8 Performance Surfaces and Optimum Points

Optimization process consists of

- Defining performance index
- Searching for optimal parameters in the index

8.1 Taylor Series

$$F(x) = F(x^*) + \frac{d}{dx}F(x)\Big|_{x=x^*}(x-x^*) + \frac{1}{2}\frac{d^2}{dx^2}F(x)\Big|_{x=x^*}(x-x^*)^2 + \cdots + \frac{1}{n!}\frac{d^n}{dx^n}F(x)\Big|_{x=x^*}(x-x^*)^n + \cdots \qquad (8.1\text{-}1)$$

8.1.1 Vector Case

$$F(\mathbf{x}) = F(x_1, x_2, \cdots, x_n) \qquad (8.1.1\text{-}1)$$

$$\begin{aligned} F(\mathbf{x}) = F(\mathbf{x}^*) &+ \frac{\partial}{\partial x_1}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*) + \frac{\partial}{\partial x_2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_2 - x_2^*) + \cdots + \frac{\partial}{\partial x_n}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_n - x_n^*) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial x_1^2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*)^2 + \frac{1}{2}\frac{\partial^2}{\partial x_1 \partial x_2}F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(x_1 - x_1^*)(x_2 - x_2^*) + \cdots \end{aligned} \qquad (8.1.1\text{-}2)$$

$$F(\mathbf{x}) = F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x} - \mathbf{x}^*) + \cdots \qquad (8.1.1\text{-}3)$$

$\nabla F(x)$: Gradient,

$$\nabla F(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} F(\mathbf{x}) & \frac{\partial}{\partial x_2} F(\mathbf{x}) & \cdots & \frac{\partial}{\partial x_n} F(\mathbf{x}) \end{bmatrix}^T \tag{8.1.1-4}$$

$\nabla^2 F(x)$: Hessian,

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} F(\mathbf{x}) & \frac{\partial^2}{\partial x_1 \partial x_2} F(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} F(\mathbf{x}) \\ \frac{\partial^2}{\partial x_2 \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_2^2} F(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_2 \partial x_n} F(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} F(\mathbf{x}) & \frac{\partial^2}{\partial x_n \partial x_2} F(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_n^2} F(\mathbf{x}) \end{bmatrix} \tag{8.1.1-5}$$

8.2 Directional Derivatives

Directional derivative along **p**: (+ : increasing, **-** : decreasing, 0 : constant)

$$\frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|} \tag{8.2-1}$$

Second derivative along **p**: (+ : convex, **-** : concave, 0 : straight)

$$\frac{\mathbf{p}^T \nabla^2 F(\mathbf{x})\mathbf{p}}{\|\mathbf{p}\|^2} \tag{8.2-2}$$

Example:

$$F(\mathbf{x}) = x_1^2 + 2x_2^2 \tag{8.2-3}$$

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 2x_1 \\ 4x_2 \end{bmatrix} \tag{8.2-4}$$

At $\mathbf{x} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$,

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{8.2-5}$$

For $\mathbf{p} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$: parallel to gradient,

$$\text{Directional derivative } \frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|} = \frac{\begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}}{\sqrt{1^2 + 2^2}} = \sqrt{5} \tag{8.2-6}$$

For $\mathbf{p} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$: orthogonal to gradient,

$$\text{Directional derivative } \frac{\mathbf{p}^T \nabla F(\mathbf{x})}{\|\mathbf{p}\|} = \frac{\begin{bmatrix} 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix}}{\sqrt{1^2 + 2^2}} = 0 \tag{8.2-7}$$

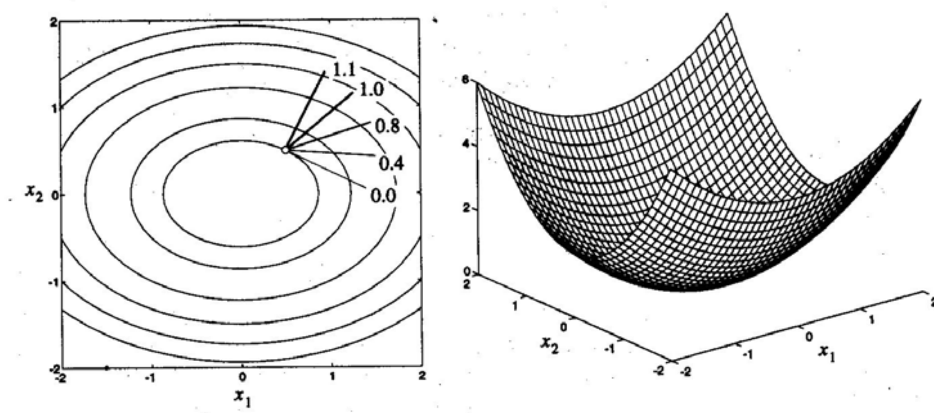


Figure 8.2-1 Quadratic Function and Directional Derivatives

8.3 Minima

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

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

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

Examples:

