2

Mathematical Background

In this chapter, we briefly review some mathematical background needed in this book, including linear algebra, mathematical analysis, and optimization theory. Through this review, most notation to be used in subsequent chapters is introduced. We then present the well-known least-squares method as an application of linear algebra and optimization theory.

2.1 Linear Algebra

We start with a review of vectors, vector spaces and matrices, and then introduce two powerful tools for matrix decomposition, namely eigendecomposition and singular value decomposition. The usefulness of matrix decomposition will become evident in the remaining parts of this book.

2.1.1 Vectors and Vector Spaces

Vectors

In this book, vectors are denoted by bold lowercase letters. For example, we denote an $N \times 1$ vector

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = (x_1, x_2, ..., x_N)^T$$

where x_n is a real or complex scalar representing the nth entry (component) of $\mathbf{x}$ and the superscript 'T' represents vector transposition. The complex-conjugate transposition, or *Hermitian*, of $\mathbf{x}$ is given by

$$\mathbf{x}^H = (\mathbf{x}^T)^* = (x_1^*, x_2^*, ..., x_N^*)$$

where the superscript 'H' denotes the Hermitian operator and the superscript '$*$' complex conjugation.

Let $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ and $\mathbf{y} = (y_1, y_2, ..., y_N)^T$ be two $N \times 1$ vectors. The *inner product* of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{n=1}^{N} x_n y_n^* = \mathbf{y}^H \mathbf{x}, \tag{2.1}$$

which is also referred to as the *Euclidean inner product* of $\mathbf{x}$ and $\mathbf{y}$. The *length*, or *norm*, of the vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\| = \left(\sum_{n=1}^{N} |x_n|^2 \right)^{1/2} = \sqrt{\mathbf{x}^H \mathbf{x}}, \tag{2.2}$$

which is also referred to as the *Euclidean norm* of $\mathbf{x}$. Other types of norms will be defined later and, for convenience, we will always use the definition (2.2) for the norm of $\mathbf{x}$ unless specified otherwise. A vector whose norm equals unity is called a *unit vector*. Furthermore, the *geometrical relationship* between two vectors $\mathbf{x}$ and $\mathbf{y}$ is given as follows: [1, p. 15]

$$\cos\phi = \begin{cases} \dfrac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \dfrac{|\mathbf{y}^H \mathbf{x}|}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, & 0 \le \phi \le \pi/2 \text{ (complex)} \\ \dfrac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \dfrac{\mathbf{y}^T \mathbf{x}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, & 0 \le \phi \le \pi \text{ (real)} \end{cases} \tag{2.3}$$

where ϕ is the angle between $\mathbf{x}$ and $\mathbf{y}$. As depicted in Fig. 2.1, the relationship can be interpreted by viewing the inner product $\langle \mathbf{x}, \mathbf{y}/\|\mathbf{y}\| \rangle$ as the projection of $\mathbf{x}$ onto the unit vector $\mathbf{y}/\|\mathbf{y}\|$. With the geometrical interpretation, $\mathbf{x}$ and $\mathbf{y}$ are said to be *orthogonal* if $\mathbf{x}^H \mathbf{y} = \mathbf{y}^H \mathbf{x} = 0$. Furthermore, if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal and have the unit norm, then they are said to be *orthonormal*.

The geometrical relationship given by (2.3) is closely related to the following *Cauchy–Schwartz inequality* or *Schwartz inequality*.[1]

Theorem 2.1 (Cauchy–Schwartz Inequality). *Let* $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ *and* $\mathbf{y} = (y_1, y_2, ..., y_N)^T$ *be real or complex nonzero vectors. Then*

$$|\mathbf{y}^H \mathbf{x}| \le \|\mathbf{x}\| \cdot \|\mathbf{y}\| \tag{2.4}$$

and the equality holds if and only if $\mathbf{x} = \alpha \mathbf{y}$ *where* $\alpha \ne 0$ *is an arbitrary real or complex scalar.*

The Cauchy–Schwartz inequality further leads to the following inequality:

[1] The Russians also refer to the Cauchy–Schwartz inequality as the *Cauchy–Schwartz–Buniakowsky inequality* [2].

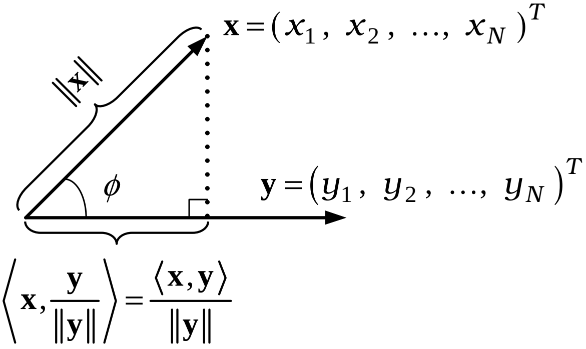
$\mathbf{x} = (x_1, x_2, \ldots, x_N)^T$
$\|\mathbf{x}\|$
ϕ
$\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$
$\left\langle \mathbf{x}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|}$

Fig. 2.1 The geometrical relationship between two vectors $\mathbf{x}$ and $\mathbf{y}$

Theorem 2.2 (Triangle Inequality). *Let* $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ *and* $\mathbf{y} = (y_1, y_2, ..., y_N)^T$ *be real or complex nonzero vectors. Then*

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \tag{2.5}$$

The proofs of the two theorems are left as exercises (Problems 2.1 and 2.2).

Vector Spaces

A *vector space* is a non-empty set of elements along with several rules for the operations of addition and scalar multiplication of elements. The elements can be vectors, sequences, functions, etc., and are also referred to as *vectors* without confusion. Let $\mathcal{V}$ denote a vector space and the vectors (elements) in $\mathcal{V}$ be also denoted by bold lowercase letters. Then for each pair of vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{V}$ there is a unique vector $\mathbf{x} + \mathbf{y}$ in $\mathcal{V}$ (the operation of addition) and for each scalar α there is a unique vector $\alpha\mathbf{x}$ in $\mathcal{V}$ (the operation of scalar multiplication). Furthermore, the operations of addition and scalar multiplication must satisfy the following axioms [3–5].

(VS1) For all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$, $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
(VS2) For all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{V}$, $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
(VS3) For all $\mathbf{x} \in \mathcal{V}$, there exists a zero vector $\mathbf{0} \in \mathcal{V}$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$.
(VS4) For each $\mathbf{x} \in \mathcal{V}$, there exists a vector $\mathbf{y} \in \mathcal{V}$ such that $\mathbf{x} + \mathbf{y} = \mathbf{0}$.
(VS5) For all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$ and for every scalar α, $\alpha(\mathbf{x} + \mathbf{y}) = \alpha\mathbf{x} + \alpha\mathbf{y}$.
(VS6) For all $\mathbf{x} \in \mathcal{V}$ and for all scalars α and β, $(\alpha + \beta)\mathbf{x} = \alpha\mathbf{x} + \beta\mathbf{x}$.
(VS7) For all $\mathbf{x} \in \mathcal{V}$ and for all scalars α and β, $(\alpha\beta)\mathbf{x} = \alpha(\beta\mathbf{x})$.
(VS8) For all $\mathbf{x} \in \mathcal{V}$, there exists a scalar 1 such that $1 \cdot \mathbf{x} = \mathbf{x}$.

A subset of a vector space $\mathcal{V}$, denoted by $\mathcal{W}$, is called a *subspace* of $\mathcal{V}$ if $\mathcal{W}$ itself is a vector space under the operations of addition and scalar multiplication defined on $\mathcal{V}$. An example is as follows.

Example 2.3
Under the operations of componentwise addition and scalar multiplication, the set of all real vectors $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ (whose entries are real) forms a real vector space, commonly denoted by $\mathcal{R}^N$. In addition, the set of all real $\mathbf{x}$ whose nth entry is zero (i.e. $x_n = 0$) is an example of a subspace of $\mathcal{R}^N$.

□

A vector space $\mathcal{V}$ is called an *inner product space* if it has a legitimate inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ defined for all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$. Note that an inner product is said to be *legitimate* if it satisfies the following axioms [3, 5].

(IPS1) For all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$ and for every scalar α, $\langle \alpha\mathbf{x}, \mathbf{y} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle$.
(IPS2) For all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z} \in \mathcal{V}$, $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$.
(IPS3) For all $\mathbf{x} \in \mathcal{V}$, $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
(IPS4) For all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$, $\langle \mathbf{x}, \mathbf{y} \rangle = (\langle \mathbf{y}, \mathbf{x} \rangle)^*$.

Similarly, a vector space $\mathcal{V}$ is called a *normed vector space* if it has a legitimate norm $\|\mathbf{x}\|$ defined for all $\mathbf{x} \in \mathcal{V}$. A norm is said to be *legitimate* if it satisfies the following axioms [3, 5].

(NVS1) For all $\mathbf{x} \in \mathcal{V}$ and for every scalar α, $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$.
(NVS2) For all $\mathbf{x}$, $\mathbf{y} \in \mathcal{V}$, $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.
(NVS3) For all $\mathbf{x} \in \mathcal{V}$, $\|\mathbf{x}\| \geq 0$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

It is important to note [5, pp. 14–15] that a legitimate inner product for a vector space $\mathcal{V}$ always induces a legitimate norm for $\mathcal{V}$ via the relation

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \quad \text{for all } \mathbf{x} \in \mathcal{V}.$$

Such a norm is referred to as an *induced norm.* An example is as follows.

Example 2.4 (Euclidean Space)
It can be easily shown that for the real vector space $\mathcal{R}^N$ (see Example 2.3), the Euclidean inner product defined as (2.1) is legitimate and induces the Euclidean norm defined as (2.2). Accordingly, $\mathcal{R}^N$ along with the Euclidean inner product is an inner product space, while $\mathcal{R}^N$ along with the Euclidean norm is a normed vector space. The former is known as the *Euclidean space* [4].

□

Let $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_N$ be the vectors in a vector space $\mathcal{V}$. Then they are said to *span* the subspace $\mathcal{W}$ if $\mathcal{W}$ consists of all linear combinations of $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_N$. Specifically, every vector $\mathbf{w}$ in $\mathcal{W}$ can be expressed as

$$\mathbf{w} = \alpha_1\mathbf{q}_1 + \alpha_2\mathbf{q}_2 + \cdots + \alpha_N\mathbf{q}_N$$

where α_k are scalars. For vectors $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_N$ in $\mathcal{V}$, one can determine their linear interdependence via the following equation:

$$c_1\mathbf{q}_1 + c_2\mathbf{q}_2 + \cdots + c_N\mathbf{q}_N = \mathbf{0}$$

where c_k are scalars and $\mathbf{0}$ is a zero vector defined by (VS3). If this equation holds true only when $c_1 = c_2 = \cdots = c_N = 0$, then $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_N$ are said to be *linearly independent*; otherwise, they are *linearly dependent*. If $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_N$ are linearly independent and span the vector space $\mathcal{V}$, they are called a *basis* for $\mathcal{V}$. A vector space $\mathcal{V}$ is said to be *finite-dimensional* if the number of linearly independent vectors in its basis is finite; otherwise, it is said to be *infinite-dimensional*.

A set S in an inner product space $\mathcal{V}$ is called an *orthogonal set* if every pair of vectors $\mathbf{q}_k, \mathbf{q}_m \in S$ is orthogonal, i.e. $\langle \mathbf{q}_k, \mathbf{q}_m \rangle = 0$ for $k \neq m$. Furthermore, if every vector $\mathbf{q}_k \in S$ has the unit norm, i.e. $\|\mathbf{q}_k\| = 1$, then the orthogonal set S is said to be *orthonormal*. In other words, an orthonormal set does not contain the zero vector $\mathbf{0}$. A basis for an inner product space $\mathcal{V}$ is said to be an *orthonormal basis* if it is an orthonormal set. For example, the set

$$\{\boldsymbol{\eta}_1 = (1,0,0,...,0)^T,\ \boldsymbol{\eta}_2 = (0,1,0,...,0)^T,\ ...,\ \boldsymbol{\eta}_N = (0,0,...,0,1)^T\} \tag{2.6}$$

is an orthonormal basis, referred to as the *standard basis*, for the Euclidean space $\mathcal{R}^N$ (see Example 2.4) where $\boldsymbol{\eta}_k$ denotes a unit vector whose kth entry equals unity and the remaining entries equal zero. Note that any basis can be transformed into an orthonormal basis via the process of *Gram–Schmidt orthogonalization* [1–3, 5].

2.1.2 Matrices

In this book, matrices are denoted by bold uppercase letters. For example,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MK} \end{pmatrix} \tag{2.7}$$

denotes an $M \times K$ matrix whose (m,k)th entry (component) is a_{mk}, a real or complex scalar. We also use the shorthand representation

$$[\mathbf{A}]_{m,k} = a_{mk}$$

to specify the matrix $\mathbf{A}$. The *transposition* of $\mathbf{A}$ is

$$[\mathbf{A}^T]_{m,k} = [\mathbf{A}]_{k,m} = a_{km} \tag{2.8}$$

and $(\mathbf{A}^T)^T = \mathbf{A}$ where the superscript 'T' stands for matrix transposition. The complex-conjugate transposition, or *Hermitian*, of $\mathbf{A}$ is

$$[\mathbf{A}^H]_{m,k} = [\mathbf{A}^*]_{k,m} = a_{km}^* \tag{2.9}$$

and $(\mathbf{A}^H)^H = \mathbf{A}$ where the superscript '$_H$' stands for the Hermitian operation. The matrix $\mathbf{A}$ is said to be *square* if $M = K$. It is further said to be *symmetric* if $\mathbf{A}^T = \mathbf{A}$ for $\mathbf{A}$ real, and *Hermitian* if $\mathbf{A}^H = \mathbf{A}$ for $\mathbf{A}$ complex. Note that $\mathbf{A}^H = \mathbf{A}^T$ as $\mathbf{A}$ is real. For matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$, $(\mathbf{AB})^H = \mathbf{B}^H\mathbf{A}^H$, $(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$, and $(\mathbf{A}+\mathbf{B})^H = \mathbf{A}^H + \mathbf{B}^H$.

Let us further represent the $M \times K$ matrix $\mathbf{A}$ given in (2.7) by

$$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K) = \begin{pmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \vdots \\ \mathbf{b}_M^T \end{pmatrix}$$

where $\mathbf{a}_k = (a_{1k}, a_{2k}, ..., a_{Mk})^T$, $k = 1, 2, ..., K$, are called the *column vectors* of $\mathbf{A}$ and $\mathbf{b}_m^T = (a_{m1}, a_{m2}, ..., a_{mK})$, $m = 1, 2, ..., M$, the *row vectors* of $\mathbf{A}$. The subspace spanned by the column vectors is called the *column space* of $\mathbf{A}$, while the subspace spanned by the row vectors is called the *row space* of $\mathbf{A}$. The number of linearly independent column vectors of $\mathbf{A}$ is equal to the number of linearly independent row vectors of $\mathbf{A}$, that is defined as the *rank* of $\mathbf{A}$, denoted by $\text{rank}\{\mathbf{A}\}$. Note that $\text{rank}\{\mathbf{A}\} = \text{rank}\{\mathbf{A}^H\} \le \min\{M, K\}$ and $\text{rank}\{\mathbf{A}^H\mathbf{A}\} = \text{rank}\{\mathbf{A}\mathbf{A}^H\} = \text{rank}\{\mathbf{A}\}$. When $\text{rank}\{\mathbf{A}\} = \min\{M, K\}$, the matrix $\mathbf{A}$ is said to be of *full rank*; otherwise, it is *rank deficient.*

The *inverse* of an $M \times M$ square matrix $\mathbf{A}$ is also an $M \times M$ square matrix, denoted by $\mathbf{A}^{-1}$, which satisfies

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{2.10}$$

where

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{2.11}$$

is the $M \times M$ *identity matrix.* If $\mathbf{A}$ is of full rank, then $\mathbf{A}^{-1}$ exists and $\mathbf{A}$ is said to be *invertible* or *nonsingular.* On the other hand, if $\mathbf{A}$ is rank deficient, then it does not have an inverse and is accordingly said to be *noninvertible* or *singular.* For nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$, $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$, $(\mathbf{A}^H)^{-1} = (\mathbf{A}^{-1})^H$, and $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Consider an $M \times M$ square matrix $\mathbf{A}$ with $[\mathbf{A}]_{m,k} = a_{mk}$. The *determinant* of $\mathbf{A}$ is commonly denoted by $\det\{\mathbf{A}\}$ or $|\mathbf{A}|$. For $M = 1$, the matrix $\mathbf{A}$ reduces to a scalar a_{11} and its determinant is defined as $\det\{a_{11}\} = a_{11}$. For $M \ge 2$, the determinant $\det\{\mathbf{A}\}$ can be defined in terms of the determinants of the associated $(M-1) \times (M-1)$ matrices as follows:

$$\begin{aligned}\det\{\mathbf{A}\} &= \sum_{m=1}^{M}(-1)^{m+k}\cdot a_{mk}\cdot\det\{\mathbf{A}_{mk}\} \quad \text{for any } k\in\{1,2..,M\}\\ &= \sum_{k=1}^{M}(-1)^{m+k}\cdot a_{mk}\cdot\det\{\mathbf{A}_{mk}\} \quad \text{for any } m\in\{1,2..,M\} \qquad (2.12)\end{aligned}$$

where $\mathbf{A}_{mk}$ is an $(M-1)\times(M-1)$ matrix obtained by deleting the mth row and kth column of $\mathbf{A}$. For example, if $M=2$, $\det\{\mathbf{A}\}$ is given by

$$\begin{aligned}\det\left\{\begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix}\right\} &= (-1)^{1+1}\cdot a_{11}\cdot\det\{a_{22}\}+(-1)^{2+1}\cdot a_{21}\cdot\det\{a_{12}\}\\ &= a_{11}a_{22}-a_{21}a_{12}.\end{aligned}$$

Note that $\det\{\mathbf{A}^T\}=\det\{\mathbf{A}\}$, $\det\{\mathbf{A}^H\}=[\det\{\mathbf{A}\}]^*$, and $\det\{\alpha\mathbf{A}\}=\alpha^M\cdot\det\{\mathbf{A}\}$ for a scalar α. For square matrices $\mathbf{A}$ and $\mathbf{B}$, $\det\{\mathbf{AB}\}=\det\{\mathbf{A}\}\det\{\mathbf{B}\}$. If $\mathbf{A}$ is nonsingular, then $\det\{\mathbf{A}\}\neq 0$ and $\det\{\mathbf{A}^{-1}\}=1/\det\{\mathbf{A}\}$. On the other hand, the *trace* of $\mathbf{A}$, denoted by $\mathrm{tr}\{\mathbf{A}\}$, is defined as

$$\mathrm{tr}\{\mathbf{A}\}=\sum_{m=1}^{M}a_{mm}, \qquad (2.13)$$

i.e. the sum of the diagonal elements of $\mathbf{A}$. As $M=1$, the matrix $\mathbf{A}$ reduces to a scalar a_{11} and its trace $\mathrm{tr}\{a_{11}\}=a_{11}$. If $\mathbf{A}$ is an $M\times K$ matrix and $\mathbf{B}$ is a $K\times M$ matrix, then $\mathrm{tr}\{\mathbf{AB}\}=\mathrm{tr}\{\mathbf{BA}\}$. As a special case, for column vectors $\mathbf{x}$ and $\mathbf{y}$, the trace $\mathrm{tr}\{\mathbf{x}\mathbf{y}^H\}=\mathrm{tr}\{\mathbf{y}^H\mathbf{x}\}=\mathbf{y}^H\mathbf{x}$.

Let $\mathbf{A}$ be an $M\times M$ Hermitian matrix and $\mathbf{x}$ be an $M\times 1$ vector, then the quadratic function

$$Q(\mathbf{x})\triangleq\mathbf{x}^H\mathbf{A}\mathbf{x} \qquad (2.14)$$

is called the *Hermitian form* of $\mathbf{A}$. The Hermitian matrix $\mathbf{A}$ is said to be *positive semidefinite* or *nonnegative definite* if $Q(\mathbf{x})\geq 0$ for all $\mathbf{x}\neq\mathbf{0}$, and is said to be *positive definite* if $Q(\mathbf{x})>0$ for all $\mathbf{x}\neq\mathbf{0}$. In the same fashion, $\mathbf{A}$ is *negative semidefinite* or *nonpositive definite* if $Q(\mathbf{x})\leq 0$ for all $\mathbf{x}\neq\mathbf{0}$, and *negative definite* if $Q(\mathbf{x})<0$ for all $\mathbf{x}\neq\mathbf{0}$.

