Chapter 7
Critical Points

Abstract At a regular point, the linear terms of a function determine its local behavior, and there is a local coordinate change that transforms the function into one of the new coordinates. At a critical point, the linear terms vanish, but there is still an analogous result for the quadratic terms, called *Morse's lemma*. However, the quadratic terms may not determine the local behavior, but when they do (the critical point is then said to be *nondegenerate*), Morse's lemma provides a local coordinate change that transforms the function into a sum of positive and negative squares of the new coordinates. In this chapter we analyze Morse's lemma and use it to characterize critical points.

7.1 Functions of one variable

Let us see how a coordinate change can transform $y = f(x)$ into a pure square near a critical point $x = a$. As happens so often in local analysis, the key tool is Taylor's theorem. We need the first-order expansion; it helps us to write the remainder using the explicit integral formula that is given in the original formulation of the theorem (Theorem 3.9, p. 79). In fact, because $f'(a) = 0$, the only nonconstant term in the expansion is the remainder:

$$f(a+\Delta x) = f(a) + h(\Delta x)\,(\Delta x)^2.$$

The variable coefficient $h(\Delta x)$ in the remainder term is the integral

$$h(\Delta x) = \int_0^1 f''(a+t\Delta x)(1-t)\,dt.$$

Because we need $h(\Delta x)$ to be continuously differentiable for all Δx near 0, we require f to have a continuous third derivative. Then

J.J. Callahan, *Advanced Calculus: A Geometric View*, Undergraduate Texts in Mathematics,
DOI 10.1007/978-1-4419-7332-0_7, © Springer Science+Business Media, LLC 2010

$$h'(\Delta x) = \int_0^1 f'''(a + t\Delta x)\, t(1-t)\, dt.$$

Substituting $\Delta x = 0$ gives

$$h(0) = f''(a)\int_0^1 (1-t)\, dt = \frac{f''(a)}{2}, \quad h'(0) = f'''(a)\int_0^1 t(1-t)\, dt = \frac{f'''(a)}{6}.$$

If a coordinate change $\Delta x \to \Delta u$ is to transform $\Delta y = f(a+\Delta x) - f(a)$ into a pure square, $\Delta y = \pm(\Delta u)^2$, then Δu must be

$$\Delta u = p(\Delta x) = \Delta x \sqrt{|h(\Delta x)|}.$$

It remains to see whether p is a valid coordinate change. Before we do this, note how the "$\pm$" comes into play in the formula for Δy. Because

$$(\Delta u)^2 = (\Delta x)^2 |h(\Delta x)| = \begin{cases} (\Delta x)^2 h(\Delta x) = \Delta y & \text{if } h(\Delta x) \geq 0, \\ -(\Delta x)^2 h(\Delta x) = -\Delta y & \text{if } h(\Delta x) < 0, \end{cases}$$

it follows that

$$\Delta y = \begin{cases} +(\Delta u)^2 & \text{if } h(\Delta x) \geq 0, \\ -(\Delta u)^2 & \text{if } h(\Delta x) < 0. \end{cases}$$

Now consider the function p. Formal differentiation gives

$$p'(\Delta x) = \sqrt{|h(\Delta x)|} \pm \frac{h'(\Delta x)}{2\sqrt{|h(\Delta x)|}} \Delta x,$$

implying p has a continuous derivative on any interval where $h(\Delta x) \neq 0$. (The sign in the formula for p' is chosen to be the sign of $h(\Delta x)$.) Moreover,

$$p'(0) = \sqrt{|h(0)|} = \sqrt{|f''(a)|/2}.$$

Thus, if $f''(a) \neq 0$, the inverse function theorem implies that p is a valid coordinate change on an open interval containing $\Delta x = 0$, and we then have

$$y = \begin{cases} f(a) + (\Delta u)^2 & \text{if } f''(a) > 0, \\ f(a) - (\Delta u)^2 & \text{if } f''(a) < 0. \end{cases}$$

If $f''(a) = 0$, our argument fails to obtain the coordinate change p, but it is natural to ask if a better argument would repair the problem. The answer is no; that is, if $f''(a) = 0$, there may be no new coordinate Δu for which $y = f(a) \pm (\Delta u)^2$. We can see this geometrically, because the equation $y = f(a) + (\Delta u)^2$ necessarily implies f has a minimum at a, and $y = f(a) - (\Delta u)^2$ implies f has a maximum there. But the function $f(x) = x^3$ has a critical point at the origin for which $f''(0) = 0$, and the origin is neither a minimum nor a maximum.

For functions of a single variable, the preceding discussion establishes two results: Morse's lemma and the more familiar second derivative test.

Theorem 7.1 (Morse's lemma). *Suppose $y = f(x)$ has a continuous third derivative on an open interval that includes a critical point $x = a$ where $f''(a) \neq 0$. Then in a sufficiently small window centered at $x = a$ there is a coordinate change $\Delta u = p(\Delta x)$ for which*

$$\Delta y = \pm(\Delta u)^2,$$

where the sign of $(\Delta u)^2$ is chosen to be the sign of $f''(a)$. □

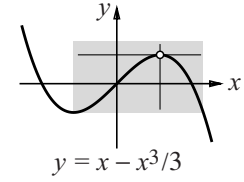
Theorem 7.2 (Second derivative test). *Suppose $y = f(x)$ has a continuous third derivative on an open interval containing a critical point $x = a$; then the critical point is*

- *A* local minimum *of f if $f''(a) > 0$*
- *A* local maximum *of f if $f''(a) < 0$*

If $f''(a) = 0$, the test is inconclusive. □

Degeneracy of a critical point

Thus, a function "looks like" its quadratic approximation near a point where the linear approximation breaks down (i.e., at a critical point), assuming the quadratic approximation does not itself break down. We already have names to distinguish between points where the linear approximation to a function breaks down and where it does not (*critical* and *regular* points, respectively). Morse's lemma suggests we make a similar distinction for critical points. Thus we say a critical point is *degenerate* if the quadratic approximation "breaks down," or "degenerates," in the sense that it fails to determine the local behavior of the function. Otherwise, we say the critical point is *nondegenerate*. For a function of one variable, the situation is clear-cut: a critical point is degenerate if and only if the second derivative vanishes. For functions of more than one variable, there are several second partial derivatives; as we show in the following sections, a critical point may be degenerate even though all of those second derivatives are nonzero. The relation between degeneracy and the second derivatives is more subtle.

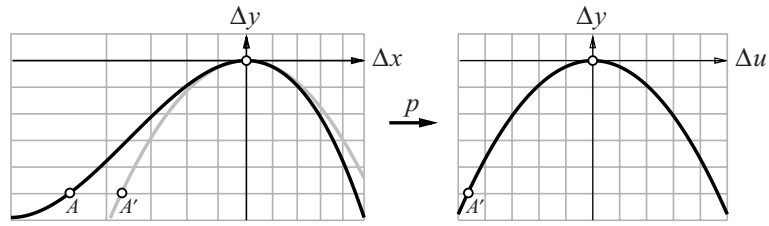


To see how Morse's lemma works, let us apply it to $f(x) = x - x^3/3$ at the critical point $x = 1$. Because $f''(1) = -2$, the point is a local maximum. In terms of window coordinates $(\Delta x, \Delta y)$ centered at $(x, y) = (1, 2/3)$, the formula for f becomes

$$\begin{aligned}\Delta y &= f(1+\Delta x) - f(1)\\ &= 1 + \Delta x - \big(1 + 3\Delta x + 3(\Delta x)^2 + (\Delta x)^3\big)/3 - (1 - 1/3)\\ &= -(\Delta x)^2 - (\Delta x)^3/3 = (\Delta x)^2\big(-1 - \Delta x/3\big).\end{aligned}$$

Thus $h(\Delta x) = -1 - \Delta x/3$, and so $\Delta y = -(\Delta u)^2$ when we set

$$\Delta u = p(\Delta x) = \Delta x\sqrt{1 + \Delta x/3}.$$



The coordinate change p maps the nonuniform grid on the left, above, to the uniform grid on the right, transforming the original cubic curve into a simple parabola. Notice that p pushes points to the left of the origin closer together horizontally; this happens because $0 < p'(\Delta x) < 1$ when $-2 < \Delta x < 0$. To the right of the origin, where $1 < p'(\Delta x)$, points are pushed apart. Finally, because $p'(0) = 1$, the two grids have essentially the same spacing near $\Delta x = 0$. That implies the cubic and the parabola "share ink" near the origin, as the gray copy of the parabola on the left makes clear.

A coordinate change can reverse concavity

The figure also shows that the coordinate change reverses the concavity of part of the graph. For example, at the point A the original cubic is concave up, but at its image A' the parabola is concave down. We associate concavity with the sign of the second derivative, so a coordinate change can reverse the sign of the second derivative. If this were to happen at a critical point (A is not a critical point), the second derivative test would be completely meaningless.

How derivatives depend on coordinates

Let us see how a coordinate change can alter the sign of the second derivative at a noncritical point. Assume $y = f(x)$ is a differentiable function with $f(0) = 0$, and let $x = h(u)$ be a coordinate change with $h(0) = 0$. Then

$$y = f(x) = f(h(u)) = g(u)$$

defines the transformed function g, and we compare $g''(0)$ with $f''(0)$. We have

$$g'(u) = f'(x) \cdot h'(u) \text{ and } g''(u) = f''(x) \cdot \big(h'(u)\big)^2 + f'(x) \cdot h''(u),$$

and our assumptions about the values of h and f at the origin give us

$$g'(0) = f'(0) \cdot h'(0) \text{ and } g''(0) = f''(0) \cdot \big(h'(0)\big)^2 + f'(0) \cdot h''(0).$$

Because h is a coordinate change near the origin, $h'(0) \neq 0$ and the first equation implies

$$g'(0) = 0 \quad \Longleftrightarrow \quad f'(0) = 0.$$

In other words, the origin is a critical point in one coordinate system if and only if it is a critical point in the other. A critical point is thus a geometric property of a function; its presence does not depend upon the coordinates used to descibe the function.

The second derivative at critical and noncritical points

Now suppose the origin is a critical point. Then the equation for $g''(0)$ reduces to

$$g''(0) = f''(0) \cdot \big(h'(0)\big)^2,$$

implying that the second derivatives of f and g have the same sign at the origin, and confirming what is implicit in the second derivative test. If, on the contrary, the origin is not a critical point, then $f'(0) \neq 0$ and the equation for $g''(0)$ now includes the term $f'(0) \cdot h''(0)$. When this additional term is taken into account, $g''(0)$ may well differ in sign from $f''(0)$.

Here is an example to illustrate the "volatility" of the sign of the second derivative under a coordinate change near a regular point. Let

$$y = f(x) = e^x, \;\; x = h(u) = \tfrac{1}{2}\ln(u+1); \;\; \text{then} \;\; y = g(u) = \sqrt{u+1}.$$

The two graphs make the point immediately: the exponential function has a graph that is everywhere concave up but the square root function has a graph that is everywhere concave down. Let us go through the analytic details. First note that the origin is not a critical point, because $f'(0) = 1$ and $g'(0) = \frac{1}{2}$. (The values of the derivatives need not agree, but one cannot be zero unless the other is.) Second, we have $h'(0) = \frac{1}{2}$ and $h''(0) = -\frac{1}{2}$. Finally, for the second derivatives we have $f''(0) = 1$ and

$$g''(0) = f''(0) \cdot \big(h'(0)\big)^2 + f'(0) \cdot h''(0) = 1 \cdot \tfrac{1}{4} + 1 \cdot -\tfrac{1}{2} = -\tfrac{1}{4}.$$

Geometry and local behavior

One of the main objects of this book is to bring to the fore the geometric character of functions and maps. The geometric attributes of a map are the ones left unchanged when the coordinates change. The eigenvalues of a linear map are geometric in this way, and so are its rank and nullity. For a nonlinear function, we tend to concentrate on local behavior, for then we can hope to bring calculus to bear. Thus, critical points are genuine geometric features of a function: if the first derivative equals zero in one coordinate system, it will be zero in every other. The concavity of a function at a critical point is likewise geometric: if the critical point is a minimum in one coordinate system, it will be a minimum in every other.

But the concavity of a function at a noncritical point is not geometric. This does not mean we cannot calculate concavity. We can; it is given by the sign of the second derivative. Concavity is nongeometric because the graphs that represent the same function in two different coordinate systems can have opposite concavities at the same (noncritical) point. An individual representative will have a particular concavity at a point, but other—equally valid—representatives will have the opposite concavity.

Significance of the various derivatives

A Taylor expansion gives us a good way to think about the role and the significance of the various derivatives of a function. Expanding $y = f(x)$ in window coordinates near $x = a$ gives

$$\Delta y = f'(a)\,\Delta x + \tfrac{1}{2} f''(a)\,(\Delta x)^2 + \tfrac{1}{6} f'''(a)\,\Delta x)^3 + \cdots.$$

The first nonzero term dominates. Thus, if $f'(a) \neq 0$, then the local behavior of f near a is entirely determined by $f'(a)$: the inverse function theorem implies there is a coordinate change $\Delta x = h(\Delta u)$ for which

$$\Delta y = f'(a)\,\Delta u.$$

The linear term dominates; all the other terms vanish, implying all the higher derivatives, including the second derivative, have become zero in the new coordinate system.

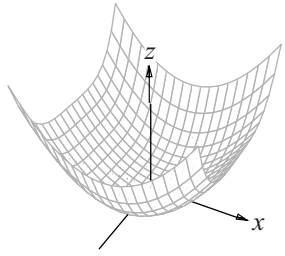
If, by contrast, the linear term is missing, $f'(a) = 0$, then dominance is transferred to the quadratic term. If $f''(a) \neq 0$, that is exactly what happens. Morse's lemma implies there is a coordinate change $\Delta x = k(\Delta v)$ for which

$$\Delta y = \tfrac{1}{2} f''(a)\,(\Delta v)^2.$$

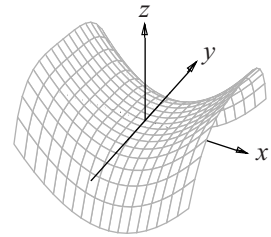
In summary: $f'(a)$ determines the local behavior of f when $f'(a) \neq 0$; $f''(a)$ is geometrically irrelevant. But if $f'(a) = 0$, then $f''(a)$ determines local behavior, at least if $f''(a) \neq 0$. If $f''(a) = 0$, then the cubic term should dominate, and so forth.

7.2 Functions of two variables

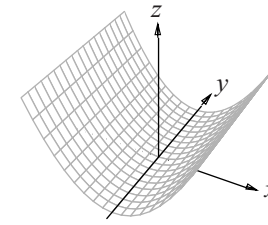
The local behavior of a function of one variable near a critical point is determined by the single quadratic term in the Taylor expansion, if that term is present. However, critical points of a function of two or more variables are more complicated: the local behavior of a function may not be determined by the quadratic terms, even when they are all present. Let us see how this can happen.



$z = f_+(x, y) = x^2 + y^4$



$z = f_-(x, y) = x^2 - y^4$



$z = Q(x, y) = x^2$

Example 1: bowls, saddles, and gutters

Consider first the pair of functions

$$f_+(x,y) = x^2 + y^4 \text{ and } f_-(x,y) = x^2 - y^4.$$

Each has a critical point at the origin, and each function serves as its own Taylor expansion there. In both cases, the quadratic part of the expansion is $Q_f(x,y) = x^2$; the y-variable is absent. If local behavior at a critical point were always determined by the quadratic terms, we would have to conclude that f_+ and f_- have the same local behavior at the origin. But they obviously do not: the graph of f_+ is a bowl, and f_+ has a minimum at the origin. The graph of f_- is a saddle, and f_- has a "minimax" there.

The crucial distinction between f_+ and f_- lies in the way the y-variable appears in their formulas, but Q_f has no y-terms so it cannot "see" that distinction. The missing terms mean that, although $z = Q_f(x,y)$ does have a minimum at the origin, the minimum is nonisolated: all points along the y-axis are minima. (By contrast, the minimum of f_+ is an *isolated* critical point.) As a result, the graph of Q_f is neither a bowl nor a saddle; it has a new shape that we call a "*gutter*." If the bottom of the gutter were to be bent up (e.g., by the addition of $+y^4$), it becomes a bowl; bent down (e.g., by adding $-y^4$), it becomes a saddle.

Example 2: rotate example 1

It appears we can attribute the degeneracy of the critical point of f_+ or f_- to this defect in Q_f. In fact, this is true, but it is not the whole story: we now show that, even if all three quadratic terms are nonzero, those terms may still not determine local behavior. Transform f_+ and f_- by rotating coordinates 45° (dilating by $\sqrt{2}$, to keep the formulas simple). That is, let

$$L : \begin{cases} x = u - v, \\ y = u + v, \end{cases}$$

and let

$$g_+(u,v) = f_+(L(u,v)) = u^2 - 2uv + v^2 + u^4 + 4u^3v + 6u^2v^2 + 4uv^3 + v^4,$$
$$g_-(u,v) = f_-(L(u,v)) = u^2 - 2uv + v^2 - u^4 - 4u^3v - 6u^2v^2 - 4uv^3 - v^4.$$

Each of the new functions still has a critical point at the origin, and each formula still serves as its own Taylor expansion there. There is no qualitative change, either: g_+, like f_+, has a minimum at the origin, and g_-, like f_-, has a saddle. Because the quadratic parts of the new functions are identical,

$$Q_g(u,v) = u^2 - 2uv + v^2,$$

the new Q_g does no better at determining local behavior than the original Q_f did, even though all three quadratic terms are present in Q_g.

Q_g has the same defect as Q_f

The formula for Q_g is different from the formula for Q_f; however, its graph is not, because the rotation–dilation that transforms $f_\pm$ into $g_\pm$ also transforms Q_f into Q_g. The graph of Q_g is just the graph of Q_f rotated 45°, a gutter whose bottom lies along the line $v = u$. Thus, without referring directly to the connection between g_+ and f_+, we can still attribute the degeneracy of the critical point of g_+ at the origin to the fact that the graph of Q_g is a gutter. In geometric terms, Q_g has the same defect as Q_f.

In analytic terms, the defect arises because there is a coordinate change that transforms Q_g into a single square, $z = \pm x^2$, so that the other variable is completely missing. To determine when a critical point is degenerate, we must therefore decide when a general function of the form

$$Q(x,y) = Ax^2 + 2Bxy + Cy^2$$

can be transformed into a single square. To do this it helps to use vector and matrix notation.

Quadratic forms

Definition 7.1 *A **quadratic form** in two variables is a function of the form*

$$Q(x,y) = Ax^2 + 2Bxy + Cy^2 = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} A & B \\ B & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{x}^\dagger M \mathbf{x} = Q(\mathbf{x}).$$

Quadratic forms and matrices

The symmetric matrix M is called the **matrix of the quadratic form**. There is a 1–1 correspondence: $Q \leftrightarrow M$. That is, every symmetric matrix determines a unique quadratic form, and every quadratic form determines a unique symmetric matrix. The symmetry is essential for uniqueness, because, for example,

$$2xy = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

This points up the fact that if we start with any 2×2 matrix A, the formula $Q(\mathbf{x}) = \mathbf{x}^\dagger A \mathbf{x}$ defines a unique quadratic form. However, if we start instead with the form Q, there is only one symmetric matrix M for which $\mathbf{x}^\dagger M \mathbf{x} = Q(\mathbf{x})$. (Thus we write the xy coefficient of Q as $2B$ to simplify splitting it into two equal parts on the "off-diagonal" of M, to make M symmetric.)

Transforming a quadratic form

Suppose L is an invertible 2×2 matrix so $\mathbf{x} = L\mathbf{u}$ is a linear coordinate change. Then, in terms of the new coordinates $\mathbf{u} = (u, v)$, the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x}$ is transformed into

$$\widehat{Q}(\mathbf{u}) = Q(L\mathbf{u}) = \big(L\mathbf{u}\big)^\dagger M \big(L\mathbf{u}\big) = \mathbf{u}^\dagger \big(L^\dagger M L\big)\, \mathbf{u}.$$

Thus $\widehat{Q}$ is also a quadratic form. Furthermore, $L^\dagger ML$ is symmetric (here $L^\dagger$ is the transpose of L) because $\big(L^\dagger ML\big)^\dagger = L^\dagger M^\dagger L^{\dagger\dagger} = L^\dagger ML$; therefore $\widehat{M} = L^\dagger ML$ is the matrix of $\widehat{Q}$. For example, if

$$Q \leftrightarrow \begin{pmatrix} 5 & 3 \\ 3 & -1 \end{pmatrix} \text{ and } L = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}, \text{ then } \widehat{Q} \leftrightarrow \begin{pmatrix} 10 & 14 \\ 14 & 7 \end{pmatrix};$$

that is, L transforms $Q = 5x^2 + 6xy - y^2$ to $\widehat{Q} = 10u^2 + 28uv + 7v^2$; see the exercises. Note that, because L is invertible by definition, $L^\dagger ML$ is invertible if and only if M is. The following theorem identifies the quadratic forms that have the defect we have come to associate with degenerate critical points. Although the theorem is a special case of Theorem 7.10 (see below, p. 244), we give it its own proof.

