influence of the field. If the parameter t is understood as time, the trajectory $(x(t), y(t))$ gives the path of the particle, namely, the location of the particle at each moment t . The derivative of this curve, $(\frac{d}{dt}x(t), \frac{d}{dt}y(t))$, given by Eqs. (B.7), defines the *velocity* of the particle (which is a vectorial entity, too) at each moment t . For other possible physical interpretations see [Needham97 pp. 451–454]. ■

Remark B.4: Although in many cases it is easy to guess intuitively the trajectories (solution curves) of a differential equation from its vector field, it is a well known fact that in some cases this task is not as easy as it sounds. More details on this subject, as well as several illustrated examples, can be found in [Schwalbe97, pp. 39–42 and 69]. ■

Remark B.5: There exists another remarkable visual difference between $\mathbf{g}(x,y)$ as a transformation and $\mathbf{g}(x,y)$ as a vector field, which concerns their behaviour under various symmetry operations: When considered as transformations, $\mathbf{g}(-x,-y)$ is a global 180° rotation of $\mathbf{g}(x,y)$ and $\mathbf{g}(x,-y)$ is a global vertical reflection of $\mathbf{g}(x,y)$ (see Sec. C.2 in Appendix C and the figures therein). However, when considered as vector fields, $\mathbf{g}(-x,-y)$ only differs from $\mathbf{g}(x,y)$ in the sense of its vectors, while $\mathbf{g}(x,-y)$ looks completely different. For example, the vector field $\mathbf{g}(x,y) = (x,y)$ consists of radial trajectories emanating from the origin, and the vector field $\mathbf{g}(x,y) = (-x,-y)$ consists of radial trajectories pointing to the origin; but the vector field $\mathbf{g}(x,y) = (x,-y)$ has a completely different shape, and it consists of hyperbolic trajectories. One should be aware of such differences in order to avoid mistakes when trying to interpret intuitively the meaning of $\mathbf{g}(-x,-y)$, $\mathbf{g}(x,-y)$, etc. ■

B.7 Relationship between the different representations of $\mathbf{g}(x,y)$

As we have seen, any transformation $\mathbf{g}(x,y)$ can be interpreted in several different ways, whose graphical representations can be very different. For example, Figs. B.4(b), B.4(d) and B.5 show the graphical representations of the same transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ when it is interpreted, respectively, as a direct transformation, as a domain transformation, and as a vector field. What are the mathematical relationships between the curve families that represent the same transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ in its various interpretations? The answer is as follows: The level lines of $\mathbf{g}(x,y)$ when it is viewed as a *domain* transformation (in our example, the two families of curves shown in Figs. B.1 or B.4(d)) are given by the curve families $g_1(x,y) = \text{const.}$ and $g_2(x,y) = \text{const.}$ The level lines of $\mathbf{g}(x,y)$ when it is viewed as a *direct* transformation (in our example, the two families of curves shown in Fig. B.4(b)) are given by the curve families $g_1^{-1}(x,y) = \text{const.}$ and $g_2^{-1}(x,y) = \text{const.}$, where $g_1^{-1}(x,y)$ and $g_2^{-1}(x,y)$ are the two components of the inverse transformation $\mathbf{g}^{-1}(x,y)$, namely: $\mathbf{g}^{-1}(x,y) = (g_1^{-1}(x,y), g_2^{-1}(x,y))$. And finally, the trajectories

(field lines) of $\mathbf{g}(x,y)$ when it is viewed as a vector field (in our example, the curve family shown in Fig. B.5(b)) are given by the family of parametric curves $(x(t),y(t))$ that are the solutions of the system of differential equations (B.7).

It is interesting to note that given a real valued function (i.e. a surface) $z = g(x,y)$ over the x,y plane, the gradient of $g(x,y)$, denoted by $\nabla g(x,y)$, gives at each point (x,y) a vector defining the maximal slope of $g(x,y)$ at this point. This is, indeed, a vector field whose definition is:

$$\nabla g(x,y) = \left(\frac{\partial}{\partial x} g(x,y), \frac{\partial}{\partial y} g(x,y) \right) \quad (\text{B.10})$$

For example, for the paraboloid $g(x,y) = x^2 + y^2$ we have: $\mathbf{g}(x,y) = \nabla g(x,y) = (2x, 2y)$. The trajectories of this vector field are the lines of maximal slope of $g(x,y)$; they are called *gradient lines* or *gradient curves* of $g(x,y)$. Note that the gradient lines of $g(x,y)$ are orthogonal to the level lines of $g(x,y)$. In our example of the paraboloid the trajectories of the vector field (B.10) are given by the following system of linear differential equations:

$$\begin{aligned} \frac{d}{dt}x(t) &= 2x(t) \\ \frac{d}{dt}y(t) &= 2y(t) \end{aligned} \quad (\text{B.11})$$

whose solution curves consist of the family of straight lines that is given in parametric form by $x(t) = c_1 e^t$, $y(t) = c_2 e^t$ for any constants c_1, c_2 , or in explicit form by $y = cx$ for any constant c [Kreyszig93 p. 168]). These lines are, indeed, the gradient lines of our paraboloid.

Note, however, that while for every reasonably well behaved surface $g(x,y)$ there exists a vector field $\mathbf{g}(x,y)$ such that $\mathbf{g}(x,y) = \nabla g(x,y)$, the converse is not necessarily true: Not every transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ can be represented as a gradient field of some surface $g(x,y)$. For example, the transformation $\mathbf{g}(x,y) = (2y, -2x)$ has no surface $g(x,y)$ such that $\frac{\partial}{\partial x} g(x,y) = 2y$ and $\frac{\partial}{\partial y} g(x,y) = -2x$, since this would imply by integration that $g(x,y) = 2xy + c_1(y)$ and $g(x,y) = -2xy + c_2(x)$, but these two conditions on $g(x,y)$ are contradictory.⁶ On the other hand, the transformation $\mathbf{g}(x,y) = (2x, 2y)$ does have a surface

⁶ This can be also explained geometrically: The trajectories of $\mathbf{g}(x,y) = (2y, -2x)$ are concentric circles about the origin; their parametric and implicit expressions are given below after Eq. (B.14). Note, however, that it is only the additional information conveyed by the parametric expression of these circles (their relative “tracing speed” as a function of the parameter t) or, equivalently, the information provided by the vector lengths within the vector field $\mathbf{g}(x,y)$, that prevents these circles from being gradient lines of any surface $g(x,y)$. For if we only consider the geometric shape of these circles, as it is conveyed by their implicit expression $(x^2 + y^2 = c^2 \text{ for any constant } c)$, there do exist surfaces $g(x,y)$ having these circles as gradient lines. For example, these circles are the gradient lines of the helicoid $g(x,y) = a \arctan(y/x)$ [Weinstein99 p. 810]: indeed, the gradient field of this surface is $\nabla g(x,y) = (-y/(x^2 + y^2), x/(x^2 + y^2))$, whose trajectories have the same implicit expression $x^2 + y^2 = c^2$ as the trajectories of our vector field $\mathbf{g}(x,y) = (2y, -2x)$. The difference lies in the evolution of the tracing speed (or of the vector lengths) between the inner and the outer circles, information which is only conveyed by the parametric expression of these circles. To see this, note that in the helicoid the steepness of the gradient curves along the surface increases as their radius gets smaller (meaning that the inner vectors of the vector field are longer), while in the surface that would have as its gradient lines the circular trajectories of $\mathbf{g}(x,y) = (2y, -2x)$, the steepness of the gradient curves along the surface would remain identical for all radiuses (since the vector lengths within the vector field increase linearly with the radius); but this is geometrically impossible.

$g(x,y)$ such that $\frac{\partial}{\partial x}g(x,y) = 2x$ and $\frac{\partial}{\partial y}g(x,y) = 2y$, since by integration we have $g(x,y) = x^2 + c_1(y)$ and $g(x,y) = y^2 + c_2(x)$, and indeed, taking $c_1(y) = y^2$ and $c_2(x) = x^2$ we obtain $g(x,y) = x^2 + y^2$.

A vector field $\mathbf{g}(x,y)$ for which there exists a surface $g(x,y)$ such that $\mathbf{g}(x,y) = \nabla g(x,y)$ is said to be a *conservative* vector field [Kreyszig93 p. 479; Weisstein99 p. 311]; in this case $g(x,y)$ is said to be a *potential function* of $\mathbf{g}(x,y)$. And it turns out [Kaplan03 pp. 326–327] that if the domain of $\mathbf{g}(x,y)$ is simply connected (a region without holes) then $\mathbf{g}(x,y)$ is conservative *iff* it is irrotational, i.e. $\text{curl } \mathbf{g} = 0$. In the 2D case $\text{curl } \mathbf{g} = 0$ means:

$$\frac{\partial}{\partial x}g_2(x,y) - \frac{\partial}{\partial y}g_1(x,y) = 0 \quad (\text{B.12})$$

which is precisely the second part of the Cauchy-Riemann condition (b) (see Eq. (B.6)) for the transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$.

Note, however, that even when such a surface $g(x,y)$ does exist it is clearly not unique, since any surface $g(x,y) + c$ has the same gradient field as $g(x,y)$. It follows, therefore, that if $g(x,y)$ exists then it is unique up to an additive constant [Ivanov95 p. 247].

In a similar way, one may also define for the same surface $z = g(x,y)$ a different vector field, that we denote here by $Hg(x,y)$, in which the trajectories coincide with the level lines of the surface, $g(x,y) = \text{const}$. In this vector field the trajectories are perpendicular at each point to the gradient of the surface. This vector field is given by (see [Birkhoff89 p. 135]):

$$Hg(x,y) = \left(\frac{\partial}{\partial y}g(x,y), -\frac{\partial}{\partial x}g(x,y) \right) \quad (\text{B.13})$$

This follows, indeed, from (B.10) if we remember that the vector $(b, -a)$ is perpendicular to the vector (a, b) . Note that we could also take the vector field $-Hg(x,y)$, whose trajectories follow the same level lines but in the opposite direction; which of the two is used is just a matter of convention.

In our present example of the paraboloid $g(x,y) = x^2 + y^2$, the vector field (B.13) is $\mathbf{h}(x,y) = Hg(x,y) = (2y, -2x)$. Its trajectories are given by the following system of linear differential equations:

$$\begin{aligned} \frac{d}{dt}x(t) &= 2y(t) \\ \frac{d}{dt}y(t) &= -2x(t) \end{aligned} \quad (\text{B.14})$$

whose solution curves consist of the family of concentric circles that is given in parametric form by $x(t) = c_1 \cos t + c_2 \sin t$, $y(t) = -c_1 \sin t + c_2 \cos t$ for any constants c_1, c_2 , or in implicit form by $x^2 + y^2 = c^2$ for any constant c [Kreyszig93 p. 170]. These circles are, indeed, the level lines of our paraboloid.

However, just as in the case of gradient fields, it turns out that while for every reasonably well behaved surface $g(x,y)$ there exists a vector field $\mathbf{g}(x,y)$ such that $\mathbf{g}(x,y) = Hg(x,y)$, the converse is not necessarily true: Not every transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ can be

represented as a vector field $Hg(x,y)$ of some surface $g(x,y)$. For example, the transformation $\mathbf{g}(x,y) = (2x, 2y)$ has no surface $g(x,y)$ such that $\frac{\partial}{\partial y}g(x,y) = 2x$ and $-\frac{\partial}{\partial x}g(x,y) = 2y$, since this would imply by integration that $g(x,y) = 2xy + c_1(y)$ and $g(x,y) = -2xy + c_2(x)$, but these two conditions on $g(x,y)$ are contradictory. On the other hand, the transformation $\mathbf{g}(x,y) = (2y, -2x)$ does have a surface $g(x,y)$ such that $\frac{\partial}{\partial y}g(x,y) = 2y$ and $-\frac{\partial}{\partial x}g(x,y) = -2x$, since this implies by integration that $g(x,y) = y^2 + c_1(x)$ and $g(x,y) = x^2 + c_2(y)$, and indeed, by taking $c_1(x) = x^2$ and $c_2(y) = y^2$ we obtain $g(x,y) = x^2 + y^2$.

It can be shown that for a given $\mathbf{g}(x,y)$ there exists a surface $g(x,y)$ such that $\mathbf{g}(x,y) = Hg(x,y)$ iff $\mathbf{g}(x,y)$ is a *solenoidal* vector field, which means (see [Kaplan03 p. 184; Weisstein99 p. 1671–1672]) that $\text{div } \mathbf{g} = 0$, or in the 2D case:

$$\frac{\partial}{\partial x}g_1(x,y) + \frac{\partial}{\partial y}g_2(x,y) = 0 \quad (\text{B.15})$$

Such a surface $g(x,y)$ is often called a *Hamiltonian potential function* of $\mathbf{g}(x,y)$ [Howse95]. Note that condition (B.15) is precisely the first part of the Cauchy-Riemann condition (b) (see Eq. (B.6)) for the transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$.

These results allow us to answer the following interesting questions: Given a transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ with two known families of level lines, can we find vector fields $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ having the same two curve families as trajectories (field lines)? And conversely, given a vector field $\mathbf{v}(x,y)$ with known trajectories, can we find a transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ whose level lines (the level lines of $g_1(x,y)$ or of $g_2(x,y)$) are identical to these trajectories? This would give us an interesting connection between the *level lines* of transformations (that are viewed as a pair of surfaces over the x,y plane) and the *trajectories* of other transformations (that are considered as vector fields).

The answer to the first question is, indeed, affirmative: As we have seen above, the two level line families of the *domain* transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ can be also regarded as the trajectories of the vector fields $Hg_1(x,y)$ and $Hg_2(x,y)$. And similarly, the two level lines of the *direct* transformation $\mathbf{g}(x,y)$ can be also regarded as the trajectories of the vector fields $Hg_1^{-1}(x,y)$ and $Hg_2^{-1}(x,y)$, where $g_1^{-1}(x,y)$ and $g_2^{-1}(x,y)$ are the two components of the inverse transformation $\mathbf{g}^{-1}(x,y)$, namely: $\mathbf{g}^{-1}(x,y) = (g_1^{-1}(x,y), g_2^{-1}(x,y))$. For example, in the case of the domain transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$, whose level lines are given by the two families of hyperbolas $2xy = m$ and $y^2 - x^2 = n$ for any m and n (see Figs. B.1 and B.4(d)), these two hyperbolic curve families are also the trajectories of the vector fields $Hg_1(x,y) = (2x, -2y)$ and $Hg_2(x,y) = (2y, 2x)$ (see Fig. B.6). However, going the other way around is not always possible: Given a vector field $\mathbf{v}(x,y)$ it is not always possible to find a transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ whose level lines (i.e. the level lines of one of the surfaces $g_1(x,y)$ or $g_2(x,y)$) correspond to the trajectories of the vector field $\mathbf{v}(x,y)$; as we have just seen, this is only possible if $\mathbf{v}(x,y)$ is solenoidal.

Any vector field $\mathbf{v}(x,y)$ can be classified as conservative, solenoidal, both conservative and solenoidal, or neither conservative nor solenoidal. For example, the vector field

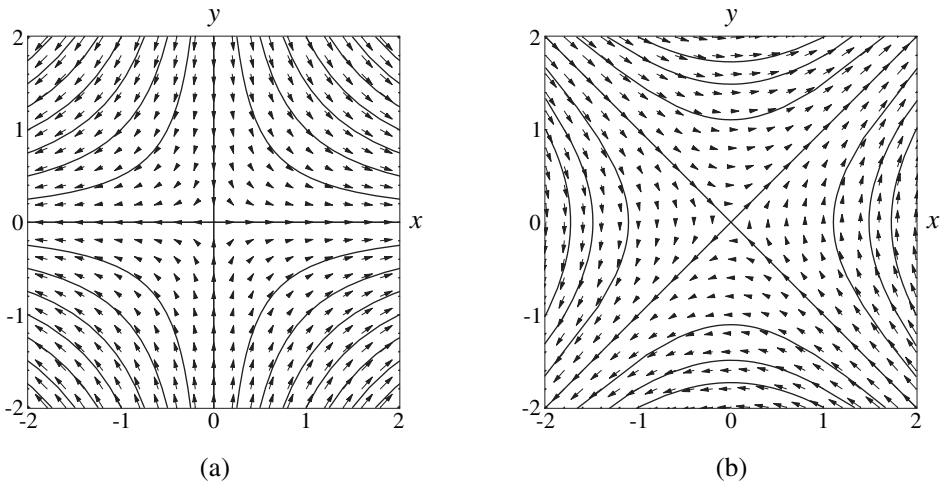


Figure B.6: The level lines of the domain transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y)) = (2xy, x^2 - y^2)$ (see Figs. B.1 or B.4(d)) are identical to the trajectories of the vector fields (a) $Hg_1(x,y) = (2x, -2y)$ and (b) $Hg_2(x,y) = (2y, 2x)$.

$\mathbf{g}(x,y) = (2xy, x^2 - y^2)$ is both conservative and solenoidal, while the vector field $\mathbf{g}(x,y) = (x, y + x)$, which corresponds to a linear shear transformation, is neither conservative nor solenoidal. A vector field that is both conservative and solenoidal is called a *harmonic* vector field [Ivanov95 p. 242]. If $\mathbf{v}(x,y)$ is a harmonic vector field, it is possible to find a transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ where $g_1(x,y)$ is a surface whose *level lines* correspond to the trajectories of the given vector field, and $g_2(x,y)$ is a surface whose *gradient lines* correspond to the trajectories of the given vector field. In this case, the level lines of the surfaces $g_1(x,y)$ and $g_2(x,y)$ of $\mathbf{g}(x,y)$ give two orthogonal sets of curves that correspond respectively to the trajectories and to the *equipotential lines*⁷ of the harmonic vector field $\mathbf{v}(x,y)$ [Kreyszig93 pp. 886–889]. Thus, a harmonic vector field $\mathbf{v}(x,y)$ can be also depicted by the net consisting of its trajectories and its equipotential lines [Ivanov95 p. 248]; but this graphic representation of the vector field $\mathbf{v}(x,y)$ should not be confused with the curvilinear net that represents $\mathbf{v}(x,y)$ as a transformation (compare Figs. B.5(b) and B.4(b); note that the equipotential lines are not shown in Fig. B.5(b), but they are a 90° rotated copy of the trajectory lines).

Clearly, if $\mathbf{v}(x,y) = (v_1(x,y), v_2(x,y))$ is harmonic it satisfies conditions (B.12) and (B.15), i.e. the two Cauchy-Riemann conditions (b) (see Eq. (F.6)), and hence it is a *conformal* transformation (see Sec. B.5). However, the converse is not necessarily true: Although the transformation $\mathbf{u}(x,y) = (v_1(x,y), -v_2(x,y))$ is conformal, since it satisfies the two Cauchy-Riemann conditions (a), it is not a harmonic vector field. For example, consider

⁷ The equipotential lines are to the trajectories (field lines) of a harmonic vector field $\mathbf{g}(x,y)$ what the level lines are to the gradient lines of the surface $g(x,y)$.

$\mathbf{v}(x,y) = (2xy, x^2 - y^2)$ and $\mathbf{u}(x,y) = (2xy, y^2 - x^2)$: Although both are conformal, $\mathbf{v}(x,y)$ is harmonic but $\mathbf{u}(x,y)$ is not (to convince oneself, it is easy to verify that there exists no surface $g(x,y)$ such that $\mathbf{u}(x,y) = \nabla g(x,y)$, and no surface $g(x,y)$ such that $\mathbf{u}(x,y) = Hg(x,y)$). On the other hand, the transformation $\mathbf{w}(x,y) = (-v_2(x,y), v_1(x,y))$ is harmonic, since it satisfies the Cauchy-Riemann conditions (b). This transformation is said to be *harmonically conjugate* to $\mathbf{v}(x,y)$ [Ivanov95 p. 242]; its equipotential lines are the trajectories of $\mathbf{v}(x,y)$, and vice versa [Ivanov95 p. 248; Needham97 p. 509].

Interestingly, these results also suggest that there may exist two different ways for representing a general 2D transformation $\mathbf{g}(x,y)$ as a pair of surfaces: Either, as explained in Sec. B.2, as the pair of surfaces $g_1(x,y)$ and $g_2(x,y)$ which are the two Cartesian components of $\mathbf{g}(x,y)$:

$$\mathbf{g}(x,y) = (g_1(x,y), 0) + (0, g_2(x,y)) = (g_1(x,y), g_2(x,y)) \quad (\text{B.16})$$

or as a pair of surfaces $g(x,y)$ and $h(x,y)$ such that:

$$\mathbf{g}(x,y) = \nabla g(x,y) + Hh(x,y) \quad (\text{B.17})$$

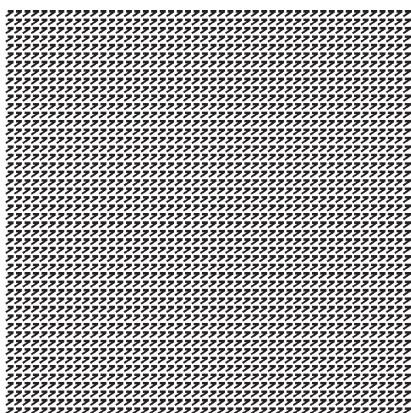
which means, in componentwise notation:

$$= \left(\frac{\partial}{\partial x}g(x,y) + \frac{\partial}{\partial y}h(x,y), \frac{\partial}{\partial y}g(x,y) - \frac{\partial}{\partial x}h(x,y) \right)$$

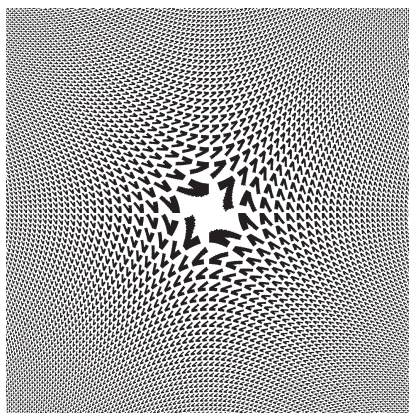
This last decomposition of $\mathbf{g}(x,y)$ is known as a *gradient-Hamiltonian decomposition*, and the functions $g(x,y)$ and $h(x,y)$ are called, respectively, a *gradient potential function* of $\mathbf{g}(x,y)$ and a *Hamiltonian potential function* of $\mathbf{g}(x,y)$. However, this representation of $\mathbf{g}(x,y)$ is not necessarily unique. More details on the gradient-Hamiltonian decomposition and on its limitations can be found in [Howse95 pp. 60, 67–68]. Note that this decomposition sheds a new light on the fact already mentioned above that if $\mathbf{g}(x,y)$ is not a conservative vector field, there is no surface $g(x,y)$ such that $\mathbf{g}(x,y) = \nabla g(x,y)$: As we can now understand, this simply means that the gradient-Hamiltonian decomposition of such transformations $\mathbf{g}(x,y)$ must have a non-vanishing Hamiltonian component. The Hamiltonian component vanishes *iff* $\mathbf{g}(x,y)$ is conservative and hence can be represented as $\mathbf{g}(x,y) = \nabla g(x,y)$, and the gradient component vanishes *iff* $\mathbf{g}(x,y)$ is solenoidal and hence can be represented as $\mathbf{g}(x,y) = Hh(x,y)$. If $\mathbf{g}(x,y)$ is harmonic, i.e. both conservative and solenoidal, its gradient-Hamiltonian decomposition cannot be unique, since in this case we have both $\mathbf{g}(x,y) = \nabla g(x,y)$ with $h(x,y) \equiv 0$, and $\mathbf{g}(x,y) = Hh(x,y)$ with $g(x,y) \equiv 0$.

B.8 Remark on the local reflection of a 2D transformation

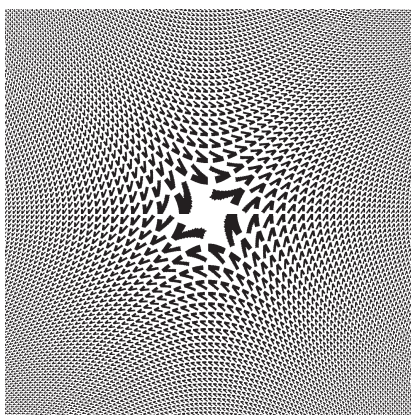
A transformation $\mathbf{g}(x,y)$ is said to be locally reflecting or locally non-reflecting around the point (x,y) according to whether the Jacobian at that point is positive or negative (see Appendix C). For example, the transformation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$ is non-reflecting throughout the plane (since $J(x,y) = 4(x^2 + y^2) > 0$), whereas the transformation $\mathbf{g}(x,y) = (2xy, x^2 - y^2)$ is reflecting throughout the plane (since $J(x,y) = -4(x^2 + y^2) < 0$). An



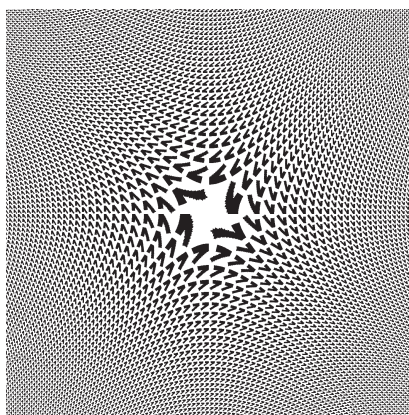
(a)



(b)



(c)



(d)

Figure B.7: Same as Fig. B.3, but this time using in the original image (a) a dot screen with the asymmetric element “1” rather than the symmetric element “•”. This allows to clearly show the local orientation of the transformed plane at each point, and thus to distinguish between transformations such as: (b) $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$, (c) $\mathbf{g}(x,y) = (2xy, x^2 - y^2)$, and (d) $\mathbf{g}(x,y) = -(2xy, y^2 - x^2)$. Note that these transformations only differ in their local orientation at each point (x,y) , but not in their global geometry.

example of a transformation that is reflecting in some parts of the plane and non-reflecting in other parts of the plane is given by Eq. (7.32), which corresponds to the moiré effect shown in Fig. 7.12(b).

The information about the local orientation of a transformation $\mathbf{g}(x,y)$ at each point (x,y) is obviously lost when we represent $\mathbf{g}(x,y)$ graphically as a coordinate change, like in Fig. B.4. But when we illustrate the effect of $\mathbf{g}(x,y)$ as we do in Figs. B.2 and B.3, there exists a simple “trick” that allows us to clearly show the local orientation of $\mathbf{g}(x,y)$ at any point of the plane. All that we need to do is to apply $\mathbf{g}(x,y)$ to a structure made of *asymmetric* rather than symmetric elements. When we do so, the local orientation of the transformation at each point (x,y) is indicated by the local orientation of the corresponding asymmetric element in the transformed plane. This is clearly illustrated in Fig. B.7, where three different transformations are applied to the same periodic dot screen (a) that consists of asymmetric “1”-shaped dots. The transformations being applied are: (b) $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$; (c) $\mathbf{g}(x,y) = (2xy, x^2 - y^2)$; and (d) $\mathbf{g}(x,y) = -(2xy, y^2 - x^2)$. Note that in (c) all the “1”-shaped elements are mirror-imaged, while in (b) and (d) they are just rotated, but not mirror-imaged. This illustrates the fact that transformation (c) is reflecting throughout the plane, while the transformations (b) and (d) are non-reflecting.

More details on the connection between the Jacobian of a transformation $\mathbf{g}(x,y)$ and the properties of the transformation can be found in Appendix C.

Appendix C

The Jacobian of a 2D transformation and its significance

C.1 Introduction

Let $\mathbf{g}(x,y)$ be the 2D transformation whose two components are:

$$\begin{aligned} u &= g_1(x,y) \\ v &= g_2(x,y) \end{aligned} \tag{C.1}$$

The *Jacobian matrix* of this transformation is the matrix:

$$\begin{pmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{pmatrix} \tag{C.2}$$

Note that the rows of this matrix correspond to the gradients of $g_1(x,y)$ and $g_2(x,y)$ (see Eq. (B.10) in Appendix B), while the columns of this matrix correspond to the directional derivatives of $\mathbf{g}(x,y)$ in the x and y directions, respectively. The *Jacobian determinant* (or simply the *Jacobian*) of the transformation $\mathbf{g}(x,y)$ is the scalar function $J: \mathbb{R}^2 \rightarrow \mathbb{R}$ that is defined as the determinant of this matrix:¹

$$J(x,y) = \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} = \frac{\partial g_1}{\partial x} \frac{\partial g_2}{\partial y} - \frac{\partial g_2}{\partial x} \frac{\partial g_1}{\partial y} \tag{C.3}$$

We have already seen in Appendix B that the Jacobian is tightly related to the mathematical properties of the transformation $\mathbf{g}(x,y)$. We will now explain the geometric interpretation of the Jacobian, and see in more detail its special role in connection with the transformation $\mathbf{g}(x,y)$. Other properties of $\mathbf{g}(x,y)$ that can be deduced from its Jacobian matrix (C.2) are discussed later in Secs. C.4–C.5.

C.2 Geometric interpretation of the Jacobian

Consider a 2D transformation $\mathbf{g}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Clearly, this transformation maps any square element of the original x,y plane into its distorted image in the destination u,v plane. The area of the new distorted element can be smaller, equal or larger than the area of the original, undistorted element, depending on the local properties of the transformation \mathbf{g} at

¹ Confusingly, some references use the term “Jacobian” for the Jacobian matrix, while other references use it, as we do, for the Jacobian determinant.

that point. In order to investigate how \mathbf{g} influences the area at each point of the plane, we consider an infinitesimal square area-element $dxdy$ within the original x,y plane, and its distorted image $dudv$ in the transformed u,v plane. It turns out that at any point (x,y) we have (see [Weisstein99 p. 950] or [Colley98 p. 347]):

$$dudv = J(x,y) dxdy \quad (\text{C.4})$$

More generally, if $A[S]$ and $A[\mathbf{g}(S)]$ denote, respectively, the area of a closed region S of the x,y plane and the area of its image $\mathbf{g}(S)$ in the u,v plane, then we have at the limit when $A[S]$ approaches zero [Spiegel63 p. 108]:

$$\lim \frac{A[\mathbf{g}(S)]}{A[S]} = J(x,y) \quad (\text{C.5})$$

This means that the Jacobian is, in fact, an infinitesimal scale factor that indicates the local area scaling caused by the transformation $\mathbf{g}(x,y)$ at each point (x,y) . Negative scaling values indicate that local scaling at (x,y) is also accompanied by local reflection (mirror imaging). If the Jacobian equals zero at a certain point (x,y) , it means that the transformation \mathbf{g} maps elements with non-zero area around that point to a zero-area image; the point (x,y) is called, then, a singular point of the transformation \mathbf{g} .

Note that the Jacobian matrix of a transformation $(u,v) = \mathbf{g}(x,y)$ and the Jacobian matrix of the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ are inverse matrices of each other [Kaplan03 p. 120; Courant89 p. 252]. This implies that the Jacobian $J_{\mathbf{g}}(x,y)$ of \mathbf{g} and the Jacobian $J_{\mathbf{g}^{-1}}(x,y)$ of \mathbf{g}^{-1} are reciprocals of each other: $J_{\mathbf{g}^{-1}}(x,y) = 1/J_{\mathbf{g}}(x,y)$. Hence, if one Jacobian is non-singular (different from zero) at the point (x,y) , so is the other. Furthermore, a transformation $\mathbf{g}(x,y)$ is invertible over \mathbb{R}^2 or a subregion thereof (meaning that there exists a transformation $\mathbf{g}^{-1}(u,v)$ such that $\mathbf{g}^{-1}(\mathbf{g}(x,y)) = (x,y)$ within that region) *iff* the Jacobian of \mathbf{g} is not identically zero within that region. Other important theorems on transformations and their Jacobians can be found, for example, in [Spiegel63 p. 108].

Remark C.1: When \mathbf{g} is a linear transformation the two functions $g_1(x,y)$ and $g_2(x,y)$ of Eq. (C.1) are linear:

$$\begin{aligned} u &= a_1x + b_1y \\ v &= a_2x + b_2y \end{aligned} \quad (\text{C.6})$$

and their slopes $\frac{\partial g_1}{\partial x}, \frac{\partial g_1}{\partial y}, \frac{\partial g_2}{\partial x}, \frac{\partial g_2}{\partial y}$ reduce into the constant values a_1, b_1, a_2, b_2 . This means that when \mathbf{g} is linear, its Jacobian matrix is simply reduced to the matrix of the linear transformation, and the Jacobian is reduced to the determinant of this matrix:

$$J(x,y) = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \quad (\text{C.7})$$

so that the Jacobian $J(x,y)$ becomes a constant number, $J(x,y) = a_1b_2 - a_2b_1$.

Viewed the other way around, non-linear transformations can be seen as a generalization of linear transformations into the case where the slopes of $g_1(x,y)$ and $g_2(x,y)$ are no longer constant but rather vary with x and y . And indeed, if the transformation is not linear, then the Jacobian matrix will vary from point to point. Nevertheless, it may still be regarded as a local linear approximation to the true mapping (plus a relocation of the origin), and the Jacobian $J(x,y)$ can still be interpreted as the local area scale factor, which may now change from point to point. In other words, the Jacobian matrix of a non-linear transformation $\mathbf{g}(x,y)$ can be viewed at any given point (x,y) as the matrix of a linear transformation $\mathbf{g}'(x,y)$ that approximates the non-linear transformation $\mathbf{g}(x,y)$ at the point (x,y) , and whose constant coefficients a_1, b_1, a_2, b_2 , are, respectively, the values of $\frac{\partial g_1}{\partial x}, \frac{\partial g_1}{\partial y}, \frac{\partial g_2}{\partial x}, \frac{\partial g_2}{\partial y}$ at that particular point.

It is interesting to note in this context that in linear algebra determinants play the role of area (or volume) functions [Lay03 pp. 204–209]. For example, the determinant (C.7) of the matrix of the linear transformation (C.6) gives the area of the parallelogram that is determined by the columns of the matrix, i.e. by the vectors (a_1, a_2) and (b_1, b_2) [Lay03 p. 205]. But these vectors are precisely the images under our linear transformation of the standard basis vectors $(1,0)$ and $(0,1)$ [Lang87 pp. 394–395]. Thus, if we denote by S the parallelogram determined by the basis vectors $(1,0)$ and $(0,1)$, and by $A[R]$ the area of the region R , then we have for any linear transformation \mathbf{g} :

$$A[\mathbf{g}(S)] = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} A[S] \quad (\text{C.8})$$

Furthermore, as long as \mathbf{g} is a linear transformation this property remains valid for any parallelogram S in the plane [Lang87 p. 457], and even for any arbitrary closed region S of the plane [Lay03 pp. 207–209]. Eq. (C.8) is, indeed, the linear equivalent of Eqs. (C.4) and (C.5).

In non-linear transformations the area scaling effect of \mathbf{g} may be different at each point (x,y) of the plane, and therefore Eq. (C.8) is no longer globally valid for all arbitrary closed regions S ; but it remains valid locally, for any *infinitesimal* region, as expressed, indeed, by Eqs. (C.4) and (C.5). ■

C.3 Properties of the transformation $\mathbf{g}(x,y)$ that can be deduced from its Jacobian

As we have seen in the previous section, a 2D transformation $\mathbf{g}(x,y)$ is closely related to its Jacobian. Therefore, it is not surprising that the Jacobian can provide precious information on the nature of the transformation in question. In the present section we provide a summary of various properties of the transformation $\mathbf{g}(x,y)$ that can be deduced directly from its Jacobian. We start with a few results concerning the global nature of the Jacobian and its effect on the global properties of the transformation $\mathbf{g}(x,y)$, and then we proceed to some of their local counterparts. Note that $\mathbf{g}(x,y)$ is considered here as a direct transformation; but if $\mathbf{g}(x,y)$ is used as a domain transformation (for example, if it is

applied to an image $r(x,y)$ to give the distorted image $r(\mathbf{g}(x,y))$, as in Figs. C.1, C.2 or B.6), then we are concerned in fact with the inverse transformation \mathbf{g}^{-1} , whose properties can be deduced from the nature of the reciprocal Jacobian $1/J(x,y)$.

- (1) If $|J(x,y)| < 1$ everywhere, then $\mathbf{g}(x,y)$ is *area-contracting*.
- (2) If $|J(x,y)| > 1$ everywhere, then $\mathbf{g}(x,y)$ is *area-expanding*.
- (3) If $J(x,y) > 0$ everywhere, then $\mathbf{g}(x,y)$ is *non-reflecting*. Such a mapping can locally represent rotations, scalings and shearing deformations and it can globally represent “rubber-sheet” distortions, but it will nowhere cause reflection. A string of text subjected to such a mapping would remain legible (although possibly highly distorted), and it would not be converted into a mirror image of itself. For example, when such a transformation \mathbf{g} is applied to a periodic dot screen consisting of asymmetric “1”-shaped elements (Fig. C.1(a)), it maps each of these elements into a distorted and possibly rotated “1”-shaped element, but none of the resulting distorted “1”s is mirror imaged (see Figs. C.1(b),(c),(f),(g),(h) and Figs. B.6(a),(d)). Note that such transformations are also called in literature *orientation-preserving* [Courant89 p. 260], but this term may be somewhat misleading since the image of our “1”-shaped elements under such transformations may still be rotated. A better term would be *sense-preserving*.
- (4) If $J(x,y) < 0$ everywhere, then $\mathbf{g}(x,y)$ is *reflecting*. Such a mapping will locally include reflection (possibly also combined with rotations, scalings and shearing deformations) and it can globally represent a “rubber-sheet” distortion combined with reflection. A string of text subjected to such a mapping would be converted into a mirror image of itself (in addition to any other distortions). For example, when such a transformation \mathbf{g} is applied to a periodic dot screen consisting of asymmetric “1”-shaped elements (Fig. C.1(a)), it maps each of these elements into a mirror imaged and possibly otherwise distorted and rotated “1”-shaped element (see Figs C.1(d),(e) and Fig. B.6(c)). Note that such transformations are also called in literature *orientation-reversing* [Courant89 p. 260], but this term may be confusing since pure rotations that are not accompanied by reflection (including a rotation by 180°) are not orientation-reversing transformations. A better term would be *sense-reversing*.
- (5) If $J(x,y) = \text{const.}$ everywhere, then $\mathbf{g}(x,y)$ has a constant scaling factor throughout. This obviously occurs if $\mathbf{g}(x,y)$ is a linear or affine transformation, but it may also occur in non-linear transformations (such as (C.9) below; see Fig. C.1(b)).
- (6) If $J(x,y) = 1$ everywhere, then $\mathbf{g}(x,y)$ is *area-preserving*. This occurs, for example, if $\mathbf{g}(x,y)$ is a rotation or a shift transformation, but it may occur also in non-linear transformations (such as (C.9) below; see Fig. C.1(b)).
- (7) If $J(x,y) = -1$ everywhere, then $\mathbf{g}(x,y)$ is *area-preserving* and *reflecting*. This occurs, for example, if $\mathbf{g}(x,y)$ is a reflection or a rotoinversion (namely, rotation combined with reflection [Cantwell02 p. 9]), but it may occur also in non-linear transformations.

- (8) If $J(x,y) = 0$ everywhere, then $\mathbf{g}(x,y)$ is *degenerate*, meaning that it maps \mathbb{R}^2 or any 2D subregion thereof onto a 1D curve, a single point, or an empty set (see Sec. B.3 in Appendix B). This occurs *iff* the two components of $\mathbf{g}(x,y)$, i.e. $g_1(x,y)$ and $g_2(x,y)$, are dependent (see Definition B.1 in Appendix B). For example, the transformation $(u,v) = \mathbf{g}(x,y) = (x, x^2)$ maps the entire x,y plane onto the 1D parabola $v = u^2$.
- (9) If $J(x,y) \neq 0$ and $J(x,y) \neq \pm\infty$ everywhere, then $\mathbf{g}(x,y)$ has no singularities, and it is one-to-one and invertible.

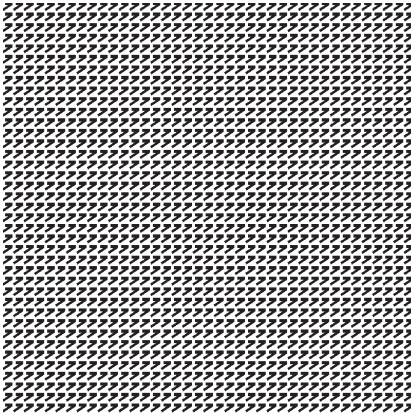
The remaining points of the list concern the *local* properties of the Jacobian around a given point (x_0, y_0) and their local influence on the transformation $\mathbf{g}(x,y)$:

- (10) If $|J(x_0, y_0)| < 1$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ is *area-contracting* near that point.
- (11) If $|J(x_0, y_0)| > 1$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ is *area-expanding* near that point.
- (12) If $J(x_0, y_0) > 0$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ is *non-reflecting* (orientation-preserving) near that point [Courant89 p. 260]. For example, an asymmetric “1”-shaped element near that point will be mapped by \mathbf{g} into a distorted but not mirror-imaged “1”.
- (13) If $J(x_0, y_0) < 0$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ is *reflecting* (orientation-reversing) near that point [Courant89 p. 260]. For example, an asymmetric “1”-shaped element near that point will be mapped by \mathbf{g} into a distorted, mirror-imaged “1”.
- (14) If $J(x_0, y_0) = 0$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ has a local area scaling of zero at that point. This means that \mathbf{g} maps elements with non-zero area around the point (x_0, y_0) to a zero area (or an almost-zero area) image. This occurs, for example, at the point (0,0) in Fig. B.2(b).
- (15) If $J(x_0, y_0) = \infty$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ has an infinitely big local area scaling at that point. This occurs, for example, at the point (0,0) in the transformation shown in Fig. B.3(b), whose expression as a *direct* transformation is given by $\mathbf{g}^{-1}(x,y) = (\sqrt{(\sqrt{u^2 + v^2} - v)}/2, \sqrt{(\sqrt{u^2 + v^2} + v)}/2)$; see Example D.5 in Appendix D.
- (16) If $J(x_0, y_0) = -\infty$ at the point (x_0, y_0) , then $\mathbf{g}(x,y)$ has an infinitely big local area scaling at that point, and, in addition, $\mathbf{g}(x,y)$ is also reflecting at that point.

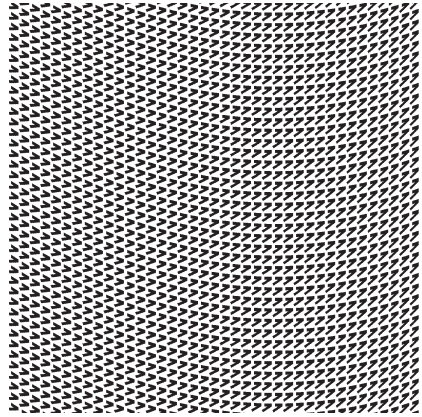
In all of the cases (14)–(16) the transformation $\mathbf{g}(x,y)$ is said to be singular at the point (x_0, y_0) , or, equivalently, to have a singular point at (x_0, y_0) .

Figs. C.1 and C.2 illustrate some of these cases by applying various domain transformations to a periodic dot screen $r(x,y)$ consisting of asymmetric “1”-shaped elements (Fig. C.1(a)). The transformations being used are all variants of the parabolic transformation $\mathbf{g}(x,y)$ that is given by:

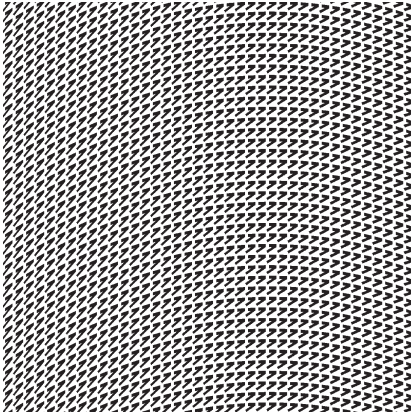
$$\begin{aligned} u &= x \\ v &= y - a(x - c)^2 \end{aligned} \tag{C.9}$$



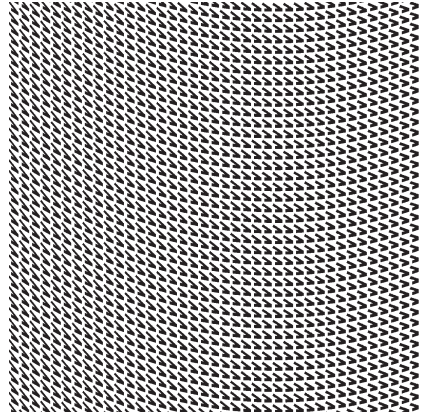
(a)



(b)



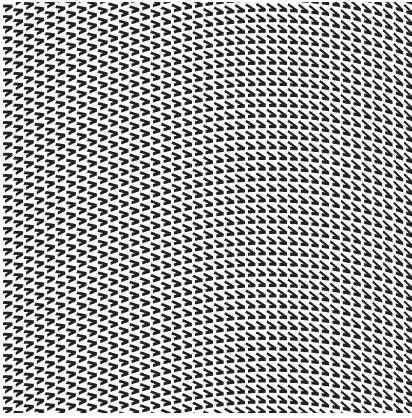
(c)



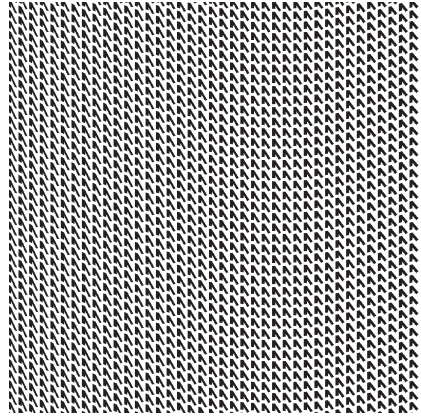
(d)

Figure C.1: A dot screen $r(x,y)$ composed of asymmetric “1”-shaped dots (a), and its deformations $r(\mathbf{g}(x,y))$ under different variants of the domain transformation (C.9). Note the influence of each of the transformations on the orientation of the parabolas, on the local orientation of the “1”-shaped elements (cells), and on the global orientation of the distorted image. The transformations being used (and their effects) are:

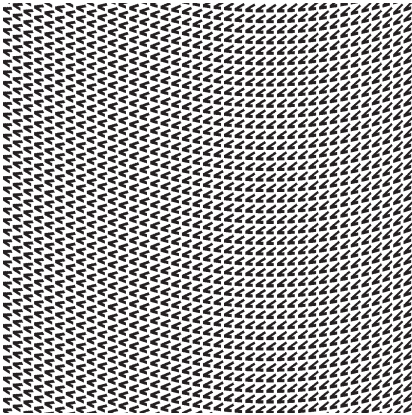
- (b) $u = x, \quad v = y - a(x - c)^2$ (top-opened parabolas, upright “1”s);
- (c) $u = x, \quad v = y + a(x - c)^2$ (bottom-opened parabolas, upright “1”s);
- (d) $u = x, \quad v = -y + a(x - c)^2$ (top-opened parabolas, reflected “1”s);
- (e) $u = x, \quad v = -y - a(x - c)^2$ (global vertical reflection of (b));
- (f) $u = y - a(x - c)^2, \quad v = -x$ (top-opened parabolas, rotated “1”s);
- (g) $u = -x, \quad v = -y + a(x - c)^2$ (top-opened parabolas, rotated “1”s);
- (h) $u = -y + a(x - c)^2, \quad v = x$ (top-opened parabolas, rotated “1”s).



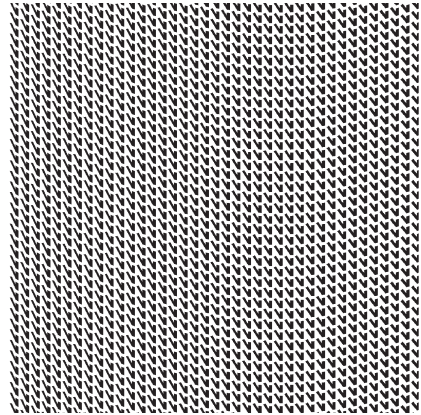
(e)



(f)



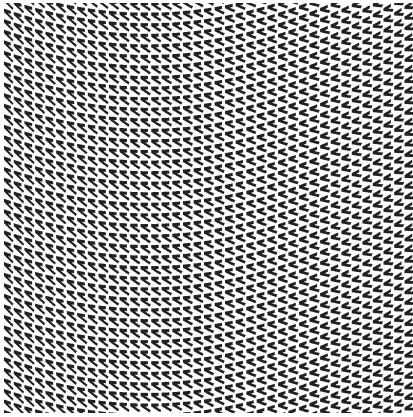
(g)



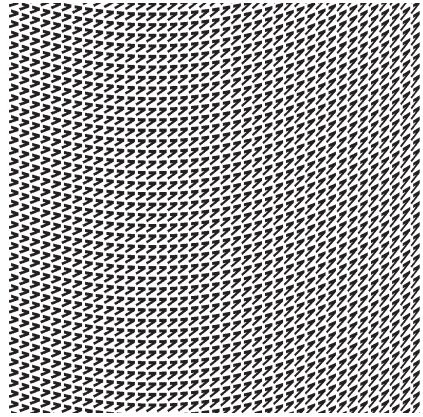
(h)

Figure C.1: (*continued.*) Note that for practical reasons the upright orientation of the “1”s in the original dot screen (a) is not vertical, but rather rotated by -45° (in order to avoid vertical collisions between the “1” elements in successive rows). This implies that each of the transformed “1”s, too, is in fact a distorted version of the rotated “1”.

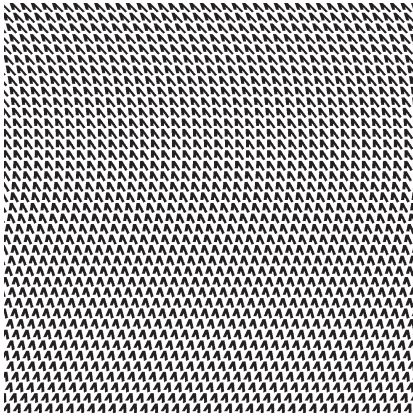
When (C.9) is applied as a domain transformation, it bends the horizontal coordinate lines of the x,y pane into equispaced top-opened parabolas that are shifted by a constant c to the right (see Fig. C.1(b)). We have chosen this transformation because of its global asymmetry with respect to the origin and with respect to the main axes. This asymmetry allows us to investigate different variants of this transformation, and to clearly visualize



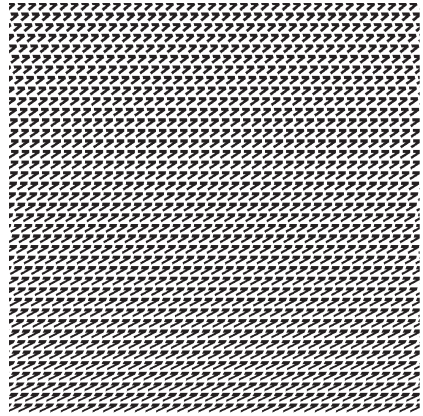
(a)



(b)



(c)



(d)

Figure C.2: Same as in Fig. C.1, with other variants of the domain transformation (C.9):

- (a) $u = -x, \quad v = y - a(x + c)^2$ (global horizontal reflection of Fig. C.1(b));
- (b) $u = x, \quad v = y - a(x + c)^2$ (horizontally reflected parabolas, upright “1”s);
- (c) $u = y, \quad v = -x - a(y - c)^2$ (global 90° rotation of Fig. C.1(b));
- (d) $u = x + a(y - c)^2, \quad v = y$ (90° rotated parabolas, upright “1”s);

their effect on the resulting transformed plane. In particular, it allows us to easily distinguish between: (1) Global reflections and rotations of the entire plane (see Figs. C.1(b) and C.1(e)); (2) reflections or rotations of the parabolic geometry alone, with no influence on the local orientation of the “1”-shaped elements (see Figs. C.1(b) and

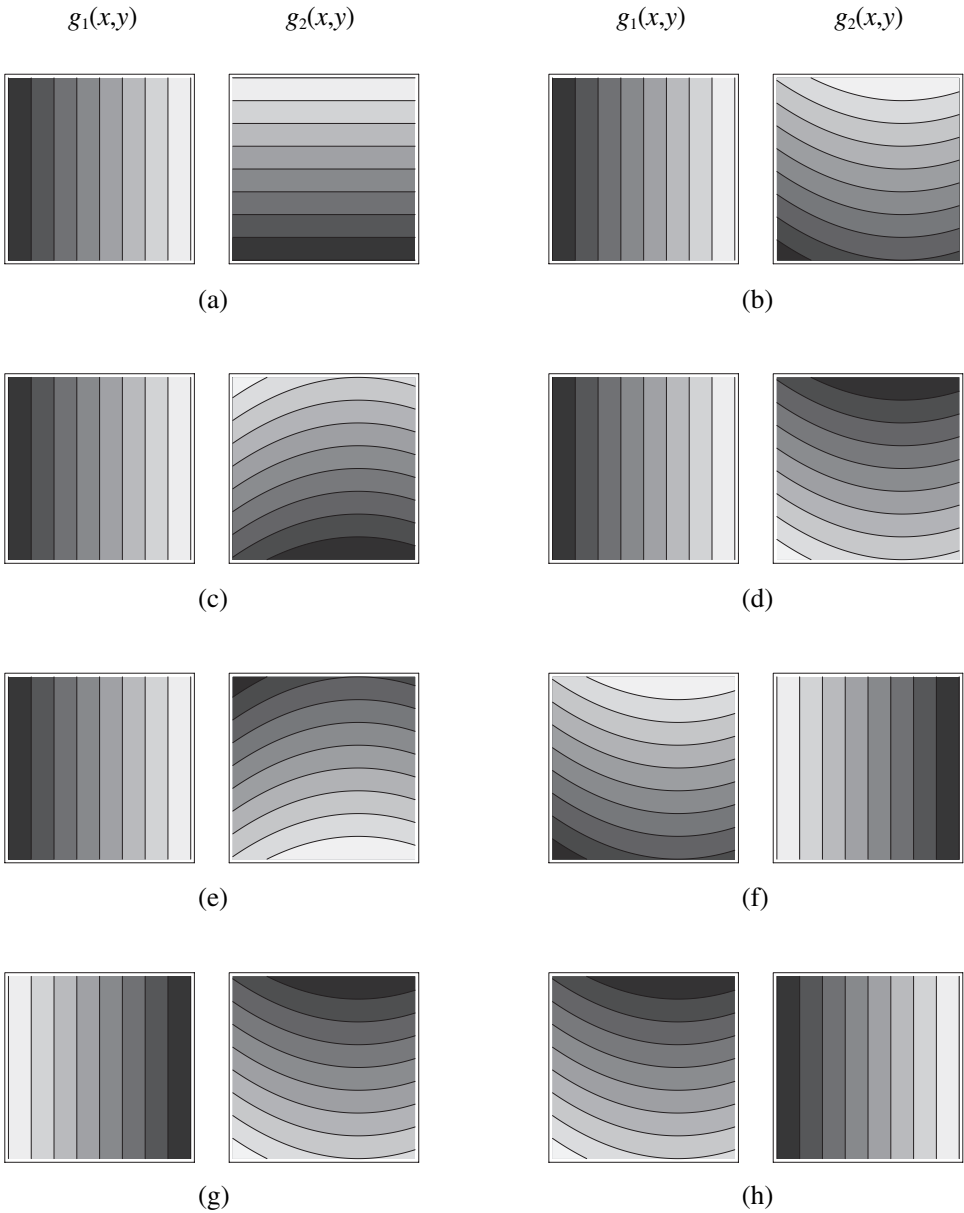


Figure C.3: The two components (surfaces) $g_1(x,y)$ and $g_2(x,y)$ of each of the transformations shown in Figs. C.1(a)–(h). The curves plotted on each of the surfaces are level lines, and the gray levels show the surface altitude: brighter shades indicate higher values and darker shades indicate lower values.

C.1(c)); and (3) local reflections or rotations of the “1”-shaped elements alone, with no global reflections or rotations (see Figs. C.1(b) and C.1(d),(f)–(h)).

The global and local orientation properties of the different variants of (C.9) are best explained by Fig. C.3, which shows for each of the cases of Fig. C.1 the two components (surfaces) $g_1(x,y)$ and $g_2(x,y)$ of the transformation being used (see Sec. B.2 in Appendix B). The cells shown in Fig. C.1 (each of which contains a “1”-shaped element) are, in fact, the curved quadrilaterals that are formed by the curvilinear coordinate lines, i.e. by the level curves of the surfaces $g_1(x,y)$ and $g_2(x,y)$. The cell orientations in the transformed plane simply reflect the orientations of these two families of level curves, where the orientation of each curve family corresponds to the direction of increasing curve altitudes. The ascending order of the level curves in each family is clearly indicated in Fig. C.3 by the gray levels that show the corresponding surface altitude, going from dark (lower levels) to bright (higher levels). Note that the difference between cases (b) and (c) is only in the *orientation* of the parabolic level curves, but not in their order; while the difference between cases (b) and (d) is only in the *order* of the parabolic level curves, but not in their orientation. The difference between cases (b) and (e) is both in the orientation and in the order of the parabolic level curves. Note also the difference between cases (d) and (h), which simply consists of interchanging $g_1(x,y)$ and $g_2(x,y)$.

Thus, as we can see by comparing Figs. C.1(a)–(h) with their corresponding families of level curves in Figs. C.3(a)–(h), the global geometry of the cells in the transformed plane is due to the *shape* of the corresponding level curves, while the local orientation of the cells is due to the *ordering* of the level curves. For example, when the level curves are identical in their shape but their order is inversed, the result is a local sense inversion (compare Figs. C.1(b) and C.1(d)). As another example, interchanging $g_1(x,y)$ and $g_2(x,y)$ does not modify the shape of the level curves, and it only causes a reflection of the cells (compare Figs. C.1(d) and C.1(h)).

Table C.1 gives the full list of the different variants of transformation (C.9) involving top-opened and bottom-opened parabolas. A similar table can be also constructed for the cases involving a global horizontal reflection (whose top-opened parabolas are shifted from the center to the left rather than to the right) and for the cases involving global rotations by 90° or 270° (which give left-opened and right-opened parabolas, respectively). Some of these cases are shown in Fig. C.2.

In the general case of a transformation $u = g_1(x,y)$, $v = g_2(x,y)$ the total number of different possible symmetric variants is 512: There exist 16 possible sign combinations for x and y (such as $u = g_1(-x,y)$, $v = g_2(x,-y)$, etc.); for each of these we have 4 possible sign variations of g_1 and g_2 themselves (such as $u = -g_1(x,y)$, $v = g_2(x,y)$); then we have 4 possible permutations of x and y within g_1 and g_2 (for example, $u = g_1(y,x)$, $v = g_2(x,y)$); and for each of the 256 combinations we have so far there still exist two possible permutations of g_1 and g_2 themselves (such as $u = g_2(x,y)$, $v = g_1(x,y)$). The total number of variants may significantly reduce in simple cases such as $g_1(x,y) = x$, but on the other hand it may further increase if we also consider the signs of the individual terms within $g_1(x,y)$

and $g_2(x,y)$, as we did in the case of $g_2(x,y) = y - a(x - c)^2$. Note, however, that if the transformation $\mathbf{g}(x,y)$ is symmetric, as was the case in Fig. B.7, then many of its different variants may give the same result. In Fig. C.1 we have intentionally chosen a highly asymmetric transformation $\mathbf{g}(x,y)$, which allows us to clearly distinguish in the figure between its different variants.

C.4 The local orientation properties of a transformation $\mathbf{g}(x,y)$

As we have seen in Sec. C.3, the Jacobian of a transformation $\mathbf{g}(x,y)$ contains in a nutshell all the information about the local *magnification* and *reflection* properties of the transformation. It does not provide, however, any information related to the local *orientation* (or rotation) properties of the transformation; for example, it does not account for the difference between the transformations that generate Figs. C.1(b) and C.1(f)–(h). What other mathematical construct or criterion can be used to provide this missing information?

In order to answer this question, suppose that each of the cells generated by the coordinate grid of the original, untransformed image (the “1”-shaped elements in Fig. C.1(a)) is drawn by a laser beam or by a plotter, that scans the entire cell line-by-line. We assume that the beam scans the cell in horizontal lines that follow the positive x direction, and that successive horizontal lines are ordered from the bottom upward, following the positive y direction. When the transformation $\mathbf{g}(x,y)$ is applied, each of the original square elements is distorted into a curvilinear quadrilateral. Consider the distorted element located at the point (x,y) . When the laser beam draws this distorted element, each of the originally straight scanlines is distorted into a curvilinear line, and the vertical step of the beam, as it advances between the successive curvilinear scanlines, is also distorted (see Fig. C.4).

When we proceed to the limit, each of the curvilinear quadrilaterals becomes infinitesimally small, and the direction of the scanlines and the direction of the steps between successive scanlines reduce into the local tangent slopes of the curvilinear coordinates at the point (x,y) . It turns out that these directions are given by the two following vectors:

$$\begin{aligned} \text{Scanline direction:} \quad \mathbf{v}_1(x,y) &= \frac{1}{J(x,y)} \left(\frac{\partial}{\partial y} g_2(x,y), -\frac{\partial}{\partial x} g_2(x,y) \right) \\ \text{Interline direction:} \quad \mathbf{v}_2(x,y) &= \frac{1}{J(x,y)} \left(-\frac{\partial}{\partial y} g_1(x,y), \frac{\partial}{\partial x} g_1(x,y) \right) \end{aligned} \tag{C.10}$$

The orientations of these two vectors give us the local scanline direction and the local interline direction at the point (x,y) , and their lengths indicate the local stretching factor of the distorted cell in these two directions. Note that the scanline direction $\mathbf{v}_1(x,y)$ and the interline direction $\mathbf{v}_2(x,y)$ are perpendicular to the gradients $\nabla g_1(x,y)$ and $\nabla g_2(x,y)$ (see Fig. C.4(b)); but they are not necessarily perpendicular to each other.

















	Transformation	$J(x,y)$	$\mathbf{v}_1(x,y)$ $\mathbf{v}_2(x,y)$	Parabola orientation	Element orientation	Fig.	Remarks
1	$u = x$ $v = y - a(x - c)^2$	1	$(1, 2a(x-c))$ $(0, 1)$	\uparrow		C.1(b)	(1)
2	$u = y - a(x - c)^2$ $v = -x$	1	$(0, 1)$ $(-1, -2a(x-c))$	\uparrow		C.1(f)	(2)
3	$u = -x$ $v = -y + a(x - c)^2$	1	$(-1, -2a(x-c))$ $(0, -1)$	\uparrow		C.1(g)	(2)
4	$u = -y + a(x - c)^2$ $v = x$	1	$(0, -1)$ $(1, 2a(x-c))$	\uparrow		C.1(h)	(2)
5	$u = y - a(x - c)^2$ $v = x$	-1	$(0, 1)$ $(1, 2a(x-c))$	\uparrow	 R		
6	$u = -x$ $v = y - a(x - c)^2$	-1	$(-1, -2a(x-c))$ $(0, 1)$	\uparrow	 R		
7	$u = -y + a(x - c)^2$ $v = -x$	-1	$(0, -1)$ $(-1, -2a(x-c))$	\uparrow	 R		
8	$u = x$ $v = -y + a(x - c)^2$	-1	$(1, 2a(x-c))$ $(0, -1)$	\uparrow	 R	C.1(d)	(3)
9	$u = x$ $v = y + a(x - c)^2$	1	$(1, -2a(x-c))$ $(0, 1)$	\downarrow		C.1(c)	(4)
10	$u = y + a(x - c)^2$ $v = -x$	1	$(0, 1)$ $(-1, 2a(x-c))$	\downarrow			
11	$u = -x$ $v = -y - a(x - c)^2$	1	$(-1, 2a(x-c))$ $(0, -1)$	\downarrow			
12	$u = -y - a(x - c)^2$ $v = x$	1	$(0, -1)$ $(1, -2a(x-c))$	\downarrow			
13	$u = y + a(x - c)^2$ $v = x$	-1	$(0, 1)$ $(1, -2a(x-c))$	\downarrow	 R		
14	$u = -x$ $v = y + a(x - c)^2$	-1	$(-1, 2a(x-c))$ $(0, 1)$	\downarrow	 R		
15	$u = -y - a(x - c)^2$ $v = -x$	-1	$(0, -1)$ $(-1, 2a(x-c))$	\downarrow	 R		
16	$u = x$ $v = -y - a(x - c)^2$	-1	$(1, -2a(x-c))$ $(0, -1)$	\downarrow	 R	C.1(e)	(5)

Table C.1: (continued on the opposite page)

Legend:

Parabolas orientation: \uparrow = top-opened parabolas; \downarrow = bottom-opened parabolas. Cells orientation: \nearrow = upright orientation; \nwarrow = rotated by 90° ; \swarrow = rotated by 180° ; \searrow = rotated by 270° ; \nearrow = reflected (i.e. mirror-imaged); \nwarrow = reflected and rotated by 90° ; \swarrow = reflected and rotated by 180° ; \searrow = reflected and rotated by 270° . All cases involving reflection are indicated by R.

Remarks:

- (1) All the “1”-shaped cells preserve their original orientation, and the parabolas have an upright orientation. This case serves us as a reference when comparing between cases.
- (2) The “1”-shaped cells are rotated, but the parabolas preserve their original upright orientation.
- (3) The “1”-shaped cells are reflected in the vertical sense, but the parabolas preserve their upright orientation.
- (4) The parabolas are reflected vertically, but the “1”-shaped cells preserve their original orientation.
- (5) A global reflection about the horizontal axis: both the parabolas and the “1”-shaped cells are vertically reflected.

Note that the upright orientation of the “1”-shaped elements is \nearrow , and all rotations and reflections are considered with respect to this original orientation. $J(x,y)$ is the Jacobian of the transformation in question, and $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ are the scanline and interline orientation vectors (see Sec. C.4).

Table C.1: (*continued.*) Different variants of the parabolic transformation (C.9) and some of their local and global properties. The legend and the remarks for the table are given above. Cases 1–4: Local rotations of the cells. Cases 5–8: Local rotoinversions of the cells. Cases 9–12: Same as cases 1–4 but with vertically reflected parabolas. Cases 13–16: Same as cases 5–8 but with vertically reflected parabolas. The different transformations $g(x,y)$ are applied to a periodic dot screen $r(x,y)$ (see Fig. C.1(a)) as *domain* (inverse) transformations, and the distorted result is $r(g(x,y))$ as shown in Figs. C.1(b)–(h).

To illustrate this result, let us return to Figs. C.1(b) and C.1(f)–(h). We see that in all of these figures the level curves are identical in their geometric shape, and what distinguishes between the 4 cases is only the order of the level curves, which determines the local orientation vectors $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ and hence the scanning order of the “1”-shaped elements. The local orientation vectors $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ are given in Table C.1 for each of the transformations in question. Another case illustrating the use of this result for the explanation of the orientation of moiré patterns is shown in Fig. 7.12.

As we can see, the vectors $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ are, in fact, the two columns of the Jacobian matrix of the *inverse* transformation $g^{-1}(x,y)$:

$$\begin{pmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{pmatrix}^{-1} = \frac{1}{J(x,y)} \begin{pmatrix} \frac{\partial g_2}{\partial y} & -\frac{\partial g_1}{\partial y} \\ -\frac{\partial g_2}{\partial x} & \frac{\partial g_1}{\partial x} \end{pmatrix} \quad (\text{C.11})$$

It should be noted, however, that Eqs. (C.10) are based on the assumption that $\mathbf{g}(x,y)$ is used as a *domain* (inverse) transformation, i.e. that it is applied to an original image $r(x,y)$ to give the distorted image $r(\mathbf{g}(x,y))$, as in Figs. C.1 and C.2. But if $\mathbf{g}(x,y)$ is used as a *direct* transformation, then the vectors $\mathbf{v}_1(x,y)$ and $\mathbf{v}_2(x,y)$ are given by the columns of the *inverse* of matrix (C.11), namely, by the columns of the Jacobian matrix (C.2) of $\mathbf{g}(x,y)$ itself:

$$\begin{aligned} \text{Scanline direction:} \quad \mathbf{v}_1(x,y) &= \left(\frac{\partial}{\partial x} g_1(x,y), \frac{\partial}{\partial x} g_2(x,y) \right) \\ \text{Interline direction:} \quad \mathbf{v}_2(x,y) &= \left(\frac{\partial}{\partial y} g_1(x,y), \frac{\partial}{\partial y} g_2(x,y) \right) \end{aligned} \quad (\text{C.12})$$

This is illustrated in Fig. C.5 for the simple case of a linear transformation. The top row of the figure shows the effect of the direct transformation $(u,v) = (x+y, 2y)$, while the bottom row shows the effect of its inverse, $(x,y) = (u - \frac{1}{2}v, \frac{1}{2}v)$. As we can clearly see, the Jacobian matrix of the direct transformation \mathbf{g} is $\begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ and the Jacobian matrix of the inverse transformation \mathbf{g}^{-1} is $\frac{1}{2} \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}$. And indeed, the scanline and interline directions of the direct transformation are given, in accordance with Eqs. (C.12), by $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (1,2)$ (see Fig. C.5(b)), while those of the inverse transformation are given, in accordance with Eqs. (C.10), by $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (-\frac{1}{2}, \frac{1}{2})$ (see Fig. C.5(d)).

C.5 Other properties of $\mathbf{g}(x,y)$ that can be deduced from its Jacobian matrix

The Jacobian matrix of a transformation $\mathbf{g}(x,y)$ can provide further information on the nature of $\mathbf{g}(x,y)$, in addition to the local scaling, reflection and orientation properties already mentioned so far. Some of these additional properties are briefly summarized below.

- (1) If all the entries of the Jacobian matrix are constant numbers (rather than functions of x and y), the first-order derivatives of $\mathbf{g}(x,y)$ are constant and they do not vary from point to point. This means that $\mathbf{g}(x,y)$ is either *affine* or *linear* (depending on whether or not it also shifts the origin). Such transformations always map straight lines into straight lines.
- (2) If there is at most one element in each row and column of the Jacobian matrix which is not identically zero, then the transformation $(u,v) = \mathbf{g}(x,y)$ is *independent*. This means that a change in each of its input variables (x or y) causes a corresponding change in only a single distinct output variable (u or v). Such a mapping will preserve the independence of the coordinate axes. A simple example would be the interchange of the two axes, $\mathbf{g}(x,y) = (y,x)$.

- (3) If the Jacobian matrix is diagonal (i.e. all its elements which are not located on its main diagonal are identically zero), then the transformation $(u,v) = g(x,y)$ is *diagonal*. This means that each output variable of the transformation (u or v) depends only on the corresponding input variable (respectively, x or y), so that the coordinate axes are preserved. A simple example would be the non-linear scaling transformation $g(x,y) = (\log x, \log y)$. Note that a diagonal mapping is more strongly constrained than an independent mapping in which the coordinate axes may be interchanged. A diagonal mapping is necessarily independent.

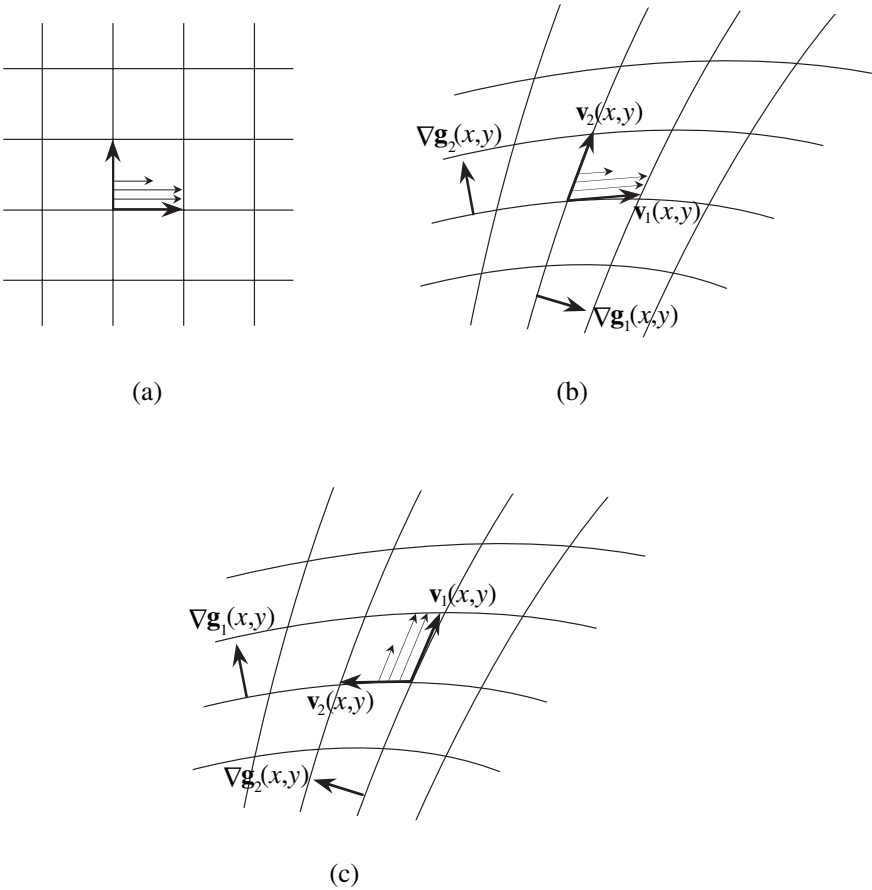


Figure C.4: Explanation of the local orientation properties of a transformation $g(x,y)$. (a) The original coordinate grid before the application of $g(x,y)$. (b) The distorted grid after the application of $g(x,y)$. Each grid cell in (a) is drawn by a succession of scanlines, which is distorted in (b) into a succession of curvilinear scanlines. The scanline direction and the interline direction in (b) determine the local cell orientations in the distorted grid. (c) Another variant of the transformation $g(x,y)$, in which the global geometry remains the same as in (b) but the cells are rotated by 90° .

