8 Performance Surfaces and Optimum Points

Objectives

This chapter lays the foundation for a type of neural network training technique called performance learning. There are several different classes of network learning laws, including associative learning (as in the Hebbian learning of Chapter 7) and competitive learning (which we will discuss in Chapter 16). Performance learning is another important class of learning law, in which the network parameters are adjusted to optimize the performance of the network. In the next two chapters we will lay the groundwork for the development of performance learning, which will then be presented in detail in Chapters 10–14. The main objective of the present chapter is to investigate performance surfaces and to determine conditions for the existence of minima and maxima of the performance surface. Chapter 9 will follow this up with a discussion of procedures to locate the minima or maxima.

Theory and Examples

Performance Learning

There are several different learning laws that fall under the category of *performance learning*. Two of these will be presented in this text. These learning laws are distinguished by the fact that during training the network parameters (weights and biases) are adjusted in an effort to optimize the "performance" of the network.

Performance Index

There are two steps involved in this optimization process. The first step is to define what we mean by "performance." In other words, we must find a quantitative measure of network performance, called the *performance index*, which is small when the network performs well and large when the network performs poorly. In this chapter, and in Chapter 9, we will assume that the performance index is given. In Chapters 10, 11 and 13 we will discuss the choice of performance index.

The second step of the optimization process is to search the parameter space (adjust the network weights and biases) in order to reduce the performance index. In this chapter we will investigate the characteristics of performance surfaces and set some conditions that will guarantee that a surface does have a minimum point (the optimum we are searching for). Thus, in this chapter we will obtain some understanding of what performance surfaces look like. Then, in Chapter 9 we will develop procedures for locating the optimum points.

Taylor Series

Taylor Series Expansion

Let us say that the performance index that we want to minimize is represented by $F(x)$, where x is the scalar parameter we are adjusting. We will assume that the performance index is an analytic function, so that all of its derivatives exist. Then it can be represented by its *Taylor series expansion* about some nominal point x^*:

$$\begin{aligned} F(x) = {} & F(x^*) + \frac{d}{dx}F(x)\Big|_{x = x^*}(x - x^*) \\ & + \frac{1}{2}\frac{d^2}{dx^2}F(x)\bigg|_{x = x^*}(x - x^*)^2 + \cdots \\ & + \frac{1}{n!}\frac{d^n}{dx^n}F(x)\bigg|_{x = x^*}(x - x^*)^n + \cdots \end{aligned} \tag{8.1}$$

We will use the Taylor series expansion to approximate the performance index, by limiting the expansion to a finite number of terms. For example, let

$$F(x) = \cos(x) . \tag{8.2}$$

The Taylor series expansion for $F(x)$ about the point $x^* = 0$ is

$$F(x) = \cos(x) = \cos(0) - \sin(0)(x-0) - \frac{1}{2}\cos(0)(x-0)^2$$
$$+ \frac{1}{6}\sin(0)(x-0)^3 + \cdots$$
$$= 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 + \cdots \tag{8.3}$$

The zeroth-order approximation of $F(x)$ (using only the zeroth power of x) is

$$F(x) \approx F_0(x) = 1 . \tag{8.4}$$

The second-order approximation is

$$F(x) \approx F_2(x) = 1 - \frac{1}{2}x^2 . \tag{8.5}$$

(Note that in this case the first-order approximation is the same as the zeroth-order approximation, since the first derivative is zero.)

The fourth-order approximation is

$$F(x) \approx F_4(x) = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4 . \tag{8.6}$$

A graph showing $F(x)$ and these three approximations is shown in Figure 8.1.

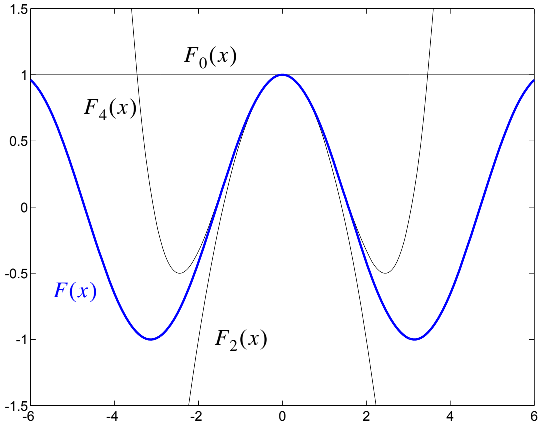


Figure 8.1 Cosine Function and Taylor Series Approximations

From the figure we can see that all three approximations are accurate if x is very close to $x^* = 0$. However, as x moves farther away from x^* only the higher-order approximations are accurate. The second-order approximation is accurate over a wider range than the zeroth-order approximation, and the fourth-order approximation is accurate over a wider range than the second-order approximation. An investigation of Eq. (8.1) explains this behavior. Each succeeding term in the series involves a higher power of $(x - x^*)$. As x gets closer to x^*, these terms will become geometrically smaller.

We will use the Taylor series approximations of the performance index to investigate the shape of the performance index in the neighborhood of possible optimum points.

To experiment with Taylor series expansions of the cosine function, use the MATLAB® Neural Network Design Demonstration Taylor Series (`nnd8ts1`).

Vector Case

Of course the neural network performance index will not be a function of a scalar x. It will be a function of all of the network parameters (weights and biases), of which there may be a very large number. Therefore, we need to extend the Taylor series expansion to functions of many variables. Consider the following function of n variables:

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n). \tag{8.7}$$

The Taylor series expansion for this function, about the point x^*, will be

$$\begin{aligned} F(\mathbf{x}) = F(\mathbf{x}^*) &+ \frac{\partial}{\partial x_1}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*) + \frac{\partial}{\partial x_2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_2 - x_2^*) \\ &+ \cdots + \frac{\partial}{\partial x_n}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_n - x_n^*) + \frac{1}{2}\frac{\partial^2}{\partial x_1^2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*)^2 \\ &+ \frac{1}{2}\frac{\partial^2}{\partial x_1 \partial x_2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*)(x_2 - x_2^*) + \cdots \end{aligned} \tag{8.8}$$

This notation is a bit cumbersome. It is more convenient to write it in matrix form, as in:

$$\begin{aligned} F(\mathbf{x}) = F(\mathbf{x}^*) &+ \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) \\ &+ \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) + \cdots \end{aligned} \tag{8.9}$$

Gradient where $\nabla F(\mathbf{x})$ is the *gradient*, and is defined as

$$\nabla F(\mathbf{x}) = \left[\frac{\partial}{\partial x_1} F(\mathbf{x}) \; \frac{\partial}{\partial x_2} F(\mathbf{x}) \; \dots \; \frac{\partial}{\partial x_n} F(\mathbf{x}) \right]^T , \tag{8.10}$$

Hessian

and $\nabla^2 F(\mathbf{x})$ is the *Hessian*, and is defined as:

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} F(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_2 \partial x_n} F(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_n \partial x_2} F(\mathbf{x}) & \dots & \frac{\partial^2}{\partial x_n^2} F(\mathbf{x}) \end{bmatrix} . \tag{8.11}$$

The gradient and the Hessian are very important to our understanding of performance surfaces. In the next section we discuss the practical meaning of these two concepts.

To experiment with Taylor series expansions of a function of two variables, use the MATLAB® Neural Network Design Demonstration Vector Taylor Series(`nnd8ts2`).

Directional Derivatives

The ith element of the gradient, $\partial F(\mathbf{x}) / \partial x_i$, is the first derivative of the performance index F along the x_i axis. The ith element of the diagonal of the Hessian matrix, $\partial^2 F(\mathbf{x}) / \partial x_i^2$, is the second derivative of the performance index F along the x_i axis. What if we want to know the derivative of the function in an arbitrary direction? We let $\mathbf{p}$ be a vector in the direction along which we wish to know the derivative. This *directional derivative* can be computed from

Directional Derivative

$$\frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|} . \tag{8.12}$$

The second derivative along $\mathbf{p}$ can also be computed:

$$\frac{\mathbf{p}^T \nabla^2 F(\mathbf{x}) \mathbf{p}}{\|\mathbf{p}\|^2} . \tag{8.13}$$

To illustrate these concepts, consider the function

$$F(\mathbf{x}) = x_1^2 + 2x_2^2 . \tag{8.14}$$

Suppose that we want to know the derivative of the function at the point $\mathbf{x}^* = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}^T$ in the direction $\mathbf{p} = \begin{bmatrix} 2 & -1 \end{bmatrix}^T$. First we evaluate the gradient at $\mathbf{x}^*$:

$$\nabla F(\mathbf{x})\Big|_{\mathbf{x} = \mathbf{x}^*} = \left.\begin{bmatrix} \frac{\partial}{\partial x_1} F(\mathbf{x}) \\ \frac{\partial}{\partial x_2} F(\mathbf{x}) \end{bmatrix}\right|_{\mathbf{x} = \mathbf{x}^*} = \left.\begin{bmatrix} 2x_1 \\ 4x_2 \end{bmatrix}\right|_{\mathbf{x} = \mathbf{x}^*} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} . \tag{8.15}$$

The derivative in the direction $\mathbf{p}$ can then be computed:

$$\frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|} = \frac{\begin{bmatrix} 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}}{\left\| \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\|} = \frac{\begin{bmatrix} 0 \end{bmatrix}}{\sqrt{5}} = 0 . \tag{8.16}$$

Therefore the function has zero slope in the direction $\mathbf{p}$ from the point $\mathbf{x}^*$. Why did this happen? What can we say about those directions that have zero slope? If we consider the definition of directional derivative in Eq. (8.12), we can see that the numerator is an inner product between the direction vector and the gradient. Therefore any direction that is orthogonal to the gradient will have zero slope.

Which direction has the greatest slope? The maximum slope will occur when the inner product of the direction vector and the gradient is a maximum. This happens when the direction vector is the same as the gradient. (Notice that the magnitude of the direction vector has no effect, since we normalize by its magnitude.) This effect is illustrated in Figure 8.2, which shows a contour plot and a 3-D plot of $F(\mathbf{x})$. On the contour plot we see five vectors starting from our nominal point $\mathbf{x}^*$ and pointing in different directions. At the end of each vector the first directional derivative is displayed. The maximum derivative occurs in the direction of the gradient. The zero derivative is in the direction orthogonal to the gradient (tangent to the contour line).

To experiment with directional derivatives, use the MATLAB® Neural Network Design Demonstration Directional Derivatives (`nnd8dd`).