Theorem 7.3. *Let $Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x}$ be a quadratic form. Suppose a linear coordinate change $\mathbf{x} = L\mathbf{u}$ can be chosen so that the variable u does not appear in the formula*

$$\widehat{Q}(u,v) = \widehat{Q}(\mathbf{u}) = \mathbf{u}^\dagger L^\dagger ML\, \mathbf{u}$$

for the transformed quadratic form. Then the matrix M of Q is noninvertible and conversely.

Proof. Let us first suppose M is noninvertible. Then there is a nonzero vector $\mathbf{r}$ in its kernel: $M\mathbf{r} = \mathbf{0}$. Choose a second vector $\mathbf{s}$ so that $\{\mathbf{r}, \mathbf{s}\}$ form a basis for $\mathbb{R}^2$, and let L be the invertible matrix whose columns are the vectors $\mathbf{r}$ and $\mathbf{s}$. If we write Q as transformed by L in the form

$$\widehat{Q}(u,v) = \mathbf{u}^\dagger L^\dagger ML\,\mathbf{u} = \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix},$$

then the variable u will be missing from this expression if $\alpha = \beta = 0$, that is, if the entries in the first row and the first column of $L^\dagger ML$ equal 0.

To show that $L^\dagger ML$ has this property, first write L and $L^\dagger$ in the form

$$L = \begin{pmatrix} \mathbf{r} & \mathbf{s} \end{pmatrix}, \quad L^\dagger = \begin{pmatrix} \mathbf{r}^\dagger \\ \mathbf{s}^\dagger \end{pmatrix}.$$

(Note that $\mathbf{r}^\dagger$ and $\mathbf{s}^\dagger$ are *row* vectors.) Then matrix multiplication allows us to write ML in a similar way, as

$$ML = \begin{pmatrix} M\mathbf{r} & M\mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & M\mathbf{s} \end{pmatrix}.$$

It follows that

$$L^\dagger ML = \begin{pmatrix} \mathbf{r}^\dagger \\ \mathbf{s}^\dagger \end{pmatrix} \begin{pmatrix} \mathbf{0} & M\mathbf{s} \end{pmatrix} = \begin{pmatrix} \mathbf{r}^\dagger \mathbf{0} & \mathbf{r}^\dagger M\mathbf{s} \\ \mathbf{s}^\dagger \mathbf{0} & \mathbf{s}^\dagger M\mathbf{s} \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{r}^\dagger M\mathbf{s} \\ 0 & \mathbf{s}^\dagger M\mathbf{s} \end{pmatrix}.$$

The entries in the first column of $L^\dagger ML$ are therefore zero, and because the matrix is symmetric, the entries in its first row must be zero as well.

To prove the converse, we suppose that one of the variables in $\mathbf{u} = (u, v)$ is missing from the expression

$$\widehat{Q}(\mathbf{u}) = \mathbf{u}^\dagger L^\dagger ML\,\mathbf{u}.$$

Then $\widehat{M} = L^\dagger ML$ has a row (and a column) of zeros, so $\det \widehat{M} = 0$, implying $\widehat{M}$ is noninvertible. Consequently, $M = (L^\dagger)^{-1}\widehat{M}L^{-1}$ is noninvertible, as well. □

As a result of Theorem 7.3, we find that the natural way to distinguish between quadratic forms is provided by the following definition.

Definition 7.2 *A quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M\mathbf{x}$ is **nondegenerate** if its matrix M is invertible, and is **degenerate** otherwise.*

Corollary 7.4 *The quadratic form $Q(x,y) = Ax^2 + 2Bxy + Cy^2$ is nondegenerate if and only if $AC \neq B^2$.*

Proof. The determinant of the matrix of Q is $AC - B^2$; Q is nondegenerate if and only if this determinant is nonzero. □

The Hessian

To connect these general results about quadratic forms back to the local behavior of a function at a critical point, we introduce the *Hessian*.

Definition 7.3 *Suppose the function $z = f(x,y)$ has continuous second derivatives on a neighborhood of a critical point $(x,y) = (a,b)$. The* ***Hessian of f at (a,b)*** *is the symmetric matrix of second derivatives*

$$H_{(a,b)} = \begin{pmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{yx}(a,b) & f_{yy}(a,b) \end{pmatrix}.$$

The ***Hessian form of f at (a,b)*** *is the quadratic form associated with the Hessian.*

Continuity of the second derivatives guarantees that $H_{(a,b)}$ is symmetric. Because there is usually no chance for confusion, we use the symbol $H_{(a,b)}$ for the Hessian form as well; thus

$$H_{(a,b)}(x,y) = f_{xx}(a,b)x^2 + 2f_{xy}(a,b)xy + f_{yy}(a,b)y^2.$$

Local behavior and the Hessian

Now assume that f has continuous third derivatives near (a,b) so we can write the second-order Taylor expansion of f at (a,b). In terms of window coordinates $\Delta x = x - a$, $\Delta y = y - b$ and $\Delta z = f(a+\Delta x, b+\Delta y) - f(a,b)$ and the Hessian form, the expansion is simply

$$\Delta z = \tfrac{1}{2} H_{(a,b)}(\Delta x, \Delta y) + O(3).$$

This tells us the local behavior of f near (a,b), so we ask: when, and how, does the Hessian determine that local behavior? In other words, when does the quadratic form $H_{(a,b)}(\Delta x, \Delta y)$ dominate the higher-order terms represented by $O(3)$? The answer is provided by Morse's lemma.

Definition 7.4 *Suppose the function $z = f(x,y)$ has continuous second derivatives near the critical point (a,b). Then* ***(a,b)*** *is* ***nondegenerate*** *if the Hessian $H_{(a,b)}$ of f at (a,b) is nondegenerate, and is* ***degenerate*** *otherwise.*

Theorem 7.5 (Morse's lemma). *Suppose $z = f(x,y)$ has continuous third derivatives in a neighborhood of a nondegenerate critical point (a,b). Then, in a sufficiently small window centered at (a,b), there is a coordinate change $(\Delta u, \Delta v) = \mathbf{h}(\Delta x, \Delta y)$ (nonlinear, in general) for which*

$$\Delta z = \pm (\Delta u)^2 \pm (\Delta v)^2.$$

The signs of $(\Delta u)^2$ and $(\Delta v)^2$ are the signs of the eigenvalues of the Hessian $H_{(a,b)}$ of f at (a,b).

Proof. See the proof of the n-variable version, Theorem 7.16 (p. 248), in the next section. □

Analogies

The two eigenvalues of the Hessian are analogous to the single second derivative in the one-variable version (Theorem 7.1, p. 221). The Hessian is symmetric; therefore its eigenvalues are real (Exercise 2.14.a, p. 60). The critical point is nondegenerate; therefore the eigenvalues are nonzero; the sign of each is either positive or negative.

Corollary 7.6 (Second derivative test) *Suppose $z = f(x,y)$ has continuous third derivatives in a neighborhood of a critical point $(x,y) = (a,b)$. Then the nature of the critical point depends on the values of the second partial derivatives of f, (all evaluated at (a,b)), as follows.*

- *A* saddle point *if* $f_{xx}f_{yy} - f_{xy}^2 < 0$
- *A* local minimum *if* $f_{xx}f_{yy} - f_{xy}^2 > 0$ *and* $f_{xx} + f_{yy} > 0$
- *A* local maximum *if* $f_{xx}f_{yy} - f_{xy}^2 > 0$ *and* $f_{xx} + f_{yy} < 0$

If $f_{xx}f_{yy} - f_{xy}^2 = 0$*, the test is inconclusive.*

Proof. According to Morse's lemma, the nature of the critical point is determined by the signs of the eigenvalues, as follows. If the eigenvalues have opposite signs, then $\Delta z = \pm\big((\Delta u)^2 - (\Delta v)^2\big)$, a saddle; if both are positive, then $\Delta z = (\Delta u)^2 + (\Delta v)^2$, a local minimum; if both are negative, then $\Delta z = -(\Delta u)^2 - (\Delta v)^2$, a local maximum; if either is zero, Morse's lemma does not apply.

If λ_1 and λ_2 are the eigenvalues of $H_{(a,b)}$, then

$$\lambda_1\lambda_2 = \det H_{\mathbf{a}} = f_{xx}f_{yy} - f_{xy}^2, \quad \lambda_1 + \lambda_2 = \operatorname{tr} H_{\mathbf{a}} = f_{xx} + f_{yy}.$$

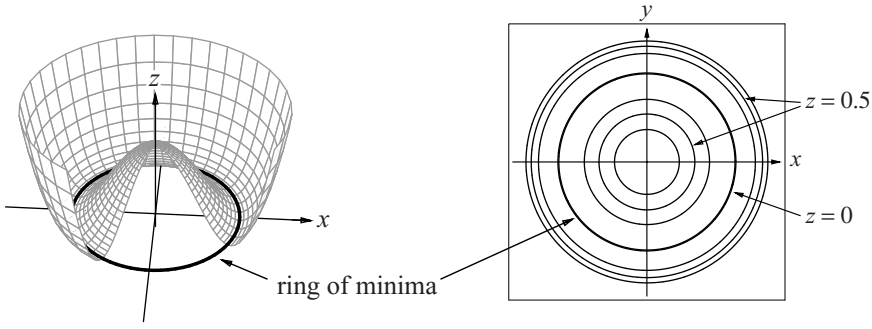
All the assertions of the test now follow, including the final one about an inconclusive result. □

Example: the wine bottle

We now work through the details of a rather rich and varied example to see how Morse's lemma applies. The example begins with the function

$$z = f(x,y) = (x^2 + y^2 - 1)^2.$$

First of all, because x and y appear only in the form $x^2 + y^2$, the graph must be rotationally symmetric around the z-axis. Furthermore, $z \geq 0$ (because z equals a positive square), and z attains its minimum value, $z = 0$, everywhere on the circle $x^2 + y^2 = 1$. (These minima are thus nonisolated critical points; cf. page 225.) If (x,y) is near the origin, but $(x,y) \neq (0,0)$, then $z < 1$. But $z = 1$ when $(x,y) = (0,0)$, so z has a local maximum at the origin. The graph of f therefore resembles the base of a wine bottle. (The sediment that precipitates out of an old wine will settle into the small space along the ring of minima.)



The level curves reflect the circular symmetry; they are all concentric with the origin. Each level $0 < z < 1$ consists of a pair of circles on either side of the level $z = 0$ (the unit circle). Each level above $z = 1$ is a single circle that lies outside the unit circle.

Analyzing the critical points of f

Let us carry out a standard analysis of the critical points of f. We have

$$\frac{\partial f}{\partial x} = 4x(x^2+y^2-1), \quad \frac{\partial f}{\partial y} = 4y(x^2+y^2-1),$$

so $(x,y) = (0,0)$ is a critical point in addition to each of the points where $x^2+y^2 = 1$. To apply the second derivative test, we need the Hessian, which equals

$$H_{(a,b)} = \begin{pmatrix} 4(a^2+b^2-1)+8a^2 & 8ab \\ 8ab & 4(a^2+b^2-1)+8b^2 \end{pmatrix}$$

at an arbitrary point (a,b). At the origin,

$$H_{(0,0)} = \begin{pmatrix} -4 & 0 \\ 0 & -4 \end{pmatrix},$$

so the test succeeds and tells us that the origin is a (nondegenerate) local maximum. At any point on $a^2+b^2 = 1$, however, the Hessian reduces to

$$H_{(a,b)} = \begin{pmatrix} 8a^2 & 8ab \\ 8ab & 8b^2 \end{pmatrix} \text{ but } \det H_{(a,b)} = 64a^2b^2 - 64a^2b^2 = 0,$$

so the test fails. All points on the ring $a^2+b^2 = 1$ of minima are degenerate critical points of f. Consider now what this means for the Hessian form:

$$H_{(a,b)}(\Delta x, \Delta y) = 8a^2(\Delta x)^2 + 16ab\Delta x\Delta y + 8b^2(\Delta y)^2 = 8(a\Delta x + b\Delta y)^2.$$

The Hessian form is degenerate

The Hessian form involves only the square of a single quantity, $a\Delta x + b\Delta y$. Thus, if we introduce the new variables

$$\Delta u = a\Delta x + b\Delta y, \quad \Delta v = -b\Delta x + a\Delta y,$$

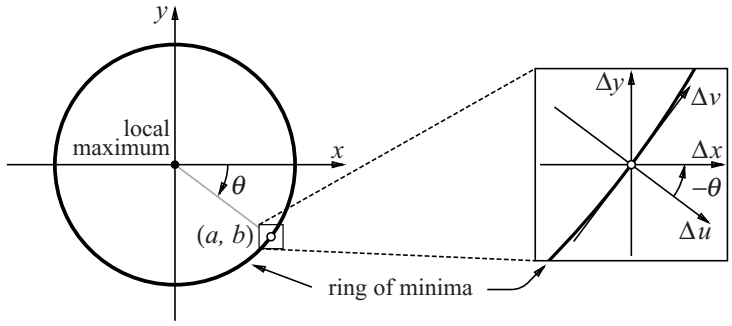
then the Hessian form is just $8(\Delta u)^2$. The variable Δv is missing here, so the Hessian is indeed degenerate in precisely the sense we have been using for quadratic forms. The formulas for Δu and Δv give us new coordinates in the window centered at $(x,y) = (a,b)$; the coordinate change is the linear map defined by the matrix

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

The missing variable in the Hessian

Because $a^2+b^2 = 1$, it follows that P is a pure rotation. Let θ be the angle from the positive x axis to the radial line from the origin to the point (a,b) (so $\theta = \arctan(b/a)$). Then P is rotation by the angle $\arctan(-b/a) = -\arctan(b/a) = -\theta$.

This means that the positive Δx-axis lies at the angle $-\theta$ from the positive Δu-axis; see Exercise 5.16, page 180. Consequently, the Δu-axis points in the same direction as the vector (a,b)—the radial direction—so the Δv-axis is tangent to the ring of minima. Compare this to our previous example: when the Hessian form had no y-component, it had a line of critical points in the direction of the y-axis.

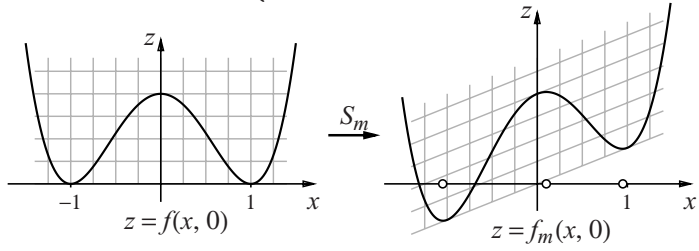


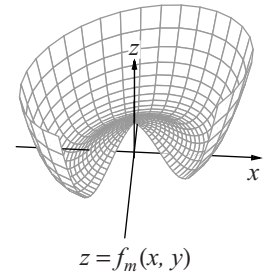
Modify the function by tipping its graph

The purpose of our extended example is to see how Morse's lemma illuminates the structure of a function near a nondegenerate critical point. But the critical points of the ring of minima are degenerate, so Morse's lemma does not apply to them. (Morse's lemma does apply to the isolated maximum at the origin, but the character of that critical point is already evident.) We have more success by first altering the function so its ring of minima "breaks up" into just two isolated critical points. We can do this by tipping the graph slightly, as in the figure in the margin, below. The base, which had been sitting on the entire ring of minima, now shifts to rest on a single point. This point is the absolute minimum of the new function. As we show presently, the opposite point on the ring will shift into a saddle point. There are no other critical points (besides the local maximum that persists near the origin). All this happens no matter how slightly the graph is tipped.

Although it is easier to think of the tipping as a rotation—for example, a rotation of the (x,z)-plane about the y-axis—the formula for the altered function will be simpler if the tipping is done by a *shear*—again, of the (x,z)-plane; see the example in the margin. A vertical shear with slope m is given by

$$S_m : \begin{cases} \text{new}\, x = x, \\ \text{new}\, y = y, \\ \text{new}\, z = mx + z. \end{cases}$$





The shear in the figure uses $m = 0.4$. We see it has the right sort of action on the grid of squares and, at the same time, we see what it does to (a vertical slice of) the graph of f. The formula for the sheared function f_m is

$$z = f_m(x,y) = mx + f(x,y) = (x^2+y^2-1)^2 + mx.$$

Notice that even the grid on the surface of the graph in the margin has been sheared. Also, the shearing has carried part of the graph below the negative x-axis, as we would expect.

The vertical slice shows that the critical points of $z = f_m(x,0)$ (marked by the open dots in the (x,z)-plane on the right) are shifted in relation to those of $z = f(x,0)$: when $m > 0$, the minima move left and the maximum moves right. In Exercise 7.3, you show that the critical points are approximately

$$\text{maximum} : x \approx \frac{m}{4}, \quad \text{minima} : x \approx -\frac{m}{8} \pm 1, \quad \text{when } m \text{ is small.}$$

Critical points of f_m

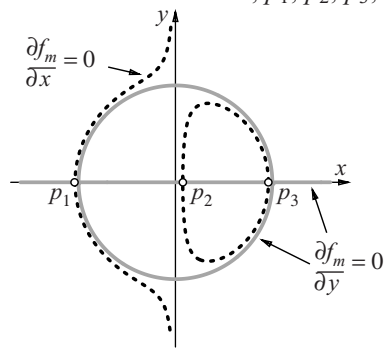
Let us now analyze the critical points of $z = f_m(x,y)$, where m is small but nonzero. We must have

$$\frac{\partial f_m}{\partial x} = 4x(x^2+y^2-1) + m = 0, \quad \frac{\partial f_m}{\partial y} = 4y(x^2+y^2-1) = 0.$$

For $\partial f_m/\partial y = 0$ to hold, either $y = 0$ or $x^2+y^2-1 = 0$. If we assume the second of these equations, then $\partial f_m/\partial x = 0$ reduces to $m = 0$; but this contradicts our assumption that $m \neq 0$. Hence, no point on the ring of minima of the original f is a critical point of the new function f_m. So let us assume instead that $y = 0$. Then $\partial f_m/\partial x = 0$ reduces to

$$4x(x^2-1) + m = 4x^3 - 4x + m = 0.$$

When m is small, this cubic has three real roots, p_1, p_2, p_3; see Exercise 7.3.



The figure above shows an alternate geometric approach to locating the critical points. They appear as the points of intersection of the critical curves on which $\partial f_m/\partial x = 0$ (shown dotted in the figure; $m = 0.3$) and $\partial f_m/\partial y = 0$ (the circle-plus-line shown in gray). The curves intersect in the three points p_1, p_2, p_3 on the x-axis.

The type of each critical point

To determine the type of each critical point, we calculate the Hessian, restricting ourselves to points of the form $(p,0)$:

$$H_{(p,0)} = \begin{pmatrix} 12p^2-4 & 0 \\ 0 & 4p^2-4 \end{pmatrix}.$$

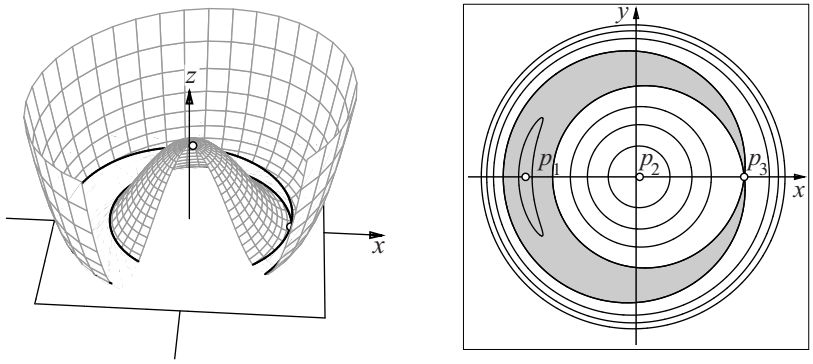
When m is sufficiently small and positive, the three critical points satisfy

$$p_1 < -1, \quad 0 < p_2 < 1/\sqrt{3}, \quad 1/\sqrt{3} < p_3 < 1.$$

This allows us to make the following inferences about their Hessians:

$$H_{(p_1,0)} = \begin{pmatrix} + & 0 \\ 0 & + \end{pmatrix}, \quad H_{(p_2,0)} = \begin{pmatrix} - & 0 \\ 0 & - \end{pmatrix}, \quad H_{(p_3,0)} = \begin{pmatrix} + & 0 \\ 0 & - \end{pmatrix}.$$

It follows that p_1 is a (local) minimum, p_2 a (local) maximum, and p_3 a saddle. (What happens if $m < 0$?)