An *eigenvector* of an $M\times M$ square matrix $\mathbf{A}$ is an $M\times 1$ nonzero vector, denoted by $\mathbf{q}$, which satisfies

$$\mathbf{A}\mathbf{q}=\lambda\mathbf{q} \qquad (2.15)$$

where λ is a scalar.[2] The scalar λ is an *eigenvalue* of $\mathbf{A}$ corresponding to the eigenvector $\mathbf{q}$. One can see from (2.15) that for any nonzero constant α,

[2] More precisely, the vector $\mathbf{q}$ is a *right eigenvector* of $\mathbf{A}$ if $\mathbf{A}\mathbf{q}=\lambda\mathbf{q}$, and a *left eigenvector* of $\mathbf{A}$ if $\mathbf{q}^H\mathbf{A}=\lambda\mathbf{q}^H$. In this book, “eigenvector” implies “right eigenvector” [6].

$\mathbf{A}(\alpha\mathbf{q}) = \lambda(\alpha\mathbf{q})$. This implies that any scaled version of $\mathbf{q}$ is also an eigenvector of $\mathbf{A}$ corresponding to the same eigenvalue λ. Eigenvectors which are orthogonal (i.e. $\mathbf{q}_m^H\mathbf{q}_n = 0$ for eigenvectors $\mathbf{q}_m$ and $\mathbf{q}_n$) and have the unit norm are referred to as *orthonormal eigenvectors*.

Special Forms of Matrices

A complex square matrix $\mathbf{U}$ is called a *unitary matrix* if it satisfies

$$\mathbf{U}\mathbf{U}^H = \mathbf{U}^H\mathbf{U} = \mathbf{I}, \tag{2.16}$$

i.e. $\mathbf{U}^H = \mathbf{U}^{-1}$ and $|\det\{\mathbf{U}\}| = 1$. Similarly, a real square matrix $\mathbf{V}$ is called an *orthogonal matrix* if it satisfies

$$\mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}, \tag{2.17}$$

i.e. $\mathbf{V}^T = \mathbf{V}^{-1}$ and $\det\{\mathbf{V}\} = 1$. Obviously, the identity matrix $\mathbf{I}$ is an orthogonal matrix.

A *diagonal matrix* is an $M \times M$ square matrix defined as

$$\mathbf{D} = \mathrm{diag}\{d_1, d_2, ..., d_M\} = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_M \end{pmatrix}. \tag{2.18}$$

If the diagonal matrix $\mathbf{D}$ is nonsingular, i.e. $\det\{\mathbf{D}\} = d_1 d_2 \cdots d_M \neq 0$, then its inverse

$$\mathbf{D}^{-1} = \mathrm{diag}\left\{\frac{1}{d_1}, \frac{1}{d_2}, ..., \frac{1}{d_M}\right\}. \tag{2.19}$$

An *upper triangular matrix* is an $M \times M$ square matrix defined as

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1M} \\ 0 & u_{22} & \cdots & u_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{MM} \end{pmatrix}, \tag{2.20}$$

and a *lower triangular matrix* defined as

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{M1} & l_{M2} & \cdots & l_{MM} \end{pmatrix}. \tag{2.21}$$

From (2.12), it follows that $\det\{\mathbf{U}\} = u_{11}u_{22}\cdots u_{MM}$ and $\det\{\mathbf{L}\} = l_{11}l_{22}\cdots l_{MM}$.

A *Toeplitz matrix* is an $M \times M$ square matrix defined as

$$\mathbf{R} = \begin{pmatrix} r_0 & r_1 & \cdots & r_{M-2} & r_{M-1} \\ r_{-1} & r_0 & \ddots & & r_{M-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r_{-M+2} & & \ddots & r_0 & r_1 \\ r_{-M+1} & r_{-M+2} & \cdots & r_{-1} & r_0 \end{pmatrix}, \tag{2.22}$$

i.e. the entries on each of the diagonals are equal. Note that a Toeplitz matrix can be completely specified by its first column and first row.

A matrix $\mathbf{A}$ is called a 2×2 *partitioned matrix* if it can be expressed as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \tag{2.23}$$

where $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are the submatrices of $\mathbf{A}$. Manipulations of the submatrices for partitioned matrices are similar to those of the entries for general matrices. In particular, the Hermitian of $\mathbf{A}$ can be written as

$$\mathbf{A}^H = \begin{pmatrix} \mathbf{A}_{11}^H & \mathbf{A}_{21}^H \\ \mathbf{A}_{12}^H & \mathbf{A}_{22}^H \end{pmatrix}. \tag{2.24}$$

Furthermore, if $\mathbf{B}$ is also a 2×2 partitioned matrix given by

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix},$$

then

$$\mathbf{AB} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{pmatrix} \tag{2.25}$$

where $\mathbf{B}_{11}$, $\mathbf{B}_{12}$, $\mathbf{B}_{21}$, and $\mathbf{B}_{22}$ are the submatrices with suitable sizes for the submatrix multiplications in $\mathbf{AB}$.

Matrix Formulas and Properties

The following theorem provides a useful formula for the derivation of matrix inverse [7,8].

Theorem 2.5 (Matrix Inversion Lemma). *Let* $\mathbf{R}$ *be a nonsingular* $M \times M$ *matrix given by*

$$\mathbf{R} = \mathbf{A} + \mathbf{BCD} \tag{2.26}$$

where $\mathbf{A}$ *is a nonsingular* $M \times M$ *matrix,* $\mathbf{B}$ *is an* $M \times K$ *matrix,* $\mathbf{C}$ *is a nonsingular* $K \times K$ *matrix, and* $\mathbf{D}$ *is a* $K \times M$ *matrix. Then the inverse of* $\mathbf{R}$ *can be expressed as*

$$\mathbf{R}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1}. \tag{2.27}$$

A special case of the matrix inversion lemma is given as follows [7].[3]

Corollary 2.6 (Woodbury's Identity). *Let* $\mathbf{R}$ *be a nonsingular* $M \times M$ *matrix given by*

$$\mathbf{R} = \mathbf{A} + \alpha\mathbf{u}\mathbf{u}^H \tag{2.28}$$

where $\mathbf{A}$ *is a nonsingular* $M \times M$ *matrix,* $\mathbf{u}$ *is an* $M \times 1$ *vector, and* α *is a scalar. Then the inverse of* $\mathbf{R}$ *can be expressed as*

$$\mathbf{R}^{-1} = \mathbf{A}^{-1} - \frac{\alpha\mathbf{A}^{-1}\mathbf{u}\mathbf{u}^H\mathbf{A}^{-1}}{1 + \alpha\mathbf{u}^H\mathbf{A}^{-1}\mathbf{u}}. \tag{2.29}$$

The proof of Theorem 2.5 is left as an exercise (Problem 2.3), while Corollary 2.6 can be proved simply by substituting $\mathbf{B} = \mathbf{u}$, $\mathbf{C} = \alpha$ and $\mathbf{D} = \mathbf{u}^H$ into (2.27).

Moreover, two theorems regarding partitioned matrices are stated as follows [7, p. 572], [9, pp. 166–168], and the proofs are left as exercises (Problems 2.4 and 2.5).

Theorem 2.7. *Let* $\mathbf{A}$ *be a square matrix given as the partitioned form of (2.23). Then the determinant of* $\mathbf{A}$ *can be expressed as*

$$\det\{\mathbf{A}\} = \det\{\mathbf{A}_{11}\} \cdot \det\{\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\} \tag{2.30}$$

provided that $\mathbf{A}_{11}$ *is a nonsingular square matrix, or equivalently*

$$\det\{\mathbf{A}\} = \det\{\mathbf{A}_{22}\} \cdot \det\{\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\} \tag{2.31}$$

provided that $\mathbf{A}_{22}$ *is a nonsingular square matrix.*

Theorem 2.8. *Let* $\mathbf{A}$ *be a nonsingular square matrix given as the partitioned form of (2.23) where* $\mathbf{A}_{11}$ *and* $\mathbf{A}_{22}$ *are also nonsingular square matrices. Then the inverse of* $\mathbf{A}$ *can be expressed as*

[3] For ease of later use, we give a slightly generalized statement of Woodbury's identity by including a scalar α. As $\alpha = 1$, it reduces to the normal statement of Woodbury's identity.

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \tag{2.32}$$

where

$$\begin{aligned}
\mathbf{B}_{11} &= (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} \\
\mathbf{B}_{12} &= -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\
\mathbf{B}_{21} &= -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} \\
\mathbf{B}_{22} &= (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}.
\end{aligned}$$

In the following, we summarize several matrix properties and leave the proofs as exercises (Problems 2.6, 2.7 and 2.8).

Property 2.9. *A positive definite matrix is nonsingular.*

Property 2.10. *The eigenvalues of a Hermitian matrix are all real.*

Property 2.11. *The eigenvalues of a positive definite (positive semidefinite) matrix are all real positive (nonnegative).*

Property 2.12. *The inverse of a positive definite matrix is also positive definite.*

Property 2.13. *For any matrix $\mathbf{A}$, both $\mathbf{A}^H\mathbf{A}$ and $\mathbf{A}\mathbf{A}^H$ are positive semidefinite.*

Property 2.14. *The eigenvectors of a Hermitian matrix corresponding to distinct eigenvalues are orthogonal.*

Although Property 2.14 is for the case of distinct eigenvalues, one can always find a complete set of orthogonal eigenvectors, or equivalently, orthonormal eigenvectors for any Hermitian matrix, no matter whether its eigenvalues are distinct or not [2, p. 297].

As a consequence, if $\mathbf{A}$ is a positive definite matrix, then its inverse $\mathbf{A}^{-1}$ exists (by Property 2.9) and is also positive definite (by Property 2.12). Furthermore, the eigenvalues of both matrices $\mathbf{A}$ and $\mathbf{A}^{-1}$ are all real positive (by Property 2.11).

2.1.3 Matrix Decomposition

Among the available tools of matrix decomposition, two representatives, *eigendecomposition* and *singular value decomposition (SVD)*, to be presented are of importance in the area of statistical signal processing. In particular, the eigendecomposition is useful in developing subspace based algorithms, while the SVD is powerful in solving least-squares problems as well as in determining the numerical rank of a real or complex matrix in the presence of roundoff errors (due to finite precision of computing machines).

Eigendecomposition

According to the foregoing discussion (the paragraph following Property 2.14), we can always find a complete set of M orthonormal eigenvectors for an $M \times M$ Hermitian matrix $\mathbf{A}$. As such, let $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ be the M orthonormal eigenvectors of $\mathbf{A}$ corresponding to the eigenvalues λ_1, λ_2, ..., λ_M. Then, by definition,

$$\mathbf{A}(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M) = (\lambda_1 \mathbf{u}_1, \lambda_2 \mathbf{u}_2, ..., \lambda_M \mathbf{u}_M)$$

or

$$\mathbf{AU} = \mathbf{U\Lambda} \tag{2.33}$$

where $\mathbf{\Lambda} = \text{diag}\{\lambda_1,\ \lambda_2,\ ...,\ \lambda_M\}$ is an $M \times M$ diagonal matrix and $\mathbf{U} = (\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M)$ is an $M \times M$ unitary matrix since $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ are orthonormal. From (2.33), it follows that

$$\mathbf{A} = \mathbf{U\Lambda U}^H = \sum_{m=1}^{M} \lambda_m \mathbf{u}_m \mathbf{u}_m^H. \tag{2.34}$$

Equation (2.34) is called the *eigendecomposition* or the *spectral decomposition* of $\mathbf{A}$. Moreover, when $\mathbf{A}$ is nonsingular, (2.34) leads to

$$\mathbf{A}^{-1} = \mathbf{U}^{-H}\mathbf{\Lambda}^{-1}\mathbf{U}^{-1} = \mathbf{U\Lambda}^{-1}\mathbf{U}^H = \sum_{m=1}^{M} \frac{1}{\lambda_m} \mathbf{u}_m \mathbf{u}_m^H. \tag{2.35}$$

Singular Value Decomposition

The SVD is stated in the following theorem and, for clarity, is illustrated in Fig. 2.2. The theorem is called the *SVD theorem*, or the *Autonne–Eckart–Young theorem* in recognition of the originators [10].[4]

Theorem 2.15 (SVD Theorem). *Let* $\mathbf{A}$ *be an* $M \times K$ *real or complex matrix with* $\text{rank}\{\mathbf{A}\} = r$. *Then there exist an* $M \times M$ *unitary matrix*

$$\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M) \tag{2.36}$$

and a $K \times K$ *unitary matrix*

$$\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K) \tag{2.37}$$

such that the matrix $\mathbf{A}$ *can be decomposed as*

[4] The SVD was established for real square matrices by Beltrami and Jordan in the 1870s, for complex square matrices by Autonne in 1902, and for general rectangular matrices by Eckart and Young in 1939 [10].

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H = \sum_{m=1}^{r} \lambda_m \mathbf{u}_m \mathbf{v}_m^H \tag{2.38}$$

where $\mathbf{u}_i$ *are* $M \times 1$ *vectors,* $\mathbf{v}_i$ *are* $K \times 1$ *vectors, and*

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{2.39}$$

is an $M \times K$ *matrix. The matrix* $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, ..., \lambda_r\}$ *is an* $r \times r$ *diagonal matrix where* λ_i *are real and* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$.

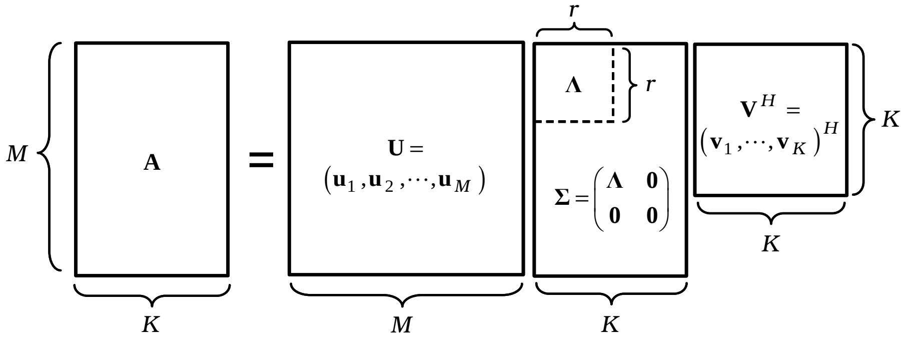

Fig. 2.2 Illustration of the SVD for an $M \times K$ matrix $\mathbf{A}$ with $M > K > r = \mathrm{rank}\{\mathbf{A}\}$ where $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, ..., \lambda_r\}$

As shown in Appendix 2A, the SVD theorem can be proved by either of the two approaches, *Approach I* and *Approach II*, where Approach I starts from the matrix $\mathbf{A}^H\mathbf{A}$ and Approach II from the matrix $\mathbf{A}\mathbf{A}^H$. Some important results regarding both approaches are summarized as follows.

- Results from Approach I. The nonnegative real numbers λ_1, λ_2, ..., λ_K are identical to the positive square roots of the eigenvalues of the $K \times K$ matrix $\mathbf{A}^H\mathbf{A}$ and the column vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_K$ of $\mathbf{V}$ are the corresponding orthonormal eigenvectors. The positive real numbers λ_1, λ_2, ..., λ_r, together with $\lambda_{r+1} = \cdots = \lambda_K = 0$ (since $\mathrm{rank}\{\mathbf{A}\} = r$), are called the *singular values* of $\mathbf{A}$, while the vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_K$ are called the *right singular vectors* of $\mathbf{A}$. With λ_m and $\mathbf{v}_m$ computed from $\mathbf{A}^H\mathbf{A}$, the column vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_r$ in $\mathbf{U}$ are accordingly determined via (see (2.204))

$$(\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r) = \left(\frac{\mathbf{A}\mathbf{v}_1}{\lambda_1}, \frac{\mathbf{A}\mathbf{v}_2}{\lambda_2}, ..., \frac{\mathbf{A}\mathbf{v}_r}{\lambda_r} \right), \tag{2.40}$$

  while the remaining column vectors $\mathbf{u}_{r+1}$, $\mathbf{u}_{r+2}$, ..., $\mathbf{u}_M$ (allowing some choices) are chosen such that $\mathbf{U}$ is unitary. The vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ are called the *left singular vectors* of $\mathbf{A}$.

- Results from Approach II. The singular values λ_1, λ_2, ..., λ_M of $\mathbf{A}$ are identical to the positive square roots of the eigenvalues of the $M \times M$ matrix $\mathbf{A}\mathbf{A}^H$ and the left singular vectors $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ are the corresponding orthonormal eigenvectors. With λ_m and $\mathbf{u}_m$ computed from $\mathbf{A}\mathbf{A}^H$, the right singular vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_r$ are accordingly determined via (see (2.213))

$$(\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r) = \left(\frac{\mathbf{A}^H \mathbf{u}_1}{\lambda_1}, \frac{\mathbf{A}^H \mathbf{u}_2}{\lambda_2}, ..., \frac{\mathbf{A}^H \mathbf{u}_r}{\lambda_r} \right), \tag{2.41}$$

  while the remaining right singular vectors $\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_K$ are chosen such that $\mathbf{V}$ is unitary.

As a result, a matrix may have numerous forms of SVD [11, p. 309]. Moreover, following the above-mentioned results, one can compute (by hand) the SVD of an $M \times K$ matrix $\mathbf{A}$ through the eigenvalues and orthonormal eigenvectors of $\mathbf{A}^H\mathbf{A}$ or $\mathbf{A}\mathbf{A}^H$, although it is generally not suggested for finite-precision computation [10]. It is also important to note that the number of nonzero singular values determines the rank of $\mathbf{A}$, revealing that the SVD provides a basis for practically determining the numerical rank of a matrix.

A special case of the SVD theorem is as follows.

Corollary 2.16 (Special Case of the SVD Theorem). *Let* $\mathbf{A}$ *be an* $M \times M$ *Hermitian matrix with* $\mathrm{rank}\{\mathbf{A}\} = r$ *and* $\mathbf{A}$ *is nonnegative definite. Then the matrix* $\mathbf{A}$ *can be decomposed as*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H = \sum_{m=1}^{r} \lambda_m \mathbf{u}_m \mathbf{u}_m^H \tag{2.42}$$

where $\mathbf{\Sigma} = \mathrm{diag}\{\lambda_1, ..., \lambda_r, \lambda_{r+1}, ..., \lambda_M\}$ *is an* $M \times M$ *diagonal matrix and* $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M)$ *is an* $M \times M$ *unitary matrix. The singular values* $\lambda_1 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_M = 0$ *are the eigenvalues of* $\mathbf{A}$ *and the singular vectors* $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ *are the corresponding orthonormal eigenvectors.*

The proof is left as an exercise (Problem 2.10). Comparing (2.42) with (2.34) reveals that for a Hermitian matrix $\mathbf{A}$, the SVD of $\mathbf{A}$ is equivalent to the eigendecomposition of $\mathbf{A}$.

2.2 Mathematical Analysis

This section briefly reviews the fundamentals of mathematical analysis, including sequences, series, Hilbert spaces, vector spaces of sequences and functions, and pays attention to the topic of Fourier series. Some of these topics need the background of functions, provided in Appendix 2B.

2.2.1 Sequences

A *sequence* is regarded as a list of real or complex numbers in a definite order:

$$a_m, a_{m+1}, ..., a_{n-1}, a_n$$

where a_k, $k = m, m+1, ..., n$, are called the *terms* of the sequence. The sequence is denoted by $\{a_k\}_{k=m}^{n}$ or, briefly, $\{a_k\}$. One should not confuse a sequence $\{a_k\}_{k=m}^{n}$ with a set $\{a_k, k = m, m+1, ..., n\}$; the order of a_k is meaningless for the latter. Moreover, a sequence $\{a_k\}$ is said to be an *infinite sequence* if it has infinitely many terms. A natural concern about a one-sided infinite sequence, $\{a_k\}_{k=1}^{\infty}$, is whether it converges or not, that is the topic to be dealt with next.

Sequences of Numbers

A real or complex sequence $\{a_k\}_{k=1}^{\infty}$ is said to *converge* to a real or complex number a if

$$\lim_{k\to\infty} a_k = a, \tag{2.43}$$

i.e. for every real number $\varepsilon > 0$ there exists an integer N such that

$$|a_k - a| < \varepsilon \quad \text{for all } k \geq N \tag{2.44}$$

where N is, in general, dependent on ε. If $\{a_k\}$ does not converge, it is called a *divergent sequence* [12]. A sequence $\{a_k\}$ is said to be *bounded* if $|a_k| \leq A$ for all k where A is a positive constant. A real sequence $\{a_k\}_{k=1}^{\infty}$ is said to be *increasing* (*decreasing*) or, briefly, *monotonic* if $a_k \leq a_{k+1}$ ($a_k \geq a_{k+1}$) for all k, and is said to be *strictly increasing* (*strictly decreasing*) if $a_k < a_{k+1}$ ($a_k > a_{k+1}$) for all k. A theorem regarding monotonic sequences is as follows [13, p. 61].