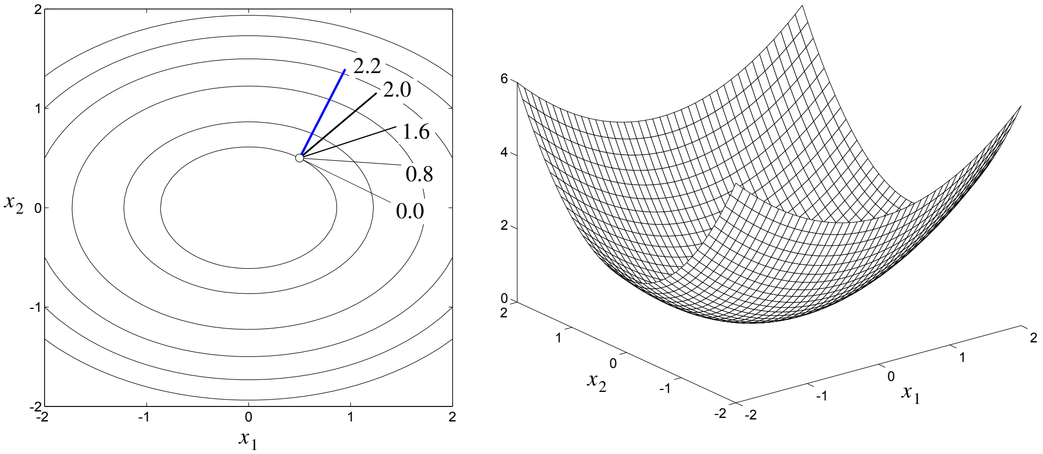


Figure 8.2 Quadratic Function and Directional Derivatives

Minima

Recall that the objective of performance learning will be to optimize the network performance index. In this section we want to define what we mean by an optimum point. We will assume that the optimum point is a minimum of the performance index. The definitions can be easily modified for maximization problems.

Strong Minimum

Strong Minimum

The point $\mathbf{x}^*$ is a strong minimum of $F(\mathbf{x})$ if a scalar $\delta > 0$ exists, such that $F(\mathbf{x}^*) < F(\mathbf{x}^* + \Delta\mathbf{x})$ for all $\Delta\mathbf{x}$ such that $\delta > \|\Delta\mathbf{x}\| > 0$.

In other words, if we move away from a strong minimum a small distance in *any* direction the function will increase.

Global Minimum

Global Minimum

The point $\mathbf{x}^*$ is a unique global minimum of $F(\mathbf{x})$ if $F(\mathbf{x}^*) < F(\mathbf{x}^* + \Delta\mathbf{x})$ for all $\Delta\mathbf{x} \neq \mathbf{0}$.

For a simple strong minimum, $\mathbf{x}^*$, the function may be smaller than $F(\mathbf{x}^*)$ at some points outside a small neighborhood of $\mathbf{x}^*$. Therefore this is sometimes called a local minimum. For a global minimum the function will be larger than the minimum point at every other point in the parameter space.

Weak Minimum

Weak Minimum

The point $\mathbf{x}^*$ is a weak minimum of $F(\mathbf{x})$ if it is not a strong minimum, and a scalar $\delta > 0$ exists, such that $F(\mathbf{x}^*) \le F(\mathbf{x}^* + \Delta\mathbf{x})$ for all $\Delta\mathbf{x}$ such that $\delta > \|\Delta\mathbf{x}\| > 0$.

No matter which direction we move away from a weak minimum, the function cannot decrease, although there may be some directions in which the function does not change.

As an example of local and global minimum points, consider the following scalar function:

$$F(x) = 3x^4 - 7x^2 - \frac{1}{2}x + 6 . \tag{8.17}$$

This function is displayed in Figure 8.3. Notice that it has two strong minimum points: at approximately -1.1 and 1.1. For both of these points the function increases in a local neighborhood. The minimum at 1.1 is a global minimum, since there is no other point for which the function is as small.

There is no weak minimum for this function. We will show a two-dimensional example of a weak minimum later.

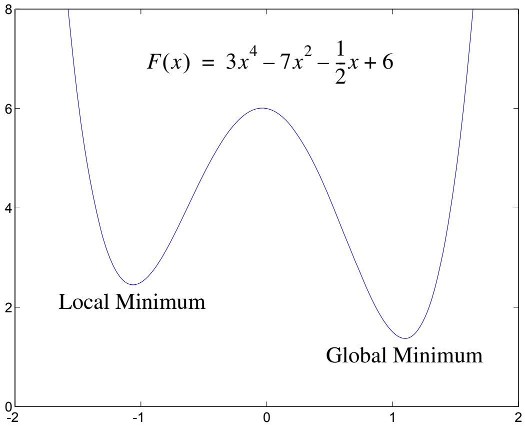


Figure 8.3 Scalar Example of Local and Global Minima

Now let's consider some vector cases. First, consider the following function:

$$F(\mathbf{x}) = (x_2 - x_1)^4 + 8x_1x_2 - x_1 + x_2 + 3 . \tag{8.18}$$

Contour Plot

In Figure 8.4 we have a *contour plot* (a series of curves along which the function value remains constant) and a 3-D surface plot for this function (for function values less than 12). We can see that the function has two strong local minimum points: one at (-0.42, 0.42), and the other at (0.55, -0.55). The global minimum point is at (0.55, -0.55).

Saddle Point

There is also another interesting feature of this function at (-0.13, 0.13). It is called a *saddle point* because of the shape of the surface in the neighborhood of the point. It is characterized by the fact that along the line $x_1 = -x_2$ the saddle point is a local maximum, but along a line orthogonal to that line it is a local minimum. We will investigate this example in more detail in Problems P8.2 and P8.5.

This function is used in the MATLAB® Neural Network Design Demonstration Vector Taylor Series (`nnd8ts2`).

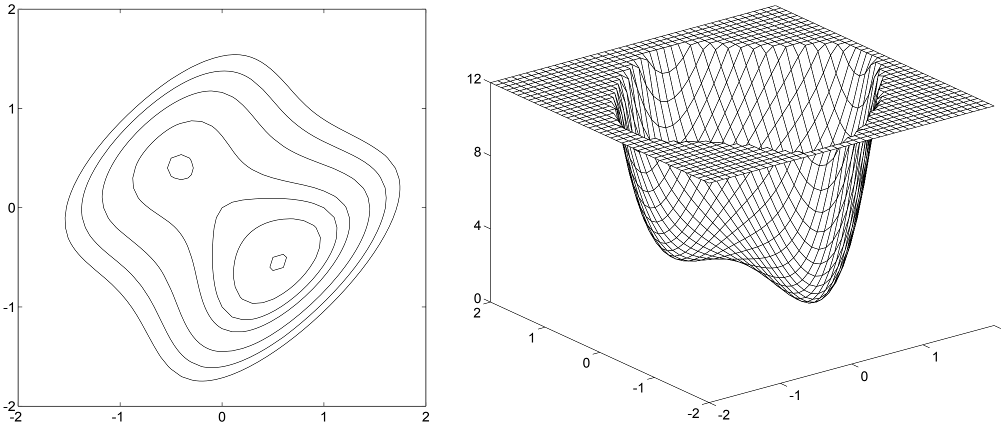


Figure 8.4 Vector Example of Minima and Saddle Point

As a final example, consider the function defined in Eq. (8.19):

$$F(\mathbf{x}) = (x_1^2 - 1.5x_1x_2 + 2x_2^2)x_1^2 \tag{8.19}$$

The contour and 3-D plots of this function are given in Figure 8.5. Here we can see that any point along the line $x_1 = 0$ is a weak minimum.

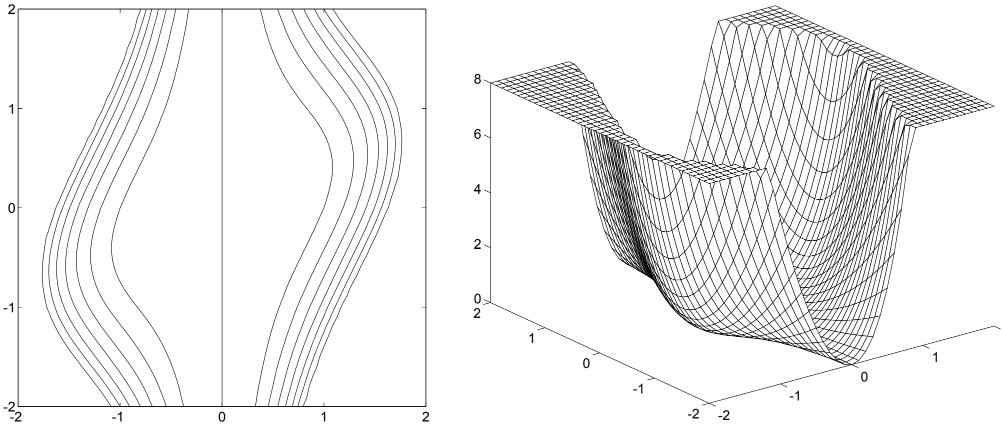


Figure 8.5 Weak Minimum Example

Necessary Conditions for Optimality

Now that we have defined what we mean by an optimum (minimum) point, let's identify some conditions that would have to be satisfied by such a point. We will again use the Taylor series expansion to derive these conditions:

$$F(\mathbf{x}) = F(\mathbf{x}^* + \Delta\mathbf{x}) = F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x}$$
$$+ \frac{1}{2}\Delta\mathbf{x}^T\nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} + \cdots, \quad (8.20)$$

where

$$\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}^* . \quad (8.21)$$

First-Order Conditions

If $\|\Delta\mathbf{x}\|$ is very small then the higher order terms in Eq. (8.20) will be negligible and we can approximate the function as

$$F(\mathbf{x}^* + \Delta\mathbf{x}) \cong F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} . \quad (8.22)$$

The point $\mathbf{x}^*$ is a candidate minimum point, which means that the function should go up (or at least not go down) if $\Delta\mathbf{x}$ is not zero. For this to happen the second term in Eq. (8.22) should not be negative. In other words

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} \geq 0 . \quad (8.23)$$

However, if this term is positive,

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} > 0 , \quad (8.24)$$

then this would imply that

$$F(\mathbf{x}^* - \Delta\mathbf{x}) \cong F(\mathbf{x}^*) - \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} < F(\mathbf{x}^*) . \quad (8.25)$$

But this is a contradiction, since $\mathbf{x}^*$ should be a minimum point. Therefore, since Eq. (8.23) must be true, and Eq. (8.24) must be false, the only alternative must be that

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}\Delta\mathbf{x} = 0 . \quad (8.26)$$

Since this must be true for any $\Delta\mathbf{x}$, we have

$$\nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0} . \quad (8.27)$$

Therefore the gradient must be zero at a minimum point. This is a first-order, necessary (but not sufficient) condition for $\mathbf{x}^*$ to be a local minimum point. Any points that satisfy Eq. (8.27) are called *stationary points*.