- (4) The Cauchy-Riemann conditions that determine whether a given transformation $\mathbf{g}(x,y)$ is conformal (see at the end of Sec. B.5 in Appendix B) can be also expressed in terms of the Jacobian matrix of $\mathbf{g}(x,y)$. In these terms, a transformation $\mathbf{g}(x,y)$ is conformal *iff* its Jacobian matrix has the form:

$$(a) \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \quad \text{or:} \quad (b) \begin{pmatrix} a & b \\ b & -a \end{pmatrix} \quad (C.13)$$

where a and b are functions $f_1(x,y)$ and $f_2(x,y)$ (note that the signs of a and b can be also negative). When the elements a and b are constant these matrices correspond to a linear (or affine) *similarity* transformation, namely, a transformation that is composed of a rotation and a uniform scaling [Casselmann04 p. 144; Lay03 p. 339], and possibly also a reflection,² and (in the affine case) a shift. And indeed, in sufficiently small regions a conformal mapping looks like a linear (or affine) similarity transformation: it locally preserves shapes and angles (possibly up to a reflection) — even though it may distort large shapes wildly. We can say, therefore, that a conformal transformation is a transformation that behaves *locally* as a linear (or affine) similarity transformation, although at each point (x,y) the similarity transformation in question may be different. Note that the Jacobian determinant of a conformal transformation $\mathbf{g}(x,y)$ can be zero at some *isolated* points (x,y) ; these are the singular points of the transformation.

A conformal transformation $\mathbf{g}(x,y)$ can be also characterised as *isotropic*. This means that $\mathbf{g}(x,y)$ may apply a local scale factor to the distances between neighbouring points, but this factor does not depend on the orientation of the line between the two points, although it may vary from point to point. An isotropic mapping $\mathbf{g}(x,y)$ will convert a circle at any point in the plane into another circle (but possibly of a different size and in a different place), whereas a non-isotropic mapping would produce an ellipse. If the mapping is also linear then circles of any size will behave in this way, whereas with a non-linear mapping this may only be true for circles of infinitely small size.

Note that an alternative condition on the Jacobian matrix \mathbf{J} in order that the transformation $\mathbf{g}(x,y)$ be conformal is that \mathbf{J} satisfies $\mathbf{J}^T \mathbf{J} = f(x,y) \mathbf{I}$, where \mathbf{J}^T denotes the transpose of the matrix \mathbf{J} , $f(x,y)$ is a function and \mathbf{I} is the identity matrix. The equivalence of this condition with the Cauchy-Riemann conditions (a) or (b) of Eq. (C.13) can be easily obtained by performing explicitly the matrix multiplication $\mathbf{J}^T \mathbf{J}$:

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{pmatrix}$$

where a, b, c, d stand for functions $f_1(x,y), \dots, f_4(x,y)$. By equating the elements of the product matrix with the elements of a diagonal matrix having equal elements on the diagonal we obtain the two identities:

² Note that case (a) in Eq. (C.13) corresponds to a *direct similarity*, that preserves shapes and angles, while case (b) corresponds to an *opposite similarity*, that preserves shapes and angles up to a reflection (meaning that angles are reversed).

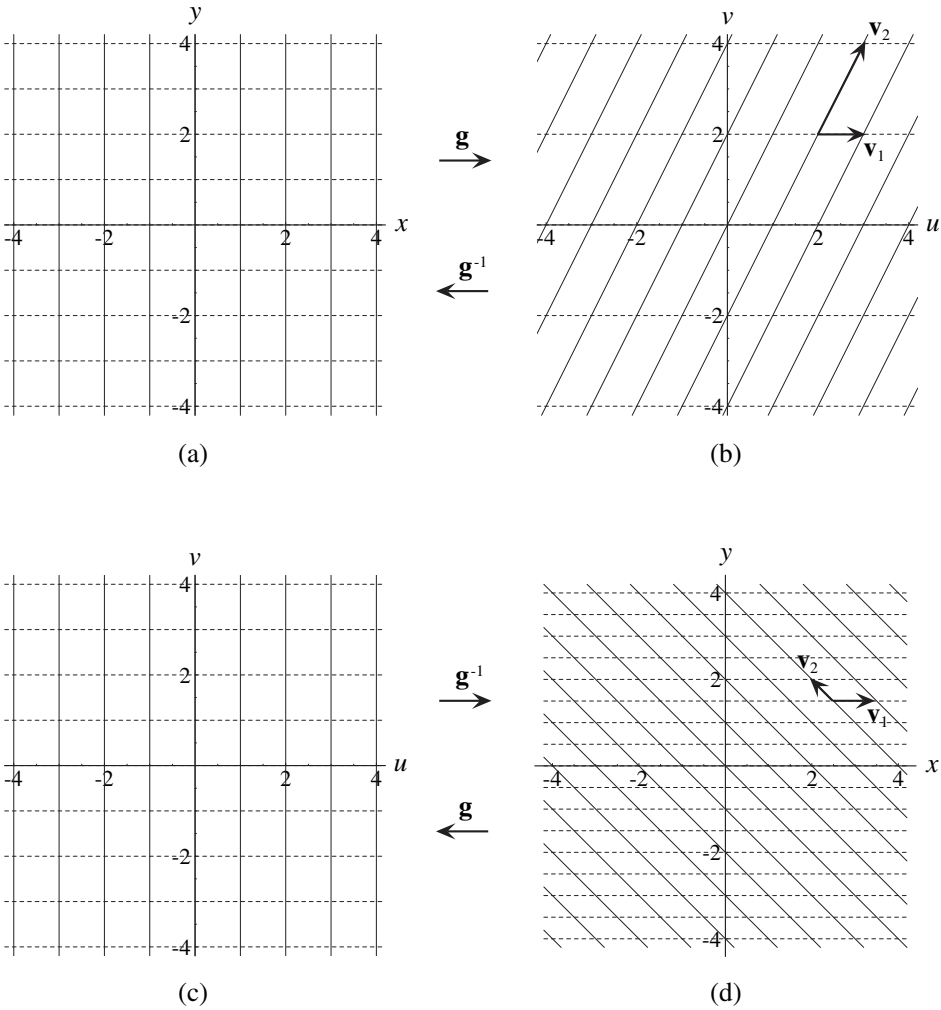


Figure C.5: Top row: the scanline and interline directions of the direct transformation \mathbf{g} given by $(u,v) = (x+y, 2y)$ are $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (1,2)$. Bottom row: the scanline and interline directions of the inverse transformation \mathbf{g}^{-1} , $(x,y) = (u - \frac{1}{2}v, \frac{1}{2}v)$, are $\mathbf{v}_1 = (1,0)$, $\mathbf{v}_2 = (-\frac{1}{2}, \frac{1}{2})$.

$$ab + cd = 0$$

$$a^2 + c^2 = b^2 + d^2$$

Substituting $d = -ab/c$ from the first identity into the second identity we get:

$$c^2(a^2 + c^2) = b^2(a^2 + c^2)$$

which means, in conjunction with the first identity, that either $c = -b$, $a = d$ or $c = b$, $a = -d$; these are, indeed, the Cauchy-Riemann conditions of Eq. (C.13).

Appendix D

Direct and inverse spatial transformations

D.1 Introduction

Because spatial transformations (i.e. 2D functions of the form $(u,v) = \mathbf{g}(x,y)$) are widely used in science and technology, one cannot overestimate the importance of their full understanding. And yet, there exist several different potential sources of confusion in the handling of such transformations. The risk of confusion is increased even further due to the existence of different notation standards, as well as different paradigms for the software algorithms which implement these transformations (or rather their discrete forms) in computer applications. It is therefore our aim in this appendix to develop an intuitive understanding of such transformations, to shed some additional light on their behaviour, and to explain the main sources of confusion and how to avoid them.

We start our discussion in Sec. D.2 with a general reminder whose aim is to put our spatial transformations $(u,v) = \mathbf{g}(x,y)$ in their right mathematical context and to help us understand their various graphical representations. In Sec. D.3 we deepen our understanding of the interconnections between the domain and range planes of a transformation \mathbf{g} . Then, in Sec. D.4 we introduce \mathbf{g}^{-1} , the inverse transformation of \mathbf{g} , and discuss the relationship between these two transformations and their respective coordinate systems. In Sec. D.5 we explain the active and passive interpretations of a transformation \mathbf{g} , and then, in Sec. D.6 we discuss the very important notions of domain and range transformations. In Sec. D.7 we present the relative point of view, which explains the relationship between object deformations and coordinate deformations. In Sec. D.8 we provide several examples that show the various graphical representations of some typical linear and non-linear transformations, to illustrate the main subjects that were discussed so far. In Sec. D.9 we proceed to the explanation of some other possible sources of confusion, including forward and backward transformations in computer applications, and the use of pre-multiplication or post-multiplication formalisms. Then we discuss the implications of all these results to the moiré theory, and in particular to the preparation of our moiré figures (in Sec. D.10), and to the fixed points between the superposed layers (in Sec. D.11). Finally, in Sec. D.12 we derive some useful approximations that allow us to formulate our main results in terms of either direct or inverse transformations.

D.2 Background and basic notions

A transformation (or mapping) from $D \subset \mathbb{R}^m$ to $R \subset \mathbb{R}^n$ is a function $\mathbf{g}: D \rightarrow R$ that returns for each point $(x_1, \dots, x_m) \in D$ a new point $(y_1, \dots, y_n) \in R$. The set D is called the *domain* of the transformation \mathbf{g} , the set R is called the *range* of \mathbf{g} (or the *image* of the

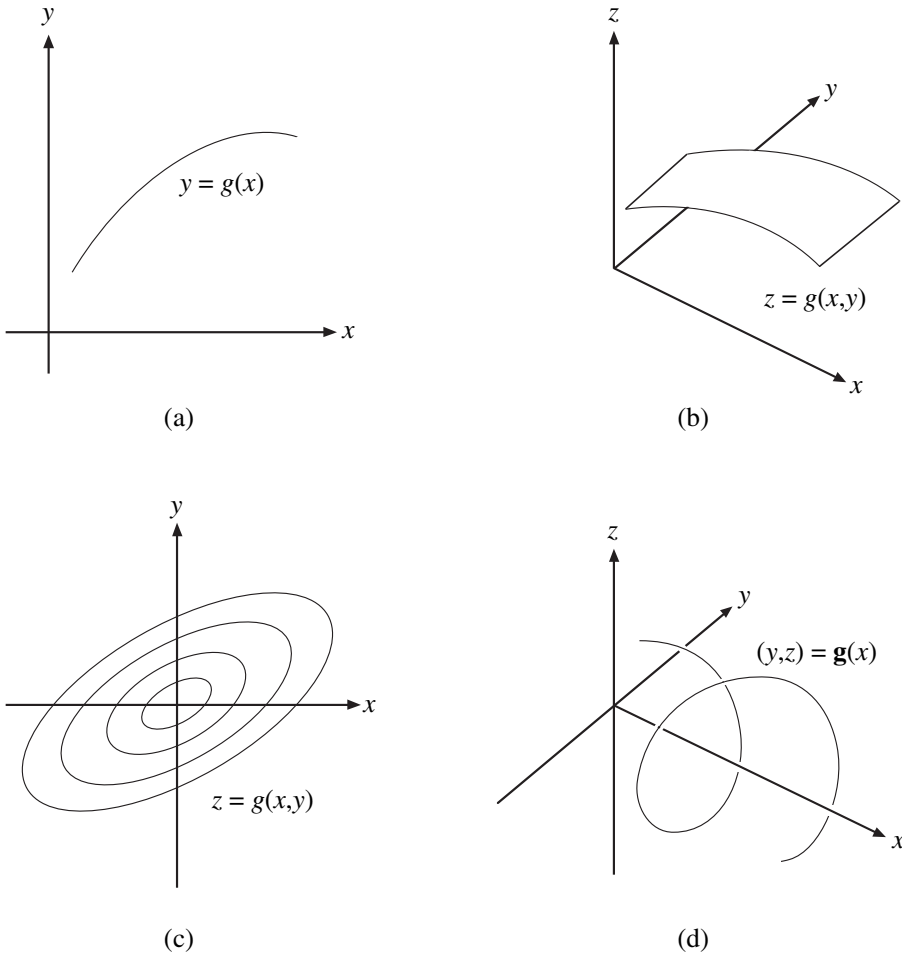


Figure D.1: (a) Schematic graphic representation of a function $y = g(x)$ as a curve in the 2D plane. (b) Schematic graphic representation of a function $z = g(x, y)$ as a surface in the 3D space. (c) Schematic graphic representation of a function $z = g(x, y)$ as a topographic map. (d) Schematic graphic representation of a function $(y, z) = \mathbf{g}(x)$ as a curve in the 3D space.

domain D under \mathbf{g}), and the new point (y_1, \dots, y_n) is said to be the *image* of the original point (x_1, \dots, x_m) under the transformation \mathbf{g} ; symbolically, this is denoted by:

$$(y_1, \dots, y_n) = \mathbf{g}(x_1, \dots, x_m)$$

In the moiré theory we usually have $m = n = 2$, and therefore we will be mainly interested in transformations of the form $(u, v) = \mathbf{g}(x, y)$. However, for didactic reasons, let us first briefly review here the simpler cases in which $m < 2$ or $n < 2$.

The simplest possible case is that of the functions of the form $y = g(x)$, which have a 1D domain and a 1D range. Such functions can be illustrated pictorially as a graph in the x,y plane that shows, for each original point x along the horizontal axis, the image y to which it is mapped by g (see Fig. D.1(a)). Thus, if g is a continuous function, it can be viewed as a curve in the x,y plane. Simple examples include $y = 2x$ or the non-linear function $y = x^2$.

The next simple case is that of the functions of the form $z = g(x,y)$, which have a 2D domain and a 1D range. Such functions can be illustrated pictorially as a 3D graph (or rather as a 2D perspective view of such a 3D graph) that shows for each original point (x,y) in the x,y plane the value z to which it is mapped by g (see Fig. D.1(b)). Thus, if g is a continuous function, it can be interpreted as a surface in the 3D x,y,z space. A function $z = g(x,y)$ can be also represented graphically as a topographic map in the x,y plane; in this case the relief of the surface is represented by level lines (see Fig. D.1(c)).

Yet another simple case, although less frequently encountered, is that of the functions $(y,z) = \mathbf{g}(x)$, which have a 1D domain and a 2D range. We denote such a function by a boldface letter \mathbf{g} since the value $\mathbf{g}(x)$ it returns for each original scalar x is a vector. Such a function can be illustrated, once again, as a 3D graph (or rather as its 2D perspective view). But this time the graph shows, for each given point along the x axis, the value (y,z) to which it is mapped by \mathbf{g} (see Fig. D.1(d)). Thus, if \mathbf{g} is a continuous function, it can be interpreted as a curve in the 3D x,y,z space.

Having reviewed the simpler cases with $m < 2$ or $n < 2$, we arrive now to our main case of interest, that of the transformations $(u,v) = \mathbf{g}(x,y)$, which have a 2D domain and a 2D range. Clearly, a full graphic representation of such a transformation requires a 4D drawing in the four coordinates x,y,u,v , which shows for each original point (x,y) in the 2D domain its corresponding image (u,v) in the 2D range. But because such a 4D drawing is not realizable, several different more-or-less tricky methods exist to allow us represent the transformations $(u,v) = \mathbf{g}(x,y)$ graphically, within the limits of the possible. These different representations of the transformation \mathbf{g} and the interconnections between them have been reviewed in detail in Appendix B; here, for the sake of our introductory survey, we only give a short reminder of the main representations, and we refer the reader to the appropriate sections in Appendix B for further details.

- (1) The most straightforward method for representing the transformation $(u,v) = \mathbf{g}(x,y)$ graphically consists of drawing its 2D domain and its 2D range separately, as two different planes (see, for example, Fig. D.4 in the next section). The 2D domain is drawn with its standard Cartesian x,y coordinate system, and the 2D range is drawn with its own standard Cartesian u,v coordinate system. This representation illustrates the action of the transformation $(u,v) = \mathbf{g}(x,y)$ by showing within its range (the u,v plane) how the original x,y coordinate grid of the domain plane has been affected. This gives in the u,v plane a distorted, curvilinear grid (the image of the original x,y grid under the transformation \mathbf{g}), in addition to the standard Cartesian u,v grid of the u,v plane itself. More details on this method can be found in Sec. B.3 of Appendix B.

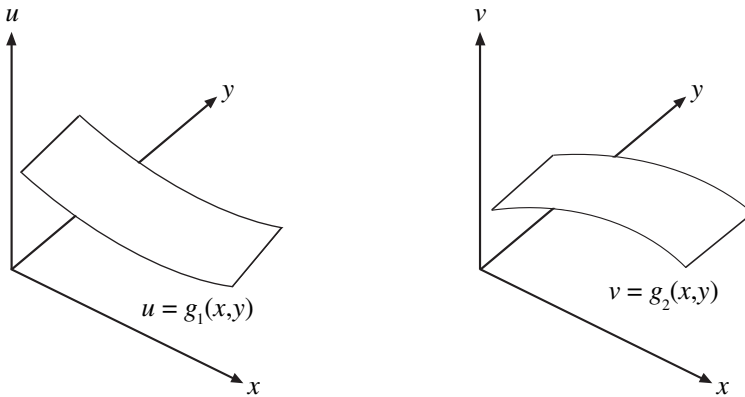


Figure D.2: Schematic graphic representation of a transformation $(u,v) = \mathbf{g}(x,y)$ as a pair of surfaces $u = g_1(x,y)$, $v = g_2(x,y)$ in the 3D space.

- (2) Another variant of method (1) consists of drawing the two planes, the 2D domain and the 2D range, superposed on top of each other within a single plot (see, for example, Fig. D.3 in the next section). Although this graphic representation of $(u,v) = \mathbf{g}(x,y)$ may become overcrowded with details from both planes, it still allows to compare easily the original x,y grid with its distorted image, and it can be used as a “dictionary” (correspondance map) between the x,y and the u,v coordinate systems. For more details on this variant see Sec. B.5 in Appendix B.
- (3) A third representation of the transformation $(u,v) = \mathbf{g}(x,y)$ is based on its component-wise notation:

$$u = g_1(x,y)$$

$$v = g_2(x,y)$$

This notation allows us to interpret the transformation $(u,v) = \mathbf{g}(x,y)$ as a pair of functions of the form $z = g(x,y)$, each of which can be illustrated, as we have seen above, as a surface in the 3D space (see Fig. D.2). More details on this method can be found in Sec. B.2 of Appendix B.

- (4) In another variant of method (3), each of the two surfaces is drawn as a separate topographic map with its own level lines (see, for example, Fig. B.1 in Appendix B). The axes in both of the maps are x and y .
- (5) A further method for representing the transformation $(u,v) = \mathbf{g}(x,y)$ graphically consists of showing its effect within a single planar plot as a *vector field*. This is done by drawing an arrow emanating from each point (x,y) of the x,y plane (or more practically, from some representative points on a given grid within the x,y plane), where the length and the orientation of each arrow indicate the length and the orientation of the vector $(u,v) = \mathbf{g}(x,y)$ that is assigned by \mathbf{g} to the point (x,y) ; see, for example, Fig. D.9(f)

below.¹ It is important to stress, however, that each such arrow does not connect the point (x,y) to its image $(u,v) = \mathbf{g}(x,y)$, but rather to the point $(x,y) + \mathbf{g}(x,y)$. The axes of the vector field remain, therefore, x and y . For more detail on this method see Sec. B.6 in Appendix B.

- (6) In another variant of this method, successive vectors of the vector field are connected into continuous curves, known as the *trajectories* (or *field lines*) of the vector field (see Fig. B.5(b) in Appendix B). This allows to illustrate the effect of the vector field $(u,v) = \mathbf{g}(x,y)$ along its trajectories, much like the graphical description of an electric or magnetic field in physics. More details can be found in Sec. B.6 of Appendix B.

There also exist other, more exotic methods for representing the transformation $(u,v) = \mathbf{g}(x,y)$. Although they are quite rarely used, it may still be interesting to mention some of them briefly:

- (7) In a different variant of the vector field (method (5)), each arrow emanates from the point (x,y) and points to its image $(u,v) = \mathbf{g}(x,y)$. This variant is rarely if ever used in the literature; but in fact it is equivalent to the vector field of the relative transformation $\mathbf{k}(x,y) = \mathbf{g}(x,y) - (x,y)$, where each arrow (before its length is possibly being scaled) connects the point (x,y) to the point $(x,y) + \mathbf{k}(x,y) = \mathbf{g}(x,y)$.
- (8) Another unusual representation of the transformation $(u,v) = \mathbf{g}(x,y)$ is a variant of method (1) in which the action of the transformation is not demonstrated by the way it affects the standard Cartesian x,y grid, but rather by the way it affects some other curves in the x,y plane. This can be advantageous when the transformation \mathbf{g} maps the straight lines of the x,y grid into too complicated curves, or in cases in which it is easier or more interesting to see how \mathbf{g} acts on some other curve families. Perhaps the most classical example of this type is the Cartesian to polar coordinate transformation $(r,\theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$, which is represented in the literature by its effect on the concentric circles and the radius lines that surround the origin of the x,y plane; we will return to this case in more detail in Sec. D.8, Example D.4. Several other examples can be found in [Needham97]; see, for example, the figures in pp. 58, 100 and 163 there. Yet another example appears in [Kreyszig93 p. 745, Fig. 304].

All these different representations of the transformation $(u,v) = \mathbf{g}(x,y)$ are in fact equivalent, although each of them focuses on some different facets of the transformation. It is therefore up to us to choose in each case the most suitable representation of \mathbf{g} , depending on the circumstances. For example, if we are mainly interested in the effect of the transformation \mathbf{g} on the original Cartesian coordinate system, the most natural representation to consider is (1); but if we want to visually detect the critical points of \mathbf{g} (the points where $\mathbf{g}(x,y) = (0,0)$), then the most suitable representation would certainly be (5) or (6). In the remaining sections of this appendix we will mainly use representations

¹ For practical reasons it is customary to scale the arrow lengths in the drawing by a constant factor, in order to avoid drawings with too short, hard-to-see arrows, or drawings with too long, overlapping arrows.

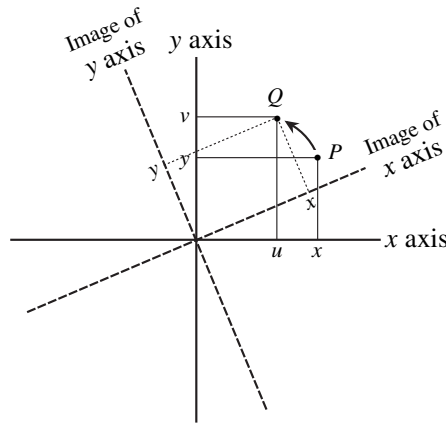


Figure D.3: The effect of a transformation $(u,v) = \mathbf{g}(x,y)$ (in the present example: a rotation by angle α), shown within the original x,y coordinate system. Domain coordinates x,y and range coordinates u,v are marked along the same original axes. Note that the dashed lines *do not* represent the u and v axes (see Remark D.3).

(1) and (2) and their counterparts for the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$. Note, however, that some of the other methods are also frequently used in this book; it is therefore important to recognize the different methods correctly and to understand which of them is being used in each case, in order to avoid confusion. In particular, we should be attentive to the fact that one and the same transformation may look completely different in each of its various graphic representations; several illustrative examples are provided in Sec. D.8 below.

D.3 A deeper look into the domain and range planes of the mapping $(u,v) = \mathbf{g}(x,y)$

Suppose we are given a transformation $(u,v) = \mathbf{g}(x,y)$ that operates on the x,y plane: $\mathbf{g}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$. As a simple example we may consider the transformation which rotates the plane by angle α about the origin. When the transformation \mathbf{g} is applied, it moves each point $P = (x,y)$ of the x,y plane to a new point Q whose location in the same x,y plane is given by $\mathbf{g}(x,y)$. This mapping effect of \mathbf{g} is often denoted in literature by $(x,y) \mapsto \mathbf{g}(x,y)$ or $(x,y) \mapsto (u,v)$ (see, for example, [Lang87 p. 386]). Fig. D.3 shows the image Q of a point P under the transformation of rotation by angle α , as well as the images of the original x and y axes under the same transformation (the dashed lines).

Note that in Fig. D.3 we have drawn the effect of the transformation $(u,v) = \mathbf{g}(x,y)$ within the original x,y coordinate system (also called the x,y space). This means that Fig. D.3 shows simultaneously the situations before and after the application of the transformation \mathbf{g} (see method (2) above). But for the sake of clarity, in order not to overload the drawing

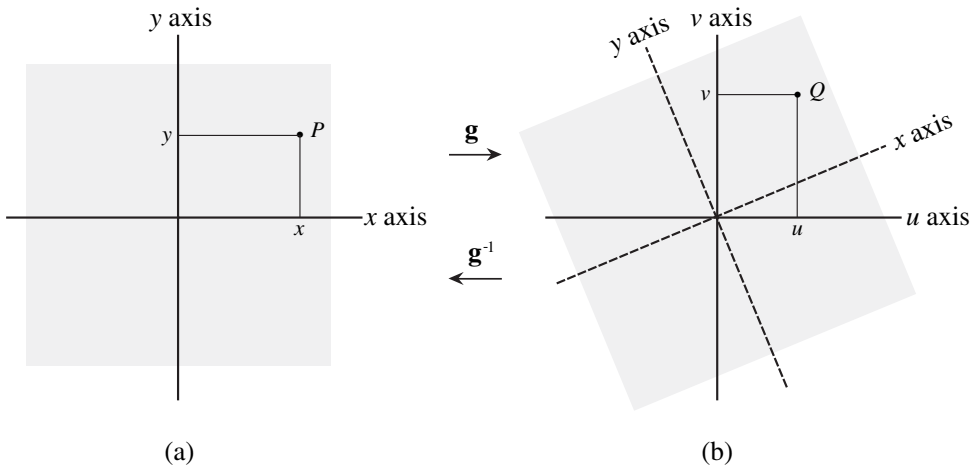


Figure D.4: The effect of a transformation $(u,v) = \mathbf{g}(x,y)$ (the same transformation as in Fig. D.3), shown in two separate plots: (a) The domain of \mathbf{g} (the x,y plane before the application of the transformation). (b) The range of \mathbf{g} (the u,v plane after the application of the transformation). A more detailed version of this figure is provided in Fig. D.5.

with details, it is often preferable to use method (1), and to illustrate the original and the transformed planes (i.e. the domain and the range of \mathbf{g}) in separate plots, as shown in Fig. D.4. In this case the standard axes of the domain of \mathbf{g} are denoted x and y , as shown in Fig. D.4(a); but after the transformation has been applied, i.e. in the range of \mathbf{g} , the x and y axes are already transformed (distorted or simply moved to new locations), and the role of the standard axes is taken over by u and v (see Fig. D.4(b)). Note that the new standard axes u and v are identical to the old standard x and y axes as they existed before the application of \mathbf{g} . A more detailed version of Fig. D.4 is shown in Fig. D.5; this figure clearly shows the effect of the rotation transformation \mathbf{g} , which simply rotates the entire plane shown in (a) into the plane shown in (b). The images of the original x and y axes under the transformation \mathbf{g} are called in Fig. D.5(b) the x and y axes, since they correspond, respectively, to the curves $y = 0$ and $x = 0$ in the u,v coordinate system of Fig. D.5(b). These axes form, indeed, the transformed coordinate system of the u,v plane, and in the general case they may be curvilinear (see, for example, Figs. D.9(a),(b) below).

It may be sometimes helpful to describe the effect of the transformation $(u,v) = \mathbf{g}(x,y)$ using the following physical interpretation (see Figs. D.9(a),(b)): Imagine that the original x,y coordinate system is printed on a flat sheet of flexible rubber, and that this sheet undergoes a planar transformation while remaining flat. Depending on the forces that are applied to that rubber sheet, the result may appear rotated, scaled, or otherwise distorted; but it always remains flat. Now, we copy the resulting distorted x,y plane on a new sheet of paper; this sheet corresponds to the range of the transformation $\mathbf{g}(x,y)$ (see Fig. D.9(b)). We draw on this sheet a new Cartesian coordinate system, *identical* to the original

untransformed x,y coordinate system of Fig. D.9(a); these new undistorted axes are the u and v coordinates of the range of our transformation (see Fig. D.9(b)).

Remark D.1: The fact that the new u,v axes are identical to the original, undistorted x,y axes will allow us later to compare any object in the plane *before* and *after* it undergoes the transformation $(u,v) = \mathbf{g}(x,y)$. ■

Remark D.2: Note the double role of the x,y coordinate system: on the one hand it refers to the original, *undistorted* coordinate system *before* the application of the transformation $(u,v) = \mathbf{g}(x,y)$, i.e. in the *domain* of $\mathbf{g}(x,y)$; but on the other hand it also refers to the *distorted* coordinate system *after* the application of the transformation, i.e. in the *range* of $\mathbf{g}(x,y)$, whose new undistorted coordinates are u,v . ■

Having understood the coordinate systems involved in the domain and in the range of the transformation \mathbf{g} , we now present the terminology that is used to refer to them.² The u,v coordinate system of the range of \mathbf{g} (see Fig. D.5(b)) is called the *u,v space*, the *target space* or the *destination space* of the transformation \mathbf{g} . The x,y coordinate system of Fig. D.5(a), showing the situation before the transformation \mathbf{g} has been applied, is called the *original x,y space*, and the distorted x,y coordinate system of Fig. D.5(b), showing the situation after the transformation \mathbf{g} has been applied, is called the *\mathbf{g} -transformed x,y space* or simply the *transformed x,y space*. Note that the values of the x,y coordinates are not affected by the transformation: If point P has coordinates (x,y) in the original x,y space, its image Q has the same (possibly curvilinear) coordinates (x,y) in the transformed x,y space (while its Cartesian coordinates in terms of the u,v space are $(u,v) = \mathbf{g}(x,y)$).³

As we can see, Fig. D.5 has an important advantage over Fig. D.4 in that it allows us to easily visualize for any given point in the plane the correspondence between its x,y coordinate values and its u,v coordinate values. We will return to this point in Sec. D.5, where we discuss the active and passive interpretations of a transformation $(u,v) = \mathbf{g}(x,y)$.

Remark D.3: Since the transformation $(u,v) = \mathbf{g}(x,y)$ maps any original point (x,y) into its image (u,v) , it also maps in Fig. D.3 the original x and y axes into the two respective dashed lines. Therefore, it may be tempting to call these dashed lines in Fig. D.3 “the u and v axes”. However, this reasoning is wrong since the points $(x,0)$, which form the x axis, are not mapped by the transformation \mathbf{g} into the points $(u,0)$, which form the u axis, but rather into the points $(u,v) = \mathbf{g}(x,0)$. And indeed, as clearly shown in Fig. D.5(a), the u and v axes are obtained by applying to the x and y axes the *inverse* transformation (in our case: a rotation in the *negative* direction). To better understand this, consider our example of a rotation by angle α . This transformation is given by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\alpha - y\sin\alpha \\ x\sin\alpha + y\cos\alpha \end{pmatrix} \quad (\text{D.1})$$

² Unfortunately, as we will see later in Sec. D.6.2, a different convention also coexists, and is being used.

³ Note that although u and v are obtained by transforming the original x,y coordinates through $(u,v) = \mathbf{g}(x,y)$, they appear in Fig. D.5(b) as *untransformed* (because they are the standard coordinates of the range of \mathbf{g}), while the transformed axes in this figure belong to the transformed x,y coordinates.

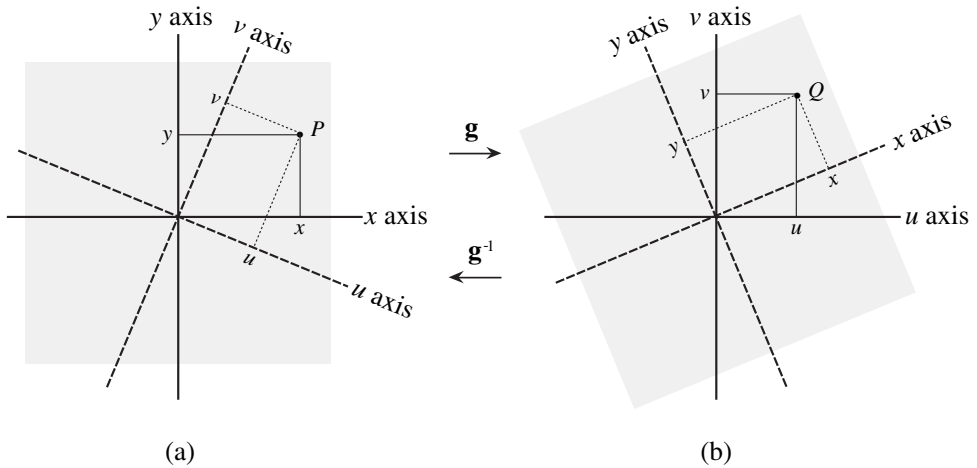


Figure D.5: Same as Fig. D.4, but this time showing both the x, y and the u, v coordinates in (a) as well as in (b). This figure clearly illustrates the active interpretation of the transformation $(u, v) = g(x, y)$, which amounts to the rotation of the entire plane of (a) into the plane shown in (b). Each of the views (a) and (b) also illustrates the passive interpretation of the same transformation, which consists of the conversion of x, y coordinates into u, v coordinates. (The active and passive interpretations are discussed in Sec. D.5.)

Its inverse transformation, $(x, y) = g^{-1}(u, v)$, is the rotation by angle $-\alpha$, namely:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u\cos\alpha + v\sin\alpha \\ -u\sin\alpha + v\cos\alpha \end{pmatrix} \quad (\text{D.2})$$

Now, the u axis is given by definition by the curve $v = 0$. This curve can be also expressed in our example, using the bottom row of Eq. (D.1), as:

$$x\sin\alpha + y\cos\alpha = 0$$

namely: $y = -x\tan\alpha = x\tan(-\alpha)$

This means that the u axis is represented in Fig. D.3 by a line whose angle is $-\alpha$, and not by the dashed line whose angle is α . This fact is clearly shown in Fig. D.5(a).

On the other hand, the dashed lines in Fig. D.5(b) do represent the x and y axes. To see this, remember that the x axis is, by definition, the line $y = 0$, which can be also expressed in our case, using the bottom row of Eq. (D.2), as:

$$-u\sin\alpha + v\cos\alpha = 0$$

namely: $v = u\tan\alpha$

This is clearly a line passing through the origin of the u,v plane at the angle of α . This means that the dashed line located above the u axis in the u,v system of Fig. D.5(b) represents, indeed, the x axis. A similar reasoning can be formulated for the y axis, too.

The possible confusion in labelling the transformed axes in Fig. D.3 occurs since Fig. D.3 attempts to show both the domain of \mathbf{g} (Fig. D.5(a)) and the range of \mathbf{g} (Fig. D.5(b)) within the same coordinate system, while the axis names in Fig. D.5(a) and Fig. D.5(b) are not the same. This fact further justifies why we prefer to illustrate the effect of a transformation \mathbf{g} in two separate plots, as shown in Figs. D.4 or D.5, rather than in a single plot, as in Fig. D.3. Drawing a single plot is, of course, allowable, and sometimes even advantageous, but it should be done with care.⁴ ■

D.4 2D transformations and their inverse

In order to further clarify the situation in the 2D case, let us describe in more detail the effect of the 2D transformation $(u,v) = \mathbf{g}(x,y)$ and the effect of its inverse $(x,y) = \mathbf{g}^{-1}(u,v)$ on each point of the plane (see Fig. D.9).⁵ Note that in the following we will sometimes need the componentwise notation of the transformation $(u,v) = \mathbf{g}(x,y)$:

$$\begin{aligned} u &= g_1(x,y) \\ v &= g_2(x,y) \end{aligned} \tag{D.3}$$

Similarly, the componentwise notation of the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ is:

$$\begin{aligned} x &= g_1^{-1}(u,v) \\ y &= g_2^{-1}(u,v) \end{aligned} \tag{D.4}$$

where $g_1^{-1}(u,v)$ and $g_2^{-1}(u,v)$ are the components of the inverse transformation $\mathbf{g}^{-1}(u,v)$. Note that the inverse of the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ is the original transformation $(u,v) = \mathbf{g}(x,y)$ itself. The original transformation \mathbf{g} is also called the *direct* transformation (as opposed to the *inverse* transformation \mathbf{g}^{-1}).

Remark D.4: Once we have defined the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$, it becomes a transformation in its own right, and we can obviously write it with any variables we may wish. Thus, if we prefer to maintain our convention of using the x,y variables for the transformation's domain and the u,v variables for its range, we may write our inverse

⁴ For example, if one insists on naming the rotated, dashed axes of Fig. D.3 by u and v , then the transition from x,y to u,v values must be expressed mathematically by the *inverse* transformation, in our case: a rotation by angle $-\alpha$. This practice can be found, for example, in [Knopp74 pp. 401–407] or in [Spiegel68 p. 36]. Note that in references using this convention the transformation $(u,v) = \mathbf{g}(x,y)$ is often called “a transition from u,v to x,y ” (see, for example, [Knopp74 p. 406]), in order to avoid the inversion effect between the figure and the formula (see also Remark D.16 in Sec. D.6.2).

⁵ For the sake of the present discussion we suppose that the transformation $(u,v) = \mathbf{g}(x,y)$ is sufficiently well behaved, and that it has a unique inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$.

transformation as: $(u,v) = \mathbf{g}^{-1}(x,y)$. This also allows us to compare \mathbf{g} and \mathbf{g}^{-1} by plotting them together in the same drawing without having to bother about the axes names. Moreover, this even allows us to define new combined transformations such as $\mathbf{h}(x,y) = \mathbf{g}(x,y) - \mathbf{g}^{-1}(x,y)$. In fact, for any rational polynomial $r(x) = \sum a_n x^n$ with positive and negative powers $n \in \mathbb{Z}$ we can define a corresponding functional polynomial in $\mathbf{g}(x,y)$, $\mathbf{r}(x,y) = \sum a_n \mathbf{g}^{[n]}(x,y)$, where the “powers” in brackets indicate composition of transformations (or inverse transformations), as follows: $\mathbf{g}^{[1]} = \mathbf{g}$, $\mathbf{g}^{[2]} = \mathbf{g} \circ \mathbf{g}$, $\mathbf{g}^{[3]} = \mathbf{g} \circ \mathbf{g} \circ \mathbf{g}$, etc., $\mathbf{g}^{[0]} = \mathbf{i}$ (the identity transformation $\mathbf{i}(x,y) = (x,y)$), $\mathbf{g}^{[-1]} = \mathbf{g}^{-1}$ (the inverse transformation of \mathbf{g}), $\mathbf{g}^{[-2]} = \mathbf{g}^{-1} \circ \mathbf{g}^{-1}$, etc. For example, the rational polynomial $r(x) = 3x^2 + 5x + 2 + \frac{4}{x}$ defines the functional polynomial:

$$\begin{aligned}\mathbf{r}(x,y) &= 3\mathbf{g}^{[2]}(x,y) + 5\mathbf{g}^{[1]}(x,y) + 2\mathbf{g}^{[0]}(x,y) + 4\mathbf{g}^{[-1]}(x,y) \\ &= 3[\mathbf{g} \circ \mathbf{g}](x,y) + 5\mathbf{g}(x,y) + 2(x,y) + 4\mathbf{g}^{-1}(x,y)\end{aligned}$$

Thus, one should not be shocked if we occasionally say that the inverse of the transformation $\mathbf{g}(x,y)$ is $\mathbf{g}^{-1}(x,y)$ (rather than $\mathbf{g}^{-1}(u,v)$). In fact, it would be desirable to stick systematically to either of these two conventions; however, it turns out that each of the two may be more suitable in some different situations. We therefore have to live with both conventions, but whenever this may cause confusion we will add an adequate remark to clarify our intentions. ■

D.4.1 The image of the standard Cartesian grid under the transformations \mathbf{g} and \mathbf{g}^{-1}

Suppose now that a moving point $P = (x,y)$ describes a curve in the domain of our transformation $\mathbf{g}(x,y)$, i.e. within the x,y plane. As we already know, the image of this point in the range of our transformation will likewise describe a curve in the u,v plane, which is called the *image curve* of the original curve.⁶ For example, to the line $x = c$, which is parallel to the y axis, there corresponds in the u,v plane the image curve given in parametric form by the pair of equations [Courant88 pp. 134–135]:

$$u = g_1(c,y)$$

$$v = g_2(c,y)$$

or, more concisely:

$$(u,v) = \mathbf{g}(c,y) \tag{D.5}$$

where y is the parameter of the curve. Similarly, to the line $y = k$ there corresponds in the u,v plane the image curve given in parametric form by the pair of equations:

$$u = g_1(x,k)$$

$$v = g_2(x,k)$$

⁶ We could also draw the resulting curve within the original x,y plane, like in Fig. D.3; but as already mentioned above, we prefer to draw it in a separate figure in order not to overload the original figure, and in order to avoid confusion in the axis names.

or, more concisely:

$$(u,v) = \mathbf{g}(x,k) \quad (\text{D.6})$$

where x is the parameter of the curve.⁷ Note that the image curves (D.5) and (D.6) can be also expressed in the implicit form, in terms of the *inverse* transformation $(x,y) = \mathbf{g}^{-1}(u,v)$. Since they are the image curves of the lines $x = c$ and $y = k$, they can be written, respectively, using Eqs. (D.4), as follows [Courant88 p. 135]:

$$g_1^{-1}(u,v) = c \quad (\text{D.7})$$

and:
$$g_2^{-1}(u,v) = k \quad (\text{D.8})$$

Now, if we assign to c and k sequences of equidistant values c_1, c_2, c_3, \dots and k_1, k_2, k_3, \dots (for instance, consecutive integer values), then the rectangular coordinate grid consisting of the lines $x = c_i$ and $y = k_i$ in the x,y plane (see Fig. D.9(a)) gives rise to a corresponding curvilinear grid consisting of two families of curves (D.7) and (D.8) in the u,v plane (see Fig. D.9(b)). This curvilinear grid furnishes a useful geometric picture of the mapping $(u,v) = \mathbf{g}(x,y)$ that clearly shows how it distorts the original x,y plane.

Remark D.5: It follows from Eqs. (D.7) and (D.8) that the two families of curvilinear lines in the u,v plane, that are the image under \mathbf{g} of the standard unit grid of the x,y plane, are simply the level lines of the two surfaces defined, respectively, by the two components of the *inverse* transformation \mathbf{g}^{-1} : The images of the lines $x = m, m \in \mathbb{Z}$ under $(u,v) = \mathbf{g}(x,y)$ are the level lines $g_1^{-1}(u,v) = m$ of the surface $z = g_1^{-1}(u,v)$ (see the plain lines in Fig. D.9(b)), and the images of the lines $y = n, n \in \mathbb{Z}$ under $(u,v) = \mathbf{g}(x,y)$ are the level lines $g_2^{-1}(u,v) = n$ of the surface $z = g_2^{-1}(u,v)$ (see the dashed lines in Fig. D.9(b)). ■

Consider now the inverse transformation, $(x,y) = \mathbf{g}^{-1}(u,v)$. Obviously, this transformation maps the curvilinear lines in the u,v plane that are defined by Eq. (D.5) or, equivalently, by Eq. (D.7) (see the plain curves in Fig. D.9(b)) back into the original straight vertical lines $x = c$ in the x,y plane (Fig. D.9(a)). Likewise, it maps the curvilinear lines in the u,v plane that are defined by Eq. (D.6) or, equivalently, by Eq. (D.8) (see the dashed curves in Fig. D.9(b)) back into the original straight horizontal lines $y = k$ in the x,y plane. Thus, the inverse transformation undoes the effects of the original transformation $\mathbf{g}(x,y)$, and brings each distorted entity in the u,v plane back into its undistorted state in the original x,y plane. This can be clearly seen by comparing Figs. D.9(b) and D.9(a).

However, it is also possible to consider the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ as a transformation in its own right, as shown in Figs. D.9(c),(d).⁸ From this point of view, the

⁷ In a similar way we can also find the image of a *curve*, say, $y = x^2 + k$, by plugging its equation (instead of the equation $y = k$) into $(u,v) = \mathbf{g}(x,y)$. For example, the image of the curve $y = x^2 + k$ in the u,v plane is given in parametric form by $(u,v) = \mathbf{g}(x, x^2 + k)$. This is further explained in Remark D.8 below.

⁸ As we have seen in Remark D.4, because \mathbf{g}^{-1} is a transformation in its own right, we could also consider it in the x,y space. But in order to avoid confusion we prefer to use here the variables x,y for the domain of \mathbf{g} (and the range of \mathbf{g}^{-1}), and the variables u,v for the range of \mathbf{g} (and the domain of \mathbf{g}^{-1}).

transformation $\mathbf{g}^{-1}(u,v)$ maps the standard Cartesian coordinate grid of its own domain, the u,v plane, into a curvilinear grid within its range, the x,y plane. Thus, to each line $u = p$, which is parallel to the v axis in the u,v plane, there corresponds in the x,y plane a curve, which is given in parametric form by the pair of equations:

$$x = g_1^{-1}(p,v)$$

$$y = g_2^{-1}(p,v)$$

or, more concisely:

$$(x,y) = \mathbf{g}^{-1}(p,v) \quad (\text{D.9})$$

where v is the parameter of the curve. Similarly, to the line $v = q$ in the u,v plane there corresponds in the x,y plane a curve, which is given in parametric form by the pair of equations:

$$x = g_1^{-1}(u,q)$$

$$y = g_2^{-1}(u,q)$$

or, more concisely:

$$(x,y) = \mathbf{g}^{-1}(u,q) \quad (\text{D.10})$$

where u is the parameter of the curve. Note that the curves (D.9) and (D.10) can be also expressed in the implicit form, in terms of the *original* transformation $\mathbf{g}(x,y)$. Since they are the image curves of the lines $u = p$ and $v = q$, they can be written, respectively, using Eqs. (D.3):

$$g_1(x,y) = p \quad (\text{D.11})$$

$$\text{and:} \quad g_2(x,y) = q \quad (\text{D.12})$$

If we assign to p and q sequences of equidistant values p_1, p_2, p_3, \dots and q_1, q_2, q_3, \dots (for instance, consecutive integer values), then the rectangular coordinate grid consisting of the lines $u = p_i$ and $v = q_i$ in the u,v plane (Fig. D.9(c)) gives rise to a corresponding curvilinear grid in the x,y plane (Fig. D.9(d)) which consists of the two families of curves (D.9) and (D.10). This curvilinear grid furnishes a useful geometric picture of the inverse mapping $(x,y) = \mathbf{g}^{-1}(u,v)$ that clearly shows how it distorts its original u,v plane.

Remark D.6: It follows from Eqs. (D.11) and (D.12) that the two families of curvilinear lines in the x,y plane, that are the image under \mathbf{g}^{-1} of the standard unit grid of the u,v plane, are simply the level lines of the two surfaces defined, respectively, by the two components of the *direct* transformation \mathbf{g} : The images of the lines $u = m$, $m \in \mathbb{Z}$ under $(x,y) = \mathbf{g}^{-1}(u,v)$ are the level lines $g_1(x,y) = m$ of the surface $z = g_1(x,y)$ (see the plain lines in Fig. D.9(d)), and the images of the lines $v = n$, $n \in \mathbb{Z}$ under $(x,y) = \mathbf{g}^{-1}(u,v)$ are the level lines $g_2(x,y) = n$ of the surface $z = g_2(x,y)$ (see the dashed lines in Fig. D.9(d)). ■

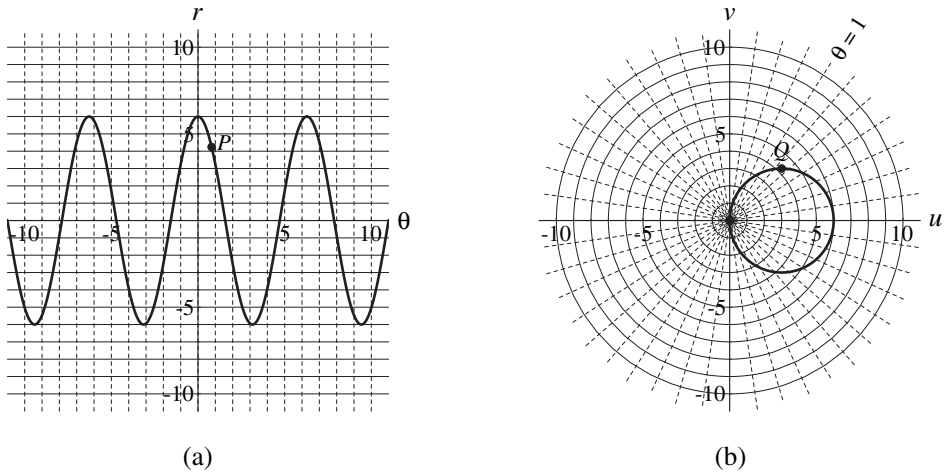


Figure D.6: Transformation (D.13) maps the cosinusoidal curve $r = 6\cos\theta$ in the θ, r plane (a) into the circle $(u-3)^2 + v^2 = 9$ in the u, v plane (b). Note that the radial and circular grid lines in (b) are the images of the vertical and horizontal θ, r grid lines in (a) under the transformation (D.13); they represent, therefore, the transformed θ, r plane. The values along the u and v axes in (b) correspond to the Cartesian coordinates of the destination u, v plane, i.e. to the vertical and horizontal grid lines $u = c$, $v = k$; but these lines are not shown in the figure in order not to overload it with details. Note that all angles are measured in radians.

Remark D.7: Note that one can express the curvilinear grid lines of both $\mathbf{g}(x, y)$ and its inverse $\mathbf{g}^{-1}(u, v)$ without knowing the *explicit* expression of the inverse transformation: The curvilinear grid lines of $\mathbf{g}(x, y)$ can be expressed in *parametric* form by Eqs. (D.5) and (D.6), and the curvilinear grid lines of $\mathbf{g}^{-1}(u, v)$ can be expressed in *implicit* form by Eqs. (D.11) and (D.12). All of these equations only require the knowledge of the direct transformation $\mathbf{g}(x, y)$. ■

This last fact allows us in Fig. D.9 to plot the effects of both \mathbf{g} and \mathbf{g}^{-1} even if the explicit expression of \mathbf{g}^{-1} is unavailable. This technique has been used, indeed, to generate the figures that accompany the examples in Sec. D.8.

D.4.2 The image of a general curve under the transformations \mathbf{g} and \mathbf{g}^{-1}

We have seen above in detail how the transformation $(u, v) = \mathbf{g}(x, y)$ acts on the straight grid lines $x = c$ and $y = k$ of the x, y plane, and how the images under \mathbf{g} of these grid lines are expressed as curves in the u, v plane. Similarly, we have also seen how the inverse transformation $(x, y) = \mathbf{g}^{-1}(u, v)$ acts on the straight grid lines $u = c$ and $v = k$ of the u, v plane, and how the images under \mathbf{g}^{-1} of these grid lines are expressed as curves back in the x, y plane. But because the transformation $(u, v) = \mathbf{g}(x, y)$ maps the x, y plane into the u, v plane, it is clear that it also maps any subset of the x, y plane into a subset of the u, v plane.

For example, the rotation transformation (D.1) maps any subset of the x,y plane into a similar subset of the u,v plane which has only been rotated by angle α about the origin. A more interesting example is given by the non-linear transformation:⁹

$$\begin{aligned} u &= r \cos \theta \\ v &= r \sin \theta \end{aligned} \tag{D.13}$$

This transformation converts polar coordinates θ, r into Cartesian coordinates u, v . Knowing the inverse of this transformation (see Example D.4 in Sec. D.8 below), $r = \sqrt{u^2 + v^2}$, $\theta = \arctan(v/u)$, we can see from Eqs. (D.7) and (D.8) how transformation (D.13) acts on the straight grid lines of the θ, r plane: It maps each horizontal line $r = c$ of the θ, r plane (where c is an arbitrary constant) into a circle centered about the origin with radius c in the u, v plane, that is expressed by $\sqrt{u^2 + v^2} = c$. Similarly, it maps each vertical line $\theta = k$ of the θ, r plane into a radial line in the u, v plane which emanates from the origin at the angle of k radians, and whose expression is $\arctan(v/u) = k$.

Proceeding with the same example, consider now the planar curve defined by the equation $r = 6 \cos \theta$. When plotted in the θ, r plane this equation gives a cosinusoidal curve (see Fig. D.6(a)).¹⁰ However, the image of this curve in the u, v plane under the transformation (D.13) is a circle tangent to the vertical axis, as shown in Fig. D.6(b) [Colley98 pp. 68–69]. In other words, when the point P in Fig. D.6(a) traces out the cosinusoidal curve, its image Q traces out the circle shown in Fig. D.6(b). We see, therefore, that the non-linear transformation (D.13) maps cosinusoidal curves in the polar θ, r plane into circles tangent to the vertical axis in the Cartesian u, v plane. Other interesting examples showing how various curves in the θ, r plane are transformed into the u, v plane under the transformation (D.13) can be found in [Lang87 pp. 254–257].

So how can we express mathematically the new curve which is obtained in the u, v plane as the image under transformation \mathbf{g} of a given curve $f(x, y) = 0$ in the x, y plane? Just as we did in the case of the straight grid lines, we replace each occurrence of x and y in the curve equation $f(x, y) = 0$ by the respective component from Eq. (D.4), and we obtain: $f(g_1^{-1}(u, v), g_2^{-1}(u, v)) = 0$, or more concisely, using vector notation: $f(\mathbf{g}^{-1}(u, v)) = 0$. This is, indeed, the implicit form of our image curve in the u, v plane.

Returning to our example, the image in the u, v plane of the original cosinusoidal curve $r = 6 \cos \theta$ under the transformation (D.13) is obtained by plugging in this curve equation the two components of the inverse transformation, $r = \sqrt{u^2 + v^2}$, $\theta = \arctan(v/u)$. We get, therefore:

⁹ Note that the names of the variables have no real importance, and they can be chosen as desired. In this case we have preferred to keep using the range coordinates u, v in accordance with our usual convention, but to use the classical polar coordinate names r, θ rather than our usual domain coordinate names x, y . Depending on the context we may prefer in other circumstances to make other choices, such as $(u, v) = (x \cos y, x \sin y)$ or $(x, y) = (r \cos \theta, r \sin \theta)$.

¹⁰ Note that because in this case r is considered as a function of θ , it is more natural to talk about the θ, r plane, in which θ is the horizontal axis and r is the vertical axis. But in other situations it is often more convenient to refer to the polar coordinates plane as the r, θ plane.

$$\sqrt{u^2 + v^2} = 6 \cos(\arctan(v/u))$$

and using the identity $\cos x = 1/\sqrt{1 + \tan^2 x}$ [Spiegel68 p. 15] and hence $\cos(\arctan(v/u)) = u/\sqrt{u^2 + v^2}$ we obtain the desired curve equation in the u, v plane:

$$u^2 + v^2 = 6u$$

or equivalently, $(u-3)^2 + v^2 = 9$

This is, indeed, a circle tangent to the vertical v axis (see Fig. D.6(b)).

Similar results can be also obtained for the image of a curve $f(u, v) = 0$ under the inverse transformation $(x, y) = \mathbf{g}^{-1}(u, v)$. We have, therefore, the following general result:

Proposition D.1: Let $f(x, y) = 0$ be a curve in the x, y plane. The image in the u, v plane of this curve after the application of the direct transformation $(u, v) = \mathbf{g}(x, y)$ is $f(\mathbf{g}_1^{-1}(u, v), \mathbf{g}_2^{-1}(u, v)) = 0$, or in vector notation, $f(\mathbf{g}^{-1}(u, v)) = 0$. Conversely, if $f(u, v) = 0$ is a curve in the u, v plane, then the image in the x, y plane of this curve under the inverse transformation $(x, y) = \mathbf{g}^{-1}(u, v)$ is $f(\mathbf{g}_1(x, y), \mathbf{g}_2(x, y)) = 0$, or in vector notation, $f(\mathbf{g}(x, y)) = 0$. ■

Remark D.8: It is interesting to note that if the curve $f(x, y) = 0$ is also known in parametric form, $x = f_1(t)$, $y = f_2(t)$, then the image in the u, v plane of this curve under the direct transformation $(u, v) = \mathbf{g}(x, y)$ can be also expressed, in parametric form, by $(u, v) = \mathbf{g}(f_1(t), f_2(t))$.¹¹ Similarly, if the curve $f(x, y) = 0$ is known in the explicit form $y = h(x)$, then the image in the u, v plane of this curve under the same direct transformation $(u, v) = \mathbf{g}(x, y)$ can be also expressed in the parametric form $(u, v) = \mathbf{g}(x, h(x))$.¹²

Conversely, if the curve $f(u, v) = 0$ is also known in parametric form, $u = f_1(s)$, $v = f_2(s)$, then the image in the x, y plane of this curve under the inverse transformation $(x, y) = \mathbf{g}^{-1}(u, v)$ can be also expressed, in parametric form, by $(x, y) = \mathbf{g}^{-1}(f_1(s), f_2(s))$. Similarly, if the curve $f(u, v) = 0$ is known in the explicit form $v = h(u)$, then the image in the x, y plane of this curve under the same inverse transformation $(x, y) = \mathbf{g}^{-1}(u, v)$ can be also expressed in the parametric form $(x, y) = \mathbf{g}^{-1}(u, h(u))$.

We have already met a few such parametric examples earlier in Sec. D.4 (see, for instance, Eqs. (D.5), (D.6) and the footnote thereafter, and Eqs. (D.9) and (D.10)). Note that in Proposition D.1 we have to plug the two components of the (inverse) transformation into the curve's definition, while here we have to plug the two components of the curve into the spatial transformation's definition. ■

¹¹ A curve in the plane can be generally defined in three equivalent forms: *explicitly* by $y = h(x)$, *implicitly* by $f(x, y) = 0$, or *parametrically* by $x = f_1(t)$, $y = f_2(t)$, where the parameter t varies continuously throughout an interval such as $-\infty < t < \infty$ [Bronshtein97 pp. 75–76]. Conversions between these forms can be done as explained in [Bronshtein97 p. 551]. Depending on the case, one or the other of these equivalent forms may have a simpler expression or be more convenient to use; note, however, that not every curve can be expressed in the explicit form.

¹² Note that this is equivalent to the parametric form $(u, v) = \mathbf{g}(f_1(t), f_2(t))$ where $f_1(t) = t$ and $f_2(t) = h(t)$.

D.5 The active and passive interpretations of a transformation¹³

A transformation $(u,v) = \mathbf{g}(x,y)$ can be interpreted in two different ways (see [Courant88 pp. 133–140]): either as a mapping (the active interpretation), or as a coordinate change (the passive interpretation).

According to the *active* interpretation, the one we have tacitly adopted thus far, the transformation \mathbf{g} moves (or maps) each point $P = (x,y)$ of the original x,y plane into its image point $Q = (u,v)$. This destination point can be represented either in the original x,y coordinate system, where the values (u,v) returned by the transformation are understood as new values along the original x and y axes (as in Fig. D.3), or in the target u,v coordinate plane, as in Fig. D.5(b). Note that \mathbf{g} moves any point (x,y) of the original x,y plane (Fig. D.5(a)) to the point having the same coordinates (x,y) in the distorted x,y plane (Fig. D.5(b)). The active point of view is also illustrated in Figs. D.9(a),(b), where the non-linear transformation \mathbf{g} maps the original axes of the x,y plane (a) into the distorted x,y axes in the destination u,v plane (b).

On the other hand, in the *passive* interpretation of the transformation \mathbf{g} we only concentrate on the plane after its deformation by \mathbf{g} has been completed (Fig. D.5(b)), but we consider this plane through two different coordinate nets: the distorted x,y coordinate net and the new undistorted u,v coordinate net.¹⁴ This allows us to interpret the transformation $(u,v) = \mathbf{g}(x,y)$ as a coordinate change in the plane, or, in other words, as a “dictionary” that translates the position of any given point in the plane from the x,y language to the u,v language, without actually moving the given point from one location in the plane to another; see Fig. D.5(b).¹⁵ If $(u,v) = \mathbf{g}(x,y)$ is a one-to-one transformation we can in general assign to each point (x,y) the corresponding values (u,v) as new coordinates, because each pair of values (x,y) uniquely determines the pair (u,v) , and vice versa. Thus, both (x,y) and (u,v) uniquely determine the position of any given point in the plane. The direct transformation $(u,v) = \mathbf{g}(x,y)$ translates from the x,y language to the u,v language, and the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ translates from the u,v language to the x,y language.¹⁶

¹³ The material in this section is only used within the present appendix, but it is not required elsewhere in the book. It is given here for the sake of completeness only, and may be skipped if desired.

¹⁴ It may be helpful to imagine that these coordinate nets are printed on two different transparencies, that can be superposed on top of the same distorted plane. This is clearly illustrated in Fig. 9: Fig. D.9(b) shows the distorted plane superposed by the distorted x,y coordinate net, and Fig. D.9(c) shows the same plane superposed by the undistorted u,v coordinate net.

¹⁵ Note that the passive interpretation of the transformation can be also considered in Fig. D.5(a) which shows the domain of \mathbf{g} , i.e. the situation before the transformation has been applied. However, this may be somewhat less natural since in order to draw the u and v axes in Fig. D.5(a) we need to apply to the x and y axes the *inverse* transformation \mathbf{g}^{-1} .

¹⁶ Note the inherent inversion that exists in the passive interpretation: The coordinates (x,y) of any given point in terms of the x,y coordinate system are translated by the transformation $(u,v) = \mathbf{g}(x,y)$ into the coordinates (u,v) of the same point in terms of the u,v coordinate system; and yet, the u,v coordinate system is obtained from the x,y system by applying the *inverse* transformation, \mathbf{g}^{-1} (see Remark D.3). For example, when \mathbf{g} represents a rotation by angle α (Fig. D.5), the u,v axes are obtained from the x,y axes by a rotation by $-\alpha$.

The active and passive interpretations of a transformation $(u,v) = \mathbf{g}(x,y)$ are concisely summarized as follows:

Proposition D.2: A transformation $(u,v) = \mathbf{g}(x,y)$ can be interpreted in two different ways: either as a mapping which actually moves each point (x,y) into a new location (u,v) within one and the same coordinate system (the *active* interpretation), or as a coordinate change which converts each point from the x,y coordinate system to the u,v coordinate system (the *passive* interpretation).¹⁷ ■

The active and passive interpretations of a transformation $(u,v) = \mathbf{g}(x,y)$ are very closely interrelated. The curves in the u,v plane that are, according to the active interpretation, the images under \mathbf{g} of straight lines parallel to the axes in the x,y plane (see Eqs. (D.7) and (D.8)), can be also regarded according to the passive interpretation as the coordinate curves for the curvilinear coordinates $x = g_1^{-1}(u,v)$, $y = g_2^{-1}(u,v)$ in the u,v plane (see Fig. D.9(b)). Similarly, the curves in the x,y plane that are, according to the active interpretation, the images under \mathbf{g}^{-1} of straight lines parallel to the axes in the u,v plane (see Eqs. (D.11) and (D.12)), can be also regarded as the coordinate curves for the curvilinear coordinates $u = g_1(x,y)$, $v = g_2(x,y)$ in the x,y plane (see Fig. D.9(d)).

Thus, the difference between the two interpretations is mainly in the point of view. If we are mainly interested in the x,y plane, we regard u and v simply as a new means of locating points in the x,y plane, and the u,v plane becomes then merely subsidiary (as in Fig. D.3). But if we are equally interested in the two planes, the x,y plane and the u,v plane, it is preferable to regard the transformation $(u,v) = \mathbf{g}(x,y)$ as specifying a correspondence between the two planes, i.e., as a mapping of one on the other (as in Fig. D.4).¹⁸ It is, however, often desirable to keep the two interpretations in mind at the same time.

Let us consider as a final example the polar to Cartesian transformation (D.13) which is illustrated in Fig. D.6. In this case we say, according to the *active* interpretation, that the transformation maps any point $P = (\theta, r)$ into its image $Q = (u, v) = (r \cos \theta, r \sin \theta)$; for example, the point $P = (\pi/4, 3\sqrt{2})$ is mapped into the point $Q = (3, 3)$. Furthermore, by considering successive points P along the curve $r = 6 \cos \theta$ shown in Fig. D.6(a), we can see that our transformation maps (or distorts) this curve from a cosinusoidal line in the original θ, r plane into a circle $(u-3)^2 + v^2 = 9$ in the transformed θ, r plane (whose new standard coordinate axes are u and v ; see Fig. D.6(b)). What is, then, the *passive* interpretation of the same transformation (D.13)? Consider again point Q in Fig. D.6(b): If we regard (D.13) as a dictionary, we see that without actually moving the point Q , our transformation $(u,v) = (r \cos \theta, r \sin \theta)$ converts the point coordinates from the θ, r language, $(\pi/4, 3\sqrt{2})$, to the u,v language, $(3, 3)$. By considering now the entire circle of Fig. D.6(b),

¹⁷ These two interpretations can be also described as “different objects viewed in the same coordinates” and “the same object viewed in different coordinates”, respectively.

¹⁸ This difference of point of view is often reflected by the notation being used to express the given transformation. For example, the notation $\mathbf{g}(x,y) = (2xy, y^2 - x^2)$, which does not explicitly mention the u,v coordinates, suggests that one is mainly interested in the x,y plane, while the equivalent notation $(u,v) = (2xy, y^2 - x^2)$ suggests that one is interested in both the x,y and u,v planes.

we see that our transformation (D.13) converts the mathematical expression of our circle (without distorting the circle itself!) from the θ, r language, $r = 6\cos\theta$, into the u, v language, $u^2 + v^2 = 6u$, or equivalently $(u-3)^2 + v^2 = 9$ (see the mathematical derivation in Sec. D.4.2). Similar results can be obtained for any other curve or object in the plane.

Remark D.9: It should be noted that there does exist one real difference between the two points of view: $(u, v) = \mathbf{g}(x, y)$ always defines a mapping, no matter how many points (x, y) it maps to one point (u, v) ; but it cannot define a meaningful coordinate change if the correspondence is not one-to-one (such a case is illustrated in Example D.7 below). ■

Remark D.10: To see how all this is related to the well-known theorems on coordinate (or basis) changes in linear algebra, consider the particular case in which $(u, v) = \mathbf{g}(x, y)$ is a linear transformation. Let $\mathbf{e}_1, \mathbf{e}_2$ be the standard basis of the original x, y space: $\mathbf{e}_1 = (1, 0)$, $\mathbf{e}_2 = (0, 1)$, and let $\mathbf{f}_1, \mathbf{f}_2$ be the standard basis of the \mathbf{g} -transformed x, y space (see Fig. D.5(b)), expressed in terms of the basis vectors \mathbf{e}_1 and \mathbf{e}_2 : $\mathbf{f}_1 = (f_{1,1}, f_{1,2})$, $\mathbf{f}_2 = (f_{2,1}, f_{2,2})$, namely:

$$\begin{aligned}\mathbf{f}_1 &= f_{1,1}\mathbf{e}_1 + f_{1,2}\mathbf{e}_2 \\ \mathbf{f}_2 &= f_{2,1}\mathbf{e}_1 + f_{2,2}\mathbf{e}_2\end{aligned}\tag{D.14}$$

Then, according to well known results in linear algebra (see, for example, [Lang87 p. 394–395] or [Lay03 p. 249]), the matrix representation of the transformation \mathbf{g} is given by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_{1,1} & f_{2,1} \\ f_{1,2} & f_{2,2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\tag{D.15}$$

where the components of \mathbf{f}_1 and \mathbf{f}_2 form the *columns* of the matrix. For instance, if $(u, v) = \mathbf{g}(x, y)$ represents rotation by angle α we have: $\mathbf{f}_1 = (\cos\alpha, \sin\alpha)$, $\mathbf{f}_2 = (-\sin\alpha, \cos\alpha)$, and hence the transformation is given by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\tag{D.16}$$

which agrees, indeed, with Eq. (D.1).

In linear algebra books transformation (D.15) is considered as a change of coordinates from the basis $\mathbf{f}_1, \mathbf{f}_2$ to the standard basis $\mathbf{e}_1, \mathbf{e}_2$ (see, for example, [Lay03 p. 249]): Given a point (x, y) in terms of the basis $\mathbf{f}_1, \mathbf{f}_2$ of the transformed x, y space, \mathbf{g} returns the coordinates (u, v) of the same point in terms of the standard basis $\mathbf{e}_1, \mathbf{e}_2$. Restated in our terms, we can say that transformation (D.15) is considered in linear algebra as a “dictionary” that translates coordinates of any given point in terms of the transformed x, y space into the coordinates of the same point in terms of the u, v space (see Fig. D.5(b)). For example, in the case of rotation, the point $(1, 0)$ in terms of the rotated x, y coordinates is converted by the transformation (D.16) into $(\cos\alpha, \sin\alpha)$, which specifies the coordinates of the same point in terms of the u, v space. This corresponds, indeed, to the *passive* interpretation of the linear transformation $(u, v) = \mathbf{g}(x, y)$.

On the other hand, it is also possible to consider the *active* interpretation of the same linear transformation: According to this point of view, the transformation moves (or maps) any point $x\mathbf{e}_1 + y\mathbf{e}_2$ in the original plane to its image point $x\mathbf{f}_1 + y\mathbf{f}_2$ in the same plane (see Fig. D.3) or, equivalently, to the point (u,v) in terms of the target plane (see Fig. D.5). Note that just as in the general case, a linear transformation \mathbf{g} moves any point (x,y) given in the original x,y coordinate system to the point having the same coordinates (x,y) in the transformed x,y coordinate system. More details on active and passive linear transformations in a given vector space can be also found in [Wolf79 Sec. 1.3].¹⁹ ■

Remark D.11: Note that the matrix in Eq. (D.15) is called in some references “the transition matrix from the old basis $\mathbf{e}_1, \mathbf{e}_2$ to the new basis $\mathbf{f}_1, \mathbf{f}_2$ ” (see, for example, [Lipschutz68 p. 153] and the remark following Theorem 7.4 there), while in other references it is called “the change-of-coordinates matrix from the basis $\mathbf{f}_1, \mathbf{f}_2$ to the basis $\mathbf{e}_1, \mathbf{e}_2$ ” (see, for example, [Lay03 pp. 249, 273]). The reason for this terminological inconsistency is that the matrix in question contains the coefficients f_{ij} that are used to convert the old basis vectors $\mathbf{e}_1, \mathbf{e}_2$ into the new basis vectors $\mathbf{f}_1, \mathbf{f}_2$ (see Eq. (D.14)); and yet, a multiplication by this matrix, as shown in Eq. (D.15), converts the coordinates (x,y) given in terms of the new basis $\mathbf{f}_1, \mathbf{f}_2$ back into coordinates (u,v) given in terms of the old basis $\mathbf{e}_1, \mathbf{e}_2$. ■

Remark D.12: Another possible source of confusion exists due to terminological inconsistencies in the literature regarding the active and passive interpretations of a transformation. In some references such as [Harris98 pp. 351–353] the active transformation has the same meaning as in our definition, i.e. moving a point from its original location P to its image location Q within the same coordinate system (like in Fig. D.3); but the passive transformation means moving the *coordinate axes* from their original location to their image location under \mathbf{g} while the point P itself does not move. In this case, the coordinates of the point P in terms of the new coordinate axes are given by the inverse transformation \mathbf{g}^{-1} . For example, if the active transformation \mathbf{g} consists of displacing any point (x,y) to the point $(x+a, y+b)$:

$$\begin{aligned} u &= x + a \\ v &= y + b \end{aligned} \tag{D.17}$$

then the passive transformation, according to this interpretation, consists of displacing the coordinate system by (a,b) , while the point (x,y) itself remains in its original location. The new coordinates of the point (x,y) are expressed, therefore, by the *inverse* transformation:

¹⁹ It is interesting to note that the active interpretation of linear transformations is rarely mentioned in linear algebra books (with [Mansfield76 pp. 202–206] being a remarkable exception), while in calculus books the active interpretation is used frequently (see, for example, [Colley98 pp. 326–327, 334–335]). The reason is that calculus begins with a single coordinate system, while linear algebra is at the other extreme: its main concern is in considering all possible bases (or linear coordinate systems) for a given vector space, and then choosing the one which is most convenient depending on the nature of the problem at hand [Mansfield76 p. 60], for example the one which gives the simplest matrix representation for a given linear transformation.

$$\begin{aligned} u &= x - a \\ v &= y - b \end{aligned} \tag{D.18}$$

[Harris98 p. 351]. This does not agree with our terminology, based on [Courant88 pp. 133–140], according to which the active and the passive points of view are merely two different interpretations of the *same* transformation $(u,v) = \mathbf{g}(x,y)$. For example, according to *our* passive interpretation the transformation (D.17) simply translates the transformed x,y coordinates of any given point back to the u,v coordinates of the same point: $u = x + a$, $v = y + b$, and no use is made of the inverse transformation (D.18).