Theorem 2.17. *If $\{a_k\}_{k=1}^{\infty}$ is a monotonic and bounded real sequence, then $\{a_k\}_{k=1}^{\infty}$ converges.*

The proof is left as an exercise (Problem 2.11).

From a sequence $\{a_k\}_{k=1}^{\infty}$, one can obtain another sequence, denoted by $\{\sigma_n\}_{n=1}^{\infty}$, composed of the arithmetic mean

$$\sigma_n = \frac{a_1 + a_2 + \cdots + a_n}{n}. \tag{2.45}$$

The arithmetic mean σ_n is also referred to as the *nth Cesàro mean* of the sequence $\{a_n\}$ [14]. A related theorem is stated as follows [15, p. 138].

Theorem 2.18. *If a real or complex sequence $\{a_k\}_{k=1}^{\infty}$ is bounded and converges to a real or complex number a, then the sequence of arithmetic mean $\{\sigma_n\}_{n=1}^{\infty}$ also converges to the number a where σ_n is defined as (2.45).*

The proof, again, is left as an exercise (Problem 2.12). When the sequence of arithmetic means $\{\sigma_n\}$ converges to a, we say that the original sequence $\{a_k\}$ is *Cesàro summable* to a. Since the average operation in (2.45) may smooth out occasional fluctuations in $\{a_k\}$, it is expected that $\{\sigma_n\}$, in general, tends to converge even if $\{a_k\}$ is divergent. An example is given as follows.

Example 2.19
Consider that $a_k = (-1)^k$. The sequence $\{a_k\}_{k=1}^{\infty}$ is bounded by 1, but it diverges since $a_k = 1$ for k even and $a_k = -1$ for k odd. On the other hand, the arithmetic mean $\sigma_n = 0$ for n even and $\sigma_n = -1/n$ for n odd. This indicates that $\lim_{n\to\infty} \sigma_n = 0$, namely, $\{a_k\}$ is Cesàro summable to zero. □

Sequences of Functions

Now consider a sequence of real or complex functions, $\{a_k(x)\}_{k=1}^{\infty}$. Since $a_k(x)$ is a function of x, the convergence of $\{a_k(x)\}_{k=1}^{\infty}$ may further depend on the value of x.

The sequence $\{a_k(x)\}_{k=1}^{\infty}$ is said to *converge pointwise* to a real or complex function $a(x)$ on an interval $[x_\mathrm{L}, x_\mathrm{U}]$ if

$$\lim_{k\to\infty} a_k(x) = a(x) \quad \text{for every point } x \in [x_\mathrm{L}, x_\mathrm{U}], \tag{2.46}$$

i.e. for every real number $\varepsilon > 0$ and every point $x \in [x_\mathrm{L}, x_\mathrm{U}]$ there exists an integer N such that

$$|a_k(x) - a(x)| < \varepsilon \quad \text{for all } k \geq N \tag{2.47}$$

where N may depend on ε and x. When the integer N is independent of x, $\{a_k(x)\}_{k=1}^{\infty}$ is said to *converge uniformly* to $a(x)$ on the interval $[x_\mathrm{L}, x_\mathrm{U}]$. In other words, a uniformly convergent sequence $\{a_k(x)\}$ exhibits similar local behaviors of convergence for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$, as illustrated in Fig. 2.3.

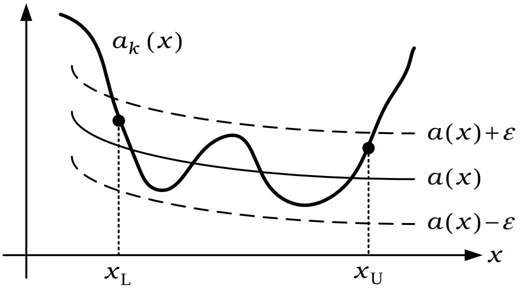

Fig. 2.3 A uniformly convergent sequence $\{a_k(x)\}_{k=1}^{\infty}$ on an interval $[x_\mathrm{L}, x_\mathrm{U}]$

It is important to emphasize that even if every $a_k(x)$ is a continuous function, a pointwise convergent sequence $\{a_k(x)\}_{k=1}^{\infty}$ may still converge to a discontinuous function $a(x)$. The following example demonstrates this fact [13, p. 320], [16, p. 171].

Example 2.20
As shown in Fig. 2.4, the sequence $\{x^k\}_{k=1}^{\infty}$ converges pointwise to the function $a(x)$ on $[0, 1]$ where

$$a(x) = \begin{cases} 0, & 0 \leq x < 1, \\ 1, & x = 1. \end{cases}$$

That is, $a(x)$ has a discontinuity at $x = 1$, although every function x^k is continuous on $[0, 1]$.

□

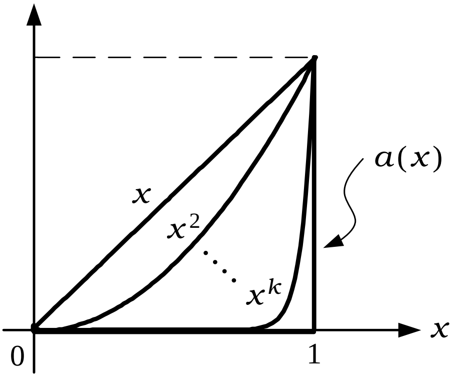

Fig. 2.4 Pointwise convergence of the sequence $\{x^k\}_{k=1}^{\infty}$ to a discontinuous function $a(x)$ on $[0, 1]$

Unlike pointwise convergence, uniform convergence ensures continuity, as the following theorem states [16, p. 174].

Theorem 2.21. *If the sequence $\{a_k(x)\}_{k=1}^{\infty}$ converges uniformly to a function $a(x)$ on an interval $[x_\mathrm{L}, x_\mathrm{U}]$ where every $a_k(x)$ is continuous on $[x_\mathrm{L}, x_\mathrm{U}]$, then the function $a(x)$ must be continuous on $[x_\mathrm{L}, x_\mathrm{U}]$.*

We leave the proof as an exercise (Problem 2.13). Theorem 2.21 implies that if every $a_k(x)$ is continuous but $a(x)$ is discontinuous, then it is not possible for the sequence $\{a_k(x)\}$ to converge uniformly to $a(x)$.

Example 2.22

Consider, again, the pointwise convergent sequence $\{x^k\}_{k=1}^{\infty}$ in Example 2.20. According to Theorem 2.21, it is clear that $\{x^k\}_{k=1}^{\infty}$ is not uniformly convergent on $[0, 1]$ since $a(x)$ has a discontinuity at $x = 1$.

□

2.2.2 Series

Closely related to a real or complex sequence $\{a_k\}_{k=m}^{n}$, a *series* is defined as $\sum_{k=m}^{n} a_k$. The series $\sum_{k=1}^{\infty} a_k$ is called a *one-sided infinite sequence* with the *nth partial sum* defined as

$$s_n = \sum_{k=1}^{n} a_k, \tag{2.48}$$

while the series $\sum_{k=-\infty}^{\infty} a_k$ is called a *two-sided infinite sequence* with the *nth partial sum* defined as

$$s_n = \sum_{k=-n}^{n} a_k. \tag{2.49}$$

Without loss of generality, we will only deal with the convergence of one-sided infinite series for brevity.

Series of Numbers

A series $\sum_{k=1}^{\infty} a_k$ is said to be *convergent* if the sequence of its partial sums, $\{s_n\}_{n=1}^{\infty}$, converges to

$$s \triangleq \sum_{k=1}^{\infty} a_k, \tag{2.50}$$

i.e. $\lim_{n\to\infty} s_n = s$ where s is called the *sum* or *value* of the series. From (2.48) and (2.50), it follows that if $\sum_{k=1}^{\infty} a_k$ is convergent, then the following condition always holds:

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} (s_n - s_{n-1}) = s - s = 0. \tag{2.51}$$

Moreover, a series $\sum_{k=1}^{\infty} a_k$ is said to be *absolutely convergent* if the series $\sum_{k=1}^{\infty} |a_k|$ is convergent.

Tests for Divergence and Convergence

In using a series $\sum_{k=1}^{\infty} a_k$, it is important to know whether $\sum_{k=1}^{\infty} a_k$ converges or diverges. The condition given by (2.51) suggests a test as follows.

Theorem 2.23 (Divergence Test). *Suppose $\sum_{k=1}^{\infty} a_k$ is a real or complex series to be tested. If the condition given by (2.51) is not satisfied, then the series $\sum_{k=1}^{\infty} a_k$ is divergent.*

Since the condition given by (2.51) is only necessary, but not sufficient, for convergence, it cannot be used for convergence testing. An example using the divergence test is as follows.

Example 2.24 (Geometric Series)
The partial sum of the geometric series $\sum_{k=1}^{\infty} \alpha r^k$ can be expressed as

$$s_n = \sum_{k=1}^{n} \alpha r^k = \alpha r \frac{1-r^n}{1-r}.$$

If $|r| < 1$, then the geometric series converges with

$$\lim_{n\to\infty} s_n = \frac{\alpha r}{1-r}.$$

On the other hand, if $|r| \geq 1$, then $\lim_{n\to\infty} \alpha r^n \neq 0$ which does not satisfy the condition given by (2.51), and thus the geometric series diverges.
□

The following test is useful for testing the convergence of a real series.

Theorem 2.25 (Integral Test). *Suppose $\sum_{k=1}^{\infty} a_k$ is a real series to be tested where $a_k \geq 0$ for all k. Find a continuous, positive, and decreasing function $f(x)$ on $[1,\infty)$ such that $f(k) = a_k$.*

- *If $\int_1^{\infty} f(x)dx$ is finite, then the series $\sum_{k=1}^{\infty} a_k$ is convergent.*
- *If $\int_1^{\infty} f(x)dx$ is infinite, then the series $\sum_{k=1}^{\infty} a_k$ is divergent.*

The proof is left as an exercise (Problem 2.14). An example using the integral test is as follows.

Example 2.26
To test the convergence of the real series $\sum_{k=1}^{\infty} 1/k^2$, let $f(x) = 1/x^2$. It is clear that $f(x)$ is continuous, positive, and decreasing on $[1,\infty)$ and $f(k) = 1/k^2$. By the integral test, $\int_1^{\infty} f(x)dx = 1$ implies that $\sum_{k=1}^{\infty} 1/k^2$ converges.
□

Series of Functions

Now consider a series $\sum_{k=1}^{\infty} a_k(x)$ whose nth partial sum is given by

$$s_n(x) = \sum_{k=1}^{n} a_k(x) \tag{2.52}$$

where $a_k(x)$ is a real or complex function of a real independent variable x. The series $\sum_{k=1}^{\infty} a_k(x)$ is said to *converge pointwise* to a real or complex function $s(x)$ if the sequence of its partial sums, $\{s_n(x)\}_{n=1}^{\infty}$, converges pointwise to $s(x)$, and is said to *converge uniformly* to $s(x)$ if the sequence $\{s_n(x)\}_{n=1}^{\infty}$ converges uniformly to $s(x)$.

According to the above-mentioned definitions, the convergence theory for sequences of functions can similarly apply to series of functions. In particular, uniform convergence of series also implies pointwise convergence of series, but the converse may not be true. Moreover, a pointwise convergent series $\sum_{k=1}^{\infty} a_k(x)$ may converge to a discontinuous function $s(x)$, even if every function $a_k(x)$ is continuous. And the following theorem directly follows from Theorem 2.21.

Theorem 2.27. *If the series $\sum_{k=1}^{\infty} a_k(x)$ converges uniformly to a function $s(x)$ on an interval $[x_\mathrm{L}, x_\mathrm{U}]$ where every $a_k(x)$ is continuous on $[x_\mathrm{L}, x_\mathrm{U}]$, then the function $s(x)$ is also continuous on $[x_\mathrm{L}, x_\mathrm{U}]$.*

As a remark, let us emphasize that there is no connection between uniform convergence and absolute convergence [4, p. 765].

Test for Uniform Convergence

The following test is most commonly used for testing the uniform convergence of series.

Theorem 2.28 (Weierstrass M-Test). *Suppose $\sum_{k=1}^{\infty} a_k(x)$ is a real or complex series to be tested on an interval $[x_\mathrm{L}, x_\mathrm{U}]$. If there exists a convergent series $\sum_{k=1}^{\infty} M_k$ such that each term $M_k \geq |a_k(x)|$ for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$, then the series $\sum_{k=1}^{\infty} a_k(x)$ is uniformly and absolutely convergent on $[x_\mathrm{L}, x_\mathrm{U}]$.*

Since the proof is lengthy and can be found, for instance, in [13], it is omitted here. An example using the Weierstrass M-test is provided as follows.

Example 2.29

Suppose $\sum_{k=1}^{\infty} e^{jkx}/k^2$ is the series to be tested on $[-\pi, \pi)$. Because $\left|e^{jkx}/k^2\right| \leq 1/k^2$ for all $x \in [-\pi, \pi)$ and $\sum_{k=1}^{\infty} 1/k^2$ converges (see Example 2.26), by the Weierstrass M-test, the series $\sum_{k=1}^{\infty} e^{jkx}/k^2$ is uniformly and absolutely convergent on $[-\pi, \pi)$.

□

2.2.3 Hilbert Spaces, Sequence Spaces and Function Spaces

Hilbert Spaces

Consider a sequence of real or complex vectors, denoted by $\{\mathbf{a}_n\}_{n=1}^{\infty}$, in a normed vector space $\mathcal{V}$. The sequence $\{\mathbf{a}_n\}_{n=1}^{\infty}$ is said to *converge in the norm* or, briefly, *converge* to a real or complex vector $\mathbf{a} \in \mathcal{V}$ if

$$\lim_{n\to\infty} \|\mathbf{a} - \mathbf{a}_n\| = 0. \tag{2.53}$$

Convergence in the norm is also often referred to as *convergence in the mean.* A sequence $\{\mathbf{a}_n\}_{n=1}^{\infty}$ in $\mathcal{V}$ is called a *Cauchy sequence* if for every real number $\varepsilon > 0$ there exists an integer N such that

$$\|\mathbf{a}_n - \mathbf{a}_m\| < \varepsilon \quad \text{for all } n > m \geq N. \tag{2.54}$$

Regarding Cauchy sequences, we have the following related theorem, whose proof is left as an exercise (Problem 2.15).

Theorem 2.30. *Every convergent sequence in a norm vector space $\mathcal{V}$ is a Cauchy sequence.*

The converse of Theorem 2.30, however, may be true for some norm vector spaces. If every Cauchy sequence in a norm vector space $\mathcal{V}$ converges to a vector in $\mathcal{V}$, then the normed vector space $\mathcal{V}$ is said to be *complete.* A complete normed vector space is also referred to as a *Banach space* [17].

Definition 2.31 (Hilbert Space). *A vector space $\mathcal{V}$ along with a legitimate norm and a legitimate inner product is said to be a Hilbert space if the normed vector space (i.e. $\mathcal{V}$ along with the legitimate norm) is complete and the inner product can induce the norm.*

As an example, the vector space $\mathcal{R}^N$ (see Example 2.3) along with the Euclidean norm and the Euclidean inner product is an N-dimensional Hilbert space [5, 14, 18], which we also refer to as the N-dimensional Euclidean space $\mathcal{R}^N$ for convenience.

Sequence Spaces

Consider a real or complex sequence $\{a_n\}_{n=1}^{\infty}$ which is bounded and satisfies

$$\left(\sum_{n=1}^{\infty} |a_n|^p\right)^{1/p} < \infty, \quad \text{for } 1 \leq p < \infty. \tag{2.55}$$

Let $\mathcal{V}$ be the set composed of all such sequences. Then, under the operations of componentwise addition and scalar multiplication of sequences, the set $\mathcal{V}$ can easily be shown to be a vector space (satisfying the axioms (VS1) through

(VS8)). The vector space $\mathcal{V}$ is a sequence space, commonly referred to as the ℓ^p *space* or, briefly, ℓ^p [5,13,17,18].

For notational simplicity, let $\mathbf{a} = (a_1, a_2, ..., a_n, ...)^T$ denote a vector corresponding to $\{a_n\}_{n=1}^{\infty} \in \ell^p$. The inner product of sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty} \in \ell^p$ is defined as

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{n=1}^{\infty} a_n b_n^*, \tag{2.56}$$

while the ℓ^p *norm* of $\{a_n\}_{n=1}^{\infty} \in \ell^p$ is defined as

$$\|\mathbf{a}\|_p = \begin{cases} \left(\sum_{n=1}^{\infty} |a_n|^p \right)^{1/p}, & \text{for } 1 \le p < \infty \\ \sup_{n=1,2,...} \{|a_n|\}, & \text{for } p = \infty \end{cases} \tag{2.57}$$

where the notation ‘sup’ stands for the *least upper bound* or the *supremum* of a set of real numbers.[5] From (2.56) and (2.57), it follows that only the ℓ^2 norm (i.e. $p = 2$) can be induced from (2.56). Furthermore, the ℓ^2 space along with the inner product defined as (2.56) and the ℓ^2 norm is known as an infinite-dimensional Hilbert space [14, p. 75]. As such, in what follows, the ℓ^2 space always refers to this Hilbert space for convenience.

Moreover, for ease of later use, we restate the Cauchy–Schwartz inequality in terms of two-sided sequences as follows.

Theorem 2.32 (Cauchy–Schwartz Inequality). *Suppose* $\{a_n\}_{n=-\infty}^{\infty}$ *and* $\{b_n\}_{n=-\infty}^{\infty}$ *are real or complex nonzero sequences with* $\sum_{n=-\infty}^{\infty} |a_n|^2 < \infty$ *and* $\sum_{n=-\infty}^{\infty} |b_n|^2 < \infty$*. Then*

$$\left| \sum_{n=-\infty}^{\infty} a_n b_n^* \right| \le \left(\sum_{n=-\infty}^{\infty} |a_n|^2 \right)^{1/2} \left(\sum_{n=-\infty}^{\infty} |b_n|^2 \right)^{1/2} \tag{2.58}$$

and the equality holds if and only if $a_n = \alpha b_n$ *for all* n *where* $\alpha \neq 0$ *is an arbitrary real or complex scalar.*

Also with regard to two-sided sequences, the following inequality is useful in development of blind equalization algorithms [19,20].

Theorem 2.33. *Suppose* $\{a_n\}_{n=-\infty}^{\infty}$ *is a real or complex nonzero sequence with* $\sum_{n=-\infty}^{\infty} |a_n|^s < \infty$ *where* s *is an integer and* $1 \le s < \infty$*. Then*

[5] One should not confuse “supremum” with “maximum.” A set which is bounded above has a supremum, but may not have a maximum (the largest element of the set) [12, p. 16]. For instance, the set $\{1 - (1/n), n = 1 \sim \infty\}$ has a supremum equal to one, but does not have any maximum.

$$\left(\sum_{n=-\infty}^{\infty} |a_n|^l\right)^{1/l} \le \left(\sum_{n=-\infty}^{\infty} |a_n|^s\right)^{1/s} \tag{2.59}$$

and the equality holds if and only if a_n has only one nonzero term where l is an integer and $l > s$.

See Appendix 2C for the proof.

Function Spaces

Consider a real or complex functions $f(x)$ which is bounded and satisfies

$$\left(\int_{x_\mathrm{L}}^{x_\mathrm{U}} |f(x)|^p dx\right)^{1/p}, \quad \text{for } 1 \le p < \infty. \tag{2.60}$$

Then the set of all such functions forms a function space (a vector space) under the operations of pointwise addition and scalar multiplication of functions. The function space is commonly referred to as the $\mathcal{L}^p[x_\mathrm{L}, x_\mathrm{U}]$ space or, briefly, $\mathcal{L}^p[x_\mathrm{L}, x_\mathrm{U}]$ [5, 13, 17, 18].