Stationary Points

Second-Order Conditions

Assume that we have a stationary point $\mathbf{x}^*$. Since the gradient of $F(\mathbf{x})$ is zero at all stationary points, the Taylor series expansion will be

$$F(\mathbf{x}^* + \Delta\mathbf{x}) = F(\mathbf{x}^*) + \frac{1}{2}\Delta\mathbf{x}^T \nabla^2 F(\mathbf{x})\big|_{\mathbf{x} = \mathbf{x}^*} \Delta\mathbf{x} + \cdots . \tag{8.28}$$

As before, we will consider only those points in a small neighborhood of $\mathbf{x}^*$, so that $\|\Delta\mathbf{x}\|$ is small and $F(\mathbf{x})$ can be approximated by the first two terms in Eq. (8.28). Therefore a strong minimum will exist at $\mathbf{x}^*$ if

$$\Delta\mathbf{x}^T \nabla^2 F(\mathbf{x})\big|_{\mathbf{x} = \mathbf{x}^*} \Delta\mathbf{x} > 0 . \tag{8.29}$$

Positive Definite Matrix

For this to be true for arbitrary $\Delta\mathbf{x} \neq \mathbf{0}$ requires that the Hessian matrix be positive definite. (By definition, a matrix $\mathbf{A}$ is *positive definite* if

$$\mathbf{z}^T \mathbf{A} \mathbf{z} > 0 \tag{8.30}$$

Positive Semidefinite

for any vector $\mathbf{z} \neq \mathbf{0}$. It is *positive semidefinite* if

$$\mathbf{z}^T \mathbf{A} \mathbf{z} \geq 0 \tag{8.31}$$

for any vector $\mathbf{z}$. We can check these conditions by testing the eigenvalues of the matrix. If all eigenvalues are positive, then the matrix is positive definite. If all eigenvalues are nonnegative, then the matrix is positive semidefinite.)

Sufficient Condition

A positive definite Hessian matrix is a second-order, *sufficient* condition for a strong minimum to exist. It is not a necessary condition. A minimum can still be strong if the second-order term of the Taylor series is zero, but the third-order term is positive. Therefore the second-order, *necessary* condition for a strong minimum is that the Hessian matrix be positive semi-definite.

To illustrate these conditions, consider the following function of two variables:

$$F(\mathbf{x}) = x_1^4 + x_2^2 . \tag{8.32}$$

First, we want to locate any stationary points, so we need to evaluate the gradient:

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 4x_1^3 \\ 2x_2 \end{bmatrix} = \mathbf{0} . \tag{8.33}$$

Therefore the only stationary point is the point $\mathbf{x}^* = \mathbf{0}$. We now need to test the second-order condition, which requires the Hessian matrix:

$$\nabla^2 F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{0}} = \left.\begin{bmatrix} 12x_1^2 & 0 \\ 0 & 2 \end{bmatrix}\right|_{\mathbf{x}=\mathbf{0}} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}. \tag{8.34}$$

This matrix is positive semidefinite, which is a necessary condition for $\mathbf{x}^* = \mathbf{0}$ to be a strong minimum point. We cannot guarantee from first-order and second-order conditions that it is a minimum point, but we have not eliminated it as a possibility. Actually, even though the Hessian matrix is only positive semidefinite, $\mathbf{x}^* = \mathbf{0}$ is a strong minimum point, but we cannot prove it from the conditions we have discussed.

Just to summarize, the necessary conditions for $\mathbf{x}^*$ to be a minimum, strong or weak, of $F(\mathbf{x})$ are:

$$\nabla F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0} \text{ and } \nabla^2 F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*} \text{ positive semidefinite.}$$

The sufficient conditions for $\mathbf{x}^*$ to be a strong minimum point of $F(\mathbf{x})$ are:

$$\nabla F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*} = \mathbf{0} \text{ and } \nabla^2 F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*} \text{ positive definite.}$$

Quadratic Functions

We will find throughout this text that one type of performance index is universal — the quadratic function. This is true because there are many applications in which the quadratic function appears, but also because many functions can be approximated by quadratic functions in small neighborhoods, especially near local minimum points. For this reason we want to spend a little time investigating the characteristics of the quadratic function.

Quadratic Function

The general form of a *quadratic function* is

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c, \tag{8.35}$$

where the matrix $\mathbf{A}$ is symmetric. (If the matrix is not symmetric it can be replaced by a symmetric matrix that produces the same $F(\mathbf{x})$. Try it!)

To find the gradient for this function, we will use the following useful properties of the gradient:

$$\nabla(\mathbf{h}^T\mathbf{x}) = \nabla(\mathbf{x}^T\mathbf{h}) = \mathbf{h}, \tag{8.36}$$

where $\mathbf{h}$ is a constant vector, and

$$\nabla \mathbf{x}^T \mathbf{Q} \mathbf{x} = \mathbf{Q}\mathbf{x} + \mathbf{Q}^T \mathbf{x} = 2\mathbf{Q}\mathbf{x} \text{ (for symmetric } \mathbf{Q}\text{)} . \tag{8.37}$$

We can now compute the gradient of $F(\mathbf{x})$:

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} , \tag{8.38}$$

and in a similar way we can find the Hessian:

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} . \tag{8.39}$$

All higher derivatives of the quadratic function are zero. Therefore the first three terms of the Taylor series expansion (as in Eq. (8.20)) give an exact representation of the function. We can also say that all analytic functions behave like quadratics over a small neighborhood (i.e., when $\|\Delta\mathbf{x}\|$ is small).

Eigensystem of the Hessian

We now want to investigate the general shape of the quadratic function. It turns out that we can tell a lot about the shape by looking at the eigenvalues and eigenvectors of the Hessian matrix. Consider a quadratic function that has a stationary point at the origin, and whose value there is zero:

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} . \tag{8.40}$$

The shape of this function can be seen more clearly if we perform a change of basis (see Chapter 6). We want to use the eigenvectors of the Hessian matrix, $\mathbf{A}$, as the new basis vectors. Since $\mathbf{A}$ is symmetric, its eigenvectors will be mutually orthogonal. (See [Brog91].) This means that if we make up a matrix with the eigenvectors as the columns, as in Eq. (6.68):

$$\mathbf{B} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_n \end{bmatrix} , \tag{8.41}$$

the inverse of the matrix will be the same as the transpose:

$$\mathbf{B}^{-1} = \mathbf{B}^T . \tag{8.42}$$

(This assumes that we have normalized the eigenvectors.)

If we now perform a change of basis, so that the eigenvectors are the basis vectors (as in Eq. (6.69)), the new $\mathbf{A}$ matrix will be

$$\mathbf{A}' = [\mathbf{B}^T \mathbf{A} \mathbf{B}] = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix} = \Lambda \quad , \tag{8.43}$$

where the λ_i are the eigenvalues of $\mathbf{A}$. We can also write this equation as

$$\mathbf{A} = \mathbf{B}\Lambda\mathbf{B}^T . \tag{8.44}$$

We will now use the concept of the directional derivative to explain the physical meaning of the eigenvalues and eigenvectors of $\mathbf{A}$, and to explain how they determine the shape of the surface of the quadratic function.

Recall from Eq. (8.13) that the second derivative of a function $F(\mathbf{x})$ in the direction of a vector $\mathbf{p}$ is given by

$$\frac{\mathbf{p}^T \nabla^2 F(\mathbf{x})\mathbf{p}}{\|\mathbf{p}\|^2} = \frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} . \tag{8.45}$$

Now define

$$\mathbf{p} = \mathbf{B}\mathbf{c} , \tag{8.46}$$

where $\mathbf{c}$ is the representation of the vector $\mathbf{p}$ with respect to the eigenvectors of $\mathbf{A}$. (See Eq. (6.28) and the discussion that follows.) With this definition, and Eq. (8.44), we can rewrite Eq. (8.45):

$$\frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} = \frac{\mathbf{c}^T\mathbf{B}^T(\mathbf{B}\Lambda\mathbf{B}^T)\mathbf{B}\mathbf{c}}{\mathbf{c}^T\mathbf{B}^T\mathbf{B}\mathbf{c}} = \frac{\mathbf{c}^T\Lambda\mathbf{c}}{\mathbf{c}^T\mathbf{c}} = \frac{\sum_{i=1}^{n} \lambda_i c_i^2}{\sum_{i=1}^{n} c_i^2} . \tag{8.47}$$

This result tells us several useful things. First, note that this second derivative is just a weighted average of the eigenvalues. Therefore it can never be larger than the largest eigenvalue, or smaller than the smallest eigenvalue. In other words,

$$\lambda_{min} \le \frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} \le \lambda_{max} . \tag{8.48}$$

Under what condition, if any, will this second derivative be equal to the largest eigenvalue? What if we choose

$$\mathbf{p} = \mathbf{z}_{max} , \tag{8.49}$$

where $\mathbf{z}_{max}$ is the eigenvector associated with the largest eigenvalue, λ_{max}? For this case the $\mathbf{c}$ vector will be

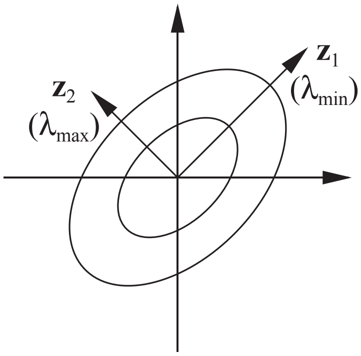
$$\mathbf{c} = \mathbf{B}^T\mathbf{p} = \mathbf{B}^T\mathbf{z}_{max} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}^T , \tag{8.50}$$

where the one occurs only in the position that corresponds to the largest eigenvalue (i.e., $c_{max} = 1$). This is because the eigenvectors are orthonormal.

If we now substitute $\mathbf{z}_{max}$ for $\mathbf{p}$ in Eq. (8.47) we obtain

$$\frac{\mathbf{z}_{max}{}^T \mathbf{A} \mathbf{z}_{max}}{\|\mathbf{z}_{max}\|^2} = \frac{\sum_{i=1}^{n} \lambda_i c_i^2}{\sum_{i=1}^{n} c_i^2} = \lambda_{max} . \tag{8.51}$$

So the maximum second derivative occurs in the direction of the eigenvector that corresponds to the largest eigenvalue. In fact, in each of the eigenvector directions the second derivatives will be equal to the corresponding eigenvalue. In other directions the second derivative will be a weighted average of the eigenvalues. The eigenvalues are the second derivatives in the directions of the eigenvectors.