$$F(x) = 3x^4 - 7x^2 - \frac{1}{2}x + 6 \tag{8.3-1}$$

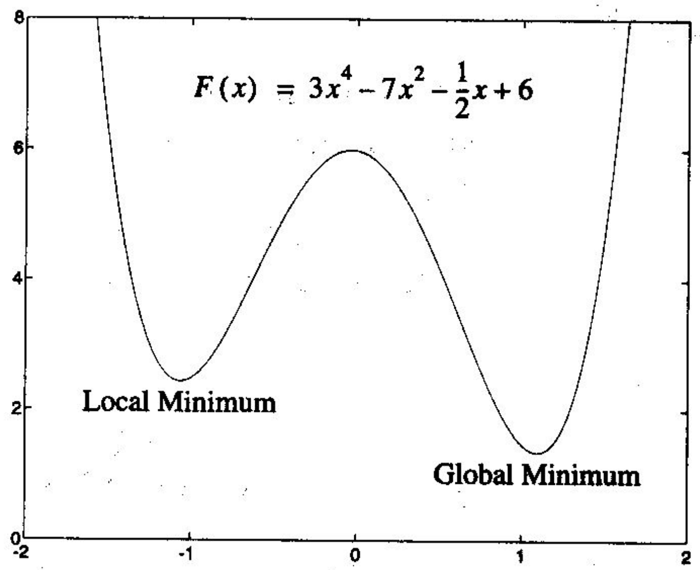


Figure 8.3-1 Scalar Example of Local and Global Minima

$$F(\mathbf{x}) = (x_2 - x_1)^4 + 8x_1x_2 - x_1 + x_2 + 3 \quad (8.3\text{-}2)$$

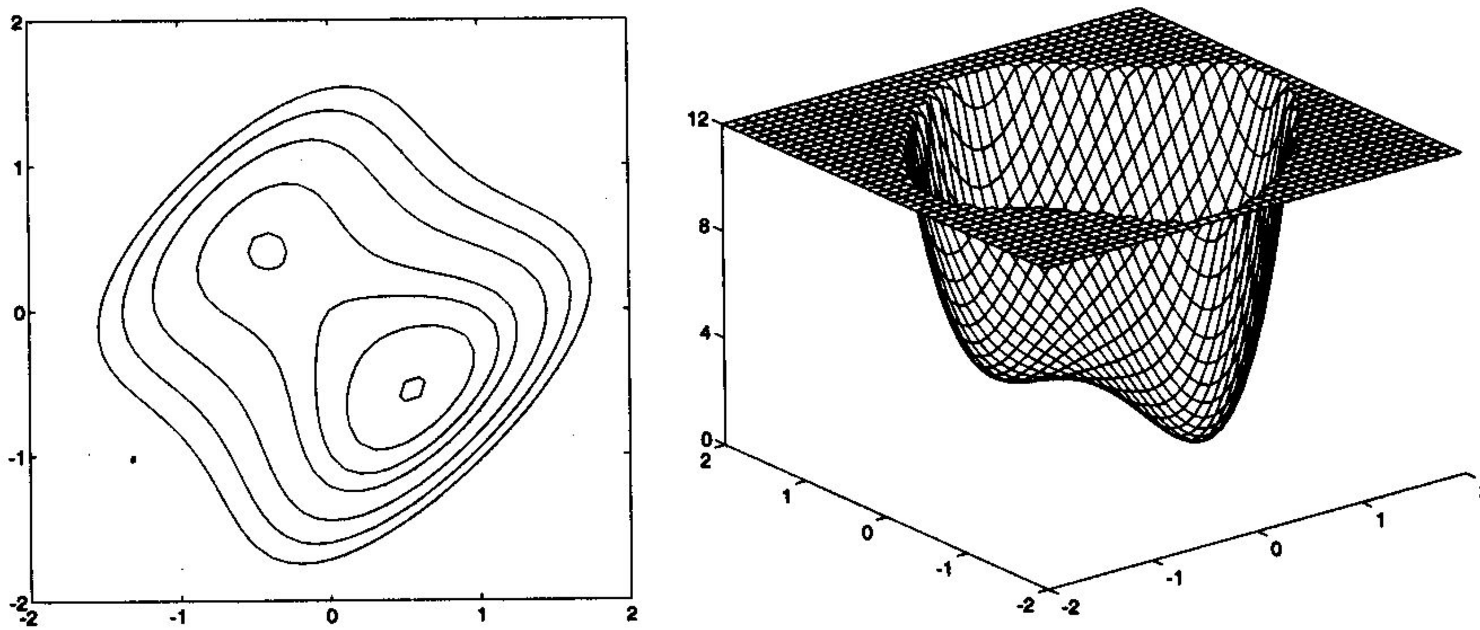


Figure 8.3-2 Vector Example of Minima and Saddle Point

$$F(\mathbf{x}) = (x_1^2 - 1.5x_1x_2 + 2x_2^2)x_1^2 \qquad (8.3\text{-}3)$$

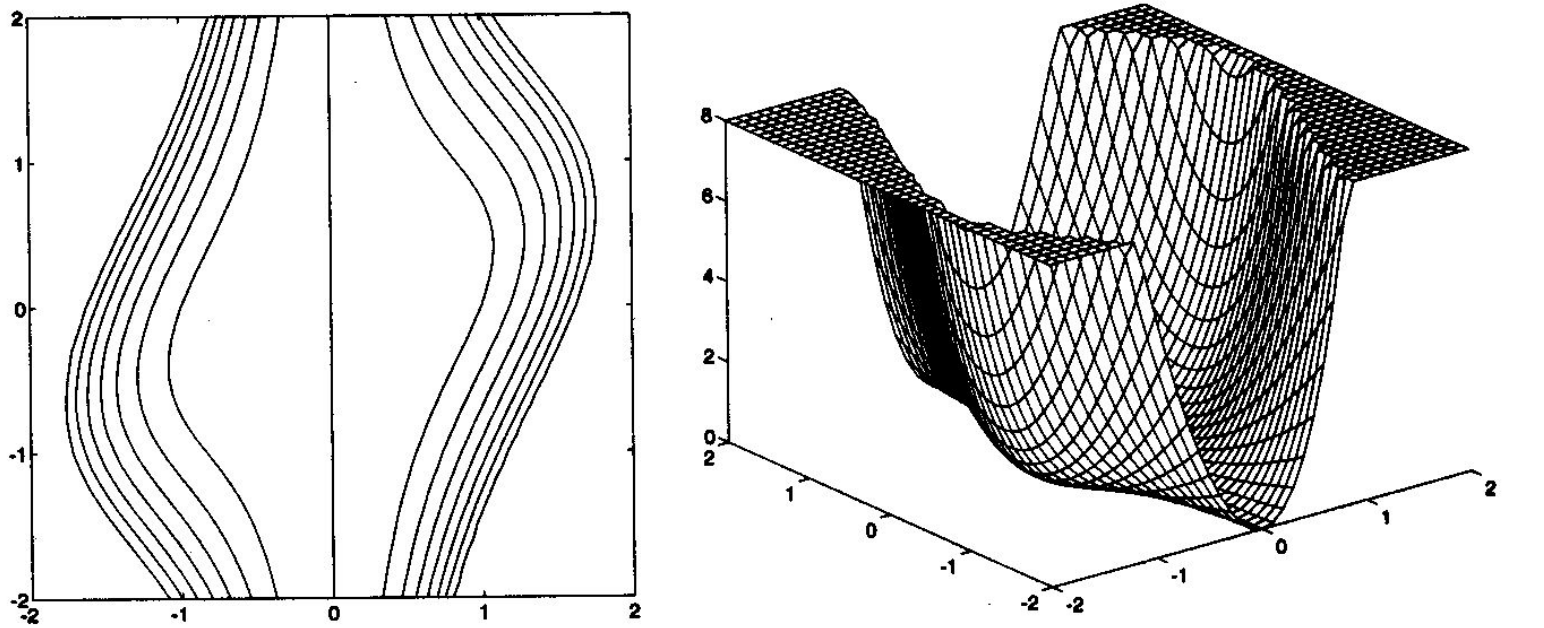


Figure 8.3-3 Weak Minimum Example

8.4 Necessary Conditions for Optimality

8.4.1 First-Order Conditions

Approximation by Taylor's series

$$F(\mathbf{x}^* + \Delta\mathbf{x}) \cong F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} \tag{8.4.1-1}$$

If **x*** is minimum point, the second term must be non negative,

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} \geq 0 \tag{8.4.1-2}$$

For positive of the second term,

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} > 0 \tag{8.4.1-3}$$

$$F(\mathbf{x}^* + \Delta\mathbf{x}) \cong F(\mathbf{x}^*) - \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} < F(\mathbf{x}^*) \tag{8.4.1-4}$$

For zero of the second term,

$$\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} = 0 \tag{8.4.1-5}$$

$$\nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} = 0 \tag{8.4.1-6}$$

- The gradient must be zero at a minimum point.
- This is a first-order, necessary (but not sufficient) condition for **x*** to be a local minimum point.
- Any points that have zero gradient are called stationary points.

8.4.2 Second-Order Conditions

Approximation by Taylor's series

$$F(\mathbf{x}^* + \Delta\mathbf{x}) = F(\mathbf{x}^*) + \frac{1}{2}\Delta\mathbf{x}^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} + \cdots \quad (8.4.2\text{-}1)$$

As If **x*** is minimum point, the second term must be non negative,

$$\Delta\mathbf{x}^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*} \Delta\mathbf{x} \geq 0 \quad (8.4.2\text{-}2)$$

Matrix **A** is positive definite (all eigenvalues are positive), if

$$\mathbf{z}^T \mathbf{A}\mathbf{z} > 0 \quad (8.4.2\text{-}3)$$

Matrix **A** is positive semidefinite (all eigenvalues are non negative) for any vector **z**≠0, if

$$\mathbf{z}^T \mathbf{A}\mathbf{z} \geq 0 \quad (8.4.2\text{-}4)$$

- 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.
- The second-order, necessary condition for a strong minimum is that the Hessian matrix be positive definite.

8.5 Quadratic Functions

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c \quad (8.5\text{-}1)$$

where, the matrix **A** is symmetric.

$$\nabla(\mathbf{h}^T\mathbf{x}) = \nabla(\mathbf{x}^T\mathbf{h}) = \mathbf{h} \quad (8.5\text{-}2)$$

where **h** is a constant vector,

$$\nabla\mathbf{x}^T\mathbf{Q}\mathbf{x} = \mathbf{Q}\mathbf{x} + \mathbf{Q}^T\mathbf{x} = 2\mathbf{Q}\mathbf{x} \text{ (for symmetric } \mathbf{Q}) \quad (8.5\text{-}3)$$

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} \quad (8.5\text{-}4)$$

$$\nabla^2 F(\mathbf{x}) = \mathbf{A} \quad (8.5\text{-}5)$$

- All higher derivatives of the quadratic function are zero.
- The first three terms of the Taylor series expansion give an exact representation of the function.
- All analytic functions behave like quadratics over a small neighborhood.

8.5.1 Eigensystem of the Hessian

Consider a quadratic function that has a stationary point at the origin, and whose value there is zero:

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} \tag{8.5.1-1}$$

The basis vectors are changed to the basis vectors of eigenvectors of the Hessian matrix, **A**.

- **A** is symmetric, its normalized eigenvectors will be mutually orthogonal.

$$\mathbf{B} = \begin{bmatrix}\mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_n\end{bmatrix} \tag{8.5.1-2}$$

$$\mathbf{B}^{-1} = \mathbf{B}^T \tag{8.5.1-2}$$

$$\mathbf{A}' = \left[\mathbf{B}^T\mathbf{A}\mathbf{B}\right] = \begin{bmatrix}\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n\end{bmatrix} = \Lambda \tag{8.5.1-3}$$

$$\mathbf{A} = \mathbf{B}\Lambda\mathbf{B}^T \tag{8.5.1-4}$$

The second derivative of a function $F(\mathbf{x})$ in the direction of a vector $\mathbf{p}$,

$$\frac{\mathbf{p}^T\nabla^2 F(\mathbf{x})\mathbf{p}}{\|\mathbf{p}\|^2} = \frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} \tag{8.5.1-5}$$

$$\mathbf{p} = \mathbf{B}\mathbf{c} \tag{8.5.1-6}$$

where **c** is the representation of the vector **p** with respect to the eigenvectors of **A**.

$$\frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} = \frac{\mathbf{c}^T\mathbf{B}^T(\mathbf{B}\Lambda\mathbf{B}^T)\mathbf{B}\mathbf{c}}{\mathbf{c}^T\mathbf{B}^T\mathbf{B}\mathbf{c}} = \frac{\mathbf{c}^T\Lambda\mathbf{c}}{\mathbf{c}^T\mathbf{c}} = \frac{\sum_{i=1}^{n}\lambda_i c_i^2}{\sum_{i=1}^{n} c_i^2} \tag{8.5.1-7}$$

- The second derivative is a weighted average of the eigenvalues.
- The second derivative can never be larger that the largest eigenvalue, or smaller than the smallest eigenvalue.

$$\lambda_{\min} \le \frac{\mathbf{p}^T\mathbf{A}\mathbf{p}}{\|\mathbf{p}\|^2} \le \lambda_{\max} \tag{8.5.1-8}$$

- The maximum second derivative occurs in the direction of the eigenvector that corresponds to the largest eigenvalue.
- 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 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.

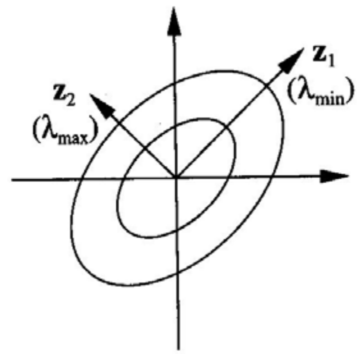


Figure 8.5.1-1 The principal Axes of the Function Contours

Example:

$$F(\mathbf{x}) = x_1^2 + x_2^2 = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \mathbf{x} \qquad (8.5.1\text{-}9)$$

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \ \lambda_1 = 2 \text{ and } \mathbf{z}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \lambda_2 = 2 \text{ and } \mathbf{z}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad (8.5.1\text{-}10)$$