Define the inner product of functions $f(x)$ and $g(x) \in \mathcal{L}^p[x_\mathrm{L}, x_\mathrm{U}]$ as

$$\langle f, g\rangle = \int_{x_\mathrm{L}}^{x_\mathrm{U}} f(x)g(x)^* dx \tag{2.61}$$

and the $\mathcal{L}^p$ *norm* of $f(x) \in \mathcal{L}^p[x_\mathrm{L}, x_\mathrm{U}]$ as

$$\|f\|_p = \left(\int_{x_\mathrm{L}}^{x_\mathrm{U}} |f(x)|^p dx\right)^{1/p}, \quad \text{for } 1 \le p < \infty. \tag{2.62}$$

Only the $\mathcal{L}^2$ norm ($p = 2$) can be induced from (2.61). More importantly, due to the operation of integration in (2.62), $\|f\|_2 = 0$ merely implies that $f(x) = 0$ *almost everywhere* on $[x_\mathrm{L}, x_\mathrm{U}]$, that is, $f(x)$ may not be identically zero on a set of points on which the integration is "negligible." [6] From this, it follows that the inner product defined as (2.61) does not satisfy the axiom (IPS3) and the $\mathcal{L}^2$ norm does not satisfy the axiom (NVS3). To get around this difficulty, we adopt the following convention: $\|f\|_2 = 0$ implies that $f(x)$ is a zero function, i.e. $f(x) = 0$ for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$. With this convention, the $\mathcal{L}^2[x_\mathrm{L}, x_\mathrm{U}]$ space along with the inner product defined as (2.61) and the $\mathcal{L}^2$ norm is also known as an infinite-dimensional Hilbert space [18, p. 193]. In what follows, the $\mathcal{L}^2[x_\mathrm{L}, x_\mathrm{U}]$ space always refers to this Hilbert space for convenience.

Moreover, the Cauchy–Schwartz inequality described in Theorem 2.1 is applicable to the $\mathcal{L}^2[x_\mathrm{L}, x_\mathrm{U}]$ space, that is restated here in terms of functions with the above convention.

[6] The set of points on which integration is "negligible" is called a set of *measure zero* [13, 14].

Theorem 2.34 (Cauchy–Schwartz Inequality). *Suppose* $f(x)$ *and* $g(x)$ *are real or complex nonzero functions on* $[x_{\mathrm{L}}, x_{\mathrm{U}}]$ *with* $\int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx < \infty$ *and* $\int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |g(x)|^2 dx < \infty$*. Then*

$$\left| \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f(x) g(x)^* dx \right| \leq \left\{ \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx \right\}^{1/2} \left\{ \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |g(x)|^2 dx \right\}^{1/2} \tag{2.63}$$

and the equality holds if and only if $f(x) = \alpha g(x)$ *for all* $x \in [x_{\mathrm{L}}, x_{\mathrm{U}}]$ *where* $\alpha \neq 0$ *is an arbitrary real or complex scalar.*

Approximations in Function Spaces

Let us emphasize that any function in $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$ is actually viewed as a vector in the vector space. As such, convergence for a sequence of functions in $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$ means convergence in the norm for a sequence of vectors, that is closely related to the problem of minimum mean-square-error (MMSE) approximation in $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$ as revealed below.

Let $\{\phi_1(x), \phi_2(x), ..., \phi_n(x)\}$ be a set of real or complex orthogonal functions in $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$ where

$$\int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} \phi_k(x) \phi_m^*(x) dx = \begin{cases} E_\phi, & k = m, \\ 0, & k \neq m. \end{cases} \tag{2.64}$$

Given a real or complex function $f(x) \in \mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$, let us consider the problem of approximating the nth partial sum

$$s_n(x) = \sum_{k=-n}^{n} \theta_k \phi_k(x) \tag{2.65}$$

to the function $f(x)$ in the MMSE sense, i.e. finding the optimal parameters $\theta_{-n}, \theta_{-n+1}, ..., \theta_n$ such that the following mean-square-error (MSE) is minimum:

$$J_{\mathrm{MSE}}(\theta_k) = \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x) - s_n(x)|^2 \, dx. \tag{2.66}$$

By substituting (2.65) into (2.66) and using (2.64), we obtain

$$\begin{aligned}
J_{\mathrm{MSE}}(\theta_k) &= \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx + E_\phi \sum_{k=-n}^{n} |\theta_k|^2 \\
&\quad - \sum_{k=-n}^{n} \left[\theta_k^* \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f(x)\phi_k^*(x)dx + \theta_k \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f^*(x)\phi_k(x)dx \right] \\
&= \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx + E_\phi \sum_{k=-n}^{n} \left| \theta_k - \frac{1}{E_\phi} \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f(x)\phi_k^*(x)dx \right|^2 \\
&\quad - E_\phi \sum_{k=-n}^{n} \left| \frac{1}{E_\phi} \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f(x)\phi_k^*(x)dx \right|^2 .
\end{aligned}$$

This implies that the optimal θ_k, denoted by $\widehat{\theta}_k$, is given by

$$\widehat{\theta}_k = \frac{1}{E_\phi} \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} f(x)\phi_k^*(x)dx \quad \text{for } k = -n, -n+1, ..., n, \tag{2.67}$$

and the corresponding minimum value of $J_{\mathrm{MSE}}(\theta_k)$ is given by

$$\min\{J_{\mathrm{MSE}}(\theta_k)\} = \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx - E_\phi \sum_{k=-n}^{n} |\widehat{\theta}_k|^2. \tag{2.68}$$

Since (2.68) holds for any n and $J_{\mathrm{MSE}}(\theta_k) \geq 0$ (see (2.66)), letting $n \to \infty$ leads to the following inequality.

Theorem 2.35 (Bessel's Inequality). *Suppose* $\{\phi_1(x), \phi_2(x), ..., \phi_n(x)\}$ *is a set of real or complex orthogonal functions in* $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$*. If* $f(x)$ *is a real or complex function in* $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$*, then optimal approximation of the series* $\sum_{k=-\infty}^{\infty} \theta_k \phi_k(x)$ *to* $f(x)$ *in the MMSE sense gives*

$$\sum_{k=-\infty}^{\infty} |\widehat{\theta}_k|^2 \leq \frac{1}{E_\phi} \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx < \infty \tag{2.69}$$

where $\widehat{\theta}_k$ *is the optimal* θ_k *and* $E_\phi = \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |\phi_k(x)|^2 dx$*.*

From (2.66) and (2.62), it follows that when the sequence of functions $\{s_n(x)\}_{n=1}^{\infty}$ converges in the norm to $f(x) \in \mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$,

$$\lim_{n\to\infty} \|f - s_n\|_2 = \lim_{n\to\infty} \sqrt{J_{\mathrm{MSE}}(\widehat{\theta}_k)} = 0. \tag{2.70}$$

Correspondingly, Bessel's inequality (2.69) becomes the equality

$$\sum_{k=-\infty}^{\infty} |\widehat{\theta}_k|^2 = \frac{1}{E_\phi} \int_{x_{\mathrm{L}}}^{x_{\mathrm{U}}} |f(x)|^2 dx, \tag{2.71}$$

which is known as *Parseval's equality* or *Parseval's relation.* Owing to (2.70), convergence in the norm for the $\mathcal{L}^2[x_{\mathrm{L}}, x_{\mathrm{U}}]$ space is also referred to as *convergence in the mean-square (MS) sense.*

2.2.4 Fourier Series

Fourier series are of great importance in developing the theory of mathematical analysis, and have widespread applications in the areas of science and engineering such as signal representation and analysis in signal processing.

Consider that $f(x)$ is a periodic function with period 2π. When $f(x)$ is real, the *Fourier series* of $f(x)$ is given by

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx) \tag{2.72}$$

where a_k and b_k are given by

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx, \quad k = 0, 1, 2, ... \tag{2.73}$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx, \quad k = 1, 2, ... \tag{2.74}$$

The real numbers a_k and b_k are called the *Fourier coefficients* of $f(x)$. Note that $\{1, \cos kx, \sin kx, k = 1 \sim \infty\}$ is a set of orthogonal functions satisfying (2.64) ($E_\phi = \pi$). From (2.73) and (2.74), one can see that if $f(x)$ is odd, then $a_k = 0$ for all k; whereas if $f(x)$ is even, $b_k = 0$ for all k. On the other hand, when $f(x)$ is complex, the Fourier series of $f(x)$ is given by

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{jkx} \tag{2.75}$$

where c_k, a Fourier coefficient of $f(x)$, is a complex number given by

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-jkx}dx. \tag{2.76}$$

Note that $\{e^{jkx}, k = -\infty \sim \infty\}$ is also a set of orthogonal functions satisfying (2.64) ($E_\phi = 2\pi$).

Next, let us discuss the existence of Fourier series. In particular, we are concerned with the sufficient conditions under which the Fourier series given by (2.75) converges.

Local Behavior of Convergence

With the nth partial sum defined as

$$s_n(x) = \sum_{k=-n}^{n} c_k e^{jkx}, \tag{2.77}$$

the convergence problem of the Fourier series given by (2.75) is the same as that of the sequence $\{s_n(x)\}_{n=1}^{\infty}$.

Pointwise Convergence

It was believed, for a long time, that if the periodic function $f(x)$ is continuous, then the Fourier series would converge to $f(x)$ for all $x \in [-\pi, \pi)$ (i.e. pointwise convergence). Actually, there do exist continuous periodic functions whose Fourier series diverge at a given point or even everywhere; see [14, pp. 83–87] for an example of such functions. This implies that pointwise convergence requires some additional conditions on $f(x)$ as follows [18].

Theorem 2.36 (Pointwise Convergence Theorem). *Suppose $f(x)$ is a real or complex periodic function of period 2π. Then, under the conditions that (i) $f(x)$ is piecewise continuous on $[-\pi, \pi)$ and (ii) the derivative $f'(x)$ is piecewise continuous on $[-\pi, \pi)$, the Fourier series of $f(x)$ given by (2.75) is pointwise convergent and*

$$\lim_{n\to\infty} s_n(x) = \frac{f(x^-) + f(x^+)}{2} \quad \textit{for all } x \in [-\pi, \pi) \tag{2.78}$$

where $s_n(x)$ is the corresponding nth partial sum given by (2.77), and $f(x^-)$ and $f(x^+)$ are the left-hand limit and the right-hand limit of $f(x)$, respectively.

See Appendix 2B for a review of terminologies of functions and see Appendix 2D for the proof of this theorem. From this theorem and (2.219), it follows that the Fourier series converges to $f(x)$ at the points of continuity and converges to $[f(x^-) + f(x^+)]/2$ at the points of discontinuity. Note that Theorem 2.36 is only a special case of the *Dirichlet Theorem*, for which the required conditions are known as the *Dirichlet conditions* [21, 22].[7]

Uniform Convergence

By using the Weierstrass M-test, we have the following theorem for uniform and absolute convergence of the Fourier series (Problem 2.17).

Theorem 2.37. *Suppose $\{c_k\}_{k=-\infty}^{\infty}$ is any absolutely summable sequence, i.e. $\sum_{k=-\infty}^{\infty} |c_k| < \infty$. Then the Fourier series $\sum_{k=-\infty}^{\infty} c_k e^{jkx}$ converges uniformly and absolutely to a continuous function of x on $[-\pi, \pi)$.*

Moreover, by using the Weierstrass M-test and the pointwise convergence theorem with more restrictive conditions on $f(x)$, we have another theorem regarding the uniform and absolute convergence [18, pp. 216–218].

Theorem 2.38. *Suppose $f(x)$ is a real or complex periodic function of period 2π. Then, under the conditions that (i) $f(x)$ is continuous on $[-\pi, \pi)$ and (ii) the derivative $f'(x)$ is piecewise continuous on $[-\pi, \pi)$, the Fourier series of $f(x)$ given by (2.75) converges uniformly and absolutely to $f(x)$ on $[-\pi, \pi)$.*

See Appendix 2E for the proof.

[7] The Dirichlet theorem due to P. L. Dirichlet (1829) was the first substantial progress on the convergence problem of Fourier series [13].

Global Behavior of Convergence

The Fourier series given by (2.75) is said to *converge in the mean-square (MS) sense* to $f(x)$ if

$$\lim_{n\to\infty}\int_{-\pi}^{\pi}|f(x)-s_n(x)|^2\,dx=0 \tag{2.79}$$

where $s_n(x)$ is the nth partial sum given by (2.77). Accordingly, with MS convergence, we can only get an overall picture about the convergence behavior over the entire interval. It reveals nothing about the detailed behavior of convergence at any point.

Recall that if $f(x)$ is in the $\mathcal{L}^2[-\pi,\pi)$ space, then MS convergence is equivalent to convergence in the norm. Correspondingly, Parseval's relation

$$\sum_{n=-\infty}^{\infty}|c_k|^2=\frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|^2dx<\infty \tag{2.80}$$

holds and thus the sequence $\{c_k\}_{k=-\infty}^{\infty}$ is square summable. The converse is stated in the following theorem (Problem 2.18).

Theorem 2.39. *Suppose $\{c_k\}_{k=-\infty}^{\infty}$ is any square summable sequence, i.e. $\sum_{k=-\infty}^{\infty}|c_k|^2<\infty$. Then the Fourier series $\sum_{k=-\infty}^{\infty}c_ke^{jkx}$ converges in the MS sense to a function in the $\mathcal{L}^2[-\pi,\pi)$ space.*

Furthermore, a more generalized theorem regarding the MS convergence is provided as follows. The proof is beyond the scope of this book; the reader can find it in [13, pp. 411–414] for the real case and [14, pp. 76–80] for the complex case.

Theorem 2.40. *Suppose $f(x)$ is a real or complex periodic function of period 2π. If the function $f(x)$ is bounded and integrable on $[-\pi,\pi)$, then the Fourier series of $f(x)$ given by (2.75) converges in the MS sense to $f(x)$ on $[-\pi,\pi)$.*

Compared with local convergence (pointwise convergence and uniform convergence), global convergence (MS convergence) requires even weaker conditions on the function $f(x)$ or the sequence $\{c_k\}_{k=-\infty}^{\infty}$ and so the existence of Fourier series is almost not an issue in practice.

Fourier Series of Generalized Functions

In some cases, we may need to deal with functions which are outside the ordinary scope of function theory. An important class of such functions is the one of *generalized functions* introduced by G. Temple (1953) [23]. Among this class, a representative is the so-called *impulse* or *Dirac delta function*, commonly denoted by $\delta(x)$.[8] It is mathematically defined by the following relations

$$\begin{cases} \delta(x) = 0 \quad \text{for } x \neq 0, \\ \int_{-\infty}^{\infty} \delta(x) dx = 1, \end{cases} \tag{2.81}$$

and possesses the following sifting property:

$$\int_{-\infty}^{\infty} \delta(x-\tau) f(x) dx = f(\tau). \tag{2.82}$$

Strictly speaking, a periodic function like

$$f(x) = \sum_{m=-\infty}^{\infty} 2\pi\delta(x + 2\pi m) \tag{2.83}$$

does not have a Fourier series. But, using (2.76) and (2.82), we can still mathematically define the Fourier series of $f(x)$ as

$$f(x) = \sum_{k=-\infty}^{\infty} e^{jkx} \quad (\text{i.e. } c_k = 1 \text{ for all } k) \tag{2.84}$$

and make use of this in many applications. In other words, the theory of Fourier series should be broadened for more extensive applications. The extended theory of Fourier series is, however, beyond the scope of this book; refer to [23, 24] for the details.

2.3 Optimization Theory

Consider that $J(\boldsymbol{\theta})$ is a real function of the $L \times 1$ vector

$$\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_L)^T \tag{2.85}$$

where θ_1, θ_2, ..., θ_L are real or complex unknown parameters to be determined. An optimization problem is to find (search for) a solution for $\boldsymbol{\theta}$ which minimizes or maximizes the function $J(\boldsymbol{\theta})$, referred to as the *objective function.* There are basically two types of optimization problems, *constrained optimization problems* and *unconstrained optimization problems* [12, 25, 26]. As the names indicate, the former type is subject to some constraints (e.g. equality constraints and inequality constraints), whereas the latter type does not involve any constraint. In the scope of the book, we are interested in unconstrained optimization problems, which along with the related theory are introduced in this section.

[8] The notation '$\delta(x)$' for Dirac delta function was first used by G. Kirchhoff, and then introduced into quantum mechanics by Dirac (1927) [23].

2.3.1 Vector Derivatives

As we will see, finding the solutions to the minima or maxima of the objective function $J(\boldsymbol{\theta})$ often involves manipulations of the following *first derivative* (with respect to $\boldsymbol{\theta}^*$)

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1^*}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_2^*}, \ldots, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_L^*} \right)^T \tag{2.86}$$

where $f(\boldsymbol{\theta})$ is an arbitrary real or complex function of $\boldsymbol{\theta}$ and $\partial f(\boldsymbol{\theta})/\partial \theta_k^*$ is the *first partial derivative* of $f(\boldsymbol{\theta})$ with respect to the conjugate parameter θ_k^*.[9] However, the first derivative $\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^*$, or equivalently the operator

$$\frac{\partial}{\partial \boldsymbol{\theta}^*} = \left(\frac{\partial}{\partial \theta_1^*}, \frac{\partial}{\partial \theta_2^*}, \ldots, \frac{\partial}{\partial \theta_L^*} \right)^T, \tag{2.87}$$

depends on whether $\boldsymbol{\theta}$ is real or complex, as discussed below.

Derivatives with Respect to a Real Vector

When $\boldsymbol{\theta}$ is real, applying the operator $\partial/\partial \boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^T$ yields

$$\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}^*} = \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial \theta_1}{\partial \theta_1} & \frac{\partial \theta_2}{\partial \theta_1} & \cdots & \frac{\partial \theta_L}{\partial \theta_1} \\ \frac{\partial \theta_1}{\partial \theta_2} & \frac{\partial \theta_2}{\partial \theta_2} & \cdots & \frac{\partial \theta_L}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \theta_1}{\partial \theta_L} & \frac{\partial \theta_2}{\partial \theta_L} & \cdots & \frac{\partial \theta_L}{\partial \theta_L} \end{pmatrix} = \mathbf{I}, \tag{2.88}$$

which is useful to the derivation of $\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. In particular, if $f(\boldsymbol{\theta}) = \mathbf{b}^T \boldsymbol{\theta} = \boldsymbol{\theta}^T \mathbf{b}$ where the vector $\mathbf{b}$ is independent of $\boldsymbol{\theta}$, then

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{b} = \mathbf{I}\mathbf{b} = \mathbf{b}. \tag{2.89}$$

Moreover, if $f(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{b}(\boldsymbol{\theta})$ where the vector $\mathbf{b}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta}$, then

$$\begin{aligned} \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{b}(\boldsymbol{\theta}) + \frac{\partial \mathbf{b}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \boldsymbol{\theta} = \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{A}\boldsymbol{\theta} + \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} \right) \mathbf{A}^T \boldsymbol{\theta} \\ &= \mathbf{I}\mathbf{A}\boldsymbol{\theta} + \mathbf{I}\mathbf{A}^T\boldsymbol{\theta} = (\mathbf{A} + \mathbf{A}^T)\boldsymbol{\theta}, \end{aligned} \tag{2.90}$$

which reduces to $\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 2\mathbf{A}\boldsymbol{\theta}$ when $\mathbf{A}$ is symmetric.

[9] Although utilization of $\partial f(\boldsymbol{\theta})/\boldsymbol{\theta}^*$ and that of $\partial f(\boldsymbol{\theta})/\boldsymbol{\theta}$ both lead to the same solutions for the optimization problems, Brandwood [27] has pointed out that the former gives rise to a slightly neater expression and thus is more convenient.

Derivatives with Respect to a Complex Vector

Now consider the case that $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_L)^T$ is complex, i.e.

$$\theta_k = x_k + jy_k, \quad k = 1, 2, ..., L, \tag{2.91}$$

where $x_k = \text{Re}\{\theta_k\}$ is the real part of θ_k and $y_k = \text{Im}\{\theta_k\}$ is the imaginary part of θ_k. Naturally, one can derive $\partial f(\boldsymbol{\theta})/\partial \theta_k^*$ in terms of x_k and y_k. Alternatively, direct derivation of $\partial f(\boldsymbol{\theta})/\partial \theta_k^*$ (without involving x_k and y_k) is more appealing, but special care should be taken for the following reason. In conventional complex-variable theory, if $f(\boldsymbol{\theta})$ cannot be expressed in terms of only θ_k (i.e. it also consists of θ_k^*), then it is nowhere *differentiable* by θ_k and we say that $f(\boldsymbol{\theta})$ is not *analytic* [28]. The analytic problem, however, can be resolved by simply treating $f(\boldsymbol{\theta}) \equiv f(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ as a function of $2L$ independent variables $\theta_1, \theta_2, ..., \theta_L, \theta_1^*, \theta_2^*, ..., \theta_L^*$ [27]; see the following illustration.