The eigenvectors define a new coordinate system in which the quadratic cross terms vanish. The eigenvectors are known as the principal axes of the function contours. The figure to the left illustrates these concepts in two dimensions. This figure illustrates the case where the first eigenvalue is smaller than the second eigenvalue. Therefore the minimum curvature (second derivative) will occur in the direction of the first eigenvector. This means that we will cross contour lines more slowly in this direction. The maximum curvature will occur in the direction of the second eigenvector, therefore we will cross contour lines more quickly in that direction.

One caveat about this figure: it is only valid when both eigenvalues have the same sign, so that we have either a strong minimum or a strong maximum. For these cases the contour lines are always elliptical. We will provide examples later where the eigenvalues have opposite signs and where one of the eigenvalues is zero.

For our first example, consider the following function:

$$F(\mathbf{x}) = x_1^2 + x_2^2 = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{x} . \tag{8.52}$$

The Hessian matrix and its eigenvalues and eigenvectors are

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} , \lambda_1 = 2 , \mathbf{z}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} , \lambda_2 = 2 , \mathbf{z}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} . \tag{8.53}$$

(Actually, any two independent vectors could be the eigenvectors in this case. There is a repeated eigenvalue, and its eigenvector is the plane.) Since all the eigenvalues are equal, the curvature should be the same in all directions, and therefore the function should have circular contours. Figure 8.6 shows the contour and 3-D plots for this function, a circular hollow.

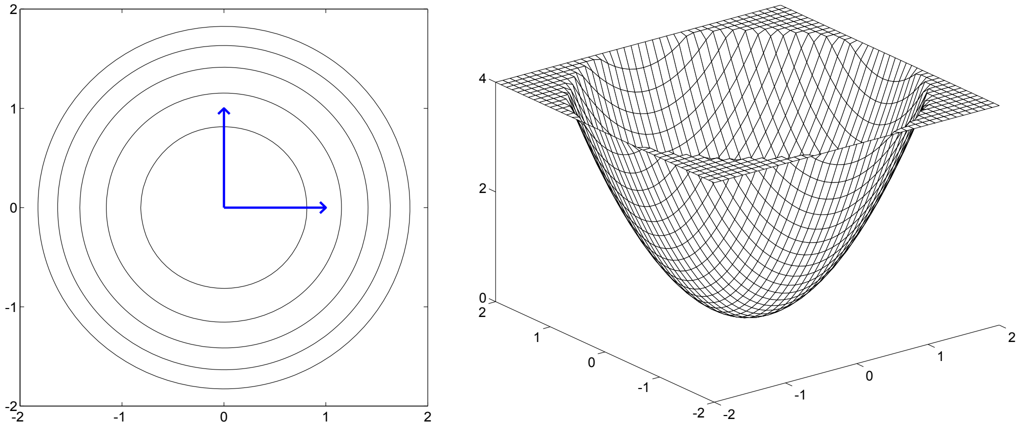


Figure 8.6 Circular Hollow

Let's try an example with distinct eigenvalues. Consider the following quadratic function:

$$F(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\mathbf{x} \tag{8.54}$$

The Hessian matrix and its eigenvalues and eigenvectors are

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},\ \lambda_1 = 1,\ \mathbf{z}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix},\ \lambda_2 = 3,\ \mathbf{z}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{8.55}$$

(As we discussed in Chapter 6, the eigenvectors are not unique, they can be multiplied by any scalar.) In this case the maximum curvature is in the direction of $\mathbf{z}_2$ so we should cross contour lines more quickly in that direction. Figure 8.7 shows the contour and 3-D plots for this function, an elliptical hollow.

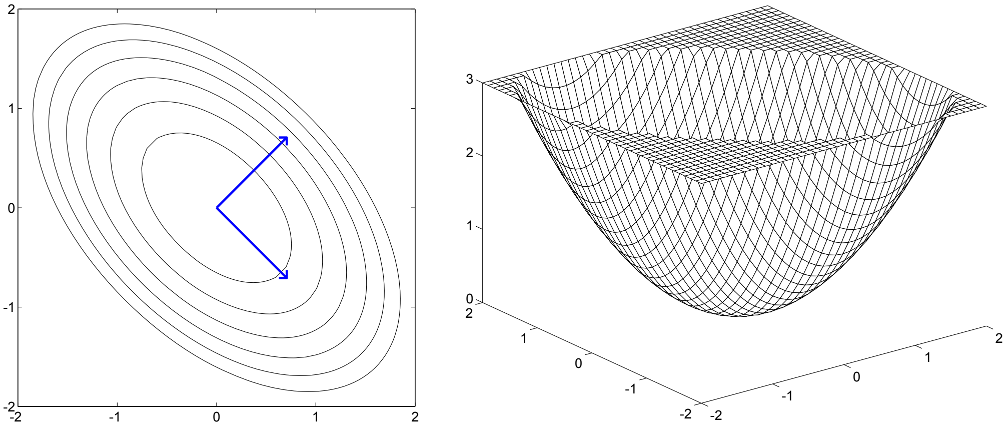


Figure 8.7 Elliptical Hollow

What happens when the eigenvalues have opposite signs? Consider the following function:

$$F(\mathbf{x}) = -\frac{1}{4}x_1^2 - \frac{3}{2}x_1x_2 - \frac{1}{4}x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}-0.5 & -1.5\\ -1.5 & -0.5\end{bmatrix}\mathbf{x}\,. \tag{8.56}$$

The Hessian matrix and its eigenvalues and eigenvectors are

$$\nabla^2F(\mathbf{x}) = \begin{bmatrix}-0.5 & -1.5\\ -1.5 & -0.5\end{bmatrix},\ \lambda_1 = 1\,,\ \mathbf{z}_1 = \begin{bmatrix}-1\\ 1\end{bmatrix},\ \lambda_2 = -2\,,\ \mathbf{z}_2 = \begin{bmatrix}-1\\ -1\end{bmatrix}. \tag{8.57}$$

The first eigenvalue is positive, so there is positive curvature in the direction of $\mathbf{z}_1$. The second eigenvalue is negative, so there is negative curvature in the direction of $\mathbf{z}_2$. Also, since the magnitude of the second eigenvalue is greater than the magnitude of the first eigenvalue, we will cross contour lines faster in the direction of $\mathbf{z}_2$.

Figure 8.8 shows the contour and 3-D plots for this function, an elongated saddle. Note that the stationary point,

$$\mathbf{x}^* = \begin{bmatrix}0\\ 0\end{bmatrix}, \tag{8.58}$$

is no longer a strong minimum point, since the Hessian matrix is not positive definite. Since the eigenvalues are of opposite sign, we know that the Hessian is indefinite (see [Brog91]). The stationary point is therefore a saddle point. It is a minimum of the function along the first eigenvector (positive eigenvalue), but it is a maximum of the function along the second eigenvector (negative eigenvalue).

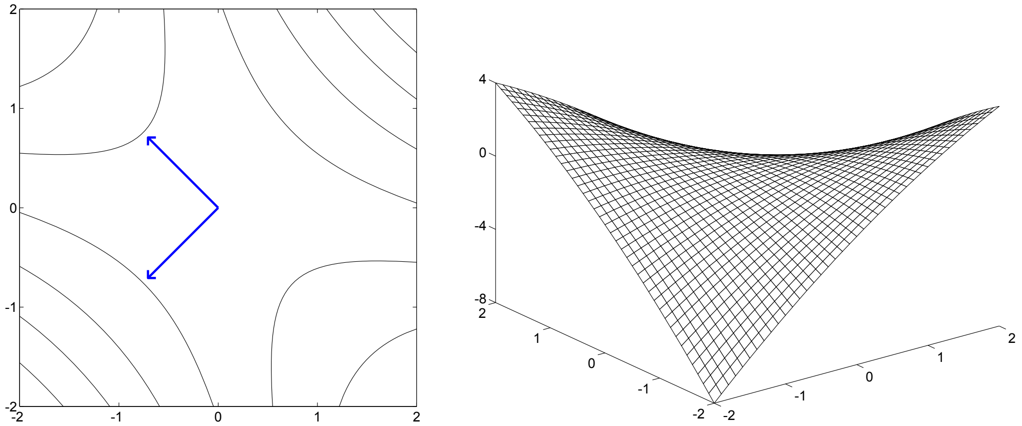


Figure 8.8 Elongated Saddle

As a final example, let's try a case where one of the eigenvalues is zero. An example of this is given by the following function:

$$F(\mathbf{x}) = \frac{1}{2}x_1^2 - x_1x_2 + \frac{1}{2}x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\mathbf{x}\,. \qquad (8.59)$$

The Hessian matrix and its eigenvalues and eigenvectors are

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},\ \lambda_1 = 2\,,\ \mathbf{z}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix},\ \lambda_2 = 0\,,\ \mathbf{z}_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}. \qquad (8.60)$$

The second eigenvalue is zero, so we would expect to have zero curvature along $\mathbf{z}_2$. Figure 8.9 shows the contour and 3-D plots for this function, a stationary valley. In this case the Hessian matrix is positive semidefinite, and we have a weak minimum along the line

$$x_1 = x_2\,, \qquad (8.61)$$

corresponding to the second eigenvector.

For quadratic functions the Hessian matrix must be positive definite in order for a strong minimum to exist. For higher-order functions it is possible to have a strong minimum with a positive semidefinite Hessian matrix, as we discussed previously in the section on minima.

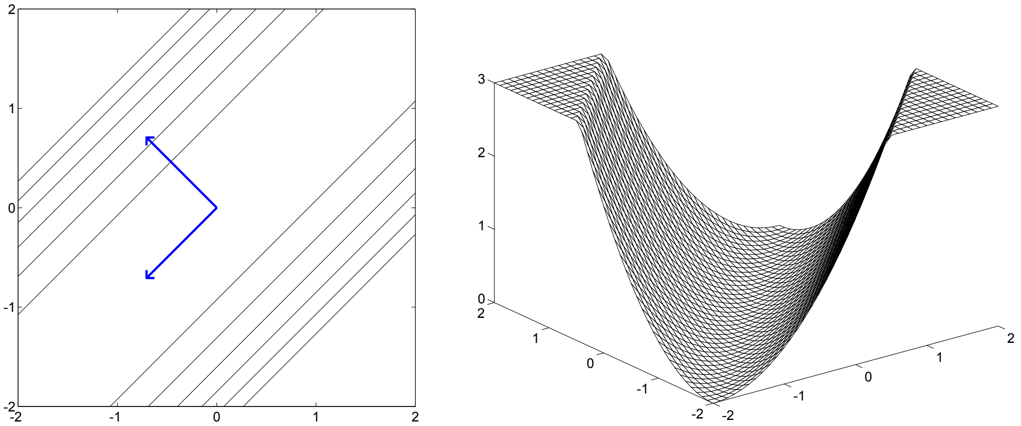


Figure 8.9 Stationary Valley

To experiment with other quadratic functions, use the MATLAB® Neural Network Design Demonstration Quadratic Function (`nnd8qf`).