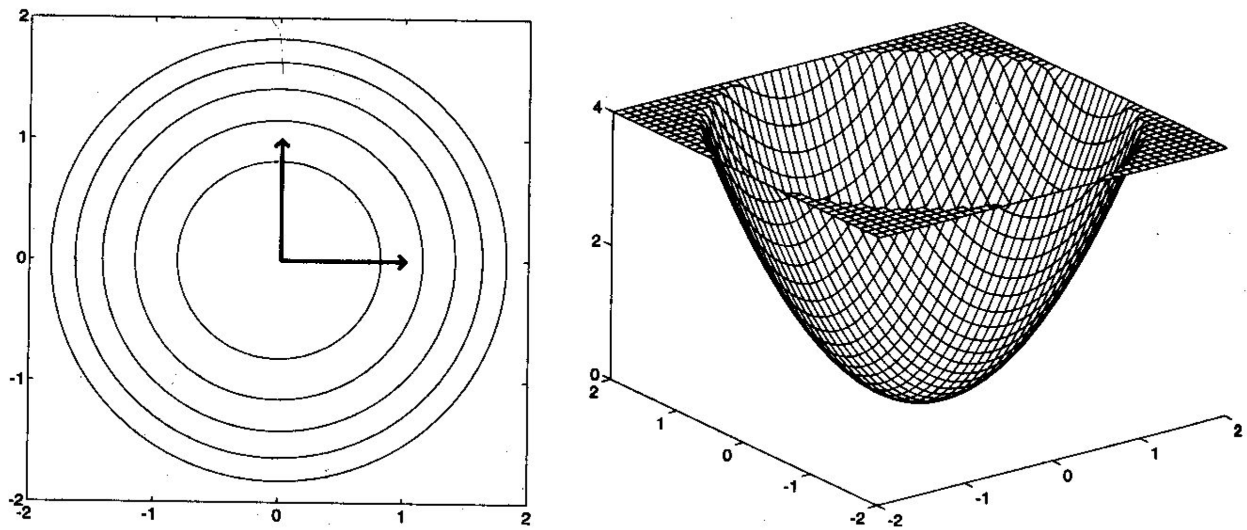


Figure 8.5.1-2 Circular Hollow

Example:

$$F(\mathbf{x}) = x_1^2 + x_1x_2 + x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}2 & 1\\1 & 2\end{bmatrix}\mathbf{x} \quad (8.5.1\text{-}11)$$

$$\nabla^2F(\mathbf{x}) = \begin{bmatrix}2 & 1\\1 & 2\end{bmatrix},\ \lambda_1 = 1 \text{ and } \mathbf{z}_1 = \begin{bmatrix}1\\-1\end{bmatrix},\ \lambda_2 = 3 \text{ and } \mathbf{z}_2 = \begin{bmatrix}1\\1\end{bmatrix} \quad (8.5.1\text{-}12)$$

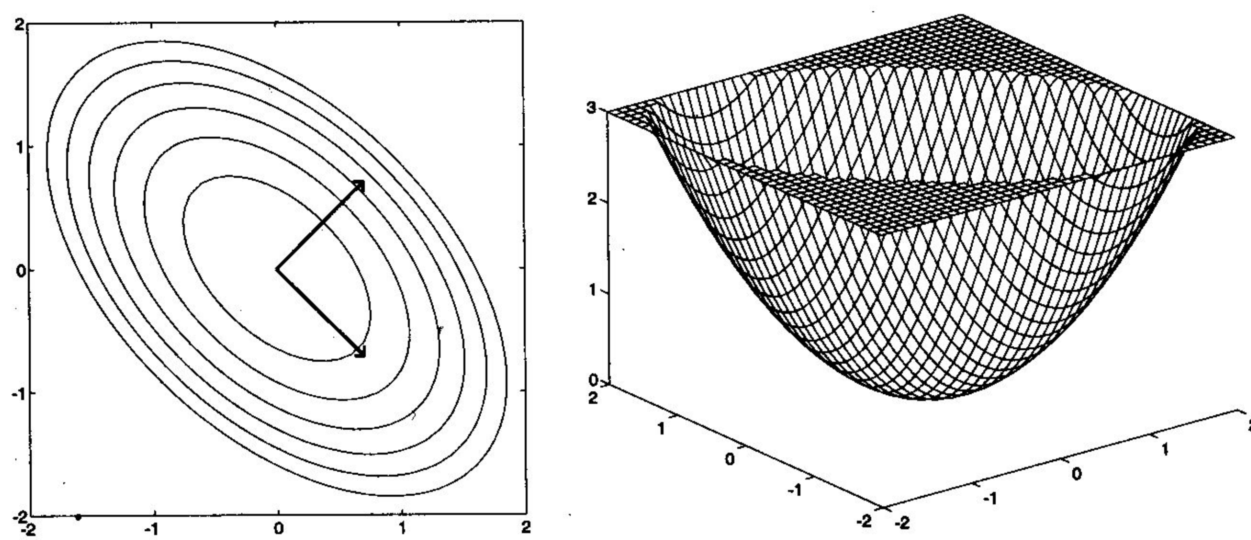


Figure 8.5.1-3 Elliptic Hollow

Example:

$$F(\mathbf{x}) = -\frac{1}{4}x_1^2 - \frac{3}{2}x_1x_2 - \frac{1}{4}x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}-0.5 & -1.5\\ -1.5 & -0.5\end{bmatrix}\mathbf{x} \quad (8.5.1\text{-}13)$$

$$\nabla^2F(\mathbf{x}) = \begin{bmatrix}-0.5 & -1.5\\ -1.5 & -0.5\end{bmatrix},\ \lambda_1 = 1 \text{ and } \mathbf{z}_1 = \begin{bmatrix}-1\\ 1\end{bmatrix},\ \lambda_2 = -2 \text{ and } \mathbf{z}_2 = \begin{bmatrix}-1\\ -1\end{bmatrix} \quad (8.5.1\text{-}14)$$

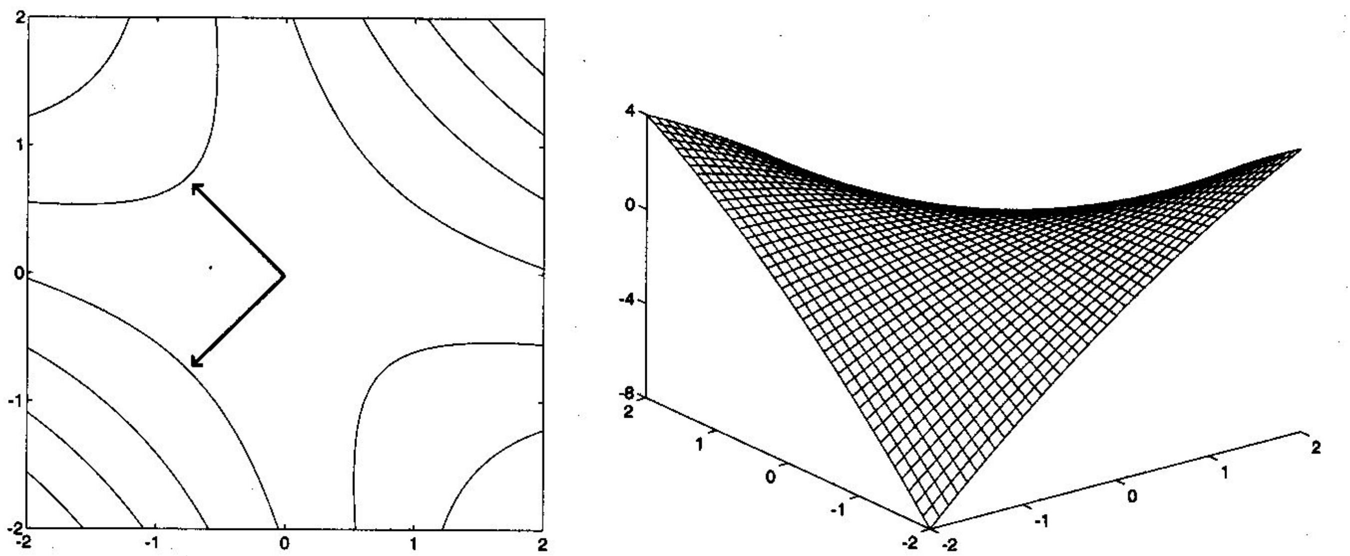


Figure 8.5.1-4 Elongated Saddle

Example:

$$F(\mathbf{x}) = \frac{1}{2}x_1^2 - x_1x_2 + \frac{1}{2}x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\mathbf{x} \tag{8.5.1-16}$$

$$\nabla^2 F(\mathbf{x}) = \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix},\ \lambda_1 = 1 \text{ and } \mathbf{z}_1 = \begin{bmatrix}-1\\ 1\end{bmatrix},\ \lambda_2 = 0 \text{ and } \mathbf{z}_2 = \begin{bmatrix}-1\\ -1\end{bmatrix} \tag{8.5.1-17}$$

A weak minimum along the line

$$x_1 = x_2 \tag{8.5.1-18}$$

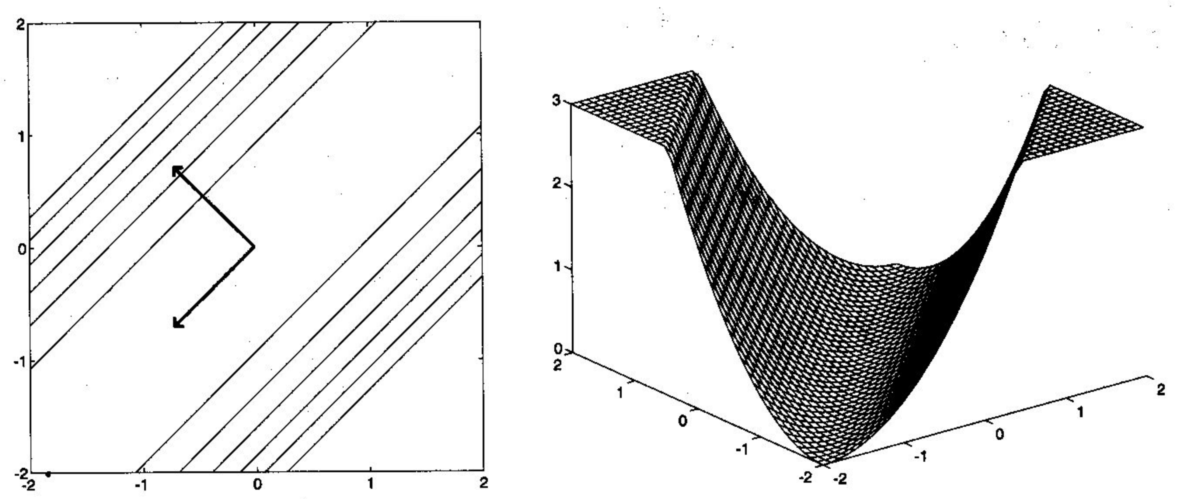


Figure 8.5.1-5 Stationary Valley

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** or will have **no stationary point**.
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**.