The difference between these two views of the passive interpretation of \mathbf{g} can be illustrated by splitting Fig. D.5(b) into two different figures, one showing the point Q superposed by the standard u,v coordinates and the other showing the same point Q superposed by the transformed x,y coordinates. According to *our* passive point of view, both figures show the range of the transformation \mathbf{g} after its action has already been completed. But according to the other point of view, the first figure shows the initial situation, and the other figure shows how the transformation \mathbf{g} actively distorts the original coordinate system, while the point Q remains in its original location. The coordinates of Q with respect to the new distorted coordinate system are obtained, therefore, by applying the inverse transformation \mathbf{g}^{-1} to the coordinates of Q in the undistorted coordinate system. ■

Remark D.13: Although the terms “mapping” and “transformation” are synonyms, they are often used in slightly different connotations. The term “mapping” usually implies the active interpretation, where points P of one space are mapped to corresponding points Q in another space. On the other hand, the term “transformation” is often utilized when the target space is the same as the source space, and it refers to a deformation or a rearrangement of that space. ■

D.6 Domain and range transformations of a function

So far we have considered the influence of transformations $(u,v) = \mathbf{g}(x,y)$ on the plane or on its 2D or 1D subsets. We now proceed to the influence of such transformations on *functions* that are defined on the plane. This will bring us to the important notions of domain and range transformations, which have a central role in our work on the moiré theory.

Let $z = f(u,v)$ be a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$. As we already know from Sec. D.2, such a function can be interpreted geometrically as a surface over the u,v plane. Since a function is always defined between two spaces, its *domain* and its *range*,²⁰ each function $z = f(u,v)$

²⁰ The domain of the function $z = f(u,v)$ is the 2D set consisting of the points (u,v) on which the function is defined, and the range of the function is the 1D set consisting of the values z it returns. The reason we use here the domain variables u,v rather than x,y is explained in Footnote 24 below.

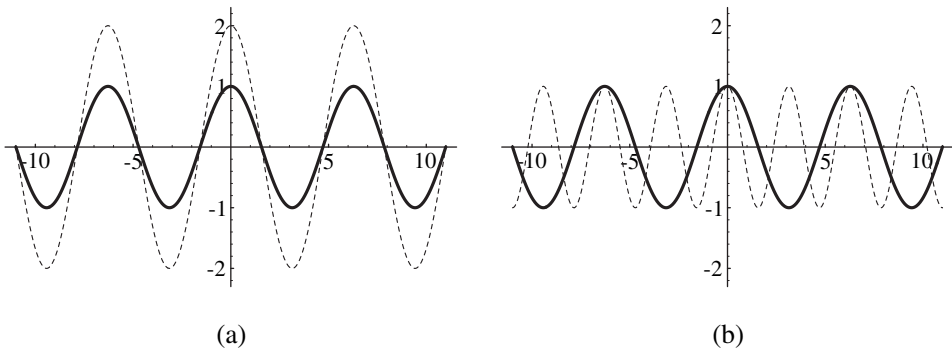


Figure D.7: (a) The influence of the range transformation $s = 2z$ on the function $z = \cos u$. The dashed curve represents the resulting function, $s = 2\cos u$. (b) The influence of the domain transformation $u = 2x$ on the same function $z = \cos u$. The dashed curve represents the resulting function, $z = \cos(2x)$. In both (a) and (b) the transformed and untransformed functions are plotted on the same axes, in order to allow the comparison between them.

can be distorted in two different ways: either by transforming its domain, or by transforming its range. We can distinguish, therefore, between domain and range transformations as follows:²¹

- (1) A *domain transformation* of the function $z = f(u, v)$ is a transformation $(u, v) = \mathbf{g}(x, y)$, $\mathbf{g}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$, which is applied to the domain of $f(u, v)$. Following this operation the original function $f(u, v)$ is transformed into the new function $f(\mathbf{g}(x, y))$.
- (2) A *range transformation* of the same function $z = f(u, v)$ is a transformation $s = t(z)$, $t: \mathbb{R} \rightarrow \mathbb{R}$, which is applied to the range of $f(u, v)$ (i.e. to the values z it returns). Following this operation the original function $f(u, v)$ is transformed into the new function $t(f(u, v))$.

Geometrically speaking, a domain transformation distorts the surface $f(u, v)$ spatially (for example: by rotation, translation, scaling along the u, v directions, etc.). In other words, it moves each point (u, v) to a new location $(x, y) = \mathbf{g}^{-1}(u, v)$, without affecting the z coordinate assigned to the point. On the other hand, a range transformation distorts the surface $f(u, v)$ vertically (for example: by scaling it by 2 in the z direction, by taking the cosine of the altitude z , etc.). In other words, it changes the z coordinate assigned to each point into $s = t(z)$, without affecting the point's u, v coordinates (see Fig. D.7(a)).

It is important to note that domain and range transformations are not different types of transformations, but rather different uses or applications of a transformation. Thus, in

²¹ Although these definitions are given here for the case of a function $z = f(u, v)$, they are, in fact, completely general, and apply to any other types of functions, including $z = f(u)$, $(x, y) = \mathbf{g}(u)$, etc. (see Sec. D.2), with the appropriate adaptations to the dimensionalities of their domain and range.

cases where the domain and the range of f have the same dimensionality (like in the 1D case $z = f(u)$ that we discuss in Sec. D.6.1 below) the very same transformation can be applied either to the domain of f or to the range of f . In the first case it will play the role of a domain transformation of the function f , while in the second case it will play the role of a range transformation of f .

Remark D.14: As we have already seen, transformations can be applied not only to functions defined over the plane, but also directly to the plane itself or to objects that are subsets of the plane. For example, if $z = f(u, v)$ is a function (surface) over the u, v plane, $f(u, v) = 0$ defines a curve in the u, v plane (which corresponds to the zero level line of the surface $z = f(u, v)$). But unlike the surface $z = f(u, v)$, which has both a domain and a range, the planar curve $f(u, v) = 0$ only has a domain (since its range is reduced to the degenerate space $\{0\}$), and therefore it can only undergo *domain* transformations. ■

While the effect of a range transformation on the original function f is rather straightforward, the effect of a domain transformation on the function f may be less obvious and sometimes even quite confusing. It is therefore our aim here to help in developing an intuitive understanding of range and domain transformations and to point out the main pitfalls in their use.

D.6.1 The 1D case

In order to better understand the situation, let us start with the simpler, 1D case. Suppose that we are given a function $z = f(u)$, for example $z = \cos u$. Note that in this case both the range and the domain of the function f are one-dimensional, so that any function $g: \mathbb{R} \rightarrow \mathbb{R}$ may be applied to f either as a range transformation, giving $g(f(u))$, or as a domain transformation, giving $f(g(x))$.

We start with the case of a range transformation. Suppose that the transformation $s = t(z)$ is applied to our function $z = f(u)$ as a range transformation. The effect of this transformation on $f(u)$ is straightforward: for example, if we apply to $f(u)$ the range transformation $s = 2z$ we obtain a vertically stretched version of $f(u)$, namely, $2f(u)$. This transformation maps each value z on the vertical axis to $2z$, without affecting the u coordinate; this can be clearly seen by plotting the two functions on top of each other (see Fig. D.7(a)).²²

Now, suppose that instead of the range transformation $s = t(z)$ we apply to our function $z = f(u)$ a domain transformation $u = g(x)$; for example, we may choose once again the same two-fold magnification transformation, $u = 2x$. Clearly, this transformation maps each value x on the horizontal axis to the value $2x$, without affecting the z coordinate. Based on our experience with range transformations it could be natural, therefore, to expect that the application of this transformation would stretch our function $z = f(u)$ laterally by a factor of 2. However, in reality $f(2x)$ is not a stretched version of $f(u)$, but

²² Note that in order to compare the two functions we must plot both of them on the same axes. Therefore, in Fig. D.7(a) the vertical axis represents both of the variables z and s (see Remark D.1).

rather a condensed version of $f(u)$ which has been squeezed laterally by a factor of 2. For example, $z = \cos(2x)$ is a laterally condensed version of $z = \cos u$; this can be clearly seen by plotting the two functions on top of each other (see Fig. D.7(b)).²³

This difference between the influence of range and domain transformations can be explained as follows. The application of a domain transformation $u = g(x)$ to the given function $z = f(u)$ gives $z = f(g(x))$, i.e. it moves each point u in the domain of the original function $z = f(u)$ into the new location x which is determined by the *inverse* of the transformation g , namely, $x = g^{-1}(u)$. But when we apply to the given function $z = f(u)$ a range transformation $s = t(z)$, it simply modifies the z coordinate assigned to each point into the new value $s = t(z)$, and no inversion is involved. This is schematically illustrated by the commutative diagram shown in Fig. D.8(a). The original function $z = f(u)$ is represented in this figure by the top horizontal arrow. After applying to this function both a domain transformation $u = g(x)$ and a range transformation $s = t(z)$ we obtain the resulting function $s = f_r(x)$, whose variables are x and s rather than u and z ; this function is represented in our figure by the bottom horizontal arrow. The new function f_r can be expressed in terms of the known functions g, f and t by:

$$f_r(x) = t(f(g(x))) \quad (\text{D.19})$$

as shown by the circular arrow in the figure. (Note the order inversion in (D.19): although t appears first in the equation, in reality it operates last, after g and f .) In particular, if g is the identity transformation, f_r is simply a range transformation of f ; while if t is the identity transformation, f_r is a domain transformation of f : $f_r(x) = f(g(x))$.

Fig. D.8(a) illustrates the fundamental difference which exists between the two transformations that we have applied to our original function $z = f(u)$ to obtain the new function $s = f_r(x)$: While the range transformation $s = t(z)$ maps the variable z of our original function into the transformed variable s of the new function $s = f_r(x)$, the domain transformation $u = g(x)$ maps the variable x (the variable of the new, distorted function $s = f_r(x)$) back into the variable u (the variable of our original, undistorted function $z = f(u)$). The transformation from the original variable u into the new, distorted variable x is given, therefore, by the *inverse* transformation, $x = g^{-1}(u)$, and not by the transformation $u = g(x)$ that we actually plug into the given function $f(u)$. And indeed, by drawing the functions $f(g(x))$ and $f(u)$ on the same coordinate system we can see that the influence of the domain transformation $u = g(x)$ on $f(u)$ is inverse. For example, $f(2x)$ is a *squeezed* version of $f(u)$; $f(x+1)$ is a unit translation of $f(u)$ to the *negative* direction; etc. If we wish to stretch our function $f(u)$ laterally by a factor of 2 within the original coordinate system, we need to take $f(x/2)$ and not $f(2x)$.

²³ Once again, in order to compare the two functions we must plot both of them on the same axes. Therefore, in Fig. D.7(b) the horizontal axis represents both of the variables u and x (see Remark D.1). In fact, we could say that “the function $z = f(2u)$ is a laterally condensed version of $z = f(u)$ ”; but in order to avoid any confusion due to the use of the original variable u in the transformed function, too, it would be better to use instead a more “neutral” variable name, such as w , and say that “the function $z = f(2w)$ is a laterally condensed version of $z = f(w)$ ”. In this case the horizontal axis would be simply named w .

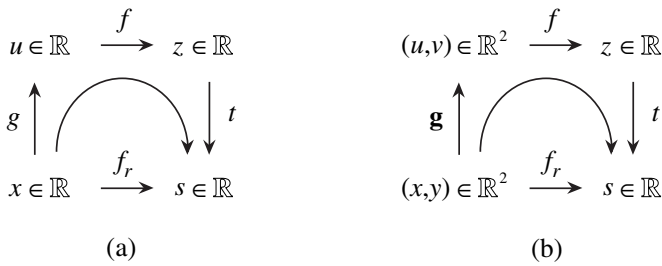


Figure D.8: (a) Commutative diagram illustrating the influence of a domain transformation $u = g(x)$ and a range transformation $s = t(z)$ on a given 1D function $z = f(u)$. (b) Commutative diagram illustrating the influence of a domain transformation $(u,v) = \mathbf{g}(x,y)$ and a range transformation $s = t(z)$ on a given 2D function $z = f(u,v)$. Note that while the range transformation t maps the original variable z of our function f to the new, distorted variable s , the domain transformation maps the new, distorted variables x,y back to the original variables u,v of our function f .

Remark D.15: Note that any expression of the form $f(g(x))$ can be interpreted in two different ways: either as the application of g (as a domain transformation) to the given function f , or as the application of f (as a range transformation) to the given function g . Although both interpretations give, of course, the same result, in each situation one or the other may be easier to understand. The same is also true for expressions of the form $f(g(x,y))$ (where $f(u)$ is still a 1D function): For example, the expression $\cos(2\pi\sqrt{x^2 + y^2})$ can be understood either as the result of bending the original straight cosinusoidal surface $\cos(2\pi u)$ into a circular cosinusoidal surface, or as the result of applying the function $\cos(2\pi z)$ to the z values (altitude) of the original conic surface $z = \sqrt{x^2 + y^2}$. ■

D.6.2 The 2D case

The same considerations hold also in the 2D case. Suppose that we are given a function $z = f(u,v)$. Just as in the 1D case, the influence of a range transformation $s = t(z)$ on this function is straightforward: it simply maps each z value (altitude) of $z = f(u,v)$ to the new value $t(z)$, without affecting the u and v coordinates. Thus, if we apply to $f(x,y)$ the range transformation $s = 2z$, we obtain a vertically stretched version of $f(u,v)$, namely, $2f(u,v)$.

Now, let us proceed to the influence of a domain transformation $(u,v) = \mathbf{g}(x,y)$ on our function $z = f(u,v)$. As a simple example we may consider the domain transformation $(u,v) = (2x, 2y)$. Clearly, this transformation maps each point (a,b) of the plane to the point $(2a, 2b)$. But in spite of this stretching effect, it turns out that plugging this transformation into the function $z = f(u,v)$ gives $z = f(2x, 2y)$, which is not a spatially stretched version of $f(u,v)$ but rather a spatially condensed version of $f(u,v)$. The reason for this inversion is that the application of a domain transformation $(u,v) = \mathbf{g}(x,y)$ to the given function $z = f(u,v)$ gives $z = f(\mathbf{g}(x,y))$, i.e. it moves each point (u,v) in the domain of the original function

$z = f(u,v)$ into the new location (x,y) which is determined by the *inverse* of the transformation \mathbf{g} , namely, $(x,y) = \mathbf{g}^{-1}(u,v)$. This is illustrated by the commutative diagram shown in Fig. D.8(b). The original function $z = f(u,v)$ is represented here by the top horizontal arrow, and the resulting function f_r obtained by applying to f both a domain transformation \mathbf{g} and a range transformation t is represented by the bottom horizontal arrow. But the effect of f_r can be also expressed, by following the circular arrow, as:

$$f_r(x,y) = t(f(\mathbf{g}(x,y))) \quad (\text{D.20})$$

Clearly, if \mathbf{g} is the identity transformation, f_r is simply a range transformation of the function f ; and similarly, if t is the identity transformation, f_r is a domain transformation of the function f : $f_r(x,y) = f(\mathbf{g}(x,y))$.

Fig. D.8(b) shows that just as in the 1D case, the influence of the transformation $(u,v) = \mathbf{g}(x,y)$ when it is applied as a domain transformation to our function $f(u,v)$ is, in fact, inverse: it maps the transformed variables x,y of the new function $s = f_r(x,y)$ back into the untransformed variables u,v of the original function $z = f(u,v)$. Hence, the transition from the original u,v space of $f(u,v)$ into the new, distorted x,y space of $f(\mathbf{g}(x,y))$ is represented by the *inverse* transformation, \mathbf{g}^{-1} , although we have plugged into f the transformation \mathbf{g} itself and not its inverse \mathbf{g}^{-1} .

To better illustrate this, let us consider again the planar transformation $(u,v) = \mathbf{g}(x,y) = (2x,2y)$, which corresponds to a two-fold expansion of the x,y plane. When we apply \mathbf{g} as a domain transformation to the function $f(u,v) = u^2 + v^2 - 1$ (a top opened paraboloid), we clearly obtain the inverse effect since $f(2x,2y) = (2x)^2 + (2y)^2 - 1$ is a spatially shrunk version of the original function $f(u,v)$. The expansion effect is obtained by applying to $f(u,v)$ the domain transformation $(u,v) = \mathbf{g}^{-1}(x,y) = (x/2,y/2)$ (see Remark D.4 on the variable names).

Note that the same rule applies also to any subsets of the function $z = f(u,v)$, for example to the curve defined by its zero level line $f(u,v) = 0$. Thus, proceeding with the same example, if we apply our two-fold magnification $(u,v) = \mathbf{g}(x,y) = (2x,2y)$ as a domain transformation to the circle $f(u,v) = u^2 + v^2 - 1 = 0$, we obtain:

$$(2x)^2 + (2y)^2 - 1 = 0$$

namely:

$$x^2 + y^2 = (\tfrac{1}{2})^2$$

which is clearly a two-fold reduced circle. Note, however, that when we consider our transformation $(u,v) = \mathbf{g}(x,y) = (2x,2y)$ as a direct transformation that operates on the original, undistorted x,y plane, it indeed magnifies the entire x,y plane, including the circle $x^2 + y^2 = (\tfrac{1}{2})^2$ that is embedded in it, into the target u,v plane and its circle $u^2 + v^2 = 1$. (Remember that according to Proposition D.1 the image of the curve $f(x,y) = 0$ under the direct transformation $(u,v) = \mathbf{g}(x,y)$ is given by the implicit equation $f(\mathbf{g}_1^{-1}(u,v), \mathbf{g}_2^{-1}(u,v)) = 0$. In our case, plugging $x = u/2$ and $y = v/2$ into the circle's equation $x^2 + y^2 = (\tfrac{1}{2})^2$ gives, indeed, $u^2 + v^2 = 1$.)

Proposition D.3: Suppose we are given a transformation $(u,v) = \mathbf{g}(x,y)$. When this transformation is applied to the original, undistorted x,y plane (or any subset thereof) as a *direct* transformation, we obtain in the target u,v plane the \mathbf{g} -transformed copy of the x,y plane (or of its subset). For example, the transformation $(u,v) = (2x,2y)$ gives us in the target u,v plane a two-fold magnified version of the original x,y plane (or any subset thereof). But when the same transformation is applied to the function $z = f(u,v)$ (or any subset thereof, such as the level line $f(u,v) = \text{const.}$) as a *domain* transformation, we obtain a \mathbf{g}^{-1} -transformed copy of the function f (or of its subset). For example, when our transformation $(u,v) = (2x,2y)$ is applied as a domain transformation to the function $z = f(u,v)$, we obtain a spatially condensed version of this function, $z = f(2x,2y)$. ■

Note that although the transformation $(u,v) = \mathbf{g}(x,y)$ acts on the original, undistorted x,y space and yields its distorted image in the target u,v space, when \mathbf{g} is applied as a domain transformation to a function $z = f(u,v)$ the roles are inverted: The original undistorted function f is given in the u,v coordinate system, while the domain-transformed function $f(\mathbf{g}(x,y))$ subsists over the x,y space.²⁴

As a further example, let $(u,v) = \mathbf{g}(x,y)$ be the planar transformation which corresponds to a rotation by 90° counterclockwise about the origin:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}$$

Clearly, this transformation rotates each point (x,y) of the plane by 90° counterclockwise (for example, it maps the point $(1,0)$ on the horizontal axis to the point $(0,1)$ on the vertical axis, etc.). Thus, the vector (u,v) is a copy rotated by $+90^\circ$ of the vector (x,y) . However, for any given function $f(u,v)$, its transformed counterpart $f(\mathbf{g}(x,y))$, that is, $f(-y,x)$, is a -90° rotated version of $f(u,v)$. This can be easily verified by plotting the original and the transformed functions.

More generally, the transformation $(u,v) = \mathbf{g}(x,y)$ consisting of a rotation by angle α counterclockwise about the origin is defined by Eq. (D.1):

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x\cos\alpha - y\sin\alpha \\ x\sin\alpha + y\cos\alpha \end{pmatrix}$$

but the function $f(x\cos\alpha - y\sin\alpha, x\sin\alpha + y\cos\alpha)$ is a rotated version of $f(u,v)$ by $-\alpha$, not by α . Thus, if we wish to rotate $f(u,v)$ by angle α within the original coordinate system, the transformation we need to apply to its domain is the *inverse* transformation which corresponds to a rotation by $-\alpha$. The rotated version of $f(u,v)$ by angle α is, therefore, $f(x\cos\alpha + y\sin\alpha, -x\sin\alpha + y\cos\alpha)$.

²⁴ This agrees with our general convention in the moiré theory, that the geometrically transformed layers subsist in the x,y space (see, for example, the transformed gratings and screens in Chapters 3, 6 and 7 of this volume or in Chapter 10 of *Vol. I*; for instance, the top-opened parabolic cosinusoidal grating of Fig. 10.1(c) in *Vol. I* is expressed by $r(x,y) = \cos(2\pi f[y - ax^2])$). Note, however, that in our discussions on the moiré theory we usually denote the original, undistorted coordinate space by the variables x',y' rather than u and v , since u and v are reserved in our work to the Fourier, spectral domain.

Similarly, applying the direct transformation (D.1) to any curve $f(x,y) = 0$ in the x,y plane gives in the u,v plane a copy rotated by α of the original curve (see Proposition D.1). But applying (D.1) as a domain transformation to the curve $f(u,v) = 0$ rotates this curve by $-\alpha$. Note that in the first case the untransformed curve is $f(x,y) = 0$, while in the second case the untransformed curve is given by $f(u,v) = 0$.

Obviously, if we wish to leave $f(u,v)$ unchanged and to consider the influence of $\mathbf{g}(x,y)$ on the *coordinate system* and on its axes, the result will be inversed. For example, in the case of rotation we will obtain a rotation of the u,v axes by angle α to the positive direction [Weisstein99 p. 1580]. Similarly, if $\mathbf{g}(x,y)$ is defined by $\mathbf{g}(x,y) = (2x,2y)$ the result can be seen as a twofold expansion of the axes while the surface $f(u,v)$ itself remains unchanged; and if $\mathbf{g}(x,y) = (x+1,y)$ the result can be seen as a unit translation of the horizontal axis to the *positive* direction. This point will be addressed in Sec. D.7 below.

Remark D.16: We may mention at this point yet another possible source of confusion due to terminological inconsistencies in the literature, this time related to the naming conventions in domain transformations. A transformation $(u,v) = \mathbf{g}(x,y)$ is usually called in the literature a transformation from the x,y space to the u,v space, because it translates the coordinates of any given point in the plane from the x,y language to the u,v language. For instance (see Example D.4 below), the transformation $(r,\theta) = (\sqrt{x^2+y^2}, \arctan(y/x))$ is known as the *Cartesian to polar* coordinate transformation, and its inverse, $(x,y) = (r\cos\theta, r\sin\theta)$, is called the *polar to Cartesian* coordinate transformation [Colley98 p. 68]. Note, however, that in some references the naming conventions are inversed, and the transformation $(u,v) = \mathbf{g}(x,y)$ is considered as a mapping from the u,v space to the x,y space (see, for example, [Spiegel63 pp. 108, 124, 182]). This naming convention can be explained by the fact that if we are given a function $z = f(u,v)$ in the u,v space, and we apply to it $(u,v) = \mathbf{g}(x,y)$ as a domain transformation, we obtain the function $z = f(\mathbf{g}(x,y))$ in the x,y space. This situation occurs, for example, when changing variables under a double integral in order to facilitate its calculation (see a few such examples in [Colley98 pp. 336–338]). Thus, the transformation $(u,v) = \mathbf{g}(x,y)$ translates the function f from the u,v language to the x,y language, although it actually converts the x,y space into the u,v space.²⁵ This is, again, a consequence of the inversion property inherent to domain transformations.²⁶ ■

Remark D.17: The extension of Remark D.15 to the case of a 2D function f is rather straightforward. Thus, any expression of the form $f(\mathbf{g}(x,y))$ can be interpreted either as the application of $(u,v) = \mathbf{g}(x,y)$ (as a domain transformation) to the given function $z = f(u,v)$, or as the application of $z = f(u,v)$ (as a range transformation) to $(u,v) = \mathbf{g}(x,y)$. Note, however, that in this 2D case the second interpretation is less useful, since it maps the 2D range of \mathbf{g} into a new 1D range. And yet, both interpretations are completely equivalent and give the same result. ■

²⁵ In computer graphics (image warping) it is often said that \mathbf{g} transforms the image f from the x,y transformed, destination space back into the u,v original, undistorted space.

²⁶ A useful trick for avoiding confusion (or lengthy explanations) in the case of polar to Cartesian or Cartesian to polar coordinate transformations is to use the ambiguous term “polar coordinate transformation”.

D.6.3 The effect of transformation g on objects and on their characteristic functions

We conclude this section with the following result, which illustrates from a slightly different angle Propositions D.1 and D.3 and the potentially confusing nature of domain transformations.

Suppose we are given a point P in the x,y plane, say, the point $P = (1,0)$ on the x axis. The *characteristic function* of this point (i.e. the function that takes the value 1 at this point and remains zero everywhere else) is defined over the x,y plane by:

$$f(x,y) = \begin{cases} 1 & (x,y) = (1,0) \\ 0 & \text{otherwise} \end{cases} \quad (\text{D.21})$$

Now, suppose that we apply to the x,y plane a 2D transformation, say, a rotation by 90° counterclockwise:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{g} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

As a result of this transformation our point P is rotated to a new location, the point $(0,1)$ on the vertical axis:

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

However, if we wish to express the very same result using the *characteristic function* of the point P , Eq. (D.21), we need to apply to it the *inverse* transformation, that is given by:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{g}^{-1} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} v \\ -u \end{pmatrix}$$

By applying \mathbf{g}^{-1} to $f(x,y)$ as a domain transformation we obtain:

$$\begin{aligned} h(u,v) &= f(\mathbf{g}^{-1}(u,v)) = f(v,-u) = \\ &= \begin{cases} 1 & (v,-u) = (1,0) \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 1 & (u,v) = (0,1) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

This is, indeed, the characteristic function in the u,v plane of our rotated point, $(0,1)$.

Since this result holds for any given point, it obviously remains true for any object on the plane. For example, suppose we draw the standard unit grid on the x,y plane. This unit grid can be represented by its characteristic function which is defined by:

$$f(x,y) = \begin{cases} 1 & x \in \mathbb{Z} \text{ or } y \in \mathbb{Z} \\ 0 & \text{otherwise} \end{cases} \quad (\text{D.22})$$

If we plot this function, we get a white grid on black background. (We could also interchange the 1 and 0 values in this definition in order to obtain a black grid on white background, in order to be consistent with our figures.) This function is, indeed, the mathematical representation of our unit grid. Now, suppose that we apply to the x,y plane a 2D transformation $\mathbf{g}(x,y)$, say, the transformation defined by $(u,v) = (2x,2y)$. As a result of this transformation the unit grid undergoes a two fold magnification, as shown in Figs. D.10(a),(b). However, if we wish to express the resulting distorted grid using its definition by Eq. (D.22), the domain transformation we have to apply to the characteristic function $f(x,y)$ is the *inverse* transformation $\mathbf{g}^{-1}(u,v)$. In other words:

Proposition D.4: If the characteristic function of an object in the x,y space is given by $f(x,y)$, then the characteristic function of the same object after it has been distorted by the transformation $\mathbf{g}(x,y)$ is given by $f(\mathbf{g}^{-1}(u,v))$. Thus, although the object itself is distorted by the original transformation $\mathbf{g}(x,y)$, its mathematical expression, $f(x,y)$, is affected by \mathbf{g}^{-1} , not by \mathbf{g} . ■

D.7 The relative point of view: object deformations vs. coordinate deformations²⁷

It may be sometimes useful to focus our attention on an object $z = f(u,v)$ which is defined on the u,v plane, or any subset thereof, such as $f(u,v) = 0$. (The 1D equivalent would be the object $z = f(u)$ which is defined on u .) We may then think of the range transformation $s = t(z)$ and of the domain transformation $(u,v) = \mathbf{g}(x,y)$ (or $u = g(x)$, in the 1D case) in three different ways:²⁸

- (1) We can consider them as mappings which affect the entire space (2D plane or 1D line), including the object and the coordinate system.
- (2) Alternatively, we can consider them as mappings which only affect the object itself, without influencing the underlying coordinate system.
- (3) Finally, we can also consider them as transformations which only affect the coordinate system, while our object itself (say, a rigid physical object) remains unchanged.

For example, in the 1D case we may consider the curve $z = \cos u$ either (1) as a subset of the z,u plane that undergoes the same domain or range transformation as the entire plane; (2) as a free, flexible curve which undergoes the distortions defined by $u = g(x)$ and $s = t(z)$ within the original, unchanged coordinate system (see Fig. D.7); or (3) as a fixed, rigid curve which remains unchanged while the transformations $u = g(x)$ and $s = t(z)$ are applied to the coordinate system. In the more interesting 2D case, we may consider $z = f(u,v)$ either (1) as an object in space which undergoes the same transformation as the

²⁷ The material in this section is only used within the present appendix, but it is not required elsewhere in the book. It is given here for the sake of completeness only, and may be skipped if desired.

²⁸ As we saw in Remark D.14, if the object in question is a planar curve or any other subset of the plane, it can only undergo *domain* transformations since its range is reduced to the degenerate space $\{0\}$.

entire space; (2) as a free, flexible object which undergoes the distortions defined by $(u,v) = \mathbf{g}(x,y)$ (rotation, spatial scaling, etc.) and by $s = t(z)$ (vertical scaling) within the original, unchanged coordinate system; or (3) as a fixed, rigid object which remains unchanged while the transformations $(u,v) = \mathbf{g}(x,y)$ and $s = t(z)$ are applied to the coordinate system. Note that in all of these cases the transformations in question are viewed as active transformations, since they distort the given object, the coordinate system, or both.

Clearly, point of view (3) implies that the new coordinates of our given object, after the transformation has been applied, are obtained by the *inverse* transformation \mathbf{g}^{-1} . For example, if the transformation \mathbf{g} consists of rotation by angle α , the application of \mathbf{g} to the axes implies that our object's coordinates with respect to the new axes are obtained by a rotation by $-\alpha$.

The inversion effect due to the relative point of view (object or coordinate distortion) also occurs in range transformations. For instance, the effect of the range transformation $s = 2z$ is inverted between points of view (2) and (3) (our object becomes larger or smaller with respect to the vertical axis).

It is important to note, however, that the inversion effect due to the relative point of view *does not* account for the fundamental inversion effect which is inherent to domain transformations (Sec. D.6). Indeed, as we have just seen, the inversion effect due to the relative point of view affects both domain and range transformations in the same way.

All of the three points of view (1)–(3) are used in the literature; for example, [Courant88 p. 135] uses the first convention, while [Cantwell02 p. 14] uses the second convention. The first convention is used, for example, when plotting data on a logarithmic paper. The third convention is mainly used in the case of linear or affine transformations, such as rotations, scaling and translations. In the case of non-linear transformations, keeping the object unchanged while the coordinate system is being distorted may seem rather unusual; and yet, this is routinely done, for example, when one wishes to consider a given physical object in terms of polar rather than Cartesian coordinates.

More details on object transformations, coordinate changes and the conversions between them, along with several illustrated examples (for linear and affine cases only), can be found in [Foley90 Sec. 5.8].

D.8 Examples

Let us consider now a few examples to better illustrate our discussion. Most of these examples show simple transformations $(u,v) = \mathbf{g}(x,y)$ that are well known and widely used. And yet, it turns out that some of these transformations are mainly known in the literature through one particular representation (which may differ from case to case), while their other representations often remain unfamiliar and sometimes even quite surprising. But because in the present book we often need to consider a given transformation through

several of its different representations, we must get acquainted with all of them, even those that are rarely if ever mentioned in the literature. These hidden facets of our transformations are revealed here through a systematic graphical comparison, which is obtained by showing each of the transformations in its different representations, keeping in all cases the same structural framework as in the model shown in Fig. D.9.

Fig. D.9 illustrates a general, arbitrary transformation $(u,v) = \mathbf{g}(x,y)$. We have intentionally chosen for this purpose a non-linear reminiscent of the rotation by angle α , in order to facilitate the understanding of the general case based on the intuition we have acquired in the case of pure rotation. Fig. D.9(a) shows the original x,y space before the application of the direct transformation \mathbf{g} , and Fig. D.9(b) shows the target u,v space and the distorted x,y space after the application of this transformation. Similarly, Figs. D.9(c) and (d) show the effect of the inverse transformation, $(x,y) = \mathbf{g}^{-1}(u,v)$, which corresponds to a “non-linear rotation” by angle $-\alpha$. Fig. D.9 serves us as a model, and all of the figures throughout this section are constructed according to this model.

The following list contains the main questions that can be answered by each of these figures (followed by the parts of the figure that illustrate the answer, with an arrow indicating the direction of the effect, when applicable):²⁹

1. What is the effect of \mathbf{g} on the standard Cartesian grid? (a) \rightarrow (b)
2. What is the effect of \mathbf{g}^{-1} on the standard Cartesian grid? (c) \rightarrow (d)
3. To what curves $x = g_1^{-1}(u,v) = m$ and $y = g_2^{-1}(u,v) = n$ in the u,v plane does \mathbf{g} map the straight lines $x = m$ and $y = n$ of the x,y plane? (a) \rightarrow (b)
4. How does \mathbf{g}^{-1} map these curves in the u,v plane back into the original straight lines $x = m$ and $y = n$ of the x,y plane? (a) \leftarrow (b)
5. To what curves $u = g_1(x,y) = m$ and $v = g_2(x,y) = n$ in the x,y plane does \mathbf{g}^{-1} map the straight lines $u = m$ and $v = n$ of the u,v plane? (c) \rightarrow (d)
6. How does \mathbf{g} map these curves in the x,y plane back into the original straight lines $u = m$ and $v = n$ of the u,v plane? (c) \leftarrow (d)
7. How do the curves $x = c$ and $y = k$ look like when plotted on the u,v plane? (b)
8. How do the curves $u = c$ and $v = k$ look like when plotted on the x,y plane? (d)
9. How do the level lines of the surfaces $z = g_1(x,y)$ and $z = g_2(x,y)$ look like? (d)
10. How do the level lines of the surfaces $z = g_1^{-1}(u,v)$ and $z = g_2^{-1}(u,v)$ look like? (b)
11. How does \mathbf{g} distort an object? (a) \rightarrow (b)
12. How does \mathbf{g}^{-1} undo this object deformation? (a) \leftarrow (b)

²⁹ Note that parts (g) and (h) of the figures are only provided in cases where the explicit expression of \mathbf{g}^{-1} is available. Consequently, they are missing in Figs. D.9 and D.18.

13. What happens when the object is rigid and can't be distorted by \mathbf{g} , and instead, to obtain an equivalent effect, the *inverse* transformation \mathbf{g}^{-1} is applied to the x, y coordinates? (a) \rightarrow (d)
14. What is the effect of $(u, v) = \mathbf{g}(x, y)$ as a domain transformation, and how does $f(\mathbf{g}(x, y))$ look like? For example, how does $\cos(g_1(x, y))$ look like? (e)
15. What is the effect of $(x, y) = \mathbf{g}^{-1}(u, v)$ as a domain transformation, and how does $f(\mathbf{g}^{-1}(u, v))$ look like? For example, how does $\cos(g_1^{-1}(x, y))$ look like? (g)
16. What is the effect of \mathbf{g} as a vector field? (f)
17. What is the effect of \mathbf{g}^{-1} as a vector field? (h)

Note that in all the following figures, both parts (a) and (d) show the same original x, y plane *before* the application of the transformation \mathbf{g} ; the only difference is that part (a) shows this plane covered by the x, y coordinate net, while part (d) shows it covered by the u, v coordinate net. In fact, we could have plotted the two coordinate nets in both figures (compare with Fig. D.5(a)), but for the sake of clarity we prefer to show them separately. In a similar way, both parts (b) and (c) show the same target u, v plane *after* the application of the transformation \mathbf{g} , but the former shows it with the x, y coordinate net while the latter shows it with the u, v coordinate net (compare with Fig. D.5(b)). Hence, in all the following figures parts (a) and (d) represent the domain of the transformation \mathbf{g} (and the range of its inverse \mathbf{g}^{-1}), while parts (b) and (c) represent the range of \mathbf{g} (and the domain of \mathbf{g}^{-1}).

Remark D.18: In the figures which accompany the following examples we provide, whenever possible, the graphic representations of both \mathbf{g} and its inverse \mathbf{g}^{-1} . However, what we want to stress by doing so is not merely the fact that a transformation and its inverse look different; this fact is, indeed, rather obvious. The important point here is that *the very same mathematical expression* $\mathbf{g}(x, y)$ (for example, $\mathbf{g}(x, y) = (2xy, y^2 - x^2)$, which is shown in Fig. D.15) may have completely different “incarnations” depending on how it is being used: When it is used as a direct transformation it bends the original unit grid into a certain curvilinear shape; when it is used as a domain (and hence, inverse) transformation it bends the unit grid (or any other rectilinear structure) into a different curvilinear shape, which corresponds, in fact, to the inverse transformation; and when it is used as a vector field, it yields yet a different graphical representation. Each of these “incarnations” simply reveals a different facet of the same transformation $\mathbf{g}(x, y)$, and one should always be sure to understand which of them is being used in any particular situation in order to avoid confusion. ■

Finally, before we proceed to the examples themselves, let us remind here that the effect of the inverse transformations, shown in parts (c), (d) in each of the following figures, has been obtained using the technique described in Remark D.7. This allows us to draw parts (c) and (d) of the figures in *all* cases, even in cases where the explicit expression of the inverse transformation \mathbf{g}^{-1} is not available (like in Figs. D.9 and D.18). However, parts (g)

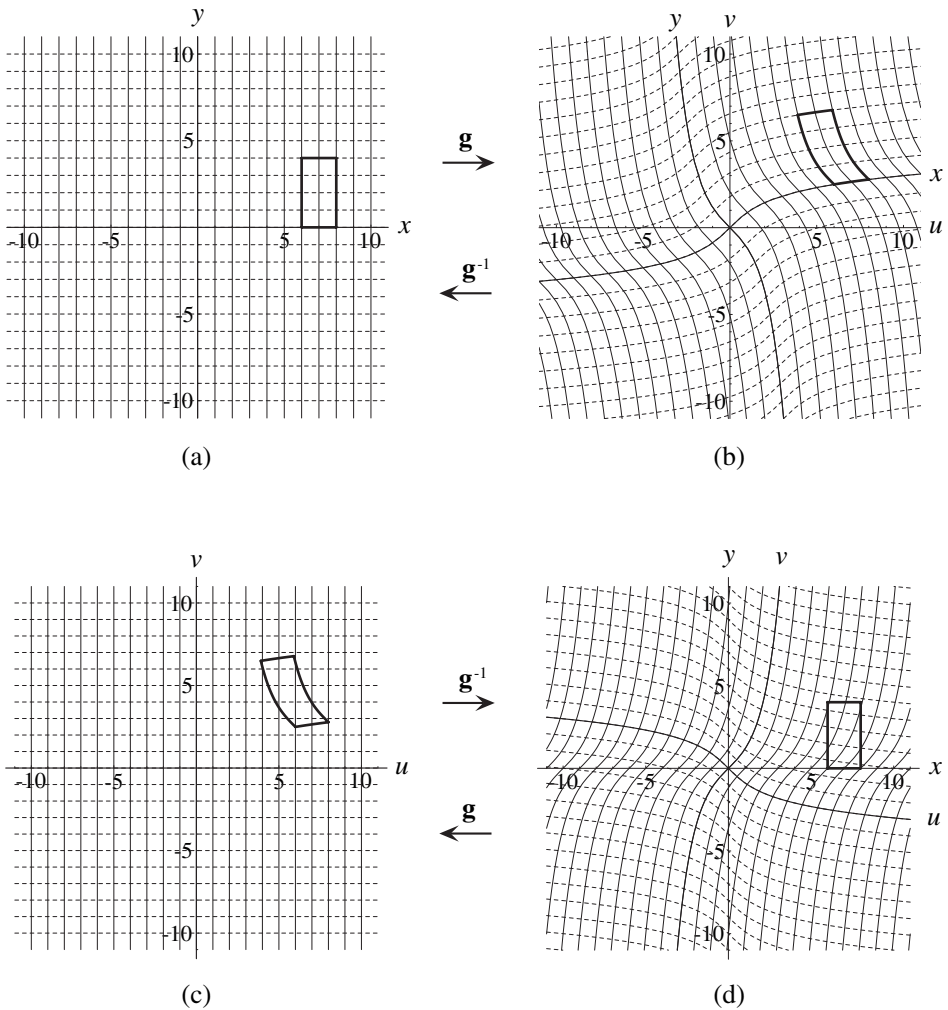


Figure D.9: A generic transformation $(u,v) = \mathbf{g}(x,y)$ (top row) and its inverse $(x,y) = \mathbf{g}^{-1}(u,v)$ (bottom row) for illustrating the general concepts. In both cases the vertical plain grid lines are mapped into the respective plain curves, and the horizontal dashed grid lines are mapped into the respective dashed curves. For the interested readers, the actual transformation used to obtain this “non-linear rotation” effect is $(u,v) = (x - \operatorname{argsinh}(y), y + \operatorname{argsinh}(x))$. Note that (b) and (c) show the *same* distorted u,v plane (the range of the transformation \mathbf{g}) which is only covered by different coordinate nets: the distorted x,y coordinate net in (b), and the standard, undistorted u,v coordinate net in (c). Similarly, (a) and (d) show the *same* undistorted x,y plane (the domain of the transformation \mathbf{g}) which is only covered by different coordinate nets: the standard, undistorted x,y coordinate net in (a), and the distorted u,v coordinate net in (d).

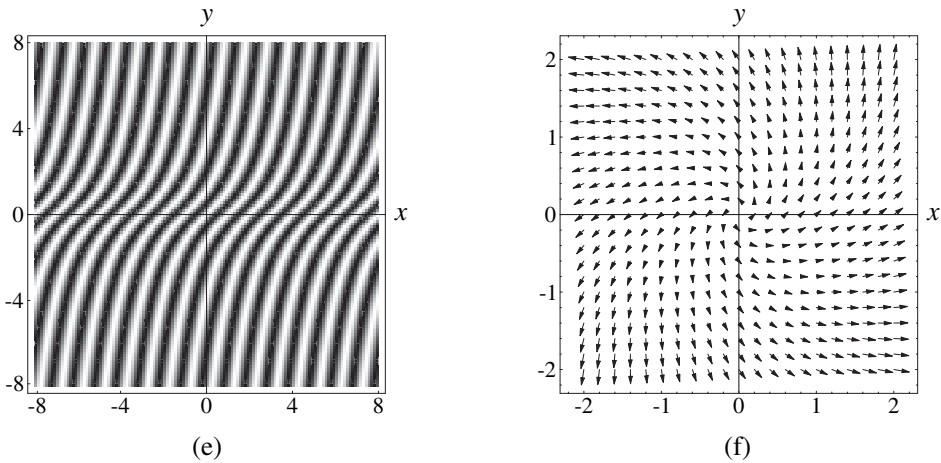


Figure D.9: (*continued.*) (e) The application of $(u,v) = (x - \operatorname{argsinh}(y), y + \operatorname{argsinh}(x))$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(4\pi[x - \operatorname{argsinh}(y)])$, a “non-linearly rotated” copy of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. Note that in the present example the explicit form of the inverse transformation is not readily available.

and (h) of the figures do not make use of this technique, and hence, as already mentioned, they are only provided when the explicit expression of \mathbf{g}^{-1} is available.

Example D.1: We start with the simple linear transformation $(u,v) = \mathbf{g}(x,y) = (2x,2y)$, that is illustrated in Figs. D.10(a),(b). Fig. D.10(a) shows the unit grid made of the lines $x = m$ and $y = n$, $m,n \in \mathbb{Z}$ before the application of this transformation, and Fig. D.10(b) shows the image of this grid after the application of the transformation. For example, the image of the vertical line $x = 1$ of Fig. D.10(a) is $u = 2$.³⁰ As we can clearly see, this transformation uniformly expands the plane by a factor of 2.

The effect of the inverse transformation \mathbf{g}^{-1} , which is given by $(x,y) = (u/2,v/2)$, is illustrated in Figs. D.10(c),(d). Fig. D.10(c) shows the unit grid made of the lines $u = m$ and $v = n$, $m,n \in \mathbb{Z}$ before the application of this inverse transformation, and Fig. D.10(d) shows the image of this grid after the application of the inverse transformation. For example, the image of the vertical line $u = 1$ of Fig. D.10(c) is $x = \frac{1}{2}$. As we can see, this transformation uniformly shrinks the u,v plane by a factor of 2.

Note, however, that when we apply the transformation $(u,v) = \mathbf{g}(x,y)$ to a given function $z = f(u,v)$ as a domain transformation, the resulting function $z = f(\mathbf{g}(x,y))$ is distorted in

³⁰ Although the present example is rather trivial, it may still be instructive to see how this result is formally obtained using Proposition D.1. Indeed, the image of the original curve $x = 1$ under \mathbf{g} is simply $u/2 = 1$, which means $u = 2$.

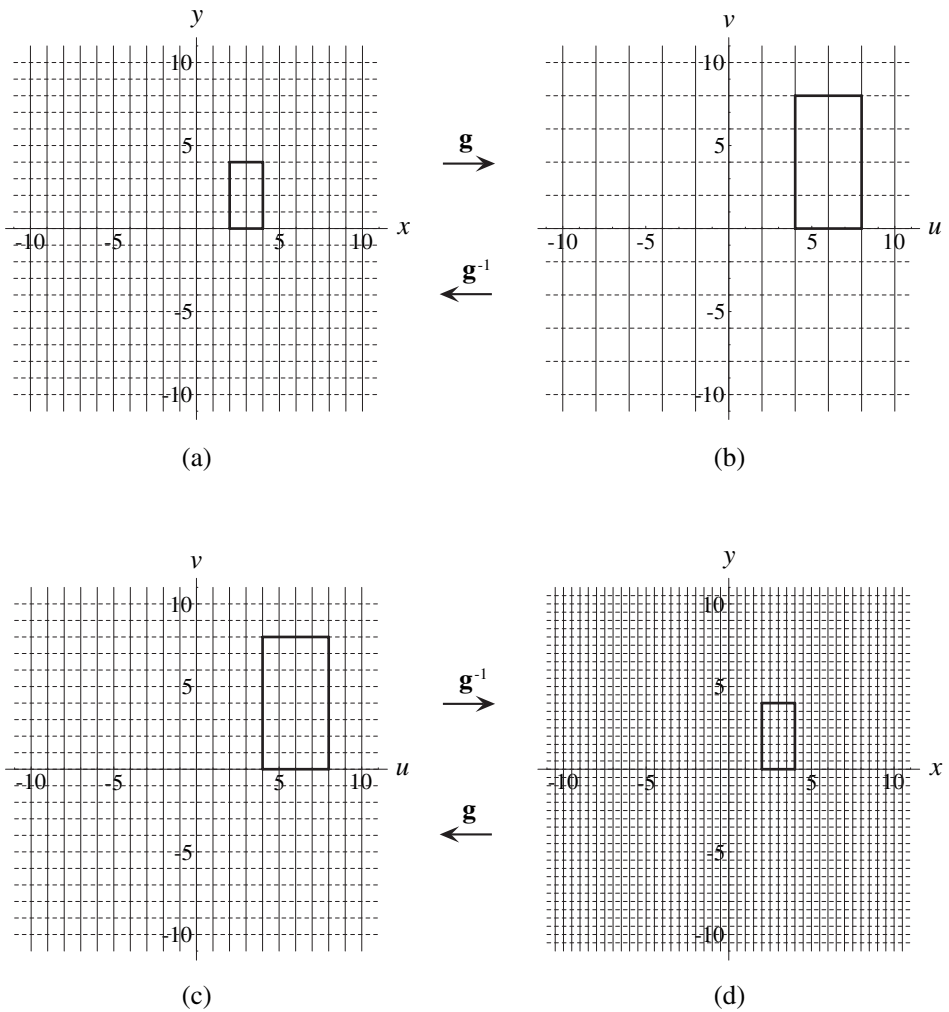
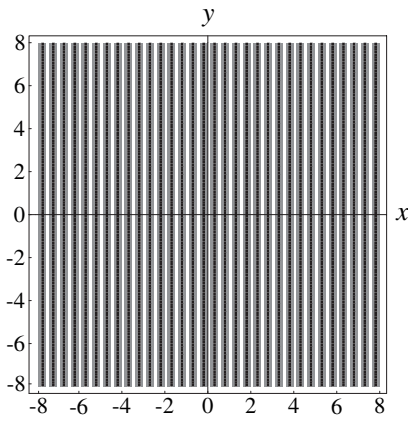
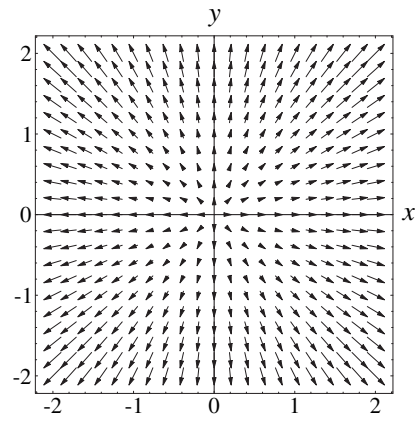


Figure D.10: (a),(b) The effect of the transformation $(u,v) = (2x,2y)$ on the unit grid consists of uniformly expanding it by a factor of 2. (c),(d) The effect of the inverse transformation $(x,y) = (u/2,v/2)$ on the unit grid consists of uniformly shrinking it by a factor of 2. In both cases the vertical plain grid lines are mapped into the respective plain lines, and the horizontal dashed grid lines are mapped into the respective dashed lines.

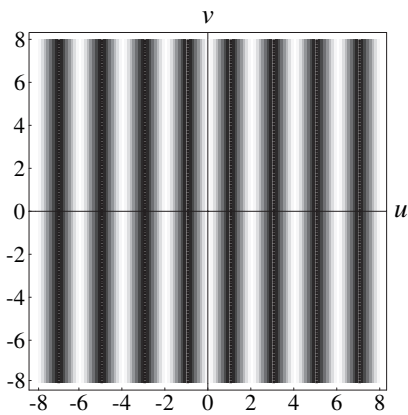
accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, considering the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$, it is clear that $z = \cos(4\pi x)$ is its *shrunk* counterpart with half the period (see Fig. D.10(e)). If we wish to *double* the period of $z = \cos(2\pi x)$, we need to apply to it the inverse transformation \mathbf{g}^{-1} , which gives $z = \cos(\pi u)$ (see Fig. D.10(g), and Remark D.4 on the variable names).



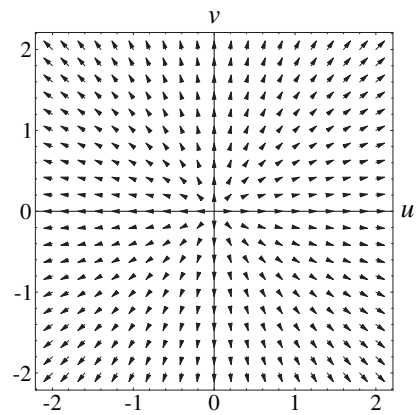
(e)



(f)



(g)



(h)

Figure D.10: (*continued.*) (e) The application of $(u,v) = (2x,2y)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(4\pi x)$, a spatially two-fold shrunk version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(x,y) = (u/2,v/2)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(\pi u)$, a spatially two-fold expanded version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field. Note that the vectors in (f) and (h) have the same radial orientations, and they only differ in their lengths; but the arrow lengths in both drawings have been scaled down in order to avoid overlappings.

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.10(f) and D.10(h), respectively. Note that both of these vector fields consist of radial vectors that emanate from the origin, and the difference is only in the vector lengths. ■

Example D.2: Consider now the linear transformation $(u,v) = \mathbf{g}(x,y)$ given by Eq. (D.1): $(u,v) = (x\cos\alpha - y\sin\alpha, x\sin\alpha + y\cos\alpha)$. The effect of this transformation on the original x,y plane is illustrated by Figs. D.11(a),(b). Fig. D.11(a) shows the unit grid made of the lines $x = m$ and $y = n$, $m,n \in \mathbb{Z}$ before the application of this transformation, and Fig. D.11(b) shows the image of this grid after the application of the transformation. As we can see, this transformation rotates the plane by angle α counterclockwise.

The effect of the inverse transformation, $(x,y) = (u\cos\alpha + v\sin\alpha, -u\sin\alpha + v\cos\alpha)$, is illustrated by Figs. D.11(c),(d). Clearly, this transformation rotates the plane by angle $-\alpha$.

Note that when we apply the transformation $(u,v) = \mathbf{g}(x,y)$ to a given function $z = f(u,v)$ as a domain transformation, the resulting function $z = f(\mathbf{g}(x,y))$ is distorted in accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, $z = \cos(2\pi[x\cos\alpha - y\sin\alpha])$ is the counterpart of the cosinusoidal surface $z = \cos(2\pi u)$ which has been rotated by angle $-\alpha$ (see Fig. D.11(e)). In order to rotate the cosinusoidal surface $z = \cos(2\pi x)$ by angle α , we need to apply to it the inverse transformation \mathbf{g}^{-1} , which gives $z = \cos(2\pi[u\cos\alpha + v\sin\alpha])$ (see Fig. D.11(g)). Similarly, the rotated version of the planar parabolic curve $y = x^2$ by angle α is given, in implicit form, by: $(-u\sin\alpha + v\cos\alpha) = (u\cos\alpha + v\sin\alpha)^2$ (see Proposition D.1).

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.11(f) and D.11(h), respectively. Note that both of these vector fields consist of vectors that follow spiral trajectories, and the difference between them is only in the direction of the spirals. ■

Example D.3: Let us consider, as our first non-linear example, the transformation $(u,v) = \mathbf{g}(x,y)$ defined by $(u,v) = (x^2, y^2)$. As we can see in Figs. D.12(a),(b), this transformation maps any x value into $u = x^2$, and any y value into $v = y^2$. The result is a non-linear expansion of the plane, whose effect increases as we move away from the origin. Note that this transformation “folds” the entire plane into the first quadrant; this is explained and illustrated in detail in [Callahan74 p. 232].

The effect of the inverse transformation, $(x,y) = (\sqrt{u}, \sqrt{v})$, is illustrated by Figs. D.12(c),(d). As we can clearly see, this transformation non-linearly shrinks the plane, where the shrinking effect becomes stronger as we move away from the origin. If we take into account both positive and negative values of the roots, the three quadrants that are “lost” when applying $\mathbf{g}(x,y)$ “reappear” under $\mathbf{g}^{-1}(u,v)$.

Hence, the range of the transformation $\mathbf{g}(x,y)$ and the domain of its inverse, $\mathbf{g}^{-1}(u,v)$, only consist here of the first quadrant of the plane. Such “losses” in the range or in the domain occur quite often in non-linear transformations.

This example can be also used to illustrate Remarks D.5 and D.6 in Sec. D.4.1: The level lines of the surfaces $z = x^2$ and $z = y^2$ (the two components of the *original* transformation $\mathbf{g}(x,y)$) are represented, respectively, by the plain and dashed lines of Fig. D.12(d), a figure which illustrates the effect of the *inverse* transformation $\mathbf{g}^{-1}(u,v)$. Note that these level lines become closer to each other as we move away from the origin, reflecting the increasing steepness of the surfaces $z = x^2$ and $z = y^2$ (remember that level lines represent constant differences in the altitude z). Similarly, the level lines of the surfaces $z = \sqrt{u}$ and $z = \sqrt{v}$ (the two components of the *inverse* transformation $\mathbf{g}^{-1}(u,v)$) are represented, respectively, by the plain and dashed lines of Fig. D.12(b), a figure which illustrates the effect of the *direct* transformation $\mathbf{g}(x,y)$.

When we apply the transformation $(u,v) = \mathbf{g}(x,y)$ to a given function $z = f(u,v)$ as a domain transformation, the resulting function $z = f(\mathbf{g}(x,y))$ is distorted in accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, the 2D function $z = \cos(2\pi x^2)$ is a non-linearly *shrunk* version of $z = \cos(2\pi u)$, whose corrugations become narrower as we move away from the origin, just like the vertical lines of Fig. D.12(d) (which are, indeed, as we have just seen above, the level lines of $z = x^2$). This is clearly shown in Fig. D.12(e). If we wish to obtain a non-linearly *expanded* version of $z = \cos(2\pi x)$, whose corrugations behave like the vertical lines of Fig. D.12(b), we need to apply to it the inverse transformation \mathbf{g}^{-1} , which gives $z = \cos(2\pi\sqrt{u})$.

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.12(f) and D.12(h), respectively. ■

Remark D.19: To visualize the level lines $z = n$ of a surface $z = f(x,y)$ one may draw the surface $\cos(2\pi f(x,y))$; its maxima, which are given by the locus $f(x,y) = n$, $n \in \mathbb{Z}$ in the x,y plane, represent these level lines. ■

Example D.4: Consider the well known Cartesian to polar coordinate transformation $(r,\theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$. (Note that we have replaced here our usual range axis names u,v by the more familiar ones, r,θ .) Although this transformation is widely used in mathematics and in engineering, it still may reserve us some surprises. To start with, let us consider the effect of this transformation on the original x,y plane. This effect is shown in Figs. D.13(a),(b): Part (a) of the figure shows the unit grid made of the lines $x = m$ and $y = n$, $m,n \in \mathbb{Z}$ before the application of this transformation, while part (b) of the figure shows the image of this grid after the application of the transformation.

The effect of the inverse transformation, the polar to Cartesian coordinate transformation $(x,y) = (r\cos\theta, r\sin\theta)$, is illustrated by Figs. D.13(c),(d). As we can see, this transformation maps the vertical lines $r = m$, $m \in \mathbb{Z}$ into the concentric circles $\sqrt{x^2 + y^2} = m$, and the horizontal dashed lines $\theta = n$, $n \in \mathbb{Z}$ into the radial dashed lines $\arctan(y/x) = n$.³¹

³¹ Note that some references use to limit the values of r and θ to $r \geq 0$ and $0 \leq \theta < 2\pi$, so that to each point (x,y) except for the origin there corresponds a unique point (r,θ) and vice versa. However, this restriction is not always advantageous [Colley98 pp. 67–69], and we prefer not to impose it here. Thus, the vertical lines $r = m$ and $r = -m$ are both mapped into the same circle; similarly, the horizontal lines $\theta = n$ and $\theta = 2\pi k + n$ for any $k \in \mathbb{Z}$ are mapped into the same radial line.

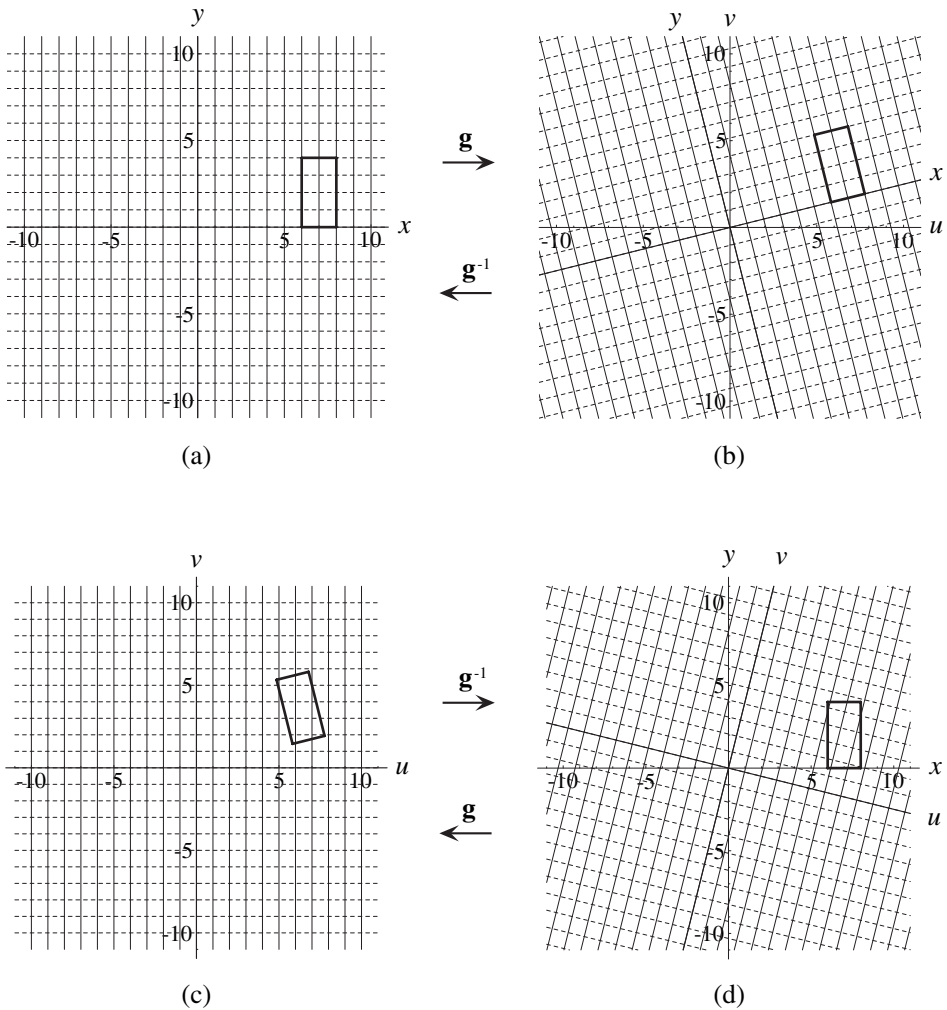
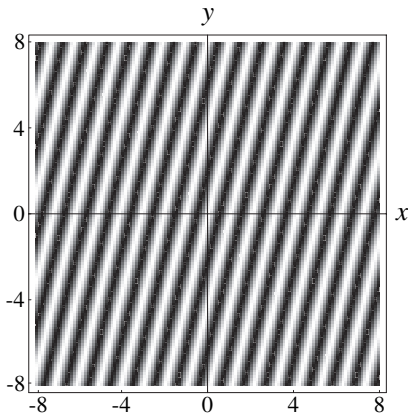
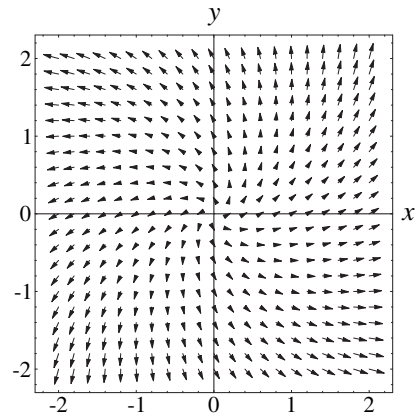


Figure D.11: (a),(b) The effect of the transformation $(u,v) = (x\cos\alpha - y\sin\alpha, x\sin\alpha + y\cos\alpha)$ on the unit grid consists of rotating it counter-clockwise by angle α . (c),(d) The effect of the inverse transformation $(x,y) = (u\cos\alpha + v\sin\alpha, -u\sin\alpha + v\cos\alpha)$ on the unit grid consists of rotating it by angle $-\alpha$. In both cases the vertical plain grid lines are mapped into the respective plain lines, and the horizontal dashed grid lines are mapped into the respective dashed lines.

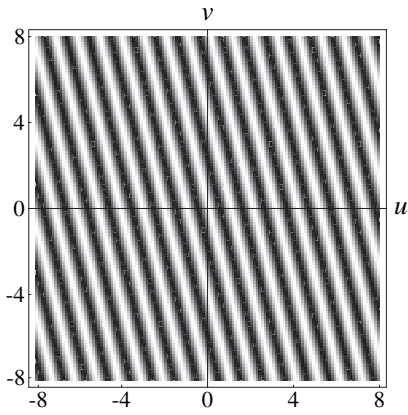
Interestingly, in this case it is the effect of the *inverse* transformation \mathbf{g}^{-1} on the unit grid that is widely known (Figs. D.13(c),(d)), while the effect of \mathbf{g} itself on the unit grid looks quite unfamiliar (Figs. D.13(a),(b)), and is rarely if ever mentioned in the literature. It can be easily verified, using Eqs. (D.7) and (D.8), that \mathbf{g} maps the vertical grid lines $x = m$, $m \in \mathbb{Z}$ in Fig. D.13(a) into the family of curves $r\cos\theta = m$, namely, the secant curves



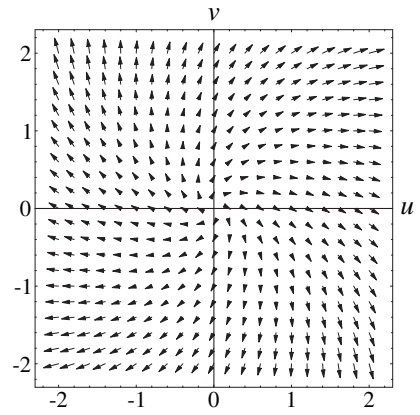
(e)



(f)



(g)



(h)

Figure D.11: (*continued.*) (e) The application of $(u,v) = (x \cos \alpha - y \sin \alpha, x \sin \alpha + y \cos \alpha)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(4\pi[x \cos \alpha - y \sin \alpha])$, a copy of the original function that is rotated by angle $-\alpha$; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field; note the left-oriented spiral shape of this vector field. (g) The application of the *inverse* transformation $(x,y) = (u \cos \alpha + v \sin \alpha, -u \sin \alpha + v \cos \alpha)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(2\pi[u \cos \alpha + v \sin \alpha])$, a copy of the original function that is rotated by angle α ; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same *inverse* transformation as a vector field; note the right-oriented spiral shape of this vector field.

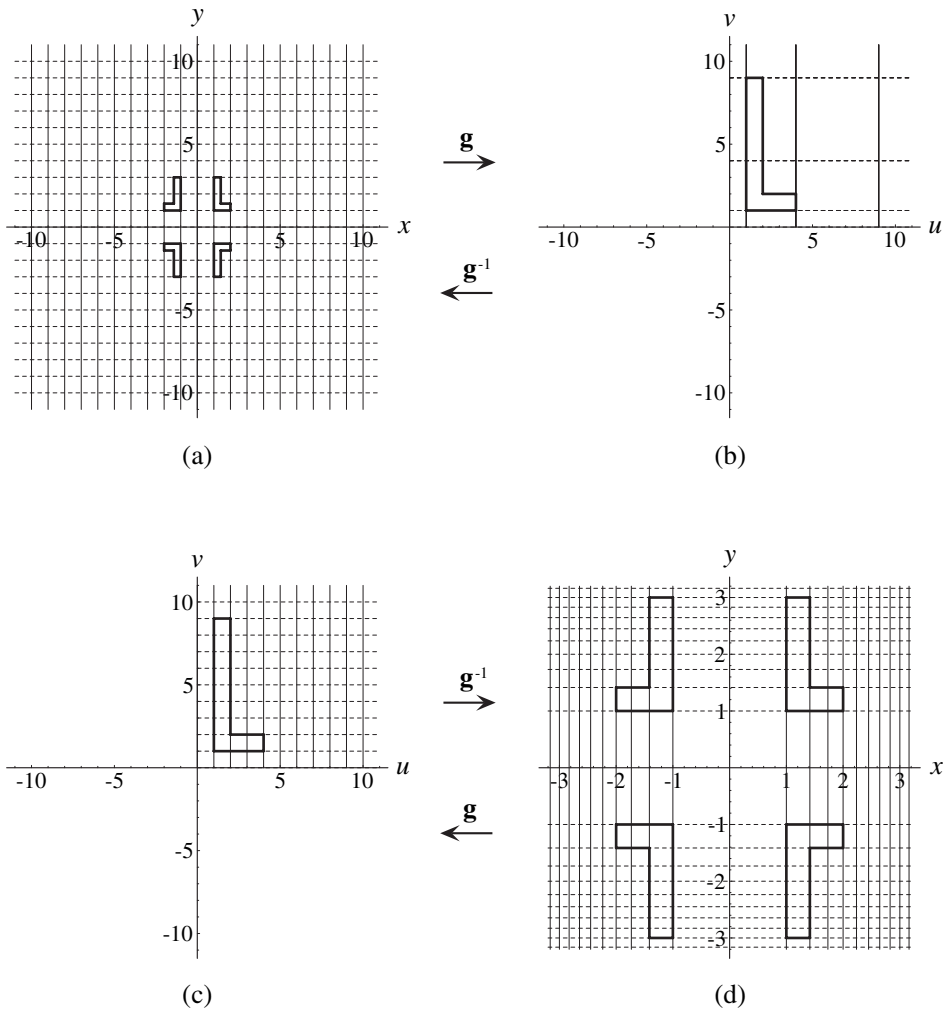


Figure D.12: (a),(b) The effect of the transformation $(u,v) = (x^2,y^2)$ on the unit grid consists of a non-linear expansion. Each of the four objects in (a) is mapped into the same image in (b). (c),(d) The effect of the inverse transformation $(x,y) = (\sqrt{u}, \sqrt{v})$ on the unit grid consists of a non-linear contraction. In both cases the vertical plain grid lines are mapped into the respective plain lines, and the horizontal dashed grid lines are mapped into the respective dashed lines. Note that part (d) has been magnified to better show details.

$r = m \sec \theta$; similarly, g maps the dashed horizontal grid lines $y = n$, $n \in \mathbb{Z}$ in Fig. D.13(a) into the family of dashed curves $r \sin \theta = n$, which are the cosecant curves $r = n \operatorname{cosec} \theta$. (The shapes of the multi-branched curves $y = \sec x$ and $y = \operatorname{cosec} x$ can be found in any mathematical handbook such as [Bronshtein79 p. 69] or [Harris98 p. 301].)

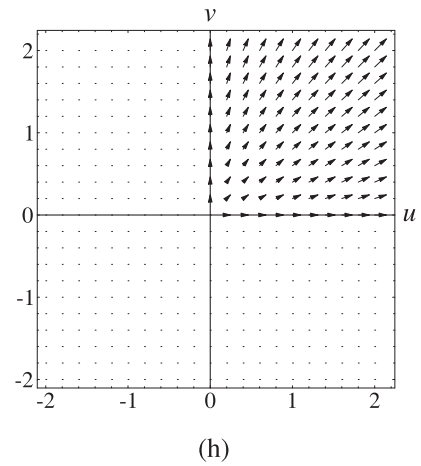
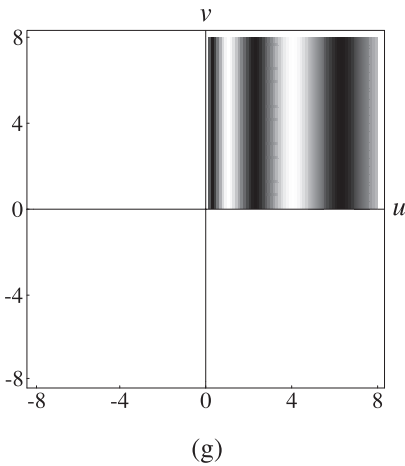
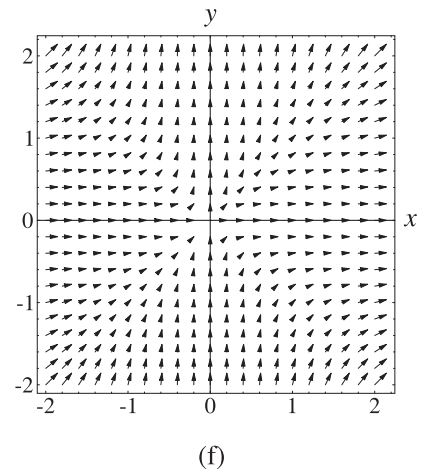
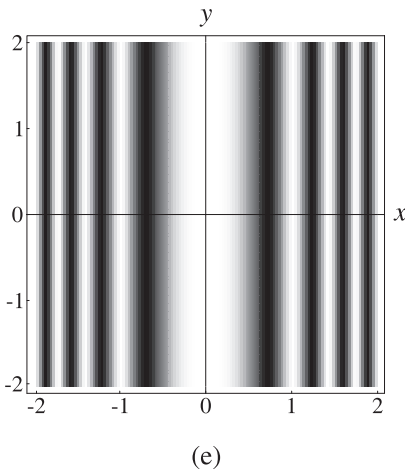


Figure D.12: (*continued.*) (e) The application of $(u, v) = (x^2, y^2)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(2\pi x^2)$, a non-linearly shrunk version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(x, y) = (\sqrt{u}, \sqrt{v})$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(2\pi \sqrt{u})$, a non-linearly expanded version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field. Note that part (e) has been magnified to better show details.