The crescent-shaped level at the saddle

Think of the figure on the left, above, as showing the graph of f_m filled with liquid up to the level of the saddle point p_3. The level curve of f_m at that level is a thin crescent that has the characteristic "X" shape (albeit elongated and bent) where it passes through the saddle point itself. In the contour plot on the right, the liquid surface is shown in gray. Outside the crescent, the spacing between successive level curves is still $\Delta z = 0.25$, as it was for the original function in the contour plot on page 229. However, at that spacing, no further level curves will be found inside the crescent; the minimum point p_1 lies only about 0.2 units below the saddle. The single curve that is shown inside (in the shaded crescent) is about 0.18 units below the level of the saddle. As $m \to 0$, the shaded crecent shape gets thinner, converging to the ring of minima when $m = 0$, and this contour plot becomes the one on page 229.

Now let us see what Morse's lemma tells us about the saddle point $(p_3,0)$. Fundamentally, it provides new curvilinear coordinates $(\Delta u, \Delta v)$ that will reduce the window equation to $(\Delta u)^2 - (\Delta v)^2$. To understand this, we begin by constructing the window equation at any point $(p,0)$ on the x-axis:

$$\begin{aligned}\Delta z &= f_m(p+\Delta x,\Delta y) - f_m(p,0)\\ &= (4p^3-4p+m)\Delta x + (6p^2-2)(\Delta x)^2 + (2p^2-2)(\Delta y)^2\\ &\quad + 4p\,(\Delta x)^3 + 4p\,\Delta x\,(\Delta y)^2 + (\Delta x)^4 + 2(\Delta x)^2(\Delta y)^2 + (\Delta y)^4.\end{aligned}$$

Completing the square

At a critical point, $4p^3-4p+m=0$, so Δz loses its linear term (as we expect). If the window equation were purely quadratic, of the form

$$\Delta z = A\,(\Delta x)^2 + 2B\,\Delta x\,\Delta y + C\,(\Delta y)^2,$$

with A, B, C constants, then we could make Δz a sum of two squares by the familiar process of completing the square (assuming $A \neq 0$):

$$\begin{aligned}\Delta z &= A\left((\Delta x)^2 + 2\frac{B}{A}\Delta x\,\Delta y + \frac{B^2}{A^2}(\Delta y)^2\right) - \frac{B^2}{A}(\Delta y)^2 + C(\Delta y)^2\\ &= A\left(\Delta x + \frac{B}{A}\Delta y\right)^2 - \left(\frac{B^2}{A} - C\right)(\Delta y)^2.\end{aligned}$$

To finish, let us suppose $A > 0$; this makes the first square positive and the second negative (we expect the squares to have different signs at a saddle). The coordinate change

$$\mathbf{h}: \begin{cases} \Delta u = \sqrt{A}\,\Delta x + \dfrac{B}{\sqrt{A}}\,\Delta y, \\ \Delta v = \Delta y\sqrt{\dfrac{B^2}{A} - C}, \end{cases}$$

then gives $\Delta z = (\Delta u)^2 - (\Delta v)^2$, a simple sum of (positive and negative) squares.

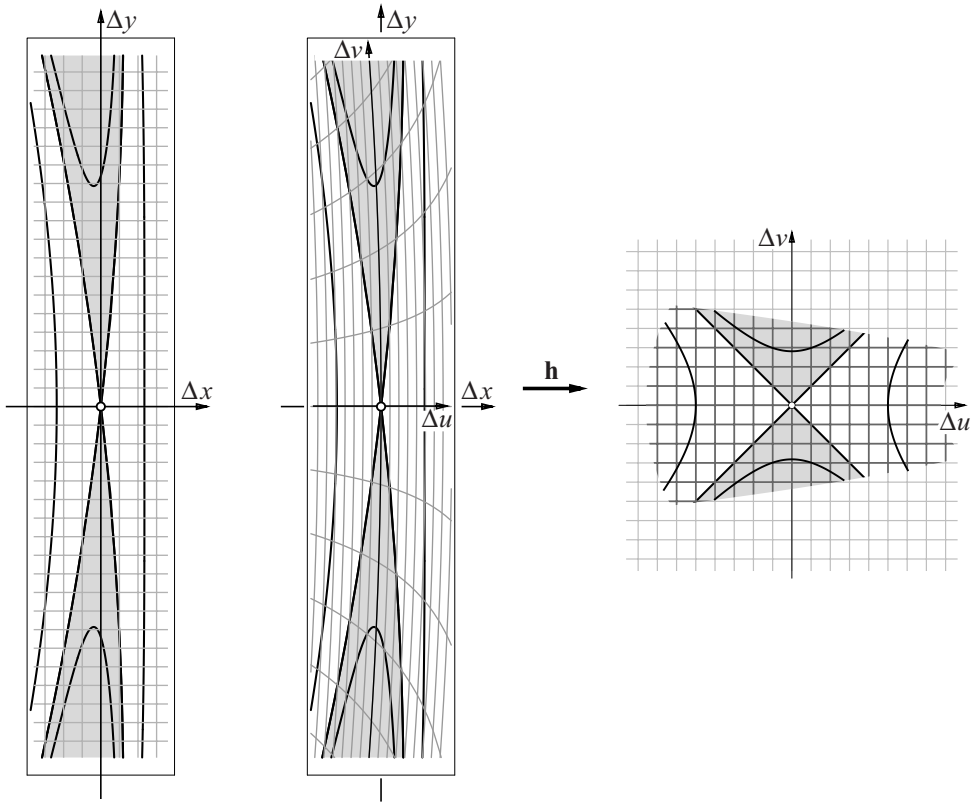
Morse's observations

However, because the given Δz is not purely quadratic, this approach seems futile. But now Morse makes two crucial observations:

- The validity of the change of coordinates **h** does not depend on the coefficients A, B, and C being constants; a quadratic form with variable coefficients can work, too.
- The window equation at any critical point can be "disassembled" properly into a quadratic form with variable coefficients.

He then provides (remarkably simple) instructions for disassembling the window equation into the proper components. We define equivalent instructions below (Lemma 7.3 p. 249), and they give us the following (see p. 251).

$$\begin{aligned}A &= 6p^2 - 2 + 4p\,\Delta x + (\Delta x)^2 + \tfrac{1}{3}(\Delta y)^2,\\ B &= \tfrac{4}{3}p\,\Delta y + \tfrac{2}{3}\Delta x\,\Delta y,\\ C &= 2p^2 - 2 + \tfrac{4}{3}p\,\Delta x + \tfrac{1}{3}(\Delta x)^2 + (\Delta y)^2.\end{aligned}$$



"Squaring up" contours near the saddle point

We see above the action of the map $\mathbf{h} : (\Delta x, \Delta y) \mapsto (\Delta u, \Delta v)$, using the expressions for A, B, and C just given, with $p = p_3 = 0.9872574766623532$. On the left is the window with its "native" $(\Delta x, \Delta y)$-coordinates. In the middle is the same $(\Delta x, \Delta y)$-window but now overlaid with the curvilinear coordinates $(\Delta u, \Delta v)$ pulled back by $\mathbf{h}$. On the right is the (curved) image of the window as pushed forward by $\mathbf{h}$ to the $(\Delta u, \Delta v)$-plane. The windows are very small; the spacing in both coordinate grids is 0.005. It is clear that $\mathbf{h}$ "squares up" the contours: the zero-level $\Delta z = 0$ becomes the pair of perpendicular straight lines $\Delta v = \pm \Delta u$. Notice that the zero-level intersects the $(\Delta u, \Delta v)$-grid lines in exactly the same places in both windows. The other two contours are not equally spaced with $\Delta z = 0$ but are instead chosen at levels (namely, $\Delta z = -0.0002$ and $\Delta z = 0.0006$) that show up well in the original (thin) window.

h is invertible

It remains to verify that $\mathbf{h}$ is indeed a valid coordinate change—that is, an invertible map—on some neighborhood of $(\Delta x, \Delta y) = (0, 0)$. The functions that appear in $\mathbf{h}$ are smooth where they are defined; thus the inverse function theorem says it is sufficient to show that the derivative $\mathrm{d}\mathbf{h}_{(0,0)}$ is invertible. This follows (cf. Exercise 7.4) from

$$\mathrm{d}\mathbf{h}_{(0,0)} = \begin{pmatrix} \sqrt{6p^2 - 2} & 0 \\ 0 & \sqrt{2 - 2p^2} \end{pmatrix} \approx \begin{pmatrix} 1.96165 & 0 \\ 0 & 0.225045 \end{pmatrix}.$$

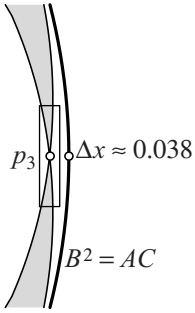
The axes are eigendirections of the derivative $d\mathbf{h}_{(0,0)}$; consequently, the image of each axis under $\mathbf{h}$ itself is tangent to the corresponding axis in the target. Moreover, $\mathbf{h}$ approximately doubles horizontal distances but compresses vertical distances to less than a quarter of their original length. The figure above shows all this quite clearly.

The domain of invertibility

Morse's lemma guarantees that there are curvilinear coordinates on some open set around the critical point on which the function appears as a sum of squares. But how large is that open set? It is the set on which the coordinate change map $\mathbf{h}$ is invertible. In this case, we can expect the invertibility to break down when the form

$$\Delta z = A\left(\Delta x + \frac{B}{A}\Delta y\right)^2 - \left(\frac{B^2}{A} - C\right)(\Delta y)^2$$

becomes degenerate. This will happen if either coefficient vanishes. Here the crucial coefficient is the second one. The figure in the margin shows that the curve $B^2 = AC$ contains points very close to $p_3 : (\Delta x, \Delta y) = (0,0)$. Thus, only by keeping the window at p_3 relatively narrow was it possible to avoid that curve and thus avoid losing the invertibility of $\mathbf{h}$.



Curvilinear coordinates near the minimum

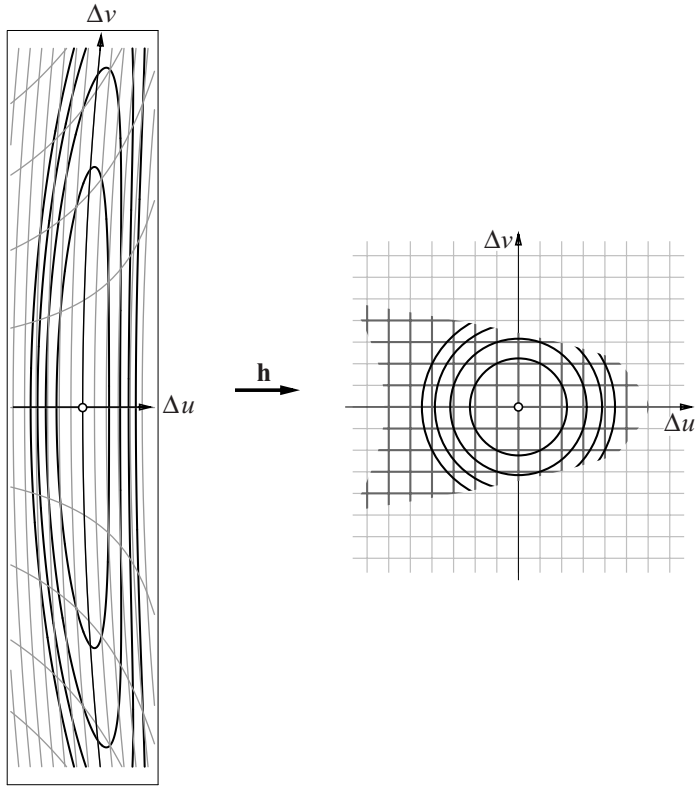
We can obtain curvilinear coordinates that "square up" the contours of f_m around its minimum point $(x,y) = (p_1, 0)$ using essentially the same coordinate transformation $\mathbf{h}$. Apart from the obvious change from $p = p_3$ to $p = p_1$, just one pair of modifications is needed. First, because the critical point is now a minimum, the window equation must be rewritten as a sum of positive squares,

$$\Delta z = A\left(\Delta x + \frac{B}{A}\Delta y\right)^2 + \left(C - \frac{B^2}{A}\right)(\Delta y)^2.$$

Note the change in the form of the coefficient of $(\Delta y)^2$; this forces a corresponding alteration in the formula for Δv:

$$\Delta v = \Delta y \sqrt{C - \frac{B^2}{A}}.$$

The result is shown below. The "native" window, on the left, has the same proportions as the one we used for the saddle point, but it is half again as large. The source and the target of $\mathbf{h}$ are drawn to the same scale (and a grid square in the $(\Delta u, \Delta v)$-plane is 0.01 units on a side), making it evident that $\mathbf{h}$ stretches the horizontal direction but compresses the vertical. What the grids actually show us are the effects of the pullback by $\mathbf{h}^{-1}$: horizontal compression and vertical elongation. The contours of f are equally spaced, at the levels $\Delta z = 0.0005, 0.0010, 0.0015, 0.0020$. Notice that each contour meets points of the $(\Delta u, \Delta v)$-grid in exactly the same places in both windows.



Invertiblility of **h**

Details for the derivative $d\mathbf{h}_{(0,0)}$ are very similar to those for the saddle point; note that the slight change in the definition of Δv has caused $\sqrt{2-2p^2}$ to be replaced by $\sqrt{2p^2-2}$.

$$d\mathbf{h}_{(0,0)} = \begin{pmatrix} \sqrt{6p^2-2} & 0 \\ 0 & \sqrt{2p^2-2} \end{pmatrix} \approx \begin{pmatrix} 2.0367 & 0 \\ 0 & 0.2222 \end{pmatrix}.$$

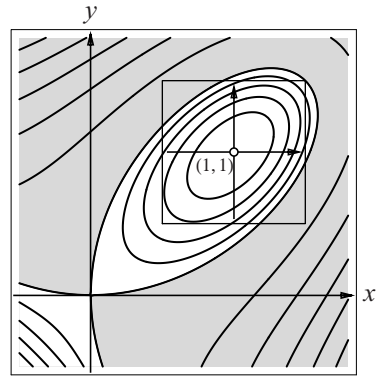
Because $d\mathbf{h}_{(0,0)}$ is once again a diagonal matrix, the image of each axis under **h** is tangent to its corresponding axis in the target. Horizontal lengths are approximately doubled and vertical ones are compressed by the factor 2/9. We conclude that **h** is locally invertible, giving valid curvilinear coordinates $(\Delta u, \Delta v)$ in some suitably restricted window centered at the minimum point $(x,y) = (p_1, 0)$.

Example: the folium of Descartes

We now consider briefly a second function that is simpler than the wine bottle but nevertheless illustrates new aspects of Morse's lemma. The function is one introduced by Descartes:

$$z = f(x,y) = x^3 + y^3 - 3xy.$$

Some of its level curves near the origin are shown in the figure below. In the shaded region, where the function takes positive values, the contour interval is $\Delta z = 1.5$; in the unshaded region, we have used a smaller interval: $\Delta z = 0.2$. The zero-level curve that separates the two regions includes a leaf-shaped loop that has led to the curve being called the *folium* ("leaf") of Descartes. We use the same name to refer to the function itself.



The level curves make it clear that $z = f(x,y)$ has a saddle at the origin and a local minimum inside the "leaf," and a quick calculation shows that the minimum is at $(x,y) = (1,1)$. In terms of window coordinates $\Delta x = x - 1$, $\Delta y = y - 1$ centered at the minimum, the formula for f becomes

$$\begin{aligned} z &= (1+\Delta x)^3 + (1+\Delta y)^3 - 3(1+\Delta x)(1+\Delta y) \\ &= 1 + 3\Delta x + 3(\Delta x)^2 + (\Delta x)^3 + 1 + 3\Delta y + 3(\Delta y)^2 + (\Delta y)^3 \\ &\quad - 3 - 3\Delta x - 3\Delta y - 3\Delta x\,\Delta y \\ &= -1 + 3(\Delta x)^2 - 3\Delta x\,\Delta y + 3(\Delta y)^2 + (\Delta x)^3 + (\Delta y)^3. \end{aligned}$$

This reduces to

$$\Delta z = A\,(\Delta x)^2 + 2B\,\Delta x\,\Delta y + C\,(\Delta y)^2,$$

with $\Delta z = z + 1$ and

$$A = 3 + \Delta x, \quad B = -3/2, \quad C = 3 + \Delta y,$$

Action of the coordinate change **h**

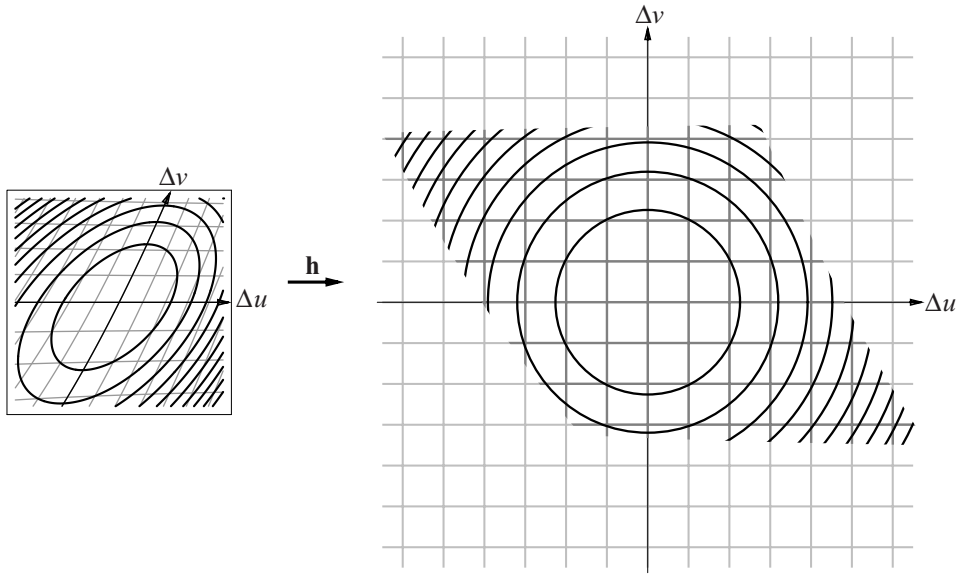
The standard coordinate change

$$\mathbf{h}: \begin{cases} \Delta u = \sqrt{A}\left(\Delta x + \dfrac{B}{A}\Delta y\right), \\ \Delta v = \Delta y \sqrt{C - \dfrac{B^2}{A}}, \end{cases}$$

in the window then transforms Δz into

$$\Delta z = (\Delta u)^2 + (\Delta v)^2.$$

The contours in the original $(\Delta x, \Delta y)$-window are roughly elliptical. The map **h** carries them to concentric circles in the target $(\Delta u, \Delta v)$-plane. The figure below helps us to follow the details. The curvilinear $(\Delta u, \Delta v)$ coordinates that are overlaid on the source on the left are the ones pulled back from the target by **h**. Therefore, the intersections between the original contours and the curvilinear grid in the source match exactly the intersections between the image circles and the square grid in the target.



Action of **h**

At this scale (the source window is a unit square), the contours are close to ellipses, and **h** looks almost linear. Its linear approximation at the origin is

$$\mathrm{d}\mathbf{h}_{(0,0)} = \begin{pmatrix} \sqrt{3} & -\sqrt{3}/2 \\ 0 & 3/2 \end{pmatrix}.$$

The map resembles a horizontal shear that pushes points that lie above the horizontal axis to the left and points below to the right. Horizontal distances are increased by a factor of about $\sqrt{3} \approx 1.7$, and vertical ones by a factor of about 1.5. The effect of the dilation is to make the ellipses both larger and somewhat wider; the effect of the shear is then to turn them into circles.

Comparing the folium and the wine bottle

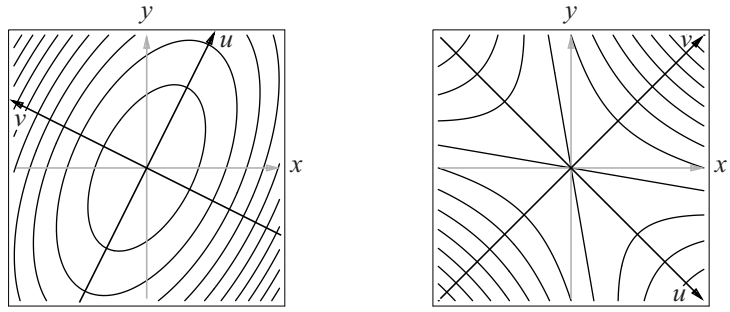
Notice that the Δu- and Δv-axes we see overlaid on the source do not line up with the major and minor axes of the nested ellipses. This points to the main difference between the folium and the wine bottle examples. At the minimum of the tipped wine bottle, the curvilinear coordinate axes were aligned with the principal axes of the (approximate) ellipses, so the coordinate change **h** had a simpler action there: to turn the ellipses into circles, it just stretched the ellipses by two different factors along their principal axes. As a consequence, the derivative $\mathrm{d}\mathbf{h}_{(0,0)}$ was a diagonal matrix, representing a pure strain in the coordinate directions.