Stationary point of a quadratic equation,

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c \qquad (8.5.1\text{-}19)$$

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} \qquad (8.5.1\text{-}20)$$

$$\mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{d} \qquad (8.5.1\text{-}21)$$

- If **A** is not invertible (has some zero eigenvalues) and **d** is nonzero then no stationary points will exist.

9 Performance Optimization

9.1 Steepest Descent

Updating to minimum point,

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k \quad (9.1\text{-}1)$$

$$F(\mathbf{x}_{k+1}) < F(\mathbf{x}_k) \quad (9.1\text{-}2)$$

$$F(\mathbf{x}_{k+1}) = F(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx F(\mathbf{x}_k) + \mathbf{g}_k^T \Delta\mathbf{x}_k \quad (9.1\text{-}3)$$

$$\mathbf{g}_k \equiv \nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_k} \quad (9.1\text{-}4)$$

$$\mathbf{g}_k^T \Delta\mathbf{x}_k = \alpha_k \mathbf{g}_k^T \mathbf{p}_k < 0 \quad (9.1\text{-}5)$$

$$\mathbf{g}_k^T \mathbf{p}_k < 0 \quad (9.1\text{-}6)$$

The direction of steepest descent,

$$\mathbf{g}_k^T \mathbf{p}_k : \text{most negative} \quad (9.1\text{-}7)$$

$$\mathbf{p}_k = -\mathbf{g}_k \quad (9.1\text{-}8)$$

The method of steepest descent,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k \quad (9.1\text{-}9)$$

Learning rate determination, α_k,

1. Stable learning rate: Using fixed learning rate or predetermined learning rate
2. Minimizing along a line: Minimize the performance index $F(\mathbf{x})$ with respect to α_k at each iteration

$$\mathbf{x}_k - \alpha_k \mathbf{g}_k \quad (9.1\text{-}10)$$

Consider

$$F(\mathbf{x}) = x_1^2 + 25x_2^2 \quad (9.1\text{-}11)$$

$$\mathbf{x}_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \quad (9.1\text{-}12)$$

$$\nabla F(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} F(\mathbf{x}) \\ \frac{\partial}{\partial x_2} F(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 50x_2 \end{bmatrix} \quad (9.1\text{-}13)$$

$$\mathbf{g}_0 = \nabla F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_0} = \begin{bmatrix} 1 \\ 25 \end{bmatrix} \quad (9.1\text{-}14)$$

Fixed learning rate of $\alpha = 0.01$,

$$\mathbf{x}_1 = \mathbf{x}_0 - \alpha \mathbf{g}_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - 0.01 \begin{bmatrix} 1 \\ 25 \end{bmatrix} = \begin{bmatrix} 0.49 \\ 0.25 \end{bmatrix} \quad (9.1\text{-}15)$$

$$\mathbf{x}_2 = \mathbf{x}_1 - \alpha \mathbf{g}_1 = \begin{bmatrix} 0.49 \\ 0.25 \end{bmatrix} - 0.01 \begin{bmatrix} 0.98 \\ 12.5 \end{bmatrix} = \begin{bmatrix} 0.4802 \\ 0.125 \end{bmatrix} \quad (9.1\text{-}16)$$

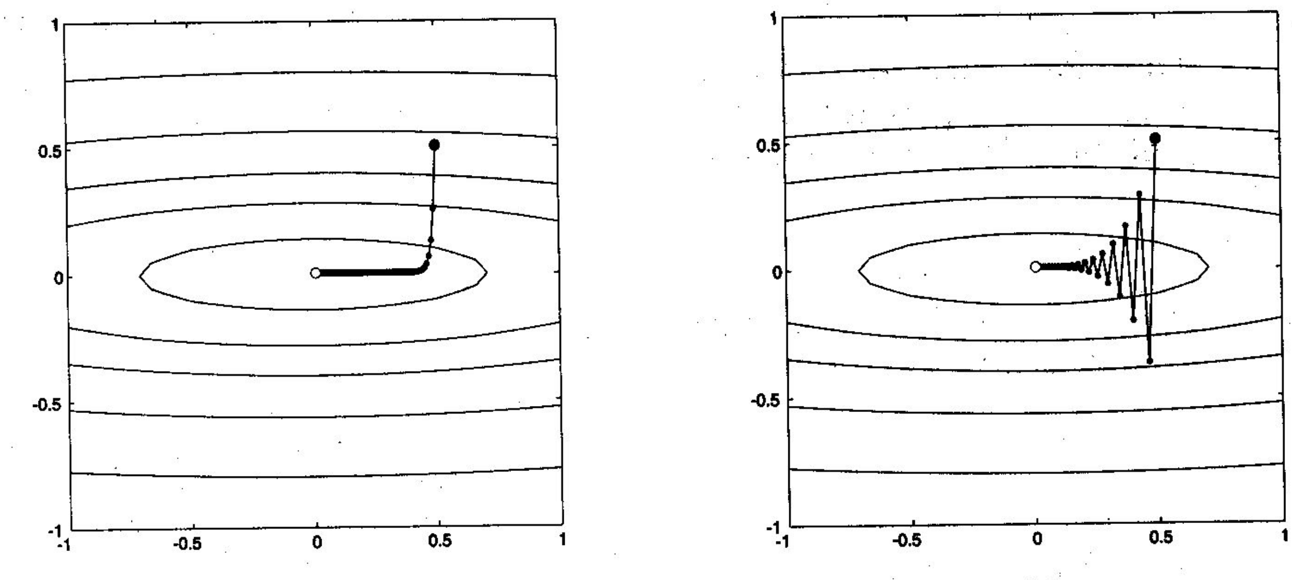


Figure 9.1-1 Trajectory for Steepest Descent with α = 0.01 and 0.035

9.1.1 Stable Learning Rates

For quadratic performance index,

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c \tag{9.1.1-1}$$

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} \tag{9.1.1-2}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\mathbf{g}_k = \mathbf{x}_k - \alpha(\mathbf{A}\mathbf{x}_k + \mathbf{d}) \tag{9.1.1-3}$$

$$\mathbf{x}_{k+1} = [\mathbf{I} - \alpha\mathbf{A}]\mathbf{x}_k - \alpha\mathbf{d} \tag{9.1.1-4}$$

- This linear dynamic system is stable if the eigenvalues of the matrix $[\mathbf{I}\text{-}\alpha\mathbf{A}]$ are less than one in magnitude.

$\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$: eigenvalues of the Hessian matrix, $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$: eigenvectors of the Hessian matrix,

$$[\mathbf{I} - \alpha\mathbf{A}]\mathbf{z}_i = \mathbf{z}_i - \alpha\mathbf{A}\mathbf{z}_i = \mathbf{z}_i - \alpha\lambda_i\mathbf{z}_i = (1 - \alpha\lambda_i)\mathbf{z}_i \tag{9.1.1-5}$$

- The eigenvectors of $[\mathbf{I}\text{-}\alpha\mathbf{A}]$ are the same as the eigenvectors of $\mathbf{A}$, and the eigenvalues of $[\mathbf{I}\text{-}\alpha\mathbf{A}]$ are $(1\text{-}\alpha\lambda_i)$.

The condition for the stability of the steepest descent algorithm,

$$\left|(1 - \alpha\lambda_i)\right| < 1 \tag{9.1.1-6}$$

$$0 < \alpha < \frac{2}{\lambda_i} \tag{9.1.1-7}$$

$$\alpha < \frac{2}{\lambda_{\max}} \tag{9.1.1-8}$$

Example:

$$F(\mathbf{x}) = x_1^2 + 25x_2^2 = \frac{1}{2}\mathbf{x}^T\begin{bmatrix}2 & 0\\0 & 50\end{bmatrix}\mathbf{x}, \mathbf{A} = \begin{bmatrix}2 & 0\\0 & 50\end{bmatrix} \qquad (9.1.1\text{-}9)$$

$$\lambda_1 = 2 \text{ and } \mathbf{z}_1 = \begin{bmatrix}1\\0\end{bmatrix},\ \lambda_2 = 50 \text{ and } \mathbf{z}_2 = \begin{bmatrix}0\\1\end{bmatrix} \qquad (9.1.1\text{-}10)$$

$$\alpha < \frac{2}{\lambda_{\max}} = \frac{2}{50} = 0.04 \qquad (9.1.1\text{-}11)$$

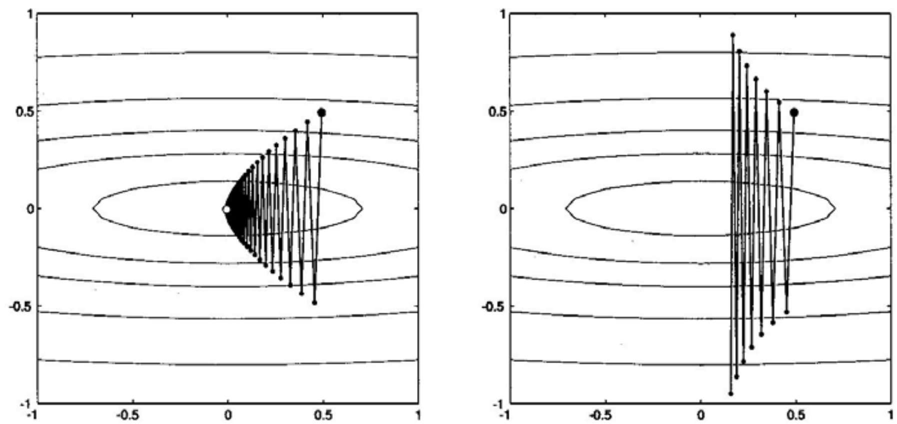


Figure 9.1.1-1 Trajectories for α = 0.039 (Left), and α = 0.041 (Right)