At this point we can summarize some characteristics of the quadratic function.

1. If the eigenvalues of the Hessian matrix are all positive, the function will have a single strong minimum.
2. If the eigenvalues are all negative, the function will have a single strong maximum.
3. If some eigenvalues are positive and other eigenvalues are negative, the function will have a single saddle point.
4. If the eigenvalues are all nonnegative, but some eigenvalues are zero, then the function will either have a weak minimum (as in Figure 8.9) or will have no stationary point (see Solved Problem P8.7).
5. If the eigenvalues are all nonpositive, but some eigenvalues are zero, then the function will either have a weak maximum or will have no stationary point.

We should note that in this discussion we have assumed, for simplicity, that the stationary point of the quadratic function was at the origin, and that it had a zero value there. This requires that the terms $\mathbf{d}$ and c in Eq. (8.35) both be zero. If c is nonzero then the function is simply increased in magnitude by c at every point. The shape of the contours do not change. When $\mathbf{d}$ is nonzero, and $\mathbf{A}$ is invertible, the shape of the contours are not changed, but the stationary point of the function moves to

$$\mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{d} . \tag{8.62}$$

If $\mathbf{A}$ is not invertible (has some zero eigenvalues) and $\mathbf{d}$ is nonzero then stationary points may not exist (see Solved Problem P8.7).

Summary of Results

Taylor Series

$$F(\mathbf{x}) = F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x}-\mathbf{x}^*)$$

$$+\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\nabla^2 F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x}-\mathbf{x}^*)+\cdots$$

Gradient

$$\nabla F(\mathbf{x}) = \left[\frac{\partial}{\partial x_1}F(\mathbf{x})\ \frac{\partial}{\partial x_2}F(\mathbf{x})\ \ldots\ \frac{\partial}{\partial x_n}F(\mathbf{x})\right]^T$$

Hessian Matrix

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2}F(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2}F(\mathbf{x}) & \ldots & \frac{\partial^2}{\partial x_1 \partial x_n}F(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1}F(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2}F(\mathbf{x}) & \ldots & \frac{\partial^2}{\partial x_2 \partial x_n}F(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1}F(\mathbf{x}) & \frac{\partial^2}{\partial x_n \partial x_2}F(\mathbf{x}) & \ldots & \frac{\partial^2}{\partial x_n^2}F(\mathbf{x}) \end{bmatrix}$$

Directional Derivatives

First Directional Derivative

$$\frac{\mathbf{p}^T\nabla F(\mathbf{x})}{\|\mathbf{p}\|}$$

Second Directional Derivative

$$\frac{\mathbf{p}^T\nabla^2 F(\mathbf{x})\mathbf{p}}{\|\mathbf{p}\|^2}$$

Minima

Strong Minimum

The point $\mathbf{x}^*$ is a strong minimum of $F(\mathbf{x})$ if a scalar $\delta > 0$ exists, such that $F(\mathbf{x}) < F(\mathbf{x} + \Delta\mathbf{x})$ for all $\Delta\mathbf{x}$ such that $\delta > \|\Delta\mathbf{x}\| > 0$.

Global Minimum

The point $\mathbf{x}^*$ is a unique global minimum of $F(\mathbf{x})$ if $F(\mathbf{x}) < F(\mathbf{x} + \Delta\mathbf{x})$ for all $\Delta\mathbf{x} \neq \mathbf{0}$.

Weak Minimum

The point $\mathbf{x}^*$ is a weak minimum of $F(\mathbf{x})$ if it is not a strong minimum, and a scalar $\delta > 0$ exists, such that $F(\mathbf{x}) \leq F(\mathbf{x} + \Delta\mathbf{x})$ for all $\Delta\mathbf{x}$ such that $\delta > \|\Delta\mathbf{x}\| > 0$.

Necessary Conditions for Optimality

First-Order Condition

$$\nabla F(\mathbf{x})\Big|_{\mathbf{x} = \mathbf{x}^*} = \mathbf{0} \quad \text{(Stationary Points)}$$

Second-Order Condition

$$\nabla^2 F(\mathbf{x})\Big|_{\mathbf{x} = \mathbf{x}^*} \geq 0 \quad \text{(Positive Semidefinite Hessian Matrix)}$$

Quadratic Functions

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c$$

Gradient

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d}$$

Hessian

$$\nabla^2 F(\mathbf{x}) = \mathbf{A}$$

Directional Derivatives

$$\lambda_{min} \leq \frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} \leq \lambda_{max}$$

Solved Problems

P8.1 In Figure 8.1 we illustrated 3 approximations to the cosine function about the point $x^* = 0$. Repeat that procedure about the point $x^* = \pi/2$.

The function we want to approximate is

$$F(x) = \cos(x) .$$

The Taylor series expansion for $F(x)$ about the point $x^* = \pi/2$ is

$$F(x) = \cos(x) = \cos\left(\frac{\pi}{2}\right) - \sin\left(\frac{\pi}{2}\right)\left(x - \frac{\pi}{2}\right) - \frac{1}{2}\cos\left(\frac{\pi}{2}\right)\left(x - \frac{\pi}{2}\right)^2 + \frac{1}{6}\sin\left(\frac{\pi}{2}\right)\left(x - \frac{\pi}{2}\right)^3 + \cdots$$

$$= -\left(x - \frac{\pi}{2}\right) + \frac{1}{6}\left(x - \frac{\pi}{2}\right)^3 - \frac{1}{120}\left(x - \frac{\pi}{2}\right)^5 + \cdots$$

The zeroth-order approximation of $F(x)$ is

$$F(x) \approx F_0(x) = 0 .$$

The first-order approximation is

$$F(\mathbf{x}) \approx F_1(x) = -\left(x - \frac{\pi}{2}\right) = \frac{\pi}{2} - x .$$

(Note that in this case the second-order approximation is the same as the first-order approximation, since the second derivative is zero.)

The third-order approximation is

$$F(\mathbf{x}) \approx F_3(x) = -\left(x - \frac{\pi}{2}\right) + \frac{1}{6}\left(x - \frac{\pi}{2}\right)^3 .$$

A graph showing $F(x)$ and these three approximations is shown in Figure P8.1. Note that in this case the zeroth-order approximation is very poor, while the first-order approximation is accurate over a reasonably wide range. Compare this result with Figure 8.1. In that case we were expanding about a local maximum point, $x^* = 0$, so the first derivative was zero.

Check the Taylor series expansions at other points using the Neural Network Design Demonstration Taylor Series (`nnd8ts1`).

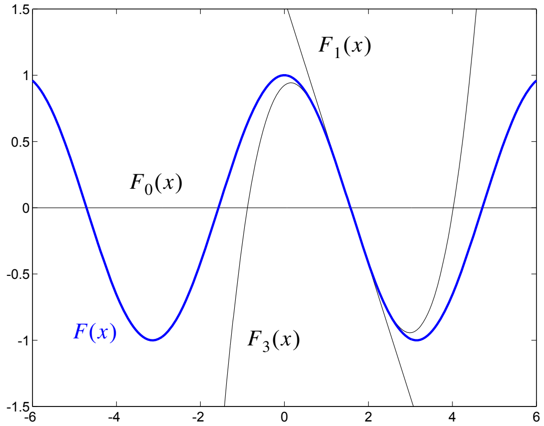


Figure P8.1 Cosine Approximation About $x = \pi/2$

P8.2 Recall the function that is displayed in Figure 8.4, on page 8-9. We know that this function has two strong minima. Find the second-order Taylor series expansions for this function about the two minima.

The equation for this function is

$$F(\mathbf{x}) = (x_2 - x_1)^4 + 8x_1x_2 - x_1 + x_2 + 3 .$$

To find the second-order Taylor series expansion, we need to find the gradient and the Hessian for $F(\mathbf{x})$. For the gradient we have

$$\nabla F(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1}F(\mathbf{x}) \\ \frac{\partial}{\partial x_2}F(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} -4(x_2 - x_1)^3 + 8x_2 - 1 \\ 4(x_2 - x_1)^3 + 8x_1 + 1 \end{bmatrix},$$

and the Hessian matrix is

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2}F(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2}F(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1}F(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2}F(\mathbf{x}) \end{bmatrix}$$

$$= \begin{bmatrix} 12(x_2 - x_1)^2 & -12(x_2 - x_1)^2 + 8 \\ -12(x_2 - x_1)^2 + 8 & 12(x_2 - x_1)^2 \end{bmatrix}$$

One strong minimum occurs at $\mathbf{x}^1 = \begin{bmatrix}-0.42 & 0.42\end{bmatrix}^T$, and the other at $\mathbf{x}^2 = \begin{bmatrix}0.55 & -0.55\end{bmatrix}^T$. If we perform the second-order Taylor series expansion of $F(\mathbf{x})$ about these two points we obtain:

$$F^1(\mathbf{x}) = F(\mathbf{x}^1) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^1}(\mathbf{x}-\mathbf{x}^1) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^1)^T\nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^1}(\mathbf{x}-\mathbf{x}^1)$$

$$= 2.93 + \frac{1}{2}\left(\mathbf{x} - \begin{bmatrix}-0.42\\0.42\end{bmatrix}\right)^T \begin{bmatrix}8.42 & -0.42\\-0.42 & 8.42\end{bmatrix}\left(\mathbf{x} - \begin{bmatrix}-0.42\\0.42\end{bmatrix}\right).$$

If we simplify this expression we find

$$F^1(\mathbf{x}) = 4.49 - \begin{bmatrix}-3.7128 & 3.7128\end{bmatrix}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\begin{bmatrix}8.42 & -0.42\\-0.42 & 8.42\end{bmatrix}\mathbf{x}.$$

Repeating this process for $\mathbf{x}^2$ results in

$$F^2(\mathbf{x}) = 7.41 - \begin{bmatrix}11.781 & -11.781\end{bmatrix}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\begin{bmatrix}14.71 & -6.71\\-6.71 & 14.71\end{bmatrix}\mathbf{x}.$$

The original function and the two approximations are plotted in the following figures.