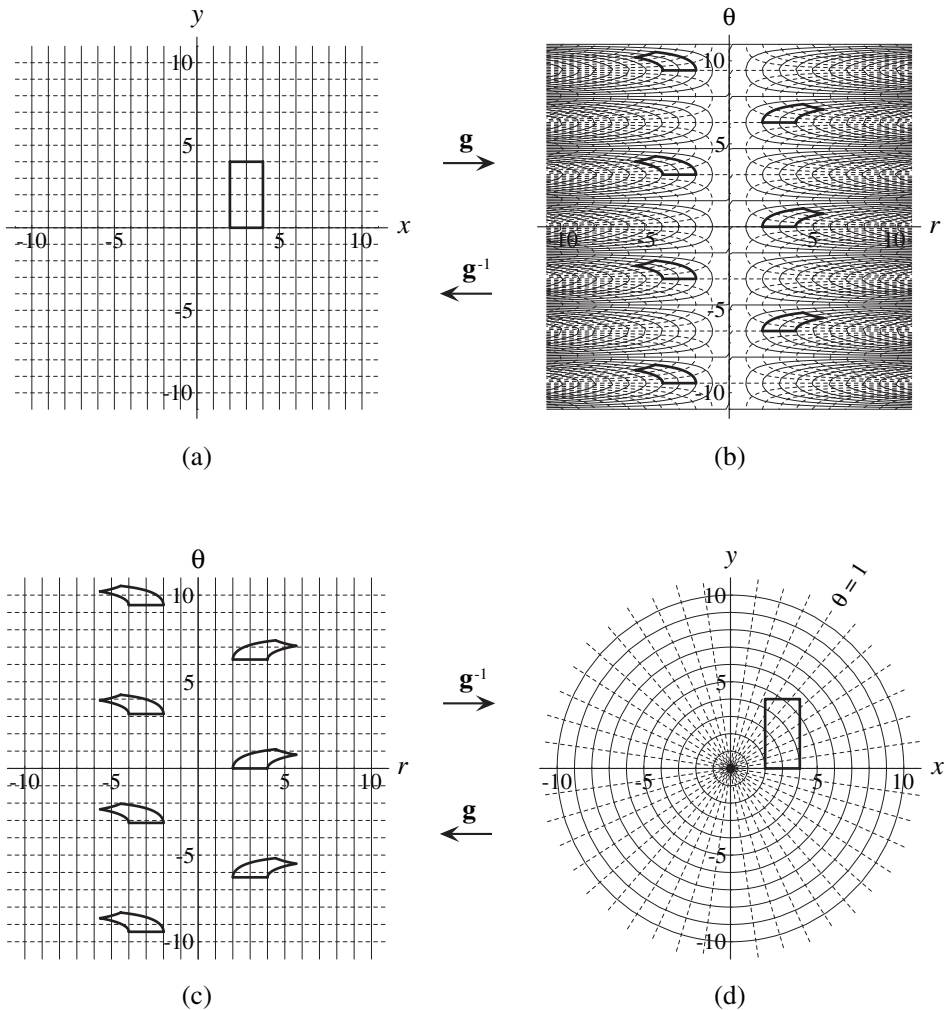
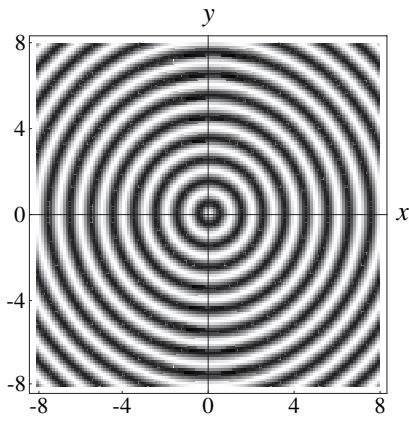
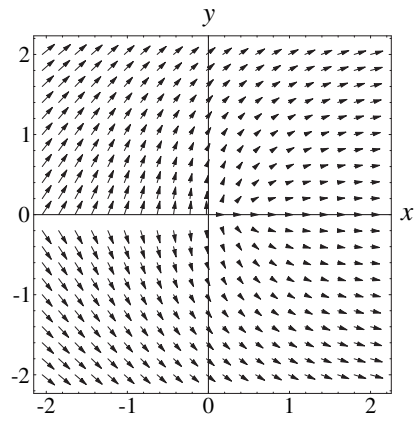


Figure D.13: (a),(b) The effect of the transformation $(r, \theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$ on the unit grid: Each vertical line $x = m$ is mapped into a secant curve $r = m \sec \theta$, and each horizontal line $y = n$ is mapped into a cosecant curve $r = n \csc \theta$. Note that the rectangular object is mapped into infinitely many images. (c),(d) The effect of the inverse transformation $(x, y) = (r \cos \theta, r \sin \theta)$ on the unit grid: Each vertical line is mapped into a circle, and each horizontal line is mapped into a radial line. Note that the horizontal line $\theta = 1$ is mapped into the radial line whose angle is 1 radian (i.e. $180^\circ/\pi \approx 57.2958^\circ$), the line $\theta = 2$ is mapped into the line at 2 radians, and so forth.³²

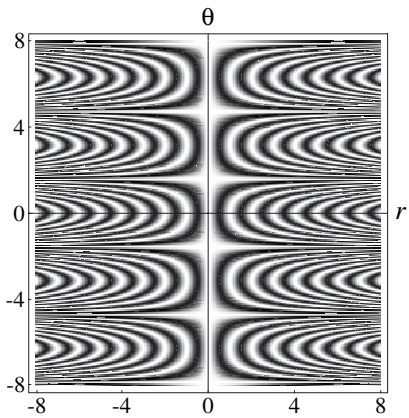
³² Note that the lines $\theta = -10, \theta = -9, \dots, \theta = 10$ of Fig. D.13(c) are mapped into almost equispaced radial lines in Fig. D.13(d). This happens since by pure coincidence 11 radians almost exactly coincide with an integer multiple of 270° (3 quadrants), so that the images of the lines $\theta = k$ and $\theta = k + 22$ fall almost precisely on the same radial line. But in reality, the image of the infinite family of lines $\theta = k$, $k \in \mathbb{Z}$ is everywhere dense in the x, y plane (see the lemma in [Arnold73 pp. 163–164]).



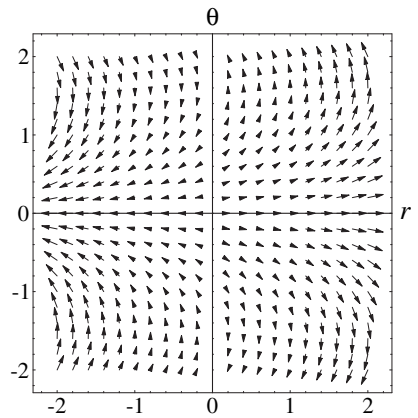
(e)



(f)



(g)



(h)

Figure D.13: (*continued.*) (e) The application of $(r, \theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi r)$ gives $z = \cos(2\pi\sqrt{x^2 + y^2})$, a circular version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(x, y) = (r \cos \theta, r \sin \theta)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(2\pi r \cos \theta)$, a curved version of the original function whose corrugations have the shape of multi-branch secant curves; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field.

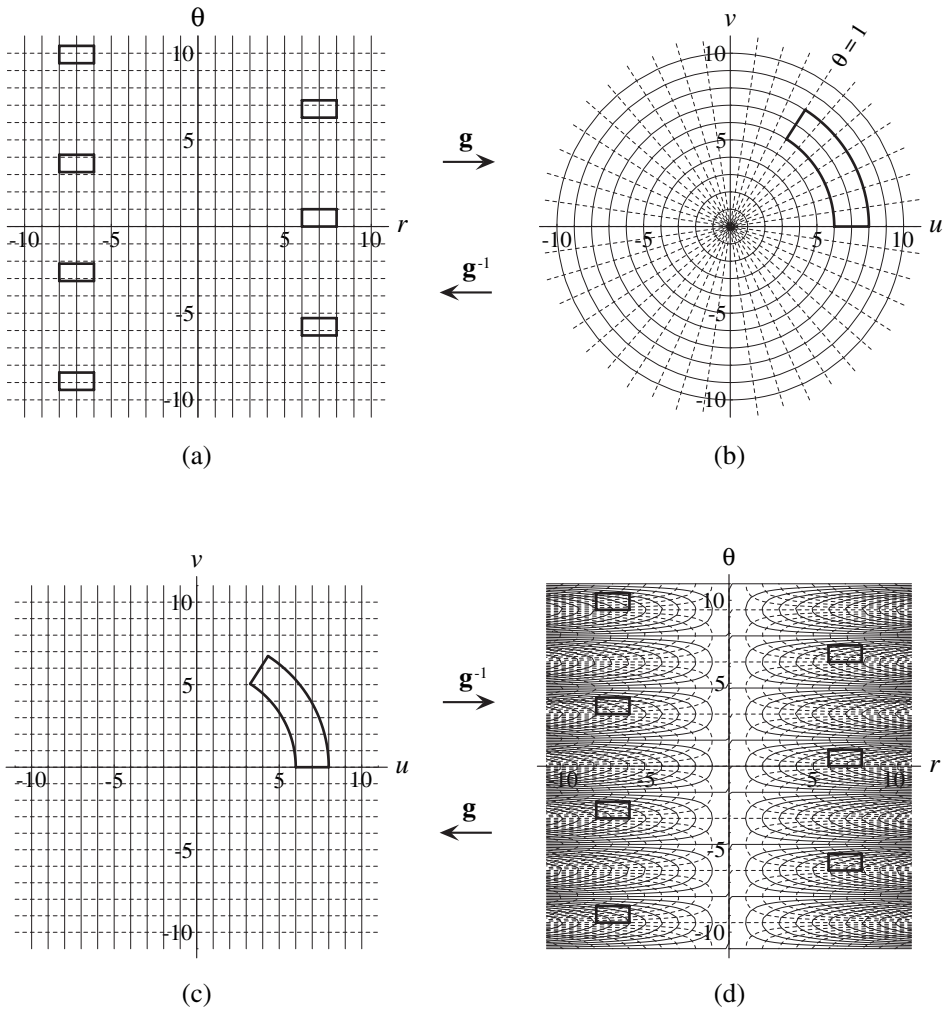


Figure D.14: (a),(b) The effect of the transformation $(u,v) = (r\cos\theta, r\sin\theta)$ on the unit grid: Each vertical line is mapped into a circle, and each horizontal line is mapped into a radial line. Note that infinitely many objects in (a) are mapped into the same object in (b). (c),(d) The effect of the inverse transformation $(r,\theta) = (\sqrt{u^2 + v^2}, \arctan(v/u))$ on the unit grid: Each vertical line is mapped into a multi-branch secant curve, and each horizontal line is mapped into a multi-branch cosecant curve. All angles are measured in radians.

So far, in all the previous examples, we described the effect of a transformation $\mathbf{g}(x,y)$ by means of its influence on the unit grid of the x,y plane. However, we can also describe the effect of $\mathbf{g}(x,y)$ by studying its influence on other curves in the x,y plane (see Sec. D.4.2).

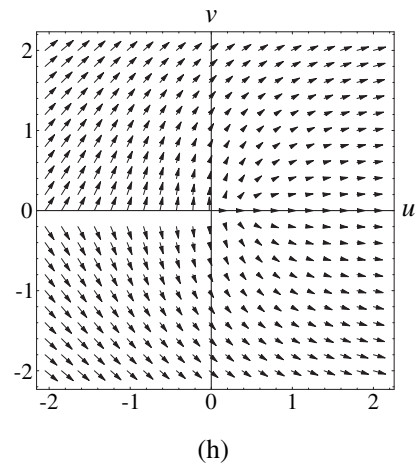
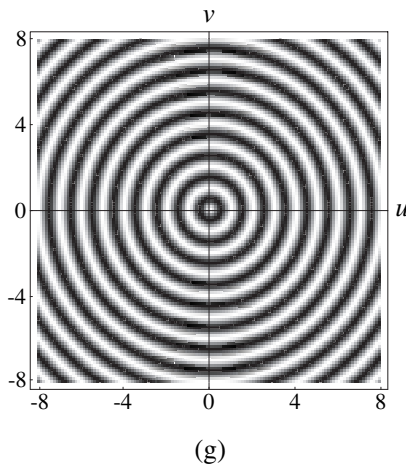
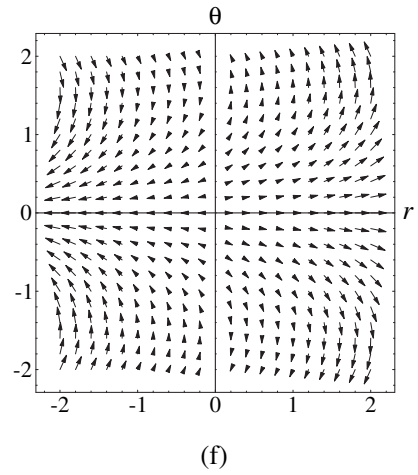
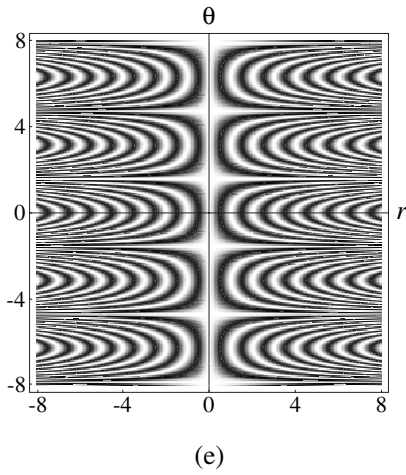


Figure D.14: (*continued.*) (e) The application of $(u, v) = (r \cos \theta, r \sin \theta)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(2\pi r \cos \theta)$, a curved version of the original function whose corrugations have the shape of multi-branch secant curves; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(r, \theta) = (\sqrt{u^2 + v^2}, \arctan(v/u))$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(2\pi \sqrt{u^2 + v^2})$, a circular version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field.

For example, our transformation $(r, \theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$, which maps vertical or horizontal straight lines into secant or cosecant curves, also maps the circles $\sqrt{x^2 + y^2} = m$, $m \in \mathbb{Z}$ into the vertical lines $r = m$, and the radial lines $\arctan(y/x) = n$, $n \in \mathbb{Z}$ into the horizontal lines $\theta = n$. This can be seen by drawing circles and radial lines on top of Fig. D.13(a) and tracing out their images in Fig. D.13(b), or, equivalently, by considering the image under \mathbf{g} of the circles and of the radial lines of Fig. D.13(d), as shown in Fig. D.13(c). Note that the mapping from Fig. D.13(d) back into Fig. D.13(c) corresponds to the inverse of \mathbf{g}^{-1} , which is \mathbf{g} itself. Interestingly, this effect of the Cartesian to polar coordinate change $(r, \theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$ is much more familiar than its effect on the Cartesian x, y grid that is shown in Figs. D.13(a), (b). This is, in fact, the way this transformation is described in the literature (see, for example, [Courant88 pp. 137–139]). Indeed, this case is a classical example of a transformation that is not illustrated in the literature by its effect on the standard x, y grid, but rather by its effect on some other curves in the x, y plane (see representation (8) in Sec. D.2).

Note that when we apply the transformation $(u, v) = \mathbf{g}(x, y)$ to a given function $z = f(u, v)$ as a domain transformation, the resulting function $z = f(\mathbf{g}(x, y))$ is distorted in accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, after the application of $\mathbf{g}(x, y)$ the straight vertical corrugations of the cosinusoidal function $z = \cos(2\pi r)$ in the r, θ space become in the x, y space, in the resulting function $z = \cos(2\pi\sqrt{x^2 + y^2})$, a family of concentric circular corrugations that surround the origin (see Fig. D.13(e), and compare with Fig. D.13(d)).³³ On the other hand, applying the inverse transformation \mathbf{g}^{-1} to the function $z = \cos(2\pi x)$ gives the unfamiliar function $z = \cos(2\pi r \cos \theta)$ which is shown in Fig. D.13(g); compare also with Fig. D.13(b).

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.13(f) and D.13(h), respectively. As we can see, both of these vector fields look rather unfamiliar, and they are rarely if ever mentioned in the literature.

Because of the particular importance of the transformation discussed in this example, we also show its inverse (the polar to Cartesian coordinate conversion) as a transformation in its own right in a separate figure, Fig. D.14. Note that Fig. D.14 shows how the inverse transformation distorts a given rectangular object, whereas Fig. D.13 shows a much less familiar result, the way in which $\mathbf{g}(x, y)$ itself distorts a rectangular object. ■

Example D.5: Figs. D.15(a), (b) illustrate the influence on the x, y plane of the transformation $\mathbf{g}(x, y)$ given by $(u, v) = (2xy, y^2 - x^2)$ [Courant88 pp. 136–137]. The vertical lines $x = m$, $m \in \mathbb{Z}$ in Fig. D.15(a) are mapped by this transformation into the family of top-opened parabolas $v = \frac{1}{4m^2}u^2 - m^2$, while the horizontal dashed lines $y = n$, $n \in \mathbb{Z}$ in Fig. D.15(a) are mapped into the family of bottom-opened parabolas $v = n^2 - \frac{1}{4n^2}u^2$ shown in Fig. D.15(b) by dashed lines.³⁴

³³ This explains, indeed, why in some references the transformation $\mathbf{g}(x, y)$ is called *polar to Cartesian* rather than *Cartesian to polar* coordinate transformation (see Remark D.16).

³⁴ Note that the vertical lines $x = m$ and $x = -m$ are both mapped into the same top-opened parabola; similarly, the horizontal lines $y = n$ and $y = -n$ are both mapped into the same bottom-opened parabola.

In this case the explicit expression of the inverse transformation is quite complicated: $(x,y) = (\sqrt{(\sqrt{u^2 + v^2} - v)/2}, \sqrt{(\sqrt{u^2 + v^2} + v)/2})$. We therefore use this example to illustrate how the influence of the inverse transformation on the u,v plane can be found without even knowing its explicit form (see Remark D.7): As shown in Figs. D.15(c),(d), the vertical lines $u = m, m \in \mathbb{Z}$ are mapped by the inverse transformation into the family of hyperbolas which are given according to Eq. (D.11) by $u = 2xy = m$, namely, $y = \frac{m}{2x}$. Similarly, the horizontal dashed lines $v = n, n \in \mathbb{Z}$ are mapped by the inverse transformation into the dashed family of hyperbolas which are given according to Eq. (D.12) by $v = y^2 - x^2 = n$, namely, $y = \pm\sqrt{n + x^2}$. Thus, we avoided using here the explicit form of \mathbf{g}^{-1} .

Note that just as we did in the previous example, we may also consider the effects of $\mathbf{g}(x,y)$ and of its inverse $\mathbf{g}^{-1}(u,v)$ on *curved* line families. For example, we can see by comparing Fig. D.15(d) with Fig. D.15(c) that $\mathbf{g}(x,y)$ (the inverse of $\mathbf{g}^{-1}(u,v)$) maps the hyperbolic lines $2xy = m, m \in \mathbb{Z}$ into the vertical lines $u = m$ and the dashed hyperbolic lines $y^2 - x^2 = n, n \in \mathbb{Z}$ into the dashed horizontal lines $v = n$. Similarly, we can see by comparing Fig. D.15(b) with Fig. D.15(a) that the inverse transformation $\mathbf{g}^{-1}(u,v)$ maps the top-opened parabolas of Fig. D.15(b) into the vertical lines of Fig. D.15(a) and the dashed, bottom-opened parabolas into horizontal lines.

This example can be also used to illustrate Remarks D.5 and D.6 in Sec. D.4.1: The level lines of the surfaces $z = 2xy$ and $z = y^2 - x^2$ (the two components of the *original* transformation $\mathbf{g}(x,y)$) are represented, respectively, by the plain and dashed hyperbolas of Fig. D.15(d), a figure which illustrates the effect of the *inverse* transformation $\mathbf{g}^{-1}(u,v)$. Similarly, the level lines of the surfaces $z = \sqrt{(\sqrt{u^2 + v^2} - v)/2}$ and $z = \sqrt{(\sqrt{u^2 + v^2} + v)/2}$ (the two components of the *inverse* transformation $\mathbf{g}^{-1}(u,v)$) are represented, respectively, by the plain and dashed parabolas of Fig. D.15(b), a figure which illustrates the effect of the *direct* transformation $\mathbf{g}(x,y)$.

Once again, note that when we apply the transformation $(u,v) = \mathbf{g}(x,y)$ to a given function $z = f(u,v)$ as a domain transformation, the resulting function $z = f(\mathbf{g}(x,y))$ is distorted in accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, following the application of $\mathbf{g}(x,y)$ the straight vertical corrugations of the function $z = \cos(2\pi u)$ become in the resulting function, $z = \cos(4\pi xy)$, a family of hyperbolic corrugations, just like the plain curves in Fig. D.15(d) (which are, indeed, as we have just seen above, the level lines of $z = 2xy$). This is clearly shown in Fig. D.15(e). On the other hand, applying the inverse transformation \mathbf{g}^{-1} to the function $z = \cos(2\pi x)$ gives a family of parabolic corrugations (see Fig. D.15(g)), just like the plain curves in Fig. D.15(b).

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.15(f) and D.15(h), respectively. While the vector field of \mathbf{g} is widely known (see, for example, Fig. 4.14(a) in Chapter 4), the vector field of \mathbf{g}^{-1} looks rather unfamiliar. ■

Example D.6: Suppose we wish to construct a transformation $\mathbf{g}(x,y)$ that bends vertical lines into right-opened parabolas. This transformation shifts each point (x,y) to the right

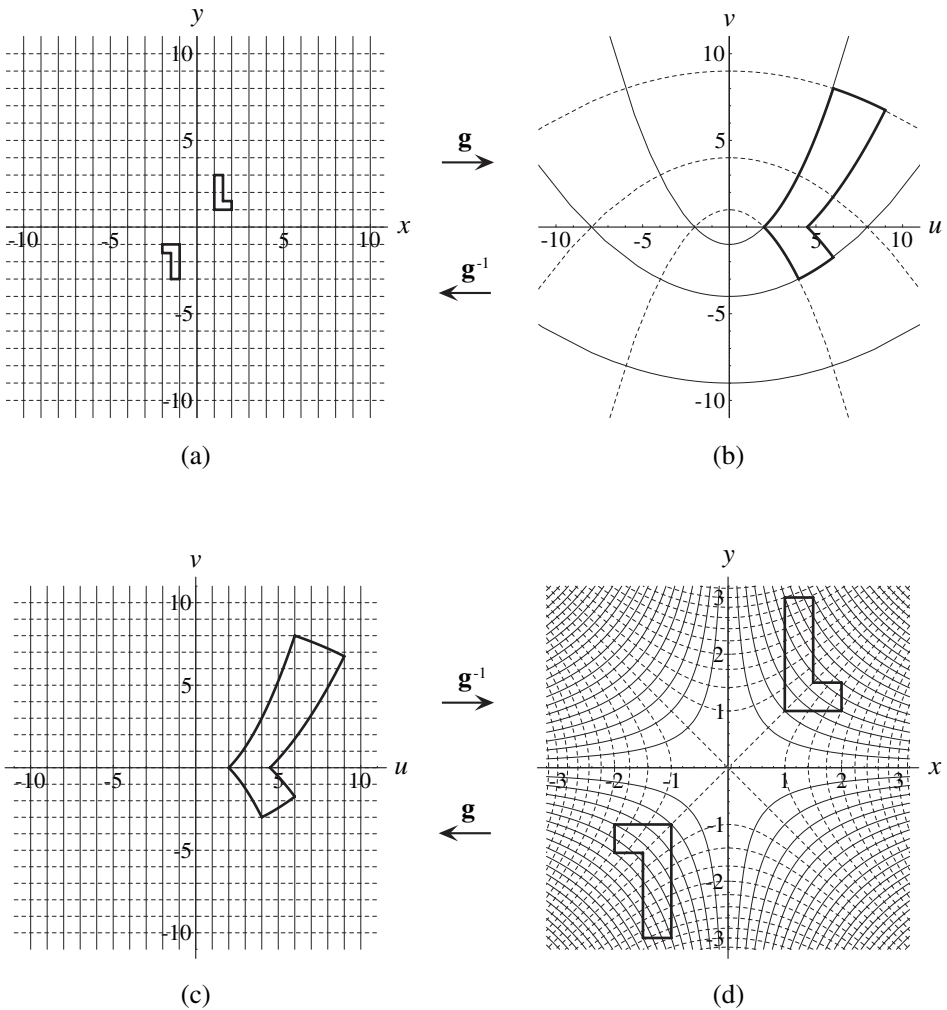
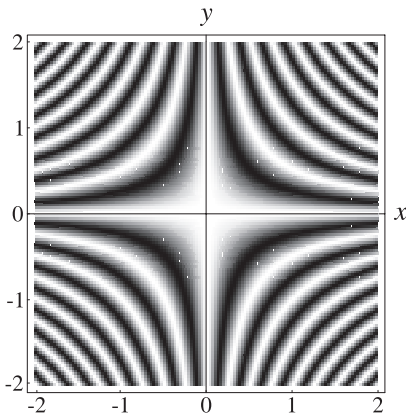
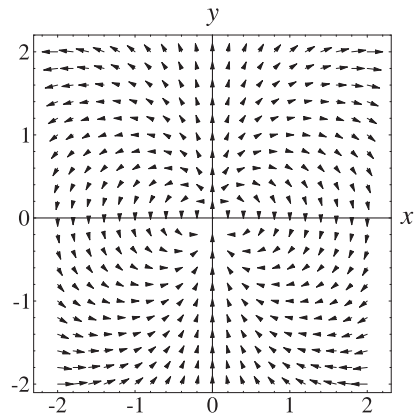


Figure D.15: (a),(b) The effect of the transformation $(u,v) = (2xy, y^2 - x^2)$ on the unit grid: Each vertical line is mapped into a top-opened parabola, while each horizontal line is mapped into a bottom-opened parabola. Both of the objects in (a) are mapped into the same object in (b). (c),(d) The effect of the inverse transformation on the unit grid: Each vertical line is mapped into a hyperbola which is asymptotic to the axes, while each horizontal line is mapped into a hyperbola which is asymptotic to the main diagonals. Note that part (d) has been magnified to better show details.

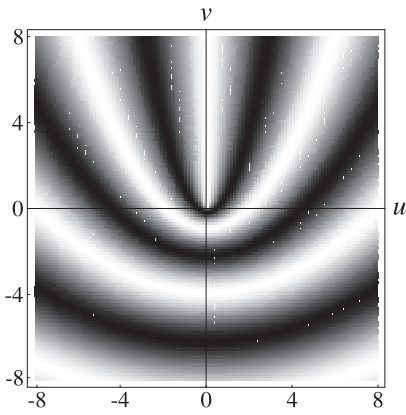
by ay^2 , a being a positive constant, while the y coordinate remains unchanged (see Fig. 3.4 in Chapter 3). It is clear, therefore, that this transformation is given by $(u,v) = (x + ay^2, y)$. And indeed, as we can clearly see by comparing Fig. D.16(a) with Fig. D.16(b), this



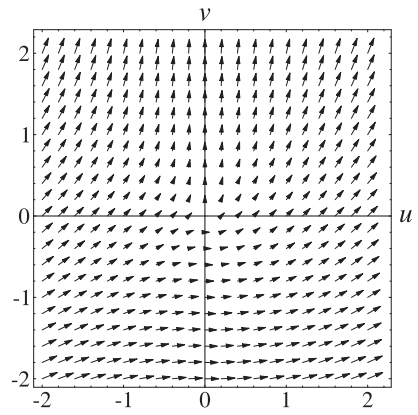
(e)



(f)



(g)



(h)

Figure D.15: (*continued.*) (e) The application of $(u,v) = (2xy, y^2 - x^2)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(4\pi xy)$, a hyperbolic version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives a parabolic version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field. Note that part (e) has been magnified to better show details.

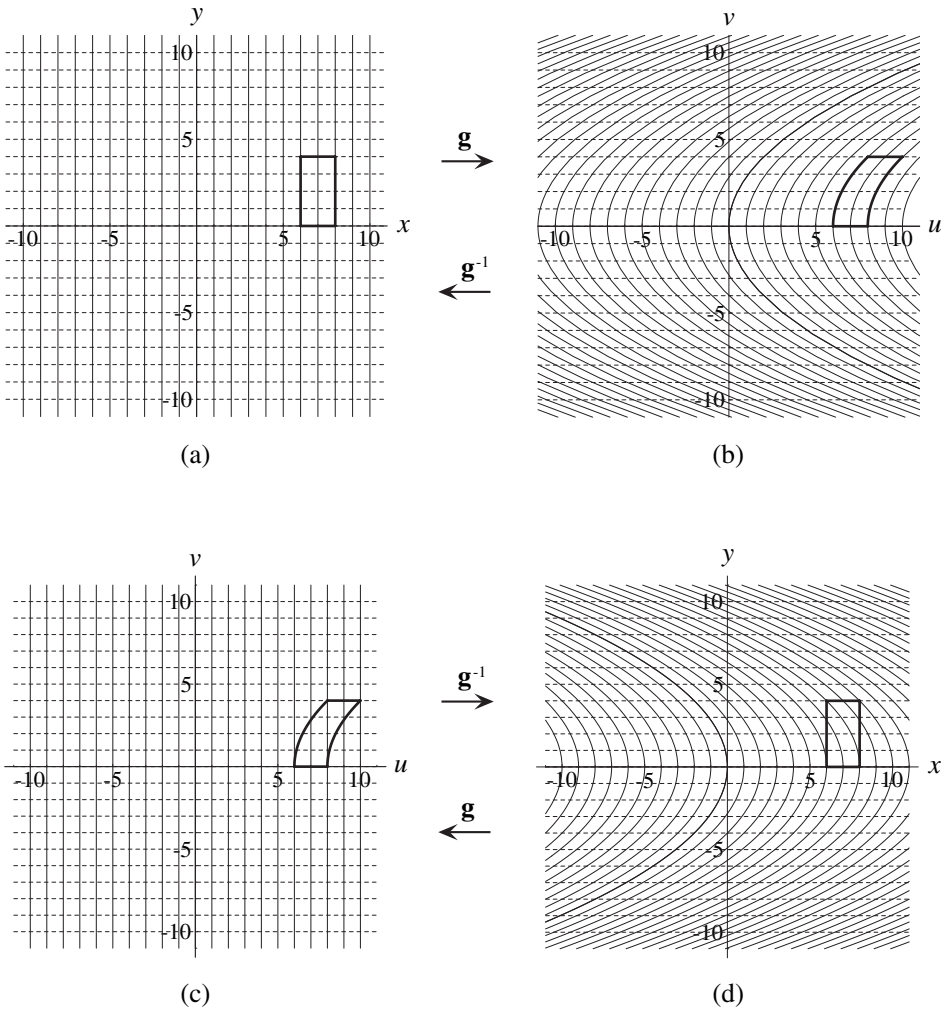


Figure D.16: (a),(b) The effect of the transformation $(u,v) = (x + ay^2, y)$ on the unit grid: Each vertical line is mapped into a right-opened parabola, while each horizontal line is simply shifted to the right. (c),(d) The effect of the inverse transformation $(x,y) = (u - av^2, v)$ on the unit grid: Each vertical line is mapped into a left-opened parabola, while each horizontal line is simply shifted to the left.

transformation bends each vertical line $x = m$, $m \in \mathbb{Z}$ into a right-opened parabola $x = ay^2 + m$, while each horizontal dashed line $y = n$, $n \in \mathbb{Z}$ is simply shifted to the right, i.e. mapped into itself. Consequently, this transformation maps the unit grid of the x,y plane (Fig. D.16(a)) into a right-opened parabolic grid (Fig. D.16(b)); this is, in fact, a non-linear horizontal shearing operation. Similarly, the inverse transformation, which is

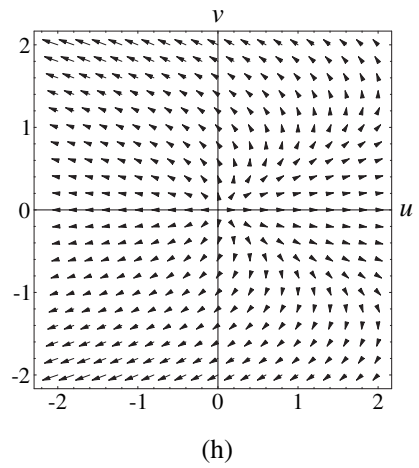
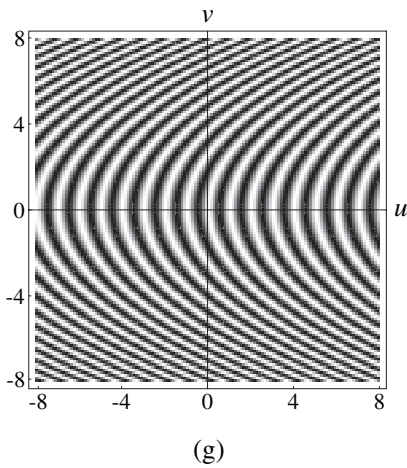
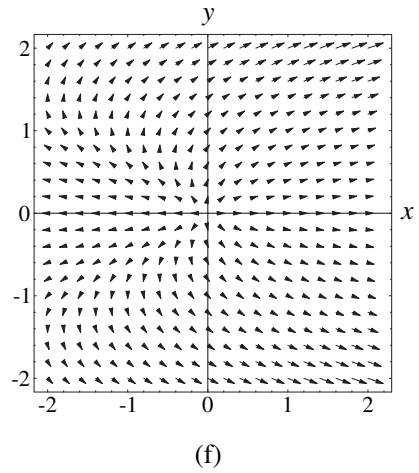
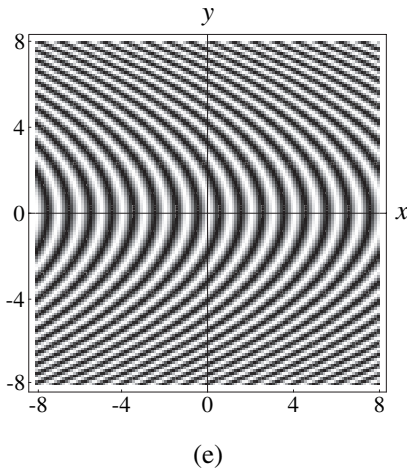


Figure D.16: (*continued.*) (e) The application of $(u, v) = (x + ay^2, y)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(x + ay^2)$, a left-opened parabolic version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(x, y) = (u - av^2, v)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(2\pi[u - av^2])$, a right-opened parabolic version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field.

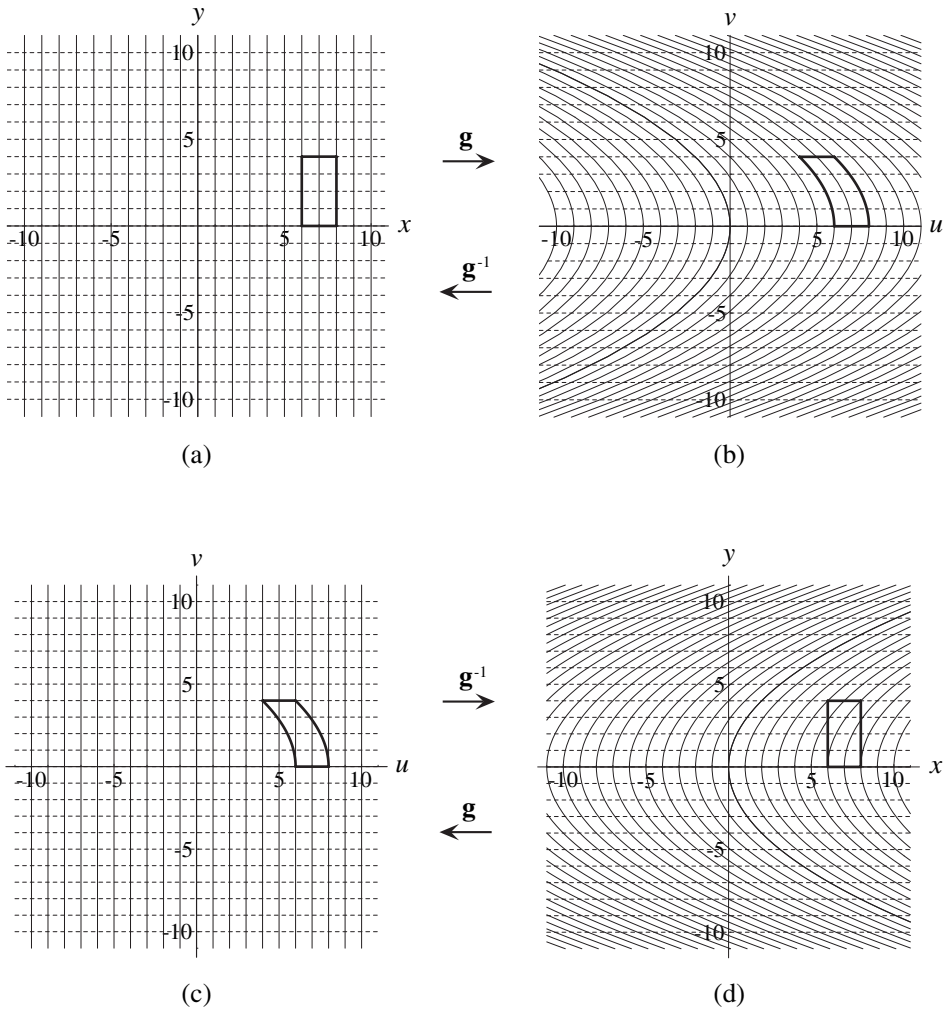
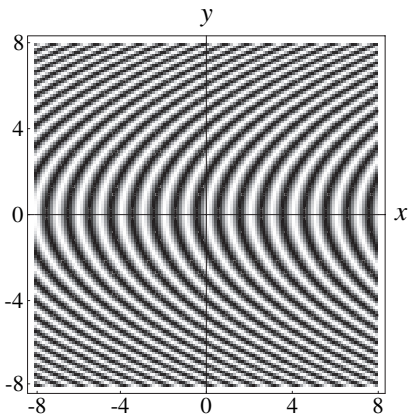


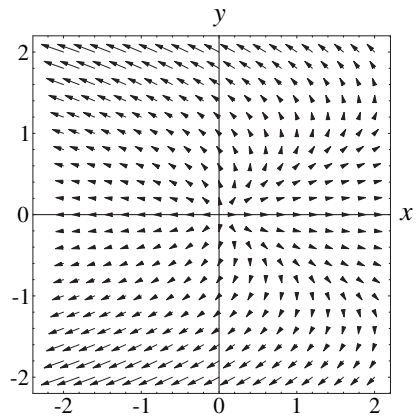
Figure D.17: (a),(b) The effect of the transformation $(u,v) = (x - ay^2, y)$ on the unit grid: Each vertical line is mapped into a left-opened parabola, while each horizontal line is simply shifted to the left. (c),(d) The effect of the inverse transformation $(x,y) = (u + av^2, v)$ on the unit grid: Each vertical line is mapped into a right-opened parabola, while each horizontal line is simply shifted to the right.

given by $(x,y) = (u - av^2, v)$, maps the unit grid (Fig. D.16(c)) into a left-opened parabolic grid (Fig. D.16(d)).

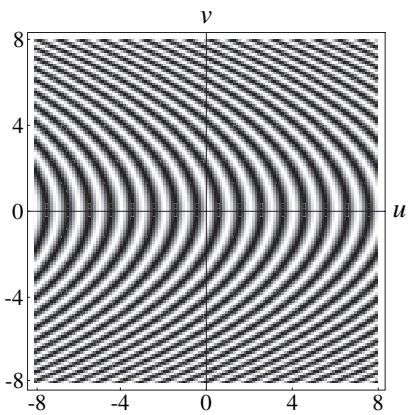
However, if our aim is to construct a transformation that bends the vertical corrugations of the function $z = \cos(2\pi u)$ into right-opened parabolas, as shown in Fig. D.16(g), then



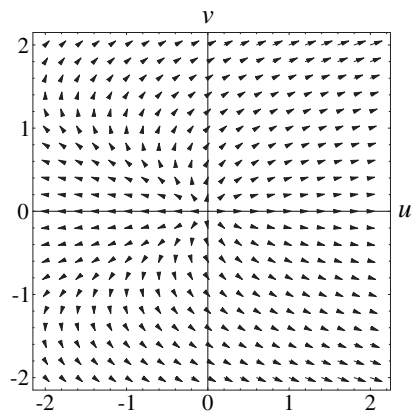
(e)



(f)



(g)



(h)

Figure D.17: (*continued.*) (e) The application of $(u, v) = (x - ay^2, y)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(2\pi[x - ay^2])$, a right-opened parabolic version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. (g) The application of the inverse transformation $(x, y) = (u + av^2, v)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi x)$ gives $z = \cos(2\pi[u + av^2])$, a left-opened parabolic version of the original function; compare with the effect of the *direct* transformation on the plain grid lines in (b). (h) The effect of the same inverse transformation as a vector field.

the transformation we need is the *inverse* of $\mathbf{g}(x,y)$, which is given, back in the original coordinate system (see Remark D.4 on the variable names), by $(u,v) = (x - ay^2, y)$. This transformation is shown as a transformation in its own right in a separate figure, Fig. D.17. As shown in Figs. D.17(a),(b), this transformation bends each vertical line $x = m$, $m \in \mathbb{Z}$ into a *left*-opened parabola, while each horizontal dashed line $y = n$, $n \in \mathbb{Z}$ is simply shifted to the left, i.e. mapped into itself. In other words, this transformation maps the unit grid of the x,y plane (Fig. D.17(a)) into a *left*-opened parabolic grid (Fig. 17(b)). But because of the inversion effect that is inherent to domain transformations, when this transformation is applied to the function $z = \cos(2\pi u)$, it bends its vertical corrugations into *right*-opened parabolas, in accordance with the effect of the inverse of this transformation, $(x,y) = (u + av^2, v)$, shown in Figs. D.17(c),(d). The resulting function, $z = \cos 2\pi(x - ay^2)$, corresponds, therefore, to our requirements, as shown in Fig. D.17(e).

Finally, the effects of the transformations \mathbf{g} and \mathbf{g}^{-1} as vector fields are shown in Figs. D.16(f) and D.16(h), respectively. ■

Example D.7: Suppose now we wish to construct a 2-fold counterpart of the previous example, which bends the vertical corrugations of $z = \cos 2\pi x$ into right-opened parabolas, as shown in Fig. D.17(e), and the horizontal corrugations of $z = \cos 2\pi y$ into top-opened parabolas. Using the same logic as at the end of the previous example, taking into account the inversion effect of domain transformations, it is clear that the transformation $\mathbf{g}(x,y)$ we seek here is defined by $(u,v) = (x - ay^2, y - ax^2)$. But although this transformation seems to be a straightforward generalization of the transformation $(u,v) = (x - ay^2, y)$ that is shown in Fig. D.17, it is in fact much more interesting. As we can see in Figs. D.18(a),(b), this transformation maps the vertical lines $x = m$, $m \in \mathbb{Z}$ into a family of shifted, left-opened parabolas, and the horizontal dashed lines $y = n$, $n \in \mathbb{Z}$ into a family of shifted, bottom-opened parabolas. However, it turns out that this transformation is quite unusual in several aspects: First, the individual parabolas within each family intersect each other. Moreover, as we can see in Fig. D.18(b), this transformation is neither surjective (its image does not cover the entire u,v plane) nor injective (it maps different points (x,y) to the same image (u,v)). And finally, as a consequence of all these pathologies, the inverse transformation does not have a simple explicit expression. Nevertheless, as we already know from Remark D.7, the effect of the inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ on the plane can be determined, without even knowing its explicit expression. And indeed, as clearly shown in Figs. D.18(c),(d), this inverse transformation maps the vertical lines $u = m$, $m \in \mathbb{Z}$ into the family of equispaced right-opened parabolas $x - ay^2 = m$, and the horizontal dashed lines $v = n$, $n \in \mathbb{Z}$ into the family of equispaced top-opened parabolas $y - ax^2 = n$.

Viewed the other way around, this also means, of course, that $\mathbf{g}(x,y)$ maps each of the parabolas of Fig. D.18(d) into the corresponding straight line in Fig. D.18(c). Note, however, that as the point $P = (x,y)$ traces out the parabola, its image $Q = (u,v)$ in Fig. D.18(c) moves along the corresponding straight line slower and slower, until it reaches

the border of the range of the transformation, where it stops and starts moving backwards along the same straight line.

In accordance with Remark D.6, the plain and dashed parabolas of Fig. D.18(d) are, respectively, the level lines of the surfaces $z = x - ay^2$ and $z = y - ax^2$ (the two components of the *direct* transformation \mathbf{g}). Similarly, in accordance with Remark D.5, the plain and dashed parabolas of Fig. D.18(b) are the level lines of the two surfaces defined by the *inverse* transformation \mathbf{g}^{-1} . However, since the level lines within each of these two surfaces intersect each other, it follows that neither of these surfaces is single valued: As shown in Fig. D.19, each of these surfaces may have 0, 1, 2, 3 or 4 z -values in different parts of the plane. In fact, each of these surfaces resembles a handkerchief that has been folded twice; see also [Callahan74 p. 232] or [Koenderink90 pp. 332–334].

Note that in this example neither \mathbf{g} nor \mathbf{g}^{-1} can be used to define a meaningful coordinate system, since these transformations are not one-to-one (see Remark D.9).

Now, as we already know, when we apply the transformation $(u,v) = \mathbf{g}(x,y)$ to a given function $z = f(u,v)$, the resulting function $z = f(\mathbf{g}(x,y))$ is distorted in accordance with the *inverse* transformation, \mathbf{g}^{-1} . For example, following the application of $\mathbf{g}(x,y)$ the straight vertical corrugations of $z = \cos(2\pi u)$ become in the resulting function, $z = \cos(2\pi[x - ay^2])$, a family of equispaced right-opened parabolic corrugations (see Fig. D.18(d)). And indeed, it is mainly for this inverse effect of the domain transformation (although we don't even know the explicit expression of \mathbf{g}^{-1}) that we are interested in the transformation $(u,v) = (x - ay^2, y - ax^2)$ in our work (see, for example, Figs. A.1 and 3.15).

Finally, the effect of the transformation \mathbf{g} as a vector field is shown in Fig. D.18(f). Note that this vector field clearly shows the two critical points of the transformation \mathbf{g} (i.e. the points where $\mathbf{g}(x,y) = (0,0)$; see Sec. H.1 in Appendix H), which are $(0,0)$ and $(1/a, 1/a)$. The critical points of a transformation \mathbf{g} are clearly revealed by its vector field representation, but they are not as easily seen in the other representations of \mathbf{g} . ■

It is interesting to note that the transformation $(u,v) = (x - ay^2, y - ax^2)$ of the last example belongs, in fact, to a larger family of transformations, whose graphical behaviour can be easily understood using the following result:

Proposition D.5: Suppose we are given a transformation $(u,v) = \mathbf{g}(x,y)$ that is defined by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x + f_1(y) \\ y + f_2(x) \end{pmatrix} \quad (\text{D.23})$$

where $z = f_1(s)$ and $z = f_2(s)$ are arbitrary 1D functions. Then, the geometric effects of the transformation \mathbf{g} and of its inverse \mathbf{g}^{-1} on the coordinate grid of the plane can be pictured as follows:

- (a) The inverse transformation $(x,y) = \mathbf{g}^{-1}(u,v)$ maps the vertical lines $u = m$, $m \in \mathbb{Z}$ of the u,v plane into a family of parallel curves $x = -f_1(y) + m$ in the x,y plane, that are unit-spaced copies of the curve $x = -f_1(y)$ shifted along the x axis. Similarly, \mathbf{g}^{-1} distorts the

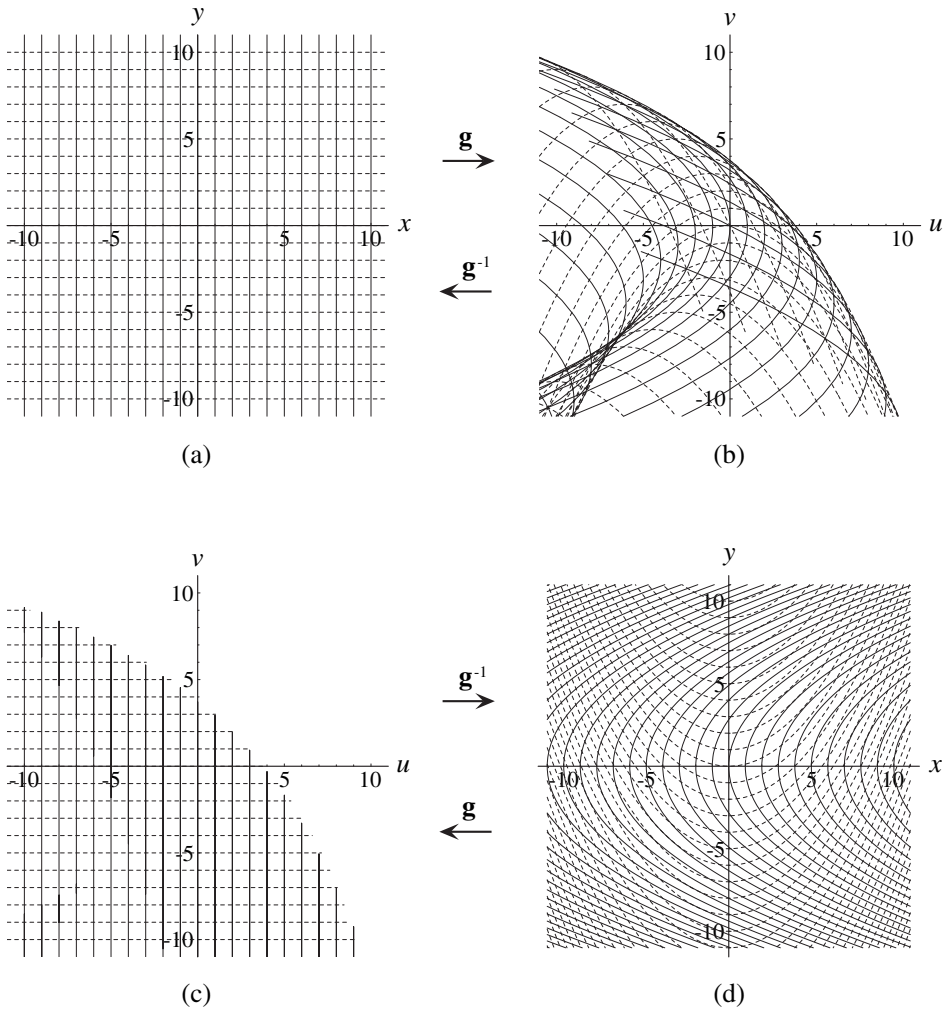


Figure D.18: (a),(b) The effect of the transformation $(u,v) = (x - ay^2, y - ax^2)$ on the unit grid: The vertical lines of the grid are mapped into a self-intersecting family of shifted, left-opened parabolas, while the horizontal lines of the grid are mapped into a self-intersecting family of shifted, bottom-opened parabolas. (c),(d) The effect of the inverse transformation on the unit grid: The vertical lines of the grid are mapped into a family of right-opened parabolas, while the horizontal lines of the grid are mapped into a family of top-opened parabolas. Note that the unprinted area of the u,v plane is never accessed by these transformations.

horizontal lines $v = n, n \in \mathbb{Z}$ of the u,v plane into a family of parallel curves $y = -f_2(x) + n$ in the x,y plane, that are unit-spaced copies of the curve $y = -f_2(x)$ shifted along the y axis.

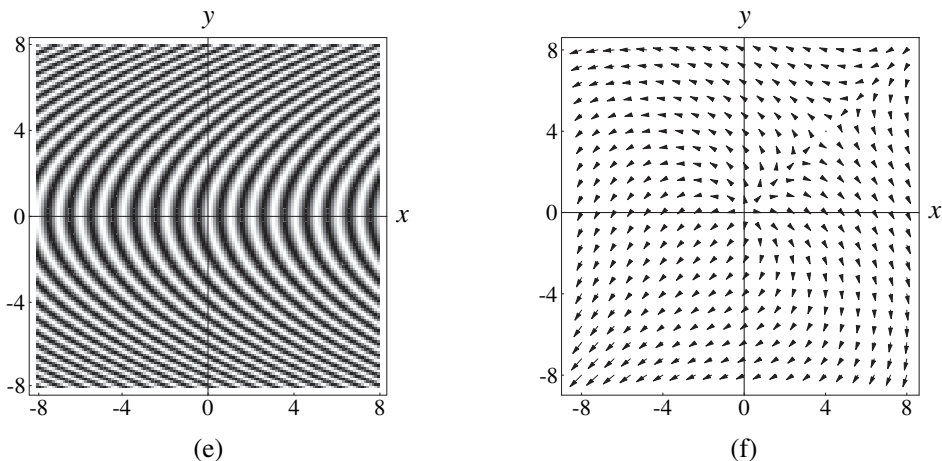


Figure D.18: (*continued.*) (e) The application of $(u,v) = (x - ay^2, y - ax^2)$ as a domain transformation to the vertical unit-period cosinusoidal grating $z = \cos(2\pi u)$ gives $z = \cos(2\pi[x - ay^2])$, a right-opened parabolic version of the original function; compare with the effect of the *inverse* transformation on the plain grid lines in (d). (f) The effect of the same transformation as a vector field. Note that in the present example the explicit form of the inverse transformation is not readily available.

- (b) The direct transformation $(u,v) = \mathbf{g}(x,y)$ maps the vertical lines $x = m$, $m \in \mathbb{Z}$ of the x,y plane into a family of curves $u = f_1(v - f_2(m)) + m$ in the u,v plane, that are parallel copies of the curve $u = f_1(v)$, and whose origins are centered at the points $(u,v) = (f_2(m), m)$ along the curve $v = f_2(u)$. Similarly, \mathbf{g} distorts the horizontal lines $y = n$, $n \in \mathbb{Z}$ of the u,v plane into a family of curves $v = f_2(u - f_1(n)) + n$ in the x,y plane, that are parallel copies of the curve $v = f_2(u)$, and whose origins are centered at the points $(u,v) = (f_1(n), n)$ along the curve $u = f_1(v)$. ■

The mathematical demonstration of this result is quite simple: Part (a) follows from the fact that for any $u = m$ we have from the first line of (D.23) $x = -f_1(y) + m$; and similarly, for any $v = n$ we have from the second line of (D.23) $y = -f_2(x) + n$. Part (b) follows from the fact that for any $x = m$ we have from (D.23) $u = f_1(y) + m$, $v = f_2(m) + y$, which gives $y = v - f_2(m)$ and hence $u = f_1(v - f_2(m)) + m$; similarly, for any $y = n$ we have from (D.23) $u = f_1(n) + x$, $v = f_2(x) + n$, which gives $x = u - f_1(n)$ and hence $v = f_2(u - f_1(n)) + n$.

Note that several of the transformations we have seen in the examples above have the form (D.23). This includes the transformations $(u,v) = (x - \arg \sinh(y), y + \arg \sinh(x))$ (see Fig. D.9), $(u,v) = (x + ay^2, y)$ (Fig. D.16), $(u,v) = (x - ay^2, y)$ (Fig. D.17) and $(u,v) = (x - ay^2, y - ax^2)$ (Fig. D.18). And indeed, the geometric behaviour of each of these transformations and of its inverse, as shown in the figures, is clearly explained by Proposition D.5 (even if the explicit form of the inverse transformation is not available).

D.9 Other possible sources of confusion

As we can see, the handling of spatial transformations or coordinate changes can be often quite confusing. In particular, it may be sometimes unclear whether one should use the direct transformation \mathbf{g} or its inverse \mathbf{g}^{-1} to obtain the desired effect. In many cases (like in Figs. D.11 and D.16) the penalty for using the wrong direction is just a sign inversion. Such sign inversions are often mistakenly attributed to a “forgotten sign”, and rectified by inverting the sign without even caring to understand why. But in other cases the difference between the results obtained by using \mathbf{g} or \mathbf{g}^{-1} is much more significant, and it cannot be simply dismissed as a “forgotten sign”. For example, the transformation $(u,v) = (2xy, y^2 - x^2)$ (see Example D.5 above) distorts the standard Cartesian grid into two families of parabolas as in Fig. D.15(b), while the inverse transformation distorts the standard Cartesian grid into two families of hyperbolas as in Fig. D.15(d). And indeed, it turns out that in some circumstances the effect of the transformation $(u,v) = (2xy, y^2 - x^2)$ is depicted by two families of parabolas (see, for example, [Spiegel68 p. 126]), while in other circumstances it is depicted by two families of hyperbolas (see, for example, [Spiegel63 p. 185] or Example 10.23 and Fig. 10.36 in Chapter 10 of *Vol. I*). In fact, both representations are correct — depending on the point of view and on the application: When $(u,v) = (2xy, y^2 - x^2)$ is used as a direct transformation it indeed distorts the original Cartesian grid into two families of parabolas, but when it is used as a domain (and hence inverse) transformation, it distorts the original Cartesian grid (or any other rectilinear structure) into two families of hyperbolas (see Proposition D.3). In fact, the same transformation may also have a third graphical representation when it is interpreted as a vector field; in this case its graphical representation resembles the physical illustration of a magnetic field (see Fig. D.15(f)). On the other hand, although polar coordinates are introduced in the literature either by $(r, \theta) = (\sqrt{x^2 + y^2}, \arctan(y/x))$ (see, for example, [Courant88 p. 138]), or by its inverse, $(x,y) = (r \cos \theta, r \sin \theta)$ (see, for example, [Kreyszig93 p. 672]), both of these transformations are systematically illustrated in the literature by Figs. D.14(a)(b), and rarely — if ever — by Figs. D.13(a),(b).

All these potential sources of confusion are intrinsic to the use of transformations and coordinate changes, and they originate from pure mathematical considerations. But as if these were not enough, there exist also several extraneous potential sources of confusion. One of these is related to the discrete representation of images on digital computers, and another is related to the existence of two different standards of notation in the literature. These additional sources of confusion are shortly described in the following subsections.

D.9.1 Forward and backward mapping algorithms in digital imaging

Although spatial transformations are basically continuous mathematical entities, in most modern applications they are implemented by digital computers, and they operate on digital, discrete images. This means that the input and the output spaces are represented as 2D (or more generally, N -dimensional) arrays of pixels lying on an integer grid. This discrete nature of digital images and of their transformations introduces several complica-

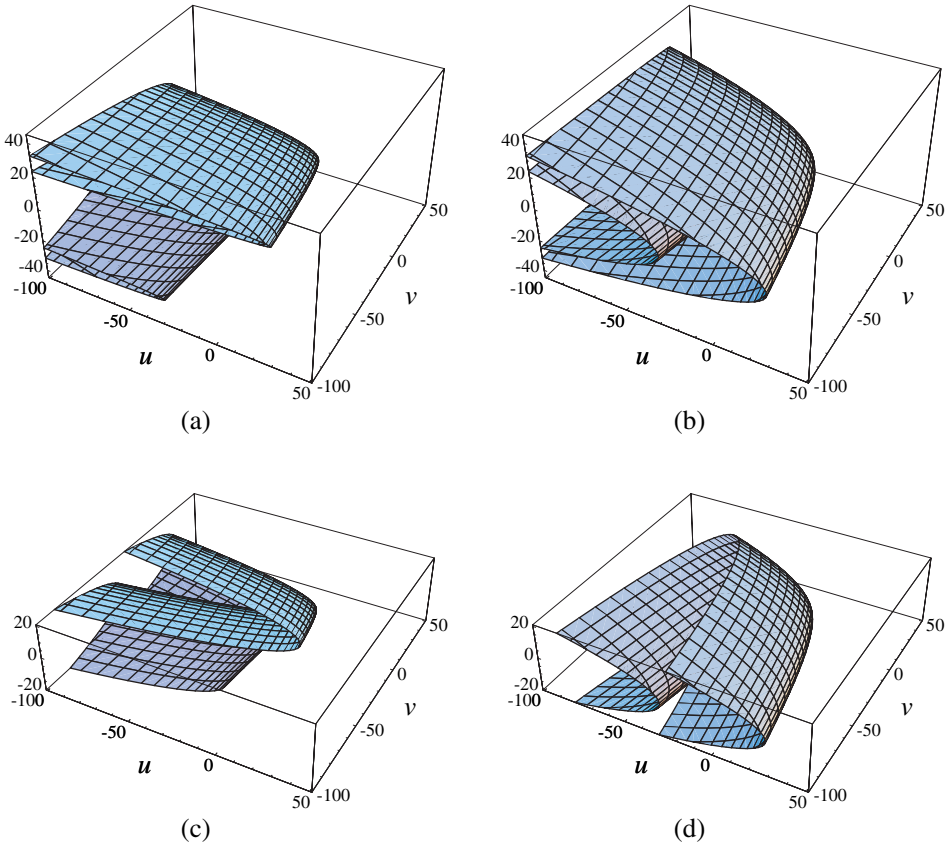


Figure D.19: The multivalued surfaces (a) $z = g_1^{-1}(u, v)$ and (b) $z = g_2^{-1}(u, v)$ of Example D.7. These are the surfaces whose level lines are shown, respectively, by the continuous and dashed parabolas of Fig. D.18(b). Note that for the sake of clarity the surfaces have been drawn for a larger span of u and v values than Fig. D.18(b). Parts (c) and (d) show the same surfaces as in the upper figures, but here they have been vertically truncated at the levels $z = 20$ and $z = -20$ in order to clearly show the parabolic shape of their level lines.

tions, that are addressed in various ways by the algorithms which perform digital transformations. These algorithms can be divided into two main families, known as forward mapping and backward mapping, each having its own advantages and shortcomings [Wolberg90 pp. 42–45].³⁵

A *forward mapping* operates by scanning the original input image pixel by pixel, and copying the value of the image at each input pixel location (x, y) onto the corresponding

³⁵ Note that these terms should not be confused with the terms *direct mapping* and *inverse mapping*, which refer to a transformation $(u, v) = \mathbf{g}(x, y)$ and its inverse $(x, y) = \mathbf{g}^{-1}(u, v)$.

position (u,v) in the output image. The output position (u,v) is determined by passing the x and y coordinates of each input pixel through the transformation:

$$(u,v) = \mathbf{g}(x,y)$$

Note, however, that in general the coordinates u and v thus obtained are real numbers even if x and y are integers. This would not be a problem if the output domain were continuous. But in our discrete case the positions u and v must be discretized (for example by rounding or truncation) in order to fit to the underlying discrete grid of the output image, so that every pixel in the input image can be copied into a pixel in the output image. This seemingly innocent discretization may give rise to two types of problems in the output image: holes and overlaps. Holes occur in the output image if contiguous pixels in the input image are mapped into sparse positions in the output grid, and overlaps occur when different input pixels happen to fall on the same destination pixel in the output image (see [Wolberg90 pp. 42–44] for a more detailed explanation).

Due to these as well as other shortcomings of forward mapping, the use of *backward mapping* is more common in most digital image applications. A backward mapping operates by scanning the target output image pixel by pixel, and copying onto each output pixel location (u,v) the value of the input pixel at the corresponding (x,y) position in the input image. The input position (x,y) that corresponds to a given output pixel location is determined by passing the u and v coordinates of each output pixel through the *inverse* transformation:

$$(x,y) = \mathbf{g}^{-1}(u,v)$$

Just as in the previous case, the continuous x and y locations thus obtained are, in general, real numbers, even though u and v are integers. This means that, here too, the values x and y must be discretized in order to fit to the underlying discrete grid of the input image.³⁶ Note, however, that unlike in the forward mapping scheme, the backward mapping scheme guarantees that all output pixels are computed. For this reason, backward mapping proves to be a much more convenient approach than forward mapping, although it requires that the explicit expression of the inverse transformation \mathbf{g}^{-1} be available. Another advantage of backward mapping that is often overlooked in the literature is that it behaves exactly as a domain transformation, meaning that if our input image corresponds to the function $z = f(x,y)$, the resulting output image after applying the transformation is exactly $f(\mathbf{g}(x,y))$.

This last point brings us back to the potential confusion which may occur due to the use of forward or backward algorithms in digital imaging. Suppose we are given two different image-processing programs, one based on a forward-mapping algorithm and the other based on a backward-mapping algorithm. If we use in both programs the *same*

³⁶ In fact, a better solution consists of interpolating between the values of the input image at the pixel locations surrounding the non-integer location (x,y) , in order to determine the theoretic value of the underlying continuous input image at the point (x,y) .

transformation $\mathbf{g}(x,y)$, say, $(u,v) = (2xy, y^2 - x^2)$, we will get two completely different transformed images: one will look like Fig. D.15(b), and the other will look like Fig. D.15(d). The situation is even worse if we do not know which type of algorithm is implemented in the software being used, since then we may obtain the wrong result without even being aware of the risk. In cases such as rotation transformations the difference is merely in the direction of the rotation, so that we may be tempted to “rectify” the error by inverting the sign of a variable within the program. But this “remedy” will not help in other cases; on the contrary, it may even worsen the situation when other transformations will be used.

Finally, as already mentioned, it may happen in some cases that the explicit formula of the inverse transformation $\mathbf{g}^{-1}(u,v)$ is unknown or unavailable. In such cases the use of a backward mapping scheme for applying \mathbf{g} to the input image is not possible. However, in such cases one can still apply \mathbf{g} to the input image by supplying \mathbf{g} itself to a program which is based on a forward mapping algorithm.

Remark D.20: The layer transformations in most of the figures throughout this volume have been obtained by a *forward-mapping algorithm*, a PostScript program that moves each dot of the given dot screen (periodic or not) from its original location (x,y) to its new location $(u,v) = \mathbf{g}(x,y)$. However, some of the figures — notably those that show the transformation of more complicated images — have been obtained by a *backward-mapping algorithm*, a C language program that takes an input image $f(x,y)$ and generates its distorted version $f(\mathbf{g}(x,y))$. In all cases care has been taken to supply to the program in question the right transformation, either \mathbf{g} or \mathbf{g}^{-1} , in order to obtain the intended effect in the figure. Note, however, that our figure legends do not usually mention the implementation-dependent technicalities that were involved in the figure-generation process, and they only concentrate on the mathematical and visual effects that are demonstrated by each figure. ■

We will return to some of these issues in more detail in Sec. D.10.

D.9.2 Pre-multiplication and post-multiplication based notations

The last source of confusion we discuss here concerns transformations that can be expressed in matrix form. In addition to linear transformations, this includes also affine, bilinear, and some other types of transformations [Wolberg90 pp. 45–61]. The problem arises here from the two different standard notations that are widely used in the literature for expressing such transformations, using pre-multiplication or post-multiplication form [Lipschutz68 p. 157]. In the first case the transformation matrix precedes the vector to which it is applied:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \quad (\text{D.24})$$

while in the second case the transformation matrix is written after the vector:

$$(u,v) = (x,y) \begin{pmatrix} e & f \\ g & h \end{pmatrix} \quad (\text{D.25})$$

Both notation standards are fully equivalent, provided that we consistently stick to the same notation. It should be noted, however, that the matrix representation of a given transformation is different in each of these two notation standards, so when we copy a formula from a book, we must carefully check whether the copied matrix agrees with our notation system. For example, in the present work we always use the pre-multiplication notation (Eq. (D.24)); in this notation the matrix form of a rotation by positive angle α is given by:

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}$$

and the inverse transformation is given by:

$$\begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}$$

However, in books using the post-multiplication notation these matrices are interchanged (see, for example, [Wolberg90 pp. 49 and 217]). The rule is that if we wish to use a matrix taken from the other notation standard, we must first take the transpose of the matrix, as illustrated below:

$$\begin{aligned} \begin{pmatrix} u \\ v \end{pmatrix} &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} ax + by \\ cx + dy \end{pmatrix} \\ (u,v) &= (x,y) \begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = (x,y) \begin{pmatrix} a & c \\ b & d \end{pmatrix} = (ax + by, cx + dy) \end{aligned}$$

D.10 Implications to the moiré theory: issues related to the figures

In the previous sections of this appendix we have presented various results about *direct*, *inverse*, *domain* and *range* transformations, as well as about *forward* and *backward-mapping* algorithms in software applications; but the reasons that we actually need all these results in our work may still be unclear. In the present section and in the following one we therefore try to show some of the actual implications of these results to the moiré theory: In the present section we will see their practical implications to the generation of our moiré figures, and in the following section we will see the implications to the fixed points between the superposed layers.

As we already know (see Remark 4.1 in Chapter 4), the classical moiré theory between periodic or repetitive layers is based, as a rule, on the interpretation of the layer transformations as *domain* (and hence, *inverse*) transformations. This rule holds, for example, in the explanation of the moiré effects that are obtained when we apply the transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ to the original periodic layers $p_1(x',y')$ and $p_2(x',y')$,

giving the distorted layers $p_1(\mathbf{g}_1(x,y))$ and $p_2(\mathbf{g}_2(x,y))$ (see Propositions 10.2 and 10.5 in *Vol. I*). In fact, domain transformations are used throughout Chapters 10 and 11 of *Vol. I*, which present the theory of curvilinear moiré effects between geometrically transformed periodic layers (see Sec. 10.2 and in particular Footnote 1 and Remark 10.2 there, as well as Secs. 11.2.2, 11.3 and 11.4, all in *Vol. I*). Domain transformations are also used in the case of aperiodic layers, as we have seen throughout the present volume, with just one outstanding exception: When it comes to dot trajectories, the transformation $\mathbf{g}(x,y)$ that is applied to a dot screen is understood as an operation that moves each point (x,y) of the original dot screen to its new destination under \mathbf{g} , namely: $(x,y) \mapsto \mathbf{g}(x,y)$. This means, as explained in Chapter 4, that in the context of dot trajectories we are interested in the effect of \mathbf{g} as a *direct* transformation. (Note that in *Vol. I*, too, there exist some circumstances where the transformations $\mathbf{g}_i(x,y)$ are not used as domain transformations, and the expressions involved are formulated in terms of direct rather than inverse transformations; see, for example, Eq. (3.1) and its vector form in Sec. 3.4.1 of *Vol. I*.)

However, the fact that we need to use the same transformation alternately as a domain transformation or as a direct transformation, depending on the context, may cause us trouble in the generation and in the interpretation of our figures. The following examples will help us understand this point.

Example D.8: Consider the following cases, which illustrate the use of the inverse mapping as a domain transformation:

- (a) Suppose we are given a layer $p(x',y')$. In order to rotate it by positive angle α counterclockwise we have to apply to $p(x',y')$ the following domain transformation:

$$x' = x \cos \alpha + y \sin \alpha$$

$$y' = -x \sin \alpha + y \cos \alpha$$

which corresponds, in fact, to the inverse transformation (rotation by angle $-\alpha$; see Example D.2 above). For instance, in order to rotate the vertical cosinusoidal grating $\cos(2\pi f x')$ by angle α (see Fig. D.11(g)) we have to plot the function:

$$r(x,y) = \cos(2\pi f [x \cos \alpha + y \sin \alpha])$$

where the original variable x' has been replaced by $x \cos \alpha + y \sin \alpha$.