At the minimum point of the folium, however, the derivative is not a diagonal matrix. Appearances to the contrary notwithstanding, it is a pure strain, though, (rather than the shear it appears to be) because its eigenvalues, $\sqrt{3}$ and $3/2$, are real and unequal (cf. Theorem 2.6, p. 40). Hence we can convert $d\mathbf{h}_{(0,0)}$ into a diagonal matrix by using a further coordinate change that will align the new coordinate axes with the strain directions, that is, with the principal axes of the ellipses. In fact, if we restrict ourselves to an ordinary quadratic form with constant coefficients, we can show that the additional coordinate change can be taken as a rotation (that aligns the coordinate axes with the symmetry axes of the level curves).

Quadratic forms under rotations

Here are two examples of typical quadratic forms with their level curves. For each, we provide a rotation that transforms the form into a sum of squares, allowing us to infer analytically the shape of its level curves.

$$Q_{\text{ell}} = 6x^2 - 4xy + 3y^2, \qquad Q_{\text{hyp}} = x^2 + 6xy + y^2.$$



Under the respective coordinate changes

$$\mathbf{h}_{\text{ell}}: \begin{cases} x = \dfrac{u - 2v}{\sqrt{5}}, \\ y = \dfrac{2u + v}{\sqrt{5}}, \end{cases} \qquad \mathbf{h}_{\text{hyp}}: \begin{cases} x = \dfrac{u + v}{\sqrt{2}}, \\ y = \dfrac{-u + v}{\sqrt{2}}, \end{cases}$$

the two quadratic forms pull back to

$$Q^*_{\text{ell}} = 2u^2 + 7v^2, \qquad Q^*_{\text{hyp}} = -2u^2 + 4v^2.$$

The map $\mathbf{h}_{\text{ell}}$ is rotation by $\theta = \arctan 2$, and $\mathbf{h}_{\text{hyp}}$ is rotation by $\theta = -45°$. The rotations cause the (u, v)-coordinates to line up with what appear to be the symmetry axes of the level curves. We say that the quadratic forms have been *transformed to principal axes*.

The ellipses

The equation $Q^*_{\text{ell}} = r$ $(r > 0)$ describes an ellipse whose principal axes are the coordinate axes. All the different ellipses (i.e., for different $r > 0$) have the same proportions; that is, they are similar figures in the sense of Euclidean geometry. Because rotation preserves lengths and angles, we conclude that the level curves of Q_{ell} are nested similar ellipses that share their principal axes.

The hyperbolas

Likewise, $Q^*_{\text{hyp}} = r$ describes a hyperbola whose principal axes are the coordinate axes. All the different hyperbolas have the same asymptotes; these are the

straight lines ("degenerate hyperbolas") defined by $Q^*_{\text{hyp}} = Q_{\text{hyp}} = 0$. In the (u,v)-coordinates, the asymptotes have the equations $v = \pm u/\sqrt{2}$; in the original (x,y)-coordinates, we can get the equations either by substitution using $\mathbf{h}_{\text{hyp}}$ or by completing the square:

$$-8x^2 + (3x+y)^2 = 0 \text{ or } y = (-3 \pm \sqrt{8})x.$$

Because rotation preserves lengths and angles, we conclude that the level curves of Q_{hyp} are hyperbolas that share asymptotes and principal axes.

The coefficients are eigenvalues

Not only do the signs of the coefficients in the formulas for Q^*_{ell} and Q^*_{hyp} have geometric meaning, their ratio does, too: it determines the "aspect ratio" of the level curves. For example, in the first figure, seven ellipses cross the v-axis in the same distance that just two cross the u-axis. In the second figure, six hyperbolas cross the v-axis in the distance that three cross the u-axis. Call the ratio of the numbers in each pair the *aspect ratio* of the curves. This ratio is the same as (the absolute value of) the ratio of the eigenvalues of the symmetric matrices that define the forms:

$$M_{\text{ell}} = \begin{pmatrix} 6 & -2 \\ -2 & 3 \end{pmatrix}, \qquad M_{\text{hyp}} = \begin{pmatrix} 1 & 3 \\ 3 & 1 \end{pmatrix},$$

$$\begin{aligned} p_{\text{ell}}(\lambda) &= \lambda^2 - 9\lambda + 14 \\ &= (\lambda - 2)(\lambda - 7), \end{aligned} \qquad \begin{aligned} p_{\text{hyp}}(\lambda) &= \lambda^2 - 2\lambda - 8 \\ &= (\lambda + 2)(\lambda - 4). \end{aligned}$$

Thus, the ratio of the eigenvalues determines the geometry of the level curves: the sign of the ratio indicates the kind of curves ($+$ for ellipses, $-$ for hyperbolas) and its magnitude indicates their aspect ratio.

As we have seen (cf. p. 226), when a linear map $\mathbf{x} = L\mathbf{u}$ is used to change coordinates in the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M\,\mathbf{x}$ defined by a symmetric 2×2 matrix M, the matrix of the transformed quadratic form is $M^* = L^\dagger M L$:

$$Q^*(\mathbf{u}) = Q(L\mathbf{u}) = (L\mathbf{u})^\dagger M\,(L\mathbf{u}) = \mathbf{u}^\dagger L^\dagger M L\,\mathbf{u} = \mathbf{u}^\dagger M^* \mathbf{u}.$$

But the rotation matrices we are now using for coordinate changes have a special property: the transpose of a rotation is its inverse:

$$R_\theta^{-1} = R_{-\theta} = R_\theta^\dagger.$$

Therefore, when a rotation R is used to transform a quadratic form, we can write the relation between the two matrices defining the forms in a new way: $M^* = R^{-1} M R$. In particular, if the transformed Q^* is a sum of squares, then its matrix M^* is a diagonal matrix, and we have the following result.

Theorem 7.7. *If the rotation $\mathbf{x} = R\mathbf{u}$ transforms the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M\,\mathbf{x}$ into $Q^*(\mathbf{u}) = \mathbf{u}^\dagger D\,\mathbf{u}$, where D is a diagonal matrix, then the diagonal elements of D are the eigenvalues of M and the columns of R are corresponding eigenvectors.*

Proof. Let the diagonal elements of D be α_1 and α_2, and let $\mathbf{e}_1 = (1,0)^\dagger$ and $\mathbf{e}_2 = (0,1)^\dagger$ be the standard basis vectors in $\mathbb{R}^2$. Then

$$D\mathbf{e}_1 = \alpha_1\mathbf{e}_1, \quad D\mathbf{e}_2 = \alpha_2\mathbf{e}_2,$$

so α_i is an eigenvalue of D with eigenvector $\mathbf{e}_i$, $i = 1,2$. By assumption, $D = R^\dagger MR = R^{-1}MR$, or $RD = MR$. Let $\mathbf{v}_i = R\mathbf{e}_i$; this is the i-th column of R. We find

$$\alpha_i\mathbf{v}_i = R(\alpha_i\mathbf{e}_i) = RD\mathbf{e}_i = MR\mathbf{e}_i = M\mathbf{v}_i, \quad i = 1,2,$$

implying that α_i is an eigenvalue of M with eigenvector $\mathbf{v}_i$. □

Transforming to principal axes

Using Theorem 7.7 as a guide, we now have a way to transform a quadratic form to principal axes, that is, a way to construct a rotation $\mathbf{x} = R\mathbf{u}$ that will align the $\mathbf{u}$-coordinate axes with the principal axes of the curves $Q(\mathbf{x}) =$ constant and reduce the form to a sum of squares.

Theorem 7.8 (Principal axes theorem). *For any quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M\mathbf{x}$, there is a rotation $\mathbf{x} = R\mathbf{u}$ that transforms Q into a sum of squares $Q^*(\mathbf{u}) = \lambda_1 u^2 + \lambda_2 v^2$, where λ_1 and λ_2 are the eigenvalues of M.*

Proof. We use a proof that extends naturally to quadratic forms in n variables. We know M has a real eigenvalue λ_1 with an eigenvector $\mathbf{v}$ that we can assume to be a unit vector. Let $\mathbf{w}$ be a unit vector orthogonal to $\mathbf{v}$, chosen so the square $\mathbf{v} \wedge \mathbf{w}$ has positive orientation (cf. p. 41). Let R be the matrix whose columns are $\mathbf{v}$ and $\mathbf{w}$, in that order:

$$R = \begin{pmatrix}\mathbf{v} & \mathbf{w}\end{pmatrix}, \quad R^\dagger = \begin{pmatrix}\mathbf{v}^\dagger \\ \mathbf{w}^\dagger\end{pmatrix}, \quad R^\dagger R = \begin{pmatrix}\mathbf{v}^\dagger\mathbf{v} & \mathbf{v}^\dagger\mathbf{w} \\ \mathbf{w}^\dagger\mathbf{v} & \mathbf{w}^\dagger\mathbf{w}\end{pmatrix} = \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix},$$

so $R^\dagger = R^{-1}$. Then $MR = \begin{pmatrix}M\mathbf{v} & M\mathbf{w}\end{pmatrix} = \begin{pmatrix}\lambda_1\mathbf{v} & M\mathbf{w}\end{pmatrix}$, and

$$R^{-1}MR = R^\dagger MR = \begin{pmatrix}\lambda_1\mathbf{v}^\dagger\mathbf{v} & \mathbf{v}^\dagger M\mathbf{w} \\ \lambda_1\mathbf{w}^\dagger\mathbf{v} & \mathbf{w}^\dagger M\mathbf{w}\end{pmatrix} = \begin{pmatrix}\lambda_1 & 0 \\ 0 & \beta\end{pmatrix} = D,$$

where $\beta = \mathbf{w}^\dagger M\mathbf{w}$. The lower-left term of D is zero because $\mathbf{v}$ and $\mathbf{w}$ are orthogonal; the upper-right term is zero because $D = R^\dagger MR$ is symmetric. The proof of Theorem 7.7 shows that $\beta = \lambda_2$, the second eigenvalue of M, and that $\mathbf{w}$ is a corresponding eigenvector.

Finally, because $\mathbf{v}$ lies on the unit circle and $\mathbf{w}$ lies 90° counterclockwise from it,

$$\mathbf{v} = \begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix}-\sin\theta \\ \cos\theta\end{pmatrix},$$

for some $0 \le \theta < 2\pi$. Thus $R = R_\theta$. □

Corollary 7.9 *Level curves of the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\dagger M\mathbf{x}$ are ellipses if $\det M > 0$ and are hyperbolas if $\det M < 0$.*

Proof. Transforming Q to principal axes gives $Q^*(\mathbf{u}) = \lambda_1 u^2 + \lambda_2 v^2$, where $\det M = \lambda_1\lambda_2$. The level curves are ellipses when this product is positive and hyperbolas when it is negative. □

7.3 Morse's lemma

In this section we show that the local behavior of a function $z = f(x_1,\dots,x_n)$ at a nondegenerate critical point is determined by its Hessian matrix of second derivatives at the critical point. The key step is Morse's lemma, which provides a coordinate change that reduces the function to a pure sum of squares near the critical point. We begin by transferring to n dimensions all the terms and concepts introduced in the previous section.

Quadratic forms

Definition 7.5 *A **quadratic form** in n variables is a function of the form*

$$Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x},$$

where $\mathbf{x} = (x_1,\dots,x_n)$ *(treated as a column vector) and M is an* $n \times n$ *matrix.*

Antisymmetric matrices

Note that the matrix M need not be symmetric, nor is it uniquely defined by Q; see the exercises. In fact, adding a antisymmetric matrix R to M does not alter Q, because R by itself defines the quadratic form that is identically zero. The next lemma is the converse; it says that only the antisymmetric matrices define the zero form. (An antisymmetric matrix is also said to be *skew-symmetric*.)

Lemma 7.1. *Suppose the quadratic form* $Q_0(\mathbf{x}) = \mathbf{x}^\dagger R \mathbf{x}$ *is identically zero; then the matrix* $R = (r_{ij})$ *is* antisymmetric*; that is,* $r_{ji} = -r_{ij}$ *for every* $i, j = 1,\dots,n$.

Proof. We evaluate $Q_0(\mathbf{x})$ for particular vectors $\mathbf{x}$. First take $\mathbf{x} = \mathbf{e}_i$, the ith standard basis vector in $\mathbb{R}^n$. Then $0 = Q_0(\mathbf{e}_i) = r_{ii}$. Next, take $\mathbf{x} = \mathbf{e}_i + \mathbf{e}_j$, $i \neq j$; then $0 = Q_0(\mathbf{e}_i + \mathbf{e}_j) = r_{ij} + r_{ji}$. □

Symmetric matrices

According to the next lemma, with each quadratic form Q we can associate a unique symmetric matrix M that defines the form: $Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x}$. We write $Q \leftrightarrow M$ to indicate this association.

Lemma 7.2. *Suppose* $Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x}$ *is a quadratic form, where M is an arbitrary* $n \times n$ *matrix. Then* $\widetilde{M} = (M + M^\dagger)/2$ *is symmetric and defines the same quadratic form. Moreover, if S is symmetric and* $Q(\mathbf{x}) = \mathbf{x}^\dagger S \mathbf{x}$*, then* $S = \widetilde{M}$.

Proof. Let $Q^\dagger(\mathbf{x}) = \mathbf{x}^\dagger M^\dagger \mathbf{x}$ be the quadratic form defined by the transpose matrix $M^\dagger$. Because $Q^\dagger(\mathbf{x})$ is just a scalar (a 1×1 matrix), it is equal to its own transpose; thus

$$Q^\dagger(\mathbf{x}) = (\mathbf{x}^\dagger M^\dagger \mathbf{x})^\dagger = \mathbf{x}^\dagger M \mathbf{x} = Q(\mathbf{x}).$$

In other words, even when M and $M^\dagger$ are different, they define the same quadratic form. Now let $\widetilde{Q}$ be the quadratic form defined by $\widetilde{M}$. Then

$$\widetilde{Q}(\mathbf{x}) = \mathbf{x}^\dagger \tfrac{1}{2}(M + M^\dagger)\,\mathbf{x} = \tfrac{1}{2}\left(\mathbf{x}^\dagger M\mathbf{x} + \mathbf{x}^\dagger M^\dagger \mathbf{x}\right) = \tfrac{1}{2}\left(Q(\mathbf{x}) + Q^\dagger(\mathbf{x})\right) = Q(\mathbf{x}).$$

If $Q(\mathbf{x}) = \mathbf{x}^\dagger S\mathbf{x}$; then the quadratic form Q_0 defined by the symmetric matrix $R = S - \widetilde{M}$ must be identically zero: $Q_0(\mathbf{x}) = \mathbf{x}^\dagger R\,\mathbf{x} \equiv 0$. By the previous lemma, R is also antisymmetric, so R must be the zero matrix, implying that $S = \widetilde{M}$. □

We now single out the degenerate quadratic forms as the ones that are either missing a variable or can be so transformed by a suitable linear coordinate change. We show, as we did in the two-variable case, that a form is degenerate in this sense precisely when its associated matrix is noninvertible.

Theorem 7.10. *Let $Q(\mathbf{x}) = \mathbf{x}^\dagger M\mathbf{x}$ be a quadratic form in n variables, where M is the symmetric matrix associated with Q. Suppose a linear coordinate change $\mathbf{x} = L\mathbf{u}$ can be chosen so that the variable u_1 does not appear in the formula*

$$\widehat{Q}(u_1, \ldots, u_n) = \widehat{Q}(\mathbf{u}) = Q(L\mathbf{u}) = \mathbf{u}^\dagger L^\dagger ML\,\mathbf{u}$$

for the transformed quadratic form. Then M is noninvertible, and conversely.

Proof. Let us first suppose M is noninvertible. Then there is a nonzero vector $\mathbf{r}$ in its kernel: $M\mathbf{r} = \mathbf{0}$. Choose additional vectors $\mathbf{s}_2, \ldots, \mathbf{s}_n$ so that the n vectors $\{\mathbf{r}, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ form a basis for $\mathbb{R}^n$, and let L be the invertible matrix whose columns are the vectors $\mathbf{r}, \mathbf{s}_2, \ldots, \mathbf{s}_n$, in that order. The variable u_1 will be missing from

$$\widehat{Q}(\mathbf{u}) = \mathbf{u}^\dagger L^\dagger ML\,\mathbf{u}$$

if all the entries in the first row and the first column of the matrix $L^\dagger ML$ are zero.

To show that $L^\dagger ML$ has this property, first write L and $L^\dagger$ in the form

$$L = \begin{pmatrix} \mathbf{r} & \mathbf{s}_2 & \cdots & \mathbf{s}_n \end{pmatrix} \text{ and } L^\dagger = \begin{pmatrix} \mathbf{r}^\dagger \\ \mathbf{s}_2^\dagger \\ \vdots \\ \mathbf{s}_n^\dagger \end{pmatrix}.$$

Then matrix multiplication allows us to write the $n \times n$ matrix ML is a similar way, as

$$ML = \begin{pmatrix} M\mathbf{r} & M\mathbf{s}_2 & \cdots & M\mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \mathbf{0} & M\mathbf{s}_2 & \cdots & M\mathbf{s}_n \end{pmatrix}.$$

In that case,

$$L^\dagger ML = \begin{pmatrix} \mathbf{r}^\dagger \\ \mathbf{s}_2^\dagger \\ \vdots \\ \mathbf{s}_n^\dagger \end{pmatrix} \begin{pmatrix} \mathbf{0} & M\mathbf{s}_2 & \cdots & M\mathbf{s}_n \end{pmatrix} = \begin{pmatrix} \mathbf{r}^\dagger\mathbf{0} & \mathbf{r}^\dagger M\mathbf{s}_2 & \cdots & \mathbf{r}^\dagger M\mathbf{s}_n \\ \mathbf{s}_2^\dagger\mathbf{0} & \mathbf{s}_2^\dagger M\mathbf{s}_2 & \cdots & \mathbf{s}_2^\dagger M\mathbf{s}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{s}_n^\dagger\mathbf{0} & \mathbf{s}_n^\dagger M\mathbf{s}_2 & \cdots & \mathbf{s}_n^\dagger M\mathbf{s}_n \end{pmatrix}.$$

Every entry in the first column is 0; but $L^\dagger ML$ is symmetric, so every entry in the first row is 0, as well. Thus the first variable, u_1, is everywhere missing from $\widehat{Q}(\mathbf{u})$.

To prove the converse, suppose that the jth variable is missing from the expression of a quadratic form. Then the jth row and jth column of the matrix associated with the form contain only zeros, implying the determinant of the matrix is zero and the matrix is noninvertible. □

Nondegeneracy

Definition 7.6 *A quadratic form Q is **nondegenerate** if its associated symmetric matrix is invertible, and is **degenerate** otherwise.*

Corollary 7.11 *A quadratic form is nondegenerate if and only if the eigenvalues of its associated symmetric matrix are all nonzero.*

Proof. The determinant of a matrix equals the product of its eigenvalues, so the matrix is invertible if and only if all its eigenvalues are nonzero. □

Eigenvalues of a symmetric matrix

There is more we must say about the eigenvalues of the symmetric matrix M of a quadratic form. Because we obtain eigenvalues as the roots of a polynomial, in general those eigenvalues are complex numbers, even when the entries of M are all real numbers. However, the eigenvalues associated with a quadratic form via its symmetric matrix are all real.

Theorem 7.12. *If M is a symmetric $n \times n$ matrix with real entries, then all the eigenvalues of M are real numbers.*

Proof. Let $p(\lambda)$ be the characteristic polynomial of M (Definition 2.1, p. 35); by the fundamental theorem of algebra, there are n (not necessarily distinct) complex numbers $\lambda_1, \ldots, \lambda_n$ that are the roots of $p(\lambda) = 0$. For each distinct root $\lambda = \alpha + \mathrm{i}\beta$ (with α and β real), there is a complex eigenvector $\mathbf{z} = \mathbf{x} + \mathrm{i}\mathbf{y}$ such that $M\mathbf{z} = \lambda\mathbf{z}$ and $\mathbf{z} \neq \mathbf{0}$. Each of these has a **complex conjugate**:

$$\overline{\lambda} = \alpha - \mathrm{i}\beta, \quad \overline{\mathbf{z}} = \mathbf{x} - \mathrm{i}\mathbf{y}.$$

If $\overline{\lambda} = \lambda$, then $\beta = 0$ so $\lambda = \alpha$, a real number.