Example 2.41
Consider the function $f(\theta) = \theta^*$ where $\theta = x + jy$, and x and y are real. According to the conventional complex-variable theory, the first derivative of $f(\theta)$ with respect to θ is given by [28]

$$\frac{df(\theta)}{d\theta} = \lim_{\Delta\theta \to 0} \frac{f(\theta + \Delta\theta) - f(\theta)}{\Delta\theta} = \lim_{\Delta\theta \to 0} \frac{\Delta\theta^*}{\Delta\theta}.$$

As illustrated in Fig. 2.5, if $\Delta\theta$ approaches zero along the real axis, i.e. $\Delta\theta = \Delta x \to 0$, then $df(\theta)/d\theta = 1$. If $\Delta\theta$ approaches zero along the imaginary axis, i.e. $\Delta\theta = j\Delta y \to 0$, then $df(\theta)/d\theta = -1$. As a result, there is no way to assign a unique value to $df(\theta)/d\theta$, and thus $f(\theta)$ is not differentiable. On the other hand, by treating $f(\theta) \equiv f(\theta, \theta^*)$ as a function of independent variables θ and θ^*, we obtain $\partial f(\theta, \theta^*)/\partial\theta = 0$ and $\partial f(\theta, \theta^*)/\partial\theta^* = 1$. That is, $f(\theta, \theta^*)$ is differentiable with respect to θ and θ^* independently.

□

With the treatment of independent variables θ_k and θ_k^*, we now proceed to derive the partial derivative $\partial f(\boldsymbol{\theta})/\partial \theta_k^*$. From (2.91), it follows that

$$x_k = \frac{1}{2}(\theta_k + \theta_k^*) \quad \text{and} \quad y_k = \frac{1}{2j}(\theta_k - \theta_k^*). \tag{2.92}$$

Differentiating x_k and y_k given by (2.92) with respect to θ_k and θ_k^* yields

$$\frac{\partial x_k}{\partial \theta_k} = \frac{1}{2}, \quad \frac{\partial x_k}{\partial \theta_k^*} = \frac{1}{2}, \quad \frac{\partial y_k}{\partial \theta_k} = \frac{1}{2j}, \quad \text{and} \quad \frac{\partial y_k}{\partial \theta_k^*} = -\frac{1}{2j}. \tag{2.93}$$

This, together with the *chain rule* [29], leads to

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k} = \frac{\partial f(\boldsymbol{\theta})}{\partial x_k}\frac{\partial x_k}{\partial \theta_k} + \frac{\partial f(\boldsymbol{\theta})}{\partial y_k}\frac{\partial y_k}{\partial \theta_k} = \frac{1}{2}\left\{\frac{\partial f(\boldsymbol{\theta})}{\partial x_k} - j\frac{\partial f(\boldsymbol{\theta})}{\partial y_k}\right\} \tag{2.94}$$

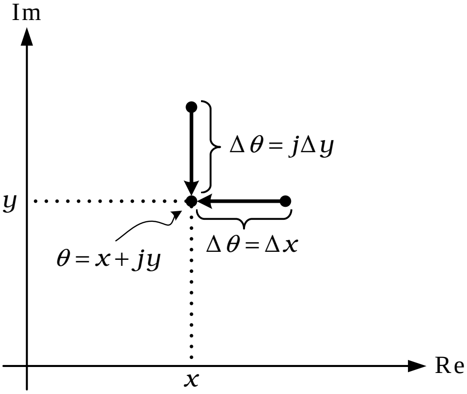

Fig. 2.5 Illustration of $\Delta\theta$ approaching zero along the real and imaginary axes

and

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_k^*} = \frac{\partial f(\boldsymbol{\theta})}{\partial x_k}\frac{\partial x_k}{\partial \theta_k^*} + \frac{\partial f(\boldsymbol{\theta})}{\partial y_k}\frac{\partial y_k}{\partial \theta_k^*} = \frac{1}{2}\left\{\frac{\partial f(\boldsymbol{\theta})}{\partial x_k} + j\frac{\partial f(\boldsymbol{\theta})}{\partial y_k}\right\}. \tag{2.95}$$

From (2.91), (2.94) and (2.95), it is clear that

$$\frac{\partial \theta_k^*}{\partial \theta_k} = \frac{\partial \theta_k}{\partial \theta_k^*} = 0 \quad \text{and} \quad \frac{\partial \theta_k}{\partial \theta_k} = \frac{\partial \theta_k^*}{\partial \theta_k^*} = 1. \tag{2.96}$$

By (2.96), we have

$$\frac{\partial \boldsymbol{\theta}^H}{\partial \boldsymbol{\theta}^*} = \mathbf{I} \quad \text{and} \quad \frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}^*} = \mathbf{0}, \tag{2.97}$$

which, again, are useful to the derivation of $\partial f(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^*$. In particular, if $f(\boldsymbol{\theta}) = \mathbf{b}^H\boldsymbol{\theta}$ where $\mathbf{b}$ is independent of $\boldsymbol{\theta}$, then $\partial f(\boldsymbol{\theta})/\boldsymbol{\theta}^* = \mathbf{0}$, and if $f(\boldsymbol{\theta}) = \boldsymbol{\theta}^H\mathbf{b}$, then

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = \left(\frac{\partial \boldsymbol{\theta}^H}{\partial \boldsymbol{\theta}^*}\right)\mathbf{b} = \mathbf{b}. \tag{2.98}$$

If $f(\boldsymbol{\theta}) = \boldsymbol{\theta}^H\mathbf{A}\boldsymbol{\theta}$, then

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = \left(\frac{\partial \boldsymbol{\theta}^H}{\partial \boldsymbol{\theta}^*}\right)\mathbf{A}\boldsymbol{\theta} + \left(\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}^*}\right)\left(\boldsymbol{\theta}^H\mathbf{A}\right)^T = \mathbf{A}\boldsymbol{\theta} \quad \text{(by (2.97))}. \tag{2.99}$$

Table 2.1 summarizes the vector derivatives for both real and complex cases.

Table 2.1 Summary of vector derivatives

Real Case				Complex Case			
$\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}} = \mathbf{I}$				$\frac{\partial \boldsymbol{\theta}^H}{\partial \boldsymbol{\theta}^*} = \mathbf{I}$ and $\frac{\partial \boldsymbol{\theta}^T}{\partial \boldsymbol{\theta}^*} = \mathbf{0}$			
$f(\boldsymbol{\theta})$	$\boldsymbol{\theta}^T\mathbf{b}$	$\mathbf{b}^T\boldsymbol{\theta}$	$\boldsymbol{\theta}^T\mathbf{A}\boldsymbol{\theta}$	$f(\boldsymbol{\theta})$	$\boldsymbol{\theta}^H\mathbf{b}$	$\mathbf{b}^H\boldsymbol{\theta}$	$\boldsymbol{\theta}^H\mathbf{A}\boldsymbol{\theta}$
$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$	$\mathbf{b}$	$\mathbf{b}$	$(\mathbf{A}+\mathbf{A}^T)\boldsymbol{\theta}$	$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}$	$\mathbf{b}$	$\mathbf{0}$	$\mathbf{A}\boldsymbol{\theta}$

2.3.2 Necessary and Sufficient Conditions for Solutions

From the foregoing discussions, we note that when the unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_L)^T$ is complex, it is also more convenient to treat the real objective function $J(\boldsymbol{\theta}) \equiv J(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ as a function of independent variables θ_k and θ_k^*. As such, for notational convenience, let us reformulate the above-mentioned optimization problem into the equivalent problem of minimizing or maximizing the real objective function $J(\boldsymbol{\vartheta})$ where $\boldsymbol{\vartheta}$ is the real or complex unknown parameter vector defined as

$$\begin{cases} \boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, ..., \vartheta_L)^T = \boldsymbol{\theta} & \text{for real } \boldsymbol{\theta}, \\ \boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, ..., \vartheta_{2L})^T = (\boldsymbol{\theta}^T, \boldsymbol{\theta}^H)^T & \text{for complex } \boldsymbol{\theta}. \end{cases} \tag{2.100}$$

Several terminologies regarding $J(\boldsymbol{\vartheta})$ are introduced as follows.

The objective function $J(\boldsymbol{\vartheta})$ is said to have a *local minimum* or a *relative minimum* at the solution point $\widehat{\boldsymbol{\vartheta}}$ if there exists a real number $\varepsilon > 0$ such that

$$J(\widehat{\boldsymbol{\vartheta}}) \leq J(\boldsymbol{\vartheta}) \quad \text{for all } \boldsymbol{\vartheta} \text{ satisfying } ||\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}|| < \varepsilon. \tag{2.101}$$

The objective function $J(\boldsymbol{\vartheta})$ is said to have a *global minimum* or an *absolute minimum* at the solution point $\widehat{\boldsymbol{\vartheta}}$ if

$$J(\widehat{\boldsymbol{\vartheta}}) \leq J(\boldsymbol{\vartheta}) \quad \text{for all } \boldsymbol{\vartheta}. \tag{2.102}$$

Similarly, the objective function $J(\boldsymbol{\vartheta})$ is said to have a *local maximum* or a *relative maximum* at the solution point $\widehat{\boldsymbol{\vartheta}}$ if there exists a real number $\varepsilon > 0$ such that

$$J(\widehat{\boldsymbol{\vartheta}}) \geq J(\boldsymbol{\vartheta}) \quad \text{for all } \boldsymbol{\vartheta} \text{ satisfying } ||\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}|| < \varepsilon, \tag{2.103}$$

and have a *global maximum* or an *absolute maximum* at the solution point $\widehat{\boldsymbol{\vartheta}}$ if

$$J(\widehat{\boldsymbol{\vartheta}}) \geq J(\boldsymbol{\vartheta}) \quad \text{for all } \boldsymbol{\vartheta}. \tag{2.104}$$

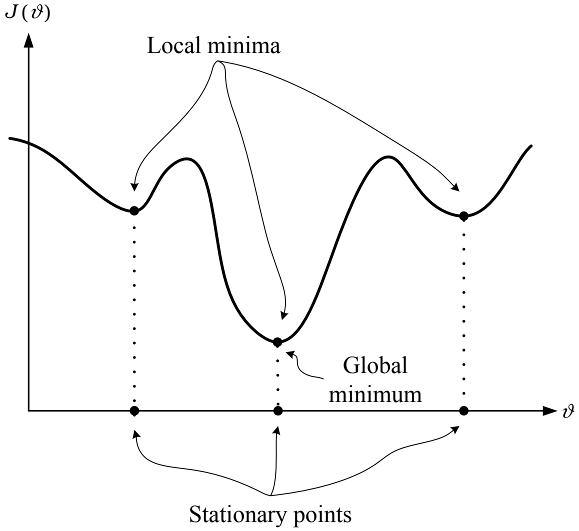

Fig. 2.6 Illustration of the solution points for the problem of minimizing $J(\vartheta)$ where ϑ is real

In other words, a global minimum (maximum) of $J(\boldsymbol{\vartheta})$ is also a local minimum (maximum) of $J(\boldsymbol{\vartheta})$. Figure 2.6 gives an illustration of these definitions.

Define the *gradient vector*, or simply the *gradient*, as[10]

$$\nabla J(\boldsymbol{\vartheta}) = \frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*} \tag{2.105}$$

(the physical meaning will be discussed later), where

$$\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*} = \begin{cases} \dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} & \text{for real } \boldsymbol{\theta}, \\ \left(\left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right]^T, \left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^T \right)^T & \text{for complex } \boldsymbol{\theta}. \end{cases} \tag{2.106}$$

A necessary condition for the local extrema (local minima or maxima) of $J(\boldsymbol{\vartheta})$ is as follows [26, p. 73].

Theorem 2.42 (Necessary Condition). *If the objective function $J(\boldsymbol{\vartheta})$ has an extremum at $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$ and if its first derivative $\partial J(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}^*$ exists at $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$, then its gradient*

[10] The gradient $\nabla J(\boldsymbol{\vartheta})$ defined as (2.105) is the same as that defined in [8, p. 894] except for a scale factor.

$$\nabla J(\widehat{\boldsymbol{\vartheta}}) \triangleq \nabla J(\boldsymbol{\vartheta})\Big|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}} = \mathbf{0}. \tag{2.107}$$

The proof is left as an exercise (Problem 2.19). When $\widehat{\boldsymbol{\vartheta}}$ satisfies (2.107), it is said to be a *stationary point* of $J(\boldsymbol{\vartheta})$. Furthermore, a stationary point $\widehat{\boldsymbol{\vartheta}}$ is said to be a *saddle point* of $J(\boldsymbol{\vartheta})$ if it corresponds to a local minimum of $J(\boldsymbol{\vartheta})$ with respect to one direction on the hypersurface of $J(\boldsymbol{\vartheta})$ and a local maximum of $J(\boldsymbol{\vartheta})$ with respect to another direction [12, 26, 30]. In other words, a saddle point of $J(\boldsymbol{\vartheta})$ corresponds to an unstable equilibrium of $J(\boldsymbol{\vartheta})$, and thus it will typically not be obtained by optimization methods.

Example 2.43 (Saddle Point)
Consider the objective function $J(\boldsymbol{\vartheta}) = J(\vartheta_1, \vartheta_2) = -\vartheta_1^2 + \vartheta_2^2$ where $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^T$, and ϑ_1 and ϑ_2 are real. Taking the first derivative of $J(\boldsymbol{\vartheta})$ with respect to $\boldsymbol{\vartheta}^*$ $(= \boldsymbol{\vartheta})$

$$\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \begin{pmatrix} \partial J(\boldsymbol{\vartheta})/\partial\vartheta_1 \\ \partial J(\boldsymbol{\vartheta})/\partial\vartheta_2 \end{pmatrix} = \begin{pmatrix} -2\vartheta_1 \\ 2\vartheta_2 \end{pmatrix}$$

and setting the result to zero, we obtain the stationary point $\widehat{\boldsymbol{\vartheta}} = (\widehat{\vartheta}_1, \widehat{\vartheta}_2)^T = (0,0)^T$. Figure 2.7 depicts the objective function $J(\vartheta_1, \vartheta_2)$ and the stationary point $(\widehat{\vartheta}_1, \widehat{\vartheta}_2) = (0,0)$. One can see from this figure that the function $J(\vartheta_1, \widehat{\vartheta}_2) = J(\vartheta_1, 0) = -\vartheta_1^2$ has a local maximum at $\vartheta_1 = \widehat{\vartheta}_1 = 0$, and the function $J(\widehat{\vartheta}_1, \vartheta_2) = J(0, \vartheta_2) = \vartheta_2^2$ has a local minimum at $\vartheta_2 = \widehat{\vartheta}_2 = 0$. This reveals that the stationary point $(\widehat{\vartheta}_1, \widehat{\vartheta}_2) = (0,0)$ is a saddle point.

□

Let us emphasize that a stationary point may correspond to a local minimum point, a local maximum point, a saddle point, or a point of some other exotic category [12, pp. 217, 218]. Some categories of stationary points may be recognized by inspecting the Hermitian matrix

$$\mathbf{J}_2(\boldsymbol{\vartheta}) \triangleq \frac{\partial}{\partial \boldsymbol{\vartheta}} \left[\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \right]^T = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^T \tag{2.108}$$

for real $\boldsymbol{\theta}$, or the Hermitian matrix

$$\mathbf{J}_2(\boldsymbol{\vartheta}) \triangleq \frac{\partial}{\partial \boldsymbol{\vartheta}^*} \left[\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*} \right]^H = \begin{pmatrix} \dfrac{\partial}{\partial \boldsymbol{\theta}^*} \left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right]^H & \dfrac{\partial}{\partial \boldsymbol{\theta}^*} \left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^H \\ \dfrac{\partial}{\partial \boldsymbol{\theta}} \left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right]^H & \dfrac{\partial}{\partial \boldsymbol{\theta}} \left[\dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^H \end{pmatrix} \tag{2.109}$$

for complex $\boldsymbol{\theta}$, where the matrix $\mathbf{J}_2(\boldsymbol{\vartheta})$ is referred to as the *Hessian matrix* of $J(\boldsymbol{\vartheta})$. In particular, the local minimum points and local maximum points can be recognized by virtue of the Hessian matrix, as stated in the following theorem.

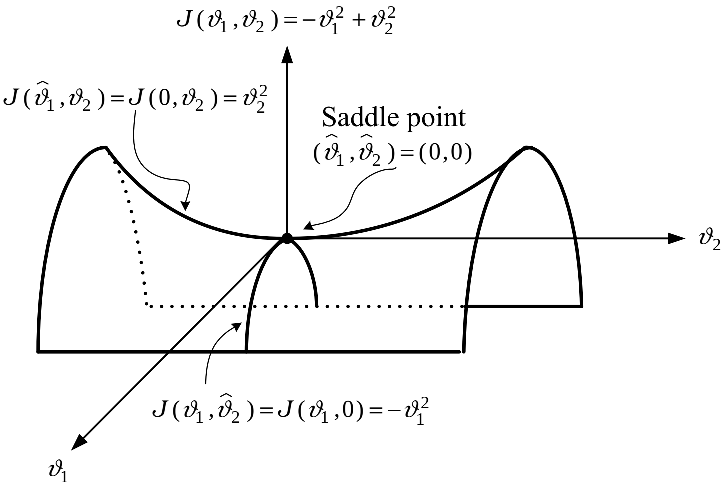
$J(\vartheta_1,\vartheta_2) = -\vartheta_1^2 + \vartheta_2^2$
$J(\widehat{\vartheta}_1,\vartheta_2) = J(0,\vartheta_2) = \vartheta_2^2$
Saddle point
$(\widehat{\vartheta}_1,\widehat{\vartheta}_2) = (0,0)$
ϑ_2
$J(\vartheta_1,\widehat{\vartheta}_2) = J(\vartheta_1,0) = -\vartheta_1^2$
ϑ_1

Fig. 2.7 Illustration of saddle point

Theorem 2.44 (Sufficient Conditions). *Suppose* $\widehat{\boldsymbol{\vartheta}}$ *is a stationary point of the objective function* $J(\boldsymbol{\vartheta})$*. If the Hessian matrix*

$$\mathbf{J}_2(\widehat{\boldsymbol{\vartheta}}) \triangleq \mathbf{J}_2(\boldsymbol{\vartheta})\Big|_{\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}} \tag{2.110}$$

is positive definite (negative definite), then $\widehat{\boldsymbol{\vartheta}}$ *corresponds to a local minimum (a local maximum) of* $J(\boldsymbol{\vartheta})$*.*

This theorem can be proved by virtue of the following *Taylor series* for $J(\boldsymbol{\vartheta})$ at $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$: (refer to [26, p. 71] for the real case)

$$J(\boldsymbol{\vartheta}) = J(\widehat{\boldsymbol{\vartheta}}) + (\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}})^H \boldsymbol{\nabla} J(\widehat{\boldsymbol{\vartheta}}) + \frac{1}{2}(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}})^H \mathbf{J}_2(\widehat{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\vartheta}}) + \cdots \tag{2.111}$$

We leave the proof of this theorem as an exercise (Problem 2.20).

2.3.3 Gradient-Type Optimization Methods

There are numerous types of optimization techniques available for solving the unconstrained optimization problem, among which we are interested in *gradient-type methods* for their efficiency as well as their wide scope of applications. Without loss of generality, we will introduce gradient-type methods in terms of the minimization problem of $J(\boldsymbol{\vartheta})$ because maximization of $J(\boldsymbol{\vartheta})$ is equivalent to minimization of $-J(\boldsymbol{\vartheta})$.

Iterative Procedure of Gradient-Type Methods

Let $\widehat{\boldsymbol{\vartheta}}$ denote a (local) minimum point of $J(\boldsymbol{\vartheta})$. Gradient-type methods are, in general, based on the following iterative procedure for searching for $\widehat{\boldsymbol{\vartheta}}$.

(S1) Set the iteration number $i = 0$.

(S2) Choose an appropriate initial condition $\boldsymbol{\vartheta}^{[0]}$ for $\widehat{\boldsymbol{\vartheta}}$ and an appropriate initial search direction $\mathbf{d}^{[0]}$.

(S3) Generate a new approximation to $\widehat{\boldsymbol{\vartheta}}$ via

$$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{d}^{[i]} \tag{2.112}$$

where $\mu^{[i]} > 0$ is the step size which should be determined appropriately to make sure of the movement along the direction of a (local) minimum of $J(\boldsymbol{\vartheta})$.

(S4) Check the convergence of the procedure. If the procedure has not yet converged, then go to Step (S5); otherwise, obtain a (local) minimum point as $\widehat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^{[i+1]}$ and stop the procedure.

(S5) Find a new search direction $\mathbf{d}^{[i+1]}$ which points towards a (local) minimum of $J(\boldsymbol{\vartheta})$ in general.

(S6) Update the iteration number i by $(i + 1)$ and go to Step (S3).

This procedure is also depicted in Fig. 2.8 for clarity.