- (b) Bending a periodic layer $p(x',y')$ into a right-opened parabolic shape is obtained by applying to $p(x',y')$ the domain transformation:

$$x' = x - ay^2$$

$$y' = y$$

which corresponds, in fact, to the inverse transformation (see Example D.6 above). Taking again the example of the vertical cosinusoidal grating $\cos(2\pi f x')$, we can bend it

into a right-opened parabolic cosinusoidal grating (see Fig. D.16(g)) by plotting the function:

$$r(x,y) = \cos(2\pi[x - ay^2])$$

Another example consists of the distorted periodic and aperiodic layers shown in Figs. 3.5(a),(b), which are expressed mathematically by $r(x,y) = p(x - ay^2, y)$.

- (c) Bending a two-fold periodic dot screen $p(x',y')$ into a 2D hyperbolic dot screen is obtained by applying to $p(x',y')$ the domain transformation:

$$x' = 2xy$$

$$y' = y^2 - x^2$$

(see Example D.5 above). The resulting geometrically transformed dot screen is expressed, therefore, by $p(2xy, y^2 - x^2)$ (see Figs. B.7(a),(b) in Appendix B). ■

Naturally, it would be best to generate all the distorted gratings and dot screens in the figures that accompany our discussions on the classical moiré theory by an algorithm based on backward mapping. As we have seen in Sec. D.9.1, backward-mapping algorithms operate on the original image exactly as a domain transformation does, and therefore the explicit expressions that we have to provide to the algorithm are identical to those used in our equations, and the resulting moirés obtained in the figures fully correspond to our mathematical results as we indeed obtain them in terms of domain transformations. However, in reality three different problems may occur:

Problem 1: Because in the context of dot trajectories (in Chapter 4) we need to express the transformed layers in terms of direct transformations while elsewhere (in Chapters 3, 6 and 7) we need to express the same transformed layers in terms of domain (and hence, inverse) transformations, we are facing a dilemma in our figure legends: Given that each figure can be used in all chapters, which of the two expressions should we present in the figure legend? For example, taking the parabolically distorted dot screens shown in Fig. 3.5, should we say in the figure legend that they were obtained by applying the transformation $\mathbf{g}(x,y) = (x - ay^2, y)$, as we would do in Chapters 3, 6 and 7 — or by applying the transformation $\bar{\mathbf{g}}(x,y) = (x + ay^2, y)$, as we would do in Chapter 4? The solution we have adopted is to provide in the figure legends just a verbal description of the transformations, without giving their explicit formulas. But when we still wish to add the formulas, the convention we adopt in the figure legends throughout this volume corresponds to the normal usage of transformations in the classical moiré theory as *domain* transformations. In cases where the transformation should be understood as a *direct* transformation, we either say it explicitly, or use the barred notation $\bar{\mathbf{g}}(x,y)$, as explained in Sec. 4.4, Remark 4.1.

Problem 2: When we generate our figures using a program based on a backward-mapping algorithm the transformations we have to supply to the program correspond to those being used as domain transformations (in Chapters 3, 6 and 7), but not to those

being used as direct transformations (in Chapter 4). Therefore, in order to generate using such a backward-mapping algorithm figures that come to illustrate dot trajectories, as in Chapter 4, we must supply to the program the *inverse* of the transformations being used in the text. Note, however, that this is just a technical problem that should be taken into account when generating the figures, but we can forget about it once the figures are done.

Of course, if we generate our figures using a program that is based on a forward-mapping algorithm, the inverse problem will occur. And indeed, as mentioned in Remark D.20, most of the transformed dot screens shown in the figures throughout the present volume were produced, for technical reasons, by an algorithm based on *forward mapping*. This algorithm uses the most straightforward and natural approach for distorting a random or periodic dot screen: it simply maps each dot of the original, undistorted dot screen from its original location (x,y) to the new, transformed location $(u,v) = \mathbf{g}(x,y)$.³⁷ This forces us, in order to obtain the correct results in the figures, to provide to the forward-mapping program the *direct* transformation of each of the layers, namely, the inverse of the transformation \mathbf{g} that is used in the mathematical expression $p(\mathbf{g}(x,y))$. But once again, this technicality in the figure generation does not affect our mathematical developments, and once the figures are ready we continue to assume that the transformed layers in question have been obtained by applying \mathbf{g} as a domain (and hence inverse) transformation, as usual.

Example D.9: The following cases illustrate the use of direct transformations in the forward-mapping program that generates most of the layer superpositions in this volume:

- (a) In order to rotate a dot screen by positive angle α counterclockwise we have to provide to the forward-mapping program the following direct transformation:

$$u = x \cos \alpha - y \sin \alpha$$

$$v = x \sin \alpha + y \cos \alpha$$

This means that a dot that should have been printed in the original, untransformed dot screen at the location (x,y) will be printed in reality at the location (u,v) as defined above (see, for example, Fig. 2.1(d)). And yet, assuming that the original, undistorted dot screen is given by $p(x',y')$, the rotated dot screen continues to be expressed mathematically by $p(x \cos \alpha + y \sin \alpha, -x \sin \alpha + y \cos \alpha)$, using the *inverse* mapping $(x',y') = (x \cos \alpha + y \sin \alpha, -x \sin \alpha + y \cos \alpha)$ as a domain transformation.

- (b) In order to bend an original dot screen into a right-opened parabolic dot screen we have to provide to the forward-mapping program the following direct transformation:

$$u = x + ay^2$$

$$v = y$$

³⁷ The reason for using a forward-mapping algorithm to generate the figures in this volume is basically practical. Note that in the case of random screens the use of a backward-mapping algorithm may be more difficult.

(see, for example, Fig. 3.5). And yet, assuming that the original, undistorted dot screen is given by $p(x',y')$, the distorted dot screen of the resulting figure is expressed mathematically, as explained in Sec. 3.4.1, by $p(x-ay^2, y)$, using the *inverse* mapping $(x',y') = (x-ay^2, y)$ as a domain transformation.

- (c) Finally, if we bend an original dot screen $p(u,v)$ by using in the forward-mapping program the transformation:

$$u = 2xy$$

$$v = y^2 - x^2$$

the resulting dot screen will be distorted into a *parabolic* shaped screen (see Fig. D.15(b)). If we wish to obtain the *hyperbolic* screen which is expressed by $p(2xy, y^2 - x^2)$, as shown in Fig. D.15(d) or in Fig. B.7 of Appendix B, we have to provide to our forward-mapping program the inverse transformation, i.e. the rather complicated expression with the nested roots that is given in Example D.5 above. ■

Problem 3: The fact that the transformations used to draw the figures are sometimes inversed with respect to the conventions used in the theoretic results may raise an additional problem. According to Proposition 5.1 in Sec. 5.3 (see also the second part of Proposition 10.5 in Sec. 10.9.2 of *Vol. I*), when we apply the domain transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$, respectively, to two periodic layers, the resulting (1,-1)-moiré in the layer superposition undergoes the domain transformation $\mathbf{g}_m(x,y) = \mathbf{g}_1(x,y) - \mathbf{g}_2(x,y)$. Thus, in order to obtain in the layer superposition a (1,-1)-moiré having the domain transformation $\mathbf{k}(x,y)$, we have to apply to the two original layers, respectively, such domain transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ that give the difference $\mathbf{k}(x,y)$; for example $\mathbf{g}_1(x,y) = (x,y) + \mathbf{k}(x,y)$ and $\mathbf{g}_2(x,y) = (x,y)$, or alternatively, $\mathbf{g}_1(x,y) = (x,y) + \frac{1}{2}\mathbf{k}(x,y)$ and $\mathbf{g}_2(x,y) = (x,y) - \frac{1}{2}\mathbf{k}(x,y)$ (see also Eqs. (3.41) and (3.42)). This would pose no problems if we were to draw the figure using a program that is based on a backward-mapping algorithm. But as we have seen in Problem 2 above, in order to draw these distorted layers using our forward-mapping algorithm we have to provide to our program the inverse expressions $[(x,y) + \mathbf{k}(x,y)]^{-1}$ and $[(x,y)]^{-1} = (x,y)$, or $[(x,y) + \frac{1}{2}\mathbf{k}(x,y)]^{-1}$ and $[(x,y) - \frac{1}{2}\mathbf{k}(x,y)]^{-1}$; but these expressions may turn out to be too complex or simply unavailable.

What happens, then, if we “forget” the inversions and provide to the forward-mapping program the known, non-inverted expressions (namely, $(x,y) + \mathbf{k}(x,y)$ and (x,y) , or $(x,y) + \frac{1}{2}\mathbf{k}(x,y)$ and $(x,y) - \frac{1}{2}\mathbf{k}(x,y)$)? The respective domain transformations become, in this case, $[(x,y) + \mathbf{k}(x,y)]^{-1}$ and $[(x,y)]^{-1} = (x,y)$, or $[(x,y) + \frac{1}{2}\mathbf{k}(x,y)]^{-1}$ and $[(x,y) - \frac{1}{2}\mathbf{k}(x,y)]^{-1}$. Hence, the differences between the layer’s domain transformations become, respectively:

$$[(x,y) + \mathbf{k}(x,y)]^{-1} - (x,y) \quad (\text{D.26})$$

$$\text{and:} \quad [(x,y) + \frac{1}{2}\mathbf{k}(x,y)]^{-1} - [(x,y) - \frac{1}{2}\mathbf{k}(x,y)]^{-1} \quad (\text{D.27})$$

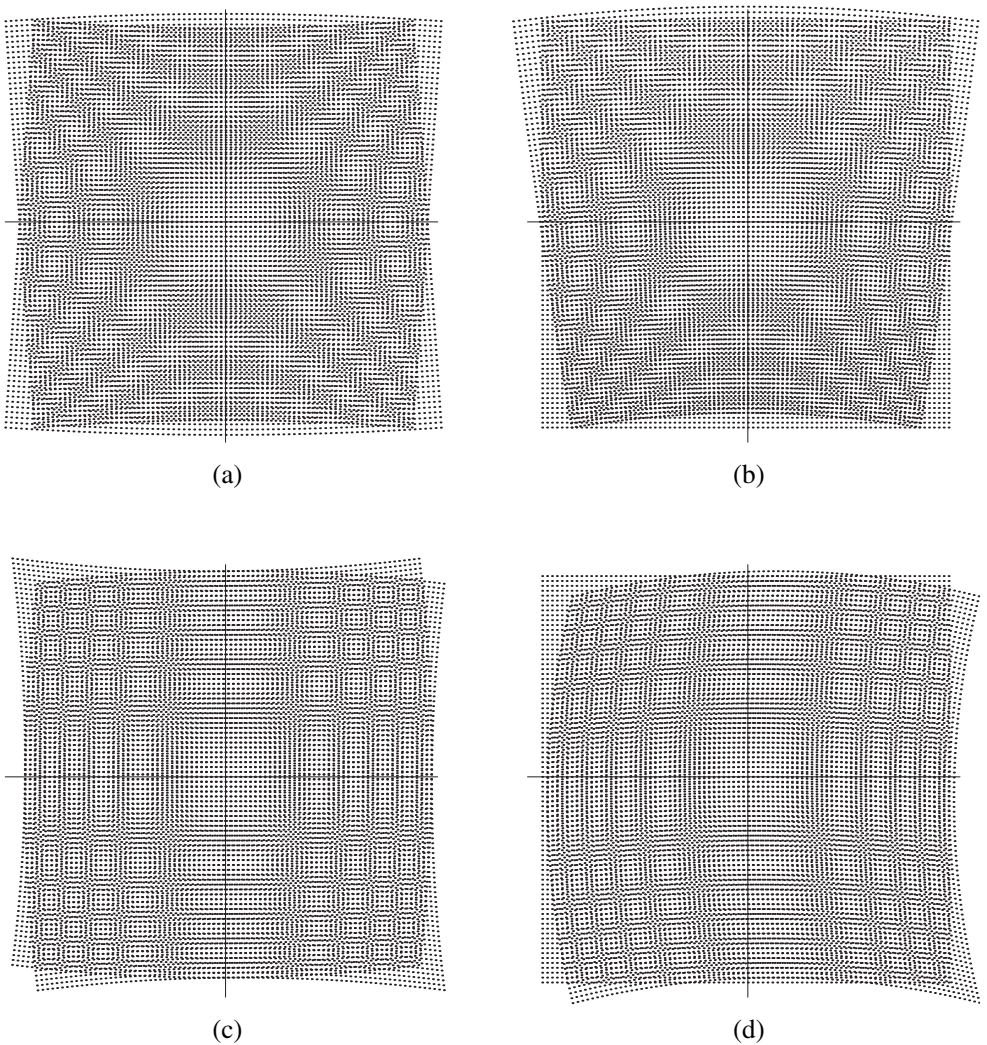


Figure D.20: The effects of using forward mapping to transform the original layers in a screen superposition. (a) The moiré effect obtained in Fig. 4.13(b), which is drawn using forward mapping, when the transformation $\mathbf{k}(x,y) = (2xy, y^2 - x^2)$ is equally distributed between the two layers. This moiré effect fully agrees with the results that are obtained when the layers are drawn by backward mapping. (b) When the transformation $\mathbf{k}(x,y)$ is entirely taken care of by the first layer, the resulting moiré is significantly distorted. (c),(d) Same as (a) and (b), but this time using the transformation $\mathbf{k}(x,y) = (ay^2 + x_0, y_0 - ax^2)$, a slightly different variant of Fig. 3.15(b). Note that the same phenomenon occurs also in the aperiodic counterparts of these superpositions, but the distortions in the aperiodic case are less visible.

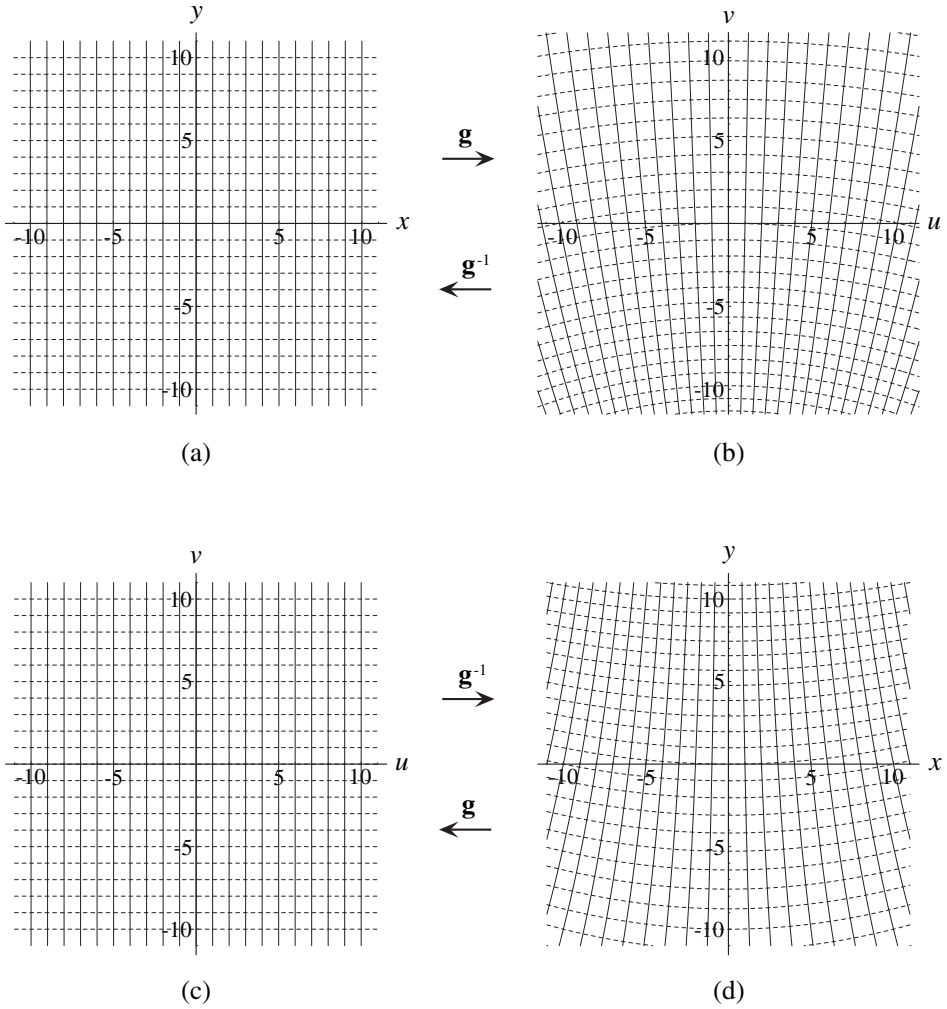


Figure D.21: The effect of the transformation $\mathbf{g}(x,y) = (x,y) + \mathbf{k}(x,y)$ where $\mathbf{k}(x,y)$ is a very weak non-linear transformation, $\mathbf{k}(x,y) = (2xy, y^2 - x^2)/100$. Note that $\mathbf{k}(x,y)$ is in fact a strongly diluted version of the non-linear transformation shown in Fig. D.15, and its influence here on the underlying linear transformation $\mathbf{g}_1(x,y) = (x,y)$ is very moderate. Therefore, the inverse of $\mathbf{g}(x,y)$, shown in (d), is almost identical to $(x,y) - \mathbf{k}(x,y)$: $[(x,y) + \mathbf{k}(x,y)]^{-1} \approx (x,y) - \mathbf{k}(x,y)$.

According to our proposition, these differences express the domain transformation of the resulting moiré in the layer superposition. But clearly, none of these differences can be expected to give the desired result, $\mathbf{k}(x,y)$, as was the case when backward mapping was used and no inversions were needed (see Eqs. (3.41) and (3.42)):

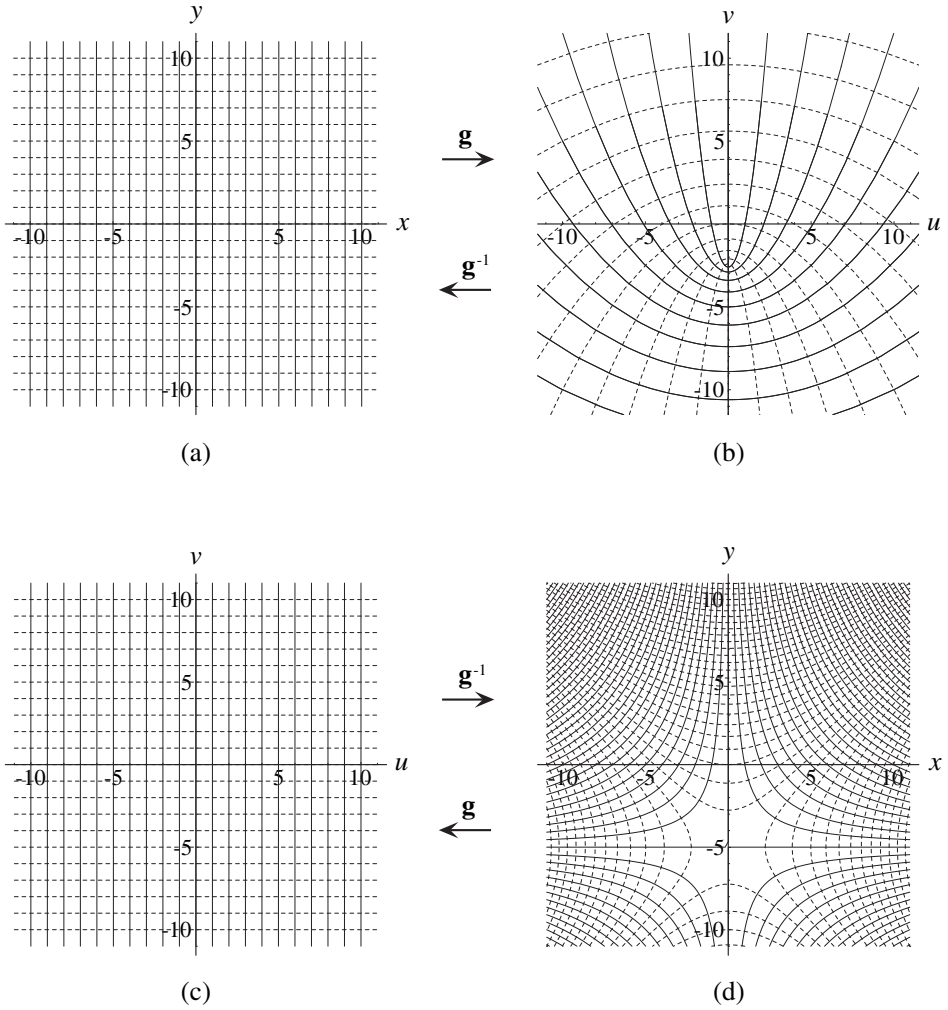


Figure D.22: Same as in Fig. D.21, but here the non-linearity introduced by $\mathbf{k}(x, y)$ is ten times higher: $\mathbf{k}(x, y) = (2xy, y^2 - x^2)/10$. Note that in this case the approximation $[(x, y) + \mathbf{k}(x, y)]^{-1} \approx (x, y) - \mathbf{k}(x, y)$ is no longer valid.

$$[(x, y) + \mathbf{k}(x, y)] - (x, y) = \mathbf{k}(x, y)$$

$$[(x, y) + \tfrac{1}{2}\mathbf{k}(x, y)] - [(x, y) - \tfrac{1}{2}\mathbf{k}(x, y)] = \mathbf{k}(x, y)$$

However, surprisingly, superposition tests with various transformations show that when forward mapping is being used to draw the layers, the difference (D.27), which corresponds to an equal distribution of the distortion between the two layers, *does* give in the layer superposition the same moiré effect $\mathbf{k}(x, y)$ as in the case of backward mapping

(up to a sign inversion). Even more surprisingly, it turns out that this occurs *only* if $\mathbf{k}(x,y)$ is equally distributed between the two layers, as in Eq. (D.27). When the distortion is not equally distributed, for instance in the case of Eq. (D.26), the moiré effect obtained with forward mapping is not $-\mathbf{k}(x,y)$, but rather a distorted version of $-\mathbf{k}(x,y)$ (see Figs. D.20(b),(d); note that in this case $-\mathbf{k}(x,y) = \mathbf{k}(x,y)$). In other words: the expected moiré $\mathbf{k}(x,y) = \mathbf{g}_1(x,y) - \mathbf{g}_2(x,y)$, as predicted by Proposition I.10.5, *can* be obtained in the superposition (up to a sign inversion) even if we perform the two layer transformations $\mathbf{g}_i(x,y)$ by means of forward mapping, but this happens only if the deformation $\mathbf{k}(x,y)$ is equally distributed between the two original, undistorted layers. This, however, seems to contradict Proposition I.10.5, which guarantees that the resulting moiré effect must have the shape of the difference $\mathbf{k}(x,y) = \mathbf{g}_1(x,y) - \mathbf{g}_2(x,y)$ in *all* cases, no matter how $\mathbf{k}(x,y)$ is distributed between $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$. How can it be?

These surprising results are explained as follows: Remember that in our case we only use weak layer transformations $\mathbf{g}(x,y)$, in order not to destroy the correlation between the superposed layers. This means that our layer transformations $\mathbf{g}(x,y)$ are of the form:

$$\mathbf{g}(x,y) = (x,y) - \mathbf{o}(x,y)$$

or:
$$\mathbf{g}(x,y) = (x,y) + \mathbf{o}(x,y)$$

where $\mathbf{o}(x,y)$ is a negligible transformation that differs only slightly, within the zone covered by the layer superposition, from the zero transformation $\mathbf{z}(x,y) = (0,0)$.

And indeed, it turns out (see Proposition D.9 in Sec. D.12 below) that if $\mathbf{g}(x,y)$ is such a weak layer transformation then we have for its inverse, $\mathbf{g}^{-1}(x,y)$, the following close approximations:

$$[(x,y) - \mathbf{o}(x,y)]^{-1} \approx (x,y) + \mathbf{o}(x,y)$$

$$[(x,y) + \mathbf{o}(x,y)]^{-1} \approx (x,y) - \mathbf{o}(x,y)$$

or, in a more compact notation, denoting the identity transformation by \mathbf{i} :

$$[\mathbf{i} - \mathbf{o}]^{-1} \approx \mathbf{i} + \mathbf{o} \quad (\text{D.28})$$

$$[\mathbf{i} + \mathbf{o}]^{-1} \approx \mathbf{i} - \mathbf{o} \quad (\text{D.29})$$

(see also Figs. D.21 and D.22). Furthermore, Proposition D.9 also asserts that the approximation error \mathbf{e}_1 in (D.28) and the approximation error \mathbf{e}_2 in (D.29) (both of which are functions of x and y , $\mathbf{e}_1(x,y)$, $\mathbf{e}_2(x,y)$) are almost identical. Therefore, denoting this common approximate error by $\mathbf{e}(x,y)$ we obtain for Eq. (D.27):

$$\begin{aligned} & [(x,y) + \tfrac{1}{2}\mathbf{k}(x,y)]^{-1} - [(x,y) - \tfrac{1}{2}\mathbf{k}(x,y)]^{-1} = \\ & = [(x,y) - \tfrac{1}{2}\mathbf{k}(x,y) + \mathbf{e}(x,y)] - [(x,y) + \tfrac{1}{2}\mathbf{k}(x,y) + \mathbf{e}(x,y)] \\ & = -\mathbf{k}(x,y) \end{aligned} \quad (\text{D.30})$$

whereas for Eq. (D.26) we have:

$$\begin{aligned}
 [(x,y) + \mathbf{k}(x,y)]^{-1} - (x,y) &= \\
 &= [(x,y) - \mathbf{k}(x,y) + \mathbf{e}(x,y)] - (x,y) \\
 &= -\mathbf{k}(x,y) + \mathbf{e}(x,y)
 \end{aligned} \tag{D.31}$$

This explains, indeed, why, when providing to the forward-mapping algorithm the *inverse* layer transformations rather than the *direct* ones, we still obtain the correct moiré effect if the distortion $\mathbf{k}(x,y)$ is equally distributed between the two superposed layers; but if the distortion is only taken care of by one of the two layers, the resulting moiré is rather distorted, as shown in Fig. D.20. Note that Eq. (D.30) also explains why the obtained moiré is identical (up to a sign inversion) to the moiré $\mathbf{k}(x,y)$ which is expected when using a backward-mapping algorithm, and not to its inverse, $[\mathbf{k}(x,y)]^{-1}$.

D.11 Fixed points of a superposition in terms of direct or inverse transformations

In this section we provide several useful results that can help us better understand our discussions on fixed points in layer superpositions. We start with the simpler case which occurs when only one of the superposed layers is transformed, and then we proceed to the more general case where both layers are transformed.

D.11.1 Fixed points when only one layer is transformed

Suppose we are given two identical dot screens $r_1(x,y)$ and $r_2(x,y)$, periodic or not, that are superposed on top of one another in full coincidence, dot on dot. We apply to one of the layers (say, the first one) a transformation $\mathbf{g}(x,y)$; as a result, the original layer $r_1(x,y)$ changes into $r_1(\mathbf{g}(x,y))$.

Let $\mathbf{g}_m(x,y)$ denote the transformation that is undergone by the resulting moiré. We have seen in Chapter 3 that:

$$\mathbf{g}_m(x,y) = \mathbf{g}(x,y) - (x,y) \tag{D.32}$$

(which is exactly the same result as in the case of the (1,-1)-moiré between periodic or repetitive layers; see Propositions 10.2 and 10.5 in Sec. 10.9 of *Vol. I*). Note that $\mathbf{g}(x,y)$ and $\mathbf{g}_m(x,y)$ are understood here as the *domain* transformations undergone by the layer $r_1(x,y)$ and by the resulting moiré, respectively; this means that both $\mathbf{g}(x,y)$ and $\mathbf{g}_m(x,y)$ are used here as *inverse* transformations.

As we have seen in Chapter 3, thanks to the layer transformation $\mathbf{g}(x,y)$ one or more Glass patterns may be generated in the layer superposition $r_1(\mathbf{g}(x,y))r_2(x,y)$. Typically, a Glass pattern occurs around each of the fixed points of the layer superposition, namely, around each point (x,y) that satisfies $\mathbf{g}_m(x,y) = (0,0)$, or equivalently:

$$\mathbf{g}(x,y) = (x,y) \quad (\text{D.33})$$

And yet, the same layer transformation $\mathbf{g}(x,y)$ can be also expressed in terms of its effect on the locations of the individual screen dots, as we do in Chapter 4. Clearly, each point (x,y) of the original layer $r_1(x,y)$ is moved by our transformation to a new location $\bar{\mathbf{g}}(x,y)$:

$$(x,y) \mapsto \bar{\mathbf{g}}(x,y)$$

However, in this case the effect of the layer transformation is understood as a *direct* transformation, meaning that $\bar{\mathbf{g}}(x,y)$ is, in fact, $\mathbf{g}^{-1}(x,y)$, the *inverse* of the domain transformation $\mathbf{g}(x,y)$ (the reasons for using the barred notation to denote direct transformations are explained in Sec. 4.4, Remark 4.1). For example, suppose that we are given the original dot screen $r_1(x,y)$. If we apply to it the domain transformation $\mathbf{g}(x,y) = (x/2, y/2)$ we obtain the transformed screen $r_1(x/2, y/2)$, which is a two-fold expanded version of the original screen $r_1(x,y)$. But the effect of this two-fold expansion on each individual dot of the screen is expressed by means of the *direct* transformation $\bar{\mathbf{g}}(x,y) = (2x, 2y)$:

$$(x,y) \mapsto (2x, 2y)$$

which is the inverse of the transformation $\mathbf{g}(x,y) = (x/2, y/2)$.

Now, as explained in Sec. 4.4, the effect of the direct transformation $\bar{\mathbf{g}}(x,y)$ on any individual dot is best described by the vector field:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}(x,y) - (x,y) \quad (\text{D.34})$$

This vector field assigns to each point (x,y) the vector $\bar{\mathbf{g}}(x,y) - (x,y)$ that connects (x,y) to its destination $\bar{\mathbf{g}}(x,y)$ under the transformation $\bar{\mathbf{g}}$.

As we can see, Eqs. (D.32) and (D.34) represent the same physical reality in two different ways: Eq. (D.32) is expressed in terms of inverse transformations, while Eq. (D.34) is expressed in terms of direct transformations. It is important to note, however, that although $\bar{\mathbf{g}}(x,y)$ and $\mathbf{g}(x,y)$ are mutually inverse transformations, $\bar{\mathbf{h}}(x,y)$ is not the inverse of $\mathbf{g}_u(x,y)$.

Now, based on Eq. (D.34), it is clear that we can also define the fixed points of the superposition as those points (x,y) that satisfy $\bar{\mathbf{h}}(x,y) = (0,0)$, or equivalently:

$$\bar{\mathbf{g}}(x,y) = (x,y) \quad (\text{D.35})$$

We have obtained, therefore, two different ways to express the fixed points in the resulting layer superposition: On the one hand we say in Eq. (D.33), just as we did in Chapter 3, that the fixed points of the superposition are those points (x,y) for which $\mathbf{g}(x,y) = (x,y)$; but on the other hand, based on different considerations, we say in Eq. (D.35) that the fixed points of the same superposition are those points for which $\bar{\mathbf{g}}(x,y) = (x,y)$. This seems to be incoherent, since obviously $\mathbf{g}(x,y)$ and $\bar{\mathbf{g}}(x,y)$ (i.e. $\mathbf{g}^{-1}(x,y)$)

are different transformations. But fortunately, the following result comes here to the rescue:

Proposition D.6: Let $\mathbf{g}(x,y)$ be a 2D transformation, and let $\mathbf{g}^{-1}(x,y)$ be its inverse (we assume here, as usual, that $\mathbf{g}(x,y)$ is sufficiently well-behaved, and that it satisfies all the conditions required in order to be invertible). Then, any fixed point of \mathbf{g} is also a fixed point of \mathbf{g}^{-1} , and any fixed point of \mathbf{g}^{-1} is also a fixed point of \mathbf{g} . ■

To see this, suppose that (x_F, y_F) is a fixed point of \mathbf{g} . This means that at the point (x_F, y_F) we have:

$$\mathbf{g}(x_F, y_F) = (x_F, y_F)$$

but then, we also have at the same point:

$$\mathbf{g}^{-1}(\mathbf{g}(x_F, y_F)) = \mathbf{g}^{-1}(x_F, y_F)$$

which means:

$$(x_F, y_F) = \mathbf{g}^{-1}(x_F, y_F)$$

so that (x_F, y_F) is, indeed, a fixed point of \mathbf{g}^{-1} , too. The other direction can be shown in a similar way.

We see, therefore, by virtue of this proposition, that the two different definitions of fixed points in our layer superposition are fully equivalent: Any point (x_F, y_F) that satisfies $\mathbf{g}(x_F, y_F) = (x_F, y_F)$ satisfies also $\mathbf{g}^{-1}(x_F, y_F) = (x_F, y_F)$, and vice versa. (Incidentally, this implies also that at any fixed point (x_F, y_F) we have $\mathbf{g}(x_F, y_F) = \mathbf{g}^{-1}(x_F, y_F)$). In conclusion:

Proposition D.7: When one of the two superposed layers is transformed by $\mathbf{g}(x,y)$ but the other layer remains unchanged, the fixed points can be expressed by either $\mathbf{g}(x,y) = (x,y)$ or $\bar{\mathbf{g}}(x,y) = (x,y)$. ■

D.11.2 Fixed points when both layers undergo transformations

Unfortunately, this happy situation does not extend to the more general case in which both of the superposed layers undergo transformations. Suppose, for example, that our two original layers $r_1(x,y)$ and $r_2(x,y)$ undergo, respectively, the domain transformations $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$. We have, therefore, the following generalization of Eq. (D.32):

$$\mathbf{g}_M(x,y) = \mathbf{g}_1(x,y) - \mathbf{g}_2(x,y) \quad (\text{D.36})$$

where $\mathbf{g}_M(x,y)$ is the domain transformation undergone by the resulting moiré.

Clearly, the fixed points of this superposition are those points (x,y) that satisfy $\mathbf{g}_M(x,y) = (0,0)$, or equivalently:³⁸

$$\mathbf{g}_1(x,y) = \mathbf{g}_2(x,y) \quad (\text{D.37})$$

³⁸ These points are the *mutual fixed points* of the transformations \mathbf{g}_1 and \mathbf{g}_2 (see Sec. 3.5).

On the other hand, just as we did in the simpler case before, we can also consider the effect of the layer transformations on the locations of the individual screen dots. In this case, each original point (x,y) in the first layer is moved by the transformation $\bar{\mathbf{g}}_1$ to a new location $\bar{\mathbf{g}}_1(x,y)$, while in the other layer the same point (x,y) is moved by the transformation $\bar{\mathbf{g}}_2$ to another destination, $\bar{\mathbf{g}}_2(x,y)$:

$$(x,y) \mapsto \bar{\mathbf{g}}_1(x,y)$$

$$(x,y) \mapsto \bar{\mathbf{g}}_2(x,y)$$

Note that $\bar{\mathbf{g}}_1$ and $\bar{\mathbf{g}}_2$ are the *direct* transformations that express the same physical layer distortions as the domain transformations \mathbf{g}_1 and \mathbf{g}_2 above, so that we have $\bar{\mathbf{g}}_1 = \mathbf{g}_1^{-1}$ and $\bar{\mathbf{g}}_2 = \mathbf{g}_2^{-1}$. For example, a two-fold expansion of the first layer can be expressed either by $r_1(x/2, y/2)$, using the domain transformation $\mathbf{g}_1(x,y) = (x/2, y/2)$, or by $(x,y) \mapsto (2x, 2y)$, using the direct transformation $\bar{\mathbf{g}}_1(x,y) = \mathbf{g}_1^{-1}(x,y) = (2x, 2y)$.

As we have seen in Sec. 4.5, the effect of the direct transformations $\bar{\mathbf{g}}_1$ and $\bar{\mathbf{g}}_2$ on any individual dot is described by the vector field:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) \quad (\text{D.38})$$

This vector field assigns to each point (x,y) the vector $\bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y)$ that connects the destination of the original point (x,y) under the transformation $\bar{\mathbf{g}}_2$ (the point $\bar{\mathbf{g}}_2(x,y)$) to the destination of the same point (x,y) under the transformation $\bar{\mathbf{g}}_1$ (the point $\bar{\mathbf{g}}_1(x,y)$).³⁹

As we can see, Eqs. (D.36) and (D.38) represent the same physical reality in two different ways: Eq. (D.36) is expressed in terms of inverse transformations, while Eq. (D.38) is expressed in terms of direct transformations. Note, however, that although $\bar{\mathbf{g}}_1$ and \mathbf{g}_1 as well as $\bar{\mathbf{g}}_2$ and \mathbf{g}_2 are mutually inverse transformations, $\bar{\mathbf{h}}(x,y)$ is not the inverse of $\mathbf{g}_m(x,y)$.

Now, based on Eq. (D.38) one may also define the fixed points of our superposition as those points that satisfy $\bar{\mathbf{h}}(x,y) = (0,0)$, or equivalently:

$$\bar{\mathbf{g}}_1(x,y) = \bar{\mathbf{g}}_2(x,y) \quad (\text{D.39})$$

Hence, just as in the simpler case with $\bar{\mathbf{g}}_2(x,y) = (x,y)$, we have obtained here two different ways to express the fixed points in the resulting layer superposition: On the one hand we say in Eq. (D.37) that the fixed points of the superposition are those points (x,y) for which $\mathbf{g}_1(x,y) = \mathbf{g}_2(x,y)$; but on the other hand, Eq. (D.39) suggests that the fixed points of the superposition are those points for which $\bar{\mathbf{g}}_1(x,y) = \bar{\mathbf{g}}_2(x,y)$. Based on our experience from the simpler case with $\bar{\mathbf{g}}_2(x,y) = (x,y)$, we might expect that conditions (D.37) and (D.39) give exactly the same fixed points. However, it turns out that this is not always true. As a simple counter-example, consider the case where the first layer undergoes a scaling by

³⁹ Note that for any points (a,b) and (c,d) in the plane, when the tail of the vector $(a,b) - (c,d)$ is attached to the point (c,d) , its head is located at the point (a,b) ; this means that the vector $(a,b) - (c,d)$ connects the point (c,d) to the point (a,b) .

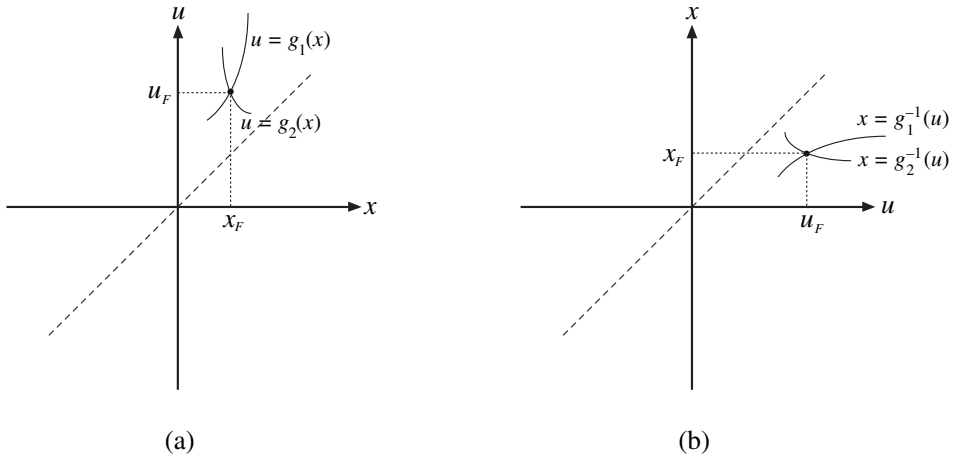


Figure D.23: (a) Two functions $g_1(x)$ and $g_2(x)$ that coincide at a certain point x_F . (b) The inverse functions g_1^{-1} and g_2^{-1} coincide at the point $u_F = g_1(x_F) = g_2(x_F)$. Note that the inverse functions are the reflections of the direct functions with respect to the main diagonal.

factor s and the second layer is shifted by (a, b) . It is easy to verify that the fixed point between the direct transformations, i.e. the point which satisfies $(sx, sy) = (x + a, y + b)$, is given by $(x_F, y_F) = (\frac{a}{s-1}, \frac{b}{s-1})$, whereas the fixed point between the inverse transformations, which satisfies $(x/s, y/s) = (x - a, y - b)$, is given by $(x_F, y_F) = (\frac{sa}{s-1}, \frac{sb}{s-1})$. (Incidentally, it is interesting to note that in this case the distance between the two fixed points is exactly $(\frac{sa}{s-1}, \frac{sb}{s-1}) - (\frac{a}{s-1}, \frac{b}{s-1}) = (a, b)$, for any values of a, b and s .)

The reason for this discrepancy between Eqs. (D.37) and (D.39) is that Proposition D.6 cannot be generalized to the case of fixed points between two transformed layers. In other words, the fact that at a given point (x_F, y_F) we have $\mathbf{g}_1(x_F, y_F) = \mathbf{g}_2(x_F, y_F)$ does not guarantee that we have at the same point $\mathbf{g}_1^{-1}(x_F, y_F) = \mathbf{g}_2^{-1}(x_F, y_F)$, too. This is clearly illustrated by the counter-example above. In fact, the reason for this failure is much easier to understand by considering the 1D counterpart of this question: As shown in Fig. D.23, the inverse of any function $g(x)$ is simply its reflection about the main diagonal. And indeed, a glance at Fig. D.23 shows that if $g_1(x)$ and $g_2(x)$ coincide at a certain point x_F , $g_1(x_F) = g_2(x_F)$, then their inverse functions g_1^{-1} and g_2^{-1} do not necessarily coincide at the same point x_F , but rather at the point $u_F = g_1(x_F) = g_2(x_F)$, giving $g_1^{-1}(u_F) = g_2^{-1}(u_F) = x_F$. Only in the particular case where the point of coincidence (x_F, u_F) happens to fall on the main diagonal, so that $u_F = x_F$, we have, indeed, $g_1^{-1}(x_F) = g_2^{-1}(x_F)$. It is obvious, therefore, that this always happens when one of the two functions is the identity function (and hence coincides with the main diagonal). As we have seen above, the same is also true in the case of 2D transformations, but in this case the visualization of the “diagonal” is less obvious since it resides in a 4D space.

But although Proposition D.6 cannot be extended to the more general case, it turns out that under the conditions mentioned below — that are anyway satisfied in our case — we can still obtain the following weaker, yet quite useful result:

Proposition D.8: If \mathbf{g}_1 and \mathbf{g}_2 are weak transformations, i.e. transformations that only slightly differ from the identity transformation $\mathbf{i}(x,y) = (x,y)$:

$$\mathbf{g}_1(x,y) = (x,y) + \mathbf{o}_1(x,y)$$

$$\mathbf{g}_2(x,y) = (x,y) + \mathbf{o}_2(x,y)$$

then $\mathbf{g}_1(x_F, y_F) = \mathbf{g}_2(x_F, y_F)$ implies that $\mathbf{g}_1^{-1}(x_F, y_F) \approx \mathbf{g}_2^{-1}(x_F, y_F)$, and $\mathbf{g}_1^{-1}(x_F, y_F) = \mathbf{g}_2^{-1}(x_F, y_F)$ implies that $\mathbf{g}_1(x_F, y_F) \approx \mathbf{g}_2(x_F, y_F)$. ■

The 1D counterpart of this result is, indeed, quite obvious, since here both $g_1(x)$ and $g_2(x)$ are close to the diagonal. In the 2D case this result follows from the fact that any weak transformation $\mathbf{g} = \mathbf{i} + \mathbf{o}$ satisfies the approximation $[\mathbf{i} + \mathbf{o}]^{-1} \approx \mathbf{i} - \mathbf{o}$ (see Proposition D.9 below). Using this result we have, therefore:

$$\mathbf{g}_1^{-1}(x,y) = [(x,y) + \mathbf{o}_1(x,y)]^{-1} \approx (x,y) - \mathbf{o}_1(x,y)$$

$$\mathbf{g}_2^{-1}(x,y) = [(x,y) + \mathbf{o}_2(x,y)]^{-1} \approx (x,y) - \mathbf{o}_2(x,y)$$

But since at the point (x_F, y_F) we have $\mathbf{g}_1(x_F, y_F) = \mathbf{g}_2(x_F, y_F)$ it follows that we also have there $\mathbf{o}_1(x_F, y_F) = \mathbf{o}_2(x_F, y_F)$, and thus we obtain, indeed:

$$\mathbf{g}_1^{-1}(x_F, y_F) \approx (x_F, y_F) - \mathbf{o}_1(x_F, y_F) = (x_F, y_F) - \mathbf{o}_2(x_F, y_F) = \mathbf{g}_2^{-1}(x_F, y_F).$$

Thus, as long as \mathbf{g}_1 and \mathbf{g}_2 are weak transformations, our two possible definitions of a fixed point between the two superposed layers can be considered as practically equivalent. But since in our case we are anyway forced to use weak layer transformations, in order not to destroy the correlation between the superposed layers, it follows that for our needs we can freely use either of the two definitions of a fixed point, (D.37) or (D.39). Interestingly, in some cases (see Remark D.22) both definitions become fully identical.

Remark D.21: Just for the sake of completeness, it is interesting to note that if we allow both of the superposed layers to be modified, there exist infinitely many different ways to distribute a given additive distortion between the two layers. Consider, for example, the case in which the first layer is scaled by factor s and the second layer remains unchanged, and suppose that the distortion we wish to add consists of a lateral shift of (a,b) . If we consider the layer transformations as *domain* transformations, the same effect will be obtained when we apply a shift of $-(a,b)$ to the first layer or when we apply a shift of (a,b) to the second layer, since:

$$\mathbf{g}_M(x,y) = (x/s + a, y/s + b) - (x,y) \tag{D.40}$$

$$\text{is identical to: } \mathbf{g}_M(x,y) = (x/s, y/s) - (x - a, y - b) \tag{D.41}$$

The same effect is also obtained when the shift is equally distributed between the two layers, i.e. when the first layer is shifted by $\frac{1}{2}(a,b)$ and the second layer is shifted by $\frac{1}{2}(a,b)$:

$$\mathbf{g}_M(x,y) = (x/s + \frac{1}{2}a, y/s + \frac{1}{2}b) - (x - \frac{1}{2}a, y - \frac{1}{2}b)$$

Clearly, there exist infinitely many different ways to distribute the distortion between the two layers, all of which give exactly the same result $\mathbf{g}_M(x,y)$, and hence, the same macroscopic moiré in the superposition (by virtue of Proposition 5.1 in Sec. 5.3, or the second part of Proposition 10.5 in Sec. 10.9.2 of *Vol. I*). However, when we consider the very same distorted layers in terms of *direct* transformations (for example, in order to study the dot trajectories in the layer superposition, as explained in Chapter 4), each of these cases gives a different result. For example, in the case of Eq. (D.40) we have:

$$\begin{aligned} \bar{\mathbf{h}}_1(x,y) &= [(x/s + a, y/s + b)]^{-1} - [(x,y)]^{-1} \\ &= (s(x-a), s(y-b)) - (x,y) \end{aligned} \quad (\text{D.42})$$

(since the inverse transformation of $u = x/s + a$, $v = y/s + b$ is $x = s(u-a)$, $y = s(v-b)$; see also Remark D.4 on the variable names), while in the case of Eq. (D.41) we obtain:

$$\begin{aligned} \bar{\mathbf{h}}_2(x,y) &= [(x/s, y/s)]^{-1} - [(x-a, y-b)]^{-1} \\ &= (sx, sy) - (x+a, y+b) \end{aligned} \quad (\text{D.43})$$

where clearly $\bar{\mathbf{h}}_2(x,y) \neq \bar{\mathbf{h}}_1(x,y)$.

Because all the different distributions of the shift between the two layers are equivalent in terms of domain transformations, it is clear that all of them have the same fixed point in accordance with Eq. (D.37); and indeed, both $(x/s, y/s) = (x-a, y-b)$ and $(x/s+a, y/s+b) = (x,y)$ give the same fixed point, $(x,y) = (\frac{sa}{s-1}, \frac{sb}{s-1})$.

However, using the definition (D.39), which is based on direct transformations, Eq. (D.42) gives the fixed point $(x,y) = (\frac{sa}{s-1}, \frac{sb}{s-1})$ while Eq. (D.43) gives the fixed point $(x,y) = (\frac{a}{s-1}, \frac{b}{s-1})$. But among all of the different possible cases, there exists only a single one that gives the same fixed point as in Eqs. (D.40) and (D.41): By virtue of Proposition D.6, this is precisely the case in which only one of the superposed layers undergoes a transformation, while the other layer remains unchanged (in our example above this is the case given in Eq. (D.42)). Nevertheless, as we have just seen above, if the transformations applied to our layers are weak then the difference between the resulting values is practically negligible. ■

D.12 Useful approximations

In this appendix we provide some approximations that are often handy to use, because they allow us to formulate our main results in terms of either direct or inverse transformations. We start with the following informal definitions:

Definition D.1: A transformation $(u,v) = \mathbf{o}(x,y)$ is said to be *almost zero* or *almost null* if it differs only slightly, within our zone of interest (i.e. within the area covered by the layer superposition), from the zero transformation $(u,v) = \mathbf{z}(x,y) = (0,0)$. ■

Definition D.2: A transformation $(u,v) = \mathbf{g}(x,y)$ is said to be *weak* or *almost identity* if it differs only slightly, within our zone of interest (i.e. within the area covered by the layer superposition), from the identity transformation $(u,v) = \mathbf{i}(x,y) = (x,y)$. In other words, $\mathbf{g}(x,y)$ is a weak transformation if it satisfies $\mathbf{g}(x,y) = (x,y) + \mathbf{o}(x,y)$, where $\mathbf{o}(x,y)$ is an almost zero transformation. ■

We now provide the following useful result:

Proposition D.9: If $\mathbf{o}(x,y)$ is an almost zero transformation, then:

$$(a) \quad [(x,y) - \mathbf{o}(x,y)]^{-1} \approx (x,y) + \mathbf{o}(x,y)$$

$$(b) \quad [(x,y) + \mathbf{o}(x,y)]^{-1} \approx (x,y) - \mathbf{o}(x,y)$$

or, in a more compact notation, denoting the identity and the zero transformations by \mathbf{i} and \mathbf{o} , respectively:

$$(a) \quad [\mathbf{i} - \mathbf{o}]^{-1} \approx \mathbf{i} + \mathbf{o}$$

$$(b) \quad [\mathbf{i} + \mathbf{o}]^{-1} \approx \mathbf{i} - \mathbf{o}$$

Furthermore, the errors in the two approximations (a) and (b) are almost identical. ■

Note that this proposition also means that if $\mathbf{g}(x,y)$ is a weak transformation, so that $\mathbf{g}(x,y) = (x,y) + \mathbf{o}(x,y)$, then $\mathbf{g}^{-1}(x,y) \approx (x,y) - \mathbf{o}(x,y)$; see also Figs. D.21 and D.22.

To derive this proposition, note that an almost zero transformation \mathbf{o} is almost linear, because in its 2D Taylor series development (see, for example, [Courant89 pp. 68–70] or [Strogatz94 p. 150]) all the non-linear terms (i.e. the terms of second and higher orders) are negligible and can be dropped out. Therefore we can use the distributive law for the composition of linear transformations [Mansfield76 p. 168], and we obtain, denoting by “ \circ ” the composition of transformations:

$$\begin{aligned} (\mathbf{i} + \mathbf{o}) \circ (\mathbf{i} - \mathbf{o}) &= \mathbf{i} \circ \mathbf{i} + \mathbf{o} \circ \mathbf{i} - \mathbf{i} \circ \mathbf{o} - \mathbf{o} \circ \mathbf{o} \\ &= \mathbf{i} - \mathbf{o} \circ \mathbf{o} \\ &\approx \mathbf{i} \end{aligned}$$

(since the influence of $\mathbf{o} \circ \mathbf{o}$ is negligible). This means, indeed, that the transformations $\mathbf{i} + \mathbf{o}$ and $\mathbf{i} - \mathbf{o}$ are approximately the inverse of each other (see [Weisstein99 p. 901] for an equivalent reasoning in terms of matrices).

An alternative approach for deriving this proposition is based on the following well-known identities [Harris89 p. 540]:

$$(1 - x)^{-1} = 1 + x + x^2 + x^3 + x^4 + x^5 + \dots \quad (|x| < 1)$$

$$(1 + x)^{-1} = 1 - x + x^2 - x^3 + x^4 - x^5 + \dots \quad (|x| < 1)$$

or rather on their matrix counterparts [Horn85 p. 301]:

$$(I - A)^{-1} = I + A + A^2 + A^3 + A^4 + A^5 + \dots \quad (\|A\| < 1)$$

$$(I + A)^{-1} = I - A + A^2 - A^3 + A^4 - A^5 + \dots \quad (\|A\| < 1)$$

where A is a square matrix (in our case 2×2), $\|A\|$ is its norm, and I denotes the unit matrix.⁴⁰

These matrix identities mean, in terms of the linear transformation $\mathbf{g}(x,y)$ that corresponds to the matrix A (see, for example, [Kreyszig78 p. 375]):⁴¹

$$(\mathbf{i} - \mathbf{g})^{-1} = \mathbf{i} + \mathbf{g} + \mathbf{g} \circ \mathbf{g} + \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} + \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} + \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} + \dots \quad (\|\mathbf{g}\| < 1)$$

$$(\mathbf{i} + \mathbf{g})^{-1} = \mathbf{i} - \mathbf{g} + \mathbf{g} \circ \mathbf{g} - \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} + \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} - \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} \circ \mathbf{g} + \dots \quad (\|\mathbf{g}\| < 1)$$

Consider now the almost-zero transformation $\mathbf{o}(x,y)$. As we have seen above, this transformation is almost linear, and therefore we have:

$$(\mathbf{i} - \mathbf{o})^{-1} \approx \mathbf{i} + \mathbf{o} + \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \dots \quad (\|\mathbf{o}\| < 1)$$

$$(\mathbf{i} + \mathbf{o})^{-1} \approx \mathbf{i} - \mathbf{o} + \mathbf{o} \circ \mathbf{o} - \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} - \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \dots \quad (\|\mathbf{o}\| < 1)$$

But since we assume that \mathbf{o} is an almost zero transformation, the influence of the mapping compositions $\mathbf{o} \circ \mathbf{o}$, $\mathbf{o} \circ \mathbf{o} \circ \mathbf{o}$, etc. is negligible; and indeed, by omitting them we obtain the two desired approximations:

$$[\mathbf{i} - \mathbf{o}]^{-1} \approx \mathbf{i} + \mathbf{o}$$

$$[\mathbf{i} + \mathbf{o}]^{-1} \approx \mathbf{i} - \mathbf{o}$$

Furthermore, it turns out that the approximation error \mathbf{e}_1 in (a) and the approximation error \mathbf{e}_2 in (b), both of which are functions of x and y , $\mathbf{e}_1(x,y)$, $\mathbf{e}_2(x,y)$, are almost identical. To see this, note that the approximation errors \mathbf{e}_1 and \mathbf{e}_2 are given by:⁴²

$$\mathbf{e}_1 = \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \dots$$

$$\mathbf{e}_2 = \mathbf{o} \circ \mathbf{o} - \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} - \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \dots$$

⁴⁰ It turns out that these matrix series converge if $\|A\| < 1$ for *any* matrix norm $\|A\|$ [Horn85 pp. 258, 301]. This condition is clearly satisfied by matrices that are very close to the zero matrix O , as in our case.

⁴¹ The norm $\|\mathbf{g}\|$ can be defined, for example, by $\|\mathbf{g}\| = \sup_{|(x,y)|=1} |\mathbf{g}(x,y)|$ [Kreyszig78 p. 92], where $|(u,v)|$ is the length of the vector (u,v) .

⁴² Note that \mathbf{e}_1 and \mathbf{e}_2 express the truncation errors in (a) and (b), and therefore they accurately represent the errors when \mathbf{o} is a linear transformation (that corresponds to the matrix A). But when \mathbf{o} is only *almost* linear, the errors \mathbf{e}_1 and \mathbf{e}_2 are only *approximate*.

It follows, therefore, because \mathbf{o} is an almost zero transformation, that the difference:

$$\mathbf{e}_1 - \mathbf{e}_2 = 2\mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + 2\mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} \circ \mathbf{o} + \dots$$

is negligible with respect to \mathbf{e}_1 and \mathbf{e}_2 , since $\mathbf{o} \circ \mathbf{o} \circ \mathbf{o}$ is of lower order of magnitude than $\mathbf{o} \circ \mathbf{o}$. This means, indeed, that $\mathbf{e}_1 \approx \mathbf{e}_2$.

Using this result we provide now the following useful approximations that connect between Eqs. (D.36) and (D.38):

Proposition D.10: If \mathbf{g}_1 and \mathbf{g}_2 are weak transformations then:

$$\bar{\mathbf{g}}_1(x, y) - \bar{\mathbf{g}}_2(x, y) \approx \mathbf{g}_2(x, y) - \mathbf{g}_1(x, y) \quad (\text{D.44})$$

This means, using our notations from Eqs. (D.36) and (D.38), that the following approximations:

$$\bar{\mathbf{h}}(x, y) \approx -\mathbf{g}_M(x, y) \quad (\text{D.45})$$

$$\bar{\mathbf{h}}(x, y) \approx \mathbf{g}_2(x, y) - \mathbf{g}_1(x, y) \quad (\text{D.46})$$

$$\mathbf{g}_M(x, y) \approx \bar{\mathbf{g}}_2(x, y) - \bar{\mathbf{g}}_1(x, y) \quad (\text{D.47})$$

also hold. ■

To see this, let $\mathbf{g}_1 = \mathbf{i} + \mathbf{o}_1$ and $\mathbf{g}_2 = \mathbf{i} + \mathbf{o}_2$ be weak transformations. We have, therefore:

$$\bar{\mathbf{g}}_1(x, y) - \bar{\mathbf{g}}_2(x, y) = [(x, y) + \mathbf{o}_1(x, y)]^{-1} - [(x, y) + \mathbf{o}_2(x, y)]^{-1}$$

$$\begin{aligned} \text{Using Proposition D.9:} \quad & \approx [(x, y) - \mathbf{o}_1(x, y)] - [(x, y) - \mathbf{o}_2(x, y)] \\ & = [(x, y) + \mathbf{o}_2(x, y)] - [(x, y) + \mathbf{o}_1(x, y)] \\ & = \mathbf{g}_2(x, y) - \mathbf{g}_1(x, y) \end{aligned}$$

The other approximations, (D.45)–(D.47), are simply different variants of (D.44). This proposition is very useful in cases where we do not have the explicit forms of the inverse (or of the direct) transformations, since it still allows us to obtain a close approximation based on those transformations whose explicit forms *are* known (see, for instance, Example 5.5 in Sec. 5.3).

Remark D.22: It is interesting to note that for some categories of transformations $\mathbf{g}_1(x, y)$ and $\mathbf{g}_2(x, y)$ the approximations given in Proposition D.10 turn into *identities*. For example, it is easy to see that if $(u, v) = \mathbf{g}_1(x, y)$ and $(u, v) = \mathbf{g}_2(x, y)$ are given, respectively, by:

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x + f_1(y) \\ y \end{pmatrix} \quad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x \\ y + f_2(x) \end{pmatrix} \quad (\text{D.48})$$

where $z = f_1(s)$ and $z = f_2(s)$ are arbitrary 1D functions, then their inverse transformations are given, respectively, by:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u - f_1(v) \\ v \end{pmatrix} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u \\ v - f_2(u) \end{pmatrix}$$

and thus we have:

$$\mathbf{g}_1(x,y) - \mathbf{g}_2(x,y) = (f_1(y), -f_2(x))$$

$$\bar{\mathbf{g}}_2(x,y) - \bar{\mathbf{g}}_1(x,y) = (f_1(y), -f_2(x))$$

meaning that the following identity holds:

$$\mathbf{g}_1(x,y) - \mathbf{g}_2(x,y) = \bar{\mathbf{g}}_2(x,y) - \bar{\mathbf{g}}_1(x,y) \quad (\text{D.49})$$

Another example consists of all the transformations $\mathbf{g}_1(x,y)$, $\mathbf{g}_2(x,y)$ having the form:

$$(u,v) = (f_1(x), f_2(y)) \qquad (u,v) = (f_1^{-1}(x), f_2^{-1}(y)) \quad (\text{D.50})$$

where $z = f_1(s)$ and $z = f_2(s)$ are arbitrary 1D functions. In this case we have: $\bar{\mathbf{g}}_1(x,y) = \mathbf{g}_2(x,y)$ and $\bar{\mathbf{g}}_2(x,y) = \mathbf{g}_1(x,y)$, and therefore identity (D.49) holds, again.

There also exist other families of transformation pairs that satisfy identity (D.49); but for all other pairs \mathbf{g}_1 and \mathbf{g}_2 that do not satisfy the identity we still have the approximation given by Proposition D.10, provided that \mathbf{g}_1 and \mathbf{g}_2 are sufficiently weak transformations. ■

Finally, note that whenever the approximations of Proposition D.10 turn into identities, this also implies that the approximations given by Proposition D.8 turn into equalities, so that a point in the plane is a mutual fixed point of $\mathbf{g}_1(x,y)$ and $\mathbf{g}_2(x,y)$ *iff* it is a mutual fixed point of $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$. The converse, however, is not necessarily true. For example, consider the simple case in which one of the two superposed layers remains unchanged: $\mathbf{g}_2(x,y) = (x,y)$. We already know from Proposition D.6 that in this case any fixed point of \mathbf{g} is also a fixed point of $\bar{\mathbf{g}}$ (and vice versa), so that any point that satisfies $\mathbf{g}_1(x,y) - (x,y) = (0,0)$ also satisfies $(x,y) - \bar{\mathbf{g}}_1(x,y) = (0,0)$. However, this does not yet imply that for *any* point (x,y) we have $\mathbf{g}_1(x,y) - (x,y) = (x,y) - \bar{\mathbf{g}}_1(x,y)$. For example, if one of the layers has undergone a two-fold magnification $\bar{\mathbf{g}}_1(x,y) = (2x, 2y)$ and the other layer remains unchanged, then:

$$\bar{\mathbf{g}}_1(x,y) - (x,y) = (2x, 2y) - (x,y) = (x,y)$$

while: $\mathbf{g}_1(x,y) - (x,y) = (x/2, y/2) - (x,y) = -(x/2, y/2)$

This clearly shows that in this case identity (D.49) is not satisfied; and yet, the fixed points of $\bar{\mathbf{g}}_1$ and \mathbf{g}_1 are, indeed, identical (the point (0,0)).

We conclude this appendix with some further approximations that prove to be useful in Chapter 4 (see Remark 4.3):

Proposition D.11: If \mathbf{g}_1 and \mathbf{g}_2 are weak transformations, i.e. transformations that only slightly differ from the identity transformation $\mathbf{i}(x,y) = (x,y)$ (at least within our zone of interest), then the following approximations:

$$\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y)) - (x,y) \approx \mathbf{g}_2(x,y) - \mathbf{g}_1(x,y) \approx \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) \quad (\text{D.51})$$

$$(x,y) - \bar{\mathbf{g}}_2(\mathbf{g}_1(x,y)) \approx \mathbf{g}_2(x,y) - \mathbf{g}_1(x,y) \approx \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) \quad (\text{D.52})$$

also hold. ■

This result can be demonstrated as follows:

Denoting $(x',y') = \mathbf{g}_2(x,y)$ we have:

$$\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y)) - (x,y) = \bar{\mathbf{g}}_1(x',y') - (x,y)$$

Now, using $\mathbf{g}_1 = \mathbf{i} + \mathbf{o}_1$ and hence $\bar{\mathbf{g}}_1 = [\mathbf{i} + \mathbf{o}_1]^{-1}$:

$$= [(x',y') + \mathbf{o}_1(x',y')]^{-1} - (x,y)$$

which gives, by virtue of Proposition D.9:

$$\approx (x',y') - \mathbf{o}_1(x',y') - (x,y)$$

and by substituting back $(x',y') = \mathbf{g}_2(x,y)$ and using $\mathbf{g}_2 = \mathbf{i} + \mathbf{o}_2$:

$$= (x,y) + \mathbf{o}_2(x,y) - \mathbf{o}_1((x,y) + \mathbf{o}_2(x,y)) - (x,y)$$

But since within our zone of interest \mathbf{o}_1 is almost linear we obtain:

$$\approx (x,y) + \mathbf{o}_2(x,y) - \mathbf{o}_1(x,y) - \mathbf{o}_1(\mathbf{o}_2(x,y)) - (x,y)$$

which gives in turn, because $\mathbf{o}_1(\mathbf{o}_2(x,y))$ is negligible:

$$\approx \mathbf{g}_2(x,y) - \mathbf{g}_1(x,y)$$

or, using (D.44):

$$\approx \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y)$$

Eq. (D.51) shows that when \mathbf{g}_1 and \mathbf{g}_2 are weak transformations Eq. (4.7) of Sec. 4.5 is indeed a close approximation to Eq. (4.8). This explains also why vector field (4.7) gives a good approximation to the dot trajectories in cases where both of the original layers are being transformed, provided that both layer transformations are rather weak. Note that we are allowed to use these approximations because in our application we must anyway restrict ourselves to weak layer transformations, in order not to destroy the correlation between the superposed layers within our zone of interest.

Appendix E

Convolution and cross correlation

E.1 Introduction

Convolution and cross correlation are operations on functions: each of these operations takes two functions $f(x)$ and $g(x)$, and produces from them a new function of the same variable x , that is customarily denoted by $h(x) = f(x) * g(x)$ or by $c_{f,g}(x) = f(x) \star g(x)$, respectively.¹ The need for the subscript in the cross correlation $c_{f,g}(x)$ will become clear shortly. In the 2D case the convolution and the cross correlation of the functions $f(x,y)$ and $g(x,y)$ are denoted by $h(x,y) = f(x,y) ** g(x,y)$ and $c_{f,g}(x,y) = f(x,y) \star \star g(x,y)$. A full description of convolution and cross correlation can be found in the literature (see, for example, Chapters 6 and 9 in [Gaskill78], Chapters 5 and 7 in [Cartwright90], or Chapter 5 in [Bracewell95]); in this appendix we only provide a general overview and discuss those properties that may be needed for our application. Because we are basically interested in 2D images this appendix is mostly formulated in terms of 2D functions; but the corresponding 1D results can be deduced from their 2D counterparts without difficulty. Note also that this appendix only deals with the continuous case. The discrete case — where the functions or images being treated are discrete (i.e. composed of pixels) — is basically obtained by replacing integrations by summations, but it remains beyond our scope here. More details on discrete convolution and cross correlation can be found in books on digital image processing or digital signal processing.

E.2 Convolution

The convolution of two real-valued functions $f(x,y)$ and $g(x,y)$ is defined to be:

$$h(x,y) = f(x,y) ** g(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g(x-x', y-y') dx'dy' \quad (\text{E.1})$$

Because this integral is clearly a function of the independent variables x and y , the resulting function $h(x,y)$ is again a function of x and y .² As we will see soon, the convolution operation may be viewed as one of finding the volume of the product of $f(x',y')$ and $g(x-x', y-y')$ as x and y are allowed to vary. In general, the resulting function $h(x,y)$ is smoother than either $f(x,y)$ or $g(x,y)$: The fine structure of the original functions

¹ More formally, convolution and cross correlation are *binary operators* that act on functions. Such binary operators are often written in the form $h = L[f,g]$. Simple examples of such operators include $S[f,g] = f + g$, $P[f,g] = fg$, $C[f,g] = f \star g$, etc. Similarly, there exist also *unary operators* that act on a single function, such as $U[f] = f^2$ or the Fourier transform, $\mathcal{F}[f]$ [Cartwright90 p. 135].

² Note that inside the integral the x and y axes are renamed x' and y' ; thus, after integrating over the plane the dummy integration variables x' and y' disappear, and we are left with a function of x and y .

tends to be washed out, the sharp peaks and valleys tend to be rounded, etc., but the amount of smoothing depends on the exact nature of $f(x,y)$ and $g(x,y)$ (see [Gaskill78 pp. 162–163]).

Definition (E.1) may seem at first sight quite obscure, but in fact it has a quite intuitive graphical interpretation that gives a good visual insight into the nature of the convolution operation — the “move and multiply” interpretation (see, for example, [Rosenfeld82 pp. 13–14] or [Gaskill78 pp. 151–154, 291–292]): In order to obtain graphically the convolution of $f(x,y)$ and $g(x,y)$, we first draw the function $g(-x',-y')$, which is simply a 180° rotation of $g(x',y')$, and then we shift it along the x and y directions on top of $f(x',y')$. For each position x,y of the moving function, the resulting value of $h(x,y)$ is simply the volume under the product of the two functions (the fixed function $f(x',y')$ and the moving function $g(-x',-y')$ when it is shifted by x,y , i.e., $g(-(x' - x), -(y' - y)) = g(x - x', y - y')$).

The convolution operation has several useful properties, such as commutativity:

$$f(x,y) ** g(x,y) = g(x,y) ** f(x,y) \quad (\text{E.2})$$

associativity:

$$[f_1(x,y) ** f_2(x,y)] ** f_3(x,y) = f_1(x,y) ** [f_2(x,y) ** f_3(x,y)] \quad (\text{E.3})$$

homogeneity:

$$\begin{aligned} [cf(x,y)] ** g(x,y) &= c[f(x,y) ** g(x,y)] \\ f(x,y) ** [cg(x,y)] &= c[f(x,y) ** g(x,y)] \end{aligned} \quad (\text{E.4})$$

distributivity over addition:³

$$\begin{aligned} [f_1(x,y) + f_2(x,y)] ** g(x,y) &= [f_1(x,y) ** g(x,y)] + [f_2(x,y) ** g(x,y)] \\ f(x,y) ** [g_1(x,y) + g_2(x,y)] &= [f(x,y) ** g_1(x,y)] + [f(x,y) ** g_2(x,y)] \end{aligned} \quad (\text{E.5})$$

and shift preservation:

$$f(x - a, y - b) ** g(x,y) = f(x,y) ** g(x - a, y - b) = h(x - a, y - b) \quad (\text{E.6})$$

Another interesting property is that the volume under the convolution $h(x,y)$ equals the product of the volumes under $f(x,y)$ and $g(x,y)$ [Bracewell95 p. 193]. Other properties of the convolution operation can be found, for example, in [Gaskill78 pp. 159–166, 292–294] and in [Poularikas96 pp. 27–32].

³ More formally, properties (E.4) and (E.5) together imply that the convolution operator is *linear* [Cartwright90 p. 130]. In general, a binary operator $L[f,g]$ is said to be linear if it is homogeneous and additive in each of its two arguments, i.e. if it has the following properties: $L[cf,g] = cL[f,g]$, $L[f_1 + f_2, g] = L[f_1,g] + L[f_2,g]$, $L[f, cg] = cL[f,g]$, and $L[f, g_1 + g_2] = L[f,g_1] + L[f,g_2]$. Similarly, a unary operator $U[f]$ is said to be linear if $U[cf] = cU[f]$ and $U[f_1 + f_2] = U[f_1] + U[f_2]$; an example of such an operator is the Fourier transform $\mathcal{F}[f]$ [Cartwright90 p. 90].

E.3 Cross correlation

Given two real-valued functions $f(x,y)$ and $g(x,y)$, we define the cross correlation of $f(x,y)$ and $g(x,y)$ to be:

$$\begin{aligned} c_{f,g}(x,y) &= f(x,y) \star\star g(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g(x' - x, y' - y) dx' dy' \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x'' + x, y'' + y) g(x'', y'') dx'' dy'' \end{aligned} \quad (\text{E.7})$$

(where the second integral is obtained from the first one by a simple change of variables, $x'' = x' - x$, $y'' = y' - y$). The cross-correlation function $c_{f,g}(x,y)$ indicates the relative amount of agreement between the two given functions $f(x,y)$ and $g(x,y)$ for all possible degrees of misalignment (displacements). In other words, the function $c_{f,g}(x,y)$ is a measure of the similarity between $f(x,y)$ and a moving copy of $g(x,y)$, and it gets its maximum at the point (x,y) representing the displacement of g for which its similarity with f is the highest.⁴

As we can see, the cross-correlation operation, too, can be presented graphically using the “move and multiply” interpretation. It is important to note, however, that although this operation is similar to convolution, there is one very significant difference: the function $g(x,y)$ is not rotated as in convolution. This fact implies that convolution and cross correlation behave very differently. An example illustrating graphically the difference between $f(x) * g(x)$ and $f(x) \star g(x)$ for some given functions $f(x)$ and $g(x)$ can be found, for instance, in [Coulon84 pp. 82–83]. In fact, it is easy to see from definitions (E.1) and (E.7) that cross correlation can be expressed in terms of convolution by:

$$f(x,y) \star\star g(x,y) = f(x,y) ** g(-x,-y) \quad (\text{E.8})$$

It follows, therefore, that unlike convolution, the cross-correlation operation is *not* commutative, meaning that in general $f(x,y) \star\star g(x,y) \neq g(x,y) \star\star f(x,y)$. For this reason we have to clearly distinguish between the two functions $c_{f,g}(x,y) = f(x,y) \star\star g(x,y)$ and $c_{g,f}(x,y) = g(x,y) \star\star f(x,y)$; the relationship between the two is given by:⁵

$$c_{g,f}(x,y) = c_{f,g}(-x,-y) \quad (\text{E.9})$$

Another consequence of Eq. (E.8) is that if $g(-x,-y) = g(x,y)$ then $f(x,y) \star\star g(x,y)$ is identical to the convolution $f(x,y) ** g(x,y)$. In this particular case the cross correlation is, of course, commutative; such exceptional commutativity may also occur in some other special cases (see Problem 5-10 in [Bracewell95 pp. 201 and 643]).