Thus, to prove the theorem, we show $\overline{\lambda} = \lambda$; to do this, we calculate the matrix product $\overline{\mathbf{z}}^\dagger M\mathbf{z}$ two ways. First,

$$\overline{\mathbf{z}}^\dagger M\mathbf{z} = \overline{\mathbf{z}}^\dagger(\lambda\mathbf{z}) = \lambda(\overline{\mathbf{z}}^\dagger\mathbf{z}).$$

In the second calculation, we use the fact that $M^\dagger = M = \overline{M}$, because M is symmetric and real, and we equate $\overline{\mathbf{z}}^\dagger M\mathbf{z}$ and $\mathbf{z}^\dagger\overline{\mathbf{z}}$ with their transposes because they are scalars:

$$\overline{\mathbf{z}}^\dagger M\mathbf{z} = (\overline{\mathbf{z}}^\dagger M\mathbf{z})^\dagger = \mathbf{z}^\dagger M^\dagger\overline{\mathbf{z}} = \mathbf{z}^\dagger\overline{M\mathbf{z}} = \mathbf{z}^\dagger\overline{\lambda\mathbf{z}} = \overline{\lambda}(\mathbf{z}^\dagger\overline{\mathbf{z}}) = \overline{\lambda}(\mathbf{z}^\dagger\overline{\mathbf{z}})^\dagger = \overline{\lambda}(\overline{\mathbf{z}}^\dagger\mathbf{z}).$$

Thus $\overline{\lambda}(\overline{\mathbf{z}}^\dagger\mathbf{z}) = \lambda(\overline{\mathbf{z}}^\dagger\mathbf{z})$, and because $\overline{\mathbf{z}}^\dagger\mathbf{z} = \mathbf{x}^\dagger\mathbf{x} + \mathbf{y}^\dagger\mathbf{y} = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 > 0$, we can divide by $\overline{\mathbf{z}}^\dagger\mathbf{z}$ and conclude $\overline{\lambda} = \lambda$. □

Eigenvectors with real eigenvalues

Although the eigenvalues of a real symmetric matrix must be real, the eigenvectors need not be. For example, every nonzero complex vector is an eigenvector of the identity matrix (with real eigenvalue 1). However, we can show that, in a sense, the complex eigenvectors are superfluous: there is always a real eigenvector associated with each real eigenvalue of a real matrix, symmetric or otherwise.

Theorem 7.13. *Suppose* $\mathbf{z}$ *is a complex eigenvector of the real matrix* M*, associated with the real eigenvalue* λ*. Then the real and imaginary parts of* $\mathbf{z}$ *are (real) eigenvectors of* M *associated with* λ*.*

Proof. Write $\mathbf{z} = \mathbf{x} + i\mathbf{y}$; then $\lambda(\mathbf{x} + i\mathbf{y}) = \lambda\mathbf{z} = M\mathbf{z} = M(\mathbf{x} + i\mathbf{y})$. The real and imaginary parts of this equation hold separately; because λ and M are real, the real and imaginary are

$$\lambda\mathbf{x} = M\mathbf{x}, \quad \lambda\mathbf{y} = M\mathbf{y}. \qquad \square$$

The Hessian

Thus all the eigenvalues of a symmetric matric are real, and each distinct eigenvalue has a corresponding real eigenvector. We are now ready to introduce the Hessian of a function of n variables and begin the local analysis of that function near a critical point.

Definition 7.7 *Suppose the function* $z = f(\mathbf{x})$ *has continuous second derivatives on a neighborhood of a critical point* $\mathbf{x} = \mathbf{a}$*. The* ***Hessian of*** f ***at*** $\mathbf{a}$ *is the symmetric matrix of second derivatives*

$$H_{\mathbf{a}} = \begin{pmatrix} f_{11}(\mathbf{a}) & \cdots & f_{1n}(\mathbf{a}) \\ \vdots & \ddots & \vdots \\ f_{n1}(\mathbf{a}) & \cdots & f_{nn}(\mathbf{a}) \end{pmatrix}.$$

The ***Hessian form of*** f ***at*** $\mathbf{a}$ *is the quadratic form associated with the Hessian.*

As we noted already in the two-variable case, continuity of the second derivatives guarantees that $H_{\mathbf{a}}$ is symmetric. Moreover, we continue to use the symbol $H_{\mathbf{a}}$ for the Hessian form as well; thus

$$H_{\mathbf{a}}(x_1, \ldots, x_n) = f_{11}(\mathbf{a})x_1^2 + 2f_{12}(\mathbf{a})x_1x_2 + \cdots + f_{nn}(\mathbf{a})x_n^2.$$

Definition 7.8 *Suppose the function* $z = f(\mathbf{x})$ *has continuous second derivatives near the critical point* $\mathbf{a}$*. Then* $\mathbf{a}$ *is* ***nondegenerate*** *if the Hessian* $H_{\mathbf{a}}$ *of* f *at* $\mathbf{a}$ *is nondegenerate, and is* ***degenerate*** *otherwise.*

The effect of coordinate changes

Our goal is to show that coordinate changes can put a function into a particularly simple form near a nondegenerate critical point. But we must ask: can a coordinate change eliminate a critical point, or can it convert a nondegenerate critical point into a degenerate one? We now show that criticality and nondegeneracy are geometric properties of functions, unaltered by coordinate changes.

Theorem 7.14. *Suppose the coordinate change* $\mathbf{x} = \mathbf{h}(\mathbf{u})$ *transforms* $f(\mathbf{x})$ *into* $g(\mathbf{u})$*:* $f(\mathbf{x}) = f(\mathbf{h}(\mathbf{u})) = g(\mathbf{u})$*. Then* $z = f(\mathbf{x})$ *has a critical point at* $\mathbf{x} = \mathbf{a} = \mathbf{h}(\mathbf{b})$ *if and only if* $z = g(\mathbf{u})$ *has a critical point at* $\mathbf{u} = \mathbf{b}$*.*

Proof. By the chain rule, $\mathrm{d}g_{\mathbf{b}} = \mathrm{d}f_{\mathbf{a}} \circ \mathrm{d}\mathbf{h}_{\mathbf{b}}$. Because $\mathrm{d}\mathbf{h}_{\mathbf{b}}$ is invertible because $\mathbf{h}$ is a coordinate change,

$$\mathrm{d}g_{\mathbf{b}} = \mathbf{0} \quad \Longleftrightarrow \quad \mathrm{d}f_{\mathbf{a}} = \mathbf{0}. \qquad \square$$

Theorem 7.15. *Suppose the coordinate change* $\mathbf{x} = \mathbf{h}(\mathbf{u})$ *transforms* $f(\mathbf{x})$ *into* $g(\mathbf{u})$*, where* f *and* g *have continuous third derivatives. Then* $\mathbf{u} = \mathbf{b}$ *is a nondegenerate critical point of* $z = g(\mathbf{u})$ *if and only if* $\mathbf{x} = \mathbf{a} = \mathbf{h}(\mathbf{b})$ *is a nondegenerate critical point of* $z = f(\mathbf{x})$.

Proof. We have $f(\mathbf{x}) = f(\mathbf{h}(\mathbf{u})) = g(\mathbf{u})$. Let $H_{\mathbf{a}}$ be the Hessian matrix of f at $\mathbf{a}$, and let $H^*_{\mathbf{b}}$ be the Hessian matrix of g at $\mathbf{b}$; we must establish a connection between $H_{\mathbf{a}}$ and $H^*_{\mathbf{b}}$ that implies one is invertible precisely when the other is.

The Hessians appear in the respective Taylor expansions of f and g:

$$f(\mathbf{a}+\Delta\mathbf{x}) - f(\mathbf{a}) = \tfrac{1}{2}\Delta\mathbf{x}^\dagger H_{\mathbf{a}}\,\Delta\mathbf{x} + O((\Delta\mathbf{x})^3),$$
$$g(\mathbf{b}+\Delta\mathbf{u}) - g(\mathbf{b}) = \tfrac{1}{2}\Delta\mathbf{u}^\dagger H^*_{\mathbf{b}}\,\Delta\mathbf{u} + O((\Delta\mathbf{u})^3).$$

However,

$$f(\mathbf{a}+\Delta\mathbf{x}) - f(\mathbf{a}) = \Delta z = g(\mathbf{b}+\Delta\mathbf{u}) - g(\mathbf{b}),$$

so we can begin to connect the two Hessians by writing

$$2\Delta z = \Delta\mathbf{x}^\dagger H_{\mathbf{a}}\,\Delta\mathbf{x} + O((\Delta\mathbf{x})^3) = \Delta\mathbf{u}^\dagger H^*_{\mathbf{b}}\,\Delta\mathbf{u} + O((\Delta\mathbf{u})^3).$$

Now express $\Delta\mathbf{x}$ in terms of $\Delta\mathbf{u}$ by using the differentiability of $\mathbf{h}$ at $\mathbf{b}$:

$$\Delta\mathbf{x} = \mathbf{x} - \mathbf{a} = \mathbf{h}(\mathbf{b}+\Delta\mathbf{u}) - \mathbf{h}(\mathbf{b}) = \mathrm{d}\mathbf{h}_{\mathbf{b}}(\Delta\mathbf{u}) + \boldsymbol{o}(\Delta\mathbf{u}) = L\Delta\mathbf{u} + \boldsymbol{o}(\Delta\mathbf{u}).$$

For visual clarity we have set $\mathrm{d}\mathbf{h}_{\mathbf{b}} = L$ here; the remainder is "little oh" of $\Delta\mathbf{u}$. By Exercise 3.28 (p. 104), $L\Delta\mathbf{u} = \boldsymbol{O}(\Delta\mathbf{u})$, so $\Delta\mathbf{x} = \boldsymbol{O}(\Delta\mathbf{u})$.

For every $\Delta\mathbf{u} \neq \mathbf{0}$, write $\Delta\mathbf{u} = s\Delta\mathbf{y}$ with $\Delta\mathbf{y}$ a unit vector and a suitable $s > 0$. Then $O((\Delta\mathbf{u})^3) = O(s^3)$,

$$\Delta\mathbf{x} = sL\Delta\mathbf{y} + \boldsymbol{o}(s) = s\big(L\Delta\mathbf{y} + \boldsymbol{o}(s)/s\big), \quad O((\Delta\mathbf{x})^3) = O(s^3),$$

and we can write the two expressions for $2\Delta z$ as

$$s^2\left(L\Delta\mathbf{y} + \boldsymbol{o}(s)/s\right)^\dagger H_{\mathbf{a}}\left(L\Delta\mathbf{y} + \boldsymbol{o}(s)/s\right) + O(s^3) = s^2\,\Delta\mathbf{y}^\dagger H^*_{\mathbf{b}}\Delta\mathbf{y} + O(s^3).$$

Now divide the equation by s^2 and take the limit as $s \to 0$, using $o(s)/s \to 0$ and $O(s^3)/s^2 \to 0$. The result is

$$(L\Delta\mathbf{y})^\dagger H_{\mathbf{a}}\,(L\Delta\mathbf{y}) = \Delta\mathbf{y}^\dagger\big(L^\dagger H_{\mathbf{a}} L\big)\Delta\mathbf{y} = \Delta\mathbf{y}^\dagger H^*_{\mathbf{b}}\,\Delta\mathbf{y}$$

for every $\Delta\mathbf{y} \neq \mathbf{0}$. This implies $L^\dagger H_{\mathbf{a}} L = H^*_{\mathbf{b}}$ and hence

$$\det H^*_{\mathbf{b}} = \det H_{\mathbf{a}}\,(\det L)^2.$$

Because $\det L = \det \mathrm{d}\mathbf{h}_{\mathbf{b}} \neq 0$ because $\mathbf{h}$ is a coordinate change, $\det H^*_{\mathbf{b}} \neq 0$ if and only if $\det H_{\mathbf{a}} \neq 0$. □

Analogous equations

The equations $dg_{\mathbf{b}} = df_{\mathbf{a}} \circ \mathrm{d}\mathbf{h}_{\mathbf{a}}$ and $\det H^*_{\mathbf{b}} = \det H_{\mathbf{a}}\,(\det \mathrm{d}\mathbf{h}_{\mathbf{b}})^2$ in the last two proofs are the multivariable analogues of the earlier equations $g'(0) = f'(0)\,h'(0)$

and $g''(0) = f''(0)(h'(0))^2$ that showed criticality and nondegeneracy were geometric properties of single-variable functions (p. 222).

Morse theory and Morse's lemma

We are now ready to state and prove the main theorem. It first appears in an important paper on the topological properties of multivariable functions that Marston Morse published in 1925 [13]. Because the theorem was just one of several technical facts he needed to establish the paper's main results (now called *Morse theory*), it was natural for him to label this fact as a lemma. For us, however, the fact is central, though it is still always called *Morse's lemma*: at a nondegenerate critical point, a function can always be converted into a sum of squares.

Theorem 7.16 (Morse's lemma). *Suppose $z = f(\mathbf{x})$ has continuous third derivatives on an open set X^n, the point $\mathbf{x} = \mathbf{a}$ in X^n is a nondegenerate critical point of f, and the Hessian matrix $H_\mathbf{a}$ has r negative eigenvalues. Then, in a sufficiently small window $W_\mathbf{a}$ centered at $\mathbf{a}$, there is a coordinate change $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$ for which*

$$\begin{aligned}\Delta z &= f(\mathbf{a}+\Delta\mathbf{x}) - f(\mathbf{a}) \\ &= -(\Delta u_1)^2 - \cdots - (\Delta u_r)^2 + (\Delta u_{r+1})^2 + \cdots + (\Delta u_n)^2.\end{aligned}$$

Because the Hessian $H_\mathbf{a}$ is symmetric and the critical point is nondegenerate, the eigenvalues of $H_\mathbf{a}$ are all real and nonzero. If all are positive (i.e., $r = 0$ in the statement of the theorem), then there are no negative squares in the sum. If all eigenvalues are negative (i.e., $r = n$), then there are no positive squares in the sum.

Survey of the proof

The proof of Morse's lemma breaks up naturally into three parts. In the first part (Theorem 7.18), a coordinate change reduces the window equation for a function at an nondegenerate critical point into a simple quadratic form with variable coefficients. In the second part (Theorem 7.19), a further coordinate change "diagonalizes" the quadratic form. This means that the form becomes a sum of positive and negative squares (and the symmetric matrix associated with it becomes a diagonal matrix). But significantly, it also means that the coefficients of the quadratic form become constants. In other words, any function "looks like" a sum of squares near a nondegenerate critical point. The third part of the proof of Morse's lemma (Theorem 7.25) shows that the number of negative squares in the sum does not depend on the way the coordinate changes were chosen, but is always equal to the number of negative eigenvalues in the Hessian of the given function at its critical point.

Morse's lemma, part 1: expanding by powers

Morse begins the proof of the Morse lemma by expanding a function into linear and quadratic terms in a way that is uncannily similar to Taylor's expansion. Taylor's formula splits the function into three simple pieces—a constant, a linear form, and a quadratic form—plus a fourth piece that contains the remaining "complexity" of the function. Morse recasts the formula so there is no separate remainder; the coefficients of the quadratic form become variable, and contain all the complexity that Taylor's formula puts into the remainder. We have already seen Morse's expansion put to use: on pages 219–220 we used it to determine the local behavior of a function of one variable near a critical point. Here, then, for the sake of comparison are the theorems that provide the expansions of Taylor and Morse.

Theorem 7.17 (Taylor). *Suppose $z = f(\mathbf{x})$ has continuous third derivatives on an open set that contains the line segment from $\mathbf{a}$ to $\mathbf{a}+\Delta\mathbf{x}$. Then*

$$f(\mathbf{a}+\Delta\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a})\,\Delta x_i + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i\,\partial x_j}(\mathbf{a})\,\Delta x_i\,\Delta x_j + O((\Delta\mathbf{x})^3). \qquad \square$$

Theorem 7.18 (Morse). *Suppose $z = f(\mathbf{x})$ has continuous third derivatives on an open set that contains the line segment from $\mathbf{a}$ to $\mathbf{a}+\Delta\mathbf{x}$. Then there are continuously differentiable functions $h_{ij}(\Delta\mathbf{x}) = h_{ji}(\Delta\mathbf{x})$ for which*

$$f(\mathbf{a}+\Delta\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a})\,\Delta x_i + \sum_{i,j=1}^{n} h_{ij}(\Delta\mathbf{x})\,\Delta x_i\,\Delta x_j,$$

and $h_{ij}(\mathbf{0}) = \dfrac{1}{2}\dfrac{\partial^2 f}{\partial x_i\,\partial x_j}(\mathbf{a})$.

Proof. For clarity, we separate the proof into a number of steps. One of our aims is to provide explicit instructions for constructing the coefficients $h_{ij}(\Delta\mathbf{x})$ of the quadratic form. Note, in what follows, similarities with the proof of Taylor's theorem.

With the following lemma, we are able to build all the terms in Morse's formula, including the crucial coefficients h_{ij}. Step 1

Lemma 7.3. *Suppose $z = F(\mathbf{x})$ has continuous derivatives of order $k+1$ on an open set that contains the line segment from $\mathbf{a}$ to $\mathbf{a}+\Delta\mathbf{x}$. Then there are functions $p_i(\Delta\mathbf{x})$ with continuous derivatives of order k for which*

$$F(\mathbf{a}+\Delta\mathbf{x}) = F(\mathbf{a}) + \sum_{i=1}^{n} p_i(\Delta\mathbf{x})\,\Delta x_i,$$

and $p_i(\mathbf{0}) = \dfrac{\partial F}{\partial x_i}(\mathbf{a})$, $i = 1,\ldots,n$.

Proof. We express the difference $\Delta z = F(\mathbf{a}+\Delta\mathbf{x}) - F(\mathbf{a})$ as an integral, as in the beginning of the proof of Taylor's theorem for a single-variable function (cf. p. 79):

$$\int_0^1 \frac{d}{dt}F(\mathbf{a}+t\Delta\mathbf{x})\,dt = F(\mathbf{a}+t\Delta\mathbf{x})\Big|_0^1 = F(\mathbf{a}+\Delta\mathbf{x}) - F(\mathbf{a}) = \Delta z.$$

In this multivariable setting, the chain rule gives us

$$\frac{d}{dt}F(\mathbf{a}+t\Delta\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial F}{\partial x_i}(\mathbf{a}+t\Delta\mathbf{x})\,\Delta x_i,$$

so

$$\Delta z = \sum_{i=1}^{n} \left(\int_0^1 \frac{\partial F}{\partial x_i}(\mathbf{a}+t\Delta\mathbf{x})\,dt \right) \Delta x_i.$$

Therefore we take

$$p_i(\Delta\mathbf{x}) = \int_0^1 \frac{\partial F}{\partial x_i}(\mathbf{a} + t\Delta\mathbf{x})\, dt.$$

Because $\partial F/\partial x_i$ has continuous derivatives of order k, so does p_i. Moreover,

$$p_i(\mathbf{0}) = \int_0^1 \frac{\partial F}{\partial x_i}(\mathbf{a} + t\,\mathbf{0})\, dt = \frac{\partial F}{\partial x_i}(\mathbf{a}) \int_0^1 dt = \frac{\partial F}{\partial x_i}(\mathbf{a}). \qquad \square$$

Step 2

Now apply Lemma 7.3 to the function f itself to obtain functions $g_i(\Delta\mathbf{x})$ for which

$$f(\mathbf{a} + \Delta\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} g_i(\Delta\mathbf{x})\, \Delta x_i.$$

According to the same lemma, each function g_i has continuous second derivatives, and

$$g_i(\mathbf{0}) = \frac{\partial f}{\partial x_i}(\mathbf{a}),$$

which gives us a start on Morse's expansion.

Step 3

Apply Lemma 7.3 again to each $g_i(\Delta\mathbf{x})$, $i = 1, \ldots, n$, this time taking $\mathbf{a} = \mathbf{0}$. We get functions $\widetilde{h}_{ij}(\Delta\mathbf{x})$, $j = 1, \ldots, n$, with continuous first derivatives for which

$$g_i(\Delta\mathbf{x}) = g_i(\mathbf{0}) + \sum_{j=1}^{n} \widetilde{h}_{ij}(\Delta\mathbf{x})\, \Delta x_j = \frac{\partial f}{\partial x_i}(\mathbf{a}) + \sum_{j=1}^{n} \widetilde{h}_{ij}(\Delta\mathbf{x})\, \Delta x_j,$$

$$\text{and } \widetilde{h}_{ij}(\mathbf{0}) = \frac{\partial g_i}{\partial x_j}(\mathbf{0}).$$

Comment: Nominally, each g_i is a function of the window variables Δx_j, but because Δx_j and x_j differ merely by a constant ($\Delta x_j = x_j - a_j$), the differential operators

$$\frac{\partial}{\partial(\Delta x_j)} \text{ and } \frac{\partial}{\partial x_j}$$

have the same action. For simplicity we therefore write

$$\frac{\partial g_i}{\partial x_j} \text{ instead of } \frac{\partial g_i}{\partial(\Delta x_j)}$$

here and in all the following work.