In Step (S3) of the iterative procedure, determination of the step size $\mu^{[i]}$ can be formulated into the problem of finding the parameter μ which minimizes the objective function $f(\mu) \triangleq J(\boldsymbol{\vartheta}^{[i]} - \mu\mathbf{d}^{[i]})$ (by (2.112)). Accordingly, this problem can be solved by using the class of *one-dimensional (1-D) minimization methods* such as the *1-D Newton method* (also known as the *Newton–Raphson method*), the *1-D quasi-Newton method*, and so on [26]. Alternatively, the step size $\mu^{[i]}$ can be simply chosen as the value of $\mu_0/2^k$ for a preassigned positive real number μ_0 and a certain (positive or negative) integer k such that $J(\boldsymbol{\vartheta}^{[i]} - (\mu_0/2^k)\mathbf{d}^{[i]}) < J(\boldsymbol{\vartheta}^{[i]})$. In Step (S4), the convergence criterion

$$\left|\frac{J(\boldsymbol{\vartheta}^{[i]}) - J(\boldsymbol{\vartheta}^{[i+1]})}{J(\boldsymbol{\vartheta}^{[i]})}\right| \le \zeta \tag{2.113}$$

can be used for testing the convergence of the iterative procedure where ζ is a small positive constant. Of course, other types of convergence criteria can also be applied. In Step (S5), the way of finding a new search direction $\mathbf{d}^{[i+1]}$ determines substantially the efficiency of gradient-type methods and thus leads to the main differences between the existing gradient-type methods. As indicated by the name "gradient-type method," the update of $\mathbf{d}^{[i+1]}$ involves the gradient $\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i+1]})$ and in some cases the Hessian matrix $\mathbf{J}_2(\boldsymbol{\vartheta}^{[i+1]})$. Note that the gradient-type methods that require only the gradient are referred to as *first-order methods*, while those requiring both the gradient and the

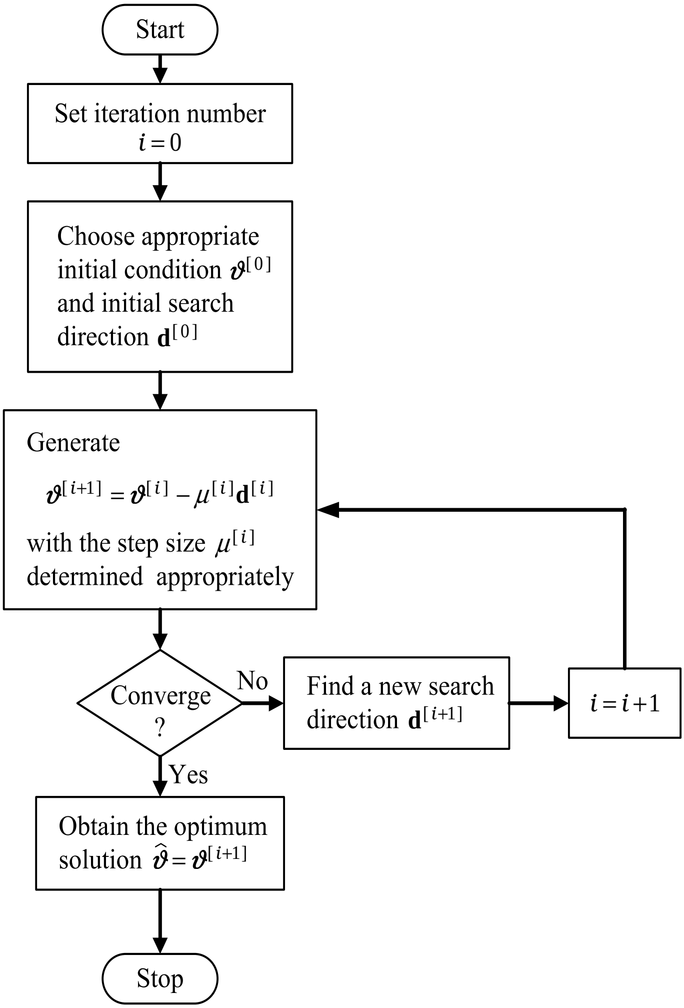
Start
Set iteration number $i=0$
Choose appropriate initial condition $\boldsymbol{\vartheta}^{[0]}$ and initial search direction $\mathbf{d}^{[0]}$
Generate $\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{d}^{[i]}$ with the step size $\mu^{[i]}$ determined appropriately
Converge ?
No
Find a new search direction $\mathbf{d}^{[i+1]}$
$i=i+1$
Yes
Obtain the optimum solution $\hat{\boldsymbol{\vartheta}} = \boldsymbol{\vartheta}^{[i+1]}$
Stop

Fig. 2.8 Flow chart for the iterative procedure of gradient-type methods

Hessian matrix are referred to as *second-order methods*. As a final remark, all gradient-type methods are only guaranteed to find local minimum solutions due to the local property of the gradient nature.

Overview of Existing Gradient-Type Methods

Among the existing gradient-type methods for minimization of $J(\boldsymbol{\vartheta})$, the simplest is the so-called *steepest descent method*, which belongs to the category of first-order methods and is extremely important from a theoretical viewpoint. Convergence of the steepest descent method is more or less insensitive to the initial condition $\boldsymbol{\vartheta}^{[0]}$, but the convergence rate is excessively slow in the vicinity of minimum solution points [31, p. 91], thereby limiting its application scope. On the other hand, a well-known second-order method, the *Newton method*, exhibits a rather fast convergence rate in the vicinity of minimum solution points. The Newton method, however, requires the initial condition $\boldsymbol{\vartheta}^{[0]}$ to be sufficiently close to any one of the minimum solution points for convergence, and also requires the inverse Hessian matrix $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})$, whose computational complexity is in general quite high. To overcome the initial-condition problem of the Newton method, the *Marquardt method*, a combination of the steepest descent method and the Newton method, tries to share the merits of both methods. It performs as the steepest descent method at first and then performs as the Newton method when a minimum solution point is approached. Obviously, like the Newton method, the Marquardt method is a second-order method and also suffers from the problem of high computational complexity.

The motivation for reducing the computational complexity of Newton method further leads to the family of *quasi-Newton methods*. The idea behind quasi-Newton methods is to approximate either the Hessian matrix $\mathbf{J}_2(\boldsymbol{\vartheta})$ or its inverse $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})$ in terms of the gradient $\boldsymbol{\nabla}J(\boldsymbol{\vartheta})$. Clearly, quasi-Newton methods also belong to the category of first-order methods. A representative which approximates $\mathbf{J}_2(\boldsymbol{\vartheta})$ iteratively is the *Broyden–Fletcher–Goldfarb–Shanno (BFGS) method*, while a representative which approximates $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})$ iteratively is the *Davidon–Fletcher–Powell (DFP) method*. Known as the best quasi-Newton method, the BFGS method performs initially as the steepest descent method and then (after a number of iterations) performs as the Newton method. Our experience of computer simulation shows that the BFGS method is very efficient and numerically stable, and thus has been used for the simulation examples in this book. Next, let us give the detailed descriptions of some selected gradient-type methods, namely, the steepest descent method, the Newton method and the BFGS method.

Steepest Descent Method

At iteration i, the steepest descent method[11] updates the parameter vector $\boldsymbol{\vartheta}$ via

$$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}), \tag{2.114}$$

i.e. the search direction $\mathbf{d}^{[i]} = \boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})$ (see (2.112)). The operation of (2.114) and the physical meaning of the gradient $\boldsymbol{\nabla}J(\boldsymbol{\vartheta})$ are interpreted as follows.

Let $\boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}$ be a neighboring point of $\boldsymbol{\vartheta}$ and $\Delta J(\boldsymbol{\vartheta}) = J(\boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}) - J(\boldsymbol{\vartheta})$ be the change in $J(\boldsymbol{\vartheta})$ due to $\Delta\boldsymbol{\vartheta}$. Then, by (2.111), we have

[11] The steepest descent method is also called the *Cauchy method* in recognition of the originator A. L. Cauchy (1847) [26].

$$\Delta J(\boldsymbol{\vartheta}) = (\Delta \boldsymbol{\vartheta})^H \, \boldsymbol{\nabla} J(\boldsymbol{\vartheta}) \quad \text{as } \Delta \boldsymbol{\vartheta} \to \mathbf{0} \tag{2.115}$$

where we have ignored the second-order and other higher-order (≥ 3) terms. From (2.115) and the Cauchy–Schwartz inequality (Theorem 2.1), it follows that

$$|\Delta J(\boldsymbol{\vartheta})| \leq \|\Delta \boldsymbol{\vartheta}\| \cdot \|\boldsymbol{\nabla} J(\boldsymbol{\vartheta})\| \quad \text{as } \Delta \boldsymbol{\vartheta} \to \mathbf{0} \tag{2.116}$$

and the equality holds only when $\Delta \boldsymbol{\vartheta} = \alpha \boldsymbol{\nabla} J(\boldsymbol{\vartheta})$ where α is a real or complex scalar. This reveals that the change rate of $J(\boldsymbol{\vartheta})$ defined as

$$\lim_{\Delta \boldsymbol{\vartheta} \to \mathbf{0}} \frac{|\Delta J(\boldsymbol{\vartheta})|}{\|\Delta \boldsymbol{\vartheta}\|} \tag{2.117}$$

is upper bounded by $\|\boldsymbol{\nabla} J(\boldsymbol{\vartheta})\|$, and that the gradient $\boldsymbol{\nabla} J(\boldsymbol{\vartheta})$ represents the direction giving the maximum change rate of $J(\boldsymbol{\vartheta})$. Moreover, when $\Delta \boldsymbol{\vartheta} = -\mu \boldsymbol{\nabla} J(\boldsymbol{\vartheta})$ for any real positive scalar μ, (2.115) reduces to

$$\Delta J(\boldsymbol{\vartheta}) = -\mu \|\boldsymbol{\nabla} J(\boldsymbol{\vartheta})\|^2 \leq 0 \tag{2.118}$$

and thus

$$J(\boldsymbol{\vartheta} - \mu \boldsymbol{\nabla} J(\boldsymbol{\vartheta})) = J(\boldsymbol{\vartheta} + \Delta \boldsymbol{\vartheta}) = J(\boldsymbol{\vartheta}) + \Delta J(\boldsymbol{\vartheta}) \leq J(\boldsymbol{\vartheta}), \tag{2.119}$$

which accounts for the operation of the update equation (2.114).

As a consequence of the preceding discussions, we come up with the following theorem to explain the physical meaning of the gradient $\boldsymbol{\nabla} J(\boldsymbol{\vartheta})$.

Theorem 2.45. *The negative of the gradient, $-\boldsymbol{\nabla} J(\boldsymbol{\vartheta})$, represents the direction giving the maximum change rate in reducing $J(\boldsymbol{\vartheta})$, i.e. the direction of steepest descent.*

Although the steepest descent method takes advantage of the gradient, the direction of steepest descent is only a local property (since $\Delta \boldsymbol{\vartheta} \to \mathbf{0}$) and thereby may vary from point to point. In fact, the steepest descent method quite often "zigzags" toward a local minimum, thereby requiring more and more steps of a smaller and smaller size when the minimum is approached [31, p. 91]. As such, it usually takes an enormous number of iterations to obtain an accurate solution.

Regarding the implementation of the update equation (2.114), it follows, from (2.105) and (2.106), that the update equation can be written as

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.120}$$

for real $\boldsymbol{\theta}$, and

$$\begin{pmatrix} \boldsymbol{\theta}^{[i+1]} \\ \boldsymbol{\theta}^{*[i+1]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}^{[i]} \\ \boldsymbol{\theta}^{*[i]} \end{pmatrix} - \mu^{[i]} \cdot \left. \begin{pmatrix} \dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \\ \dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{pmatrix} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.121}$$

for complex $\boldsymbol{\theta}$. One can easily see, from (2.121), that the update equation for $\boldsymbol{\theta}^{[i+1]}$ is equivalent to that for $\boldsymbol{\theta}^{*[i+1]}$ since $\mu^{[i]}$ is real, and thus only the former is actually needed. Table 2.2 summarizes the steepest descent method.

Table 2.2 Steepest descent method

	Update Equation
Generic form	At iteration i, update the parameter vector $\boldsymbol{\vartheta}$ via $$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]} \boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$$ where $\mu^{[i]} > 0$ is the step size and $\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$ is the gradient at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{[i]}$.
Real case	At iteration i, update the real parameter vector $\boldsymbol{\theta}$ via $$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} .$$
Complex case	At iteration i, update the complex parameter vector $\boldsymbol{\theta}$ via $$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} .$$

Newton Method

Suppose that $\boldsymbol{\vartheta}_0$ is a guess for the parameter vector $\boldsymbol{\vartheta}$ and the Hessian matrix $\mathbf{J}_2(\boldsymbol{\vartheta}_0)$ is nonsingular. Replacing $\widehat{\boldsymbol{\vartheta}}$ in (2.111) by $\boldsymbol{\vartheta}_0$ and taking the first derivative of (2.111) with respect to $\boldsymbol{\vartheta}^*$ yields

$$\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}^*} = \boldsymbol{\nabla} J(\boldsymbol{\vartheta}_0) + \alpha \mathbf{J}_2(\boldsymbol{\vartheta}_0)(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) \tag{2.122}$$

where all the higher-order terms (order ≥ 3) have been ignored and

$$\alpha = \begin{cases} 1 & \text{for real } \boldsymbol{\theta}, \\ 1/2 & \text{for complex } \boldsymbol{\theta}. \end{cases} \tag{2.123}$$

Setting (2.122) to zero, we obtain

$$\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 - \frac{1}{\alpha}\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}_0)\boldsymbol{\nabla}J(\boldsymbol{\vartheta}_0), \tag{2.124}$$

which reveals that $\boldsymbol{\vartheta}$ can be obtained from $\boldsymbol{\vartheta}_0$. However, since the higher-order terms that we have ignored may induce some errors in (2.124), it is suggested that (2.124) be used iteratively as follows: [26, pp. 389–391]

$$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i]})\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}) \tag{2.125}$$

where $\boldsymbol{\vartheta}^{[i]}$ denotes the parameter vector $\boldsymbol{\vartheta}$ obtained at iteration i and $\mu^{[i]} > 0$ is the step size included to avoid divergence. As a result, the search direction for the Newton method is $\mathbf{d}^{[i]} = \mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i]})\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})$.

To further analyze the Newton method, let $\boldsymbol{\vartheta}^{[i]} = \boldsymbol{\vartheta}$, $\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}$ and $\mu^{[i]} = \mu$ in (2.125). Then we have

$$\Delta\boldsymbol{\vartheta} = -\mu\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})\boldsymbol{\nabla}J(\boldsymbol{\vartheta}), \quad \mu > 0. \tag{2.126}$$

Once again, by using (2.111), we have

$$J(\boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}) = J(\boldsymbol{\vartheta}) + (\Delta\boldsymbol{\vartheta})^H \boldsymbol{\nabla}J(\boldsymbol{\vartheta}) + \frac{1}{2}(\Delta\boldsymbol{\vartheta})^H \mathbf{J}_2(\boldsymbol{\vartheta})\Delta\boldsymbol{\vartheta} + \cdots \tag{2.127}$$

where the higher-order (≥ 3) terms can be neglected as $\Delta\boldsymbol{\vartheta} \to \mathbf{0}$; this, in turn, requires that the step size μ be sufficiently small according to (2.126). From (2.126) and (2.127), it follows that the change $\Delta J(\boldsymbol{\vartheta}) \triangleq J(\boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}) - J(\boldsymbol{\vartheta})$ can be written as

$$\Delta J(\boldsymbol{\vartheta}) = -\mu(1 - \frac{\mu}{2})\left[\boldsymbol{\nabla}J(\boldsymbol{\vartheta})\right]^H \mathbf{J}_2^{-1}(\boldsymbol{\vartheta})\boldsymbol{\nabla}J(\boldsymbol{\vartheta}) \quad \text{as } \Delta\boldsymbol{\vartheta} \to \mathbf{0}. \tag{2.128}$$

Accordingly, if $\mathbf{J}_2(\boldsymbol{\vartheta})$ is positive definite and $\mu < 2$, then the change $\Delta J(\boldsymbol{\vartheta}) \leq 0$ and

$$J(\boldsymbol{\vartheta}^{[i+1]}) = J(\boldsymbol{\vartheta} + \Delta\boldsymbol{\vartheta}) = J(\boldsymbol{\vartheta}) + \Delta J(\boldsymbol{\vartheta}) \leq J(\boldsymbol{\vartheta}) = J(\boldsymbol{\vartheta}^{[i]}). \tag{2.129}$$

That is, the search direction always points towards a (local) minimum of $J(\boldsymbol{\vartheta})$ when the Hessian matrix $\mathbf{J}_2(\boldsymbol{\vartheta})$ is positive definite, or equivalently $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})$ is positive definite (by Property 2.12), and the step size μ is chosen small enough. However, due to utilization of only the lower-order terms of the Taylor series in the derivation, the Newton method requires the initial condition $\boldsymbol{\vartheta}^{[0]}$ to be sufficiently close to the solution point. Moreover, it is generally difficult and sometimes almost impossible to compute $\mathbf{J}_2(\boldsymbol{\vartheta})$ as well as $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta})$.

Regarding the implementation of the update equation (2.125), we note, from (2.105) and (2.106), that for real $\boldsymbol{\theta}$ the update equation is given by

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \mathbf{J}_2^{-1}(\boldsymbol{\theta}^{[i]}) \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.130}$$

where $\mathbf{J}_2(\boldsymbol{\theta}^{[i]}) \equiv \mathbf{J}_2(\boldsymbol{\vartheta}^{[i]})$, and for complex $\boldsymbol{\theta}$ it is given by

$$\begin{pmatrix} \boldsymbol{\theta}^{[i+1]} \\ \boldsymbol{\theta}^{*[i+1]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\theta}^{[i]} \\ \boldsymbol{\theta}^{*[i]} \end{pmatrix} - \mu^{[i]} \begin{pmatrix} \mathbf{A}^{[i]} & \mathbf{B}^{[i]} \\ \left(\mathbf{B}^{[i]}\right)^* & \left(\mathbf{A}^{[i]}\right)^* \end{pmatrix}^{-1} \cdot \left. \begin{pmatrix} \dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \\ \dfrac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{pmatrix} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.131}$$

where

$$\mathbf{A}^{[i]} = \left(\mathbf{A}^{[i]}\right)^H = \left. \frac{\partial}{\partial \boldsymbol{\theta}^*} \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right]^H \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}, \tag{2.132}$$

$$\mathbf{B}^{[i]} = \left(\mathbf{B}^{[i]}\right)^T = \left. \frac{\partial}{\partial \boldsymbol{\theta}^*} \left[\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^H \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}. \tag{2.133}$$

Similar to the complex case of the steepest descent method, by (2.131), (2.132), (2.133) and Theorem 2.8, one can show that only the following update equation is needed for complex $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \mathbf{C}^{[i]} \left\{ \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} - \mathbf{D}^{[i]} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \right\} \tag{2.134}$$

where

$$\mathbf{C}^{[i]} = \left\{ \mathbf{A}^{[i]} - \mathbf{B}^{[i]} \left[\left(\mathbf{A}^{[i]}\right)^* \right]^{-1} \left(\mathbf{B}^{[i]}\right)^* \right\}^{-1}, \tag{2.135}$$

$$\mathbf{D}^{[i]} = \mathbf{B}^{[i]} \left[\left(\mathbf{A}^{[i]}\right)^* \right]^{-1}. \tag{2.136}$$

Furthermore, one can simplify the update equation (2.134) by forcing $\mathbf{B}^{[i]} = \mathbf{0}$ for all iterations, and obtain the following "approximate" update equation for complex $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \left(\mathbf{A}^{[i]}\right)^{-1} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}. \tag{2.137}$$

We refer to the Newton method based on (2.137) as the *approximate Newton method*. Note that for the approximate Newton method, if the matrix $\mathbf{A}^{[i]}$ is positive definite, then the corresponding Hessian matrix approximated as

$$\mathbf{J}_2(\boldsymbol{\vartheta}^{[i]}) \approx \begin{pmatrix} \mathbf{A}^{[i]} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{A}^{[i]}\right)^* \end{pmatrix} \tag{2.138}$$

is positive definite, too. Accordingly, the above-mentioned interpretation for the operation of Newton method (see explanation of (2.129)) also applies to the approximate Newton method. Table 2.3 summarizes the Newton method and the approximate Newton method. Note that the approximate Newton method exists only for the complex case.