Note that in the case of *complex-valued* functions $f(x,y)$ and $g(x,y)$ one can also define the *complex cross correlation* of $f(x,y)$ and $g(x,y)$:

⁴ Note that in order to facilitate comparisons of the correlation between different functions it may be useful to normalize the cross correlation by dividing it by the square root of the product of the volume under $[f(x,y)]^2$ and the volume under $[g(x,y)]^2$. This ensures that the values of the normalized cross correlation are always between -1 and 1 [Cartwright90 p. 174].

⁵ More formally, this means that the cross correlation operator $C[f,g]$ is anti-symmetric: $C[g(x), f(x)] = C[f(-x), g(-x)]$ [Cartwright90 pp. 174, 176].

$$\begin{aligned}
\gamma_{f,g}(x,y) &= f(x,y) \star\star g^*(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g^*(x'-x, y'-y) dx' dy' \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x''+x, y''+y) g^*(x'',y'') dx'' dy''
\end{aligned} \tag{E.10}$$

where $g^*(x,y)$ is the complex conjugate of the function $g(x,y)$. This definition is largely used in the literature, but because in our case we are only interested in real-valued functions (images), it is clear that $g^*(x,y) = g(x,y)$ and therefore the complex cross correlation (E.10) reduces into the cross correlation as defined in Eq. (E.7).

Remark E.1: There exists in the literature an alternative convention for the definition of convolution and cross correlation, in which the roles of f and g under the integral are inversed (see, for example, [Bracewell95 pp. 176–179]). Although this does not have any effect on the resulting function in the case of convolution (due to the commutativity of this operation), in the case of cross correlation this alternative definition inverses the meanings of $c_{f,g}(x,y)$ and $c_{g,f}(x,y)$, and gives instead of (E.8): $f(x,y) \star\star g(x,y) = f(-x,-y) ** g(x,y)$. In order to avoid confusion we will stick here to our definitions (E.1) and (E.7), following the notations in [Gaskill78]. ■

Cross correlation is less generous than convolution in terms of nice mathematical properties. We have already mentioned its lack of commutativity, but in fact, it also lacks associativity (as one can see by rewriting the expressions $[f_1(x,y) \star\star f_2(x,y)] \star\star f_3(x,y)$ and $f_1(x,y) \star\star [f_2(x,y) \star\star f_3(x,y)]$ in terms of convolution, using Eq. (E.8)). And yet, cross correlation still does have some other useful properties, such as homogeneity [Cartwright90 p. 174]:

$$\begin{aligned}
[c f(x,y)] \star\star g(x,y) &= c [f(x,y) \star\star g(x,y)] \\
f(x,y) \star\star [c g(x,y)] &= c [f(x,y) \star\star g(x,y)]
\end{aligned} \tag{E.11}$$

distributivity over addition [Cartwright90 p. 174]:⁶

$$\begin{aligned}
[f_1(x,y) + f_2(x,y)] \star\star g(x,y) &= [f_1(x,y) \star\star g(x,y)] + [f_2(x,y) \star\star g(x,y)] \\
f(x,y) \star\star [g_1(x,y) + g_2(x,y)] &= [f(x,y) \star\star g_1(x,y)] + [f(x,y) \star\star g_2(x,y)]
\end{aligned} \tag{E.12}$$

an asymmetric shift-preservation property:

$$f(x+a, y+b) \star\star g(x,y) = f(x,y) \star\star g(x-a, y-b) = c_{f,g}(x-a, y-b) \tag{E.13}$$

and in spite of its non-commutativity it still satisfies the identity [Weisstein99 p. 352]:

$$(f \star\star g) \star\star (f \star\star g) = (f \star\star f) \star\star (g \star\star g) \tag{E.14}$$

Further properties of cross correlation are mentioned at the end of this section.

⁶ Once again, like in the case of convolution, properties (E.11) and (E.12) together mean that cross correlation is a linear operator. Note that the lack of commutativity, $C[g,f] \neq C[f,g]$, does not contradict the linearity of the operator (i.e., its homogeneity and additivity in each of its two arguments).

An interesting particular case of cross correlation occurs when the functions f and g are identical; in this case the operation (E.7) is called the autocorrelation of $f(x,y)$, and denoted by $c_{f,f}(x,y) = f(x,y) \star\star f(x,y)$. The autocorrelation operation is described in detail in [Bracewell95 pp. 181–193]. It has several interesting properties some of which do not, in general, hold for either cross correlation or convolution, not even for self convolution $f(x,y) \star\star f(x,y)$ [Gaskill78 pp. 174–176]:

- (1) The autocorrelation $c_{f,f}(x,y)$ has a twofold rotational symmetry about the origin (see a few examples in [Bracewell95 p. 185, Fig. 5-10]).⁷
- (2) Its value is maximum at the origin: $c_{f,f}(x,y) \leq c_{f,f}(0,0)$.
- (3) Its central value $c_{f,f}(0,0)$ is the volume under the function $[f(x,y)]^2$ [Bracewell95 p. 192].
- (4) The volume under the autocorrelation function is the square of the volume under $f(x,y)$ [Bracewell95 p. 192]. (Note that this is not the same as the volume under $[f(x,y)]^2$ in the previous property.)
- (5) The autocorrelation is invariant under shifts of $f(x,y)$:
 $f(x-a, y-b) \star\star f(x-a, y-b) = f(x,y) \star\star f(x,y)$ [Bracewell95 pp. 185–196].

Note that some of these properties of autocorrelation can be *partially* generalized into the case of cross correlation between two different functions $f(x,y)$ and $g(x,y)$. Property (2) can be generalized as follows [Coulon84 p. 82], [Bendat93 p. 48]:

$$[c_{f,g}(x,y)]^2 \leq c_{f,f}(0,0) c_{g,g}(0,0) \quad (\text{E.15})$$

Note, however, that unlike in autocorrelation this does not mean that the absolute value of cross correlation is maximum at the origin. The counterpart of property (3) says in the case of cross correlation that $c_{f,g}(0,0)$ is the volume under the product $f(x,y)g(x,y)$. The counterpart of property (4) says that the volume under the cross correlation $f(x,y) \star\star g(x,y)$ equals the product of the volumes under $f(x,y)$ and $g(x,y)$; this is obtained from the analogous property of convolution (see at the end of Sec. E.2) by using Eq. (E.8) and noting that the area under $g(-x,-y)$ equals the area under $g(x,y)$. Finally, the counterpart of property (5) for cross correlation is already given in Eq. (E.13) above; note that this shift-preservation property is weaker than the invariance under shifts in autocorrelation, since it only asserts that shifting $f(x,y)$ or $g(x,y)$ gives a shifted version of $c_{f,g}(x,y)$, but it does not give $c_{f,g}(x,y)$ itself.

E.4 Extension to more general cases

Definitions (E.1) and (E.7) are only appropriate in cases where the integrals have finite values. This includes cases where one (or both) of the functions is a *finite-energy signal*

⁷ This also means that the unary operator of autocorrelation $U[f]$ is symmetric: $U[f(x,y)] = U[f(-x,-y)]$.

(i.e. square integrable) [Champeney73 pp. 59, 68], but it obviously excludes cases with functions of constant character, periodic functions, or stationary random functions. But there exists a way for generalizing the notions of convolution and cross correlation to functions that are *finite-power signals* [Champeney73 pp. 59, 63, 68]; those include also constant functions, step functions, periodic functions and stationary random functions. This generalization is done by replacing the integral by a limiting process (see, for example, [Cartwright90 pp. 174–175], [Gaskill78 p. 158] or [Coulon84 pp. 34, 91–92]):⁸

$$f(x,y) ** g(x,y) = \lim_{a \rightarrow \infty} \frac{1}{a^2} \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} f(x',y') g(x-x', y-y') dx'dy' \quad (\text{E.16})$$

$$f(x,y) \star\star g(x,y) = \lim_{a \rightarrow \infty} \frac{1}{a^2} \int_{-a/2}^{a/2} \int_{-a/2}^{a/2} f(x',y') g(x'-x, y'-y) dx'dy' \quad (\text{E.17})$$

In the periodic case, where both of the given functions f and g are assumed to have the same period T , the limits given in Eqs. (E.16) and (E.17) are identical to the mean values calculated on a single period. Therefore, in this case we have [Coulon84 p. 99]:

$$f(x,y) ** g(x,y) = \frac{1}{T^2} \int_T \int_T f(x',y') g(x-x', y-y') dx'dy' \quad (\text{E.18})$$

$$f(x,y) \star\star g(x,y) = \frac{1}{T^2} \int_T \int_T f(x',y') g(x'-x, y'-y) dx'dy' \quad (\text{E.19})$$

This gives, of course, periodic functions having the same period T as f and g . These functions are known, respectively, as T -convolution and T -cross correlation (or *cyclic* convolution and *cyclic* cross-correlation). An extension to cases where the functions f and g have different periods is also possible, as shown in [Gaskill78 p. 158].

E.5 The Fourier transform of convolution and of cross correlation

The operations of convolution and cross correlation play a major role in the Fourier theory thanks to two fundamental theorems, that are known as the convolution theorem and the cross-correlation theorem. Given two real-valued functions $f(x,y)$ and $g(x,y)$, the convolution theorem asserts that the Fourier transform of the convolution $h(x,y) = f(x,y) ** g(x,y)$ is given by the product of the individual Fourier transforms:

$$H(u,v) = F(u,v)G(u,v) \quad (\text{E.20})$$

As an immediate result, the autoconvolution theorem says that the Fourier transform of the autoconvolution $h(x,y) = f(x,y) ** f(x,y)$ is given by:

$$H(u,v) = F(u,v)^2 \quad (\text{E.21})$$

⁸ In fact, this is a generalization of the normalized versions of Eqs. (E.1) and (E.7) that are divided by the area a^2 in which the integration is taking place (assuming that the functions f and g have a finite spatial extent).

On the other hand, given the same real-valued functions $f(x,y)$ and $g(x,y)$, the cross-correlation theorem says that the Fourier transform of the cross correlation $c_{f,g}(x,y) = f(x,y) \star \star g(x,y)$ is given by [Gaskill78 p. 200]:

$$C_{f,g}(u,v) = F(u,v)G(-u,-v) \quad (\text{E.22})$$

As a particular case, the autocorrelation theorem says that the Fourier transform of the autocorrelation $c_{f,f}(x,y) = f(x,y) \star \star f(x,y)$ is given by:

$$C_{f,f}(u,v) = F(u,v)F(-u,-v) \quad (\text{E.23})$$

but since in our case $f(x,y)$ is real it follows that $F(u,v)$ is Hermitian, and therefore $F(-u,-v)$ is just the complex conjugate $F^*(u,v)$ [Bracewell95 p. 208], and we obtain:

$$C_{f,f}(u,v) = F(u,v)F^*(u,v) = |F(u,v)|^2 \quad (\text{E.24})$$

This means that the Fourier transform of the autocorrelation function $c_{f,f}(x,y)$ is the *power spectrum* of $f(x,y)$.⁹ This result is also known as the Wiener-Khintchine theorem.

In the case of *finite-power signals* (see Sec. E.4), the Fourier transform does not always exist; for example, stationary random functions do not have Fourier transforms [Champeney73 p. 59].¹⁰ In such cases Eqs. (E.20) and (E.22) do not hold. And yet, it turns out that Eq. (E.24) does have a valid counterpart: As explained in [Coulon84 p. 93] for the 1D case, if $f(x)$ is a finite-power signal, then $r(x,a) = f(x) \text{rect}(x/a)$ is a finite-energy signal that satisfies $f(x) = \lim_{a \rightarrow \infty} r(x,a)$. Let $R(u,a)$ be the Fourier transform of $r(x,a)$; by the 1D analog of Eq. (E.24) we have $C_{r,r}(u,a) = |R(u,a)|^2$; and it turns out that:¹¹

$$C_{f,f}(u) = \lim_{a \rightarrow \infty} \frac{1}{a} |R(u,a)|^2 \quad (\text{E.25})$$

In the particular case of finite-power signals where the given functions f and g are periodic with the same period T the convolution and cross-correlation theorems do hold, but they must be adapted accordingly, replacing convolution and cross correlation by T -convolution and T -cross correlation. This gives the T -convolution and the T -cross-correlation theorems. More details can be found in Secs. 4.2 and 4.3 of *Vol. I* and in [Champeney87 p. 166].

⁹ Note the major difference between this result and its counterpart for *autoconvolution* which is given in Eq. (E.21). The difference is that unlike $F(u,v)^2$, the power spectrum $|F(u,v)|^2$ contains no phase information; this also means that the autocorrelation function $c_{f,f}(x,y)$, too, unlike the autoconvolution function $h(x,y)$, contains no information about the phase of $f(x,y)$ [Bracewell86 p. 115]. This difference originates from the fact that for any complex number $z = |z|e^{i\theta}$, the value $|z|^2$ is purely real and has no phase component, whereas the value z^2 equals $|z|^2 e^{i2\theta}$ and in general has a non-zero phase component. Remark that the loss of phase information occurs either while taking the cross correlation ($f(x,y) \Rightarrow c_{f,f}(x,y)$) or while taking the power spectrum ($f(x,y) \Rightarrow |F(u,v)|^2$), but not while taking the Fourier transform ($c_{f,f}(x,y) \Rightarrow C_{f,f}(u,v)$). The Fourier transform does not cause any loss of information.

¹⁰ Note that this fact concerns *theoretical* stationary random functions, that extend throughout the range $-\infty < x < \infty$, but not their *practical* approximations whose spatial extent is finite.

¹¹ Note that in some references such as [Champeney73, Chapter 4] it is customary to call this limit the *power spectrum* of the finite-power signal $f(x)$, and to use the term *energy spectrum* for $|F(u)|^2$ in cases where $f(x)$ is a finite-energy signal. However, we do not follow this convention, and prefer to use the same term, power spectrum, in all circumstances.

E.6 Methods for quantifying the correlation; similarity measures

It is often desirable to compare the degree of correlation between functions; for example, one may want to know if the functions $f_1(x,y)$ and $f_2(x,y)$ are more correlated (i.e. more similar to each other) than $f_1(x,y)$ and $f_3(x,y)$. How can we formally formulate the degree of correlation between two functions (or in our application, between two images)? As we have seen, the cross correlation operation between two functions $f(x,y)$ and $g(x,y)$ may be a useful starting point. However, the cross correlation is again a 2D function, and the question remains how we can use it to compare the degree of correlation between the two images; in other words, how can we extract from the 2D cross correlation (or even better, from the normalized 2D cross correlation) a single number, known as a matching score, that can be used to evaluate the similarity between $f(x,y)$ and $g(x,y)$? In fact, this question does not have a unique answer, and one could think of several ways of doing so. Let us evaluate here some of the most plausible methods that may come to one's mind.

- (1) The volume under the cross correlation. This seems to be a promising method, but in fact it turns out to be useless. Suppose, for example, that $g(x,y)$ is a rotated version of $f(x,y)$. Clearly, the more $g(x,y)$ is rotated, the lower the similarity between $g(x,y)$ and $f(x,y)$. However, as we have seen, the volume under the cross correlation equals the product of the volumes under $f(x,y)$ and under $g(x,y)$; but the volumes under both $f(x,y)$ and $g(x,y)$ are independent of the rotation angle. This means that the area under the cross correlation remains constant when we rotate $g(x,y)$, and therefore it cannot be used to determine the degree of correlation between the two images.
- (2) The maximum value or the maximum absolute value of the cross correlation. This method is used, indeed, in *template matching*, where the problem is to find the closest match between a given unknown image and a set of known images.¹² In this approach one computes the cross correlation between the unknown and each of the known images, and the *closest match* is then found by selecting the image that yields the cross-correlation function with the largest value. Since the resultant cross correlations are 2D functions, this involves searching for the largest amplitude of each such function [Gonzalez87 p. 92]. Note that if we only look for the position of $g(x,y)$ in which it is most similar to $f(x,y)$ (for example, if we search the position of a letter "M" within an image consisting of some given text) then the procedure is simpler: In this case the desired location is simply the point (x,y) for which the cross correlation of the two images has the highest value [Gonzalez87 pp. 425–427]. If $g(x,y)$ is identical to $f(x,y)$ the highest value is located at the origin, and it equals the volume under the function $[f(x,y)]^2$ (see properties (2) and (3) of autocorrelation in Sec. E.3).

One may also think of other matching scores based on cross correlation to quantify the similarity between two functions. It should be remembered, however, that the cross-correlation operation (and hence, any similarity score that is based on it) is only capable of detecting similarities between functions $f(x)$ and $g(x)$ that are basically related by a simple

¹² Such situations frequently occur in digital image processing when the images in question are discrete, but the principles are the same for both discrete and continuous cases.

relationship of the type: $g(x) = cf(x - b)$, namely, *amplitude scaling* and *lateral shift*. Even simple linear relationships such as $g(x) = f(ax)$, let alone non-linear relationships such as $g(x) = f(x^2)$, cannot be uncovered by cross correlation [Cartwright90 p. 174]. The 2D counterpart of this fact is, indeed, nicely illustrated in our discussion on Glass patterns in Sec. 7.8. In order to overcome this significant restriction, it is customary to *undo* any geometric transformations that were undergone by the signals (or images) to be compared — of course, provided that these transformations are known in advance, or that they can be estimated (for example, there exist methods based on log-polar mapping or on the Mellin transform for recovering scale and rotation transformations that were undergone by an image; see, for instance, [Feitelson88 Sec. 3.2], [Hotta99] and [Casasent76]).

It should be noted that methods based on cross-correlation are not the only ones that can be used to estimate the degree of similarity between two functions $f(x)$ and $g(x)$. Other similarity metrics include the *coherence* [Cartwright90 pp. 179–180; Coulon84 p. 146]; the *scalar product* of f and g , $s[f, g] = \int f(x)g(x) dx$ (or its normalized version, that is called in [Cartwright90 pp. 169–171] the *correlation coefficient* of the functions f and g); the Euclidean distance $d_2[f, g] = [\int |f(x) - g(x)|^2 dx]^{1/2}$; and other distance functions that are based on different norms such as $d_1[f, g] = \int |f(x) - g(x)| dx$ or $d_\infty[f, g] = \sup |f(x) - g(x)|$ [Lipschutz01 pp. 252–254].¹³ An example showing a graphical comparison between some of these distances can be found in [Coulon84 pp. 44–46]. Among the distance functions the Euclidean distance is the most useful, but the other distances are occasionally used, too, either because they are better adapted for a given context, or simply because they allow easier calculation [Coulon84 p. 43].

Note that unlike cross correlation all of these techniques yield the distances between the given functions as a single number and not as a function, and they do not take into account displacements between the given functions.

Remark E.2: Some of the similarity measures mentioned above are, indeed, interrelated. For example, the Euclidean distance between two functions f and g is related to their scalar product by the relation [Coulon84 p. 48]:

$$(d_2[f, g])^2 = s[f, f] + s[g, g] - 2s[f, g] \quad (\text{E.26})$$

In particular, if f and g are orthogonal, meaning that $s[f, g] = 0$, this relation reduces into the Pythagoras theorem:

$$\begin{aligned} (d_2[f, g])^2 &= s[f, f] + s[g, g] \\ &= (d_2[f, 0])^2 + (d_2[g, 0])^2 \end{aligned}$$

There also exists a relationship between the cross correlation of f and g and their scalar product: If we denote by g_x the counterpart of g which has been translated by x , namely,

¹³ Note that the entities we denote here by lower case, such as $s[f, g]$, $d[f, g]$ etc. are not *operators*, like $C[f, g]$, but *functionals*: for any two functions f and g they yield a *number*, and not another function.

$g_x(x') = g(x' - x)$, then the cross-correlation function $c_{f,g}(x)$ expresses the evolution of the scalar product as a function of the translation x :

$$c_{f,g}(x) = s[f, g_x] \quad (\text{E.27})$$

Similarly, there also exists a relationship between the cross correlation of f and g and the Euclidean distance between them as a function of the translation x [Coulon84 p. 81]:

$$(d_2[f, g_x])^2 = c_{f,f}(0) + c_{g,g}(0) - 2c_{f,g}(x) \quad (\text{E.28})$$

(using property (3) of autocorrelation). Note that Eq. (E.26) is simply a particular case of this relationship where the translation is $x = 0$. ■

Remark E.3: At each value of the translation x where the cross-correlation function $c_{f,g}(x)$ equals zero, the functions f and g are *non-correlated*, and the functions f and g_x are *orthogonal*. ■

Remark E.4: It may be theoretically possible to define other variants of cross correlation (and convolution) that would account for other misregistrations of $g(x,y)$ than simple translation. For example, a variant of cross correlation that accomodates *scalings* of $g(x,y)$ rather than translations could be defined by:

$$s_{f,g}(a,b) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g(ax', by') dx'dy'$$

and a variant that accomodates both scalings and translations could be defined by:

$$s_{f,g}(x,y,a,b) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g(ax' - x, by' - y) dx'dy'$$

This last variant would give peaks in the 4D space spanned by the x , y , a and b axes at those values of (x,y,a,b) for which the best correlation is found between $f(x,y)$ and the scaled and shifted variants of $g(x,y)$. Similar definitions could be also provided to allow for rotations, general linear or affine transformations, or even non-linear transformations. Such an approach has been presented, for the discrete case, in [Pratt91 pp. 669–671].

However, the interest in such generalizations is usually limited, because the same effects can be obtained by first undoing the transformations undergone by $g(x,y)$ (and possibly also by $f(x,y)$), and then applying standard cross correlation to the untransformed $f(x,y)$ and $g(x,y)$. This approach is more economical in terms of computational load since it does not require to run throughout a four- or even higher-dimentional space, and moreover, it allows us to use the large body of already existing theoretical results (including the convolution and the cross-correlation theorems) without having to adapt them (if at all possible) to each particular family of transformations we might wish to consider. ■

Appendix F

The Fourier treatment of random images and of their superpositions

F.1 Introduction

The Fourier theory is very well adapted to the study of random structures, too (see, for example, [Bracewell86 Chapters 15–16], [Champeney73 Chapter 6] or [Coulon84 Chapters 5–6]). The main difference between the Fourier treatment of deterministic signals and that of random signals is that in the latter case all the phase information is lost, and we no longer have the full spectral representation of the signals but only their power spectra (see, for example, [Castelman79 pp. 199–201], [Bracewell86 p. 381]).

And yet, as we will see below, given a *fully known* random image such as the random dot screen shown in Fig. 2.1(a), we can still consider it as being deterministic, and treat it just like any other image. For example, we can apply to it the Fourier transform and obtain its full spectral representation, including all the phase information. In this appendix we evaluate both the stochastic and the deterministic approaches in the particular context of the moiré theory (the investigation of Glass patterns between random layers). We first provide in Secs. F.2–F.4 a short review of the stochastic approach, and then, in Sec. F.5 we explain why we have chosen to use the deterministic rather than the stochastic approach.

Note that whenever we wish to cover both the 1D and the 2D cases we use the generic term “signal” rather than our usual term “image”, which has a strong 2D connotation.

F.2 Stochastic processes and their power spectra

Signals can be classified into two distinct categories: deterministic signals and random (or stochastic) signals. A signal is said to be *deterministic* if its values are fully known throughout its domain of definition, or if its values can be predicted by an appropriate mathematical formula or model. On the other hand, a signal is said to be *random* (or *stochastic*) if its precise values are not fully predictable and they depend, at least partly, on the laws of probability; such a signal does not have an analytic representation and it can only be described using statistical considerations.¹ As we clearly see from this definition, a signal does not need to be completely unknown in order to be random; for example, a signal whose shape is fully known and only its position along the axes is unknown (such

¹ Stated in other words, the values of such a signal $f(x)$ do not depend in a completely definite way on the independent variable x , as in a deterministic signal; instead, the evolution of the signal depends also on chance, so one gets in different observations different realizations of $f(x)$.

as a sinusoidal signal whose initial phase is unknown) is already considered as a random signal. Note, however, that although a random signal does not have an analytic representation, it still can be characterized by its *statistical* properties (such as the probability distribution of its values, its average, its standard deviation, etc.) and by its *frequential* properties (its spectral decomposition in terms of its power spectrum).

Any observed random signal should be considered as one particular case among all the similar signals that could be produced by the same phenomenon or random process. Mathematically, a *random process* (or a *stochastic process*) is defined as an ensemble of signals, $\{f_{\zeta}(\mathbf{x})\}$, where the variable \mathbf{x} represents a point in the signal's domain of definition (for example, along the time axis in the 1D case or within the x,y plane in the 2D case), and the variable ζ represents an element of the ensemble, i.e. one among all the possible signals that may result from the same statistical experiment [Papoulis65 pp. 279–281]. Note that each of the variables ζ and \mathbf{x} may be either continuous or discrete. A stochastic process is said to be *continuous* or *discrete* depending on whether ζ is continuous or discrete; if the variable \mathbf{x} is discrete the process is called a *random sequence*, and if both ζ and \mathbf{x} are discrete the process is called a *point process*. Each member of the ensemble $\{f_{\zeta}(\mathbf{x})\}$ is called a *particular instance* or a *particular realization* of the random process, and is, in itself, a random signal.²

The concept of a stochastic process can be illustrated pictorially as a series of random functions $f_{\zeta}(\mathbf{x})$ that are stacked along the ζ axis. Assume that we draw the 3D x,y,z space such that the x axis points to the right within the paper's plane, the z axis points upward within the paper's plane, and the y axis is perpendicular to this plane and points toward the observer. If the variable \mathbf{x} is 1D, we can identify the ζ axis with our y direction, so that each of the functions $z = f_{\zeta}(\mathbf{x})$ of the random process can be drawn within the paper's plane or parallel to it along the ζ axis (see Fig. 5.1 in [Coulon84 p. 112] or Fig. 1.1 in [Bendat93 p. 2]). If the variable \mathbf{x} is 2D, the random process is best represented as a stack of planar gray level plots $f_{\zeta}(x,y)$ that is viewed from a lateral perspective, where the ζ axis (the stacking direction) coincides with our z direction (see Fig. 8 in [Rosenfeld82 p. 40]).

Clearly, statistical properties of a stochastic process such as its mean value can be approached in two different ways: they can be either computed along the x axis (or the x,y axes in the 2D case), or along the ζ axis. In the first case the mean value we obtain is called a *time average* (or *spatial average*), while in the second case, when the mean is computed over the different realizations of the random process, the value we obtain is called an *ensemble average*. A stochastic process is said to be *ergodic* if (1) the time averages of all its member functions are equal; (2) the ensemble average is constant with time; and (3) the time average and the ensemble average are equal [Castelman79 p. 200].

² It should be stressed, however, that once a random signal has been observed or recorded, all its values are fully known, and it should be therefore considered as a deterministic signal, albeit of random origin. This fact may cause some terminological confusion, because such a signal is still very often called a random signal. For example, all the random dot screens and line gratings in the figures throughout this volume are, in fact, deterministic signals of random origin, because they are fully known. But in most cases the intended meaning can be understood from the context without difficulty.

(Note that this definition is formulated here for the 1D case, but its 2D counterpart is easily obtained by replacing the term “time average” by “spatial average”). Thus, for ergodic processes, time averages (or spatial averages) and ensemble averages are interchangeable. This is also true for higher-order statistical averages such as the mean quadratic value of the process, its standard deviation, etc. (see Table 5.2 in [Coulon84 p. 117]). Hence, whenever ergodicity can be assumed one may compute the statistical properties of the stochastic process by analyzing the temporal (or spatial) behaviour of a single signal (a single member of the ensemble), which is obviously much easier to do in practice. Fortunately, ergodic processes model commonly encountered random signals quite well [Castelman79 p. 201]. In the following we will assume, indeed, that all our stochastic processes are ergodic.

We now proceed to the spectral properties of a random process; but before doing so, let us first consider its autocorrelation function. The autocorrelation function of a signal is defined as a time (or spatial) average by:

$$c_{f,f}(x,y) = f(x,y) \star \star f(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x',y') g(x' - x, y' - y) dx'dy'$$

(see Sec. E.3 in Appendix E). In an ergodic stochastic process the autocorrelation function is the same for all member signals, and thus it characterizes the ensemble. Therefore, although the precise values of an ergodic stochastic process are unknown, its autocorrelation function *is fully known*; this function reflects, in fact, our partial knowledge of the stochastic process. Now, since the autocorrelation $c_{f,f}(x,y)$ of the process is known, its Fourier transform $C_{f,f}(u,v)$ is also known; but according to the autocorrelation theorem (see Sec. E.5 in Appendix E), this is precisely the power spectrum of the process. This means that the power spectrum of a stochastic process is also fully known, just like the autocorrelation function; in fact, both of them contain the same information, which is only presented in a different way, in the spectral domain or in the image domain. It is important to note, however, that although we know the *power spectrum* of the stochastic process, and hence its *amplitude spectrum*, too (which is simply the square root of the power spectrum), we do not know the *phase spectrum* of the stochastic process.³ This means that unlike in deterministic signals, random signals do not have a full Fourier spectrum, and one can only deal with their power (or amplitude) spectra [Castelman79 p. 201; Bracewell86 p. 381]. Another explanation why random signals only have power spectra is provided in [Champeney73 pp. 79–80]: it turns out that any attempt to generalize the Fourier transform to random signals is doomed to be unsuccessful, while the power spectrum *can* be generalized to such cases.

The power spectrum of a stochastic process can be seen as the mean (i.e. expectance) of the power spectra of all the infinitely many individual signals that make up the stochastic process [Coulon84 pp. 135–136]. Under the assumption of *ergodicity*, this is also equal

³ Remember that the Fourier transform $F(u)$ of a function $f(x)$ is a complex-valued entity, so that it can be presented either in terms of its real part $\text{Re}[F(u)]$ and its imaginary part $\text{Im}[F(u)]$, where $F(u) = \text{Re}[F(u)] + i\text{Im}[F(u)]$, or, equivalently, in the polar representation, in terms of its amplitude spectrum $\text{Abs}[F(u)]$ and its phase spectrum $\text{Arg}[F(u)]$, where $F(u) = \text{Abs}[F(u)] \cdot e^{i\text{Arg}[F(u)]}$.

to the mean power spectrum (E.25) of any individual signal (particular instance) of the process [Papoulis65 p. 343]. Due to the averaging effect the power spectrum of a stochastic process is much smoother than the power spectra of the individual signals of the process: Usually, because the structure of each individual signal of the process is rather irregular, its power spectrum may be jumpy or noisy, and it often admits a typical *diffuse* appearance (see [Bracewell95 pp. 586–590, 600–601] and Problem 2-2). However, in the power spectrum of a stochastic process the fluctuations that are inherent to each of the individual spectra are smoothed out, so that we are left with a “clean” representation of the net spectral behaviour of the process. This ideal average power spectrum, in which all random effects have been averaged out, may typically consist of smooth curves and isolated impulses, but it is no longer buried in a diffuse random background noise. Several pictorial illustrations of such power (or rather amplitude) spectra can be found in the central column of the figure in [Coulon84 p. 500].

This clean look of the theoretic, average power spectrum of the stochastic process certainly facilitates the understanding of the underlying spectral information, and is therefore more attractive to use than the noisy spectra of the individual random signals.⁴ However, if an individual signal of the process is fully known, we still may prefer to consider it as a *deterministic* signal, even though it originates from a random process. This would allow us to take the Fourier transform of the signal and obtain its full spectral information, i.e. both its amplitude spectrum and its phase spectrum (although both of them would usually contain some diffuse noise, too). This would also allow us to freely pass between the original signal in the image domain and its spectrum in the frequency domain and vice versa, and to benefit from fundamental results such as the convolution theorem (which allows us to consider products in one domain as convolutions in the other domain and vice versa). On the other hand, when we are dealing with a random process we have no longer access to the Fourier spectrum but only to the amplitude (or power) spectrum, and all the phase information (the phase spectrum) is lost. This means that we can no longer freely pass between the image and spectral domains, and furthermore, we can no longer use the convolution theorem (note that power spectra have no equivalent to the convolution theorem; more about the properties and non-properties of power spectra can be found in [Champeney73 pp. 64–65]). We will return to these considerations in more detail in Sec. F.5, where we evaluate the stochastic and the deterministic approaches in the particular context of the moiré theory.

Remark F.1: Note that in a deterministic signal $f(x)$ the power spectrum $\mathcal{P}_f(u)$ can be obtained in two different ways: Either directly from the Fourier transform $F(u)$ of the signal: $\mathcal{P}_f(u) = |F(u)|^2$, or indirectly, using the autocorrelation theorem (see Sec. E.5 in Appendix E). In the indirect way we first have to find the autocorrelation function $c_{f,f}(x)$ of the given signal $f(x)$, and then we apply to it the Fourier transform: $\mathcal{P}_f(u) = C_{f,f}(u)$. However, if $f(x)$ is not a deterministic signal but a random process the direct approach is no longer available, because a random process does not have a Fourier transform. In this

⁴ A more detailed discussion on the averaged, limit power spectrum and its possible uses can be found in [Gardner88 pp. 5–6, 67–72].

case the power spectrum can only be found in the indirect method, using the autocorrelation theorem; this also explains why the power spectrum of a random process is often defined in the literature as the Fourier transform of the autocorrelation (see, for example, [Papoulis65 p. 338]). ■

F.3 Possible stochastic modelizations of random screens and gratings

Stochastic processes may be used to modelize many different types of random images; a non-exhaustive, illustrated survey of several different types of such random images can be found in Chapter 17 of [Bracewell95]. In the present section we briefly review some of the stochastic processes that can be used for the modelization of random images of the types we mostly use, such as dot screens, line gratings, etc. The different models are presented in increasing order of their suitability to our needs.

F.3.1 Point processes

The simplest and most natural approach for the stochastic modelization of random structures such as random dot screens is based on *spatial point processes*. A spatial point process is any stochastic mechanism which generates a *point pattern*, i.e. a countable set of points \mathbf{x}_i in the plane [Diggle83 Chapter 4]. Clearly, such a model only takes into account the random locations of our screen elements, but not the elements themselves (their shapes, sizes, etc.). In fact, we can think of each particular instance of the spatial point process as a nailbed consisting of randomly located impulses; our random dot screen can be then considered as the convolution of this random nailbed with a single dot, as we will see in the following subsection.

The simplest point process is the *Poisson process*. This is, indeed, the cornerstone on which the theory of spatial point processes is built. This model is used in applications as an idealized standard of *complete spatial randomness*; although unattainable in practice, it often provides a useful approximate description of an observed pattern [Diggle83 p. 50]. Informally, a Poisson process can be seen as a 1D or 2D impulse train (i.e. a comb or a nailbed) whose individual impulses are positioned at random:

$$q(x) = \sum_n \delta(x - x_n) \quad (\text{F.1})$$

The formal requirements on (F.1) for being a Poisson process (i.e. the formal requirements for the dot locations x_n to be fully random) are given, for example, in [Diggle83 p. 50].

The statistical properties of a Poisson process are well established, and they can be found in the literature (see, for example, [Diggle83 pp. 50–51] or [Papoulis65 pp. 284–287]; note that the latter uses the term “Poisson impulses” for what we call here a Poisson process). Let us briefly consider now the *spectral* properties of such a process.

The power spectrum of a Poisson process (i.e. the power spectrum of a random nailed) is given by (see, for example, [Champeney73 pp. 82–83]):

$$\mathcal{P}_q(u) = \eta + \eta^2 \delta(u) \quad (\text{F.2})$$

where η is the average number of dots per unit interval.⁵ Upon Fourier inversion this also gives us the autocorrelation function of the Poisson process:

$$c_{q,q}(x) = \eta \delta(x) + \eta^2 \quad (\text{F.3})$$

As we can see from Eq. (F.2), the power spectrum of a random impulse train is evenly distributed over all frequencies, and the Poisson process is therefore an example of “white noise” (with an additional DC impulse at the spectrum origin). Note, however, that this power spectrum is not unique to random combs, and it is shared with infinitely many other random processes, including continuous, non-impulsive ones (remember that unlike the Fourier transform $F(u)$, the power spectrum $\mathcal{P}_f(u)$ does not uniquely originate from a single function $f(x)$).

Apart from the Poisson process there exist many other types of point processes, having different statistical distribution rules for the random points and different spatial and spectral properties. These processes include, for example, inhomogeneous Poisson processes, clustered processes, inhibition processes, Markov point processes, lattice-based processes, etc. A short survey on some of the main types of point processes and their properties can be found, for example, in [Diggle83, Chapter 4].

It should be noted, however, that all point processes are based on the simplifying assumption that we are dealing with a random arrangement of *impulses*; as already mentioned above, this means that point processes can only modelize the locations of our screen dots, but they do not take into account the actual geometric shapes and sizes of the individual dots. If we wish to obtain a better approximation by taking also into consideration the shapes of the individual elements of our random layers (dot screens, line gratings, etc.), we may use another type of random process which is known as *shot noise*.

F.3.2 Shot noise

A *shot noise* random process ([Champeney73 p. 82], [Papoulis65 p. 288]) is a random process whose individual signals are produced as a sum of randomly located copies of a given function (for example, a dot or a pulse). In the 1D case, if the individual dot has the profile $d(x)$ then the shot noise process is given by:

$$f(x) = \sum_n d(x - x_n) \quad (\text{F.4})$$

where the values x_n form a random sequence. As we can easily see, this is precisely the convolution of the random nailed (F.1) with the dot $d(x)$:

⁵ This last value is usually denoted in the literature by λ , but we prefer to use here η since we have already used λ in a different context, that of colour layers (see Chapter 9 in *Vol. I*).

$$f(x) = d(x) * q(x) \quad (\text{F.5})$$

Remark that we denote here an individual element of the shot noise process by $d(x)$ and call it a *dot* in anticipation of the 2D case where $d(x)$ represents a 2D dot (like in a dot screen). It should be understood, however, that in our application a general random dot screen can only be *approximated* by shot noise, since in shot noise overlapping dots are summed up (as clearly indicated by Eq. (F.4); see also Fig. 9–8 in [Papoulis65 p. 288]), whereas in random screens overlapping black dots remain black and overlapping white dots remain while. This means, indeed, that a multiplicative version of shot noise would be better adapted to our needs; but under the assumption that the random screen dots do not overlap, a random screen made of white dots on black background can be aptly modeled by the shot noise $f(x)$, and its inverse video, consisting of black dots on white background, can be modeled by $1 - f(x)$.

Proceeding now to the spectral domain, it is clear that if the signal $f(x)$ is deterministic, then we immediately have its Fourier transform (using the shift and addition theorems):

$$F(u) = D(u) \sum_n e^{-i2\pi x_n u}$$

On the other hand, if we consider $f(x)$ as a random process we no longer have its Fourier transform, since all the phase information is lost. And yet, we do have a simple expression for the *power spectrum* of the random process $f(x)$: Using the fact that the power spectrum of a convolution of two functions is the product of their individual power spectra [Champeney73 pp. 64–65] we get from Eq. (F.5):

$$\mathcal{P}_f(u) = \mathcal{P}_d(u) \mathcal{P}_q(u) \quad (\text{F.6})$$

where $\mathcal{P}_f(u)$, $\mathcal{P}_d(u)$ and $\mathcal{P}_q(u)$ are, respectively, the power spectra of the random process $f(x)$, of the individual dot $d(x)$ and of the Poisson process $q(x)$. Therefore, using Eq. (F.2) we obtain (see also [Papoulis65 p. 358]):

$$\begin{aligned} \mathcal{P}_f(u) &= \eta \mathcal{P}_d(u) + \eta^2 \mathcal{P}_d(u) \delta(u) \\ &= \eta \mathcal{P}_d(u) + \eta^2 [D(0)]^2 \delta(u) \end{aligned}$$

where $\delta(u)$ is an impulse at the origin, and $D(u)$ is the Fourier transform of the dot $d(x)$. There also exist simple expressions for the autocorrelation $c_{f,f}(x)$, the mean value μ and the variance σ^2 of the shot noise process $f(x)$ (see, for example, [Champeney73 pp. 82–83]):

$$\begin{aligned} c_{f,f}(x) &= \eta c_{d,d}(x) + \mu^2 \\ \mu &= \eta \int d(x) dx = \eta D(0) \\ \sigma^2 &= \eta \int [d(x)]^2 dx \end{aligned}$$

Using the above expression for μ the power spectrum of the shot noise process becomes (see also [Champeney73 pp. 82, 230–231]):

$$\mathcal{P}_f(u) = \eta \mathcal{P}_d(u) + \mu^2 \delta(u) \quad (\text{F.7})$$

It may be instructive to notice the smoothness of this power spectrum: indeed, it is composed of a smooth curve plus an impulse at the origin, but it includes no diffuse random noise. As explained above in Sec. F.2, this property is common to the power spectra of all random processes due to the averaging effect that is inherent to their definition.

A particular case of Eq. (F.7) occurs when the mean value μ of $f(x)$ is zero. In this case the impulse at the origin disappears and we obtain (see [Champeney73 pp. 82, 230–231]):

$$\mathcal{P}_f(u) = \eta \mathcal{P}_d(u) \quad (\text{F.8})$$

It should be remembered, however, that all the above power spectra are based on the power spectrum $\mathcal{P}_q(u)$ of the Poisson process; this means that they are based on the assumption that the dot locations are fully random and hence uncorrelated. If the dot locations (i.e. the impulse locations of the underlying point process) *are* correlated, so that for a dot which has occurred at x_1 the probability of having another dot in the infinitesimal interval between $x_1 + x$ and $x_1 + x + dx$ is, say, $p(x)dx$, then Eq. (F.7) becomes [Champeney73 p. 231]:

$$\mathcal{P}_f(u) = \eta \mathcal{P}_d(u) (1 + P(u)) \quad (\text{F.9})$$

where $P(u)$ is the Fourier transform of the probability density function $p(x)$. This means that if the dot locations *are* correlated the power spectrum is no longer determined by $\mathcal{P}_d(u)$ alone, and it includes an additional term which depends on $P(u)$, too. Consider, for example, the 2D case of a random dot screen in which the distances between neighbouring dots are highly correlated, with a characteristic nearest-neighbour distance of r . Such dot screens tend to give a ring-like power spectrum where the mean radius of the ring is $1/r$ (see, for example, the figures in [Yellott82] and [Yellott83], Plates 15–16 in [Harburn75], or Fig. 10.25 in [Glassner95 p. 433]; in all of these cases the ring-like envelope of the power spectrum is not explained by the shape of the power spectrum $\mathcal{P}_d(u, v)$ of the individual dot $d(x, y)$, but rather by the shape of $P(u, v)$, the Fourier transform of the probability density function $p(x, y)$).

Further generalizations of Eq. (F.9) are also possible. For example, [Heiden69] derives the power spectrum of shot noise processes whose pulse widths and pulse amplitudes are not constant but rather random entities that are correlated with the pulse locations.

Finally, it should be mentioned that the above power spectra are only valid when the process $f(x)$ is not periodic. If $f(x)$ is periodic its power spectrum no longer contains the continuous component $\nu \mathcal{P}_d(u)$, and it becomes purely impulsive. This can be easily seen by reconsidering Eqs. (F.5) and (F.6): Suppose, for example, that $f(x)$ is periodic with period 1. If we denote by $\text{III}(x)$ the unit-period comb of impulses, we have in this case instead of Eq. (F.5):

$$f(x) = d(x) * \text{III}(x)$$

and the power spectrum of $f(x)$ becomes:

$$\mathcal{P}_f(u) = \mathcal{P}_d(u) \mathcal{P}_{\text{III}}(u) \quad (\text{F.10})$$

where the power spectrum $\mathcal{P}_{\text{III}}(u)$ of $\text{III}(x)$ is the unit-frequency impulse comb $\text{III}(u)$. This means, indeed, that in this case $\mathcal{P}_f(u)$ is a unit-frequency comb whose impulse amplitudes are modulated by $\mathcal{P}_d(u)$, the power spectrum of the isolated dot $d(x)$.

A further generalization of Eq. (F.10) is provided in [Williams86] and more recently in [Ridolfi04 p. 67]. It is shown there that if the location of each dot in the periodic process $f(x)$ is slightly randomized (or “jittered”) where each dot is perturbed independently of the other dots, and the probability of each of the dots to lie in an infinitesimal area dx is given by $p(x)dx$, then the power spectrum of $f(x)$ becomes:

$$\mathcal{P}_f(u) = \mathcal{P}_d(u) \mathcal{P}_p(u) \mathcal{P}_{\text{III}}(u) + \eta \mathcal{P}_d(u) (1 - \mathcal{P}_p(u)) \quad (\text{F.11})$$

where $\mathcal{P}_p(u)$ is the power spectrum of the probability density function $p(x)$. The first term in (F.11) is a unit-frequency comb whose impulse amplitudes are modulated by the product $\mathcal{P}_d(u) \mathcal{P}_p(u)$, and the second term is a continuous function. This means that the power spectrum (F.11) is no longer purely impulsive: in addition to the impulses of Eq. (F.10) (whose amplitude is modulated here by $\mathcal{P}_p(u)$, too), it also contains a new continuous part whose shape depends on both $\mathcal{P}_d(u)$ and $\mathcal{P}_p(u)$. Note that in the particular case where the dot locations are fully periodic (no random perturbation in the dot locations) we have $p(x) = \delta(x)$, whose Fourier transform is $P(u) = 1$ and whose power spectrum is therefore $\mathcal{P}_p(u) = 1$; and indeed, putting this back in Eq. (F.11) gives us again, as expected, the purely impulsive power spectrum of Eq. (F.10).

F.3.3 Random fields

A random field is a stochastic process $\{f_\zeta(\mathbf{x})\}$ whose argument \mathbf{x} varies in a continuous fashion over some subset of \mathbb{R}^n , the n -dimensional Euclidean space [Adler81 p. ix]. Of course, in our application we will be most interested in the case of $n = 2$; in this case a random field is a family of 2D functions $\{f_\zeta(x,y)\}$ defined over the x,y plane (or a subset thereof). Typical examples of random fields include, for instance, the infinite family of functions $\{z = f_\zeta(x,y)\}$ that describe the height z of a wavy sea surface, or the surface of any rough plate [Adler81 pp. 1–4]. Unlike point processes and shot noise, which are based on a discrete distribution of points in the plane (the locations of the screen dots), random fields are completely general, and they can take into account all the properties of our original random layers, including the dot shapes, sizes, intensities, etc.

Because the study of random fields is, by definition, the study of random functions over some Euclidean space, this study can cover an extremely wide area, since any question that can be asked about an ordinary non-random function, or class of functions, can just as readily be asked about their random counterparts. Hence, the general theory of random fields is certainly at least as large as the general theory of functions. And indeed, adding a

random component to the theory of functions makes it much larger, more interesting, and often more complex subject [Adler81 p. 1].

Due to its vast extent, we will not attempt here to go into details of the random field theory. Interested readers can find further information on random fields and on their spectral representation in references such as [Rosenfeld82 pp. 38–47] or [Adler81]. A general overview on random fields is also provided in [EncStat82, Vol. 7, pp. 508–512].

F.4 Stochastic modelization of layer superpositions

Having understood how random layers (random dot screens, random line gratings, etc.) can be modeled as stochastic processes, we now proceed to the modelization of superpositions of such layers.

As we know from the deterministic case, the superposition of layers is usually best modeled as a *product* of the individual layers (although in some particular cases other models may be more appropriate; see Sec. 2.2). This can be extended to the stochastic case, too. Intuitively, the probability of seeing white at a point (x,y) of the layer superposition is the product of the probabilities of seeing white at the point (x,y) in each of the two original layers. However, this is only true if the two superposed layers are statistically independent of each other. But in our application, when we come to study the moiré or Glass patterns in the superposition of two layers, we cannot make the simplifying assumption that the two layers are independent; on the contrary, we clearly know that they *are* dependent (remember the high correlation that is required between the two layers in order that a Glass pattern be generated).

Although there exist in the literature models that allow the treatment of such layer superpositions (see, for example, the superposition of two Poisson point processes in [Stoyan95 pp. 152–153]), the fact that our original layers are not independent of each other considerably complicates things and may render a detailed investigation intractable. For example (see [Coulon84 p. 152]), if two stochastic signals are independent then the autocorrelation function of their product is the product of the individual autocorrelation functions, and the power spectrum of the product is the convolution of the individual power spectra; but if the signals are correlated these results do not necessarily hold.⁶

F.5 Evaluation of the stochastic vs. deterministic approaches for our application

The need for a statistical treatment may arise in several situations. The most obvious situation occurs when we do not have full information about the phenomenon that is to be

⁶ This is true for both random and deterministic signals; indeed, the counterpart of the convolution theorem for power spectra only holds for functions that are not correlated. Note, however, that the converse direction of the convolution theorem *does* hold, meaning that the power spectrum of the convolution of two signals is the product of the two power spectra [Champeney73 pp. 64–65].

treated (1D signal, 2D image, etc.), and we only know its statistical properties. This may be the case when we are dealing with a process whose physics is not fully understood, or with a process too complicated to analyze in detail. In other situations a statistical treatment may be needed even though the signal is deterministic (fully known). This may happen, for example, when we are given an ensemble of similar signals, and we simply do not know in advance which of them will occur. In this situation, too, we can only treat the general properties of the ensemble (such as the *average* power spectrum), but not the properties of an individual signal. The need to use a statistical approach may also arise when one is specifically interested in the statistical properties (distribution, mean value, standard deviation, etc.) of the given signals — which may be either deterministic or not — and in the statistical properties of various combinations of these signals (for example, in our case, the layer superposition).

It should be remembered, however, that even if the given signals were originally generated by a random process, once they are *fully known*, it is no longer *needed* to treat them statistically, and we can treat them like any other deterministic signals. In our case, for example, every given dot screen that we fully know can be considered as a deterministic signal and undergo a standard Fourier treatment even if its elements have been positioned in the plane in a random manner (see, for example, Problems 2-1–2-13 in Chapter 2).⁷ The advantage of doing so is that this way we have full access to all the information related to the signal, and we do not lose its phase information.

Let us try to see more closely what this means in our particular case of interest, the study of moiré effects in the superposition of dot screens or line gratings whose individual elements, dots or lines, are randomly positioned. For this end, let us consider Fig. 7.7 that illustrates the core of our Fourier-based explanation of the macroscopic moiré or Glass patterns which may occur in a layer superposition. If we choose to treat our given layers statistically, we no longer have access to their Fourier transforms, but only to their power spectra. However, in this case we lose all the Fourier considerations that we have used in Chapter 7: First of all, we can no longer pass freely between the image and spectral domains of Fig. 7.7, since the transition from the power spectrum back to the original images is not possible. But even worse, in this case we can no longer use the convolution theorem, since the power spectrum of the product is not necessarily equal to the convolution of the individual power spectra [Champeney73 pp. 64–65].

As an alternative approach, we may consider altering *both* rows of Fig. 7.7, namely, replacing the original layers shown in the top row of the figure (the image domain) by their respective autocorrelation functions, and the spectra of the original layers shown in the bottom row of the figure (the spectral domain) by the respective power spectra. In this case, the bidirectional Fourier relationship between the entities in the image and in the spectral domains remains fully available, thanks to the autocorrelation theorem which

⁷ In other words, when we superpose two random dot screens we are not actually superposing two *random processes*, but rather two *specific realizations* of these random processes; and the moiré (or Glass) pattern we obtain is, again, a *specific realization*, not an ensemble of moiré (or Glass) patterns.

states that the autocorrelation function and the power spectrum are a Fourier pair. This guarantees, indeed, that we can freely move from the image domain to the spectral domain and vice versa without any loss of information; furthermore, in this case we can also use the convolution theorem. However, this happy situation should not mislead us to believe that by passing to statistical reasoning we do not lose information: True, the Fourier transform itself does not cause any loss of information when we pass from one domain to the other; but it simply operates on entities in the image and in the spectral domains that *have already lost* all the phase information. Indeed, it is the passage from the original images to their autocorrelation functions (in the image domain), or equivalently, the passage from the Fourier transforms of the original images to the power spectra of the original images (in the spectral domain), that causes the loss of information. Specifically, returning to Fig. 7.7, we would no longer have in its upper row the original layers, their superpositions and their moiré phenomena, but only the corresponding autocorrelations. But this loss of information in the image domain along with the loss of the direct contact with the original layers themselves make our Fourier-based reasoning less effective in the explanation of the moiré phenomenon.

Another drawback of the stochastic approach is that it requires a full, prior mathematical knowledge of the underlying stochastic process, and if the problem at hand does not fall within the framework of a known stochastic process that has already been worked out in the known literature, one will have to invest some efforts in order to find its various properties. In fact, this is precisely the situation that we are facing in our application, since the statistical processes representing our individual layers may be quite complex (remember, for example, that overlapping dots in a random dot screen are not really summed up as in shot noise but rather multiplied). Furthermore, when we wish to study the superposition of two layers we cannot make the simplifying assumption that the two layers are not correlated; on the contrary, we do know that they are correlated, due to the high correlation that is required between the two layers in order that a Glass pattern be generated.

In conclusion, we see that in our application the use of the stochastic approach is not quite appropriate, because it is not easily tractable (at least if we require a good approximating model), and also because it causes the loss of important information. Therefore, since in our case the given random images and their superpositions are fully known, we have all good reasons to treat them as deterministic images instead of using a stochastic approach. This has also the important advantage of providing a unified treatment for all types of images, periodic, aperiodic or random.

Appendix G

Integral transforms

G.1 Introduction

In Chapter 7 we have presented the fundamental Glass pattern theorem for the superposition of geometrically transformed aperiodic layers. But although we mentioned there that this result is valid for both linear and non-linear layer transformations, we only explained it, via spectral domain considerations, for *linear* transformations (see Fig. 7.7). For the more general case involving non-linear transformations we only presented the image-domain results (see Fig. 7.12), without giving their spectral-domain interpretation as we did in Fig. 7.7 for the case of linear transformations. The reason is that in the case of non-linear transformations a general spectral-domain analysis is not possible, since there exists no general expression for the Fourier transform of $f(g(x))$ when $g(x)$ is non-linear. In the present appendix we present a generalized fourier decomposition of $f(g(x))$ that allows us to obtain our main results (including the fundamental Glass pattern theorem and its counterpart for periodic structures, the fundamental moiré theorem) even if $g(x)$ is non-linear. Although in this approach we lose the direct connection with the spectral, frequency domain (since the transformed domain no longer coincides with the Fourier, spectral domain), this approach still proves to be very useful.

G.2 Fourier decomposition of periodic and aperiodic structures

Let $f(x)$ be an aperiodic function whose Fourier transform is $F(u)$. This means, using the definition of the Fourier transform, that:

$$F(u) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi ux} dx \quad (\text{G.1})$$

Similarly, using the definition of the *inverse* Fourier transform we also have:

$$f(x) = \int_{-\infty}^{\infty} F(u) e^{i2\pi ux} du \quad (\text{G.2})$$

Note that Eq. (G.2) expresses the decomposition of the image-domain function $f(x)$ into its spectral components. This is, indeed, the continuous-frequency counterpart of the Fourier series decomposition of a *periodic* function $p(x)$, which is given by:

$$p(x) = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi f_n x} \quad (\text{G.3})$$

with the Fourier coefficients:

$$c_n = \frac{1}{T} \int_T p(x) e^{-i2\pi f_n x} dx \quad (\text{G.4})$$

where $f_n = n/T$, $n \in \mathbb{Z}$ (see Eqs. (A.5) and (A.6) in Appendix A of *Vol. I*). As explained in Sec. B.6 of *Vol. I*, $F(u)$ in Eq. (G.2) plays the same role as the Fourier coefficients c_n in Eq. (G.3), namely, it assigns the proper amplitudes (weights) to the various frequencies in the spectral decomposition of $f(x)$. The difference between the two cases is that in Eq. (G.3), where the given image-domain function $p(x)$ is periodic, the spectral decomposition only consists of a denumerable set of frequencies, $f_n = n/T$, $n \in \mathbb{Z}$, and the spectrum is impulsive; while in Eq. (G.2), where the image-domain function $f(x)$ is aperiodic, the spectral decomposition consists of a continuum of frequencies $u \in \mathbb{R}$. In this case the summation over the frequencies is no longer denumerable, and it turns into integration, the continuous counterpart of the discrete summation.

G.3 Generalized Fourier decomposition of geometrically transformed structures

Suppose now that the image-domain function $f(x)$ undergoes a mapping (coordinate transformation) $g(x)$, so that we obtain a new, transformed version of $f(x)$, that we denote by $r(x)$:

$$r(x) = f(g(x))$$

How does the application of the mapping $g(x)$ to $f(x)$ affect the spectrum of $f(x)$? As mentioned in Sec. 10.3 of *Vol. I*, when the transformation $g(x)$ is linear or affine, the spectrum $R(u)$ of the transformed function $r(x)$ can be readily expressed in terms of the original spectrum $F(u)$. However, in the more general case where $g(x)$ is non-linear, no general rule exists which tells us how the spectrum will be influenced. This renders the Fourier approach in the general case intractable unless the non-linear mapping $g(x)$ is particularly simple.

However, in Chapter 10 of *Vol. I*, where we studied geometric transformations of periodic functions $p(x)$, we have found a way to bypass this problem by representing the transformed function $r(x) = p(g(x))$ as a *generalized* Fourier series. As explained in Sec. 10.5 of *Vol. I*, instead of considering the spectrum $R(u)$ of $r(x) = p(g(x))$, whose analytical expression may be unknown or hard to find, we make the following two-step reasoning: We start with the Fourier decomposition (G.3) of the original periodic function $p(x')$, using here the variable x' rather than x :

$$p(x') = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi f_n x'} \quad (\text{G.5})$$

and then we make in this Fourier series the formal substitution $x' = g(x)$, keeping the coefficients c_n unchanged:

$$r(x) = p(g(x)) = \sum_{n=-\infty}^{\infty} c_n e^{i2\pi f_n g(x)} \quad (\text{G.6})$$

Although the resulting generalized Fourier series (G.6) is not the actual spectral decomposition of $r(x) = p(g(x))$, it still turns out to be extremely useful. As we have seen in Sec. 10.9 of *Vol. I*, this approach allowed us to analyze the superposition of

geometrically transformed periodic layers and the resulting moiré effects, without having to resort to the spectra $R(u)$ of these curvilinear structures; and indeed, this approach led us in *Vol. I* to the main results of Chapter 10, the fundamental moiré theorems for the superposition of geometrically-transformed periodic layers.

Based on the success of this approach in the case of periodic functions $p(x)$, it may be asked, therefore, whether a similar reasoning could be also used in our present case of interest, where the original functions $f(x)$ are *aperiodic*.

To answer this question, let us return to the spectral decomposition of the image-domain function $f(x')$ as given by Eq. (G.2), using the variable x' rather than x :

$$f(x') = \int_{-\infty}^{\infty} F(u) e^{i2\pi ux'} du \quad (\text{G.7})$$

Now, just as we did above in the case of discrete spectral decompositions, let us *formally* replace x' in this expression with $x' = g(x)$, keeping the “coefficients” (i.e. the Fourier transform) $F(u)$ unchanged:

$$r(x) = f(g(x)) = \int_{-\infty}^{\infty} F(u) e^{i2\pi ug(x)} du \quad (\text{G.8})$$

Obviously, this formal construct is not the spectral decomposition of $r(x)$, but some kind of generalization thereof. The actual spectral decomposition of $r(x)$ would be given, according to Eq. (G.2), by:

$$r(x) = f(g(x)) = \int_{-\infty}^{\infty} R(u) e^{i2\pi ux} du \quad (\text{G.9})$$

where $R(u)$ is the Fourier transform of $r(x)$; but as we have seen above, the problem here is that when the mapping $g(x)$ is non-linear, we do not always know the Fourier transform $R(u)$, and therefore we do not have the spectral decomposition (G.9), either. But although the generalized decomposition provided by Eq. (G.8) is *not* the spectral decomposition of $r(x)$, it still proves to be very useful for our needs, i.e. for understanding the behaviour of our aperiodic layers, their superpositions and their moiré (or Glass) patterns.

G.4 Integral transforms and their kernels

Before we go any further into these considerations, let us try to better understand the meaning of the construct provided by Eq. (G.8). So far we already know what this construct is *not* — it is not the Fourier spectral decomposition of the transformed function $r(x) = f(g(x))$ — but we would like now to see what exactly it *is*. For this end, we recall that the Fourier transform is, in fact, just one member from a larger class of transforms that are known as *integral transforms*. An integral transform $F_k(u)$ of a function $f(x)$ is defined by the integral:

$$F_k(u) = \int_a^b f(x) k(x, u) dx \quad (\text{G.10})$$

where the function $k(x,u)$ is called the *kernel* of the integral transform, and a and b are the integration limits. The class of functions to which $f(x)$ may belong and the range of the variable u are to be prescribed in each case; in particular, they must be so prescribed that the integral (G.10) converges [Churchill72 pp. 2, 317]. An integral transform is uniquely determined by its kernel $k(x,u)$ and by its integration limits a and b .¹ Note that an integral transform is in fact an operator that maps a given function $f(x)$ into another function $F_k(u)$; this is more clearly expressed by the notation $\mathcal{F}_k[f(x)] = F_k(u)$. The resulting function $F_k(u)$ is called the \mathcal{F}_k -transform of $f(x)$, and its variable u is known as the variable of the *transformed domain*.² For example, in the case where $k(x,u) = e^{-i2\pi ux}$, $a = -\infty$ and $b = \infty$ the resulting function $F_k(u)$ is the Fourier transform of $f(x)$, and its variable u represents the frequency in the Fourier, spectral domain.

In principle, any function $k(x,u)$ of two variables gives rise to an integral transform, but in practice, few such kernels yield useful transforms [Cartwright90 p. 195]. Some of the most useful kernels are listed in Table G.1 along with the integral transforms they provide. Note also that for any function $g(x)$, if $k(x,u) = g(u-x)$ then the integral transform $F_k(u)$ is simply the operation of convolution with $g(x)$, since for any given function $f(x)$ it yields the function $f(x) * g(x)$:

$$F_k(u) = \int f(x) g(u-x) dx$$

In particular, if $g(x) = \delta(x)$ we get the kernel $k(x,u) = \delta(u-x)$, and the resulting integral transform is the so-called “identity transform”, usually denoted by I , which maps any given function $f(x)$ into itself: $I[f(x)] = f(x)$.

As we can see, each kernel provides a different integral transform, which maps the given function $f(x)$ into a different function $F_k(u)$. While some of these integral transforms have very specialized uses and are rarely encountered, others have found very important applications in mathematics or in other fields. The most widely known integral transform is the Fourier transform, but in certain applications other integral transforms may prove to be more suitable. For example, because many functions have a Laplace transform but not a Fourier transform,³ the Laplace transform turns out to be more useful in numerous applications, such as in the solution of certain classes of differential equations [Cartwright90 p. 193]. However, the essential advantage of the Fourier transform over all the other integral transforms is its physical interpretability as a *frequency spectrum*

¹ Some references such as [Andrews03 p. 496] use an alternative notation, in which the integration limits are incorporated within the kernel itself, by multiplying $k(x,u)$ with suitable functions that take the value 1 within the integration range and the value 0 everywhere else. In this case the integration is always performed between $-\infty$ and ∞ , and the integral transform is uniquely determined by its kernel.

² Note the slight terminological ambiguity due to the use of the term “transform” for both the resulting function $F_k(u)$ and the operator \mathcal{F}_k itself. This ambiguity could be avoided, as done in [Churchill72 pp. 2–3], by using the term *transformation* for the operator, while keeping the term *transform* for the resulting function. But this convention would simply shift the ambiguity elsewhere, since we already use the term *transformation* in the context of coordinate transformations, or more generally, as a synonym for a mapping $g(x)$.

³ This happens, for instance, in functions with exponential growth such as $f(x) = e^x$ [Cartwright90 pp. 192–193].

[Bracewell86 p. 220]. Although the Laplace and the other transforms can be used as efficient mathematical tools for solving various problems, they do not provide the spectral decomposition (or the frequency content) of the given function $f(x)$. In such integral transforms the “transformed” domain is not the spectral domain, and its variable u does not represent frequencies as it does in the case of the Fourier transform. Therefore, in such cases we no longer speak of the “image domain” and the “frequency domain”, but rather of the “original domain” and the “transformed domain”.

Integral transform	$k(x,u)$	a	b	Ref.
Fourier transform	$e^{-i2\pi ux}$	$-\infty$	∞	p. 7
Laplace transform	e^{-ux}	$-\infty$	∞	p. 219
Cosine transform	$2\cos(2\pi ux)$	0	∞	p. 17
Sine transform	$2\sin(2\pi ux)$	0	∞	p. 17
Hankel transform	$2\pi x J_0(2\pi ux)$	0	∞	p. 248
Mellin transform	x^{u-1}	0	∞	p. 254
Abel transform	$\frac{2x}{\sqrt{x^2 - u^2}}$	0	∞	p. 262
Hilbert transform	$\frac{1}{\pi(x - u)}$	$-\infty$	∞	p. 267

Table G.1: Some of the most useful integral transforms, their kernels and their integration limits. The page numbers in the last column refer to [Bracewell86]. Note that many of these transforms have in the literature several different variants that differ from each other in their kernel $k(x,u)$ or in the integration limits a, b . For example, different variants of the Fourier transform are given in [Bracewell86 p. 7]. Obviously, each of these variants maps the given function $f(x)$ into a different function $F_k(u)$, and it could be therefore included in the table as a new entry. Note also that many of the transforms in the table have a distinct inverse transform, which is also an integral transform on its own right and could be added to the table. For example, the inverse Fourier transform has the kernel $k(x,u) = e^{i2\pi ux}$, and the other variants of the Fourier transform, too, have their respective inverse transforms with their own kernels [Bracewell86 p. 7].

It is interesting to note that although the properties of the different integral transforms vary widely, they still have some properties in common. For example, every integral transform \mathcal{F}_k is a linear operator, meaning that it satisfies:

$$\mathcal{F}_k[c_1 f_1(x) + c_2 f_2(x)] = c_1 \mathcal{F}_k[f_1(x)] + c_2 \mathcal{F}_k[f_2(x)]$$

for any two functions $f_1(x), f_2(x)$ and constants c_1, c_2 ; this property is a straightforward consequence of the fact that the integral is a linear operator [Debnath95 p. 4]. In fact, if the kernel is allowed to be a generalized function then the converse is also true, meaning that all linear operators are integral transforms. A properly formulated version of this statement is the Schwartz kernel theorem [Wikipedia05; Ehrenpreis56].⁴ Furthermore, it turns out that each integral transform \mathcal{F} has an inverse integral transform \mathcal{F}^{-1} such that $\mathcal{F}^{-1}[\mathcal{F}(u)] = f(x)$; accordingly, $\mathcal{F}^{-1}\mathcal{F} = \mathcal{F}\mathcal{F}^{-1} = I$, where I is the identity transform mentioned above [Debnath95 p. 4]. It can be also proved that integral transforms are *unique*, meaning that if $\mathcal{F}_k[f_1(x)] = \mathcal{F}_k[f_2(x)]$ then $f_1(x) = f_2(x)$ under suitable conditions; this is known as the *uniqueness theorem* [Debnath95 p. 4].

The usefulness of integral transforms lies in the simplification that they bring about, for example in dealing with differential equations. A proper choice of the transform often makes it possible to convert an intractable problem in the original domain into a much simpler problem in the transformed domain that can be easily solved. The solution obtained is, of course, the transform of the solution of the original problem, and if the solution is required in the original domain, one still needs to apply the inverse transform to complete the operation. Hence, the typical procedure in such cases is to transform the original problem, solve the transformed problem, and then use the inverse transform to obtain a solution to the original problem. In many situations the type of the integral transform to be used is determined by the nature of the problem at hand; for example, the Fourier transform is the natural choice whenever we wish to make use of spectral considerations in the frequency domain. But in other situations, the art of choosing the best integral transform (or the best kernel) is often the key to a successful solution of the given problem. Some insights into the question of how to construct the best kernel can be found, for example, in [Ferraro88] and [Rubinstein91], in the context of pattern recognition in deformed images. A method for constructing an integral transform that solves a given differential equation can be found in [Churchill72, Chapter 10 and pp. 24–25, 384].

Finally, it should be noted that integral transforms can be easily generalized to functions of several variables. In this case Eq. (G.10) becomes [Debnath95 p. 4]:

$$F_k(\mathbf{u}) = \int_S f(\mathbf{x}) k(\mathbf{x}, \mathbf{u}) d\mathbf{x}$$

where $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{u} = (u_1, \dots, u_n)$ and $S \subset \mathbb{R}^n$.

⁴ A nice explanation showing how the integral transform (G.10) can be seen as the continuous-space counterpart of the linear vector transformation $\mathbf{f}_k = \mathbf{K} \cdot \mathbf{f}$ can be found in [Churchill72 pp. 25–26]. It is based on the interpretation of a function $f(x)$ as the continuous generalization of a vector $\mathbf{f} = (f_1, \dots, f_n)$.

Returning now to the question that opened this section, concerning the meaning of the generalized decomposition of $r(x)$ that is provided by Eq. (G.8), we can clearly see now the answer: This expression is simply an integral transform of $F(u)$, whose kernel is given by:

$$k(x, u) = e^{i2\pi u g(x)}$$

In the particular case where $g(x) = x$ (the identity transformation), this gives back the kernel of the inverse Fourier transform, and Eq. (G.8) becomes the spectral decomposition of the function $f(x)$, as in Eq. (G.2). However, when $g(x)$ is not the identity transformation, Eq. (G.8) is no longer the spectral decomposition of $r(x) = f(g(x))$, but simply another integral transform, that we may call the *g-Fourier transform*. But although this mathematical construct no longer has a spectral interpretation (formally, instead of the spectral decomposition of $r(x)$ we obtain here its *g*-spectral decomposition), it still may be used as a mathematical tool for solving our particular problems, just as the Laplace and the other integral transforms are used for solving various problems without necessarily providing a spectral decomposition of the functions in question.⁵

G.5 The use of generalized Fourier transforms in the moiré theory

Let us see now how the generalized *g*-Fourier formulation can help us to understand the superposition of two transformed, aperiodic layers and the resulting moiré effects. Note that the following discussion simply extends to the continuous case the generalized Fourier series approach we have already introduced in Chapter 10 of *Vol. I* for the case of transformed periodic layers; the main difference is that in our present case summation is replaced by integration.