Step 4

Now substitute the expression for $g_i(\Delta\mathbf{x})$ into the formula for $f(\mathbf{a} + \Delta\mathbf{x})$ in Step 2:

$$f(\mathbf{a} + \Delta\mathbf{x}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{a})\, \Delta x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \widetilde{h}_{ij}(\Delta\mathbf{x})\, \Delta x_i\, \Delta x_j.$$

This looks like Morse's expansion; in particular, the last term is a quadratic form with variable coefficients $\widetilde{h}_{ij}(\Delta\mathbf{x})$. But nothing in Lemma 7.3 ensures that $\widetilde{h}_{ji}(\Delta\mathbf{x}) =$

$\widetilde{h}_{ij}(\Delta\mathbf{x})$ for every $i,j=1,\dots,n$, as required by the theorem. (In other words, the matrix $\widetilde{H}(\Delta\mathbf{x})=\big(\widetilde{h}_{ij}(\Delta\mathbf{x})\big)$ that defines the quadratic form need not be symmetric.)

Step 5

But we can use Lemma 7.2 to replace the matrix $\widetilde{H}$ by the symmetric matrix $H=(\widetilde{H}+\widetilde{H}^{\dagger})/2$ without altering the quadratic form. That is, if we let

$$h_{ij}(\Delta\mathbf{x})=\frac{\widetilde{h}_{ij}(\Delta\mathbf{x})+\widetilde{h}_{ji}(\Delta\mathbf{x})}{2},$$

then

$$f(\mathbf{a}+\Delta\mathbf{x})=f(\mathbf{a})+\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}(\mathbf{a})\,\Delta x_i+\sum_{i=1}^{n}\sum_{j=1}^{n}h_{ij}(\Delta\mathbf{x})\,\Delta x_i\,\Delta x_j$$

and $h_{ji}(\Delta\mathbf{x})=h_{ij}(\Delta\mathbf{x})$ for all $i,j=1,\dots,n$.

Step 6

In remains only to verify that $h_{ij}(\mathbf{0})=\dfrac{1}{2}\dfrac{\partial^2 f}{\partial x_i\,\partial x_j}(\mathbf{a})$. We claim that, in fact,

$$\widetilde{h}_{ij}(\mathbf{0})=\frac{1}{2}\frac{\partial^2 f}{\partial x_j\,\partial x_i}(\mathbf{a})=\frac{1}{2}\frac{\partial^2 f}{\partial x_i\,\partial x_j}(\mathbf{a})=\widetilde{h}_{ji}(\mathbf{0}).$$

To prove the claim, note (Step 3) that $\widetilde{h}_{ij}(\mathbf{0})=\dfrac{\partial g_i}{\partial x_j}(\mathbf{0})$. Therefore, because

$$g_i(\Delta\mathbf{x})=\int_0^1\frac{\partial f}{\partial x_i}(\mathbf{a}+t\Delta\mathbf{x})\,dt,$$

we can link $\widetilde{h}_{ij}$ to f by calculating the appropriate partial derivative of g_i. This involves differentiation under the integral sign, a delicate matter but one that is allowed here because the integrand is continuously differentiable; see an introductory text on real analysis. We have (by the chain rule)

$$\frac{\partial g_i}{\partial x_j}(\Delta\mathbf{u})=\int_0^1\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}(\mathbf{a}+t\Delta\mathbf{x})\right)dt=\int_0^1\frac{\partial^2 f}{\partial x_j\,\partial x_i}(\mathbf{a}+t\Delta\mathbf{x})\,t\;dt,$$

from which it follows that

$$\begin{aligned}\widetilde{h}_{ij}(\mathbf{0})=\frac{\partial g_i}{\partial x_j}(\mathbf{0})&=\int_0^1\frac{\partial^2 f}{\partial x_j\,\partial x_i}(\mathbf{a})\,t\;dt\\&=\frac{\partial^2 f}{\partial x_j\,\partial x_i}(\mathbf{a})\int_0^1 t\;dt=\frac{1}{2}\frac{\partial^2 f}{\partial x_j\,\partial x_i}(\mathbf{a}).\end{aligned}$$

This completes the proof of Theorem 7.18, and incidentally shows that, even though the matrix $\widetilde{H}(\Delta\mathbf{x})$ may not be symmetric in general, at least it is when $\Delta\mathbf{x}=\mathbf{0}$. □

Example: constructing the h_{ij}

Let us see how the instructions provided in this proof (in Steps 2 and 3) give us formulas for the functions h_{ij} at a critical point of the "tipped wine bottle" function

$f_m(x,y) = (x^2+y^2-1)^2 + mx$ (Chapter 7.2, pages 231–237). These are the formulas for A, B, and C that appear on page 234.

To find the h_{ij}, we must first construct the g_i, and these involve partial derivatives of the expression $f_m(\mathbf{x})$ (evaluated at $\mathbf{x} = \mathbf{a} + t\Delta\mathbf{x}$). But the window function

$$\Delta z = f_m(\mathbf{x}) - f_m(\mathbf{a})$$

has the same partial derivatives; the two functions merely differ by a constant. Furthermore, when we restrict $\mathbf{a}$ to the form $(p,0)$ (because we are interested only in critical points), we can use the expression for Δz we have already computed on page 234:

$$\begin{aligned}\Delta z = {} & (6p^2-2)(\Delta x)^2 + (2p^2-2)(\Delta y)^2 \\ & + 4p\,(\Delta x)^3 + 4p\,\Delta x\,(\Delta y)^2 + (\Delta x)^4 + 2(\Delta x)^2(\Delta y)^2 + (\Delta y)^4.\end{aligned}$$

Finally, keeping in mind the comment in Step 3 of the last proof, that derivatives with respect to Δx_i and x_i are interchangeable, we can now compute

$$\begin{aligned}\frac{\partial(\Delta z)}{\partial(\Delta x)} &= 2(6p^2-2)\Delta x + 12p(\Delta x)^2 + 4p(\Delta y)^2 + 4(\Delta x)^3 + 4\Delta x\,(\Delta y)^2, \\ \frac{\partial(\Delta z)}{\partial(\Delta y)} &= 2(2p^2-2)\Delta y + 8p\,\Delta x\,\Delta y + 4(\Delta x)^2\Delta y + 4(\Delta y)^3.\end{aligned}$$

Thus,

$$\begin{aligned}g_1(\Delta x,\Delta y) &= \int_0^1 \frac{\partial(\Delta z)}{\partial(\Delta x)}(t\Delta x, t\Delta y)\,dt \\ &= \int_0^1 \left\{2t(6p^2-2)\Delta x + 4t^2\left[3p(\Delta x)^2 + p(\Delta y)^2\right] + 4t^3\left[(\Delta x)^3 + \Delta x\,(\Delta y)^2\right]\right\} dt \\ &= (6p^2-2)\Delta x + 4p(\Delta x)^2 + \tfrac{4}{3}p(\Delta y)^2 + (\Delta x)^3 + \Delta x\,(\Delta y)^2.\end{aligned}$$

In a similar way,

$$\begin{aligned}g_2(\Delta x,\Delta y) &= \int_0^1 \frac{\partial(\Delta z)}{\partial(\Delta y)}(t\Delta x, t\Delta y)\,dt \\ &= (2p^2-2)\Delta y + \tfrac{8}{3}p\,\Delta x\,\Delta y + (\Delta x)^2\Delta y + (\Delta y)^3.\end{aligned}$$

We are now ready to compute the four functions h_{ij}. By definition, $h_{11} = \widetilde{h}_{11}$, so

$$\begin{aligned}h_{11}(\Delta x,\Delta y) &= \int_0^1 \frac{\partial g_1}{\partial(\Delta x)}(t\Delta x, t\Delta y)\,dt \\ &= \int_0^1 \left\{[6p^2-2] + t\,[8p\,\Delta x] + t^2[3(\Delta x)^2 + (\Delta y)^2]\right\} dt \\ &= (6p^2-2) + 4p\,\Delta x + (\Delta x)^2 + \tfrac{1}{3}(\Delta y)^2.\end{aligned}$$

(This is the function $A(\Delta x, \Delta y)$ given on page 234). Next, notice that

$$\frac{\partial g_1}{\partial(\Delta y)} = \tfrac{8}{3}p\,\Delta y + 2\,\Delta x\,\Delta y = \frac{\partial g_2}{\partial(\Delta x)},$$

implying $\widetilde{h}_{12} = \widetilde{h}_{21} = h_{12}$. We have

$$\begin{aligned} h_{12}(\Delta x, \Delta y) &= \int_0^1 \frac{\partial g_1}{\partial(\Delta y)}(t\Delta x, t\Delta y)\,dt \\ &= \int_0^1 \left\{ t\left[\tfrac{8}{3}p\,\Delta y\right] + t^2\left[2\,\Delta x\,\Delta y\right]\right\} dt = \tfrac{4}{3}p\,\Delta y + \tfrac{2}{3}\,\Delta x\,\Delta y \\ &= B(\Delta x, \Delta y). \end{aligned}$$

Finally, because $h_{22} = \widetilde{h}_{22}$, we have

$$\begin{aligned} h_{22}(\Delta x, \Delta y) &= \int_0^1 \frac{\partial(\Delta g_2)}{\partial(\Delta y)}(t\Delta x, t\Delta y)\,dt \\ &= \int_0^1 \left\{ [2p^2 - 2] + t\left[\tfrac{8}{3}p\,\Delta x\right] + t^2\left[(\Delta x)^2 + 3(\Delta y)^2\right]\right\} dt \\ &= (2p^2 - 2) + \tfrac{4}{3}p\,\Delta x + \tfrac{1}{3}(\Delta x)^2 + (\Delta y)^2. \end{aligned}$$

Because this is the function $C(\Delta x, \Delta y)$ given earlier, we have completed the example.

Morse's lemma, part 2: diagonalizing the quadratic form

We now move on to the next part of the proof of Morse's lemma. We can assume, by Theorem 7.18, that our function is already written in window coordinates as a quadratic form with variable coefficients:

$$\Delta z = f(\mathbf{a} + \Delta\mathbf{x}) - f(\mathbf{a}) = \sum_{i,j=1}^{n} h_{ij}(\Delta\mathbf{x})\,\Delta x_i\,\Delta x_j,$$

where $h_{ji}(\Delta\mathbf{x}) = h_{ij}(\Delta\mathbf{x})$ and

$$h_{ij}(\mathbf{0}) = \frac{1}{2}\frac{\partial^2 f}{\partial u_i\,\partial u_j}(\mathbf{a}).$$

Our goal is to "diagonalize" this quadratic form. If the coefficients were constants instead of functions, then linear algebra would provide a standard diagonalization method that involves changing coordinates, one variable at a time, by "completing the square." We actually use this method because, as Morse pointed out, it works just as well with variable coefficients.

The leading coefficient

The first step in completing the square is to divide by the leading coefficient (this is h_{11} in the quadratic form we are dealing with, and was A in the example we worked through on pages 234–237); therefore that coefficient must be nonzero. Of course, we have no reason a priori to expect $h_{11} \neq 0$. Even in the simple example

$$Q(\Delta x_1, \Delta x_2) = 2\,\Delta x_1\,\Delta x_2,$$

the leading coefficient is zero ($h_{11} = h_{22} = 0$, $h_{12} = h_{21} = 1$). In this case, though, we can fix the problem with an obvious coordinate change:

$$\Delta x_1 = \Delta y_1 - \Delta y_2, \quad \Delta x_2 = \Delta y_1 + \Delta y_2.$$

Then

$$Q(\Delta x_1, \Delta x_2) = 2\,\Delta x_1\,\Delta x_2 = (\Delta y_1)^2 - (\Delta y_2)^2 = Q^*(\Delta y_1, \Delta y_2),$$

Making the leading coefficient nonzero

so the coefficients of the form Q^* that results from the coordinate change are $h^*_{11} = 1$, $h^*_{12} = h^*_{21} = 0$, $h^*_{22} = -1$. In fact, the following lemma says we can always make the leading coefficient nonzero, at least if the form is nondegenerate. The lemma concerns a quadratic form with variable coefficients

$$Q(\Delta\mathbf{x}) = \sum_{i,j=1}^{n} h_{ij}(\Delta\mathbf{x})\,\Delta x_i\,\Delta x_j, \quad h_{ji}(\Delta\mathbf{x}) = h_{ij}(\Delta\mathbf{x}).$$

Lemma 7.4. *Suppose the matrix $h_{ij}(\mathbf{0})$ is invertible. Then there is a linear coordinate change $\Delta\mathbf{x} = \mathbf{L}(\Delta\mathbf{y})$ for which*

$$Q(\mathbf{L}(\Delta\mathbf{y})) = Q^*(\Delta\mathbf{y}) = \sum_{i,j=1}^{n} h^*_{ij}(\Delta\mathbf{y})\,\Delta y_i\,\Delta y_j,$$

with $h^*_{11}(\mathbf{0}) \neq 0$.

Proof. Let $H(\Delta\mathbf{x})$ be the symmetric matrix with entry $h_{ij}(\Delta\mathbf{x})$ in the ith row and jth column. Suppose first that one of the diagonal elements of $H(\mathbf{0})$ is nonzero, say $h_{JJ}(\mathbf{0}) \neq 0$. Define

$$\mathbf{L} : \Delta x_1 = \Delta y_J, \quad \Delta x_J = \Delta y_1, \quad \Delta x_k = \Delta y_k, \quad k \neq 1, J.$$

This is a transposition permutation and is its own inverse (and if $J = 1$, it is the identity). In terms of the new variables, $h^*_{11}(\Delta\mathbf{y}) = h_{JJ}(\Delta\mathbf{x})$, so $h^*_{11}(\mathbf{0}) \neq 0$, and we are done.

The alternative is that all diagonal elements of $H(\mathbf{0})$ are zero. In that case, some other element $h_{1j}(\mathbf{0})$, $j = 2, \ldots, n$ in the first row of $H(\mathbf{0})$ must be nonzero. Otherwise, $\det H(\mathbf{0}) = 0$, which is contrary to hypothesis. So suppose $h_{1J}(\mathbf{0}) = h_{J1}(\mathbf{0}) \neq 0$, where $J \neq 1$. Define

$$\mathbf{L} : \Delta x_1 = \Delta y_1 - \Delta y_J, \quad \Delta x_J = \Delta y_1 + \Delta y_J, \quad \Delta x_k = \Delta y_k, \qquad k \neq 1, J.$$

This $\mathbf{L}$ is also invertible; in fact, it is a rotation–dilation of the $(\Delta x_1, \Delta x_J)$-plane. To determine $h^*_{11}(\Delta\mathbf{y})$, we need to determine all places where $(\Delta y_1)^2$ appears as a quadratic factor in the form $Q^*(\Delta\mathbf{y}) = Q(\Delta\mathbf{x})$. There are three such places: in $(\Delta x_1)^2$, in $(\Delta x_J)^2$, and in $\Delta x_1\,\Delta x_J$. We find

$$h^*_{11}(\Delta\mathbf{y}) = h_{11}(\Delta\mathbf{x}) + h_{JJ}(\Delta\mathbf{x}) + 2h_{1J}(\Delta\mathbf{x}),$$

and thus

$$h^*_{11}(\mathbf{0}) = h_{11}(\mathbf{0}) + h_{JJ}(\mathbf{0}) + 2h_{1J}(\mathbf{0}).$$

The first two terms are diagonal elements of $H(\mathbf{0})$ and hence, by assumption, are zero; the remaining term gives $h^*_{11}(\mathbf{0}) = 2h_{1J}(\mathbf{0}) \neq 0$. □

Reducing the function to a sum of squares

The following theorem carries out the diagonalization process that gives Δz as a sum of positive and negative squares. It is the heart of Morse's lemma but does not complete the proof because it does not determine how many of the squares are positive and how many are negative.

Theorem 7.19. *Suppose $z = f(\mathbf{x})$ has continuous third derivatives on an open set X^n, and the point $\mathbf{a}$ in X^n is a nondegenerate critical point of f. Then, in a sufficiently small window $W_\mathbf{a}$ centered at $\mathbf{a}$, there is a coordinate change $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$ so that*

$$\Delta z = f(\mathbf{a}+\Delta\mathbf{x}) - f(\mathbf{a}) = \pm(\Delta u_1)^2 \pm \cdots \pm (\Delta u_n)^2.$$

Proof. We can assume by Theorem 7.18 (p. 249) that we have already written Δz as a quadratic form with variable coefficients:

$$\Delta z = \sum_{i,j=1}^{n} h_{ij}(\Delta\mathbf{x})\,\Delta x_i \Delta x_j,$$

where $h_{ji}(\Delta\mathbf{x}) = h_{ij}(\Delta\mathbf{x})$ and

$$h_{ij}(\mathbf{0}) = \frac{1}{2}\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{a}).$$

Moreover, because f has continuous third derivatives, the same theorem tells us that the coefficients $h_{ij}(\Delta\mathbf{x})$ have continuous first derivatives.

Proof by induction

This proof also goes in stages; at each stage, a coordinate change "splits off" one more variable as a perfect square. In other words, we claim that after k stages, the window equation will look like

$$\Delta z = \pm(\Delta v_1)^2 \pm \cdots \pm (\Delta v_k)^2 + \sum_{i,j=k+1}^{n} h^*_{ij}(\Delta\mathbf{v})\,\Delta v_i \Delta v_j,$$

and that the new coefficients h^*_{ij} are rational functions of the coefficients from the previous stage. Because each stage is like every other, the proof is a mathematical induction. Thus, we assume that we have already reached the stage where $k = M-1$ squares have been "split off" from the quadratic form, and deduce that the next stage, $k = M$, also holds. (The initial step in the induction is just the one where $M = 1$, so we do not need to prove it separately.)

Thus we focus on the residual quadratic form

$$Q_M(\Delta\mathbf{v}) = \sum_{i,j=M}^{n} h^*_{ij}(\Delta\mathbf{v})\,\Delta v_i \Delta v_j.$$

Note that this is a quadratic form in just the variables $\Delta v_M, \ldots, \Delta v_n$, although the coefficients h^*_{ij} remain functions of all the variables $\Delta\mathbf{v} = (\Delta v_1, \ldots, \Delta v_n)$. The lead-

ing coefficient is h^*_{MM}, and we can assume (by Lemma 7.4) that $h^*_{MM}(\mathbf{0}) \neq 0$. If we separate out all appearances of Δv_M as a quadratic factor, we get

$$Q_M(\Delta\mathbf{v}) = h^*_{MM}(\Delta\mathbf{v}) \left((\Delta v_M)^2 + 2\,\Delta v_M \sum_{j=M+1}^{n} \frac{h^*_{Mj}(\Delta\mathbf{v})}{h^*_{MM}(\Delta\mathbf{v})} \Delta v_j \right) + \sum_{i,j=M+1}^{n} h^*_{ij}(\Delta\mathbf{v})\,\Delta v_i \Delta v_j.$$

Completing the square

Completing the square then gives us

$$Q_M(\Delta\mathbf{v}) = h^*_{MM}(\Delta\mathbf{v}) \left(\Delta v_M + \sum_{j=M+1}^{n} \frac{h^*_{Mj}(\Delta\mathbf{v})}{h^*_{MM}(\Delta\mathbf{v})} \Delta v_j \right)^2 - h^*_{MM}(\Delta\mathbf{v}) \left(\sum_{j=M+1}^{n} \frac{h^*_{Mj}(\Delta\mathbf{v})}{h^*_{MM}(\Delta\mathbf{v})} \Delta v_j \right)^2 + \sum_{i,j=M+1}^{n} h^*_{ij}(\Delta\mathbf{v})\,\Delta v_i \Delta v_j.$$

Together, the terms on the second line constitute a new quadratic form in the variables $\Delta v_{M+1}, \ldots, \Delta v_n$ alone, but with variable coefficients that still depend on all the Δv_i, in general. We write that new form as

$$Q_{M+1}(\Delta\mathbf{v}) = \sum_{i,j=M+1}^{n} \widehat{h}_{ij}(\Delta\mathbf{v})\,\Delta v_i \Delta v_j.$$

The formulas show that the new coefficients $\widehat{h}_{ij}$ are rational functions of the h^*_{ij} (in which only h^*_{MM} appears in the denominator). Therefore, the new $\widehat{h}_{ij}$ have continuous first derivatives wherever the denominator $h^*_{MM}(\Delta\mathbf{v})$ does not vanish. This confirms the assertion about the coefficients that is part of the induction.