Check the Taylor series expansions at other points using the Neural Network Design Demonstration *Vector Taylor Series* (`nnd8ts2`).

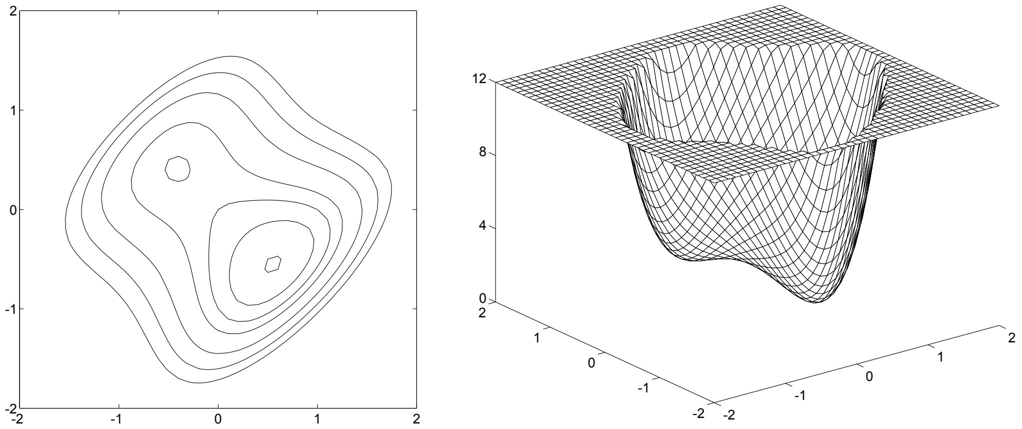


Figure P8.2 Function $F(\mathbf{x})$ for Problem P8.2

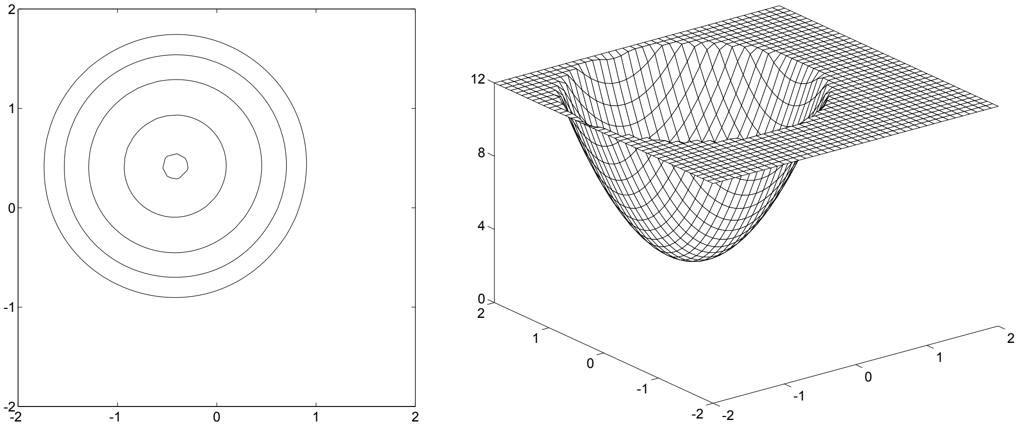


Figure P8.3 Function $F^1(\mathbf{x})$ for Problem P8.2

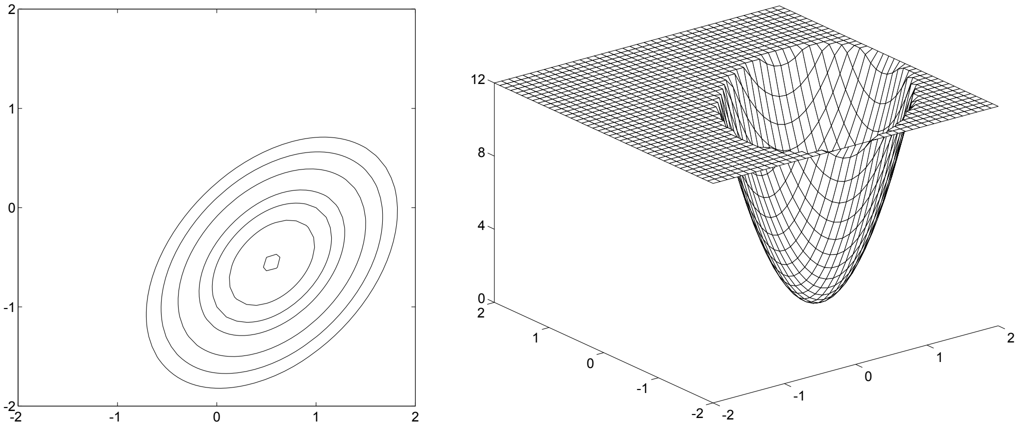


Figure P8.4 Function $F^2(\mathbf{x})$ for Problem P8.2

P8.3 For the function $F(\mathbf{x})$ given below, find the equation for the line that is tangent to the contour line at $\mathbf{x} = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$.

$$F(\mathbf{x}) = (2 + x_1)^2 + 5(1 - x_1 - x_2^2)^2$$

To solve this problem we can use the directional derivative. What is the derivative of $F(\mathbf{x})$ along a line that is tangent to a contour line? Since the contour is a line along which the function does not change, the derivative of $F(\mathbf{x})$ should be zero in the direction of the contour. So we can get the equation for the tangent to the contour line by setting the directional derivative equal to zero.

First we need to find the gradient:

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 2(2+x_1) + 10(1-x_1-x_2^2)(-1) \\ 10(1-x_1-x_2^2)(-2x_2) \end{bmatrix} = \begin{bmatrix} -6 + 12x_1 + 10x_2^2 \\ -20x_2 + 20x_1x_2 + 20x_2^3 \end{bmatrix}.$$

If we evaluate this at $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$, we obtain

$$\nabla F(\mathbf{x}^*) = \begin{bmatrix} -6 \\ 0 \end{bmatrix}.$$

Now recall that the equation for the derivative of $F(\mathbf{x})$ in the direction of a vector $\mathbf{p}$ is

$$\frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|}.$$

Therefore if we want the equation for the line that passes through $\mathbf{x}^* = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ and along which the derivative is zero, we can set the numerator of the directional derivative in the direction of $\Delta\mathbf{x}$ to zero:

$$\Delta\mathbf{x}^T \nabla F(\mathbf{x}^*) = 0,$$

where $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}^*$. For this case we have

$$\mathbf{x}^T \begin{bmatrix} -6 \\ 0 \end{bmatrix} = 0, \text{ or } x_1 = 0.$$

This result is illustrated in Figure P8.5.

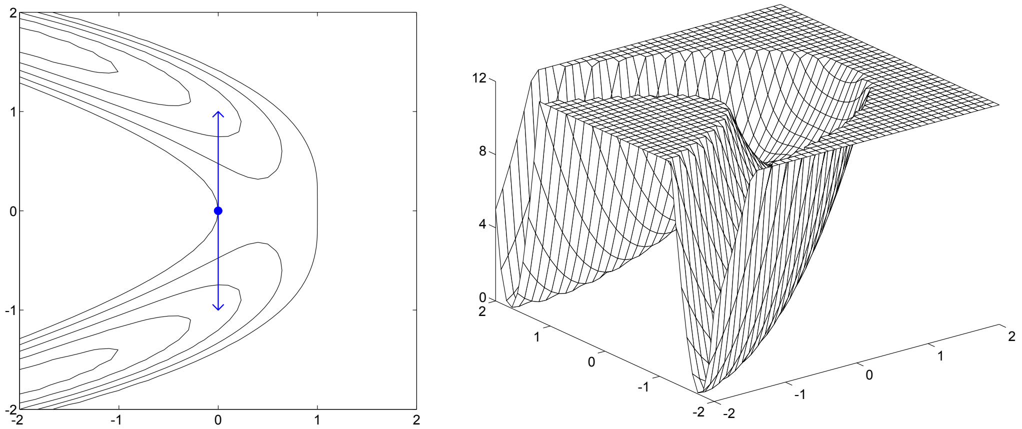


Figure P8.5 Plot of $F(\mathbf{x})$ for Problem P8.3

P8.4 Consider the following fourth-order polynomial:

$$F(x) = x^4 - \frac{2}{3}x^3 - 2x^2 + 2x + 4 .$$

Find any stationary points and test them to see if they are minima.

To find the stationary points we set the derivative of $F(x)$ to zero:

$$\frac{d}{dx}F(x) = 4x^3 - 2x^2 - 4x + 2 = 0 .$$

We can use MATLAB to find the roots of this polynomial:

```
coef=[4 -2 -4 2];

stapoints=roots(coef);

stapoints'

ans =

    1.0000    -1.0000     0.5000
```

Now we need to check the second derivative at each of these points. The second derivative of $F(x)$ is

$$\frac{d^2}{dx^2}F(x) = 12x^2 - 4x - 4 .$$

If we evaluate this at each of the stationary points we find

$$\left(\frac{d^2}{dx^2}F(1) = 4\right), \left(\frac{d^2}{dx^2}F(-1) = 12\right), \left(\frac{d^2}{dx^2}F(0.5) = -3\right) .$$

Therefore we should have strong local minima at 1 and -1 (since the second derivatives were positive), and a strong local maximum at 0.5 (since the second derivative was negative). To find the global minimum we would have to evaluate the function at the two local minima:

$$(F(1) = 4.333), (F(-1) = 1.667) .$$

Therefore the global minimum occurs at -1. But are we sure that this is a global minimum? What happens to the function as $x \to \infty$ or $x \to -\infty$? In this case, because the highest power of x has a positive coefficient and is an even power (x^4), the function goes to ∞ at both limits. So we can safely say that the global minimum occurs at -1. The function is plotted in Figure P8.6.

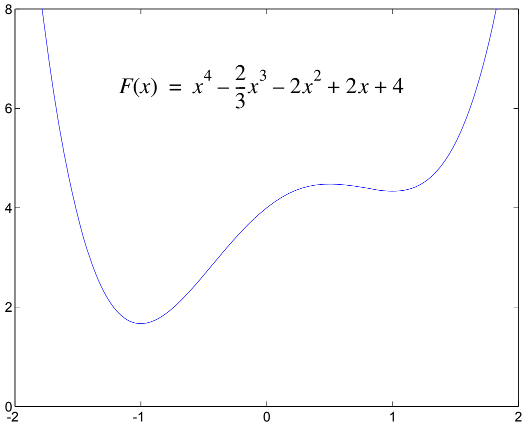


Figure P8.6 Graph of $F(x)$ for Problem P8.4

P8.5 Look back to the function of Problem P8.2. This function has three stationary points:

$$\mathbf{x}^1 = \begin{bmatrix} -0.41878 \\ 0.41878 \end{bmatrix}, \mathbf{x}^2 = \begin{bmatrix} -0.134797 \\ 0.134797 \end{bmatrix}, \mathbf{x}^3 = \begin{bmatrix} 0.55358 \\ -0.55358 \end{bmatrix}.$$

Test whether or not any of these points could be local minima.