9.1.2 Minimizing Along a Line

$$F(\mathbf{x}_{k+1}) = F(\mathbf{x}+\Delta\mathbf{x}) = F(\mathbf{x}_k + \alpha_k \mathbf{p}_k) = F(\mathbf{x}) \approx F(\mathbf{x}^*) + \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x}-\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x}-\mathbf{x}^*) \quad (9.1.2\text{-}1)$$

$$\frac{d}{d\alpha_k} F(\mathbf{x}_k + \alpha_k \mathbf{p}_k) = \nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}_k}\mathbf{p}_k + \alpha_k \mathbf{p}_k^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_k}\mathbf{p}_k \quad (9.1.2\text{-}2)$$

$$\alpha_k = -\frac{\nabla F(\mathbf{x})^T\Big|_{\mathbf{x}=\mathbf{x}_k}\mathbf{p}_k}{\mathbf{p}_k^T \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_k}\mathbf{p}_k} = -\frac{\mathbf{g}_k^T \mathbf{p}_k}{\mathbf{p}_k^T \mathbf{A} \mathbf{p}_k} \quad (9.1.2\text{-}3)$$

$$\mathbf{A}_k = \nabla^2 F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_k} \quad (9.1.2\text{-}4)$$

Example:

$$F(\mathbf{x}) = x_1^2 + x_1 x_2 + x_2^2 = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \mathbf{x} \quad (9.1.2\text{-}5)$$

$$\mathbf{x}_0 = \begin{bmatrix} 0.8 \\ -0.25 \end{bmatrix} \quad (9.1.2\text{-}6)$$

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{bmatrix} \quad (9.1.2\text{-}7)$$

$$\mathbf{p}_0 = -\mathbf{g}_0 = -\nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_0} = \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix} \quad (9.1.2\text{-}8)$$

$$\alpha_0 = -\frac{\begin{bmatrix}1.35 & 0.3\end{bmatrix}\begin{bmatrix}-1.35 \\ -0.3\end{bmatrix}}{\begin{bmatrix}-1.35 & -0.3\end{bmatrix}\begin{bmatrix}2 & 1 \\ 1 & 2\end{bmatrix}\begin{bmatrix}-1.35 \\ -0.3\end{bmatrix}} = 0.413 \qquad (9.1.2\text{-}9)$$

$$\mathbf{x}_1 = \mathbf{x}_0 - \alpha_0 \mathbf{g}_0 = \begin{bmatrix}0.8 \\ -0.25\end{bmatrix} - 0.413\begin{bmatrix}1.35 \\ 0.3\end{bmatrix} = \begin{bmatrix}0.24 \\ -0.37\end{bmatrix} \qquad (9.1.2\text{ -}10)$$

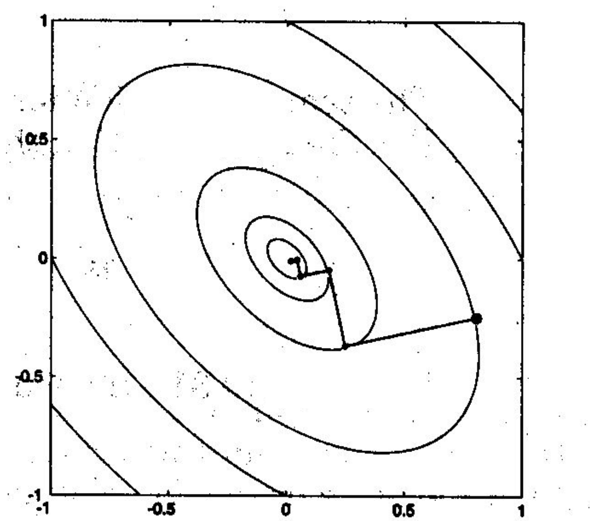


Figure 9.1.2-1 Steepest Descent with Minimization Along a Line

9.2 Newton's Method

$$F(\mathbf{x}_{k+1}) = F(\mathbf{x}_k + \Delta\mathbf{x}_k) \approx F(\mathbf{x}_k) + \mathbf{g}_k^T \Delta\mathbf{x}_k + \frac{1}{2}\Delta\mathbf{x}_k^T \mathbf{A}_k \Delta\mathbf{x}_k \qquad (9.2\text{-}1)$$

For quadratic function of

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{d}^T \mathbf{x} + c \qquad (9.2\text{-}2)$$

$$\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{d} \qquad (9.2\text{-}3)$$

$$\mathbf{g}_k + \mathbf{A}_k \Delta\mathbf{x}_k = 0 \qquad (9.2\text{-}4)$$

$$\Delta\mathbf{x}_k = -\mathbf{A}_k^{-1}\mathbf{g}_k \qquad (9.2\text{-}5)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{A}_k^{-1}\mathbf{g}_k \qquad (9.2\text{-}6)$$

Example:

$$F(\mathbf{x}) = x_1^2 + 25x_2^2 \qquad (9.2\text{-}7)$$

$$\nabla F(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} F(\mathbf{x}) \\ \frac{\partial}{\partial x_2} F(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 50x_2 \end{bmatrix}, \ \nabla^2 F(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 50 \end{bmatrix} \qquad (9.2\text{-}8)$$

$$\mathbf{x}_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \qquad (9.2\text{-}9)$$

$$\mathbf{x}_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 50 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 25 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (9.2\text{-}10)$$

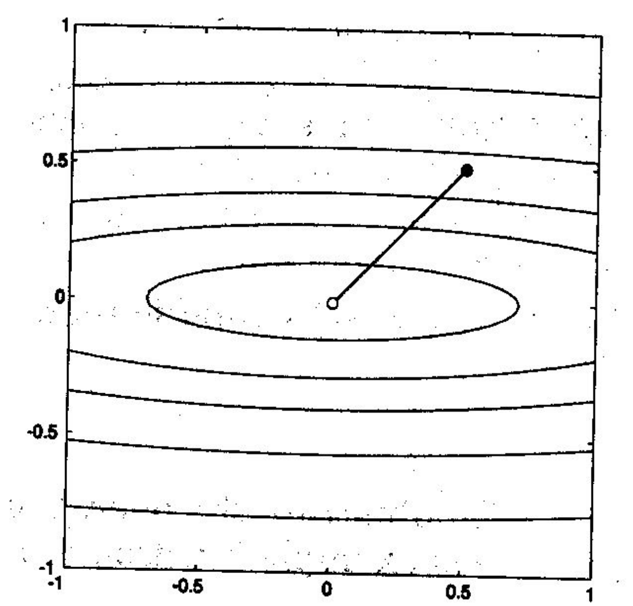


Figure 9.2-1 Trajectory for Newton's Method

- If the function $F(\mathbf{x})$ is not quadratic, then Newton's method will not generally converge in one step.

Example:

$$F(\mathbf{x}) = (x_2 - x_1)^4 + 8x_1x_2 - x_1 + x_2 + 3 \tag{9.2-11}$$

Three stationary points:

$$\mathbf{x}^1 = \begin{bmatrix} -0.42 \\ 0.42 \end{bmatrix},\ \mathbf{x}^2 = \begin{bmatrix} -0.13 \\ 0.13 \end{bmatrix}, \text{ and } \mathbf{x}^3 = \begin{bmatrix} 0.55 \\ -0.55 \end{bmatrix} \tag{9.2-12}$$

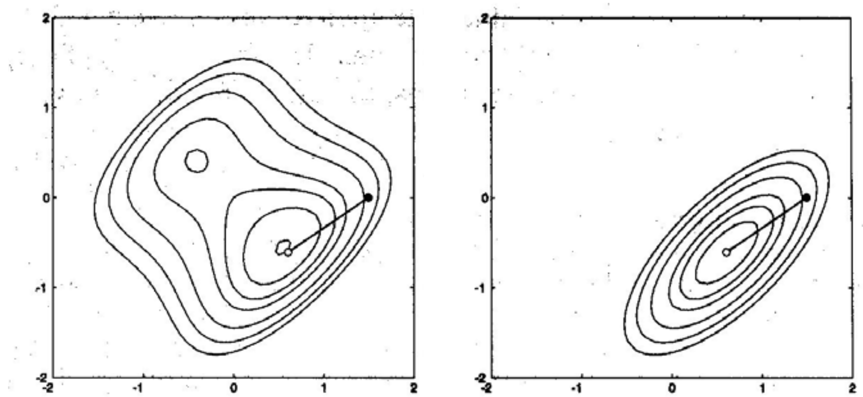


Figure 9.2-2 One Iteration of Newton's Method from $\mathbf{x}_0 = [1.5\ 0]^T$

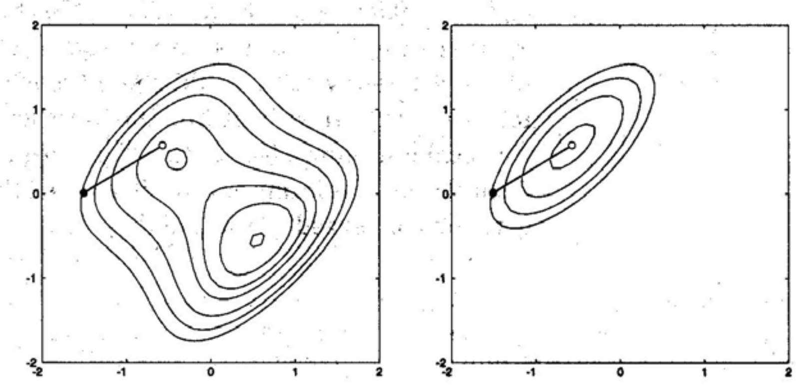


Figure 9.2-3 One Iteration of Newton's Method from $\mathbf{x}_0 = [-1.5\ 0]^T$

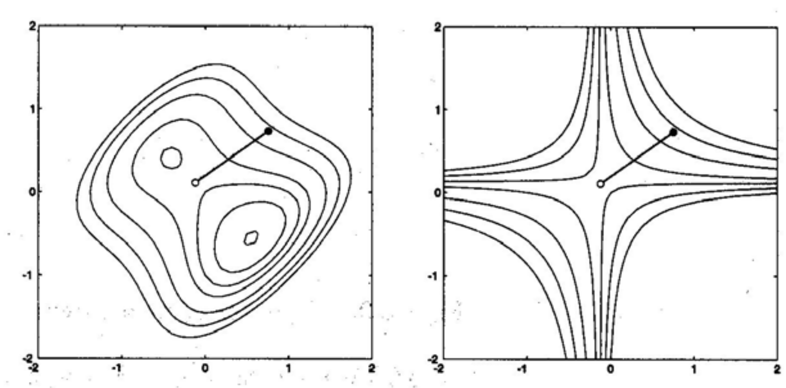


Figure 9.2-4 One Iteration of Newton's Method from $\mathbf{x}_0 = [0.75\ 0.75]^T$

9.3 Conjugate Gradient

For a quadratic function,

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c \qquad (9.3\text{-}1)$$

Definition: A set of vectors $\{\mathbf{p}_k\}$ is **mutually conjugate** with respect to a positive definite Hessian matrix **A** if and only if

$$\mathbf{p}_k^T\mathbf{A}\mathbf{p}_j = 0 \text{ for } k \neq j \qquad (9.3\text{-}2)$$

- If we make a sequence of exact linear searches along any set of conjugate directions $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$, then the exact minimum of any quadratic function, with n parameters, will be reached in at most n searches.