Table 2.3 Newton and approximate Newton methods

	Update Equation for the Newton Method
Generic form	At iteration i, update the parameter vector $\boldsymbol{\vartheta}$ via $\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i]})\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$ where $\mu^{[i]} > 0$ is the step size, $\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$ is the gradient at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{[i]}$, and $\mathbf{J}_2(\boldsymbol{\vartheta}^{[i]})$ is the Hessian matrix at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{[i]}$.
Real case	At iteration i, update the real parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \cdot \mathbf{J}_2^{-1}(\boldsymbol{\theta}^{[i]}) \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}$ where $\mathbf{J}_2(\boldsymbol{\theta}^{[i]}) = \mathbf{J}_2(\boldsymbol{\vartheta}^{[i]})$.
Complex case	At iteration i, update the complex parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]}\mathbf{C}^{[i]} \left\{ \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} - \mathbf{D}^{[i]} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \right\}$ where $\mathbf{C}^{[i]}$ and $\mathbf{D}^{[i]}$ are given by (2.135) and (2.136), respectively.
	Update Equation for the Approximate Newton Method
Complex case	At iteration i, update the complex parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \left(\mathbf{A}^{[i]}\right)^{-1} \cdot \left. \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} \right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}$ where $\mathbf{A}^{[i]}$ is given by (2.132).

Broyden–Fletcher–Goldfarb–Shanno Method

Recall that the idea behind the BFGS method is to approximate the inverse Hessian matrix $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i]})$ in (2.125) by virtue of the gradient $\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})$. Let $\mathbf{Q}^{[i]}$ be a Hermitian matrix, which will be obtained as an approximation to $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i]})$. Then, from (2.125), it follows that the update equation for the BFGS method is given by

$$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{Q}^{[i]}\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}), \tag{2.139}$$

i.e. the search direction $\mathbf{d}^{[i]} = \mathbf{Q}^{[i]}\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})$. Next, let us present how to update $\mathbf{Q}^{[i+1]}$ from $\mathbf{Q}^{[i]}$, as well as how to choose an appropriate initial condition for $\mathbf{Q}^{[0]}$.

Update Equation for $\mathbf{Q}^{[i+1]}$

Let $\mathbf{P}^{[i]} = \left(\mathbf{Q}^{[i]}\right)^{-1}$, that is, $\mathbf{P}^{[i]}$ (a Hermitian matrix) is an approximation to $\mathbf{J}_2(\boldsymbol{\vartheta}^{[i]})$. We will first derive the update equation for $\mathbf{P}^{[i+1]}$ and then convert it to the one for $\mathbf{Q}^{[i+1]}$. By substituting $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{[i]}$ and $\boldsymbol{\vartheta}_0 = \boldsymbol{\vartheta}^{[i+1]}$ into (2.122), we obtain

$$\mathbf{s}_{i+1} = \alpha\mathbf{J}_2(\boldsymbol{\vartheta}^{[i+1]})\mathbf{r}_{i+1} \tag{2.140}$$

where α is given by (2.123) and

$$\mathbf{r}_{i+1} = \boldsymbol{\vartheta}^{[i+1]} - \boldsymbol{\vartheta}^{[i]}, \tag{2.141}$$

$$\mathbf{s}_{i+1} = \boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i+1]}) - \boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}). \tag{2.142}$$

It follows that $\mathbf{P}^{[i+1]}$ should also satisfy (2.140) as follows:

$$\mathbf{s}_{i+1} = \alpha\mathbf{P}^{[i+1]}\mathbf{r}_{i+1}. \tag{2.143}$$

We note, from (2.100), (2.106), (2.141) and (2.142), that $\mathbf{r}_{i+1}$ and $\mathbf{s}_{i+1}$ are both $L \times 1$ vectors for real $\boldsymbol{\theta}$ and $(2L) \times 1$ vectors for complex $\boldsymbol{\theta}$. Also note, from (2.108) and (2.109), that $\mathbf{P}^{[i+1]}$ is an $L \times L$ symmetric matrix for real $\boldsymbol{\theta}$ and a $(2L) \times (2L)$ Hermitian matrix for complex $\boldsymbol{\theta}$. Therefore, the number of unknowns (to be determined) in $\mathbf{P}^{[i+1]}$ is more than the number of linear equations in (2.143), meaning that the solution satisfying (2.143) is not unique.

The general formula for updating $\mathbf{P}^{[i+1]}$ iteratively can be written as

$$\mathbf{P}^{[i+1]} = \mathbf{P}^{[i]} + \Delta\mathbf{P}^{[i]} \tag{2.144}$$

where, in theory, the matrix $\Delta\mathbf{P}^{[i]}$ can have rank as high as L for real $\boldsymbol{\theta}$ and $2L$ for complex $\boldsymbol{\theta}$, but rank 1 or rank 2 are more suitable in practice. By adopting the rank 2 update $\Delta\mathbf{P}^{[i]} = c_1\mathbf{z}_1\mathbf{z}_1^H + c_2\mathbf{z}_2\mathbf{z}_2^H$ (see [26, p. 398] for the real case), we have

$$\mathbf{P}^{[i+1]} = \mathbf{P}^{[i]} + c_1 \mathbf{z}_1 \mathbf{z}_1^H + c_2 \mathbf{z}_2 \mathbf{z}_2^H \tag{2.145}$$

where c_1 and c_2 are real or complex constants, and $\mathbf{z}_1$ and $\mathbf{z}_2$ are real or complex vectors to be determined. Substituting (2.145) into (2.143) yields

$$\mathbf{s}_{i+1} = \alpha \mathbf{P}^{[i]} \mathbf{r}_{i+1} + \alpha c_1 (\mathbf{z}_1^H \mathbf{r}_{i+1}) \mathbf{z}_1 + \alpha c_2 (\mathbf{z}_2^H \mathbf{r}_{i+1}) \mathbf{z}_2. \tag{2.146}$$

Equation (2.146) can be satisfied by choosing

$$\alpha c_1 (\mathbf{z}_1^H \mathbf{r}_{i+1}) \mathbf{z}_1 = \mathbf{s}_{i+1} \quad \text{and} \quad c_2 (\mathbf{z}_2^H \mathbf{r}_{i+1}) \mathbf{z}_2 = -\mathbf{P}^{[i]} \mathbf{r}_{i+1}, \tag{2.147}$$

which further leads to the following choice:

$$\mathbf{z}_1 = \mathbf{s}_{i+1} \tag{2.148}$$

$$\mathbf{z}_2 = \mathbf{P}^{[i]} \mathbf{r}_{i+1} \tag{2.149}$$

$$c_1 = \frac{1}{\alpha \mathbf{z}_1^H \mathbf{r}_{i+1}} = \frac{1}{\alpha \mathbf{s}_{i+1}^H \mathbf{r}_{i+1}} \tag{2.150}$$

$$c_2 = -\frac{1}{\mathbf{z}_2^H \mathbf{r}_{i+1}} = -\frac{1}{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{r}_{i+1}}. \tag{2.151}$$

Substituting (2.148) through (2.151) into (2.145) gives rise to the following update equation for $\mathbf{P}^{[i+1]}$:

$$\mathbf{P}^{[i+1]} = \mathbf{P}^{[i]} + \frac{\mathbf{s}_{i+1} \mathbf{s}_{i+1}^H}{\alpha \mathbf{s}_{i+1}^H \mathbf{r}_{i+1}} - \frac{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H}{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{r}_{i+1}}, \tag{2.152}$$

which is called the *Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula* (refer to [26] for the real case).

To convert the update equation (2.152) into the one for $\mathbf{Q}^{[i+1]}$, let us re-express (2.152) as

$$\mathbf{P}^{[i+1]} = \mathbf{R} - \frac{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H}{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{r}_{i+1}} \tag{2.153}$$

where

$$\mathbf{R} = \mathbf{P}^{[i]} + \frac{\mathbf{s}_{i+1} \mathbf{s}_{i+1}^H}{\alpha \mathbf{s}_{i+1}^H \mathbf{r}_{i+1}}. \tag{2.154}$$

Applying Woodbury's identity (Corollary 2.6) to (2.153) and (2.154) yields

$$\mathbf{Q}^{[i+1]} = \mathbf{R}^{-1} + \frac{\mathbf{R}^{-1} (\mathbf{P}^{[i]} \mathbf{r}_{i+1})(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{R}^{-1}}{(\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{r}_{i+1} - (\mathbf{P}^{[i]} \mathbf{r}_{i+1})^H \mathbf{R}^{-1} (\mathbf{P}^{[i]} \mathbf{r}_{i+1})} \tag{2.155}$$

and

$$\mathbf{R}^{-1} = \mathbf{Q}^{[i]} - \frac{\mathbf{Q}^{[i]}\mathbf{s}_{i+1}\mathbf{s}_{i+1}^H\mathbf{Q}^{[i]}}{\alpha\mathbf{s}_{i+1}^H\mathbf{r}_{i+1} + \mathbf{s}_{i+1}^H\mathbf{Q}^{[i]}\mathbf{s}_{i+1}}, \tag{2.156}$$

respectively. By substituting (2.156) into (2.155) and after some algebraic manipulations, we obtain

$$\mathbf{Q}^{[i+1]} = \mathbf{Q}^{[i]} + \frac{1}{\mathbf{r}_{i+1}^H\mathbf{s}_{i+1}}\left\{(\alpha+\beta_i)\,\mathbf{r}_{i+1}\mathbf{r}_{i+1}^H - \mathbf{r}_{i+1}\mathbf{s}_{i+1}^H\mathbf{Q}^{[i]} - \mathbf{Q}^{[i]}\mathbf{s}_{i+1}\mathbf{r}_{i+1}^H\right\} \tag{2.157}$$

where

$$\beta_i = \frac{\mathbf{s}_{i+1}^H\mathbf{Q}^{[i]}\mathbf{s}_{i+1}}{\mathbf{s}_{i+1}^H\mathbf{r}_{i+1}} \tag{2.158}$$

is a real number. In the derivation of (2.157), we have used the facts that $\mathbf{P}^{[i]} = (\mathbf{P}^{[i]})^H$ and that $\mathbf{r}_{i+1}^H\mathbf{s}_{i+1} = \mathbf{s}_{i+1}^H\mathbf{r}_{i+1}$ is real (by (2.141), (2.142), (2.100), and (2.106)).

As a consequence, the BFGS method employs the update equation (2.139) for $\boldsymbol{\vartheta}^{[i+1]}$ along with the update equation (2.157) for $\mathbf{Q}^{[i+1]}$ to obtain the (local) minimum solution $\boldsymbol{\vartheta}$ without involving any second partial derivatives of $J(\boldsymbol{\vartheta})$.

Suggestion for the Initial Condition $\mathbf{Q}^{[0]}$

Since $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i+1]})$ is required to be positive definite in the Newton method, the Hermitian matrix $\mathbf{Q}^{[i+1]}$, as an approximation to $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i+1]})$, should also maintain the positive definite property. The following theorem reveals the conditions for maintaining the positive definite property of $\mathbf{Q}^{[i+1]}$ (refer to [32] for the real case).

Theorem 2.46. *If the matrix* $\mathbf{Q}^{[i]}$ *is positive definite and the step size* $\mu^{[i]} > 0$ *used in (2.139) is optimum, then the matrix* $\mathbf{Q}^{[i+1]}$ *generated from (2.157) is also positive definite where* $\mathbf{r}_{i+1}$ *and* $\mathbf{s}_{i+1}$ *defined as (2.141) and (2.142) are both nonzero vectors before convergence.*

See Appendix 2F for the proof. Theorem 2.46 suggests that $\mathbf{Q}^{[0]}$ be chosen as a positive definite matrix, in addition to utilization of an appropriate step size $\mu^{[i]}$. Usually, $\mathbf{Q}^{[0]} = \mathbf{I}$ is used. As such, the BFGS method performs initially as the steepest descent method because (2.139) reduces to (2.114) when $\mathbf{Q}^{[i]} = \mathbf{I}$. After a number of iterations, it performs as the Newton method because $\mathbf{Q}^{[i+1]}$ then appears as a good approximation to $\mathbf{J}_2^{-1}(\boldsymbol{\vartheta}^{[i+1]})$. On the other hand, numerical experience indicates that the BFGS method is less influenced by the error in determining $\mu^{[i]}$ [26, p. 406]. Nevertheless, in case the positive definite property of $\mathbf{Q}^{[i+1]}$ is violated due to this error, one may periodically reset $\mathbf{Q}^{[i+1]}$ via $\mathbf{Q}^{[i+1]} = \mathbf{I}$. The corresponding BFGS method then reverts to the steepest descent method at iteration $(i+1)$, but this time it has a much better initial condition $\boldsymbol{\vartheta}^{[i+1]}$.

Implementation of the BFGS Method

For the case of real $\boldsymbol{\theta}$, the update equation (2.139) can be written as

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]}\mathbf{Q}^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.159}$$

and the update equation (2.157) for $\mathbf{Q}^{[i+1]}$ reduces to

$$\mathbf{Q}^{[i+1]} = \mathbf{Q}^{[i]} + \frac{(1+\beta_i)\,\mathbf{r}_{i+1}\mathbf{r}_{i+1}^T - \mathbf{r}_{i+1}\mathbf{s}_{i+1}^T\mathbf{Q}^{[i]} - \mathbf{Q}^{[i]}\mathbf{s}_{i+1}\mathbf{r}_{i+1}^T}{\mathbf{r}_{i+1}^T\mathbf{s}_{i+1}} \tag{2.160}$$

(since $\alpha = 1$) where the initial condition $\mathbf{Q}^{[0]} = \mathbf{I}$ is suggested and

$$\beta_i = \frac{\mathbf{s}_{i+1}^T\mathbf{Q}^{[i]}\mathbf{s}_{i+1}}{\mathbf{s}_{i+1}^T\mathbf{r}_{i+1}}, \tag{2.161}$$

$$\mathbf{r}_{i+1} = \boldsymbol{\theta}^{[i+1]} - \boldsymbol{\theta}^{[i]}, \tag{2.162}$$

$$\mathbf{s}_{i+1} = \left\{ \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i+1]}} \right\} - \left\{ \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \right\}. \tag{2.163}$$

On the other hand, for the case of complex $\boldsymbol{\theta}$, the vectors $\mathbf{r}_{i+1}$ and $\mathbf{s}_{i+1}$ defined as (2.141) and (2.142) can be written as

$$\mathbf{r}_{i+1} = \left(\widetilde{\mathbf{r}}_{i+1}^T, \widetilde{\mathbf{r}}_{i+1}^H\right)^T \quad \text{and} \quad \mathbf{s}_{i+1} = \left(\widetilde{\mathbf{s}}_{i+1}^T, \widetilde{\mathbf{s}}_{i+1}^H\right)^T \tag{2.164}$$

where

$$\widetilde{\mathbf{r}}_{i+1} = \boldsymbol{\theta}^{[i+1]} - \boldsymbol{\theta}^{[i]}, \tag{2.165}$$

$$\widetilde{\mathbf{s}}_{i+1} = \left\{ \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i+1]}} \right\} - \left\{ \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \right\}. \tag{2.166}$$

By (2.152), (2.164) and Theorem 2.8, one can show that if the initial condition $\mathbf{Q}^{[0]} = \mathbf{I}$ is used, then the matrix $\mathbf{Q}^{[i]}$ obtained from the update equation (2.157) will have the following form:

$$\mathbf{Q}^{[i]} = \begin{pmatrix} \mathbf{Q}_A^{[i]} & \mathbf{Q}_B^{[i]} \\ \left(\mathbf{Q}_B^{[i]}\right)^* & \left(\mathbf{Q}_A^{[i]}\right)^* \end{pmatrix} \tag{2.167}$$

where $\mathbf{Q}_A^{[i]} = \left(\mathbf{Q}_A^{[i]}\right)^H$ and $\mathbf{Q}_B^{[i]} = \left(\mathbf{Q}_B^{[i]}\right)^T$ since $\mathbf{Q}^{[i]}$ is Hermitian. From (2.139), (2.157), (2.164) and (2.167), it follows that we need only the following update equation for complex $\boldsymbol{\theta}^{[i+1]}$:

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \left\{ \mathbf{Q}_A^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} + \mathbf{Q}_B^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \right\} \tag{2.168}$$

and the following update equations for $\mathbf{Q}_A^{[i+1]}$ and $\mathbf{Q}_B^{[i+1]}$:

$$\begin{aligned} \mathbf{Q}_A^{[i+1]} = \mathbf{Q}_A^{[i]} + \frac{1}{2\mathrm{Re}\left\{\widetilde{\mathbf{r}}_{i+1}^H \widetilde{\mathbf{s}}_{i+1}\right\}} \Bigg\{ & (\alpha + \beta_i)\, \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^H - \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_A^{[i]} \\ & - \mathbf{Q}_A^{[i]} \widetilde{\mathbf{s}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^H - \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{s}}_{i+1}^T \left(\mathbf{Q}_B^{[i]}\right)^* - \mathbf{Q}_B^{[i]} \widetilde{\mathbf{s}}_{i+1}^* \widetilde{\mathbf{r}}_{i+1}^H \Bigg\}, \end{aligned} \tag{2.169}$$

$$\begin{aligned} \mathbf{Q}_B^{[i+1]} = \mathbf{Q}_B^{[i]} + \frac{1}{2\mathrm{Re}\left\{\widetilde{\mathbf{r}}_{i+1}^H \widetilde{\mathbf{s}}_{i+1}\right\}} \Bigg\{ & (\alpha + \beta_i)\, \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^T - \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{s}}_{i+1}^T \left(\mathbf{Q}_A^{[i]}\right)^* \\ & - \mathbf{Q}_A^{[i]} \widetilde{\mathbf{s}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^T - \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_B^{[i]} - \mathbf{Q}_B^{[i]} \widetilde{\mathbf{s}}_{i+1}^* \widetilde{\mathbf{r}}_{i+1}^T \Bigg\} \end{aligned} \tag{2.170}$$

where $\mathbf{Q}_A^{[0]} = \mathbf{I}$, $\mathbf{Q}_B^{[0]} = \mathbf{0}$, and

$$\beta_i = \frac{\mathrm{Re}\left\{\widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_A^{[i]} \widetilde{\mathbf{s}}_{i+1} + \widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_B^{[i]} \widetilde{\mathbf{s}}_{i+1}^*\right\}}{\mathrm{Re}\left\{\widetilde{\mathbf{s}}_{i+1}^H \widetilde{\mathbf{r}}_{i+1}\right\}}. \tag{2.171}$$

Furthermore, one can simplify the above update equations by forcing $\mathbf{Q}_B^{[i]} = \mathbf{0}$ for all iterations. The corresponding update equation for complex $\boldsymbol{\theta}^{[i+1]}$ is given by

$$\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]} \mathbf{Q}_A^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} \tag{2.172}$$

and the corresponding update equation for $\mathbf{Q}_A^{[i+1]}$ is given by

$$\begin{aligned} \mathbf{Q}_A^{[i+1]} = \mathbf{Q}_A^{[i]} + \frac{1}{2\mathrm{Re}\left\{\widetilde{\mathbf{r}}_{i+1}^H \widetilde{\mathbf{s}}_{i+1}\right\}} \Bigg\{ & (\alpha + \beta_i)\, \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^H \\ & - \widetilde{\mathbf{r}}_{i+1} \widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_A^{[i]} - \mathbf{Q}_A^{[i]} \widetilde{\mathbf{s}}_{i+1} \widetilde{\mathbf{r}}_{i+1}^H \Bigg\} \end{aligned} \tag{2.173}$$

where $\mathbf{Q}_A^{[0]} = \mathbf{I}$ and

$$\beta_i = \frac{\mathrm{Re}\left\{\widetilde{\mathbf{s}}_{i+1}^H \mathbf{Q}_A^{[i]} \widetilde{\mathbf{s}}_{i+1}\right\}}{\mathrm{Re}\left\{\widetilde{\mathbf{s}}_{i+1}^H \widetilde{\mathbf{r}}_{i+1}\right\}}. \tag{2.174}$$

Similarly, we refer to the BFGS method that is based on (2.172), (2.173) and (2.174) as the *approximate BFGS method*. As a result of the aforementioned discussions, the approximate BFGS method also maintains the positive definite property of the corresponding $\mathbf{Q}^{[i]}$, provided that the step size $\mu^{[i]}$ is chosen appropriately. Table 2.4 summarizes the BFGS method and the approximate BFGS method where the latter is only for the complex case.