From Problem P8.2 we know that the Hessian matrix for the function is

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} 12(x_2 - x_1)^2 & -12(x_2 - x_1)^2 + 8 \\ -12(x_2 - x_1)^2 + 8 & 12(x_2 - x_1)^2 \end{bmatrix}.$$

To test the definiteness of this matrix we can check the eigenvalues. If the eigenvalues are all positive, the Hessian is positive definite, which guarantees a strong minimum. If the eigenvalues are nonnegative, the Hessian is positive semidefinite, which is consistent with either a strong or a weak minimum. If one eigenvalue is positive and the other eigenvalue is negative, the Hessian is indefinite, which would signal a saddle point.

If we evaluate the Hessian at $\mathbf{x}^1$, we find

$$\nabla^2 F(\mathbf{x}^1) = \begin{bmatrix} 8.42 & -0.42 \\ -0.42 & 8.42 \end{bmatrix}.$$

The eigenvalues of this matrix are

$$\lambda_1 = 8.84, \lambda_2 = 8.0,$$

therefore $\mathbf{x}^1$ must be a strong minimum point.

If we evaluate the Hessian at $\mathbf{x}^2$, we find

$$\nabla^2 F(\mathbf{x}^2) = \begin{bmatrix} 0.87 & 7.13 \\ 7.13 & 0.87 \end{bmatrix}.$$

The eigenvalues of this matrix are

$$\lambda_1 = -6.26\,,\ \lambda_2 = 8.0\,,$$

therefore $\mathbf{x}^2$ must be a saddle point. In one direction the curvature is negative, and in another direction the curvature is positive. The negative curvature is in the direction of the first eigenvector, and the positive curvature is in the direction of the second eigenvector. The eigenvectors are

$$\mathbf{z}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \text{ and } \mathbf{z}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

(Note that this is consistent with our previous discussion of this function on page 8-8.)

If we evaluate the Hessian at $\mathbf{x}^3$, we find

$$\nabla^2 F(\mathbf{x}^3) = \begin{bmatrix} 14.7 & -6.71 \\ -6.71 & 14.7 \end{bmatrix}.$$

The eigenvalues of this matrix are

$$\lambda_1 = 21.42\,,\ \lambda_2 = 8.0\,,$$

therefore $\mathbf{x}^3$ must be a strong minimum point.

Check these results using the Neural Network Design Demonstration Vector Taylor Series (`nnd8ts2`).

P8.6 Let's apply the concepts in this chapter to a neural network problem. Consider the linear network shown in Figure P8.7. Suppose that the desired inputs/outputs for the network are

$$\{(p_1 = 2), (t_1 = 0.5)\}, \{(p_2 = -1), (t_2 = 0)\}\,.$$

Sketch the following performance index for this network:

$$F(\mathbf{x}) = (t_1 - a_1(\mathbf{x}))^2 + (t_2 - a_2(\mathbf{x}))^2\,.$$

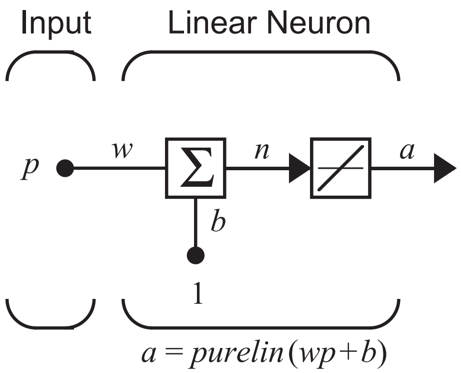


Figure P8.7 Linear Network for Problem P8.6

The parameters of this network are w and b, which make up the parameter vector

$$\mathbf{x} = \begin{bmatrix} w \\ b \end{bmatrix}.$$

We want to sketch the performance index $F(\mathbf{x})$. First we will show that the performance index is a quadratic function. Then we will find the eigenvectors and eigenvalues of the Hessian matrix and use them to sketch the contour plot of the function.

Begin by writing $F(\mathbf{x})$ as an explicit function of the parameter vector $\mathbf{x}$:

$$F(\mathbf{x}) = e_1^2 + e_2^2,$$

where

$$(e_1 = t_1 - (wp_1 + b)), (e_2 = t_2 - (wp_2 + b)).$$

This can be written in matrix form:

$$F(\mathbf{x}) = \mathbf{e}^T\mathbf{e},$$

where

$$\mathbf{e} = \mathbf{t} - \begin{bmatrix} p_1 & 1 \\ p_2 & 1 \end{bmatrix}\mathbf{x} = \mathbf{t} - \mathbf{G}\mathbf{x}.$$

The performance index can now be rewritten:

$$F(\mathbf{x}) = [\mathbf{t} - \mathbf{G}\mathbf{x}]^T[\mathbf{t} - \mathbf{G}\mathbf{x}] = \mathbf{t}^T\mathbf{t} - 2\mathbf{t}^T\mathbf{G}\mathbf{x} + \mathbf{x}^T\mathbf{G}^T\mathbf{G}\mathbf{x}.$$

If we compare this with Eq. (8.35):

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c,$$

we can see that the performance index for this linear network is a quadratic function, with

$$c = \mathbf{t}^T\mathbf{t}, \ \mathbf{d} = -2\mathbf{G}^T\mathbf{t}, \text{ and } \mathbf{A} = 2\mathbf{G}^T\mathbf{G}.$$

The gradient of the quadratic function is given in Eq. (8.38):

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} = 2\mathbf{G}^T\mathbf{G}\mathbf{x} - 2\mathbf{G}^T\mathbf{t}.$$

The stationary point (also the center of the function contours) will occur where the gradient is equal to zero:

$$\mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{d} = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{t}.$$

For

$$\mathbf{G} = \begin{bmatrix} p_1 & 1 \\ p_2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix} \text{ and } \mathbf{t} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

we have

$$\mathbf{x}^* = [\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T\mathbf{t} = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0.167 \\ 0.167 \end{bmatrix}.$$

(Therefore the optimal network parameters are $w = 0.167$ and $b = 0.167$.)

The Hessian matrix of the quadratic function is given by Eq. (8.39):

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} = 2\mathbf{G}^T\mathbf{G} = \begin{bmatrix} 10 & 2 \\ 2 & 4 \end{bmatrix}.$$

To sketch the contour plot we need the eigenvectors and eigenvalues of the Hessian. For this case we find

$$\left\{(\lambda_1 = 10.6), \left(\mathbf{z}_1 = \begin{bmatrix} 1 \\ 0.3 \end{bmatrix}\right)\right\}, \left\{(\lambda_2 = 3.4), \left(\mathbf{z}_2 = \begin{bmatrix} 0.3 \\ -1 \end{bmatrix}\right)\right\}.$$

Therefore we know that $\mathbf{x}^*$ is a strong minimum. Also, since the first eigenvalue is larger than the second, we know that the contours will be elliptical and that the long axis of the ellipses will be in the direction of the second eigenvector. The contours will be centered at $\mathbf{x}^*$. This is demonstrated in Figure P8.8.

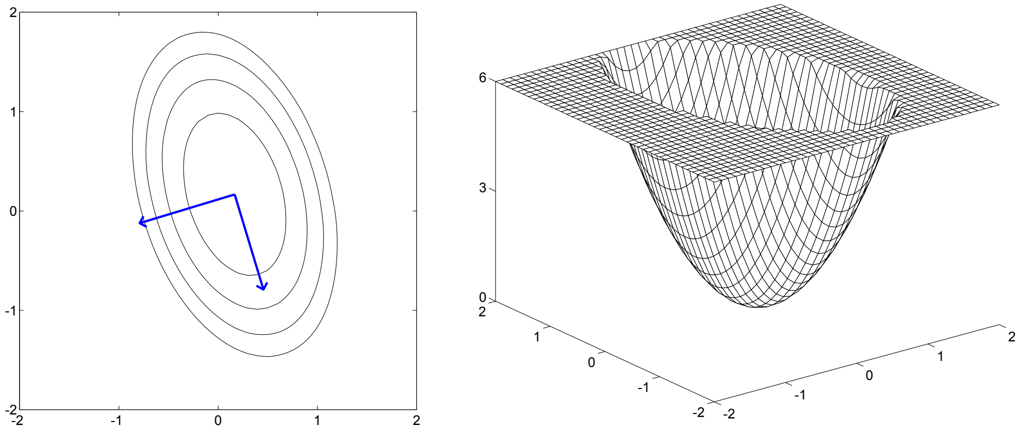


Figure P8.8 Graph of Function for Problem P8.6

P8.7 There are quadratic functions that do not have stationary points. This problem illustrates one such case. Consider the following function:

$$F(\mathbf{x}) = \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{x} .$$

Sketch the contour plot of this function.

As with Problem P8.6, we need to find the eigenvalues and eigenvectors of the Hessian matrix. By inspection of the quadratic function we see that the Hessian matrix is

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} . \tag{8.63}$$

The eigenvalues and eigenvectors are

$$\left\{ (\lambda_1 = 0), \left(\mathbf{z}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) \right\}, \left\{ (\lambda_2 = 2), \left(\mathbf{z}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \right\} .$$

Notice that the first eigenvalue is zero, so there is no curvature along the first eigenvector. The second eigenvalue is positive, so there is positive curvature along the second eigenvector. If we had no linear term in $F(\mathbf{x})$, the plot of the function would show a stationary valley, as in Figure 8.9. In this case we must find out if the linear term creates a slope in the direction of the valley (the direction of the first eigenvector).

The linear term is

$$F_{lin}(\mathbf{x}) = \begin{bmatrix} 1 & -1 \end{bmatrix} \mathbf{x} .$$

From Eq. (8.36) we know that the gradient of this term is

$$\nabla F_{lin}(\mathbf{x}) = \begin{bmatrix} 1 \\ -1 \end{bmatrix} ,$$

which means that the linear term is increasing most rapidly in the direction of this gradient. Since the quadratic term has no curvature in this direction, the overall function will have a linear slope in this direction. Therefore $F(\mathbf{x})$ will have positive curvature in the direction of the second eigenvector and a linear slope in the direction of the first eigenvector. The contour plot and the 3-D plot for this function are given in Figure P8.9.