The change in the gradient at iteration k+1 for quadratic functions,

$$\Delta\mathbf{g}_k = \mathbf{g}_{k+1} - \mathbf{g}_k = (\mathbf{A}\mathbf{x}_{k+1} + \mathbf{d}) - (\mathbf{A}\mathbf{x}_k + \mathbf{d}) = \mathbf{A}\Delta\mathbf{x}_k \qquad (9.3\text{-}3)$$

$$\Delta\mathbf{x}_k = (\mathbf{x}_{k+1} - \mathbf{x}_k) = \alpha_k\mathbf{p}_k \qquad (9.3\text{-}4)$$

$$\mathbf{p}_k^T\mathbf{A}\mathbf{p}_j = \alpha_k\mathbf{p}_k^T\mathbf{A}\mathbf{p}_j = \Delta\mathbf{x}_k^T\mathbf{A}\mathbf{p}_j = \Delta\mathbf{g}_k^T\mathbf{p}_j = 0 \text{ for } k \neq j \qquad (9.3\text{-}5)$$

- The search directions will be conjugate if they are orthogonal to the changes in the gradient.

$$\mathbf{p}_0 = -\mathbf{g}_0 \qquad (9.3\text{-}6)$$

$$\mathbf{p}_k = -\mathbf{g}_k + \beta_k\mathbf{p}_{k-1} \qquad (9.3\text{-}7)$$

By Hestenes and Steifel,

$$\beta_k = \frac{\Delta \mathbf{g}_{k-1}^T \mathbf{g}_k}{\Delta \mathbf{g}_{k-1}^T \mathbf{p}_k} \tag{9.3-8}$$

Fletcher and Reeves,

$$\beta_k = \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}} \tag{9.3-9}$$

Polak and Ribiere,

$$\beta_k = \frac{\Delta \mathbf{g}_{k-1}^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}} \tag{9.3-10}$$

The conjugate gradient method consists of the following steps:

1. Select the first search direction to be the negative of the gradient, $\mathbf{p}_0 = -\mathbf{g}_0$.
2. Take a step according to $\Delta \mathbf{x}_k = \alpha_k \mathbf{p}_k$, selecting the learning rate α_k to minimize the function along the search direction.
3. Select the next search direction according to $\mathbf{p}_k = -\mathbf{g}_k + \beta_k \mathbf{p}_{k-1}$, using $\frac{\Delta \mathbf{g}_{k-1}^T \mathbf{g}_k}{\Delta \mathbf{g}_{k-1}^T \mathbf{p}_k}$, $\frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}}$, or $\frac{\Delta \mathbf{g}_{k-1}^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}}$ to calculate β_k..
4. If the algorithm has not converged, return to step 2.

Example:

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \mathbf{x} \tag{9.3-11}$$

$$\mathbf{x}_0 = \begin{bmatrix} 0.8 \\ -0.25 \end{bmatrix} \tag{9.3-12}$$

$$\nabla F(\mathbf{x}) = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 2x_2 \end{bmatrix} \tag{9.3-13}$$

$$\mathbf{p}_0 = -\mathbf{g}_0 = -\nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_0} = \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix} \tag{9.3-14}$$

$$\alpha_0 = -\frac{\begin{bmatrix} 1.35 & 0.3 \end{bmatrix} \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix}}{\begin{bmatrix} -1.35 & -0.3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix}} = 0.413 \tag{9.3-15}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + \alpha_0 \mathbf{p}_0 = \begin{bmatrix} 0.8 \\ -0.25 \end{bmatrix} + 0.413 \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix} = \begin{bmatrix} 0.24 \\ -0.37 \end{bmatrix} \tag{9.3-16}$$

$$\mathbf{g}_1 = \nabla F(\mathbf{x})\Big|_{\mathbf{x}=\mathbf{x}_1} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 0.24 \\ -0.3 \end{bmatrix} = \begin{bmatrix} 0.11 \\ -0.5 \end{bmatrix} \tag{9.3-17}$$

$$\beta_1 = \frac{\mathbf{g}_1^T \mathbf{g}_1}{\mathbf{g}_0^T \mathbf{g}_0} = \frac{\begin{bmatrix} 0.11 & -0.5 \end{bmatrix} \begin{bmatrix} 0.11 \\ -0.5 \end{bmatrix}}{\begin{bmatrix} 1.35 & 0.3 \end{bmatrix} \begin{bmatrix} 1.35 \\ 0.3 \end{bmatrix}} = \frac{0.2621}{1.9125} = 0.137 \tag{9.3-18}$$

Using the method of Fletcher and Reeves,

$$\mathbf{p}_1 = -\mathbf{g}_1 + \beta_1 \mathbf{p}_0 = \begin{bmatrix} -0.11 \\ 0.5 \end{bmatrix} + 0.137 \begin{bmatrix} -1.35 \\ -0.3 \end{bmatrix} = \begin{bmatrix} -0.295 \\ 0.459 \end{bmatrix} \tag{9.3-19}$$

$$\alpha_1 = -\frac{\begin{bmatrix} 0.11 & -0.5 \end{bmatrix} \begin{bmatrix} -0.295 \\ 0.459 \end{bmatrix}}{\begin{bmatrix} -0.295 & 0.459 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -0.295 \\ 0.459 \end{bmatrix}} = 0.807 \tag{9.3-20}$$

$$\mathbf{x}_2 = \mathbf{x}_1 + \alpha_1 \mathbf{p}_1 = \begin{bmatrix} 0.24 \\ -0.37 \end{bmatrix} + 0.807 \begin{bmatrix} -0.295 \\ 0.459 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{9.3-21}$$

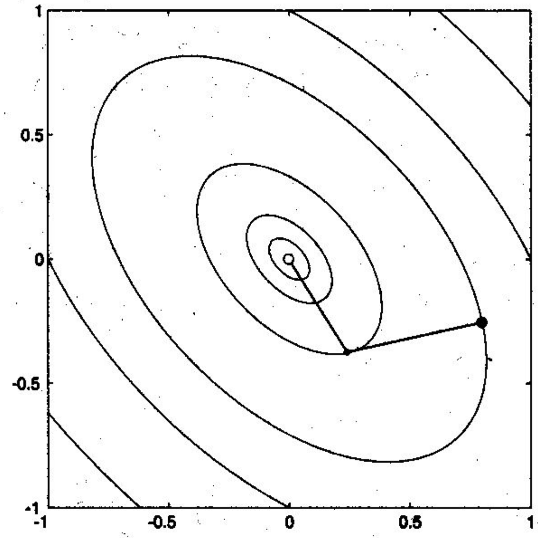


Figure 9.3-1 Conjugate Gradient Algorithm

10 Widrow-Hoff Learning

10. 1 ADALINE (ADAptive LInear NEuron) Network

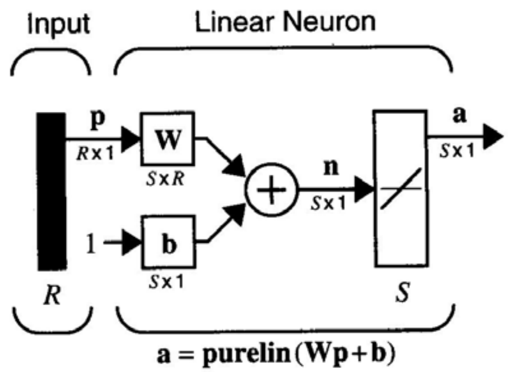


Figure 10.1-1 ADALINE Network

$$\mathbf{a} = \mathbf{purelin}(\mathbf{Wp}+\mathbf{b}) = \mathbf{Wp}+\mathbf{b} \tag{10.1-1}$$

$$a_i = purelin(n_i) = purelin({}_i\mathbf{w}^T\mathbf{p}+b_i) = {}_i\mathbf{w}^T\mathbf{p}+b_i \tag{10.1-2}$$

$${}_i\mathbf{w} = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ \vdots \\ w_{i,R} \end{bmatrix} \tag{10.1-3}$$

10.1.1 Single ADALINE

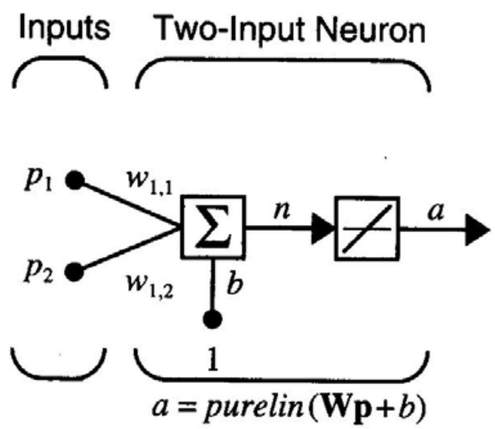


Figure 10.1.1-1 Two-Input Linear Neuron

$$a = purelin(n) = purelin({}_1\mathbf{w}^T\mathbf{p}+b) = {}_1\mathbf{w}^T\mathbf{p}+b = w_{1,1}p_1+w_{1,2}p_2+b \quad (10.1.1\text{-}1)$$