Let $f_1(x', y')$ and $f_2(x', y')$ be two 2D aperiodic layers whose Fourier transforms are $F_1(u, v)$ and $F_2(u, v)$, respectively. We therefore have, using the more compact vector notation $\mathbf{x}' = (x', y')$ and $\mathbf{u} = (u, v)$, the following Fourier spectral decompositions, just like in Eq. (G.7):

$$\begin{aligned} f_1(\mathbf{x}') &= \int F_1(\mathbf{u}) e^{i2\pi \mathbf{u} \cdot \mathbf{x}'} d\mathbf{u} \\ f_2(\mathbf{x}') &= \int F_2(\mathbf{u}) e^{i2\pi \mathbf{u} \cdot \mathbf{x}'} d\mathbf{u} \end{aligned} \tag{G.11}$$

Suppose, now, that we apply to the layers $f_1(\mathbf{x}')$ and $f_2(\mathbf{x}')$ the geometric transformations $\mathbf{x}' = \mathbf{g}_1(\mathbf{x})$ and $\mathbf{x}' = \mathbf{g}_2(\mathbf{x})$, respectively. By substituting these transformations in Eqs. (G.11) we obtain the formal expressions:

⁵ As clearly expressed by Eq. (G.8), *g*-Fourier transforms decompose any given function $f(x)$ into a continuous set of basis functions $b(u) = e^{i2\pi u g(x)}$, whose proper weights are assigned by $F(u)$, the Fourier transform of $f(x)$. If the mapping $g(x)$ is rather weak, meaning that it differs just slightly from the identity mapping $g(x) = x$, then the resulting *g*-Fourier transform is still close to the Fourier transform, and its basis functions $b(u)$ are closely sinusoidal, and are therefore strongly localized in the frequency spectrum. In the case of the Fourier transform itself (when $g(x) = x$), each basis function is perfectly sinusoidal and corresponds to a single frequency component in the spectrum.

$$\begin{aligned}
r_1(\mathbf{x}) &= f_1(\mathbf{g}_1(\mathbf{x})) = \int F_1(\mathbf{u}) e^{i2\pi\mathbf{u}\cdot\mathbf{g}_1(\mathbf{x})} d\mathbf{u} \\
r_2(\mathbf{x}) &= f_2(\mathbf{g}_2(\mathbf{x})) = \int F_2(\mathbf{u}) e^{i2\pi\mathbf{u}\cdot\mathbf{g}_2(\mathbf{x})} d\mathbf{u}
\end{aligned} \tag{G.12}$$

which are simply the decompositions of $r_1(\mathbf{x})$ and $r_2(\mathbf{x})$ into their respective \mathbf{g}_1 - and \mathbf{g}_2 -Fourier transform components. The superposition of the transformed layers is therefore expressed by:

$$\begin{aligned}
r_1(\mathbf{x}) r_2(\mathbf{x}) &= \left(\int F_1(\mathbf{u}) e^{i2\pi\mathbf{u}\cdot\mathbf{g}_1(\mathbf{x})} d\mathbf{u} \right) \left(\int F_2(\mathbf{w}) e^{i2\pi\mathbf{w}\cdot\mathbf{g}_2(\mathbf{x})} d\mathbf{w} \right) \\
&= \iint F_1(\mathbf{u}) F_2(\mathbf{w}) e^{i2\pi[\mathbf{u}\cdot\mathbf{g}_1(\mathbf{x}) + \mathbf{w}\cdot\mathbf{g}_2(\mathbf{x})]} d\mathbf{u} d\mathbf{w}
\end{aligned} \tag{G.13}$$

Note that in the particular case where $\mathbf{g}_1(\mathbf{x}) = \mathbf{g}_2(\mathbf{x}) = \mathbf{x}$ Eq. (G.13) is simply the inverse Fourier transform of the convolution $F_1(\mathbf{u}) ** F_2(\mathbf{u})$, as indeed predicted by the convolution theorem, since in this case:

$$r_1(\mathbf{x}) r_2(\mathbf{x}) = \iint F_1(\mathbf{u}) F_2(\mathbf{w}) e^{i2\pi(\mathbf{u} + \mathbf{w})\cdot\mathbf{x}} d\mathbf{u} d\mathbf{w}$$

and by substituting $\mathbf{z} = \mathbf{u} + \mathbf{w}$:

$$\begin{aligned}
&= \int \left[\int F_1(\mathbf{u}) F_2(\mathbf{z} - \mathbf{u}) d\mathbf{u} \right] e^{i2\pi\mathbf{z}\cdot\mathbf{x}} d\mathbf{z} \\
&= \int [F_1(\mathbf{z}) ** F_2(\mathbf{z})] e^{i2\pi\mathbf{z}\cdot\mathbf{x}} d\mathbf{z} \\
&= \mathcal{F}^{-1}[F_1(\mathbf{z}) ** F_2(\mathbf{z})]
\end{aligned}$$

Returning to Eq. (G.13), let us consider now, just as we did in the discrete case of Fourier series in Sec. 10.7 of *Vol. I*, the partial sum (or rather the partial integral) that consists of all the terms in which $\mathbf{w} = -\mathbf{u}$:

$$m_{1,-1}(\mathbf{x}) = \int F_1(\mathbf{u}) F_2(-\mathbf{u}) e^{i2\pi\mathbf{u}\cdot[\mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})]} d\mathbf{u} \tag{G.14}$$

This partial integral corresponds to a sub-structure that is present in the superposition $r_1(\mathbf{x}) r_2(\mathbf{x})$ of Eq. (G.13), but is not present in either of the original layers $r_1(\mathbf{x})$ and $r_2(\mathbf{x})$ themselves. And indeed, just as in the discrete case of Sec. 10.7 in *Vol. I*, this structure is simply the substructure (1,-1)-moiré that is generated in the superposition. Note that in Sec. 10.7 of *Vol. I* we defined, more generally, the substructures of the superposition which correspond to its (k_1, k_2) -moirés; but as we already know (see Sec. 7.5), in the case of aperiodic layers no moirés other than the (1,-1)-moiré can exist, since in such cases there is no correlation between the superposed layers (for example, the correlation between a random screen $r(x, y)$ and its scaled version $r(2x, 2y)$ is practically zero). Therefore, although we *can* technically define $m_{k_1, k_2}(\mathbf{x})$ for any (k_1, k_2) (and in fact, even for non-integer values of k_1 and k_2), the only visible sub-structure in the superposition of aperiodic layers corresponds to the (1,-1)-moiré. (As we have seen in Sec. 10.7.1 of *Vol. I*, in the discrete case, too, we can technically define a (k_1, k_2) -moiré $m_{k_1, k_2}(\mathbf{x})$ for any integers (k_1, k_2) ,

but in practice only a few of them are really visible in the given superposition, depending on the case).

Now, if we define the normalized profile of $m_{1,-1}(\mathbf{x})$ by:

$$p_{1,-1}(\mathbf{x}') = \int F_1(\mathbf{u})F_2(-\mathbf{u}) e^{i2\pi\mathbf{u}\cdot\mathbf{x}'} d\mathbf{u}$$

we see that $p_{1,-1}(\mathbf{x}')$ is simply the inverse Fourier transform of the product $F_1(\mathbf{u})F_2(-\mathbf{u})$, and hence, by the convolution theorem we have:

$$p_{1,-1}(\mathbf{x}') = p_1(\mathbf{x}') ** p_2(-\mathbf{x}') = p_1(\mathbf{x}') \star\star p_2(\mathbf{x}')$$

where $p_1(\mathbf{x}')$ and $p_2(\mathbf{x}')$ are the normalized profiles of the original, untransformed layers $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, and $\mathbf{x}' = \mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})$ is the coordinate transformation which brings $p_{1,-1}(\mathbf{x}')$ into the geometric layout of the (1,-1)-moiré. This leads us, indeed, to the fundamental Glass pattern theorem (see Sec. 7.8). Note that this way we obtain, indeed, the generalized version of Proposition 7.7 for any linear or non-linear transformations $\mathbf{g}_i(\mathbf{x})$, whereas in Sec. 7.4.1 the proposition was only justified for the case of linear transformations (see the footnote at the end of Sec. 7.4.1).

Note that in the particular case where $\mathbf{g}_1(\mathbf{x}) = \mathbf{g}_2(\mathbf{x})$ the two superposed layers are identical (up to their intensity profiles) and the moiré (or Glass) pattern becomes singular and hence invisible. On the other hand, when \mathbf{g}_1 and \mathbf{g}_2 are completely different, the moiré effect is too weak to be visible due to the lack of correlation between the superposed layers. Thus, the most interesting cases occur when \mathbf{g}_1 and \mathbf{g}_2 are just slightly different, so that the moiré effect is not yet singular, but not too weak, either.

Eq. (G.14) can be considered, in fact, as a $(\mathbf{g}_1 - \mathbf{g}_2)$ -Fourier transform. This formal construct allows us, indeed, to extract mathematically the moiré (or Glass) pattern from the global structure of the layer superposition — without really knowing their Fourier spectra in the frequency domain.

As we can see, the g -Fourier formalism is simply a working tool that we use here to simplify our problem by considering it in the transformed domain. It is clear that in the particular case where $g(x) = x$ the g -Fourier transform simply reduces into the Fourier transform; but in fact, a similar situation occurs for any linear or affine mapping $g(x) = ax + b$, since in such cases Eq. (G.8) becomes:

$$r(x) = f(ax + b) = \int_{-\infty}^{\infty} F(u) e^{i2\pi u(ax + b)} du$$

which gives after some short manipulations (see, for example, [Gaskill78 pp. 194–195]):

$$= \int_{-\infty}^{\infty} R(u) e^{i2\pi ux} du$$

where $R(u)$ is the Fourier transform of $r(x)$. It follows, therefore, that whenever $g(x)$ is linear (or affine) the g -Fourier transform can be seen as a Fourier transform, and therefore in such cases we can also interpret our transformed-domain reasoning in terms of the

classical Fourier spectral domain. But even when the mapping g is non-linear and the Fourier spectral interpretation is no longer possible, the g -Fourier formalism still remains extremely useful, as we have seen above.

Appendix H

Miscellaneous issues and derivations

H.1 Classification of the dot trajectories

Suppose we are given two identical aperiodic dot screens that are superposed on top of each other dot on dot, and that we apply to the two layers the direct transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$, respectively. As we have seen in Chapter 4, if the layer transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ are not too violent, the microstructure of the resulting superposition may give rise to visible dot trajectories whose shapes are determined, to a close approximation, by the trajectories (field lines) of the vector field:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y)$$

As we already know from Sec. B.6 of Appendix B, the trajectories of a vector field $\bar{\mathbf{h}}(x,y)$ are given by the solution curves of the system of differential equations:

$$\begin{aligned}\frac{d}{dt}x(t) &= \bar{h}_1(x(t),y(t)) \\ \frac{d}{dt}y(t) &= \bar{h}_2(x(t),y(t))\end{aligned}\tag{H.1}$$

where $\bar{h}_1(x,y)$ and $\bar{h}_2(x,y)$ are the two cartesian components of $\bar{\mathbf{h}}(x,y)$, namely, $\bar{\mathbf{h}}(x,y) = (\bar{h}_1(x,y), \bar{h}_2(x,y))$. Although the system of differential equations (H.1) is relatively easy to solve when $\bar{\mathbf{h}}(x,y)$ is linear (see, for example, Chapter 4 in [Kreyszig93]), its solution in non-linear cases may present a more difficult challenge. However, it turns out that it is often possible to get a qualitative idea about the shape of these solution curves (trajectories) without even having to solve the system of differential equations (H.1). This follows as a straightforward outcome of the characterization and classification of the critical points of the system of differential equations, since the behaviour of the trajectories surrounding a critical point highly depends on the properties of the critical point itself.

A point (x,y) is called a *critical point* of the vector field $\bar{\mathbf{h}}(x,y)$ or of the system of differential equations (H.1) if it satisfies $\bar{h}_1(x,y) = 0$ and $\bar{h}_2(x,y) = 0$ [Birkhoff89 p. 133; Kreyszig93 p. 176].¹ At such a point we have $\frac{d}{dt}x(t) = 0$ and $\frac{d}{dt}y(t) = 0$, and hence the direction of the solution curves of Eq. (H.1) there is indeterminate: $\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{0}{0}$. (Equivalently, we may say that at a critical point the direction of the trajectories of the vector field $\bar{\mathbf{h}}(x,y)$ is indeterminate; another way to see this is that the vector field $\bar{\mathbf{h}}(x,y)$ assigns to this point the null vector $(0,0)$, whose direction is obviously undefined.) In the

¹ Confusingly, some references use the term “fixed point” for a critical point (see, for example, [Strogatz94 pp. 124, 150; Weisstein99 p. 652]). Note that a point (x,y) for which we have $\bar{\mathbf{h}}(x,y) = (0,0)$ is a *zero* of $\bar{\mathbf{h}}(x,y)$, and not a fixed point of $\bar{\mathbf{h}}(x,y)$; it *is*, however, a mutual fixed point of $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$, since it satisfies, of course, $\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (0,0)$.

following subsections we will see how the classification of the critical points can give us clues to the qualitative behaviour of the trajectories of the system (H.1), and hence to the shape of the dot trajectories in our layer superpositions. We start, as usual, with the simplest case, in which $\bar{\mathbf{h}}(x,y)$ is linear.

H.1.1 Classification of the dot trajectories in the linear case

Suppose first that the vector field $\bar{\mathbf{h}}(x,y)$ is linear. This obviously occurs when both of the layer transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ are linear, but it may also happen when $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ are non-linear, and their non-linear components are mutually cancelled out in the difference $\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y)$. In both of these cases the linearity of $\bar{\mathbf{h}}(x,y)$ implies, of course, that the set of differential equations (H.1) is linear, too:

$$\begin{aligned}\frac{d}{dt}x(t) &= a_1x(t) + b_1y(t) \\ \frac{d}{dt}y(t) &= a_2x(t) + b_2y(t)\end{aligned}\tag{H.2}$$

Because of its linearity, $\bar{\mathbf{h}}(x,y)$ is clearly zero at the origin, and thus the origin is a critical point of the system of differential equations (H.2). (Note that if the linear transformation $\bar{\mathbf{h}}(x,y)$ is singular it may have a full critical line passing through the origin, i.e. a line all of whose points satisfy $\bar{\mathbf{h}}(x,y) = (0,0)$. However, the origin itself is always a zero of any linear $\bar{\mathbf{h}}(x,y)$, be it singular or regular, and hence Eq. (H.2) always has a critical point at the origin.)

Now, in the theory of differential equations there exists a simple technique that allows us to characterize and classify the critical point of the linear system (H.2), and hence to find qualitatively the behaviour of the trajectories (solution curves) that surround it, without having to solve the system. This technique is based on the two eigenvalues of the matrix of the linear transformation $\bar{\mathbf{h}}(x,y)$ (see, for example, Sec. 16.1 in [Gray97]). For instance, if the matrix of $\bar{\mathbf{h}}(x,y)$ has two real eigenvalues with opposite signs then the critical point is a saddle point and the trajectories surrounding it are hyperbolic, and if the eigenvalues are purely imaginary then the critical point is a *center* with circular or elliptical trajectories surrounding it. The full classification of the critical points is given in Table H.1 below, which is based on [Gray97 p. 566].² As we can see from this table, eigenvalues with a positive real part always cause repulsion from the origin, whereas eigenvalues with a negative real part cause attraction to the origin;³ the imaginary part of the eigenvalues indicates rotation of the trajectories about the origin, and a zero eigenvalue corresponds to degenerate cases with an entire critical line.

² Note that the nomenclature used to designate the different types of critical points significantly varies from reference to reference. For example, center points, spiral points and degenerate nodes according to the terminology used in [Gray97] are called in [Birkhoff89] vortex points, focal points and star points, respectively; similarly, improper nodes in [Gray97] correspond to degenerate nodes in [Strogatz94 p. 136], while degenerate nodes in [Gray97] refer to stars in [Strogatz94 p. 135]. In Table H.1 we have chosen the names that best suit our needs and our general conventions.

³ Note that this is immaterial for our needs, since the dot trajectories in the layer superposition do not show the sense of the arrows along the curves.

Remark H.1: This classification can be further simplified thanks to the fact that the eigenvalues of a linear transformation $\bar{\mathbf{h}}(x,y)$ depend only on the trace t and the determinant d of the matrix of that linear transformation [Strogatz94 p. 130; Kreyszig93 p. 176]. This eliminates the need for calculating the eigenvalues of the matrix, and allows an alternative, elegant classification of the different types of critical points and the trajectories surrounding them in terms of the two real numbers t and d . This classification can be represented graphically in the t,d plane, as shown, for example, in figure 5.2.8 in [Strogatz94 p. 137]. Note, however, that this graphical classification still may require the calculation of $t^2 - 4d$ in order to determine in which region of the graph our particular case (t,d) is situated; furthermore, this graphical method may be ambiguous in some particular cases such as stars and improper nodes, which are both located in the t,d plane along the border of the parabola $t^2 - 4d = 0$. For these reasons we prefer to stick here to the original classification of Table H.1, that is based on the eigenvalues of the matrix. ■

Example H.1: Consider the superposition shown in Fig. 2.1(e). In this case the layer transformations are given by:

$$\bar{\mathbf{g}}_1(x,y) = ((1+\varepsilon)x, (1+\varepsilon)y)$$

$$\bar{\mathbf{g}}_2(x,y) = (x,y)$$

where ε is a small positive fraction, and therefore we have:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (\varepsilon x, \varepsilon y)$$

In this case the matrix of the linear transformation $\bar{\mathbf{h}}(x,y)$ is $\begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$, and its two eigenvalues are simply $\lambda_1 = \lambda_2 = \varepsilon$ where $\varepsilon > 0$, meaning that the critical point at the origin is a star (see case 6 in Table H.1). And indeed, this fully agrees with the dot trajectories that surround the origin in the layer superposition that is shown in Fig. 2.1(e). ■

Example H.2: Consider now the superposition shown in Fig. 2.2(a). In this case one layer is obtained from the other by a slight scaling of $s_x = 1 - \varepsilon$ in the x direction, and a slight scaling of $s_y = 1 + \varepsilon$ in the y direction (ε being a small positive fraction). Therefore the layer transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ are given by:

$$\bar{\mathbf{g}}_1(x,y) = ((1-\varepsilon)x, (1+\varepsilon)y)$$

$$\bar{\mathbf{g}}_2(x,y) = (x,y)$$

so that:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (-\varepsilon x, \varepsilon y)$$

In this case the matrix of the linear transformation $\bar{\mathbf{h}}(x,y)$ is $\begin{pmatrix} -\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$, and its two eigenvalues are $\lambda_1 = -\varepsilon$ and $\lambda_2 = \varepsilon$. As we can see in case 3 of Table H.1 the critical point in this case is a saddle point, that is surrounded by hyperbolic trajectories. And indeed, this fully agrees with the dot trajectories that surround the origin in the layer superposition that is shown in Fig. 2.2(a). ■

	Eigenvalues	Type of critical points	Fig.	Remarks
1	$0 < \lambda_1 < \lambda_2$	Repelling node		
2	$\lambda_1 < \lambda_2 < 0$	Attracting node		
3	$\lambda_1 < 0 < \lambda_2$	Saddle	2.2(a)	(1)
4	$0 = \lambda_1 < \lambda_2$	Repelling line	2.3(a)	(2)
5	$\lambda_1 < \lambda_2 = 0$	Attracting line		(3)
6	$\lambda_1 = \lambda_2 > 0$ Two eigenvectors	Repelling star	2.1(e)	
7	$\lambda_1 = \lambda_2 < 0$ Two eigenvectors	Attracting star		
8	$\lambda_1 = \lambda_2 > 0$ One eigenvector	Repelling improper node		
9	$\lambda_1 = \lambda_2 < 0$ One eigenvector	Attracting improper node		
10	$\lambda_1 = \lambda_2 = 0$ One eigenvector	Linear center	2.3(c)	(4)
11	$\lambda_1, \lambda_2 = \alpha \pm i\beta$ $\alpha > 0, \beta \neq 0$	Repelling spiral	2.1(g)	
12	$\lambda_1, \lambda_2 = \alpha \pm i\beta$ $\alpha < 0, \beta \neq 0$	Attracting spiral		
13	$\lambda_1, \lambda_2 = \pm i\beta$ $\beta \neq 0$	Center	2.1(c), 2.2(c)	(5)

Table H.1: (continued on the opposite page)

Example H.3: Consider the superposition shown in Fig. 2.3(a). In this case the layer transformations are given by:

$$\bar{\mathbf{g}}_1(x, y) = (x, (1+\varepsilon)y)$$

$$\bar{\mathbf{g}}_2(x, y) = (x, y)$$

Remarks:

- (1) The critical point is surrounded by hyperbolic trajectories.
- (2) An entire critical line, with parallel straight trajectories emanating from it.
- (3) An entire critical line, with parallel straight trajectories pointing to it.
- (4) An entire critical line, with parallel straight trajectories parallel to it.
- (5) The critical point is surrounded by circular or elliptic trajectories.

Note that for our own needs (the characterization of the dot trajectories in a dot screen superposition) the sense of the trajectories is immaterial, and hence we do not need to distinguish between the repelling and attracting variants in each type of critical points. Nevertheless, we still maintain this distinction in the table for the sake of completeness.

Table H.1: (*continued.*) Summary of all the different types of critical points that may occur in a 2D linear system of differential equations, and the conditions on the eigenvalues of the matrix of the system that give rise to each of these types. Note that in cases 1–3, 6–9, 11–13 the critical point is isolated and located at the origin, while in all the other cases (which are called *degenerate* or *singular* cases) there exists a full line of critical points passing through the origin.

where ε is a small positive fraction, and therefore we have:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (0, \varepsilon y)$$

The matrix of the linear transformation $\bar{\mathbf{h}}(x,y)$ is $\begin{pmatrix} 0 & 0 \\ 0 & \varepsilon \end{pmatrix}$, and its two eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = \varepsilon > 0$, meaning that the critical point at the origin is in this case a repelling line (case 4 in Table H.1). And indeed, this fully agrees with the dot trajectories that surround the origin in the layer superposition that is shown in Fig. 2.3(a). ■

It should be noted, however, that the classification provided by Table H.1 is only valid for linear cases, and even in affine cases, where a mere constant shift has been added (see, for example, Figs. 2.3(e),(g)), the table can no longer be used to determine the shape of the trajectories.

H.1.2 Classification of the dot trajectories in the non-linear case

What happens now when $\bar{\mathbf{h}}(x,y)$ is not linear? Unlike a linear vector field that always has a critical point at the origin (or, in singular cases, an entire critical line passing through the origin), a non-linear vector field may have, depending on the case, no critical points at all, one or more isolated critical points, or even one or more straight or curved critical lines. But because the behaviour of the trajectories surrounding a critical point highly depends on the properties of the critical point itself, in cases where critical points do exist it should be possible to get a qualitative idea about the shape of the trajectories of $\bar{\mathbf{h}}(x,y)$ (i.e., the solution curves of Eq. (H.1)) simply by identifying the critical points and studying their properties, without having to solve the system (H.1). As said in [Tabor89 p. 20], critical

points can be thought of as the “organizing centers” of a system’s dynamics; thus, by identifying them and their properties one can build up a fairly global picture of the system’s behaviour.

In Sec. H.1.1 we have seen that if $\bar{\mathbf{h}}(x,y)$ is linear it has a critical point at the origin, and that we can characterize and classify this critical point by studying the eigenvalues of the matrix of the linear transformation $\bar{\mathbf{h}}(x,y)$. And indeed, it turns out that this technique can be also extended to the case of non-linear $\bar{\mathbf{h}}(x,y)$, i.e. to cases where the system (H.1) is non-linear.

This extended technique is based on the *linearization* of the non-linear system about each of its critical points separately (see, for example, [Kreyszig93 Sec. 4.5], [Strogatz94 Sec. 6.3], or [Gray97 pp. 588–590]). That is, instead of considering the original non-linear system (H.1) itself, we study its close approximation by the following linear system, which is simply its first-order Taylor approximation about the critical point (x_c, y_c) :

$$\begin{pmatrix} \frac{d}{dt}x(t) \\ \frac{d}{dt}y(t) \end{pmatrix} = \begin{pmatrix} \frac{\partial \bar{h}_1(x,y)}{\partial x} & \frac{\partial \bar{h}_1(x,y)}{\partial y} \\ \frac{\partial \bar{h}_2(x,y)}{\partial x} & \frac{\partial \bar{h}_2(x,y)}{\partial y} \end{pmatrix}_{\substack{x=x_c \\ y=y_c}} \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \quad (\text{H.3})$$

Note that the matrix of the system (H.3) consists of constant coefficients, meaning that (H.3) is, indeed, a linear system of differential equations. But since $\bar{\mathbf{h}}(x,y)$ is non-linear, this linear approximation obviously varies from point to point, and we have to recalculate the constant coefficients of its matrix for each critical point (x_c, y_c) separately. Thus, for each critical point of the non-linear system (H.1) we obtain a separate linear system of the form (H.2) that approximates the non-linear system (H.1) about that critical point. Note that each of these approximating linear systems is, in fact, shifted so as to bring the critical point (x_c, y_c) to the origin; this can be easily seen in the Taylor development that leads to Eq. (H.3), as explained in each of the references mentioned above.

Therefore, all that we have to do in the non-linear case is to solve the system of equations $\bar{h}_1(x,y) = 0$, $\bar{h}_2(x,y) = 0$ in order to identify the *isolated* critical points of the system, and then, for each of these points (x_c, y_c) , to calculate the numeric values of the partial derivatives of $\bar{\mathbf{h}}(x,y)$ at this point. This gives us for each of the critical points a linear system (H.2), and thus we can characterize and classify each of the critical points separately using the method of Sec. H.1.1 for linear cases. Note that the matrix we use in the linear approximation (H.3) is simply the Jacobian matrix of $\bar{\mathbf{h}}(x,y)$, evaluated at the critical point (x_c, y_c) .

It should be emphasized that for each critical point we have a different linearized system, and the properties of the linearized system in each critical point may be radically different. But when we “piece together” the different linearized systems, we may obtain a fairly accurate picture of the nonlinear system and its trajectories. Note, however, that this technique can only be used for isolated critical points of the non-linear system, but not for critical lines or for non-linear systems having no critical points at all.

Example H.4: As we have seen in Sec. 3.4, when two originally identical aperiodic dot screens undergo non-linear transformations, several Glass patterns may be simultaneously generated in their superposition. Let us consider, as an illustration, Fig. 3.15(a). This figure shows two originally identical aperiodic dot screens that give in their superposition, after each of them has undergone a different non-linear transformation, 4 distinct Glass patterns. The two domain transformations undergone in this case by the original layers are (see Eq. (3.35)):

$$\mathbf{g}_1(x,y) = (x, y + y_0 - ax^2)$$

$$\mathbf{g}_2(x,y) = (x + x_0 - ay^2, y)$$

where $a > 0$, $x_0 > 0$ and $y_0 > 0$. As explained in Examples 3.6 and 5.3, these transformations have 4 different mutual fixed points that are located at:

$$(x,y) = (\pm\sqrt{y_0/a}, \pm\sqrt{x_0/a})$$

(see Eq. (3.39)); and indeed, each of the 4 Glass patterns in the layer superposition is generated about one of these 4 fixed points.

However, as we can clearly see in Fig. 3.15(a), it turns out that the dot trajectories surrounding these Glass patterns are not identical: While the two Glass patterns that are located along the main diagonal are surrounded by *circular* dot trajectories, the two Glass patterns that are located along the other diagonal are surrounded by *hyperbolic* dot trajectories. How can we explain this fact?

According to Proposition 4.3, the dot trajectories that are generated in the superposition due to the application of the direct layer transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ are closely approximated by the vector field:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y)$$

In our present case (see Example 5.3) the direct layer transformations are given by:

$$\bar{\mathbf{g}}_1(x,y) = \mathbf{g}_1^{-1}(x,y) = (x, y - y_0 + ax^2)$$

$$\bar{\mathbf{g}}_2(x,y) = \mathbf{g}_2^{-1}(x,y) = (x - x_0 + ay^2, y)$$

where $a > 0$, $x_0 > 0$ and $y_0 > 0$, and therefore we have:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (x_0 - ay^2, ax^2 - y_0)$$

But because $\bar{\mathbf{h}}(x,y)$ is clearly non-linear, we need in order to find the properties of its trajectories to use the linearization technique. As we can see by solving the system of equations $\bar{h}_1(x,y) = 0$, $\bar{h}_2(x,y) = 0$, $\bar{\mathbf{h}}(x,y)$ has 4 critical points that are precisely located at:

$$(x,y) = (\pm\sqrt{y_0/a}, \pm\sqrt{x_0/a})$$

Let us therefore evaluate the matrix of the linearized system (H.3) for each of these 4 critical points separately; this will allow us to characterize each of these critical points using the method of Sec. H.1.1.

It is easy to see that the Jacobian matrix of Eq. (H.3) is in our case:

$$\begin{pmatrix} \frac{\partial \bar{h}_1(x,y)}{\partial x} & \frac{\partial \bar{h}_1(x,y)}{\partial y} \\ \frac{\partial \bar{h}_2(x,y)}{\partial x} & \frac{\partial \bar{h}_2(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 & -2ay \\ 2ax & 0 \end{pmatrix}$$

We now evaluate this matrix and its eigenvalues at each of the 4 critical points of $\bar{\mathbf{h}}(x,y)$:

(a) At the critical point $(+\sqrt{y_0/a}, +\sqrt{x_0/a})$ the matrix becomes:

$$\begin{pmatrix} 0 & -2a\sqrt{x_0/a} \\ 2a\sqrt{y_0/a} & 0 \end{pmatrix}$$

The two eigenvalues of this matrix are purely imaginary:

$$\lambda_1, \lambda_2 = \pm \sqrt{-4a\sqrt{x_0 y_0}}$$

(because $x_0, y_0, a > 0$), and therefore, according to case 13 in Table H.1, this critical point is a center. This explains, indeed, why the dot trajectories about this point in Fig. 3.15(a) are circular.

(b) At the second critical point on the main diagonal, $(-\sqrt{y_0/a}, -\sqrt{x_0/a})$ the matrix is:

$$\begin{pmatrix} 0 & 2a\sqrt{x_0/a} \\ -2a\sqrt{y_0/a} & 0 \end{pmatrix}$$

and the two eigenvalues are the same as in case (a). And indeed, the dot trajectories about this point in Fig. 3.15(a) are, again, circular.

(c) At the third critical point, $(+\sqrt{y_0/a}, -\sqrt{x_0/a})$, the matrix is:

$$\begin{pmatrix} 0 & 2a\sqrt{x_0/a} \\ 2a\sqrt{y_0/a} & 0 \end{pmatrix}$$

The two eigenvalues of this matrix are:

$$\lambda_1, \lambda_2 = \pm \sqrt{4a\sqrt{x_0 y_0}}$$

but because $x_0, y_0, a > 0$ these eigenvalues are purely real, with opposite signs. This corresponds to case 3 in Table H.1, meaning that this critical point is a saddle point that

is surrounded by hyperbolic trajectories. And indeed, this result fully agrees with the dot trajectories about this point in Fig. 3.15(a).

(d) Finally, at the fourth critical point, $(-\sqrt{y_0/a}, +\sqrt{x_0/a})$, the matrix is:

$$\begin{pmatrix} 0 & -2a\sqrt{x_0/a} \\ -2a\sqrt{y_0/a} & 0 \end{pmatrix}$$

and the two eigenvalues are the same as in case (c). And indeed, the dot trajectories about this point in Fig. 3.15(a) are, again, hyperbolic. ■

Example H.5: Consider the superposition shown in Fig. 2.3(e). This case is identical to that of Example H.3, except that a horizontal shift has been added to one of the layers:

$$\bar{\mathbf{g}}_1(x,y) = (x + x_0, (1+\varepsilon)y)$$

$$\bar{\mathbf{g}}_2(x,y) = (x,y)$$

and therefore we have:

$$\bar{\mathbf{h}}(x,y) = \bar{\mathbf{g}}_1(x,y) - \bar{\mathbf{g}}_2(x,y) = (x_0, \varepsilon y)$$

Note that because of the shift of x_0 this transformation is no longer linear but rather affine, and therefore the classification of Table H.1 no longer holds for it. And indeed, in this case the vector field $\bar{\mathbf{h}}(x,y)$ and its corresponding system of differential equations have no critical points at all, and we cannot derive the behaviour of the trajectories from properties of the critical points. ■

H.2 The connection between the vector fields $\bar{\mathbf{h}}_1(x,y)$ and $\bar{\mathbf{h}}_2(x,y)$ in Sec. 4.5

We have seen in Sec. 4.5 of Chapter 4 that in cases where both of the superposed layers are distorted by transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ the vector field which accurately represents the dot trajectories is given by Eq. (4.8):

$$\bar{\mathbf{h}}_1(x,y) = \bar{\mathbf{g}}_1(\bar{\mathbf{g}}_2(x,y)) - (x,y)$$

However, we have seen there that a similar reasoning may lead us to a different vector field, which also represents accurately the dot trajectories in the same superposition, and which is given by Eq. (4.9):

$$\bar{\mathbf{h}}_2(x,y) = (x,y) - \bar{\mathbf{g}}_2(\bar{\mathbf{g}}_1(x,y))$$

Clearly, vector fields (4.8) and (4.9) are not necessarily identical (see, for example, Figs. 4.11(b)–(d) and 4.12(b)–(d)). In fact, these vector fields happen to be identical only if our transformations satisfy the identity:

$$\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y)) - (x,y) \equiv (x,y) - \bar{\mathbf{g}}_2(\mathbf{g}_1(x,y))$$

which means:

$$\frac{1}{2} [\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y)) + \bar{\mathbf{g}}_2(\mathbf{g}_1(x,y))] \equiv (x,y) \quad (\text{H.4})$$

but this identity is certainly not satisfied by all transformation pairs $\bar{\mathbf{g}}_1, \bar{\mathbf{g}}_2$. So how is it possible that the two different vector fields (4.8) and (4.9) represent the same dot trajectories in the layer superposition?

The answer to this question is given, indeed, by Proposition 4.2: As we can see from this proposition, the dot trajectories obtained by applying the transformations $\bar{\mathbf{g}}_1(x,y)$ and $\bar{\mathbf{g}}_2(x,y)$ to our original screens are not uniquely obtained by this specific pair of layer transformations, and there exist in fact infinitely many transformation pairs that give the same dot trajectories. Among these transformation pairs there exist precisely two, the pair $\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y))$ and (x,y) and the pair (x,y) and $\bar{\mathbf{g}}_2(\mathbf{g}_1(x,y))$, in which only *one* of the two superposed layers is transformed, and for which we know, therefore, by virtue of Proposition 4.1, the precise vector field representations. These two vector fields have, indeed, different mathematical expressions — but as we have just seen, by virtue of Proposition 4.2, both of them represent the same dot trajectories in the superposition.

Incidentally, it is interesting to note that the two transformations $\bar{\mathbf{g}}_1(\mathbf{g}_2(x,y))$ and $\bar{\mathbf{g}}_2(\mathbf{g}_1(x,y))$ are, in fact, the inverse of each other. This can be seen by applying the general rule $[\mathbf{f}_1 \circ \mathbf{f}_2]^{-1} = \mathbf{f}_2^{-1} \circ \mathbf{f}_1^{-1}$ [Bernstein05 p. 6; Halmos74 p. 40], where “ \circ ” indicates the composition of transformations, to $\bar{\mathbf{g}}_2$ and \mathbf{g}_1 ; this gives us $[\bar{\mathbf{g}}_2 \circ \mathbf{g}_1]^{-1} = \bar{\mathbf{g}}_1 \circ \mathbf{g}_2$, which means, indeed, $[\bar{\mathbf{g}}_2(\mathbf{g}_1(x,y))]^{-1} = \bar{\mathbf{g}}_1(\mathbf{g}_2(x,y))$. And in fact, it can be shown that any two vector fields $\mathbf{h}_1(x,y) = \bar{\mathbf{g}}(x,y) - (x,y)$ and $\mathbf{h}_2(x,y) = (x,y) - \bar{\mathbf{g}}^{-1}(x,y)$ with an arbitrary $\bar{\mathbf{g}}(x,y)$ represent equivalent dot trajectories: Clearly, the vector field $\mathbf{h}_1(x,y)$ connects by an arrow the point (x,y) to its destination under $\bar{\mathbf{g}}$, the point $\bar{\mathbf{g}}(x,y)$, while the second vector field $\mathbf{h}_2(x,y)$ connects the point $\bar{\mathbf{g}}^{-1}(x,y)$ to its destination under $\bar{\mathbf{g}}$, the point (x,y) .⁴ This means that both vector fields correspond to the application of the same transformation $\bar{\mathbf{g}}$, although the departure points to which $\bar{\mathbf{g}}$ is applied are not the same in both cases. But because in both superpositions the same transformation $\bar{\mathbf{g}}$ is applied to one of two identical random screens, the resulting dot trajectories in the superposition are, indeed, equivalent. This can be seen from Proposition 4.2 by taking $\bar{\mathbf{g}}_1(x,y) = \bar{\mathbf{g}}(x,y)$, $\bar{\mathbf{g}}_2(x,y) = (x,y)$ and $\bar{\mathbf{f}}(x,y) = \bar{\mathbf{g}}^{-1}(x,y)$.

H.3 Hybrid (1,-1)-moiré effects whose moiré bands have 2D intensity profiles

As we have seen in Chapters 6 and 7, the transition from periodic line gratings to aperiodic (yet correlated) line gratings results in the loss of repetitivity in the resulting moiré effect: Instead of having infinitely many moiré bands, we are left in the aperiodic

⁴ Remember that for any points (a,b) and (c,d) in the plane, when the tail of the vector $(a,b) - (c,d)$ is attached to the point (c,d) , its head is located at the point (a,b) ; this means that the vector $(a,b) - (c,d)$ connects the point (c,d) to the point (a,b) .

case with a single moiré band, i.e. a Glass pattern (see, for example, Figs. 6.1, 6.2 and 7.13). This typical behaviour is general to both 1D (1,-1)-moirés between line gratings and 2D (1,0,-1,0)-moirés between dot screens, independently of their intensity profiles (see, for example, Figs. 2.1 and 7.1).⁵

It is not surprising, therefore, that the same behaviour subsists in the (1,-1)-moiré even when the intensity of each individual line in the original gratings is modulated by some given 1D or 2D information. Part (b) of Fig. H.1 shows the superposition of two such periodic line gratings: The first grating consists of lines whose individual profiles are modulated by some 2D information (in the present example: a flattened version of the letters “EPFL”, as clearly seen in the leftmost part of the figure), while the second grating consists of narrow white lines or “slits” on a black background. In such cases, if both gratings have similar periods and angles, the resulting superposition shows a (1,-1)-moiré effect that is a largely magnified version of the first grating; for example, in the case shown in our figure each of the moiré bands consists of a largely stretched-out (and possibly sheared) version of the letters “EPFL”. In other words, in such cases the 2D information that modulates the intensity profile of each of the moiré bands in the superposition is a largely stretched-out (and possibly sheared) version of the 2D information that modulates the intensity of each individual line of the first grating.⁶

Part (a) of Fig. H.1 shows, on its part, the aperiodic counterpart of Fig. H.1(b). Here, the individual lines in each of the two gratings are still the same as in Fig. H.1(b), but their locations have been randomized (using the same random numbers for both of the superposed layers). And indeed, as expected, this aperiodic superposition contains only one moiré band, the Glass pattern, whose intensity profile remains modulated by the same 2D information (the letters “EPFL”) as each of the moiré bands in Fig. H.1(b).

Note that replacing in Figs. H.1(a),(b) the modulated lines of both of the superposed gratings by simple black lines on a white background brings us back to the classical moiré (or Glass) patterns of Figs. 6.2(a),(b). The difference between Figs. H.1 and 6.2 in terms of the moiré intensity profile is, in fact, the 1D equivalent of the difference between Figs. 7.1 and 2.1 (see Secs. 7.2.4 and 7.4.2).

A similar phenomenon occurs also in superpositions of curvilinear line gratings. For example, if we replace the simple curved lines in the grating shown in Fig. 6.7 by modulated curved lines having the same profiles as in Fig. H.1, the intensity profile of the resulting curvilinear moiré (or Glass) pattern will be simply modulated by the letters “EPFL”. The same result can be also obtained, of course, by applying to the straight gratings of Fig. H.1 the geometric layer transformations that are used in Fig. 6.7. This is, indeed, the 1D equivalent of the transition between Figs. 7.1 and 7.12.

⁵ Note, however, that this behaviour does not extend to moirés of higher orders, as shown in Sec. 7.5.

⁶ This phenomenon, originally discovered in the 1980s by Joe Huck in his artistic work [Huck03], has been investigated in depth in [Hersch04] and [Chosson06]. A Fourier-based explanation of this phenomenon is provided in Sec. C.14 of Appendix C in the second edition of *Vol. I*.

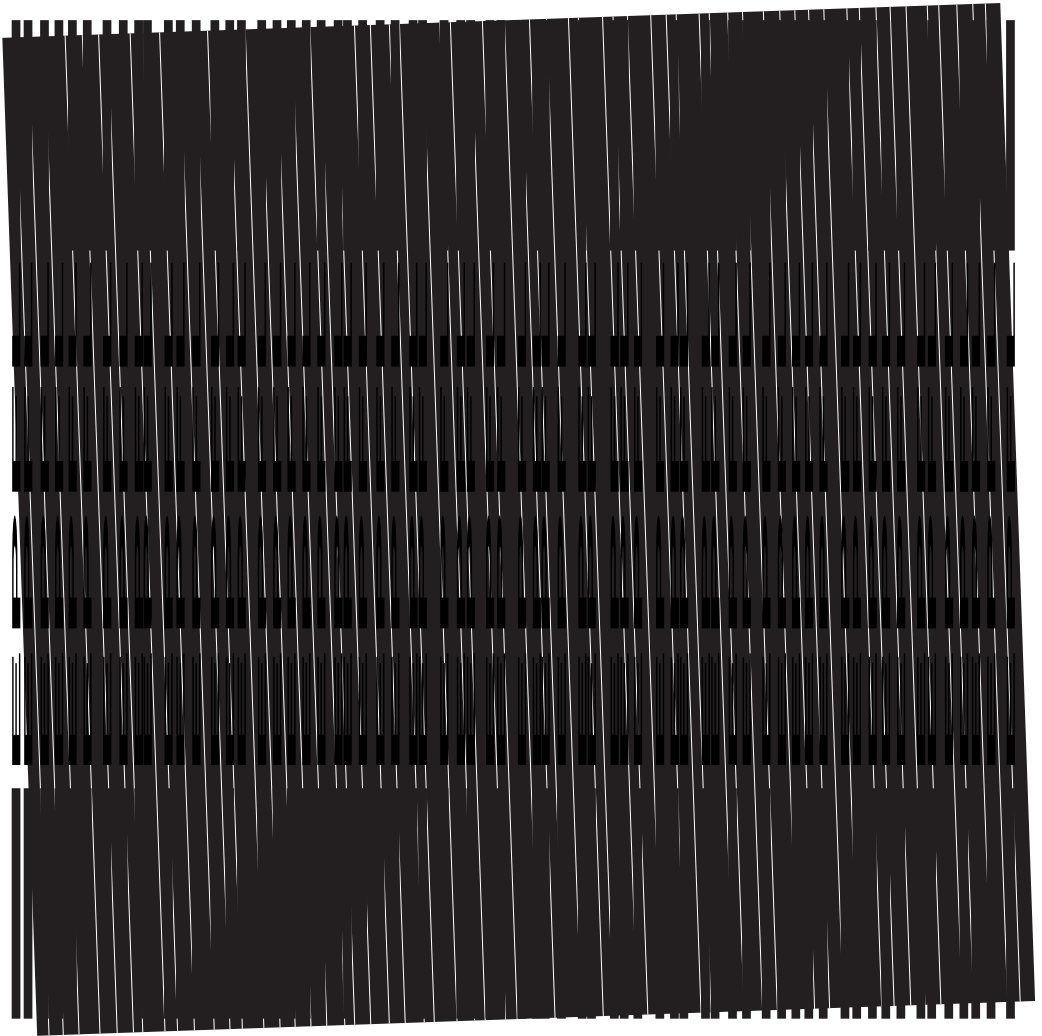


Figure H.1: (a) A superposition of two aperiodic (yet fully correlated) line gratings, one of which is composed of straight lines that contain some given 2D information (the flattened letters “EPFL”, as clearly seen in the leftmost line), while the other consists of narrow slits on a black background. As expected, a single moiré band (Glass pattern) appears in the center of the superposition, and its intensity profile contains a largely stretched-out (and possibly sheared) version of the 2D information that appears in each of the individual lines of the first grating (the letters “EPFL”). Compare with Fig. 6.2(a), in which both of the superposed gratings consist of simple black lines. The difference between Figs. H.1(a) and 6.2(a) is similar to the difference between Figs. 7.1(a) and 2.1(c) in the 2D case.

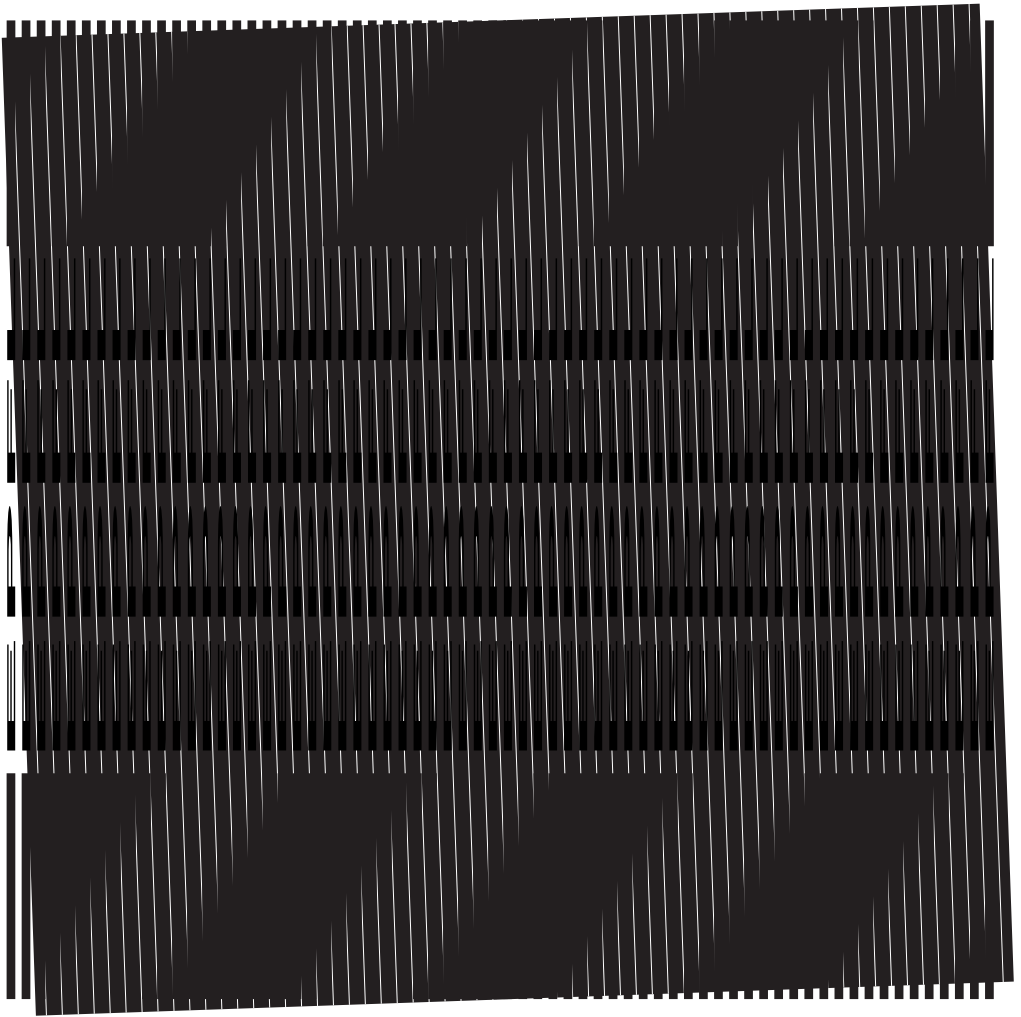


Figure H.1: (*continued.*) (b) The periodic counterpart of (a): The superposition of two *periodic* line gratings, one of which is composed of straight lines that contain some given 2D information (the flattened letters “EPFL”, as clearly seen in the leftmost line), while the other consists of narrow slits on a black background. Each of the bands of the resulting (1,-1)-moiré is a duplicate of the central band, and contains the same 2D information. Compare with Fig. 6.2(b), in which both of the superposed gratings consist of simple black lines. The difference between Figs. H.1(b) and 6.2(b) is similar to the difference between Figs. 7.1(b) and 2.1(d) in the 2D case. Note that due to some particularities of the human visual system the effects shown in Figs. H.1(a),(b) are more easily perceived when the figures are rotated by 90° (try and see!).

Finally, it is interesting to see how the fundamental Glass-pattern theorem (Sec. 7.8 in Chapter 7) applies to our present hybrid case, in which the individual lines in one of the superposed gratings have a 2D profile, while the lines in the other grating (the slits) only have a 1D profile. To see this, consider the full componentwise notation of the layer transformations $\mathbf{g}_1(\mathbf{x})$ and $\mathbf{g}_2(\mathbf{x})$ and of the moiré (or Glass pattern) transformation $\mathbf{g}(\mathbf{x})$:

$$\mathbf{g}_1(\mathbf{x}) = \begin{pmatrix} g_{1,1}(x,y) \\ g_{1,2}(x,y) \end{pmatrix}, \quad \mathbf{g}_2(\mathbf{x}) = \begin{pmatrix} g_{2,1}(x,y) \\ g_{2,2}(x,y) \end{pmatrix}, \quad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix}$$

According to Eq. (7.37) of the fundamental Glass-pattern theorem, the transformation $\mathbf{g}(\mathbf{x})$ undergone by the moiré (or Glass pattern) is given by $\mathbf{g}(\mathbf{x}) = \mathbf{g}_1(\mathbf{x}) - \mathbf{g}_2(\mathbf{x})$. In the 2D case, where both of the original, untransformed layers are dot screens, this simply means:

$$\begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} = \begin{pmatrix} g_{1,1}(x,y) \\ g_{1,2}(x,y) \end{pmatrix} - \begin{pmatrix} g_{2,1}(x,y) \\ g_{2,2}(x,y) \end{pmatrix} \quad (\text{H.5})$$

However, in the 1D case, i.e. when both of the original, undeformed layers consist of lines with a purely 1D profile, the second component in each of the above transformations becomes irrelevant, and we obtain:

$$\mathbf{g}_1(\mathbf{x}) = \begin{pmatrix} g_{1,1}(x,y) \\ 0 \end{pmatrix}, \quad \mathbf{g}_2(\mathbf{x}) = \begin{pmatrix} g_{2,1}(x,y) \\ 0 \end{pmatrix}, \quad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(x,y) \\ 0 \end{pmatrix}$$

or even, more simply, by dropping the unused components and indices:

$$\mathbf{g}_1(\mathbf{x}) = g_1(x,y), \quad \mathbf{g}_2(\mathbf{x}) = g_2(x,y), \quad \mathbf{g}(\mathbf{x}) = g(x,y)$$

In this case Eq. (7.37) reduces into its single-component counterpart:

$$g(x,y) = g_1(x,y) - g_2(x,y) \quad (\text{H.6})$$

which is precisely Eq. (7.35) of the fundamental Glass-pattern theorem for *line gratings*, as we have seen in Sec. 7.8.

We now return to our present hybrid case. In this case, only one of the two original, untransformed gratings (the second one) has a purely 1D profile, and therefore we have:

$$\mathbf{g}_1(\mathbf{x}) = \begin{pmatrix} g_{1,1}(x,y) \\ g_{1,2}(x,y) \end{pmatrix}, \quad \mathbf{g}_2(\mathbf{x}) = \begin{pmatrix} g_{2,1}(x,y) \\ 0 \end{pmatrix}, \quad \mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix}$$

Hence, Eq. (7.37) of the fundamental Glass-pattern theorem simply becomes here:

$$\begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} = \begin{pmatrix} g_{1,1}(x,y) \\ g_{1,2}(x,y) \end{pmatrix} - \begin{pmatrix} g_{2,1}(x,y) \\ 0 \end{pmatrix} \quad (\text{H.7})$$

which is, indeed, intermediate between the 2D case of Eq. (H.5) and the 1D case of Eq. (H.6). This suggests that our hybrid superposition could be considered, in fact, as a “ $1\frac{1}{2}$ D case”. A more detailed discussion on this case can be found in the second edition of *Vol. I* in Sec. C.14 of Appendix C.

Appendix I

Glossary of the main terms

I.1 About the glossary

Several thousands of publications on the moiré phenomenon have appeared during the last decades, in many different fields and applications. However, as already mentioned in *Vol. I*, the terminology used in this vast literature is very far from being consistent and uniform. Different authors use different terms for the same entities, and what is even worse, the same terms are often used in different meanings by different authors. To mention just one example, Glass patterns have been also called in literature *moiré fringes* [Glass69], *random-dot interference patterns* [Glass73], *quasi-moiré patterns* [Garavaglia01], and even *flash correlation artifacts*, or in short, *FCA*s [Prokoski99].

Obviously, in such an interdisciplinary domain as the moiré theory it would be quite impossible to adopt a universally acceptable standardization of the terms, because of the different needs and traditions in the various fields involved (optics, mechanics, mathematics, printing, etc.). Nevertheless, even without having any far-reaching pretensions, we were obliged to make our own terminological choices in a systematic and coherent way, in order to prevent confusion and ambiguity in our own work. We tried to be consistent in our terminology throughout this work, even if it forced us to assign to some terms a somewhat different meaning than one would expect (depending on his own background, of course).

We included in the present glossary all the terms for which we felt a clear definition was desirable to avoid any risk of ambiguity. But although this glossary is basically devoted to the terms that are being used in the present volume, we have also included, for the sake of completeness, some entries from the glossary of *Vol. I*, often with some additions or adaptations to our needs in the present volume. Just as in *Vol. I*, this glossary is not ordered alphabetically; rather, we preferred to group the various terms according to subjects. We hope this should help the reader not only to clearly see the meaning of each individual term by itself, but also to put it in relation with other closely related terms (which would be completely dispersed throughout the glossary if an alphabetical order were preferred). Note that terms in the glossary can be found alphabetically through the general index at the end of the book.

I.2 Terms in the image domain

grating (or *line-grating*) —

A pattern consisting of parallel lines. A grating can be periodic or aperiodic.

curvilinear grating —

A pattern consisting of parallel curvilinear lines. A curvilinear grating can be seen as a non-linear transformation of an initially uncurved periodic or aperiodic grating of straight lines.

cosinusoidal grating (not to be confused with *cosine-shaped grating*) —

A grating with a cosinusoidal periodic-profile; for example, a cosinusoidal circular grating is a circular grating with a cosinusoidal periodic-profile. Note, however, that since reflectance and transmittance functions always take values ranging between 0 and 1, the cosinusoidal grating is normally “raised” and rescaled into this range of values. For example, a reflectance function in the form of a vertical straight cosinusoidal grating is expressed by: $r(x,y) = \frac{1}{2} \cos(2\pi fx) + \frac{1}{2}$.

cosine-shaped grating (not to be confused with *cosinusoidal grating*) —

A grating (with any periodic-profile form) whose corrugations in the x,y plane are bent into a cosinusoidal shape, like in Fig. 6.17.

grid (or *line-grid*; also called in literature *cross-line grating*) —

A pattern consisting of two superposed line-gratings, periodic or aperiodic, crossing each other at a non-zero angle. Unless otherwise mentioned it will be assumed that a grid consists of two binary straight line gratings. Note that every grid can be also seen as a screen (whose dot-elements are the spaces left between the lines of the grid).

regular grid (or *square grid*) —

A 2-fold periodic grid composed of two superposed straight line-gratings that are identical but perpendicular to each other.

curved grid —

A pattern obtained by applying a non-linear transformation to a periodic or to an aperiodic grid.

screen (or *dot-screen*) —

A pattern consisting of dots. A screen can be periodic (as in Fig. 2.1(b)) or aperiodic (as in Fig. 2.1(a)).

regular screen —

A 2-fold periodic screen whose dot arrangement is orthogonal and whose periods (or frequencies) to both orthogonal directions are equal.

curved screen —

A pattern obtained by applying a non-linear transformation to a periodic or to an aperiodic screen (see, for example, Figs. 3.5(a),(b)).

halftone screen —

A binary screen in which the size (and the shape) of the screen dots may vary (typically, according to the gray level of a given original continuous-tone image). A

halftone screen may be periodic or not. Halftone screens are used in the printing world for the reproduction of continuous-tone images on bilevel printing devices.

screen gradation (or *wedge*) —

A halftone screen whose screen dots vary gradually (in their size and possibly also in their shape) across the image, generating a halftoned image with a smooth and uniform tone gradation.

image (has nothing to do with the image of a transformation) —

The most general term we use to cover “anything” in the image domain. It may be periodic or not, binary or continuous, etc. In principle, a monochrome (black-and-white) image has reflectance (or transmittance) values that vary between 0 (black) and 1 (white); similarly, a colour image has reflectance (or transmittance) values varying between 0 and 1 for each wavelength λ of its colour spectrum.

period (or *repetition-period* of a function p) —

A number $T \neq 0$ such that for any $x \in \mathbb{R}$, $p(x+T) = p(x)$. Note that the set of all the periods of $p(x)$ forms a lattice in \mathbb{R} . In the case of a 2-fold periodic function $p(x,y)$, a double period (or period parallelogram) of $p(x,y)$ is any parallelogram A which tiles the x,y plane so that $p(x,y)$ repeats itself identically on any of these tiles (see also Sec. A.3.4 in Appendix A of *Vol. I*).

period-vector (of a periodic function $p(x,y)$) —

A non-zero vector $\mathbf{P} = (x_0, y_0)$ such that for any $(x,y) \in \mathbb{R}^2$, $p(x+x_0, y+y_0) = p(x,y)$. If there exist two non-collinear vectors $\mathbf{P}_1, \mathbf{P}_2$ having this property, $p(x,y)$ is said to be 2-fold periodic; in this case, for any point $\mathbf{x} \in \mathbb{R}^2$ the points $\mathbf{x}, \mathbf{x} + \mathbf{P}_1, \mathbf{x} + \mathbf{P}_2, \mathbf{x} + \mathbf{P}_1 + \mathbf{P}_2$ define a period parallelogram of $p(x,y)$.

periodic function —

A function having a period. Note that a 2D function $p(x,y)$ can be 2-fold periodic (such as $p(x,y) = \cos x + \cos y$) or only 1-fold periodic (such as $p(x,y) = \cos x$).

almost-periodic function —

See Secs. B.3, B.5 in Appendix B of *Vol. I*.

aperiodic function —

A function which is not included in the class of almost-periodic functions. This also implies that the function in question is not periodic (see Fig. B.3 in Appendix B of *Vol. I*). Note, however, that although repetitive functions formally fall within the scope of this definition of aperiodic functions, we prefer to exclude them from our present definition, because they are still structurally ordered, and they have already been investigated in Chapters 10 and 11 of *Vol. I*. We therefore adopt the definition saying that a function is aperiodic if it is neither periodic (or almost-periodic) nor a geometrically transformed version thereof.

repetitive structure (or *repetitive function*) —

A structure (or a function) which is repetitive according to a certain rule, but which is not necessarily periodic (or almost-periodic). For example: concentric circles; gratings with logarithmic line-distances; screen gradations; etc. Note that such structures are sometimes called in literature *quasi-periodic* (like in [Bryngdahl74 p. 1290]); however, we reserve the term quasi-periodic only to its meaning in the context of the theory of almost-periodic functions (see Sec. B.5 in Appendix B of *Vol. I*).

random structure —

A structure consisting of randomly positioned elements. Note that every random structure is aperiodic, but the converse is not necessarily true. See also Sec. 2.4

coordinate-transformed structure —

A structure $r(x,y)$ which is obtained by the application of a non-linear coordinate transformation $\mathbf{g}(x,y)$ to a certain initial periodic or aperiodic structure $p(x,y)$. Note that $\mathbf{g}(x,y)$ is applied to the original structure $p(x,y)$ as a *domain transformation*; more formally, using vector notation, $r(\mathbf{x}) = p(\mathbf{g}(\mathbf{x}))$. Curvilinear gratings (such as parabolic or circular gratings) and gratings with a varying frequency (such as a grating with logarithmic line-distances) are coordinate-transformed structures. Examples of coordinate-transformed gratings (both periodic and aperiodic) are shown in the top row of Fig. 6.3, and examples of coordinate-transformed dot screens (both periodic and aperiodic) are shown in the top row of Fig. 3.5.

profile-transformed structure —

A structure $r(x,y)$ which is obtained by the application of a non-linear transformation $t(z)$ to the profile of a certain initial periodic or aperiodic structure $p(x,y)$. Note that $t(z)$ is applied to the original structure $p(x,y)$ as a *range transformation*; more formally, using vector notation, $r(\mathbf{x}) = t(p(\mathbf{x}))$. Screen gradations are an example of profile-transformed structures.

coordinate-and-profile transformed structure —

A structure $r(x,y)$ which is obtained from a certain initial periodic or aperiodic structure $p(x,y)$ by the application of both a non-linear coordinate-transformation $\mathbf{g}(x,y)$ and a non-linear profile-transformation $t(z)$. More formally, using vector notation, $r(\mathbf{x}) = t(p(\mathbf{g}(\mathbf{x})))$. An example of a coordinate-and-profile-transformed structure is given in Remark 2 of Sec. 10.2 in *Vol. I*.

intensity profile (of a structure $r(x,y)$) —

A function over the x,y plane whose value at each point (x,y) indicates the intensity (or more precisely, the reflectance or the transmittance) of the structure $r(x,y)$.

periodic profile (of a curvilinear grating, curved grid, etc.) —

The periodic profile of a curvilinear grating or a curved screen $r(x,y)$ is defined as the intensity profile of the original, uncurved periodic grating (or screen), before

the non-linear transformation has been applied to it (see Sec. 10.2 in *Vol. I*). Note that in periodic structures the periodic profile coincides with the intensity profile.

normalized periodic profile (of a curvilinear grating, curved grid, etc.) —

See Sec. 10.2 in *Vol. I*.

geometric layout (of a curvilinear grating, curved grid, etc.) —

The geometric layout of a curvilinear grating $r(x,y)$ is the locus of the centers of its curvilinear corrugations in the x,y plane; it is defined by the bending transformation of the curvilinear grating (see Sec. 10.2 in *Vol. I*). Similarly, the geometric layout of a curved grid or a curved screen is defined by its two bending functions.

bending transformation (of a curvilinear grating, curved grid, etc.) —

The bending transformation of a curvilinear grating $r(x,y) = p(g(x,y))$ is the non-linear coordinate transformation $g(x,y)$ which bends the original, uncurved grating $p(x')$ into the curvilinear grating $r(x,y)$. The bending transformation of a curved grid or a curved screen $r(x,y) = p(g_1(x,y), g_2(x,y))$ is the non-linear coordinate transformation $\mathbf{g}(x,y) = (g_1(x,y), g_2(x,y))$ which bends the original, uncurved grid or screen $p(x',y')$ into the curved structure $r(x,y)$. We usually assume that the bending transformation is smooth (a diffeomorphism), so that it has no abrupt jumps or other troublesome singularities.

I.3 Terms in the spectral domain

spectrum (or *frequency spectrum*; not to be confused with *colour spectrum*) —

The frequency decomposition of a given function, which specifies the contribution of each frequency to the function in question. The frequency spectrum is obtained by taking the Fourier transform of the given function.

visibility circle —

A circle around the spectrum origin whose radius represents the cutoff frequency, i.e., the threshold frequency beyond which fine detail is no longer detected by the eye. Obviously, its radius depends on several factors such as the viewing distance, the light conditions, etc. It should be noted that the visibility circle is just a first-order approximation. In fact, the sensitivity of the human eye is a continuous 2D bell-shaped function [Daly92 p. 6], with a steep “crater” in its center (representing frequencies which are too small to be perceived), and “notches” in the diagonal directions (owing to the drop in the eye sensibility in the diagonal directions [Ulichney87 pp.79–84]).

frequency vector —

A vector in the u,v plane of the spectrum which represents the geometric location of an impulse in the spectrum (see Sec. 2.2 and Fig. 2.1 in *Vol. I*).

DC impulse —

The impulse that is located on the spectrum origin. This impulse represents the frequency of zero, which corresponds to the constant component in the Fourier decomposition of the image; the amplitude of the DC impulse corresponds to the intensity of this constant component. This impulse is traditionally called the *DC impulse* because it represents in electrical transmission theory the direct current component, i.e., the constant term in the frequency decomposition of an electric wave; we are following here this naming convention.

comb (or *impulse-comb*, *Dirac-comb*, *impulse-train*) —

An infinite train of equally spaced impulses located on a straight line in the spectrum. Any 1D periodic function is represented in the spectrum by a comb centered on the spectrum origin. The step and the direction of this comb represent the frequency and the orientation of the periodic function; its impulse amplitudes, which are given by the Fourier series development of the periodic function, determine its intensity profile.

nailbed (or *impulse-nailbed*) —

An infinite 2D train of equally spaced impulses located in the spectrum on a dot-lattice (either square-angled or skewed). Any 2D periodic function is represented in the spectrum by an impulse nailbed centered on the spectrum origin. The steps and the two main directions of this nailbed represent the frequency and the orientation of the two main directions of the function's 2D periodicity; the impulse amplitudes, which are given by the 2D Fourier series development of the periodic function, determine its intensity profile.

support (of a *comb*, a *nailbed*, a *spectrum*, etc.) —

The set of the geometric locations on the u,v plane of all the impulses of the specified comb, nailbed, or spectrum.

line-impulse —

A generalized function which is impulsive along a 1D line through the plane, and null everywhere else. A line-impulse can be graphically illustrated as a “blade” whose behaviour is continuous along its 1D line support but impulsive in the perpendicular direction. For example, the spectrum of an aperiodic line grating is a Hermitian line impulse (see Sec. 7.3.1 and Figs. 7.7, 7.8). As another example, the spectrum of a parabolic cosinusoidal grating consists of two parallel line-impulses (see Example 10.5 in Sec. 10.3 of *Vol. I*). Note that the amplitude of a line-impulse does not necessarily die out away from its center, and it may even rapidly oscillate between two constant values.

curvilinear impulse —

A generalized function which is impulsive along a 1D curvilinear path through the plane, and null everywhere else. A curvilinear impulse can be graphically illustrated

as a curvilinear “blade” whose behaviour is continuous along its 1D curvilinear support but impulsive in the perpendicular direction.

hump —

A 2D continuous surface, often bell-shaped, elliptic or hyperbolic, which is defined around a given center on the plane. For example, the spectrum of an aperiodic dot screen is a Hermitian hump (see Sec. 7.4.1 and Fig. 2.10(e)). As another example, the convolution of two non-parallel line-impulses gives a hump (see Sec. 7.3.1, or Sec. 10.7.3 and Fig. 10.13 in *Vol. D*). Note that the amplitude of a hump does not necessarily die out away from its center, and in some cases it may even rapidly oscillate between two constant values.

impulsive spectrum —

A spectrum which only consists of impulses, i.e., whose support consists of a finite or at most denumerably infinite number of points. All periodic and almost-periodic functions have impulsive spectra.

line-spectrum —

A spectrum which consists of line impulses (see, for example, Fig. 10.11 in *Vol. D*).

hybrid spectrum —

A spectrum which contains any combination of impulses, line-impulses and continuous humps (as opposed to a purely impulsive spectrum, a purely line-spectrum or a purely continuous spectrum). See, for instance, Example 10.14 of Sec. 10.7.4 and Fig. 10.13 in *Vol. I*.

singular support (of a *spectrum*, etc.; distinguish from a *singular locus of a moiré*) —

The subset of the spectrum support over which the spectrum is impulsive. The singular support of a given spectrum includes the support of all the impulsive elements which are included in the spectrum (impulses, line-impulses, etc.), but not the support of continuous elements such as humps or wakes.

I.4 Terms related to moiré

moiré effect (or *moiré phenomenon*) —

A visible phenomenon which occurs when repetitive structures (such as line-gratings, dot-screens, etc.) are superposed. It consists of a new pattern which is clearly observed in the superposition, although it does not appear in any of the original structures. Moiré effects may occur also in the superposition of correlated (or at least partially correlated) aperiodic layers; such moiré effects are called *Glass patterns*.

Glass pattern —

A moiré pattern that occurs between two aperiodic layers (which are correlated or at least partially correlated; note that uncorrelated aperiodic layers do not generate

Glass patterns). Unlike moiré patterns between periodic or repetitive layers, a Glass pattern typically consists of a single structure (see, for example, Figs. 2.1–2.2). Often (but not always, as clearly shown in Chapter 7) a Glass pattern is brighter in its center (where the correlation between the layers is maximal and their individual elements almost coincide), and farther away it stabilizes at a darker gray level (because the layers are no longer correlated and their elements more often fall between each other, leaving less white area in the superposition). The Glass pattern reflects, therefore, the macroscopic gray level variations in the layer superposition. Usually it is also surrounded by typical geometric dot alignments in the microstructure, which are known as dot trajectories. Glass patterns are called after Leon Glass, who described them in the late 1960s [Glass69, Glass73].

linear Glass pattern —

A Glass pattern that is generated in the layer superposition along a straight line (a fixed line). See, for example, Figs. 2.3 and 6.1.

curvilinear Glass pattern —

A Glass pattern that is generated in the layer superposition along a curvilinear line (a curvilinear fixed line). See, for example, Figs. 3.16 and 6.7.

(k_1, \dots, k_m) -moiré —

The 1-fold periodic structure in the image domain which corresponds to the (k_1, \dots, k_m) -comb in the spectrum convolution (the spectrum of the superposition); see Sec. 2.8 in *Vol. I*. In other words, this is the moiré which is generated due to the interaction between the k_i harmonic frequencies of the respective layers in the superposition. This moiré may be visible if at least its fundamental impulse, the (k_1, \dots, k_m) -impulse, is located inside the visibility circle. More details on our moiré notational system can be found in Sec. 2.8 of *Vol. I*.

singular moiré (or *singular state*, *singular superposition*) —

A configuration of the superposed layers in which the period of the moiré in question becomes infinitely large (i.e., its frequency becomes 0), and hence it can no longer be seen in the superposition. Singular moirés are unstable moiré-free states, since the slightest deviation in the angle or scaling of any of the superposed layers may cause the moiré in question to “come back from infinity” and to reappear with a large, visible period. See Sec. 2.3.2.

stable moiré-free state —

A moiré-free configuration of the superposed layers in which no moiré becomes visible even when small deviations occur in the angle or in the scaling (or frequency) of any of the layers (see Sec. 2.3.2).

unstable moiré-free state —

A moiré-free configuration of the superposed layers in which any slight deviation in the angle or in the scaling (or frequency) of any of the layers causes the

reappearance of a moiré with a large, visible period (see Fig. 2.5). Any singular state is an unstable moiré-free state.

moiré profile (or *moiré intensity profile*; *moiré intensity surface*) —

A function which defines the intensity level (the macroscopic gray level) of the moiré at any point of the image (see Sec. 7.1, or Secs. 2.10 and 4.1 in *Vol. I*).

macrostructure, microstructure (within a superposition) —

The superposition of two or more layers (gratings, screens, etc.) may generate new structures which appear in the superposition but not in the original layers. These new structures can be classified into two categories: the *macrostructures*, i.e., the moiré effects proper, which are much coarser than the detail of the original layers; and the *microstructures*, i.e., the tiny geometric forms which are almost as small as the periods of the original layers, and are normally visible only from a close distance or through a magnifying glass. In the periodic case the microstructure may consist of *rosettes* (see Chapter 8 in *Vol. I*), while in the aperiodic case the microstructure may consist of *dot trajectories* (see Secs. 2.2 and 2.3.3).

rosettes —

The various tiny flower-like shapes which are often present in the microstructure of periodic dot-screen or grid superpositions (see Chapter 8 in *Vol. I*).

dot trajectories (not to be confused with *trajectories*) —

This term is reserved to the typical microstructure dot alignments that appear in the superposition of correlated aperiodic layers. These dot trajectories may have various geometric shapes, depending on the transformations undergone by the superposed layers. In the case of simple linear transformations such as layer rotations, layer scalings, etc. the resulting dot trajectories are rather simple (circular, radial, spiral, elliptic, hyperbolic, linear, etc.); see Figs. 2.1–2.3. But when the layer transformations are more complex, the resulting dot trajectories may have more interesting shapes (see, for example, Fig. 5.1). Dot trajectories are a typical property of the superposition of aperiodic layers, and they are not visible in superpositions of periodic layers.

additive / subtractive moiré (not to be confused with *additive superposition*) —

Classical terms often used in literature to designate moirés which are generated by frequency sums or frequency differences, respectively, in the spectrum. For example, the (1,-1)-moiré is subtractive, while the (1,1)-moiré is additive. Note, however, that these terms cannot be generalized to more complex cases such as the (1,1,-1)-moiré between three gratings. These terms are mostly useful in the superposition of two curvilinear gratings, where both the additive and the subtractive moiré are often observed simultaneously, each of them having a different shape and location (see, for example, Fig. 10.31 in *Vol. I*). In this case the most convenient way to define them is based on their indicial equations (see Sec. 11.2 in *Vol. I*): the additive moiré is the system of moiré fringes which

corresponds to the indicial equation $m+n=p$, while the subtractive moiré is the system of moiré fringes which is selected by the indicial equation $m-n=p$. Note that additive moirés are not generated between aperiodic layers (see Sec. 7.5).