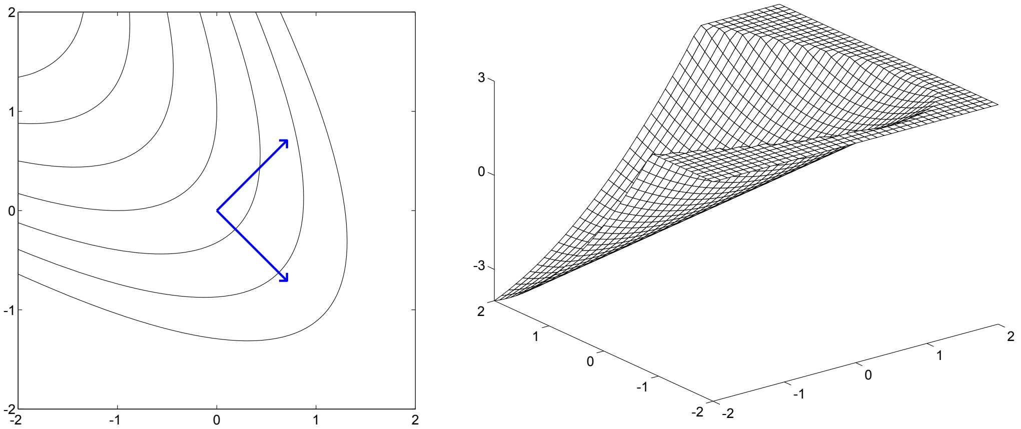


Figure P8.9 Falling Valley Function for Problem P8.7

Whenever any of the eigenvalues of the Hessian matrix are zero it is impossible to solve for the stationary point of the quadratic function using

$$\mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{d} ,$$

since the Hessian matrix does not have an inverse. This lack of an inverse could mean that we have a weak minimum point, as illustrated in Figure 8.9, or that there is no stationary point, as this example shows.

Epilogue

Performance learning is one of the most important classes of neural network learning rules. With performance learning, network parameters are adjusted to optimize network performance. In this chapter we have introduced tools that we will need to understand performance learning rules. After reading this chapter and solving the exercises, you should be able to:

i. Perform a Taylor series expansion and use it to approximate a function.

ii. Calculate a directional derivative.

iii. Find stationary points and test whether they could be minima.

iv. Sketch contour plots of quadratic functions.

We will be using these concepts in a number of succeeding chapters, including the chapters on performance learning (9–14), the radial basis network chapter (17) and the chapters on stability and Hopfield networks (20–21). In the next chapter we will build on the concepts we have covered here, to design algorithms that will optimize performance functions. Then, in succeeding chapters, we will apply these algorithms to the training of neural networks.

Further Reading

[Brog91] W. L. Brogan, *Modern Control Theory,* 3rd Ed., Englewood Cliffs, NJ: Prentice-Hall, 1991.

This is a well-written book on the subject of linear systems. The first half of the book is devoted to linear algebra. It also has good sections on the solution of linear differential equations and the stability of linear and nonlinear systems. It has many worked problems.

[Gill81] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, New York: Academic Press, 1981.

As the title implies, this text emphasizes the practical implementation of optimization algorithms. It provides motivation for the optimization methods, as well as details of implementation that affect algorithm performance.

[Himm72] D. M. Himmelblau, *Applied Nonlinear Programming*, New York: McGraw-Hill, 1972.

This is a comprehensive text on nonlinear optimization. It covers both constrained and unconstrained optimization problems. The text is very complete, with many examples worked out in detail.

[Scal85] L. E. Scales, *Introduction to Non-Linear Optimization*, New York: Springer-Verlag, 1985.

A very readable text describing the major optimization algorithms, this text emphasizes methods of optimization rather than existence theorems and proofs of convergence. Algorithms are presented with intuitive explanations, along with illustrative figures and examples. Pseudo-code is presented for most algorithms.

Exercises

E8.1 Consider the following scalar function:

$$F(x) = \frac{1}{x^3 - \frac{3}{4}x - \frac{1}{2}} .$$

i. Find the second-order Taylor series approximation for $F(x)$ about the point $x = -0.5$.

ii. Find the second-order Taylor series approximation for $F(x)$ about the point $x = 1.1$.

iii. Plot $F(x)$ and the two approximations and discuss their accuracy.

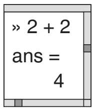
E8.2 Consider the following function of two variables:

$$F(\mathbf{x}) = e^{(2x_1^2 + 2x_2^2 + x_1 - 5x_2 + 10)} .$$

i. Find the second-order Taylor series approximation for $F(\mathbf{x})$ about the point $\mathbf{x} = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$.

ii. Find the stationary point for this approximation.

iii. Find the stationary point for $F(\mathbf{x})$. (Note that the exponent of $F(\mathbf{x})$ is simply a quadratic function.)



```
» 2 + 2
ans =
     4
```

iv. Explain the difference between the two stationary points. (Use MATLAB to plot the two functions.)

E8.3 For the following functions find the first and second directional derivatives from the point $\mathbf{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ in the direction $\mathbf{p} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$.

i. $F(\mathbf{x}) = \frac{7}{2}x_1^2 - 6x_1x_2 - x_2^2$

ii. $F(\mathbf{x}) = 5x_1^2 - 6x_1x_2 + 5x_2^2 + 4x_1 + 4x_2$

iii. $F(\mathbf{x}) = \frac{9}{2}x_1^2 - 2x_1x_2 + 3x_2^2 + 2x_1 - x_2$

iv. $F(\mathbf{x}) = -\frac{1}{2}(7x_1^2 + 12x_1x_2 - 2x_2^2)$

v. $F(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^2 + 3x_1 + 3x_2$

vi. $F(\mathbf{x}) = \frac{1}{2}x_1^2 - 3x_1x_2 + \frac{1}{2}x_2^2 - 4x_1 + 4x_2$

vii. $F(\mathbf{x}) = \frac{1}{2}x_1^2 - 2x_1x_2 + 2x_2^2 + x_1 - 2x_2$

viii. $F(\mathbf{x}) = \frac{3}{2}x_1^2 + 2x_1x_2 + 4x_1 + 4x_2$

ix. $F(\mathbf{x}) = -\frac{3}{2}x_1^2 + 4x_1x_2 + \frac{3}{2}x_2^2 + 5x_1$

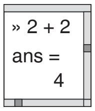
x. $F(\mathbf{x}) = 2x_1^2 - 2x_1x_2 + \frac{1}{2}x_2^2 + x_1 + x_2$

E8.4 For the following function,

$$F(x) = x^4 - \frac{1}{2}x^2 + 1 \,,$$

i. find the stationary points,

ii. test the stationary points to find minimum and maximum points, and



```
» 2 + 2
ans =
     4
```

iii. plot the function using MATLAB to verify your answers.

E8.5 Consider the following function of two variables:

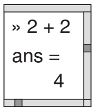
$$F(\mathbf{x}) = (x_1 + x_2)^4 - 12x_1x_2 + x_1 + x_2 + 1 \,.$$

i. Verify that the function has three stationary points at

$$\mathbf{x}^1 = \begin{bmatrix} -0.6504 \\ -0.6504 \end{bmatrix}, \; \mathbf{x}^2 = \begin{bmatrix} 0.085 \\ 0.085 \end{bmatrix}, \; \mathbf{x}^3 = \begin{bmatrix} 0.5655 \\ 0.5655 \end{bmatrix}.$$

ii. Test the stationary points to find any minima, maxima or saddle points.

iii. Find the second-order Taylor series approximations for the function at each of the stationary points.



```
» 2 + 2
ans =
     4
```

iv. Plot the function and the approximations using MATLAB.

E8.6 For the functions of Exercise E8.3:

i. find the stationary points,

ii. test the stationary points to find minima, maxima or saddle points,

iii. provide rough sketches of the contour plots, using the eigenvalues and eigenvectors of the Hessian matrices, and

iv. plot the functions using MATLAB to verify your answers.

E8.7 Consider the following quadratic function:

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 1 & -3 \\ -3 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 4 & -4 \end{bmatrix} \mathbf{x} + 2 .$$

i. Find the gradient and Hessian matrix for $F(\mathbf{x})$.

ii. Sketch the contour plot for $F(\mathbf{x})$.

iii. Find the directional derivative of $F(\mathbf{x})$ at the point $\mathbf{x}_0 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$ in the direction $\mathbf{p} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$.

iv. Is your answer to part iii. consistent with your contour plot of part ii.? Explain.

E8.8 Repeat Exercise E8.7 with the following quadratic function:

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 3 & -2 \\ -2 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 4 & 4 \end{bmatrix} \mathbf{x} + 2 .$$

E8.9 Consider the following function:

$$F(\mathbf{x}) = (1 + x_1 + x_2)^2 + \frac{1}{4}x_1^4 .$$

i. Find the quadratic approximation to $F(\mathbf{x})$ about the point $\mathbf{x}_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$

ii. Sketch the contour plot of the quadratic approximation in part i.

E8.10 Consider the following function:

$$F(\mathbf{x}) = \frac{3}{2}x_1^2 + 2x_1x_2 + x_2^3 + 4x_1 + 4x_2 .$$

i. Find the quadratic approximation to $F(\mathbf{x})$ about the point $\mathbf{x}_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$.

ii. Locate the stationary point of the quadratic approximation you found in part i.

iii. Is the answer to part ii a minimum of $F(\mathbf{x})$?

E8.11 Consider the following function:

$$F(\mathbf{x}) = x_1 x_2 - x_1 + 2x_2 .$$

i. Locate any stationary points.

ii. For each answer to part i., determine, if possible, whether the stationary point is a minimum point, a maximum point, or a saddle point.

iii. Find the directional derivative of the function at the point $\mathbf{x}_0 = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$ in the direction $\mathbf{p} = \begin{bmatrix} -1 & 1 \end{bmatrix}^T$.

E8.12 Consider the following function:

$$F(\mathbf{x}) = x_1^2 + 2x_1 x_2 + x_2^2 + (x_1 - x_2)^3 .$$

i. Find the quadratic approximation to $F(\mathbf{x})$ about the point $\mathbf{x}_0 = \begin{bmatrix} 2 & 1 \end{bmatrix}^T$.

ii. Sketch the contour plot of the quadratic approximation.

E8.13 Recall the function in Problem P8.7. For that function there was no stationary point. It is possible to modify the function, by changing only the $\mathbf{d}$ vector, so that a stationary point will exist. Find a new nonzero $\mathbf{d}$ vector that will create a weak minimum.