10.2 Mean Square Error

LMS algorithm: supervised training

$$\{\mathbf{p}_1, \mathbf{t}_1\}, \{\mathbf{p}_2, \mathbf{t}_2\}, \ldots, \{\mathbf{p}_Q, \mathbf{t}_Q\} \quad (10.2\text{-}1)$$

$$\mathbf{x} = \begin{bmatrix} {}_1\mathbf{w} \\ b \end{bmatrix} \quad (10.2\text{-}2)$$

$$\mathbf{z} = \begin{bmatrix} \mathbf{p} \\ 1 \end{bmatrix} \quad (10.2\text{-}3)$$

$$a = {}_1\mathbf{w}^T\mathbf{p}+b \quad (10.2\text{-}4)$$

$$a = \mathbf{x}^T\mathbf{z} \quad (10.2\text{-}5)$$

$$F(\mathbf{x}) = E[e^2] = E[(t-a)^2] = E[(t-\mathbf{x}^T\mathbf{z})^2] \quad (10.2\text{-}6)$$

$$F(\mathbf{x}) = E[t^2 - 2t\mathbf{x}^T\mathbf{z} + \mathbf{x}^T\mathbf{z}\mathbf{z}^T\mathbf{x}] = E[t^2] - 2\mathbf{x}^T E[t\mathbf{z}] + \mathbf{x}^T E[\mathbf{z}\mathbf{z}^T]\mathbf{x} \quad (10.2\text{-}7)$$

$$F(\mathbf{x}) = c - 2\mathbf{x}^T\mathbf{h} + \mathbf{x}^T\mathbf{R}\mathbf{x} \quad (10.2\text{-}8)$$

where

$$c = E[t^2],\ \mathbf{h} = E[t\mathbf{z}],\ \text{and}\ \mathbf{R} = E[\mathbf{z}\mathbf{z}^T] \quad (10.2\text{-}9)$$

General form quadratic function,

$$F(\mathbf{x}) = c + \mathbf{d}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} \quad (10.2\text{-}10)$$

$$\mathbf{d} = -2\mathbf{h}\ \text{and}\ \mathbf{A} = 2\mathbf{R} \quad (10.2\text{-}11)$$

The gradient,

$$\nabla F(\mathbf{x}) = \nabla\left(c + \mathbf{d}^T\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} \right) = \mathbf{d} + \mathbf{A}\mathbf{x} = -2\mathbf{h} + 2\mathbf{R}\mathbf{x} \tag{10.2-12}$$

The stationary point of the performance index,

$$-2\mathbf{h} + 2\mathbf{R}\mathbf{x} = 0 \tag{10.2-13}$$

$$\mathbf{x}^* = \mathbf{R}^{-1}\mathbf{h} \tag{10.2-14}$$

Example: $\left\{\mathbf{z}_1 = \begin{bmatrix}1\\2\end{bmatrix}, t_1 = 1\right\} \rightarrow 50\%, \left\{\mathbf{z}_2 = \begin{bmatrix}3\\4\end{bmatrix}, t_2 = -1\right\} \rightarrow 50\%,$

$$c = E\left[t^2\right] = 0.5(1)^2 + 0.5(-1)^2 = 1 \tag{10.2-15}$$

$$\mathbf{h} = E[t\mathbf{z}] = 0.5\left[1\begin{bmatrix}1\\2\end{bmatrix}\right] + 0.5\left[-1\begin{bmatrix}3\\4\end{bmatrix}\right] = \begin{bmatrix}-1\\-1\end{bmatrix} \tag{10.2-16}$$

$$\mathbf{R} = E\left[\mathbf{z}\mathbf{z}^T\right] = 0.5\begin{bmatrix}1\\2\end{bmatrix}\begin{bmatrix}1 & 2\end{bmatrix} + 0.5\begin{bmatrix}3\\4\end{bmatrix}\begin{bmatrix}3 & 4\end{bmatrix} = \begin{bmatrix}5 & 7\\7 & 10\end{bmatrix} \tag{10.2-17}$$

$$F(\mathbf{x}) = 1 - 2\mathbf{x}^T\begin{bmatrix}-1\\-1\end{bmatrix} + \mathbf{x}^T\begin{bmatrix}5 & 7\\7 & 10\end{bmatrix}\mathbf{x} \tag{10.2-18}$$

$$\nabla F(\mathbf{x}) = 2\begin{bmatrix}5 & 7\\7 & 10\end{bmatrix}\mathbf{x} - 2\begin{bmatrix}-1\\-1\end{bmatrix} = 0 \tag{10.2-19}$$

$$\mathbf{x} = \begin{bmatrix} 5 & 7 \\ 7 & 10 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -3 \\ 2 \end{bmatrix} = \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} \tag{10.2-20}$$

$\left\{ \mathbf{z}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, t_1 = 1 \right\}$,

$$a_1 = \begin{bmatrix} -3 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 1 = t_1 \tag{10.2-21}$$

$\left\{ \mathbf{z}_2 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, t_2 = -1 \right\}$,

$$a_2 = \begin{bmatrix} -3 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = -1 = t_2 \tag{10.2-22}$$

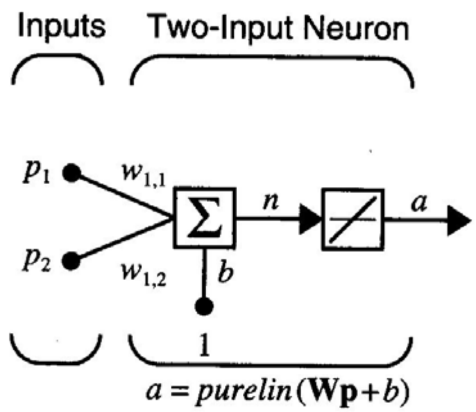


Figure 10.2-1 Network

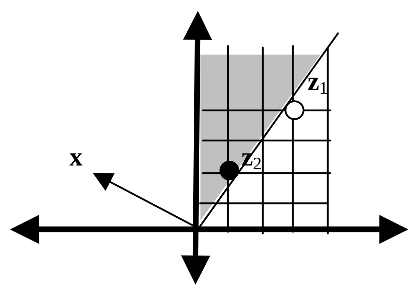


Figure 10.2-2 Weight Vector and Decision Boundary

10.3 LMS Algorithm

In Widrow and Hoff learning rule, the mean square error, $F(\mathbf{x})$, is estimated by

$$\hat{F}(\mathbf{x}) = (t(k) - a(k))^2 = e^2(k) \tag{10.3-1}$$

$$\hat{\nabla} F(\mathbf{x}) = \nabla e^2(k) \tag{10.3-2}$$

$$[\nabla e^2(k)]_j = \frac{\partial e^2(k)}{\partial w_{1,j}} = 2e(k)\frac{\partial e(k)}{\partial w_{1,j}} \text{ for } j = 1, 2, \ldots, R, \tag{10.3-3}$$

$$[\nabla e^2(k)]_{R+1} = \frac{\partial e^2(k)}{\partial b} = 2e(k)\frac{\partial e(k)}{\partial b} \tag{10.3-4}$$

$$\frac{\partial e(k)}{\partial w_{1,j}} = \frac{\partial [t(k) - a(k)]}{\partial w_{1,j}} = \frac{\partial}{\partial w_{1,j}}[t(k) - ({}_1\mathbf{w}^T \mathbf{p}(k) + b)] = \frac{\partial}{\partial w_{1,j}}\left[t(k) - \left(\sum_{i=1}^{R} w_{1,i} p_i(k) + b\right)\right] \tag{10.3-5}$$

$$\frac{\partial e(k)}{\partial w_{1,j}} = -p_j(k) \tag{10.3-6}$$

$$\frac{\partial e(k)}{\partial b} = -1 \tag{10.3-7}$$

$$\hat{\nabla} F(\mathbf{x}) = \nabla e^2(k) = -2e(k)\mathbf{z}(k) \tag{10.3-8}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla F(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_k} \tag{10.3-9}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + 2\alpha e(k)\mathbf{z}(k) \tag{10.3-10}$$

$$ {}_1\mathbf{w}(k+1) = {}_1\mathbf{w}(k) + 2\alpha e(k)\mathbf{p}(k) \tag{10.3-11} $$

$$ b(k+1) = b(k) + 2\alpha e(k) \tag{10.3-12} $$

- Widrow-Hoff learning algorithm is also called delta rule.

$$ {}_i\mathbf{w}(k+1) = {}_i\mathbf{w}(k) + 2\alpha e_i(k)\mathbf{p}(k) \tag{10.3-13} $$

$$ b_i(k+1) = b_i(k) + 2\alpha e_i(k) \tag{10.3-14} $$

$$ \mathbf{W}(k+1) = \mathbf{W}(k) + 2\alpha \mathbf{e}(k)\mathbf{p}^T(k) \tag{10.3-15} $$

$$ \mathbf{b}(k+1) = \mathbf{b}(k) + 2\alpha \mathbf{e}(k) \tag{10.3-16} $$

10.4 Example on Apple/Orange Recognition

For simplicity, a zero bias is used in the ADALINE network.

$$\left\{\mathbf{p}_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, t_1 = [-1]\right\} \rightarrow 50\%, \quad \left\{\mathbf{p}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, t_2 = [1]\right\} \rightarrow 50\% \tag{10.4-1}$$

$$c = E[t^2] = 0.5(-1)^2 + 0.5(1)^2 = 1 \tag{10.4-2}$$

$$\mathbf{h} = E[t\mathbf{z}] = 0.5\left[-1\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}\right] + 0.5\left[1\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}\right] = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{10.4-3}$$

$$\mathbf{R} = E[\mathbf{p}\mathbf{p}^T] = \frac{1}{2}\mathbf{p}_1\mathbf{p}_1^T + \frac{1}{2}\mathbf{p}_2\mathbf{p}_2^T = \frac{1}{2}\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & -1 & -1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \tag{10.4-4}$$