The new residual form

Completion of the square leads us to the coordinate change

$$\mathbf{h}_M : \begin{cases} \Delta w_M = \sqrt{|h^*_{MM}(\Delta\mathbf{v})|} \left(\Delta v_M + \sum_{j=M+1}^{n} \frac{h^*_{Mj}(\Delta\mathbf{v})}{h^*_{MM}(\Delta\mathbf{v})} \Delta v_j \right), \\ \Delta w_i = \Delta v_i, \qquad i \neq M, \end{cases}$$

that transforms Q_M into

$$Q_M(\Delta\mathbf{v}) = \pm(\Delta w_M)^2 + Q_{M+1}(\Delta\mathbf{w}).$$

The new coordinates split off one more variable as a perfect square, and leave a new residue Q_{M+1} that is again a quadratic form, but with one less quadratic variable. The coefficients of the new residual form are continuously differentiable functions of the new coordinates.

The coordinate change

The induction is therefore completed as soon as we prove that the map $\mathbf{h}_M$ is a valid coordinate change. We use the inverse function theorem. First, note that the

components of $\mathbf{h}_M$ have continuous first derivatives near the origin, because they are rational functions of $\sqrt{|h^*_{MM}(\Delta\mathbf{v})|}$ and $h^*_{Mj}(\Delta\mathbf{v})$, $j = M+1,\dots,n$. Next,

$$d(\mathbf{h}_M)_{\mathbf{0}} = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \sqrt{|h^*_{MM}(\mathbf{0})|} & \dfrac{h^*_{M,M+1}(\mathbf{0})}{\sqrt{|h^*_{MM}(\mathbf{0})|}} & \cdots & \dfrac{h^*_{M,n}(\mathbf{0})}{\sqrt{|h^*_{MM}(\mathbf{0})|}} \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix},$$

so $\det d(\mathbf{h}_M)_{\mathbf{0}} = \sqrt{|h^*_{MM}(\mathbf{0})|} \neq 0$ and the linear map $d(\mathbf{h}_M)_{\mathbf{0}}$ is invertible. The inverse function theorem then implies that $\mathbf{h}_M$ itself is invertible on some window W_M centered at $\mathbf{a}$, and the inverse is continuously differentiable on the image $U_M = \mathbf{h}_M(W_M)$.

The coordinate change $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$ that will carry out the entire diagonalization is the composite $\mathbf{h} = \mathbf{h}_n \circ \cdots \circ \mathbf{h}_1$ that carries out the individual changes, one after another. The proof makes it reasonably clear that the composite is well defined, that is, that we can always carry out the next coordinate change in the sequence. Alternatively, note that each successive pair $\mathbf{h}_{i+1} \circ \mathbf{h}_i$ of changes is defined on the open set $U_i \cap W_{i+1}$, which is certainly nonempty because it contains $\mathbf{0} = \mathbf{h}_i(\mathbf{0})$. Finally, by the chain rule, the composite $\mathbf{h}$ is continuously differentiable. □

Morse's lemma, part 3: role of the Hessian

We now come to the final part of the proof of Morse's lemma, where we show that the number of negative squares in the new formula

$$\Delta z = \pm(\Delta u_1)^2 \pm \cdots \pm (\Delta u_n)^2$$

for f is equal to the number of negative eigenvalues in the Hessian of f at $\mathbf{x} = \mathbf{a}$. In particular, it follows that this number does not depend on the choices we made in constructing the coordinate change $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$.

In terms of the new coordinates, Δz is a particularly simple quadratic form. For clarity, let us assume there are s negative squares and the coordinates have been rearranged so all the negative squares come first in the new formula; then

$$\Delta z = Q_K(\Delta\mathbf{u}) = \Delta\mathbf{u}^\dagger K \Delta\mathbf{u},$$

where the symmetric matrix K representing the form is

$$K = \begin{pmatrix} -I_s & \\ & I_{n-s} \end{pmatrix},$$

and the off-diagonal entries are all zero. We can also write Δz as the Taylor expansion

$$\Delta z = \tfrac{1}{2} H_{\mathbf{a}}(\Delta\mathbf{x}) + O((\Delta\mathbf{x})^3) = \tfrac{1}{2}\Delta\mathbf{x}^\dagger H_{\mathbf{a}}\,\Delta\mathbf{x} + O((\Delta\mathbf{x})^3),$$

in which $H_{\mathbf{a}}$ is the Hessian of f at the critical point in terms of the original $\mathbf{x}$ coordinates. The next theorem provides the first link between the matrices K and $H_{\mathbf{a}}$.

Theorem 7.20. *Let $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$ be the coordinate change of Theorem* 7.19*, and let $L = \mathrm{d}\mathbf{h}_{\mathbf{0}}$ be the derivative of $\mathbf{h}$ at $\Delta\mathbf{x} = \mathbf{0}$; then*

$$L^{\dagger}KL = \tfrac{1}{2}H_{\mathbf{a}}.$$

Proof. The proof is left as an exercise; it is similar to the earlier proof (Theorem 7.15, p. 247) that connects the Hessians of equivalent functions at corresponding critical points. □

Corollary 7.21 *The matrices $\frac{1}{2}H_{\mathbf{a}}$ and K represent the same quadratic form in the coordinates $\Delta\mathbf{x}$ and $\Delta\mathbf{u}$, respectively.*

Proof. Let $L = \mathrm{d}\mathbf{h}_{\mathbf{0}}$, as in Theorem 7.20; then the linear coordinate change $\Delta\mathbf{u} = L\Delta\mathbf{x}$ converts $Q_K(\Delta\mathbf{u}) = \Delta\mathbf{u}^{\dagger}K\Delta\mathbf{u}$ into

$$\widehat{Q}_K(\Delta\mathbf{x}) = Q_K(L\Delta\mathbf{x}) = (L\Delta\mathbf{x})^{\dagger}K(L\Delta\mathbf{x}) = \mathbf{x}^{\dagger}L^{\dagger}KL\,\Delta\mathbf{x} = \tfrac{1}{2}\,\Delta\mathbf{x}^{\dagger}H_{\mathbf{a}}\,\Delta\mathbf{x}. \qquad \square$$

Different coordinate representations of a quadratic form

The corollary leads us to regard a quadratic form as a fixed geometric object that has different representations in different coordinate systems. In other words, if we give a "geometric" vector $\mathbf{v}$ coordinates in two different ways,

$$\Delta\mathbf{x} \longleftrightarrow \mathbf{v} \longleftrightarrow \Delta\mathbf{u},$$

then there is a function $\mathbf{Q}$ (the "geometric" quadratic form) defined on such vectors for which

$$\widehat{Q}_K(\Delta\mathbf{x}) = \mathbf{Q}(\mathbf{v}) = Q_K(\Delta\mathbf{u}).$$

What properties do Q_K and $\widehat{Q}_K$ have in common? These are the geometric properties of the underlying function $\mathbf{Q}$.

Definition 7.9 *We say the quadratic form Q is **negative definite** (respectively, **positive definite**) on a set S in $\mathbb{R}^n$ if $Q(\mathbf{v}) < 0$ (respectively, $Q(\mathbf{v}) > 0$) for every $\mathbf{v} \neq \mathbf{0}$ in S.*

Definition 7.10 *The **index** of the quadratic form Q is the maximum dimension of a subspace N of $\mathbb{R}^n$ on which Q is negative definite.*

The index is a geometric property

The index of a quadratic form is defined without reference to its representation in a particular coordinate system, so it is a geometric property of the form.

Theorem 7.22. *The index of the quadratic form*

$$Q_K(\Delta\mathbf{u}) = -(\Delta u_1)^2 - \cdots - (\Delta u_s)^2 + (\Delta u_{s+1})^2 + \cdots (\Delta u_n)^2$$

is equal to s.

Proof. Let N^s be the s-dimensional linear subspace of vectors of the form

$$\Delta\mathbf{u} = (\Delta u_1, \ldots, \Delta u_s, 0, \ldots, 0).$$

Then

$$Q_K(\Delta\mathbf{u}) = -(\Delta u_1)^2 - \cdots - (\Delta u_s)^2 < 0$$

for every nonzero $\Delta\mathbf{u}$ in N^s, implying that the index of Q_K is at least equal to s.

We are done if we show that the index of Q_K is at most equal to s. First note that, by a similar argument, Q_K is positive definite on the $(n-s)$-dimensional linear subspace P^{n-s} of vectors of the form

$$\Delta\mathbf{u} = (0, \ldots, 0, \Delta u_{s+1}, \ldots, \Delta u_n).$$

Now suppose the index of Q_K were greater than s. Then there would be a linear subspace N^{s+1} of dimension $s+1$ on which Q_K were negative definite. But then the intersection $P^{n-s} \cap N^{s+1}$ would be a linear subspace of dimension at least 1, and would thus contain nonzero vectors. The value of Q_K on such a vector would be both positive and negative, an absurdity we attribute to the assumption that the index could be greater than s. We reject the assumption and conclude that the index is exactly s. □

Q_K and $\widehat{Q}_K$ both have index r

By definition, the index of a quadratic form is independent of the coordinate representation. Therefore, because $\widehat{Q}_K$ and Q_K represent the same form in different coordinates, $\widehat{Q}_K$ and Q_K must have the same index. Consequently, the Hessian form

$$\widehat{Q}_K(\Delta\mathbf{x}) = \tfrac{1}{2}\Delta\mathbf{x}^\dagger H_{\mathbf{a}}\,\Delta\mathbf{x}$$

must have index s. It remains to show that s equals the number of negative eigenvalues of $H_{\mathbf{a}}$. This involves "transforming $H_{\mathbf{a}}$ to principal axes" (cf. p. 242): using an n-dimensional rotation to reduce the Hessian form $\widehat{Q}_K$ to a sum of squares in which the eigenvalues of $H_{\mathbf{a}}$ appear as coefficients.

Definition 7.11 *An $n \times n$ invertible matrix P is **orthogonal** if its transpose equals its inverse: $P^\dagger = P^{-1}$.*

An orthogonal matrix gets its name from the fact that its columns are mutually orthogonal unit vectors. That is, if we write

$$P^\dagger = \begin{pmatrix} \mathbf{w}_1^\dagger \\ \vdots \\ \mathbf{w}_n^\dagger \end{pmatrix}, \quad P = \begin{pmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_n \end{pmatrix},$$

then the condition $P^\dagger P = I$ implies $\|\mathbf{w}_i\|^2 = \mathbf{w}_i^\dagger\,\mathbf{w}_i = 1$ for every $i = 1, \ldots, n$, and $\mathbf{w}_i^\dagger\,\mathbf{w}_j = 0$ for every $i \neq j$.

Let $\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n$ be the unit n-cube whose edges are the standard basis vectors in $\mathbb{R}^n$. Then

$$P(\mathbf{e}_1 \wedge \cdots \wedge \mathbf{e}_n) = P(\mathbf{e}_1) \wedge \cdots \wedge P(\mathbf{e}_n) = \mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_n,$$

so (Definition 2.5, p. 46) $\operatorname{vol} \mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_n = \det P = \pm 1$ for every orthogonal matrix P.

Definition 7.12 *An orthogonal matrix P is a **rotation** if* $\det P = +1$.

Rotations and orthogonal matrices

Thus, a rotation is an orthogonal matrix that preserves orientation. If P is orthogonal but $\det P = -1$, P can be converted into a rotation by changing the signs of the entries in any one of its columns.

Theorem 7.23 (Principal axes theorem). *If $Q(\mathbf{x}) = \mathbf{x}^\dagger M \mathbf{x}$ is a quadratic form in n variables, then there is a rotation $\mathbf{x} = R\mathbf{u}$ of $\mathbb{R}^n$ that transforms Q into a sum of squares*

$$Q(R\mathbf{u}) = Q^*(\mathbf{u}) = \lambda_1 u_1^2 + \cdots + \lambda_n u_n^2,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of M.

The theorem stated in terms of matrices

Proof. The theorem asserts that, for any $n \times n$ symmetric matrix M, there is a rotation R for which

$$R^{-1} M R = R^\dagger M R = D$$

is a diagonal matrix whose diagonal elements are the eigenvalues of M. We prove the theorem in this form by using mathematical induction on n.

If $n = 1$ there is nothing to do; we can take R to be the 1×1 identity matrix. Now assume that any $(n-1) \times (n-1)$ symmetric matrix can be diagonalized by a suitable rotation on $\mathbb{R}^{n-1}$, and consider an $n \times n$ symmetric matrix M.

Let $\mathbf{u}_1$ be an eigenvector of M, $M\mathbf{u}_1 = \lambda_1 \mathbf{u}_1$, and take $\mathbf{u}_1$ to be a unit vector. Extend $\mathbf{u}_1$ to an orthonormal basis $\{\mathbf{u}_1, \mathbf{w}_1, \ldots, \mathbf{w}_{n-1}\}$ of $\mathbb{R}^n$. We may assume (by changing the sign of $\mathbf{w}_{n-1}$, if necessary) that the n-cube $\mathbf{u_1} \wedge \mathbf{w}_1 \wedge \cdots \wedge \mathbf{w}_{n-1}$ has positive orientation, and thus that the matrix

$$R_1 = \begin{pmatrix} \mathbf{u}_1 & \mathbf{w}_1 & \cdots & \mathbf{w}_{n-1} \end{pmatrix}$$

is a rotation. Then

$$MR_1 = \begin{pmatrix} M\mathbf{u}_1 & M\mathbf{w}_1 & \cdots & M\mathbf{w}_{n-1} \end{pmatrix} = \begin{pmatrix} \lambda_1 \mathbf{u}_1 & M\mathbf{w}_1 & \cdots & M\mathbf{w}_{n-1} \end{pmatrix},$$

$$R_1^\dagger M R_1 = \begin{pmatrix} \lambda_1 \mathbf{u}_1^\dagger \mathbf{u}_1 & \mathbf{u}_1^\dagger M \mathbf{w}_1 & \cdots & \mathbf{u}_1^\dagger M \mathbf{w}_{n-1} \\ \lambda_1 \mathbf{w}_1^\dagger \mathbf{u}_1 & \mathbf{w}_1^\dagger M \mathbf{w}_1 & \cdots & \mathbf{w}_1^\dagger M \mathbf{w}_{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 \mathbf{w}_{n-1}^\dagger \mathbf{u}_1 & \mathbf{w}_{n-1}^\dagger M \mathbf{w}_1 & \cdots & \mathbf{w}_{n-1}^\dagger M \mathbf{w}_{n-1} \end{pmatrix}$$

$$= \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & m_{11}^* & \cdots & m_{1,n-1}^* \\ \vdots & \vdots & \ddots & \vdots \\ 0 & m_{n-1,1}^* & \cdots & m_{n-1,n-1}^* \end{pmatrix}, \quad m_{ij}^* = \mathbf{w}_i^\dagger M \mathbf{w}_j.$$

The zeros appear in the first column because every $\mathbf{w}_i \perp \mathbf{u}_1$, and in the first row because $R_1^\dagger M R_1$ is symmetric. Also,

$$m_{ij}^* = \mathbf{w}_i^\dagger M \mathbf{w}_j = (\mathbf{w}_i^\dagger M \mathbf{w}_j)^\dagger = \mathbf{w}_j^\dagger M \mathbf{w}_i = m_{ji}^*,$$

so $M^* = \left(m_{ij}^*\right)$ is an $(n-1)\times(n-1)$ symmetric matrix.

By the induction hypothesis, there is a rotation R^* that diagonalizes M^*; that is, $(R^*)^\dagger M^* R^* = D^*$. Let

$$R_2 = \begin{pmatrix} 1 & \\ & R^* \end{pmatrix};$$

this is the $n \times n$ matrix with 1 and R^* on the diagonal, and with all off-diagonal elenents not shown equal to zero. Then

$$R_2^\dagger = \begin{pmatrix} 1 & \\ & (R^*)^\dagger \end{pmatrix} \text{ and } R_2^\dagger R_2 = \begin{pmatrix} 1 & \\ & (R^*)^\dagger R^* \end{pmatrix} = \begin{pmatrix} 1 & \\ & I_{n-1} \end{pmatrix} = I_n,$$

the $n \times n$ identity matrix, thus R_2 is orthogonal. Moreover, $\det R_2 = 1 \times \det R^* = 1$, so R_2 is a rotation.

Lemma 7.5. *The matrix $R = R_2 R_1$ is a rotation and diagonalizes M.*

Proof. We know R is a rotation because it is a product of rotations. Moreover,

$$\begin{aligned} R^\dagger M R = R_2^\dagger R_1^\dagger M R_1 R_2 &= R_2^\dagger \begin{pmatrix} \lambda_1 & \\ & M^* \end{pmatrix} R_2 = \begin{pmatrix} 1 & \\ & (R^*)^\dagger \end{pmatrix} \begin{pmatrix} \lambda_1 & \\ & M^* \end{pmatrix} \begin{pmatrix} 1 & \\ & R^* \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \\ & (R^*)^\dagger M^* R^* \end{pmatrix} = \begin{pmatrix} \lambda_1 & \\ & D* \end{pmatrix}; \end{aligned}$$

this is an $n \times n$ diagonal matrix. □

Lemma 7.6. *If M is symmetric, P is orthogonal, and $P^\dagger M P$ is a diagonal matrix with diagonal elements α_i, $i = 1,\ldots,n$, then α_i is an eigenvalue of M and the i-th column of P is a corresponding eigenvector.*

Proof. This is the n-dimensional version of Theorem 7.7, page 241, and has a similar proof. The key is that $P^\dagger M P = D$ implies $MP = PD$ because $P^\dagger = P^{-1}$; see the exercises. □

Thus the diagonal elements in the diagonal matrix of Lemma 7.5 are the eigenvalues of M, and the proof of the principal axes theorem is complete. □

Interlude

Our proof of the principal axes theorem indicates that the eigenvectors of a symmetric matrix have properties not shared by matrices in general. Before returning to the analysis of critical points, we pause to establish some of those properties.

Definition 7.13 *The eigenvalue α of the matrix M has* ***multiplicity k*** *if the factor $\lambda - \alpha$ appears k times in a factorization of the characteristic polynomial of M.*

Repeated eigenvalues and eigenspaces

Although each distinct real eigenvalue of an $n \times n$ real matrix M has a real eigenvector associated with it (Theorem 7.13, p. 246), in general a repeated eigenvalue may not possess additional linearly independent eigenvectors. Consequently, the eigenvectors of such an $n \times n$ matrix may not span $\mathbb{R}^n$. The first such examples we saw were the shears M_7 and M_8 of Chapter 2 (pp. 38ff.). However, a symmetric matrix has a "complete" set of eigenvectors.

Corollary 7.24 *Suppose M is a symmetric matrix with an eigenvalue α of multiplicity k. Then the eigenvectors associated with α form a k-dimensional subspace E_α of $\mathbb{R}^n$.*

Proof. Sums and scalar multiples of eigenvectors associated with α are again eigenvectors associated with α, so they form a subspace E_α of $\mathbb{R}^n$. Because α has multiplicity k, precisely k columns of the orthogonal matrix P of Lemma 7.6 are eigenvectors associated with α. Those columns are linearly independent, so the dimension of E_α is at least k.

We must now show the dimension of E_α is not greater than k. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the columns of P; assume, by rearranging them if necessary, that the first k columns, $\mathbf{v}_1, \ldots, \mathbf{v}_k$, are the eigenvectors associated with α. Suppose $\mathbf{w}$ is in E_α; because $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is an orthonormal basis, we can write

$$\mathbf{w} = v_1\mathbf{v}_1 + \cdots + v_n\mathbf{v}_n,$$

where $v_j = \mathbf{w}^\dagger \mathbf{v}_j = \mathbf{v}_j^\dagger \mathbf{w}$ by orthonormality of the basis. Let α_j be the eigenvector associated with $\mathbf{v}_j$; then $\alpha_j = \alpha$ if and only if $j = 1, \ldots, k$. We have

$$\alpha_j v_j = \alpha_j \mathbf{w}^\dagger \mathbf{v}_j = \mathbf{w}^\dagger M\mathbf{v}_j = (\mathbf{w}^\dagger M\mathbf{v}_j)^\dagger = \mathbf{v}_j^\dagger M\mathbf{w} = \alpha \mathbf{v}_j^\dagger \mathbf{w} = \alpha v_j;$$

$M\mathbf{w} = \alpha\mathbf{w}$ because $\mathbf{w}$ is in E_α. Thus $(\alpha_j - \alpha)v_j = 0$. This forces $v_j = 0$ for $j > k$, implying that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ spans E_α. Hence $\dim E_\alpha = k$. □

The subspace E_α is sometimes called the **eigenspace** associated with α. Thus, for a *symmetric* matrix, the dimension of the eigenspace associated with an eigenvalue equals the multiplicity of that eigenvalue.