I.5 Terms related to light and colour

colour spectrum (not to be confused with *frequency spectrum*) —

The wavelength decomposition of a given light, which specifies the contribution of each visible light wavelength λ (approximately between $\lambda = 380$ nm for violet and $\lambda = 750$ nm for red) to the given light. The colour spectrum determines the visible colour of the light in question. See Chapter 9 in *Vol. I* for more details.

monochrome (or *black-and-white*; not to be confused with *monochromatic*) —

Achromatic light, image, etc. involving only black, white and all the intermediate gray levels. The colour spectrum of an ideal monochrome light is flat, i.e., it has a constant value (between 0 and 1) for all wavelengths λ of the visible light.

monochromatic (not to be confused with *monochrome*) —

Chromatic light, image, etc. involving only a single pure wavelength λ of the visible light. The colour spectrum of an ideal monochromatic light consists of a single impulse of intensity 1 at the wavelength λ .

reflectance (or *reflectance function*) —

A function $r(x,y)$ which assigns to any point (x,y) of a monochrome image viewed by reflection a value between 0 and 1 representing its light reflection: 0 for black (or no reflected light), 1 for white (or full light reflection), and intermediate values for in-between shades. More formally, reflectance is defined at any point (x,y) as the ratio of reflected to incident radiant power [Wysecki82 p. 463].

transmittance (or *transmittance function*) —

A function $r(x,y)$ which assigns to any point (x,y) of a monochrome image viewed by transmission (such as a transparency, a film, etc.) a value between 0 and 1 representing its light transmission: 0 for black (or no transmitted light), 1 for white (or full light transmission), and intermediate values for in-between shades. More formally, transmittance is defined at any point (x,y) as the ratio of transmitted to incident radiant power [Wysecki82 p. 463].

chromatic reflectance (or *chromatic reflectance function*) —

A function $r(x,y;\lambda)$ which assigns to any point (x,y) of a colour image viewed by reflection its full colour spectrum. In other words, it gives for every wavelength λ of the visible light (approximately between $\lambda = 380$ nm and $\lambda = 750$ nm) a value between 0 and 1, which represents the reflectance of light of wavelength λ at the point (x,y) of the image. This is a straightforward generalization of the reflectance function $r(x,y)$ in the monochrome case (see Chapter 9 in *Vol. I*).

chromatic transmittance (or *chromatic transmittance function*) —

A function $r(x,y;\lambda)$ which assigns to any point (x,y) of a colour image viewed by transmission its full colour spectrum. In other words, it gives for every wavelength λ of the visible light (approximately between $\lambda = 380$ nm and $\lambda = 750$ nm) a value between 0 and 1, which represents the transmittance of light of wavelength λ at the point (x,y) of the image. This is a straightforward generalization of the transmittance function $r(x,y)$ in the monochrome case (see Chapter 9 in *Vol. I*).

I.6 Miscellaneous terms**binary** (grating, etc.) —

A structure which contains only two transmittance (or reflectance) levels: 0 and 1.

discrete —

A subset D of \mathbb{R}^n is called *discrete* if there exists a number $d > 0$ so that for any points $a, b \in D$ the distance between a and b is larger than d . Note, however, that the term *discrete* is also used as the opposite of *continuous*. For example: periodic functions have *discrete spectra*, while aperiodic functions have *continuous spectra*.

domain / range transformation —

Any image $r(x,y)$ (or function $r: \mathbb{R}^2 \rightarrow \mathbb{R}$) can undergo two types of coordinate transformations: Either a transformation of its *domain*, $r(x,y) \mapsto r(\mathbf{g}(x,y))$, or a transformation of its *range*, $r(x,y) \mapsto t(r(x,y))$. As explained in Sec. D.6 of Appendix D, in the first case $\mathbf{g}(x,y)$ is applied as an inverse transformation, while in the second case $t(x)$ is used as a direct transformation. Similarly, any mapping $\mathbf{f}(x,y)$, $\mathbf{f}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$, can undergo two types of coordinate transformations: Either a transformation of its *domain*, $\mathbf{f}(x,y) \mapsto \mathbf{f}(\mathbf{g}(x,y))$, or a transformation of its *range*, $\mathbf{f}(x,y) \mapsto \mathbf{g}(\mathbf{f}(x,y))$. Again, in the first case $\mathbf{g}(x,y)$ is applied as an inverse transformation, while in the second case it is used as a direct transformation.

direct / inverse mapping (not to be confused with *forward / backward mapping algorithm*) —

The mathematical terms used to designate a mapping (geometric transformation) \mathbf{g} and its inverse \mathbf{g}^{-1} . We have, therefore, $\mathbf{g} \circ \mathbf{g}^{-1} = \mathbf{g}^{-1} \circ \mathbf{g} = \mathbf{i}$, where \mathbf{i} is the identity transformation $\mathbf{i}(x,y) = (x,y)$, and where “ \circ ” is the operation of mapping composition: $(\mathbf{f} \circ \mathbf{g})(x,y) = \mathbf{f}(\mathbf{g}(x,y))$. Note that the designations *direct* and *inverse* are interchangeable, and they depend on our point of view; thus, if we focus our attention to the mapping $\mathbf{h} = \mathbf{g}^{-1}$, we may consider \mathbf{h} as the direct mapping and $\mathbf{h}^{-1} = (\mathbf{g}^{-1})^{-1} = \mathbf{g}$ as its inverse. For example, if $\mathbf{g}(x,y) = (2x, 2y)$ then $\mathbf{g}^{-1}(x,y) = (x/2, y/2)$; and if $\mathbf{g}(x,y) = (x/2, y/2)$ then $\mathbf{g}^{-1}(x,y) = (2x, 2y)$. Note that the terms *direct / inverse transformations* are also used in the same meaning, whereas the terms *direct / inverse transforms* are reserved to operations such as the Fourier transform.

forward / backward mapping algorithm (not to be confused with *direct / inverse mapping*) —

Digital imaging terms used to designate two different types of graphical algorithms or computer programs: A *forward mapping* operates by scanning the original input image pixel by pixel, and copying the value of the image at each input pixel location (x,y) onto the corresponding position (u,v) in the output image. The output position (u,v) is determined by passing the x and y coordinates of each input pixel through the transformation $(u,v) = \mathbf{g}(x,y)$. On the other hand, a *backward mapping* operates by scanning the target output image pixel by pixel, and copying onto each output pixel location (u,v) the value of the input pixel at the corresponding (x,y) position in the input image. The input position (x,y) that corresponds to a given output pixel location is determined by passing the u and v coordinates of each output pixel through the *inverse* transformation, $(x,y) = \mathbf{g}^{-1}(u,v)$. For more details on forward and backward mapping algorithms see Sec. D.9.1 in Appendix D.

trajectories (not to be confused with *dot trajectories*) —

This term is reserved to the *solution curves* of a system of differential equations, or, equivalently, to the *field lines* of the corresponding vector field. A system of differential equations $\frac{d}{dt}x(t) = g_1(x(t),y(t))$, $\frac{d}{dt}y(t) = g_2(x(t),y(t))$ has for solutions a family of curves in the x,y plane, whose parametric representation is $(x(t),y(t))$, and whose members differ from each other by some constants c [Kreyszig93 180–186]. Each of these solution curves is called a *trajectory* since it traces out the evolution of the curve as the parameter t is being varied. These trajectories are also the field lines of the mapping $\mathbf{g}(x,y) = (g_1(x,y),g_2(x,y))$ where this mapping is regarded as a 2D vector field that assigns to each point (x,y) the vector $\mathbf{g}(x,y)$. Note that the trajectories (field lines) of a vector field $\mathbf{g}(x,y)$ are the curves that are tangent to the vectors of the vector field at any point in the plane. For more details on the connections between the vector field $\mathbf{g}(x,y)$, the corresponding system of differential equations and their trajectories see Sec. B.6 in Appendix B.

critical point (not to be confused with a *fixed point*) —

A point (x,y) is called a critical point of the vector field $\mathbf{g}(x,y)$ or of the system of differential equations $\frac{d}{dt}x(t) = g_1(x(t),y(t))$, $\frac{d}{dt}y(t) = g_2(x(t),y(t))$ if it satisfies $g_1(x,y) = 0$, $g_2(x,y) = 0$ [Birkhoff89 p. 133; Kreyszig93 p. 176]. For more details see Sec. H.1 in Appendix H. Note that the term *critical point* is also used in mathematics to designate a point (x,y) where a surface $g(x,y)$ has an extremum or a horizontal inflection point, i.e. where $\frac{\partial}{\partial x}g(x,y) = 0$ and $\frac{\partial}{\partial y}g(x,y) = 0$ (see, for example, Sec. 2.19 in [Kaplan03]).

fixed point (not to be confused with a *critical point*) —

A point (x_F,y_F) is said to be a fixed point of transformation $\mathbf{g}(x,y)$ if it is not affected by the transformation, meaning that $\mathbf{g}(x_F,y_F) = (x_F,y_F)$. Note, however, that some references use the term “fixed point” for a critical point (see, for example, [Strogatz94 pp. 124, 150; Weisstein99 p. 652]).

fixed locus —

Transformation $\mathbf{g}(x,y)$ is said to have a fixed locus if it has a locus consisting of fixed points. For example, the transformation shown in Fig. 3.17 has a fixed locus consisting of an isolated point at the origin and a family of equispaced concentric circles surrounding it. Note that each point in a fixed locus is mapped by $\mathbf{g}(x,y)$ to itself; it is not sufficient that each point of the locus be mapped by $\mathbf{g}(x,y)$ to another point within the locus.

fixed line —

A fixed locus consisting of a straight line. Transformation $\mathbf{g}(x,y)$ has a fixed line if it has an entire straight line of fixed points. For example, each of the two transformations shown in Fig. 2.3(a)–(d) has a fixed line along the x axis, while each of the transformations shown in Fig. 3.6 has two fixed lines parallel to the x axis.

mutual fixed point —

A point (x_F, y_F) is said to be a mutual fixed point of transformations \mathbf{g}_1 and \mathbf{g}_2 if $\mathbf{g}_1(x_F, y_F) = \mathbf{g}_2(x_F, y_F)$. Note that the term *common fixed point* of \mathbf{g}_1 and \mathbf{g}_2 is already used in the mathematical literature for a point (x_F, y_F) that satisfies $\mathbf{g}_1(x_F, y_F) = (x_F, y_F) = \mathbf{g}_2(x_F, y_F)$, but this definition is too restrictive for our needs.

mutual fixed locus —

Transformations \mathbf{g}_1 and \mathbf{g}_2 are said to have a mutual fixed locus if there exists a locus in the x,y plane that consists of mutual fixed points of \mathbf{g}_1 and \mathbf{g}_2 .

almost fixed point —

A point (x_F, y_F) is said to be an almost fixed point of transformation $\mathbf{g}(x,y)$ if it is only very slightly affected by the transformation, meaning that $\mathbf{g}(x_F, y_F) \approx (x_F, y_F)$. Similarly, (x_F, y_F) is an almost mutual fixed point of \mathbf{g}_1 and \mathbf{g}_2 if $\mathbf{g}_1(x_F, y_F) \approx \mathbf{g}_2(x_F, y_F)$.

correlation (between two signals, images, etc.) —

A general term referring to the agreement or similarity between the two entities in question. Confusingly, this term is routinely used in several different meanings, notably as an abbreviation to the terms *local correlation*, *global correlation* or *cross correlation* (see below).

local correlation (between two signals, images, etc.) —

Two entities (signals, images, functions, etc.) are said to be locally correlated (or to be well correlated in a certain location) if they highly agree with each other in the specified location. As its name indicates, this is a local property, so that two given images can be highly correlated in some areas but not at all correlated in other areas. Local correlation between two superposed layers is the reason for the appearance of a Glass pattern in that area of the superposition. For example, the Glass patterns shown in Fig. 7.12 consist of four dark “2”-shaped areas; within these areas the two superposed layers are well correlated (the pinholes of the

second layer coincide with the tiny “2”-shaped dots of the first layer), while in the remaining areas of the figure the correlation between the two layers is low. See also Secs. 2.2, 7.8 and Problem 7-13.

global correlation (between two signals, images, etc.) —

Two entities (signals, images, functions, etc.) are said to be globally correlated if they highly agree with each other throughout their entire domain of definition. For example, two identical copies $r_1(x)$ and $r_2(x)$ of the same aperiodic signal (one or both of which may have undergone some amplitude transformation or have some additive random noise) are globally correlated. However, the signal $r_1(x)$ is no longer globally correlated with the shifted signal $r_2(x - x_0)$ or with the stretched signal $r_2(ax)$. See also Secs. 2.2, 7.8 and Problem 7-13.

cross correlation (between two signals, images, etc.) —

The cross correlation between two functions $f(x)$ and $g(x)$ is a third function $c_{f,g}(x)$ that indicates the relative amount of agreement between $f(x)$ and $g(x)$ for all possible degrees of misalignments (shifts). At each point x the value of the function $c_{f,g}(x)$ is defined as the area under the product of f and g after g has been shifted by x . For example, if $r_1(x)$ and $r_2(x)$ are defined as above (see under “global correlation”) then their cross correlation function consists of a high peak about $x = 0$, and low values everywhere else. The reason is that r_1 and r_2 are globally correlated when the shift between them is $x = 0$, but for any other shift they are no longer globally correlated. In the case of 2D images, points (x,y) in which the value of the cross correlation function is high indicate displacements of x,y between the two images in which the volume under the product of the two entire layers is high. Obviously, the cross correlation can detect displacements x,y in which the two *entire* images are *globally* well correlated, but it cannot detect isolated zones of *local* correlation between the two images, since the contribution of such local zones is relatively small, and it may be buried and lost within the global volume under the product of the entire images. The formal definition of cross correlation as well as its main mathematical properties are given in Appendix E. See also Secs. 2.2, 7.8 and Problem 7-13.

finite energy function (or signal) (not to be confused with *finite power function*) —

A function $f(x)$ is called a finite energy function (or a finite energy signal) if it is square integrable, i.e. if it satisfies:

$$\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$$

The class of finite energy functions includes all physically realizable functions, but it does not include many other useful functions such as constant functions, step functions, periodic functions or stationary random functions [Champeney73 p. 59; Coulon84 pp. 33–35].

finite power function (or signal) (not to be confused with *finite energy function*) —

A function $f(x)$ is called a finite power function (or a finite power signal) if it satisfies:

$$\lim_{a \rightarrow \infty} \frac{1}{a} \int_{-a/2}^{a/2} |f(x)|^2 dx < \infty$$

The class of finite power functions is larger than that of finite energy functions, and it includes, for example, constant functions, step functions, periodic functions and stationary random functions [Champeney73 p. 59; Coulon84 pp. 33–35].

almost-zero transformation (or *almost-null transformation*) —

A transformation $(u,v) = \mathbf{o}(x,y)$ is said to be *almost zero* or *almost null* if it differs only slightly, within our zone of interest (i.e. within the area covered by the layer superposition), from the zero transformation $(u,v) = \mathbf{z}(x,y) = (0,0)$. See Sec. D.12 in Appendix D.

weak transformation (or *almost-identity transformation*) —

A transformation $(u,v) = \mathbf{g}(x,y)$ is said to be *weak* or *almost identity* if it differs only slightly, within our zone of interest (i.e. within the area covered by the layer superposition), from the identity transformation $(u,v) = \mathbf{i}(x,y) = (x,y)$. In other words, $\mathbf{g}(x,y)$ is a weak transformation if it satisfies $\mathbf{g}(x,y) = (x,y) + \mathbf{o}(x,y)$, where $\mathbf{o}(x,y)$ is an almost-zero transformation. See Sec. D.12 in Appendix D.

diffeomorphism —

A diffeomorphism (in our case, on \mathbb{R}^2) is a one-to-one continuously differentiable mapping of \mathbb{R}^2 onto itself whose inverse mapping is also continuously differentiable.

scaling (of a *comb*, *nailbed*, etc.) —

We distinguish between *amplitude* scalings, and *period* or *frequency* scalings (in which the expansion or contraction occurs along the x,y axes in the image, or the u,v axes in the spectrum).

phase (of a periodic function, a periodic moiré, etc.) —

See Sec. C.4 in *Vol. I* and Chapter 7 Secs. 7.1–7.5 in *Vol. I*.

separable function (of two variables) —

A function $f(x,y)$ is said to be *separable* if it can be presented as (or separated into) a product of a function of x and a function of y : $f(x,y) = g(x) \cdot h(y)$ [Gaskill78 pp. 16–17; Cartwright90 p. 117]. Note, however, that we use this term in a slightly larger sense: A 2D function $f(x,y)$ is separable if it can be presented as a product of two independent 1D functions. Therefore, although $f(x,y) = g(x) \cdot h(y)$ may no longer be separable (in the narrower sense) after it has undergone a rotation or a skewing transformation, we will still consider it as separable (with respect to the rotated or skewed axes x' and y' : $f(x',y') = g(x') \cdot h(y')$).

inseparable function (of two variables) —

A function that is not *separable*. For example, the function representing a square white dot is separable: $\text{rect}(x,y) = \text{rect}(x) \cdot \text{rect}(y)$, while the function representing a circular white dot is inseparable.

spatially separable (not to be confused with a *separable function*) —

Two functions $F(u,v)$ and $G(u,v)$ in the spectrum are called spatially separable if their supports in the u,v plane are not overlapping. Spatially separable elements in the spectrum can be separated and extracted by means of filtering, i.e., by multiplying the spectrum with an appropriate 2D low-pass or band-pass filter.

spatially inseparable (not to be confused with an *inseparable function*) —

Two functions $F(u,v)$ and $G(u,v)$ in the spectrum are called spatially inseparable if their supports in the u,v plane are at least partially overlapping. Spatially inseparable elements in the spectrum cannot be separated or extracted by multiplying the spectrum with 2D low-pass or band-pass filters.

dots per inch (dpi) (or *dots per centimeter*; not to be confused with *lines per inch*) —

A term used to specify the resolution of a digital device such as a printer, a scanner, etc. For example, a device whose resolution is 300 dpi can only address points on an underlying pixel-grid whose period is $1/300$ of an inch, and no in-between points or pixel-fractions can be addressed. Some devices have different resolutions in the horizontal and in the vertical directions.

lines per inch (lpi) (or *lines per centimeter*; not to be confused with *dots per inch*) —

A term used to specify the frequency of gratings, dot-screens, etc. This term specifies the number of periods per inch. For example: the finest grating that can be produced on a 300 dpi device, namely: a sequence of alternating one-pixel wide black and white lines, is a grating of 150 lpi (since one period consists here of two device pixels).

List of notations and symbols

This list consists of the main symbols used in the present volume. They appear with a very brief description and a reference to the page in which they are first used or defined. Obvious symbols such as '+', '−', etc. have not been included.

Symbol	Short description	Page
x, y	The coordinates (axes) of the image plane	15
u, v	The coordinates (axes) of the spectral plane	15
x', y'	Rotated coordinates (axes) in the image plane	56
$r(x,y)$	A 2D reflectance (or transmittance) function	15
$R(u,v)$	The spectrum of $r(x,y)$	15
$d(x,y)$	A single dot (of a dot-screen)	32
$D(u,v)$	The spectrum of $d(x,y)$	32
*	1D convolution (or T -convolution)	245, 436
**	2D convolution (or T -convolution)	15, 251, 411
★	1D cross correlation	246, 268
★★	2D cross correlation	258, 269, 413
$c_{f,g}$	The cross correlation between the functions f and g	17, 413
α, θ, \dots	Angles	56, 58
φ_M	Angle of a moiré effect	231
$\alpha \rightarrow 0$	The angle α tends to 0	76
f, f_1, \dots	Frequencies of 1D periodic functions $p(x), p_1(x), \dots$	228, 231
f_M	Frequency of a moiré effect	231
T, T_1, \dots	Periods of 1D periodic functions $p(x), p_1(x), \dots$	59, 231
T_M	Period of a moiré effect	59, 231

F, P	Matrices	62
F^{-1}	The inverse of matrix F	63, 322
F^T	The transpose of matrix F	390
$ F $	The determinant of matrix F	255, 282
$J(x,y)$	The Jacobian determinant of a transformation	255, 309
a, b, r, s, t	Real numbers (sometimes also used as integer numbers)	58
i, j, m, n, p, q	Integer numbers	159, 189
i	(In complex numbers): the imaginary unit, $\sqrt{-1}$	32, 38
$\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{u}$	Vectors	62, 231, 438
$\mathbf{f}_1, \dots, \mathbf{f}_m$	Frequency vectors in the u, v plane of the spectrum	230
\mathbb{Z}	The set of all integer numbers (positive, negative, and 0)	159
\mathbb{R}	The set of all real numbers	16, 48
\mathbb{R}^n	The n -dimensional Euclidean space	47, 429, 463
$\dim V$	Dimension of vector space V	294
$\text{Im } \mathbf{g}$	The image of linear transformation \mathbf{g}	294
$\text{Ker } \mathbf{g}$	The kernel of linear transformation \mathbf{g}	294
$\text{Re}[\]$	The real-valued part of a complex entity	34, 38
$\text{Im}[\]$	The imaginary-valued part of a complex entity	34, 38
$\text{Abs}[\]$	The magnitude of a complex entity	36, 38
$\text{Arg}[\]$	The phase of a complex entity	36, 38
$\mathbf{g}(x,y)$	A 2D transformation	48
$\overline{\mathbf{g}}(x,y)$	A 2D transformation (used as direct mapping)	107, 110
$(x,y) \mapsto \mathbf{g}(x,y)$	The transformation \mathbf{g} maps the point (x,y) to $\mathbf{g}(x,y)$	110
$\mathbf{g}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$	A transformation \mathbf{g} from \mathbb{R}^2 to \mathbb{R}^2	289, 327
$a \cdot b$	Multiplication	15
$\mathbf{v} \cdot \mathbf{w}$	Scalar product of two vectors	246, 439

$\mathbf{f}_1 \circ \mathbf{f}_2$	The composition of transformations \mathbf{f}_1 and \mathbf{f}_2 , i.e. $\mathbf{f}_1(\mathbf{f}_2(\mathbf{x}))$	337, 406, 452
ε, δ	Arbitrarily small, positive real numbers	83, 88
$ a $	The absolute value of the number a (real or complex)	42, 179, 417
$F(u,v) = \mathcal{F}[f(x,y)]$	$F(u,v)$ is the Fourier transform of $f(x,y)$	32
\approx	Approximately equal	131, 406
\ll	Much smaller than	76
m	The number of superposed layers	15
(k_1, \dots, k_m)	An index-vector: an m -tuple of integers	231
$\mathbf{f}_{k_1, \dots, k_m}$	The frequency-vector of the (k_1, \dots, k_m) -impulse	230
a_{k_1, \dots, k_m}	The amplitude of the (k_1, \dots, k_m) -impulse	230, 235
(k_1, \dots, k_m) -moiré	The 1D moiré corresponding to the (k_1, \dots, k_m) -comb	231, 235
$\delta(u)$	The impulse symbol	426
$\delta(u,v)$	The 2D impulse symbol	32
$\text{rect}(x)$	A square pulse: 1 in the range $-0.5 \leq x \leq 0.5$, and 0 elsewhere	214
$\text{rect}(x,y)$	A 2D square pulse: 1 in the range $-0.5 \leq x, y \leq 0.5$, and 0 elsewhere	32
$\text{sinc}(x)$	$\frac{\sin(\pi x)}{\pi x}$ for $x \neq 0$, and 1 for $x = 0$	32
■	End of example, proof, etc.	16

List of abbreviations

Symbol	Short description	Page
1D	1-dimensional	48, 187
2D	2-dimensional	48, 187
DC	The impulse at the spectrum origin (i.e., at frequency zero)	438

CMYK	The four process ink colours: Cyan, Magenta, Yellow, black	45
DFT	Discrete Fourier transform	34
FFT	Fast Fourier transform	46
dpi	Dots per inch (printer resolution)	448
lpi	Lines per inch (frequency of a grating or a screen)	448
<i>iff</i>	If and only if	22

References

- [Adler81] R. J. Adler, *The Geometry of Random Fields*. John Wiley & Sons, Chichester, 1981.
- [Ahumada83] A. J. Ahumada, Jr., D. C. Nagel and A. B. Watson, "Reduction of display artifacts by random sampling," in *Applications of Digital Image Processing VI*, Proceedings of the SPIE, Vol. 432, 1983, pp. 216–221.
- [Allebach76] J. P. Allebach and B. Liu, "Random quasiperiodic halftone process," *Journal of the Optical Society of America*, Vol. 66, No. 9, September 1976, pp. 909–917.
- [Amidor00] I. Amidror, *The Theory of the Moiré Phenomenon*. Kluwer Academic Publishers, Dordrecht, 2000.
- [Amidor02] I. Amidror, "A new print-based security strategy for the protection of valuable documents and products using moiré intensity profiles," in *Optical Security and Counterfeit Deterrence Techniques IV*, R. L. Van Renesse (Ed.), Proceedings of SPIE Vol. 4677, SPIE, Bellingham, 2002, pp. 89–100.
- [Amidor02a] I. Amidror, "Scattered data interpolation methods for electronic imaging systems: a survey," *Journal of Electronic Imaging*, Vol. 11, No. 2, April 2002, pp. 157–176.
- [Amidor03] I. Amidror, "Glass patterns as moiré effects: new surprising results," *Optics Letters*, Vol. 28, No. 1, January 2003, pp. 7–9.
- [Amidor03a] I. Amidror, "Unified approach for the explanation of stochastic and periodic moirés," *Journal of Electronic Imaging*, Vol. 12, No. 4, October 2003, pp. 669–681.
- [Amidor03b] I. Amidror, "Glass patterns in the superposition of random line gratings," *Journal of Optics A: Pure and Applied Optics*, Vol. 5, No. 3, May 2003, pp. 205–215.
- [Amidor03c] I. Amidror, "Moiré patterns between aperiodic layers: quantitative analysis and synthesis," *Journal of the Optical Society of America A*, Vol. 20, No. 10, October 2003, pp. 1900–1919.
- [Amidor04] I. Amidror, "Dot trajectories in the superposition of random screens: analysis and synthesis," *Journal of the Optical Society of America A*, Vol. 21, No. 8, August 2004, pp. 1472–1487.
- [Amidor04a] I. Amidror, "Authentication with built-in encryption by using moire intensity profiles between random layers," Published US Patent Application No. 20040001604, 2004; issued as US Patent No. 7,058,202, 2006.
- [Andrews03] L. C. Andrews and R. L. Phillips, *Mathematical Techniques for Engineers and Scientists*. SPIE Press, Bellingham, Washington, 2003.
- [Anstis70] S. M. Anstis, "Phi movement as a subtraction process," *Vision Research*, Vol. 10, 1970, pp. 1411–1430.

- [Arnold73] V. I. Arnold, *Ordinary Differential Equations*. MIT Press, Cambridge, Massachusetts, 1973.
- [Bahuguna88] R. D. Bahuguna, A. B. Western and S. Lee, "Young's double-slit experiment using speckle photography", *American Journal of Physics*, Vol. 56, No. 6, June 1988, pp. 531–533.
- [Barlow04] H. B. Barlow and B. A. Olshausen, "Convergent evidence for the visual analysis of optic flow through anisotropic attenuation of high spatial frequencies," *Journal of Vision*, Vol. 4, No. 6, 2004, pp. 415–426.
- [Barrett97] K. E. Barrett, "Intersection of conics and a rounding error problem," *International Journal of Mathematical Education in Science and Technology*, Vol. 28, No. 1, 1997, pp. 141–145.
- [Bendat93] J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis*. John Wiley & Sons, NY, 1993 (second edition).
- [Bernstein05] D. S. Bernstein, *Matrix Mathematics*. Princeton University Press, Princeton, 2005.
- [Birkhoff89] G. Birkhoff and G-C. Rota, *Ordinary Differential Equations*. John Wiley & Sons, NY, 1989 (fourth edition).
- [Bracewell86] R. N. Bracewell, *The Fourier Transform and its Applications*. McGraw-Hill Publishing Company, Reading, NY, 1986 (second edition).
- [Bracewell95] R. N. Bracewell, *Two Dimensional Imaging*. Prentice Hall, NJ, 1995.
- [Bronshtein97] I. N. Bronshtein and K. A. Semendyayev, *Handbook of Mathematics*. Springer, Berlin, 1997 (third edition).
- [Bronstein90] I. N. Bronstein and K. A. Semendiaev, *Aide-Mémoire de Mathématiques*. Eyrolles, Paris, 1990 (9th edition; French translation).
- [Bryngdahl74] O. Bryngdahl, "Moiré: formation and interpretation," *Jour. of the Optical Society of America*, Vol. 64, No. 10, 1974, pp. 1287–1294.
- [Callahan74] J. Callahan, "Singularities and plane maps," *The American Mathematical Monthly*, Vol. 81, 1974, pp. 211–240.
- [Cantwell02] B. J. Cantwell, *Introduction to Symmetry Analysis*. Cambridge University Press, Cambridge, 2002.
- [Cardinal03] K. S. Cardinal and D. C. Kiper, "The detection of colored Glass patterns," *Journal of Vision*, Vol. 3, 2003, pp. 199–208.
- [Cartwright90] M. Cartwright, *Fourier Methods for Mathematicians, Scientists and Engineers*. Ellis Horwood, UK, 1990.
- [Casasent76] D. Casasent and D. Psaltis, "Position, rotation, and scale invariant optical correlation," *Applied Optics*, Vol. 15, No. 7, July 1976, pp. 1795–1799.
- [Casselmann04] W. Casselman, *Mathematical Illustrations: A Manual of Geometry and PostScript*. Cambridge University Press, Cambridge, 2004. Full text available on the Internet at: <http://www.math.ubc.ca/people/faculty/cass/graphics/text/www/index.html>
- [Castelman79] K. R. Castelman, *Digital Image Processing*. Prentice-Hall, New Jersey, 1979.
- [Champeney73] D. C. Champeney, *Fourier Transforms and their Physical Applications*. Academic Press, London, 1973.

- [Champeney87] D. C. Champeney, *A Handbook of Fourier Theorems*. Cambridge University Press, Cambridge, 1987.
- [Chosson06] S. Chosson, *Synthèse d'Images Moiré*. Ph.D. Thesis No. 3434, EPFL, Lausanne, 2006 (in French).
- [Churchill72] R. V. Churchill, *Operational Mathematics*. McGraw-Hill Kogakusha, Tokyo, 1972 (third edition).
- [Cloud95] G. L. Cloud, *Optical Methods of Engineering Analysis*. Cambridge University Press, Cambridge, 1995.
- [Colley98] S. J. Colley, *Vector Calculus*. Prentice Hall, New Jersey, 1998.
- [Coulon84] F. de Coulon, *Theorie et Traitement des Signaux*. Presses Polytechniques Romandes, Lausanne, 1984 (in French).
- [Courant88] R. Courant, *Differential and Integral Calculus (Vol. II)*. Wiley-Interscience, USA, 1988.
- [Courant89] R. Courant and F. John, *Introduction to Calculus and Analysis (Vol. II)*. Springer, NY, 1989.
- [Dakin97] S. C. Dakin, "The detection of structure in Glass patterns: psychophysics and computational models," *Vision Research*, Vol. 37, 1997, pp. 2227–2246.
- [Daly92] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," *Human Vision, Visual Processing and Digital Display III*, SPIE proceedings, Vol. 1666, USA, 1992, pp. 2–15.
- [Debnath95] L. Debnath, *Integral Transforms and Their Applications*. CRC Press, Boca Raton, 1995.
- [Dey91] T. W. Dey, "Optical alignment method using arbitrary geometric figures," US Patent No. 5,054,929, 1991.
- [Diggle83] P. J. Diggle, *Statistical Analysis of Spatial Point Processes*. Academic Press, London, 1983.
- [Ehrenpreis56] L. Ehrenpreis, "On the theory of kernels of Schwartz," *Proceedings of the American Mathematical Society*, Vol. 7, 1956, pp. 713–718.
- [Einstein02] *The Collected Papers of Albert Einstein (Vol. 7)*. Princeton University Press, NJ, 2002.
- [EncMath88] *Encyclopaedia of Mathematics*, Vols. 1–10. Kluwer Academic Publishers, Dordrecht, 1988–1994.
- [EncStat82] *Encyclopedia of Statistical Sciences*, Vols. 1–9. John Wiley & sons, NY, 1982–1988.
- [Erf78] R. K. Erf (Ed.), *Speckle Metrology*. Academic Press, NY, 1978.
- [Feitelson88] D. G. Feitelson, *Optical Computing: A Survey for Computer Scientists*. MIT Press, Cambridge, 1988.
- [Ferraro88] M. Ferraro and T. M. Caelli, "Relationship between integral transform invariances and Lie group theory," *Journal of the Optical Society of America A*, Vol. 5, No. 5, May 1988, pp. 738–742.
- [Foley90] J. D. Foley, A. van Dam, S. K. Feiner and J. F. Hughes, *Computer Graphics: Principles and Practice*. Addison-Wesely, Reading, Massachusetts, 1990 (second edition).

- [Garavaglia01] M. Garavaglia and P. F. Melián, "Making operations with superposed black and white and colors transparencies: Does quasi-moiré patterns observation suggest two forms to correlate retinal visual signals?" in *Proceedings of SPIE Vol. 4419*, SPIE, Bellingham, 2001, pp. 581–584.
- [Gardner88] W. A. Gardner, *Statistical Spectral Analysis: A Nonprobabilistic Theory*. Prentice Hall, Englewood Cliffs, 1988.
- [Gaskill78] J. D. Gaskill, *Linear Systems, Fourier Transforms, and Optics*. John Wiley & Sons, NY, 1978.
- [Gåsvik95] K. J. Gåsvik, *Optical Metrology*. John Wiley & Sons, NY, 1995 (second edition).
- [Glass69] L. Glass, "Moiré effect from random dots," *Nature*, Vol. 223, August 1969, pp. 578–580.
- [Glass73] L. Glass and R. Pérez, "Perception of random dot interference patterns," *Nature*, Vol. 246, December 1973, pp. 360–362.
- [Glass76] L. Glass and E. Switkes, "Pattern recognition in humans: correlations which cannot be perceived," *Perception*, Vol. 5, 1976, pp. 67–72.
- [Glass02] L. Glass, "Looking at dots," *The Mathematical Intelligencer*, Vol. 24, 2002, pp. 37–43.
- [Glassner95] A. S. Glassner, *Principles of Digital Image Synthesis (Vol. 1)*. Morgan Kaufmann, San Francisco, 1995.
- [Gonzalez87] R. C. Gonzalez and P. Wintz, *Digital image processing*. Addison-Wesley, Reading, Massachusetts, 1987 (second edition).
- [Gray97] A. Gray, M. Mezzino and M. A. Pinsky, *Introduction to Ordinary Differential Equations with Mathematica*. Springer, NY, 1997.
- [Halmos74] P. R. Halmos, *Naive Set Theory*. Springer, NY, 1974.
- [Harburn75] G. Harburn, C. A. Taylor and T. R. Welberry, *Atlas of Optical Transforms*. G. Bell & Sons, London, 1975.
- [Hardy48] A. C. Hardy and F. L. Wurzburg, Jr., "A photoelectric method for preparing printing plates," *Journal of the Optical Society of America*, Vol. 38, No. 4, pp. 295–300, 1948.
- [Harris98] J. W. Harris and H. Stocker, *Handbook of Mathematics and Computational Science*. Springer, NY, 1998.
- [Heiden69] C. Heiden, "Power spectrum of stochastic pulse sequences with correlation between the pulse parameters," *Physical Review*, Vol. 188, No. 1, pp. 319–326, 1969.
- [Hersch04] R. D. Hersch and S. Chosson, "Band moiré images," *Proceedings of SIGGRAPH 2004, ACM Transactions on Graphics*, Vol. 23, No. 3, 2004, pp. 239–248.
- [Hines75] M. E. Hines, "Image scrambling technique," US Patent No. 3,914,877, 1975.
- [Horn85] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.
- [Hotta99] K. Hotta, T. Kurita and T. Mishima, "Scale invariant face recognition method using spectral features of log-polar image," in *Proceedings of the SPIE Conference on Applications of Digital Image Processing XXII*, Denver, Colorado, July 1999, SPIE Vol. 3808, pp. 33–43.

- [Howse95] J. W. Howse IV, *Gradient and Hamiltonian Dynamics: Some Applications to Neural Network Analysis and System Identification*. Ph.D. thesis, The University of New Mexico, USA, 1995. <http://www.eece.unm.edu/controls/theses/Howse.pdf>
- [Huck03] J. Huck, *Mastering Moirés: Investigating Some of the Fascinating Properties of Interference Patterns*. Private publication by J. Huck, 2003. <http://pages.sbcglobal.net/joehuck>
- [Indebetouw92] G. Indebetouw and R. Czarnek (Eds.), *Selected Papers on Optical Moiré and Applications*. SPIE Milestone Series, Vol. MS64, SPIE Optical Engineering Press, Washington, 1992.
- [Ivanov95] V. I. Ivanov and M. K. Trubetskov, *Handbook of Conformal Mapping with Computer Aided Visualization*. CRC Press, Boca Raton, 1995.
- [Kang99] H. R. Kang, *Digital Color Halftoning*. SPIE, Bellingham, Washington, 1999.
- [Kaplan03] W. Kaplan, *Advanced Calculus*. Addison-Wesley, Boston, 2003 (5th edition).
- [Kipphan01] H. Kipphan, *Handbook of Print Media*. Springer, Berlin, 2001.
- [Knopp74] P. J. Knopp, *Linear Algebra: an Introduction*. Hamilton Publishing Company, California, 1974.
- [Koenderink90] J. J. Koenderink, *Solid Shape*. The MIT Press, Cambridge, Massachusetts, 1990.
- [Kreyszig78] E. Kreyszig, *Introductory Functional Analysis with Applications*. John Wiley & Sons, NY, 1978.
- [Kreyszig93] E. Kreyszig, *Advanced Engineering Mathematics*. John Wiley & Sons, NY, 1993 (7th edition).
- [Lang87] S. Lang, *Calculus of Several Variables*. Springer, New York, 1987 (3rd edition).
- [Lau98] D. L. Lau, G. R. Arce and N. C. Gallagher, "Green-noise digital halftoning," *Proc. of the IEEE*, Vol. 86, pp. 2424–2444, 1998.
- [Lau01] D. L. Lau, A. M. Khan and G. R. Arce, "Stochastic moiré," in Proceedings of the IS&T's PICS Conference 2001, Montréal, April 22–25 2001, pp. 96–100.
- [Lau02] D. L. Lau, "Stochastic moiré II," in Proceedings of the IS&T's PICS Conference 2002, Portland, April 7–10 2002, pp. 250–255.
- [Lau02a] D. L. Lau, A. M. Khan and G. R. Arce, "Minimizing stochastic moiré in frequency-modulated halftones by means of green-noise masks," *Journal of the Optical Society of America A*, Vol. 19, November 2002, pp. 2203–2217.
- [Lau03] D. L. Lau, R. Ulichney and G. R. Arce, "Blue- and green-noise halftoning models: a review of the spatial and spectral characteristics of halftone textures," *IEEE Signal Processing Magazine*, Vol. 20, July 2003, pp. 28–38.
- [Lay03] D. C. Lay, *Linear Algebra and its Applications*. Addison-Wesley, Boston, 2003 (third edition).
- [Lipschutz68] S. Lipschutz, *Theory and Problems of Linear Algebra*. Schaum's Outline Series, McGraw-Hill, New York, 1968.
- [Lipschutz01] S. Lipschutz and M. Lipson, *Theory and Problems of Linear Algebra*. Schaum's Outline Series, McGraw-Hill, New York, 2001 (third edition).

- [Mansfield76] L. E. Mansfield, *Linear Algebra with Geometric Applications*. Marcel Dekker, NY, 1976.
- [McGrew95] S. P. McGrew, "Anticounterfeiting method and device using holograms and pseudo-random dot patterns," US Patent No. 5,396,559, 1995.
- [Needham97] T. Needham, *Visual Complex Analysis*. Oxford University Press, Oxford 1997.
- [Nishijima64] Y. Nishijima and G. Oster, "Moiré patterns: their application to refractive index and refractive index gradient measurements," *Journal of the Optical Society of America*, Vol. 54, No. 1, January 1964, pp. 1–5.
- [Olver86] P. J. Olver, *Applications of Lie Groups to Differential Equations*. Springer, NY, 1986.
- [Oster63] G. Oster and Y. Nishijima, "Moiré patterns," *Scientific American*, Vol. 208, May 1963, pp. 54–63.
- [Papoulis65] A. Papoulis, *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Kogakusha, Tokyo, 1965.
- [Patorski93] K. Patorski, *Handbook of the Moiré Fringe Technique*. Elsevier, Amsterdam, 1993.
- [Pocheć95] P. Pocheć, "Moiré based stereo matching technique," in Proceedings of the IEEE International Conference on Image Processing (Vol. 2), Washington DC, October 23–26 1995, pp. 370–373.
- [Poston78] T. Poston and I. Stewart, *Catastrophe Theory and its Applications*. Dover, NY, 1978.
- [Poularikas96] A. D. Poularikas (Ed.), *The Transforms and Applications Handbook*. CRC Press, Boca Raton, 1996.
- [Pratt91] W. K. Pratt, *Digital Image Processing*. John Wiley & Sons, NY, 1991 (second edition).
- [Prokoski99] F. J. Prokoski, "Method and apparatus for flash correlation," US Patent No. 5,982,932, 1999.
- [Renesse98] R. L. van Renesse (Ed.), *Optical Document Security*. Artech House, Boston, 1998 (second edition).
- [Renesse05] R. L. van Renesse, *Optical Document Security*. Artech House, Boston, 2005 (third edition).
- [Ridolfi04] A. Ridolfi, *Power Spectra of Random Spikes and Related Complex Signals with Application to Communications*. Ph.D. Thesis No. 3157, EPFL, Lausanne, 2004. http://lcavwww.epfl.ch/~ridolfi/research/thesis_andrea.pdf
- [Rodriguez94] M. Rodriguez, "Promises and pitfalls of stochastic screening in the graphics arts industry," in Proceedings of the IS&T 47th Annual Conference, Rochester, NY, May 15–20 1994; also reprinted in *Recent Progress in Digital Halftoning*, R. Eschbach (Ed.), IS&T, 1994, pp. 34–37.
- [Rosenfeld82] A. Rosenfeld and A. C. Kak, *Digital Picture Processing (Vol.1)*. Academic Press, Florida, USA, 1982 (second edition).
- [Rubinstein91] J. Rubinstein, J. Segman and Y. Zeevi, "Recognition of distorted patterns by invariance kernels," *Pattern Recognition*, Vol. 24, No. 10, 1991, pp. 959–967.
- [Schläpfer94] K. Schläpfer, "Are fine screens an alternative to frequency modulation screening?" Proceedings of the TAGA Conference, Baltimore, May 1994, pp. 34–41.

- [Schuette97] W. Schuette, "Glass patterns in image alignment and analysis," US Patent No. 5,613,013, 1997.
- [Schwalbe97] D. Schwalbe and S. Wagon, *VisualDSolve: Visualizing Differential Equations with Mathematica*. Springer, NY, 1997.
- [Sharma03] G. Sharma (Ed.), *Digital Color Imaging Handbook*. CRC Press, Boca Raton, 2003.
- [Spiegel63] M. R. Spiegel, *Theory and Problems of Advanced Calculus*. Schaum's Outline Series, McGraw-Hill, New York, 1963.
- [Spiegel68] M. R. Spiegel, *Handbook of Formulas and Tables*. Schaum's Outline Series, McGraw-Hill, New York, 1968.
- [Steeb96] W-H. Steeb, *Continuous Symmetries, Lie Algebras, Differential Equations and Computer Algebra*. World Scientific, Singapore, 1996.
- [Stoyan95] D. Stoyan, W. S. Kendall and J. Mecke, *Stochastic Geometry and its Applications*. John Wiley & Sons, Chichester, 1995 (second edition).
- [Strogatz94] S. H. Strogatz, *Nonlinear Dynamics and Chaos*. Addison-Wesely, Reading, Massachusetts, 1994.
- [Svanbro04] A. Svanbro, *Speckle Interferometry and Correlation Applied to Large-Displacement Fields*. Ph.D. thesis, Luleå University of Technology, Sweden, 2004. <http://www.sirius.luth.se/expmek/Angelica/Thesis.pdf>
- [Tabor89] M. Tabor, *Chaos and Integrability in Nonlinear Dynamics: an Introduction*. John Wiley & Sons, NY, 1989.
- [Ulichney87] R. Ulichney, *Digital Halftoning*. MIT Press, USA, 1987.
- [Ulichney88] R. Ulichney, "Dithering with blue noise," *Proc. of the IEEE*, Vol. 76, pp. 56–79, 1988.
- [Vargady64] L. O. Vargady, "Moiré fringes as visual position indicators," *Applied Optics*, Vol. 3, No. 5, May 1964, pp. 631–636; reprinted in [Indebetouw92, pp. 604–609].
- [Walker80] J. Walker, "The amateur scientist: Visual illusions in random-dot patterns and television 'snow'," *Scientific American*, Vol. 242, No. 4, April 1980, pp. 136–140.
- [Walker80a] J. Walker, "The amateur scientist: More about random-dot displays, plus computer programs to generate them," *Scientific American*, Vol. 243, No. 5, November 1980, pp. 158–168.
- [Weisstein99] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics*. CRC, Boca Raton, 1999.
- [Wesner74] J. W. Wesner, "Screen patterns used in reproduction of continuous-tone graphics," *Applied Optics*, Vol. 13, No. 7, July 1974, pp. 1703–1710.
- [Widmer92] E. Widmer, K. Schlöpfer, V. Humbel and S. Persiev, "The benefits of frequency modulation screening," Proceedings of the TAGA Conference, Vancouver, April 1992, pp. 28–43.
- [Wikipedia05] http://en.wikipedia.org/wiki/Integral_transform, April 2005.
- [Williams86] P. P. Williams, D. H. Davies and G. Harburn, "On randomization techniques for the suppression of unwanted moiré patterns in images generated by a scanning system with a periodic amplitude defect," *Optica Acta*, Vol. 33, No. 10, 1986, pp. 1311–1319.
- [Williams92] D. R. Williams, "Photoreceptor sampling and aliasing in human vision," Chapter 2 in *Tutorials in Optics*, D. T. Moore (Ed.), OSA Annual Meeting, Rochester, NY, 1992.

- [Wilson98] H. R. Wilson and F. Wilkinson, "Detection of global structure in Glass patterns: implications for form vision," *Vision Research*, Vol. 38, 1998, pp. 2933–2947.
- [Wolberg90] G. Wolberg, *Digital Image Warping*. IEEE Press, California, 1990.
- [Wolf79] K. B. Wolf, *Integral Transforms in Science and Engineering*. Plenum Press, NY, 1979.
- [Wyszecki82] G. Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & Sons, NY, 1982 (second edition).
- [Yellott82] J. I. Yellott Jr., "Spectral analysis of spatial sampling by photoreceptors: topological disorder prevents aliasing," *Vision Research*, Vol. 22, 1982, pp. 1205–1210.
- [Yellott83] J. I. Yellott Jr., "Spectral consequences of photoreceptor sampling in the rhesus retina," *Science*, Vol. 221, July 1983, pp. 382–385.

Index

Page numbers followed by the letter “g” indicate entries in the glossary (Appendix I).

A

additive moiré, 260, 465g
additive superposition, *see under* superposition rules
affine transformation, 51–63, 281–284
almost fixed point, *see under* fixed point
almost-identity transformation, *see under* transformation
almost-null transformation, *see under* transformation
almost-periodic function, 459g
almost-zero transformation, *see under* transformation
amplitude spectrum, *see under* spectrum
anti-counterfeiting, *see* applications of Glass patterns: document security
aperiodic function, 16, 459g
aperiodic layer, *see under* layer
applications of Glass patterns, 4–5
 art, 4
 comparison of patterns, 4, 46
 comparing the scale of two copies of an image, 96
 detection and measurement of slight displacements or deformations, 4, 42, 97, 221, 259
 detection of periodic noise (or periodic residues) in an aperiodic structure, 275
 determination of the axis of rotation, 4, 96
 determination of the similarity or the degree of correlation between patterns, 4, 19
 document security (authentication, anti-counterfeiting), 4, 97, 259, 270
 high-precision registration, 4, 96, 221, 259
 human visual system, 4, 259
 identification of a given pattern within an image, 4, 46

latent images, 97, 224
optical alignment, 4, 96, 221, 259
speckle interferometry, 43–44
speckle metrology, 4, 42–43
stereo matching, 44, 48
visualizing the curve family of an indefinite integral of a function, 4, 151
visualizing the flow lines of a vector field, 4, 105, 148
visualizing the solutions of differential equations, 105, 148
visualizing the solutions of equations, 4, 98
visually locating (or illustrating) the fixed points of a transformation, 4, 63, 97
approaches for investigating Glass patterns:
 indicial equations, 159, 193–196
 probabilistic (or stochastic) approach, 279–280, 422–432
 spectral, Fourier-based approach, 15, 225, 238–280
authentication, *see* applications of Glass patterns: document security
autocorrelation, *see under* correlation
autocorrelation theorem, 417

B

backward mapping algorithm (in digital imaging), 386–389, 468g
bending rate, 65
bending transformation, 215, 461g
binary, 467g
blade, *see* line-impulse

C

cardioid, 100
Cartesian coordinates, *see under* coordinates
characteristic function, 355

cluster, 228, 235
 colour printing, 4, 24, 45–46
 colour spectrum, 466g
 comb, 228, 462g
 conformal transformation (or mapping), 296–298, 305, 324
 congruent layers, 239, 243, 249–250
 constraint, 191–193, 208–214
 contour lines, *see* level lines
 convolution, 15, 271, 411–420
 of combs, 230
 of line-impulses (blades), 240–242
 of nailbeds, 234–235
 convolution theorem, 15, 226, 230, 235, 240, 244, 249–250, 416–417
 coordinate-and-profile transformed structure, 460g
 coordinate transformation, 295–298
 see also transformation
 coordinate-transformed structure, 460g
 coordinates:
 Cartesian, 85–86, 100–102, 230–231, 329, 332–336, 341, 354, 365–374
 curvilinear, 295–298, 332–336
 polar, 85–86, 100–102, 228, 341, 354, 365–374
 correlation, 17, 418–420, 469g
 autocorrelation, 415, 423
 cross correlation, 17, 246, 258, 264–269, 271–273, 411–420, 470g
 global correlation, 17, 264–269, 272–273, 470g
 local correlation, 17, 264–269, 272–273, 469g
 cosinusoidal grating, *see under* grating
 counterfeit deterrents, *see* applications of Glass
 patterns: document security
 critical point:
 of a system of differential equations, 443, 468g
 of a transformation, 331, 468g
 of a vector field, 443, 468g
 cross correlation, *see under* correlation
 cross-correlation theorem, 416–417
 curved grid, 458g
 curved screen, 65–82, 254–258, 458g
 curvilinear grating, 196–208, 453, 458g
 curvilinear impulse, 462g
 cutoff frequency, 228
 cyclic convolution, *see* *T*-convolution

D

DC impulse, 462g
 decorrelation, 96

deterministic signal or image, 32, 421, 430–432
 diffeomorphism, 471g
 difference moiré, *see* subtractive moiré
 diffraction, 6
 direct transformation, 109–110, 133, 151–152, 160–161, 183–186, 292, 295, 327–410, 467g
 discrete, 467g
 displacement, *see* shift
 document security, *see under* applications of
 Glass patterns
 domain, 327, 332–336, 347
 domain transformation, 56, 65, 82, 109–110, 133, 151–152, 158–160, 183–186, 218, 273, 293–295, 347–356, 460, 467g
 dot-screen, *see* screen
 dot shape, 156, 236–238
 dot trajectories, 8, 12–14, 19–29, 105–156, 158–161, 465g
 classification, 443–451
 curve equations, 114–123, 152
 invariance properties, 175–181
 morphology, 106–107
 synthesis, 134–137, 149–151, 259, 280
 under different superposition rules, 139–140, 156
 visual interpretation, 140–148
 dots per inch (dpi), 472g

E

eigenvalues of a matrix, 444–454
 element distribution, *see under* layer
 extraction of a moiré, *see* moiré extraction

F

field line, *see* trajectory
 finite-energy signal (or function), 415–416, 470g
 finite-power signal (or function), 416–417, 471g
 first order moiré, 158, 180, 222
 fixed dot shape, 240, 252
 fixed element shape, 240, 248
 fixed line, 60–62, 65, 254–258, 469g
 almost fixed line, 90–95, 104
 fixed line shape, 240
 fixed locus, 47–72, 469g
 almost fixed locus, 64, 82, 87–96, 104
 mutual fixed locus, 72–82, 469g
 synthesis, 83–87, 100–102, 280
 fixed point, 47–72, 189–193, 399–405, 443, 468g
 almost fixed point, 64, 82, 87–96, 104, 469g
 mutual fixed point, 72–82, 154–156, 258, 288, 401–405, 469g

- multiple fixed points, 65, 80–82, 254–258, 287
- fixed point theorem, 47–49, 281–288
 - fixed point theorem for affine mappings, 48, 281–284
 - fixed point theorem for second-order polynomial mappings, 284–288
- flash correlation artifacts, 46, 457
- flow line, *see* trajectory
- forward mapping algorithm (in digital imaging), 386–389, 468g
- Fourier-based approach, *see also under* approaches for investigating Glass patterns
 - generalized, 252, 269, 433–442
 - in the aperiodic case, 238–269
 - in the periodic case, 226–238
- Fourier decomposition, 433–434
 - generalized, 434–435
- Fourier series, 228, 433
 - coefficients, 433
 - generalized, 434
- Fourier spectrum, *see* spectrum
- Fourier transform, 15, 32, 421, 423, 433
 - g-Fourier transform, 439
 - generalized, 433–442
 - inverse, 232, 236, 244, 250, 433
- frequency domain, *see* spectral domain; spectrum
- frequency vector, 228, 230, 233, 461g
- functional, 419
- fundamental Glass-pattern theorem:
 - for the superposition of aperiodic gratings, 268, 456
 - for the superposition of aperiodic screens, 269, 456

G

- generalized Fourier series, *see under* Fourier series
- generalized Fourier transform, *see under* Fourier transform
- geometric layout, 159, 174, 280, 461g
- geometric location of an impulse, *see* impulse: location
- Glass pattern, 1–5, 11–14, 19, 51–53, 64, 82, 158–161, 180, 222, 457, 463g
 - applications, *see* applications of Glass patterns
 - artificial, 275–277
 - behaviour under affine layer transformations, 59–63
 - behaviour under layer rotations, 51–53
 - behaviour under layer scalings, 53–54
 - behaviour under layer shifts, 54–59, 96

- behaviour under non-linear layer transformations, 63–72
- between aperiodic line gratings, 187–224, 238–248
- curvilinear, 464g
- historical background, 4–5
- hybrid (1,-1)-Glass pattern whose band carries 2D information, 452–456
- in multilayer superpositions, 30–31, 104
- intensity profile, 214–220, 224, 225, 238–269
- invariance properties, 175–181, 222, 224
- linear, 66–67, 70, 149, 240, 244, 288, 464g
- radius of, 277–278
- singular, 272
- synthesis, 47, 83–87, 100–102, 222–223, 259, 270, 280
- theorem, *see* fundamental Glass pattern theorem
- global correlation, *see under* correlation
- gradient lines (or curves), 302
- grating, 187–224, 457g
 - aperiodic, 196
 - cosine shaped, 458g
 - cosinusoidal, 458g
 - curvilinear, *see* curvilinear grating
 - periodic, 196
 - repetitive, 196
- gray-level surface, 215
- grid, 458g
 - curved, *see* curved grid
 - regular (or square), 458g

H

- halftone screen, 458g
 - see also* halftoning
- halftoning, 4, 24, 45
- Hermitian, 240, 249
- higher order moiré, 54, 158
- human visual system, 4, 7, 228, 259
- hump, 242, 249, 463g
- hybrid between Glass and moiré pattern, *see* superposition: of partly periodic layers
- hybrid (1,-1)-Glass pattern whose band carries 2D information, 452–456
- hybrid (1,-1)-moiré whose bands carry 2D information, 452–456
- hybrid spectrum:
 - continuous and impulsive, 463g

I

- image, 15, 459g
- image domain, 15
- image of a linear transformation, 294

impulse, 228
 amplitude, 228
 DC, *see* DC impulse
 line, *see* line-impulse
 location, 228
 impulsive spectrum, *see under* spectrum
 indicial equations, *see under* approaches for
 investigating Glass patterns
 inseparable, 472g
 integral transform, 433–442
 intensity profile, 214–220, 224, 225, 238–259,
 452–456, 460g
 see also under moiré, Glass pattern
 inverse Fourier transform, *see* Fourier transform:
 inverse
 inverse transformation, 109–110, 327–410, 467g
 isometric layers, 238, 244–246, 250, 258
 isotropic mapping, 324

J

Jacobian, 255, 284, 286, 291, 309
 geometric interpretation, 309–311
 Jacobian matrix, 309, 322–326, 448–451

K

kernel of a linear transformation, 294
 kernel of an integral transform, 435–439

L

laser speckle, *see* speckle
 latent images, *see under* applications of Glass
 patterns
 layer:
 aperiodic, 1–3, 11, 15–19, 63–64
 constrained, *see* constraint
 deterministic, 421, 430–432
 element distribution within a layer, 28–30
 intermediate between periodic and aperiodic,
 17, 50–51, 98–101, 172–174, 260–262
 periodic, 1–3, 11, 15–19, 63–64
 pseudo-random, 28, 275–276
 random, 1–3, 11, 15–19, 28–29, 421–432
 repetitive, 1–3, 11, 15–19, 63–64
 stochastic, 16
 level lines, 100, 159, 162, 164, 167, 170, 174,
 289–291, 295, 303–305, 330, 338–339, 358
 limaçon, 100
 line-grating, *see* grating
 line-grid, *see* grid
 line impulse, 240–248, 462g
 line of coincidence, 189–193
 linear algebra, 345–346
 lines per inch (lpi), 472g
 local correlation, *see under* correlation

locally reflecting transformation, *see* reflecting
 transformation
 lpi (lines per inch), 472g

M

macrostructure, 18, 26–28, 180, 214, 222, 465g
 and microstructure, 18, 26–28
 magnitude spectrum, *see* spectrum: amplitude
 spectrum
 mapping, *see* transformation
 measuring (displacements, deformations, etc.),
 see various entries under applications of
 Glass patterns
 metrology, *see under* applications of Glass
 patterns
 microstructure, 18, 19, 26–28, 105–156, 180,
 465g
 and macrostructure, 18, 26–28
 artifacts, 28, 142–143
 morphology, 106–107
 under different superposition rules, 139–140,
 156
 invariance properties, 175–181
 visual interpretation, 140–148
 moiré, 1–3, 11, 24, 64, 158–161, 463g
 additive, *see* additive moiré
 between aperiodic layers, *see* Glass pattern
 between periodic and aperiodic layers, 46
 between periodic layers, 1, 11, 64
 between random layers, *see* Glass patterns
 hybrid (1,-1)-moiré whose bands carry 2D
 information, 452–456
 indexing, *see* moiré notational system
 intensity profile, 214–220, 225–238, 465g
 invariance properties, 175–181
 notational system for, *see* moiré notational
 system
 of higher order, *see* higher order moiré
 of the first order, *see* first order moiré
 period, 230–231
 profile, *see* moiré: intensity profile
 profile extraction:
 in superposed gratings, 232–233, 244–
 245
 in superposed screens, 235–236, 249–250
 singular, 464g
 subtractive, *see* subtractive moiré
 temporal, 6
 theorem, *see* fundamental moiré theorem
 unwanted, *see* unwanted moirés
 moiré analysis:
 qualitative, 225
 quantitative, 225
 moiré extraction, 232, 236, 244, 249–250

moiré-free superposition, 24–26
 stable (non-singular), 24–28, 464g
 unstable (singular), 24–26, 464g
 moiré notational system, 231
 (1,-1)-moiré, 231–233, 242–243, 259–260, 440
 (1,1)-moiré, 260
 (1,0,-1,0)-moiré, 235, 249, 254, 259–260
 (1,0,1,0)-moiré, 260
 (k_1, k_2)-moiré, 231, 260, 440
 (k_1, k_2, k_3, k_4)-moiré, 235, 260
 (k_1, \dots, k_m)-moiré, 464g
 monochromatic, 466g
 monochrome, 15, 259, 466g
 multichromatic, *see* polychromatic
 multilayer superposition, *see under*
 superposition
 multiplicative superposition, *see under*
 superposition rules
 mutual fixed locus, *see under* fixed locus
 mutual fixed point, *see under* fixed point

N

nailbed, 38, 228, 230, 233, 462g
 non-reflecting transformation, 255, 306–308, 312–313
 normalization, 232, 236, 244–245, 250–251
 notational system for moirés, *see* moiré
 notational system

O

operator:
 binary, 411
 linear, 412
 unary, 411
 optical alignment, *see under* applications of
 Glass patterns

P

parallax, 6
 period, 459g
 period-vector, 459g
 periodic function, 16, 459g
 1-fold periodic function, 459g
 2-fold periodic function, 459g
 periodic layer, *see under* layer
 periodic profile, 460g
 normalized, 461g
 phase, 471g
 phase spectrum, *see under* spectrum
 pinhole screen, *see under* screen
 point of coincidence, 52, 70, 77
 point process, 422, 425–426
 Poisson process, 425–426

polar coordinates, *see under* coordinates
 polychromatic, 15, 259
 power spectrum, *see under* spectrum
 precision alignment, *see* applications of Glass
 patterns: optical alignment
 precision measurement, *see various entries*
 under applications of Glass patterns
 profile, *see* intensity profile; periodic profile
 profile-transformed structure, 460g
 pseudo-random, *see under* layer

Q

qualitative moiré analysis, *see under* moiré
 analysis
 quantitative moiré analysis, *see under* moiré
 analysis

R

random scanning, 45
 random field, 429–430
 random image, *see* layer: random
 random layer, *see* layer: random
 random process, 32, 422
 random sampling, 44–45
 random structure, 460g
 range, 327, 332–336, 347
 range transformation, 218, 347–356, 460, 467g
 reflectance function, 15, 466g
 chromatic, 466g
 reflecting transformation, 255, 306–308, 312–313
 repetitive layer, *see under* layer
 repetitive, non-periodic structure, 460g
 rosettes, 465g

S

sampling moiré, 44–45
 scaling, 411, 471g
 scanning moiré, *see* sampling moiré
 screen, 458g
 curved, *see* curved screen
 halftone, *see* halftone screen; halftoning
 pinhole, 227, 236, 252–253
 regular, 458g
 screen gradation, 459g
 separable, 471g
 see also spatially separable
 shift, 54–59
 shot noise, 426–429
 similar layers, 239, 249–250
 similarity matrix (or transformation), 58, 324
 singular:
 Glass pattern, *see* Glass pattern: singular
 linear transformation, 282

- moiré, *see* moiré: singular
 - state, *see* superposition: singular
 - superposition, *see* superposition: singular
 - support, 463g
 - slits, 453
 - spatially separable, 472g
 - speckle, 42–44
 - interferometry, *see under* applications of
 - Glass patterns
 - metrology, *see under* applications of
 - Glass patterns
 - spectral approach, *see under* approaches for
 - investigating Glass patterns
 - in the periodic case, 226–238
 - in the aperiodic case, 238–269
 - spectral domain, 15, 226
 - see also* spectrum
 - spectrum, 15, 461g
 - amplitude spectrum, 32–42, 423
 - colour, *see* colour spectrum
 - continuous, 16, 38, 226, 230, 248
 - diffuse, 16, 32–33, 38, 226, 230, 424
 - hybrid, *see* hybrid spectrum
 - imaginary part, 32–42, 423
 - impulsive, 16, 39, 226, 230, 248, 463g
 - magnitude spectrum, *see* spectrum: amplitude spectrum
 - of a (k_1, k_2) -moiré between two periodic gratings, 231
 - of a (k_1, k_2, k_3, k_4) -moiré between two periodic screens, 235
 - of a partially periodic function (intermediate between periodic and aperiodic), 17, 38
 - of a periodic function, 16, 32–42, 226, 228
 - of a random screen, 32–42
 - of a superposition, 15
 - of an aperiodic function, 16, 32–42, 226, 228, 230
 - phase spectrum, 32–42, 423
 - power spectrum, 32, 417, 421–432, 423–425
 - real part, 32–42, 423
 - stable moiré-free state, *see under* moiré-free superposition
 - stochastic process, *see* random process
 - subtractive moiré, 260, 465g
 - superposition, 1, 15
 - macrostructure of, *see* macrostructure
 - microstructure of, *see* microstructure
 - of aperiodic layers, 1–5, 11, 19–29, 50, 157
 - of correlated aperiodic gratings, 240–246
 - of correlated aperiodic screens, 249–252
 - of line-gratings, 187–224, 230–233
 - of partly correlated layers, 21, 262–264
 - of partly periodic (or partly random) layers, 50–51, 98–101, 172–174, 260–262
 - of periodic layers, 1–2, 11, 50, 157
 - of random layers, 1–5, 11
 - of repetitive, non-periodic layers, 1–2, 11, 63
 - of screens, 12–31, 233–238, 248–259
 - of several layers, 30–31, 104
 - of uncorrelated aperiodic gratings, 246–248
 - of uncorrelated aperiodic screens, 258–259
 - singular, 24–26
 - superposition moiré in colour printing, 45–46
 - superposition rules, 15–16, 139–140, 156, 280
 - additive, 15–16
 - inverse additive, 16
 - multiplicative, 15
 - support (of a comb, a nailbed, a spectrum etc.), 462g
 - singular, 463g
 - synthesis of dot trajectories, *see* dot trajectories: synthesis
 - synthesis of Glass patterns, *see* Glass pattern: synthesis
- T**
- T -convolution, 233, 236, 416
 - T -convolution theorem:
 - in one dimension, 233
 - in two dimensions, 236
 - T -cross-correlation, 416
 - Talbot effect, 6
 - temporal moiré, 6
 - trajectory, 105, 107, 148, 298, 331, 468g
 - transform, 436, 467
 - see also* Fourier transform, integral transform
 - transformation, 289, 327, 347, 436
 - see also* direct transformation, inverse transformation, domain transformation, range transformation
 - active and passive interpretations, 343–347
 - affine, *see* affine transformation
 - almost identity, 406, 471g
 - almost null, 406, 471g
 - almost zero, 406–410, 471g
 - Cartesian to polar coordinate transformation, 340–341, 354, 365–374
 - conformal, *see* conformal transformation
 - direct, *see* direct transformation
 - identity, 337, 404, 406
 - interpretations of a 2D transformation, 289–308, 329–331
 - inverse, *see* inverse transformation
 - linear, 281–282
 - parabolic, 64–72

polar to Cartesian coordinate transformation,
340–341, 354, 365–374
second-order polynomial, 65, 284–288
weak, 98, 123, 132–133, 155, 161, 167, 174–
175, 182, 406–410, 471g
zero, 406

transformation surface, 215

translation, *see* shift

transmittance function, 15, 466g
chromatic, 467g

U

unstable moiré-free state, *see under* moiré-free
superposition

unwanted moirés, 4–5
avoiding, 4–5, 44–45

V

vector field, 102, 109–111, 298–301, 330

visibility circle, 228, 461g

W

weak transformation, *see under* transformation

wedge, *see* screen gradation

Z

zero transformation, *see under* transformation

Computational Imaging and Vision

1. B.M. ter Haar Romeny (ed.): *Geometry-Driven Diffusion in Computer Vision*. 1994
ISBN 0-7923-3087-0
2. J. Serra and P. Soille (eds.): *Mathematical Morphology and Its Applications to Image Processing*. 1994
ISBN 0-7923-3093-5
3. Y. Bizais, C. Barillot, and R. Di Paola (eds.): *Information Processing in Medical Imaging*. 1995
ISBN 0-7923-3593-7
4. P. Grangeat and J.-L. Amans (eds.): *Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*. 1996
ISBN 0-7923-4129-5
5. P. Maragos, R.W. Schafer and M.A. Butt (eds.): *Mathematical Morphology and Its Applications to Image and Signal Processing*. 1996
ISBN 0-7923-9733-9
6. G. Xu and Z. Zhang: *Epipolar Geometry in Stereo, Motion and Object Recognition. A Unified Approach*. 1996
ISBN 0-7923-4199-6
7. D. Eberly: *Ridges in Image and Data Analysis*. 1996
ISBN 0-7923-4268-2
8. J. Sporring, M. Nielsen, L. Florack and P. Johansen (eds.): *Gaussian Scale-Space Theory*. 1997
ISBN 0-7923-4561-4
9. M. Shah and R. Jain (eds.): *Motion-Based Recognition*. 1997
ISBN 0-7923-4618-1
10. L. Florack: *Image Structure*. 1997
ISBN 0-7923-4808-7
11. L.J. Latecki: *Discrete Representation of Spatial Objects in Computer Vision*. 1998
ISBN 0-7923-4912-1
12. H.J.A.M. Heijmans and J.B.T.M. Roerdink (eds.): *Mathematical Morphology and its Applications to Image and Signal Processing*. 1998
ISBN 0-7923-5133-9
13. N. Karssemeijer, M. Thijssen, J. Hendriks and L. van Erning (eds.): *Digital Mammography*. 1998
ISBN 0-7923-5274-2
14. R. Highnam and M. Brady: *Mammographic Image Analysis*. 1999
ISBN 0-7923-5620-9
15. I. Amidror: *The Theory of the Moiré Phenomenon*. 2000
ISBN 0-7923-5949-6;
Pb: ISBN 0-7923-5950-x
16. G.L. Gimel'farb: *Image Textures and Gibbs Random Fields*. 1999
ISBN 0-7923-5961
17. R. Klette, H.S. Stiehl, M.A. Viergever and K.L. Vincken (eds.): *Performance Characterization in Computer Vision*. 2000
ISBN 0-7923-6374-4
18. J. Goutsias, L. Vincent and D.S. Bloomberg (eds.): *Mathematical Morphology and Its Applications to Image and Signal Processing*. 2000
ISBN 0-7923-7862-8
19. A.A. Petrosian and F.G. Meyer (eds.): *Wavelets in Signal and Image Analysis. From Theory to Practice*. 2001
ISBN 1-4020-0053-7
20. A. Jaklič, A. Leonardis and F. Solina: *Segmentation and Recovery of Superquadrics*. 2000
ISBN 0-7923-6601-8
21. K. Rohr: *Landmark-Based Image Analysis. Using Geometric and Intensity Models*. 2001
ISBN 0-7923-6751-0
22. R.C. Veltkamp, H. Burkhardt and H.-P. Kriegel (eds.): *State-of-the-Art in Content-Based Image and Video Retrieval*. 2001
ISBN 1-4020-0109-6
23. A.A. Amini and J.L. Prince (eds.): *Measurement of Cardiac Deformations from MRI: Physical and Mathematical Models*. 2001
ISBN 1-4020-0222-X

Computational Imaging and Vision

24. M.I. Schlesinger and V. Hlaváč: *Ten Lectures on Statistical and Structural Pattern Recognition*. 2002 ISBN 1-4020-0642-X
25. F. Mokhtarian and M. Bober: *Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization*. 2003 ISBN 1-4020-1233-0
26. N. Sebe and M.S. Lew: *Robust Computer Vision*. Theory and Applications. 2003 ISBN 1-4020-1293-4
27. B.M.T.H. Romeny: *Front-End Vision and Multi-Scale Image Analysis*. Multi-scale Computer Vision Theory and Applications, written in Mathematica. 2003 ISBN 1-4020-1503-8
28. J.E. Hilliard and L.R. Lawson: *Stereology and Stochastic Geometry*. 2003 ISBN 1-4020-1687-5
29. N. Sebe, I. Cohen, A. Garg and S.T. Huang: *Machine Learning in Computer Vision*. 2005 ISBN 1-4020-3274-9
30. C. Ronse, L. Najman and E. Decencière (eds.): *Mathematical Morphology: 40 Years On*. Proceedings of the 7th International Symposium on Mathematical Morphology, April 18–20, 2005. 2005 ISBN 1-4020-3442-3
31. R. Klette, R. Kozera, L. Noakes and J. Weickert (eds.): *Geometric Properties for Incomplete Data*. 2006 ISBN 1-4020-3857-7
32. K. Wojciechowski, B. Smolka, H. Palus, R.S. Kozera, W. Skarbek and L. Noakes (eds.): *Computer Vision and Graphics*. International Conference, ICCVG 2004, Warsaw, Poland, September 2004, Proceedings. 2006 ISBN 1-4020-4178-0
33. K. Daniilidis and R. Klette (eds.): *Imaging Beyond the Pinhole Camera: Proceedings of the Twelfth Workshop Theoretical Foundations of Computer Vision*. 2006 ISBN 1-4020-4893-9
34. I. Amidror: *The Theory of the Moiré Phenomenon*. Volume 2. 2007 ISBN 1-4020-5457-2