The eigenvalues of **R,**

$$\lambda_1 = 1.0, \qquad \lambda_2 = 0.0, \qquad \lambda_3 = 2.0 \tag{10.4-3}$$

$$\alpha < \frac{1}{\lambda_{\max}} = \frac{1}{2.0} = 0.5 \tag{10.4-4}$$

Select $\alpha = 0.2$,

Presenting orange, $\left\{ \mathbf{p}_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, t_1 = [-1] \right\}$,

$$a(0) = \mathbf{W}(0)\mathbf{p}(0) = \mathbf{W}(0)\mathbf{p}_1 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} = 0 \tag{10.4-5}$$

$$e(0) = t(0) - a(0) = t_1 - a(0) = -1 - 0 = -1 \tag{10.4-6}$$

$$\mathbf{W}(1) = \mathbf{W}(0) + 2\alpha e(0)\mathbf{p}^T(0) = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} + 2(0.2)(-1) \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}^T = \begin{bmatrix} -0.4 & 0.4 & 0.4 \end{bmatrix} \tag{10.4-7}$$

Presenting apple, $\left\{ \mathbf{p}_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, t_2 = [1] \right\}$,

$$a(1) = \mathbf{W}(1)\mathbf{p}(1) = \mathbf{W}(1)\mathbf{p}_2 = \begin{bmatrix} -0.4 & 0.4 & 0.4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} = -0.4 \tag{10.4-8}$$

$$e(1) = t(1) - a(1) = t_2 - a(1) = 1 - (-0.4) = 1.4 \tag{10.4-9}$$

$$\mathbf{W}(2) = \mathbf{W}(1) + 2\alpha e(1)\mathbf{p}^T(1) = \begin{bmatrix} -0.4 & 0.4 & 0.4 \end{bmatrix} + 2(0.2)(1.4)\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}^T = \begin{bmatrix} 0.16 & 0.96 & -0.16 \end{bmatrix} \quad (10.4\text{-}10)$$

If we continue this procedure, the algorithm converges to

$$\mathbf{W}(\infty) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \quad (10.4\text{-}11)$$

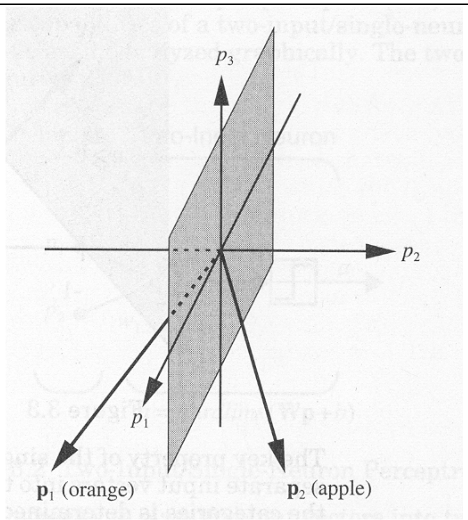


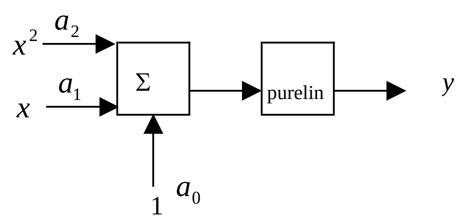
Figure 10.4-1 Prototype Vectors and Decision Boundary

- Decision boundary generated by ADALINE falls halfway between the two reference patterns.
- The perceptron rule does not produce halfway decision boundary since the perceptron rule stops as soon as the patterns are correctly classified.

10.5 Example on Linear Regression

ADALINE network is used to determine parameters of a quadratic function, a_0, a_1, and a_2, of the relation $y = a_2 x^2 + a_1 x + a_0$ when the data of *x* and *y* are as the following.

x	-3	-2	-1	0	1	2	3
y	6	3	2	3	6	11	18



By LMS algorithm,

$$F(x) = E[t^2] - 2x^T E[tz] + x^T E[zz^T]x = c - 2x^T h + x^T Rx \qquad (10.5\text{-}1)$$

$$c = \frac{1}{7}\left(6^2 + 3^2 + 2^2 + 3^2 + 6^2 + 11^2 + 18^2\right) = \frac{1}{7}539 = 77 \qquad (10.5\text{-}2)$$

$$h=\frac{1}{7}\left(6\begin{bmatrix}9\\-3\\1\end{bmatrix}+3\begin{bmatrix}4\\-2\\1\end{bmatrix}+2\begin{bmatrix}1\\-1\\1\end{bmatrix}+3\begin{bmatrix}0\\0\\1\end{bmatrix}+6\begin{bmatrix}1\\1\\1\end{bmatrix}+11\begin{bmatrix}4\\2\\1\end{bmatrix}+18\begin{bmatrix}9\\3\\1\end{bmatrix}\right)=\frac{1}{7}\begin{bmatrix}280\\56\\49\end{bmatrix}=\begin{bmatrix}40\\8\\7\end{bmatrix} \tag{10.5-3}$$

$$R=\frac{1}{7}\left(\begin{bmatrix}9\\-3\\1\end{bmatrix}\begin{bmatrix}9&-3&1\end{bmatrix}+\begin{bmatrix}4\\-2\\1\end{bmatrix}\begin{bmatrix}4&-2&1\end{bmatrix}+\begin{bmatrix}1\\-1\\1\end{bmatrix}\begin{bmatrix}1&-1&1\end{bmatrix}+\begin{bmatrix}0\\0\\1\end{bmatrix}\begin{bmatrix}0&0&1\end{bmatrix}+\begin{bmatrix}1\\1\\1\end{bmatrix}\begin{bmatrix}1&1&1\end{bmatrix}+\begin{bmatrix}4\\2\\1\end{bmatrix}\begin{bmatrix}4&2&1\end{bmatrix}+\begin{bmatrix}9\\3\\1\end{bmatrix}\begin{bmatrix}9&3&1\end{bmatrix}\right)=\frac{1}{7}\begin{bmatrix}196&0&28\\0&28&0\\28&0&7\end{bmatrix}=\frac{1}{7}\begin{bmatrix}28&0&4\\0&4&0\\4&0&1\end{bmatrix} \tag{10.5-4}$$

$$x=\begin{bmatrix}a_2\\a_1\\a_0\end{bmatrix}=R^{-1}h=\begin{bmatrix}28&0&4\\0&4&0\\4&0&1\end{bmatrix}\begin{bmatrix}40\\8\\7\end{bmatrix}=\begin{bmatrix}1\\2\\3\end{bmatrix} \tag{10.5-5}$$

By Widrow-Hoff learning rule with 0.02 learning rate,

$$W_{new}=W_{old}+2\alpha ep \tag{10.5-6}$$

Present (-3, 6),

$$y=0(-3)^2+0(-3)+0=0; e=6-0=6 \tag{10.5-7}$$

$$\begin{bmatrix}w_{11}\\w_{12}\\b\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}+2(0.02)(6)\begin{bmatrix}(-3)^2\\-3\\1\end{bmatrix}=\begin{bmatrix}2.16\\-0.72\\0.24\end{bmatrix} \tag{10.5-8}$$

Present (-2, 3),

$$y = 2.16(-2)^2 - 0.72(-2) + 0.24 = 10.32; e = 3 - 10.32 = -7.32 \quad (10.5\text{-}9)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 2.16 \\ -0.72 \\ 0.24 \end{bmatrix} + 2(0.02)(-7.32)\begin{bmatrix} (-2)^2 \\ -2 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.99 \\ -0.13 \\ -0.05 \end{bmatrix} \quad (10.5\text{-}10)$$

Present (-1, 2),

$$y = 0.99(-1)^2 - 0.13(-1) - 0.05 = 1.07; e = 2 - 1.07 = 0.93 \quad (10.5\text{-}11)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 0.99 \\ -0.13 \\ -0.05 \end{bmatrix} + 2(0.02)(0.93)\begin{bmatrix} (-1)^2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.03 \\ -0.17 \\ -0.01 \end{bmatrix} \quad (10.5\text{-}12)$$

Present (0, 3),

$$y = 1.03(0)^2 - 0.17(0) - 0.01 = -0.01; e = 3 - -0.01 = 3.01 \quad (10.5\text{-}13)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 1.03 \\ -0.17 \\ -0.01 \end{bmatrix} + 2(0.02)(3.01)\begin{bmatrix} (0)^2 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.03 \\ -0.17 \\ 0.11 \end{bmatrix} \quad (10.5\text{-}14)$$

Present (1, 6),

$$y = 1.03(1)^2 - 0.17(1) + 0.11 = 0.97; e = 6 - 0.97 = 5.03 \quad (10.5\text{-}15)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 1.03 \\ -0.17 \\ 0.11 \end{bmatrix} + 2(0.02)(5.03)\begin{bmatrix} (1)^2 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.23 \\ 0.03 \\ 0.31 \end{bmatrix} \quad (10.5\text{-}16)$$

Present (2, 11),

$$y = 1.23(2)^2 + 0.03(2) + 0.31 = 5.29; e = 11 - 5.29 = 5.71 \quad (10.5\text{-}17)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 1.23 \\ 0.03 \\ 0.31 \end{bmatrix} + 2(0.02)(5.71)\begin{bmatrix} (2)^2 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2.14 \\ 0.49 \\ 0.54 \end{bmatrix} \quad (10.5\text{-}18)$$

Present (3, 18),

$$y = 2.14(3)^2 + 0.49(3) + 0.54 = 21.27; e = 18 - 21.27 = -3.27 \quad (10.5\text{-}19)$$

$$\begin{bmatrix} w_{11} \\ w_{12} \\ b \end{bmatrix} = \begin{bmatrix} 1.23 \\ 0.03 \\ 0.31 \end{bmatrix} + 2(0.02)(-3.27)\begin{bmatrix} (3)^2 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.05 \\ -0.36 \\ 0.18 \end{bmatrix} \quad (10.5\text{-}20)$$