Sylvester's *law of inertia*

To complete the third part of the proof of Morse's lemma, we must show that, whenever a coordinate change reduces Δz to a sum of squares, the number of negative squares in that sum is always equal to the number of negative eigenvalues of the Hessian matrix $H_{\mathbf{a}}$. For a quadratic form with constant coefficients under a linear coordinate change, the invariance of the number of negative squares and positive squares was shown by J. J. Sylvester in 1852 [18]. He characterized the result as "... a law to which my view of the physical meaning of quantity of matter inclines me, upon the ground of analogy, to give the name of the Law of Inertia for Quadratic Forms, as expressing the fact of the existence of an invariable number inseparably attached to such forms."

Index of inertia

Sometimes, to underscore the invariant nature of the index of a quadratic form, we add Sylvester's term and call it the **index of inertia**.

Theorem 7.25. *Suppose $z = f(\mathbf{x})$ has continuous third derivatives on an open set X^n, $\mathbf{x} = \mathbf{a}$ is a nondegenerate critical point of f in X^n, and the Hessian matrix $H_\mathbf{a}$ of f at $\mathbf{a}$ has r negative eigenvalues. If $\Delta\mathbf{u} = \mathbf{h}(\Delta\mathbf{x})$ is a coordinate change in a window centered at $\mathbf{a}$ that reduces Δz to a sum of squares,*

$$\Delta z = -(\Delta u_1)^2 - \cdots - (\Delta u_s)^2 + (\Delta u_{s+1})^2 + \cdots + (\Delta u_n)^2,$$

then $s = r$.

Proof. By Theorem 7.20 (p. 258), the linear map $\Delta\mathbf{u} = \mathrm{d}\mathbf{h}_\mathbf{0}(\Delta\mathbf{x})$ converts the quadratic form

$$\Delta z = Q(\Delta\mathbf{u}) = -(\Delta u_1)^2 - \cdots - (\Delta u_s)^2 + (\Delta u_{s+1})^2 + \cdots + (\Delta u_n)^2$$

into

$$\Delta z = \widehat{Q}(\Delta\mathbf{x}) = \tfrac{1}{2}\Delta\mathbf{x}^\dagger H_\mathbf{a}\,\Delta\mathbf{x}.$$

Therefore, Q and $\widehat{Q}$ are just different coordinate representations of the same (geometric) quadratic form $\mathbf{Q}$, and must therefore have the same index. By Theorem 7.22, the index of Q is s. By transforming $\widehat{Q}$ to principal axes (Theorem 7.23), we see the index of $\widehat{Q}$ is r. Thus $s = r$. □

Proof is complete

This completes the third part of the proof of Morse's lemma, and thus completes the entire proof. □

One of the consequences of Morse's lemma is that the second derivatives of a function at a nondegenerate critical point determine the type of that point. In fact, the type is completely characterized by a single number: the *index of inertia* of its Hessian form. This leads to the following definition and theorem.

Index of a critical point

Definition 7.14 *The **index**, or **index of inertia**, of a nondegenerate critical point of a function is the index of its Hessian, that is, the number of negative eigenvalues of the Hessian matrix at that point.*

Theorem 7.26 (Second derivative test). *Suppose $\mathbf{x} = \mathbf{a}$ is a nondegenerate critical point of a function $z = f(\mathbf{x})$ that possesses continuous third derivatives. If r is the index of $\mathbf{a}$, then*

- $\mathbf{a}$ *is a **local minimum** if $r = 0$.*
- $\mathbf{a}$ *is a **local maximum** if $r = n$.*
- $\mathbf{a}$ *is a **saddle** if $0 < r < n$.* □

Classifying critical points

Morse's lemma and the second derivative test classify nondegenerate critical points: there are $n+1$ classes, one for each possible index. Two critical points are in the same class if a coordinate change will transform one into the other. For degenerate critical points, the situation is very different. There are infinitely many classes, and no complete classification exists, although there are partial results. The analysis of (degenerate) critical points is part of the larger study of *singularities of mappings*, an active area of current research.

Nonisolated critical points are degenerate

One useful observation we can make is that a nonisolated critical point—for example, a point on the ring of minima of the wine bottle—is necessarily degenerate. The proof (by Morse) is a nice application of the inverse function theorem.

Theorem 7.27. *Suppose* $\mathbf{x} = \mathbf{a}$ *is a nondegenerate critical point of a function* $z = f(\mathbf{x})$ *that has continuous second derivatives in some neighborhood* X^n *of* $\mathbf{a}$*. Then* $\mathbf{a}$ *is isolated, in the sense that there is some nonempty open ball* B_ε *centered at* $\mathbf{a}$ *that contains no other critical point of* f*.*

Proof. The gradient of f defines a map $\nabla f : X^n \to \mathbb{R}^n$,

$$\nabla f : \begin{cases} u_1 = \dfrac{\partial f}{\partial x_1}(\mathbf{x}), \\ \quad \vdots \\ u_n = \dfrac{\partial f}{\partial x_n}(\mathbf{x}), \end{cases}$$

that is continuously differentiable, because f is twice continuously differentiable.

By construction, a point $\mathbf{b}$ is a critical point of f if and only if $\nabla f(\mathbf{b}) = \mathbf{0}$; in particular, $\nabla f(\mathbf{a}) = \mathbf{0}$. Furthermore, the matrix of the derivative $d(\nabla f)_\mathbf{a}$ coincides with the Hessian matrix $H_\mathbf{a}$, so the nondegeneracy of $\mathbf{a}$ implies that $d(\nabla f)_\mathbf{a}$ is invertible. The inverse function theorem then implies that the map ∇f itself is invertible on some open ball B_ε centered at $\mathbf{a}$. In particular, ∇f is 1–1 there, so no point $\mathbf{b} \neq \mathbf{a}$ is mapped to $\mathbf{0}$. That is, no point $\mathbf{b} \neq \mathbf{a}$ in B_ε is a critical point of f. □

Volatility of the second derivatives

Earlier (see pp. 222–224), we observed that the value of the second derivative at a regular point of a single-variable function could be transformed into any new value whatsoever by a suitable coordinate change. At a critical point, this degree of volatility does not occur: the sign of the second derivative cannot be changed. In effect, the convexity of a function graph is a geometric invariant at a critical point but not at a regular point. There is a similar distinction between the regular and the critical points of a function of several variables. For suppose $\mathbf{x} = \mathbf{a}$ is a regular point of $z = f(\mathbf{x})$. By the implicit function theorem (in particular, Corollary 6.8, p. 198), local coordinates $(\Delta u_1, \ldots, \Delta u_n)$ can be chosen near $\mathbf{a}$ so that $\Delta z = \Delta u_n$. Thus, in terms of the new variables, the function is linear, and all of its second derivatives are identically zero. Whatever information we thought might be conveyed by the original derivatives $\partial^2 f / \partial x_i \, \partial x_j$ has vanished with the coordinate change.

By contrast, suppose $\mathbf{x} = \mathbf{a}$ is a critical point of $z = f(\mathbf{x})$. When $\mathbf{a}$ is nondegenerate, Morse's lemma tells us the index of inertia of the Hessian of f at $\mathbf{a}$ is a geometric invariant. That is, if $\mathbf{x} = \mathbf{h}(\mathbf{u})$ is a local coordinate change near $\mathbf{a} = \mathbf{h}(\mathbf{b})$ that transforms f into $g(\mathbf{u}) = f(\mathbf{h}(\mathbf{u})) = f(\mathbf{x})$, then

- $\mathbf{u} = \mathbf{b}$ is a nondegenerate critical point of $z = g(\mathbf{u})$.
- The index of inertia of the Hessian of g at $\mathbf{b}$ equals the index of inertia of f at $\mathbf{a}$.

If $\mathbf{a}$ is degenerate, then the rank of the Hessian is not maximal. In this case, though, an extension of our methods can show that the rank and the index of inertia are both geometric invariants.

Exercises

7.1. Suppose $y = f(x)$ has a continuous second derivative and $f'(0) \neq 0$.

a. For any value of B, the function $x = h_B(u) = u + Bu^2$ is a coordinate change near the origin; explain why.

b. Let $g(u) = f(h_B(u))$ represent f under the coordinate change. Show that B can be chosen so that $g''(0) = A$, where A is an arbitrary number. Write a formula that expresses how B depends on A.

7.2. Construct the symmetric matric that corresponds to each of the following quadratic forms.

a. $Q(x,y) = 5x^2 + 18xy - 2y^2$.

b. $Q(x,y) = xy - x^2 - y^2$

c. $Q(x,y) = (2x-y)(2y-x)$;

d. $Q(x,y) = \begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} 1 & 5 \\ -1 & -5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$.

7.3. Show that, when m is small, the roots of $4x^3 - 4x + m = 0$ (cf. p. 232) are approximately $m/4$ and $\pm 1 - m/8$. Specifically, you can show that

$$\frac{\partial f_m}{\partial x}\left(\pm 1 - \frac{m}{8}, 0\right) = O(m^2), \qquad \frac{\partial f_m}{\partial x}\left(\frac{m}{4}, 0\right) = O(m^3).$$

7.4. Verify that the derivative $d\mathbf{h}_{(0,0)}$ of the coordinate change map given on page 234 has the form shown on page 235.

7.5. Construct the symmetric matric that corresponds to each of the following quadratic forms.

a. $Q(x,y,z) = 10xy - 2yz + zx$.

b. $Q(x,y,z) = (x - y + z)(x + y - z)$.

c. $Q(x_1,\ldots,x_n) = \sum_{i=1}^{n} (i-5)x_i^2$.

d. $Q(x_1,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} (i+j)x_i x_j$.

e. $Q(x_1,\ldots,x_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} (i-j)x_i x_j$.

7.6. Let $f(x,y) = 3x^2 - x^3 - y^2$.

a. Verify that $(0,0)$ and $(2,0)$ are critical points of f.

b. Find the second-order Taylor polynomial for f at $(2,0)$; call it $P(x,y)$. Graph together $f(x,y)$ and $P(x,y)$ on a small neighborhood of $(2,0)$; specifically, use $1.9 \leq x \leq 2.1$, $-0.1 \leq y \leq 0.1$.

c. Does P have a critical point at $(2,0)$? What kind? Do the graphs show that P and f have the same type of critical point at $(2,0)$? What kind of critical point does f have at $(2,0)$?

d. Find the second-order Taylor polynomial for f at $(0,0)$; call it $Q(x,y)$. Graph together $f(x,y)$ and $Q(x,y)$ on a small neighborhood of $(0,0)$; specifically, use $-0.1 \leq x \leq 0.1$, $-0.1 \leq y \leq 0.1$.

e. What kind of critical point does Q have at $(0,0)$. Do the graphs show that Q and f have the same type of critical point at $(0,0)$? What kind of critical point does f have at $(0,0)$?

7.7. a. Find all critical points of $f(x,y) = x^3 + y^3 - 3x - 12y$.
b. At each critical point P of f, construct the second-order Taylor polynomial T_P of f. Does $T_P(x,y)$ also have a critical point at P? What kind?
c. In a small neighborhood of each of the critical points P, sketch the graph of f together with the Taylor polynomial T_P. Does f resemble T_P near P? Is P the same type of critical point for f that it is for T_P?
d. Conclusion: List the critical points of f, and indicate the type of each.

7.8. a. Find all critical points of $f(x,y) = x^3 - 3xy^2 - x^2 + 3y^2$.
b. At each critical point P of f, construct the second-order Taylor polynomial T_P of f. Does $T_P(x,y)$ also have a critical point at P? What kind?
c. In a small neighborhood of each of the critical points P, sketch the graph of f together with the Taylor polynomial T_P. Does f resemble T_P near P? Is P the same type of critical point for f that it is for T_P?
d. Conclusion: List the critical points of f, and indicate the type of each.

7.9. Locate the critical point of $Q(x,y) = ax^2 + 2bxy + cy^2 + dx + ey + k$ and determine its type. On which of the six parameters does the location depend, and on which does the type depend?

7.10. Locate all the critical points of $\Phi(\theta, v) = 1 - \cos\theta + \frac{1}{2}v^2$, and determine the type of each.

7.11. Let $z = f(x,y) = p^2x^2 + q^2y^2$, $0 < p^2 < q^2$, and let $D_a(x,y)$ be the square of the distance from the point $(0,0,a)$ on the z-axis to the point $(x,y,f(x,y))$ on the graph of f.

a. Make a sketch.
b. Show that D_a has a critical point at the origin, for every a.
c. For two values of a, that critical point is degenerate; determine those values.
d. At all other points a, determine how the type of the critical point depends on a.

7.12. Show that the formula originally used for the quadratic terms in Taylor's expansion (see Theorem 3.18, p. 94, and the discussion leading up to it) gives the same value as the formula using the Hessian. That is, show

$$(\Delta\mathbf{u}\cdot\nabla)^2 f(\mathbf{a}) = (\Delta\mathbf{u})^\dagger H_{\mathbf{a}}\,\Delta\mathbf{u},$$

where ∇ is the gradient differential operator.

7.13. Let M be an $n \times n$ real symmetric matrix, and let λ be an eigenvalue of M. The purpose of this exercise is to prove that λ is real. So suppose the contrary; let $\lambda = a + b\mathrm{i}$ (with $b \neq 0$), and let $Z = X + \mathrm{i}Y$ (where X and Y are real $n \times 1$ vectors) be a complex eigenvector for λ: $MZ = \lambda Z$ and $Z \neq 0$.

a. Let $\overline{\lambda} = a - b\mathrm{i}$ be the complex conjugate of λ, and let $\overline{Z} = X - \mathrm{i}Y$ be the complex conjugate of Z. Show that $\overline{\lambda}$ is also an eigenvalue of M (Hint: What is $\overline{M}$?) with eigenvector $\overline{Z}$.

b. Show that the 1×1 matrix $\overline{Z}^{\dagger} MZ$ equals $\lambda \|Z\|^2$.

c. The **conjugate transpose** of the $k \times l$ matrix A is the $l \times k$ matrix $\overline{A}^{\dagger}$. Show that the conjugate transpose of $\overline{Z}^{\dagger} MZ$ equals $\overline{\lambda}\|Z\|^2$.

d. Compare $\overline{Z}^{\dagger} MZ$ and its conjugate transpose to conclude that $\overline{\lambda} = \lambda$, showing λ is real.

7.14. Let M be an $n \times n$ real symmetric matrix. The pupose of this exercise is to show that eigenvectors of different eigenvalues must be orthogonal. So suppose X_1 is an eigenvector of M with eigenvalue λ_1 and X_2 is an eigenvector with eigenvalue $\lambda_2 \neq \lambda_1$.

a. Show that $X_2^{\dagger} M X_1 = \lambda_1 (X_2 \cdot X_1)$, where $X_2 \cdot X_1$ is the ordinary "dot product" of vectors.

b. Use $(MX_2)^{\dagger} = X_2^{\dagger} M$ (why is this true?) to show that $X_2^{\dagger} M X_1 = \lambda_2 (X_2 \cdot X_1)$. Conclude that $X_2 \cdot X_1 = 0$.

7.15. a. Find the functions $p_i(\Delta x, \Delta y)$, $i = 1, 2$ provided by Lemma 7.3 for the function $F(x,y) = e^x \sin y$ when $(a,b) = (0,0)$.

b. Verify that $e^x \sin y = p_1(x,y)x + p_2(x,y)y$.

7.16. The folium of Descartes $f(x,y)$ (p. 237) evidently has a saddle point at the origin. This exercise provides new local coordinates (u,v) in which the folium takes the form $-u^2 + v^2$.

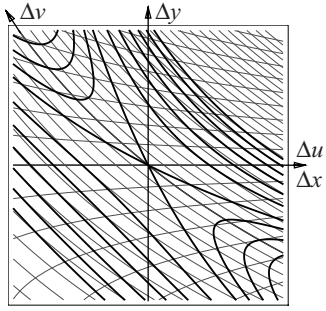
a. Determine $\varphi(\xi, \eta) = f(\xi - \eta, \xi + \eta)$; this is the form the folium takes under a (global) 45° rotation and dilation $\mathbf{c}(\xi, \eta)$.

b. Show that φ can be writen in the form $\alpha(\xi)\xi^2 + \beta(\xi)\eta^2$ and determine $\alpha(\xi)$ and $\beta(\xi)$.

c. Introduce a local coordinate change $(u,v) = \mathbf{k}(\xi, \eta)$ near $(\xi, \eta) = (0,0)$ that reduces φ to $-u^2 + v^2$. Prove that $\mathbf{k}$ is a coordinate change near the origin.

d. Let $\mathbf{h} = \mathbf{k} \circ \mathbf{c}^{-1}$. Use a suitable graphing uility to sketch the pullback of a coordinate grid in the (u,v)-plane by $\mathbf{h}$ to show that the pullback carries level curves of $-u^2 + v^2$ to level curves of $f(x,y)$. Compare your result with the figure on page 239.

7.17. a. Sketch the zero-level of the function $f(x,y) = (xy^2 - 1)(x^2 y - 1)$ in the first quadrant and infer that f has a saddle point at $(x,y) = (1,1)$.



b. Express f in terms of window coordinates at $(1,1)$; that is, determine $\Delta z = f(1+\Delta x, 1+\Delta y) - f(1,1)$ as a (sixth-degree) polynomial in Δx and Δy.

c. Show that the functions of Morse's decomposition (Theorem 7.18) at the saddle point are

$$\begin{aligned} h_{11} &= 2 + \Delta x + \tfrac{8}{3}\Delta y + \tfrac{3}{2}\Delta x\Delta y + \tfrac{3}{2}\Delta y^2 + \tfrac{9}{10}\Delta x\Delta y^2 + \tfrac{3}{10}\Delta y^3 + \tfrac{1}{5}\Delta x\Delta y^3, \\ h_{12} &= \tfrac{5}{2} + \tfrac{8}{3}\Delta x + \tfrac{8}{3}\Delta y + \tfrac{3}{4}\Delta x^2 + 3\Delta x\Delta y + \tfrac{3}{4}\Delta y^2 + \tfrac{9}{10}\Delta x^2\Delta y + \tfrac{9}{10}\Delta x\Delta y^2 \\ &\quad + \tfrac{3}{10}\Delta x^2\Delta y^2, \\ h_{22} &= 2 + \tfrac{8}{3}\Delta x + \Delta y + \tfrac{3}{2}\Delta x^2 + \tfrac{3}{2}\Delta x\Delta y + \tfrac{3}{10}\Delta x^3 + \tfrac{9}{10}\Delta x^2\Delta y + \tfrac{1}{5}\Delta x^3\Delta y. \end{aligned}$$

d. Verify, by direct computation, that $\Delta z = h_{11}\Delta x^2 + 2h_{12}\Delta x\Delta y + h_{22}\Delta y^2$.

e. Complete the square to obtain the coordinate change $(\Delta u, \Delta v) = \mathbf{h}(\Delta x, \Delta y)$ that reduces Δz to the simple diagonal form $\Delta z = \Delta u^2 - \Delta v^2$. Prove that $\mathbf{h}$ is a coordinate change near the origin in the $(\Delta x, \Delta y)$ window.

f. Use a suitable graphing utility to sketch the pullback of a coordinate grid in the $(\Delta u, \Delta v)$ window to show that level curves of $\Delta u^2 - \Delta v^2$ pull back to level curves of Δz in the $(\Delta x, \Delta y)$ window. The figure in the margin shows the $(\Delta u, \Delta v)$ coordinate grid in the $(\Delta x, \Delta y)$ window, together with level curves of f. (Levels in the Δu direction are twice as far apart as those in the Δv direction.)

7.18. a. The function $g(x,y) = (x^2 - y^2)^3 - 2x(x^2 - y^2) + 1$ has a saddle point at $(x,y) = (1,0)$. (A 45° rotation–dilation converts the function f of the preceding exercise into g.) Carry out all the steps of the preceding exercise to obtain the local coordinate change $(\Delta u, \Delta v) = \mathbf{h}(\Delta x, \Delta y)$ given by Morse's decomposition that reduces g to $\Delta u^2 - \Delta v^2$ in a window centered at $(1,0)$.

b. Prove that the local coordinate change provided by Morse's decomposition in this exercise is not a rotation–dilation of the local coordinate change of the preceding exercise. Suggestion: Consider the derivative of each coordinate change at the origin.

7.19. Find the functions $h_{ij}(x_1, x_2)$ in Morse's decomposition (Theorem 7.18) for the function $f(x_1, x_2) = \cos x_1 \cos x_2$ at the point $(a_1, a_2) = (0,0)$. Verify that

$$h_{ij}(0,0) = \frac{1}{2}\frac{\partial^2 f}{\partial x_i \partial x_j}(0,0)$$

for every i, j.