2.4 Least-Squares Method

Many science and engineering problems require solving the following set of M linear equations in K unknowns:

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b} \tag{2.175}$$

where $\mathbf{A}$ is an $M \times K$ matrix, $\mathbf{b} = (b_1, b_2, ..., b_M)^T$ is an $M \times 1$ vector, and $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_K)^T$ is a $K \times 1$ vector of unknown parameters to be solved. Let $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K)$ where $\mathbf{a}_k$, $k = 1, 2, ..., K$, are the column vectors of $\mathbf{A}$. Then the set of linear equations (2.175) can be written as

$$\mathbf{b} = \sum_{k=1}^{K} \theta_k \mathbf{a}_k. \tag{2.176}$$

Usually, (2.175) has no exact solution because $\mathbf{b}$ is not ordinarily located in the column space of $\mathbf{A}$ [6, p. 221], i.e. $\mathbf{b}$ cannot be expressed as a linear combination of $\mathbf{a}_k$, $k = 1, 2, ..., K$, for any $\boldsymbol{\theta}$ (see (2.176)). The column space of $\mathbf{A}$ is often referred to as the *range space* of $\mathbf{A}$, whose dimension is equal to rank$\{\mathbf{A}\}$. On the other hand, when $\mathbf{b} = \mathbf{0}$ (i.e. $\mathbf{A}\boldsymbol{\theta} = \mathbf{0}$), the corresponding set of solutions spans another subspace, referred to as the *null space* of $\mathbf{A}$. The dimension of the nullspace of $\mathbf{A}$, called the *nullity* of $\mathbf{A}$, is equal to $K -$ rank$\{\mathbf{A}\}$.

In practical applications, however, an approximate solution to (2.175) is still desired. Hence, let us change the original problem into the following approximation problem:

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{b} - \boldsymbol{\varepsilon} \tag{2.177}$$

where

$$\boldsymbol{\varepsilon} = \mathbf{b} - \mathbf{A}\boldsymbol{\theta} = \mathbf{b} - \sum_{k=1}^{K} \theta_k \mathbf{a}_k \tag{2.178}$$

is the $M \times 1$ vector of approximation errors (equation errors). For the approximation problem, a widely used approach is to find $\boldsymbol{\theta}$ such that

$$\widehat{\mathbf{b}} = \mathbf{A}\boldsymbol{\theta} = \sum_{k=1}^{K} \theta_k \mathbf{a}_k \tag{2.179}$$

Table 2.4 BFGS and approximate BFGS methods

	Update Equation for the BFGS Method
Generic form	At iteration i, update the parameter vector $\boldsymbol{\vartheta}$ via $\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]}\mathbf{Q}^{[i]}\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$ and update the Hermitian matrix $\mathbf{Q}^{[i]}$ via (2.157) where $\mu^{[i]} > 0$ is the step size, $\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$ is the gradient at $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{[i]}$, and $\mathbf{Q}^{[0]} = \mathbf{I}$ is suggested. In the update equation (2.157), the parameters α and β_i are given by (2.123) and (2.158), respectively, and the vectors $\mathbf{r}_{i+1}$ and $\mathbf{s}_{i+1}$ are given by (2.141) and (2.142), respectively.
Real case	At iteration i, update the real parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]}\mathbf{Q}^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}$ and update the symmetric matrix $\mathbf{Q}^{[i]}$ via (2.160), in which β_i, $\mathbf{r}_{i+1}$ and $\mathbf{s}_{i+1}$ are given by (2.161), (2.162) and (2.163), respectively.
Complex case	At iteration i, update the complex parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]}\left\{\mathbf{Q}_A^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}} + \mathbf{Q}_B^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}\right\}$ and update the Hermitian matrix $\mathbf{Q}_A^{[i]}$ and the matrix $\mathbf{Q}_B^{[i]}$ via (2.169) and (2.170), in which $\mathbf{Q}_A^{[0]} = \mathbf{I}$, $\mathbf{Q}_B^{[0]} = \mathbf{0}$, and β_i, $\widetilde{\mathbf{r}}_{i+1}$ and $\widetilde{\mathbf{s}}_{i+1}$ are given by (2.171), (2.165) and (2.166), respectively.
	Update Equation for the Approximate BFGS Method
Complex case	At iteration i, update the complex parameter vector $\boldsymbol{\theta}$ via $\boldsymbol{\theta}^{[i+1]} = \boldsymbol{\theta}^{[i]} - \mu^{[i]}\mathbf{Q}_A^{[i]} \cdot \left.\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*}\right\|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{[i]}}$ and update the matrix $\mathbf{Q}_A^{[i]}$ via (2.173), in which $\mathbf{Q}_A^{[0]} = \mathbf{I}$ and β_i, $\widetilde{\mathbf{r}}_{i+1}$ and $\widetilde{\mathbf{s}}_{i+1}$ are given by (2.174), (2.165) and (2.166), respectively.

approximates $\mathbf{b}$ in the sense of minimizing the objective function

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \|\boldsymbol{\varepsilon}\|^2 = \sum_{m=1}^{M} |\varepsilon_m|^2 \tag{2.180}$$

where ε_m is the mth entry of $\boldsymbol{\varepsilon}$. The problem of minimizing the sum of squared errors given by (2.180) is called the *least-squares (LS) problem* and the corresponding solution is called the *least-squares (LS) solution.*

2.4.1 Full-Rank Overdetermined Least-Squares Problem

Consider that $M \geq K$ and $\mathbf{A}$ is of full rank, i.e. $\mathrm{rank}\{\mathbf{A}\} = K$. The LS solution is derived as follows. Taking the first derivative of $J_{\mathrm{LS}}(\boldsymbol{\theta})$ given by (2.180) with respect to $\boldsymbol{\theta}^*$ and setting the result to zero yields

$$\frac{\partial J_{\mathrm{LS}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^*} = -\mathbf{A}^H (\mathbf{b} - \mathbf{A}\boldsymbol{\theta}) = -\mathbf{A}^H \boldsymbol{\varepsilon} = \mathbf{0}, \tag{2.181}$$

which gives rise to

$$\mathbf{A}^H \mathbf{A}\boldsymbol{\theta} = \mathbf{A}^H \mathbf{b}. \tag{2.182}$$

From (2.181), it follows that

$$\mathbf{a}_k^H \boldsymbol{\varepsilon} = 0 \qquad \text{for } k = 1, 2, ..., K. \tag{2.183}$$

That is, the error vector $\boldsymbol{\varepsilon}$ is orthogonal to the column vectors $\mathbf{a}_k$, thereby leading to the name "*normal equations*" for the set of equations (2.182) [2]. As illustrated in Fig. 2.9 (by (2.178) and (2.179)), $\boldsymbol{\varepsilon}$ has the minimum norm only when it is orthogonal (perpendicular) to the range space of $\mathbf{A}$ (the plane). This observation indicates that the solution obtained from (2.182) corresponds to the global minimum of $J_{\mathrm{LS}}(\boldsymbol{\theta})$.

Since $\mathbf{A}$ is of full rank, $\mathbf{A}^H\mathbf{A}$ is a nonsingular $K \times K$ matrix and thus there is only a unique solution to (2.182) as follows:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \mathbf{b} \tag{2.184}$$

where $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ represents the LS solution for $\boldsymbol{\theta}$. Substituting (2.184) into (2.179) gives

$$\widehat{\mathbf{b}} = \mathbf{P}_A \mathbf{b} \tag{2.185}$$

where

$$\mathbf{P}_A = \mathbf{A}(\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \tag{2.186}$$

is an $M \times M$ matrix. From Fig. 2.9, one can observe that $\widehat{\mathbf{b}}$ corresponds to the projection of $\mathbf{b}$ onto the range space of $\mathbf{A}$. For this reason, $\mathbf{P}_A$ is called the *projection matrix* of $\mathbf{A}$. It has the following properties.

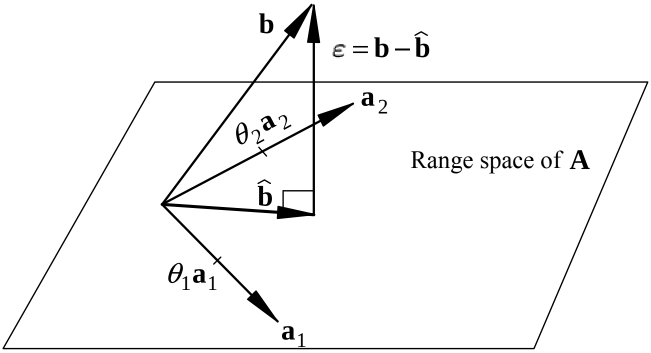

Fig. 2.9 Geometrical explanation of the LS method for $K = 2$

- Idempotent property: $\mathbf{P}_A\mathbf{P}_A = \mathbf{P}_A$.
- Hermitian property: $\mathbf{P}_A^H = \mathbf{P}_A$.

The idempotent property implies that $\mathbf{P}_A\widehat{\mathbf{b}} = \widehat{\mathbf{b}}$, i.e. the projection of $\widehat{\mathbf{b}}$ onto the range space is still $\widehat{\mathbf{b}}$. This can also be observed from Fig. 2.9 where $\widehat{\mathbf{b}}$ is already in the range space. When $M = K$, (2.184) reduces to

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{A}^{-1}\mathbf{b} \tag{2.187}$$

and the corresponding $\mathbf{P}_A = \mathbf{I}$. That is, there is no need for any projection because $\mathbf{b}$ is already in the range space for this case.

2.4.2 Generic Least-Squares Problem

Now consider the general case that M can be less than K and $\mathrm{rank}\{\mathbf{A}\} = r \leq \min\{M, K\}$, i.e. $\mathbf{A}$ can be rank deficient. The SVD of $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H = \mathbf{U}\begin{pmatrix}\boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}^H \tag{2.188}$$

where $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_M)$ and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K)$ are $M \times M$ and $K \times K$ unitary matrices, respectively, and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, ..., \lambda_r\}$. The vectors $\mathbf{u}_k$ and $\mathbf{v}_k$ are the left and right singular vectors of $\mathbf{A}$, respectively, and λ_1, λ_2, ..., λ_r are the real positive singular values. By (2.178), (2.180) and (2.188),

$$\begin{aligned} J_{\mathrm{LS}}(\boldsymbol{\theta}) &= \|\mathbf{b} - \mathbf{A}\boldsymbol{\theta}\|^2 = \|\mathbf{U}^H(\mathbf{b} - \mathbf{A}\boldsymbol{\theta})\|^2 \\ &= \|\mathbf{U}^H\mathbf{b} - \mathbf{U}^H\mathbf{A}\mathbf{V}\mathbf{V}^H\boldsymbol{\theta}\|^2 = \|\mathbf{U}^H\mathbf{b} - \boldsymbol{\Sigma}\widetilde{\boldsymbol{\theta}}\|^2 \end{aligned} \tag{2.189}$$

where

$$\widetilde{\boldsymbol{\theta}} = (\widetilde{\theta}_1, \widetilde{\theta}_2, ..., \widetilde{\theta}_K)^T \triangleq \mathbf{V}^H\boldsymbol{\theta}. \tag{2.190}$$

Equation (2.189) can be further expressed as follows:

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \sum_{k=1}^{r} |\mathbf{u}_k^H \mathbf{b} - \lambda_k \widetilde{\theta}_k|^2 + \sum_{k=r+1}^{M} |\mathbf{u}_k^H \mathbf{b}|^2. \tag{2.191}$$

Clearly, the minimum value

$$\min\{J_{\mathrm{LS}}(\boldsymbol{\theta})\} = \sum_{k=r+1}^{M} |\mathbf{u}_k^H \mathbf{b}|^2$$

is attained when the first r entries of $\widetilde{\boldsymbol{\theta}}$ satisfy

$$\widetilde{\theta}_k = \frac{\mathbf{u}_k^H \mathbf{b}}{\lambda_k} \quad \text{for } k = 1, 2, ..., r, \tag{2.192}$$

regardless of what the remaining entries $\widetilde{\theta}_k$, $k = r+1, r+2, ..., K$, are. This indicates that there are infinitely many solutions to the generic LS problem.

Among these solutions, the LS solution $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ is always chosen as the one with the minimum norm. It is therefore also referred to as the *minimum-norm solution*. Because $||\widetilde{\boldsymbol{\theta}}||^2 = \boldsymbol{\theta}^H \mathbf{V}\mathbf{V}^H \boldsymbol{\theta} = ||\boldsymbol{\theta}||^2$, the minimum-norm solution $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ corresponds to letting $\widetilde{\theta}_k = 0$ for $k = r+1, r+2, ..., K$. This, together with (2.190) and (2.192), therefore gives

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{V}\widetilde{\boldsymbol{\theta}} = \sum_{k=1}^{K} \mathbf{v}_k \widetilde{\theta}_k = \sum_{k=1}^{r} \mathbf{v}_k \left(\frac{\mathbf{u}_k^H \mathbf{b}}{\lambda_k} \right). \tag{2.193}$$

The solution given by (2.193) is also equivalent to the form

$$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{A}^+ \mathbf{b} \tag{2.194}$$

where

$$\mathbf{A}^+ = \sum_{k=1}^{r} \frac{1}{\lambda_k} \mathbf{v}_k \mathbf{u}_k^H = \mathbf{V}\boldsymbol{\Sigma}^+ \mathbf{U}^H \tag{2.195}$$

is a $K \times M$ matrix in which

$$\boldsymbol{\Sigma}^+ = \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{2.196}$$

is also a $K \times M$ matrix. The matrix $\mathbf{A}^+$ is called the *Moore–Penrose generalized inverse* or the *pseudoinverse* of $\mathbf{A}$. By substituting (2.194) into (2.179), we also obtain $\widehat{\mathbf{b}}$ as given by (2.185) with the generic form of projection matrix

$$\mathbf{P}_A = \mathbf{A}\mathbf{A}^+. \tag{2.197}$$

Table 2.5 gives a summary of the LS method. When $\mathbf{A}$ is of full rank and $M \geq K$ (i.e. the full-rank overdetermined LS problem), the pseudoinverse $\mathbf{A}^+ = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$ (Problem 2.21), and thus the LS solution given by (2.194) reduces to the one given by (2.184). Nevertheless, if computational complexity is not of major concern, it is preferred to use (2.194) to obtain the LS solution due to the better numerical properties of SVD. On the other hand, for the case of $\mathbf{A}$ having a special structure such as the Toeplitz structure, it may be better to use other algorithms that take advantage of the special structure to solve the LS problem.

Table 2.5 Least-squares (LS) method

Problem	Find a $K \times 1$ vector $\boldsymbol{\theta}$ to solve the set of linear equations $$\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$$ by minimizing the sum of squared errors $$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \lVert\boldsymbol{\varepsilon}\rVert^2$$ where $\mathbf{A}$ is an $M \times K$ matrix with the SVD $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H$ and $\boldsymbol{\varepsilon} = \mathbf{b} - \mathbf{A}\boldsymbol{\theta}$ is the error vector.
Generic solution	The (minimum-norm) LS solution $$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{A}^+\mathbf{b}$$ where $\mathbf{A}^+$ is the pseudoinverse of $\mathbf{A}$ given by $$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^H = \sum_{k=1}^{r} \frac{1}{\lambda_k}\mathbf{v}_k\mathbf{u}_k^H.$$
Special cases	(i) $M \geq K$ and $\mathrm{rank}\{\mathbf{A}\} = K$: $$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H\mathbf{b}$$ (ii) $M = K$ and $\mathrm{rank}\{\mathbf{A}\} = K$: $$\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{A}^{-1}\mathbf{b}$$

2.5 Summary

We have reviewed the definitions of vectors, vector spaces, matrices, and some special forms of matrices. Several useful formulas and properties of matrices as well as matrix decomposition including eigendecomposition and the SVD were described. The SVD was then applied to the derivation of a minimum-norm solution to the generic LS problem. Regarding the mathematical analysis, we have dealt with the convergence of sequences and series including the Fourier series, as well as sequence and function spaces. As for the optimization theory, we have introduced the necessary and sufficient conditions for solutions, carefully dealt with the first derivative of the objective function with respect to a complex vector, and provided an overview of gradient-type optimization methods. Three popular gradient-type methods were introduced in terms of a complex-valued framework since they are often applied to blind equalization problems. Vector differentiation was then applied to find the solution to the full-rank overdetermined LS problem.

Appendix 2A
Proof of Theorem 2.15

The theorem can be proved by either of the following two approaches.

Approach I: Derivation from the Matrix $\mathbf{A}^H\mathbf{A}$

According to Properties 2.13 and 2.11, the eigenvalues of the $K \times K$ matrix $\mathbf{A}^H\mathbf{A}$ are all real nonnegative. Therefore, let λ_1, λ_2, ..., λ_K be nonnegative real numbers, and λ_1^2, λ_2^2, ..., λ_K^2 be the eigenvalues of $\mathbf{A}^H\mathbf{A}$. Furthermore, let these eigenvalues be arranged in the following order:

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2 > 0$$

and

$$\lambda_{r+1}^2 = \lambda_{r+2}^2 = \cdots = \lambda_K^2 = 0 \tag{2.198}$$

where the second equation follows from the fact that $\text{rank}\{\mathbf{A}^H\mathbf{A}\} = \text{rank}\{\mathbf{A}\} = r$. Let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K$ be the orthonormal eigenvectors of $\mathbf{A}^H\mathbf{A}$ corresponding to the eigenvalues λ_1^2, λ_2^2, ..., λ_K^2, respectively. That is,

$$\mathbf{A}^H\mathbf{A}\mathbf{v}_i = \lambda_i^2\mathbf{v}_i, \quad i = 1, 2, ..., K \tag{2.199}$$

and

$$\mathbf{v}_i^H\mathbf{v}_j = \begin{cases} 1, & \text{for } i = j, \\ 0, & \text{for } i \neq j. \end{cases} \tag{2.200}$$

Let $\mathbf{V} = (\mathbf{V}_1 \ \mathbf{V}_2)$ be a $K \times K$ matrix where $\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r)$ is a $K \times r$ matrix and $\mathbf{V}_2 = (\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_K)$ is a $K \times (K - r)$ matrix. Then, from (2.200), it follows that

$$\mathbf{V}^H\mathbf{V} = \begin{pmatrix} \mathbf{V}_1^H \\ \mathbf{V}_2^H \end{pmatrix} (\mathbf{V}_1 \ \mathbf{V}_2) = \begin{pmatrix} \mathbf{V}_1^H\mathbf{V}_1 & \mathbf{V}_1^H\mathbf{V}_2 \\ \mathbf{V}_2^H\mathbf{V}_1 & \mathbf{V}_2^H\mathbf{V}_2 \end{pmatrix} = \mathbf{I},$$

i.e. $\mathbf{V}$ is a unitary matrix. By (2.198) and (2.199), we have

$$\begin{aligned} \mathbf{A}^H\mathbf{A}\mathbf{V}_2 &= (\mathbf{A}^H\mathbf{A}\mathbf{v}_{r+1}, \mathbf{A}^H\mathbf{A}\mathbf{v}_{r+2}, ..., \mathbf{A}^H\mathbf{A}\mathbf{v}_K) \\ &= (\lambda_{r+1}^2\mathbf{v}_{r+1}, \lambda_{r+2}^2\mathbf{v}_{r+2}, ..., \lambda_K^2\mathbf{v}_K) = \mathbf{0}, \end{aligned}$$

implying that

$$(\mathbf{A}\mathbf{V}_2)^H(\mathbf{A}\mathbf{V}_2) = \mathbf{V}_2^H(\mathbf{A}^H\mathbf{A}\mathbf{V}_2) = \mathbf{0}$$

or

$$\mathbf{A}\mathbf{V}_2 = \mathbf{0}. \tag{2.201}$$

In the same way, by (2.199), we have

$$\mathbf{A}^H\mathbf{A}\mathbf{V}_1 = (\lambda_1^2\mathbf{v}_1, \lambda_2^2\mathbf{v}_2, ..., \lambda_r^2\mathbf{v}_r) = \mathbf{V}_1\mathbf{\Lambda}^2 \tag{2.202}$$

where $\mathbf{\Lambda}^2 = \text{diag}\{\lambda_1^2, \lambda_2^2, ..., \lambda_r^2\}$. Equation (2.202) gives rise to

$$\mathbf{V}_1^H\mathbf{A}^H\mathbf{A}\mathbf{V}_1 = \mathbf{V}_1^H\mathbf{V}_1\mathbf{\Lambda}^2 = \mathbf{\Lambda}^2,$$

implying that

$$(\mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1})^H(\mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1}) = \mathbf{\Lambda}^{-1}(\mathbf{V}_1^H\mathbf{A}^H\mathbf{A}\mathbf{V}_1)\mathbf{\Lambda}^{-1} = \mathbf{I} \tag{2.203}$$

where we have used the fact that $\mathbf{\Lambda}^{-H} = \mathbf{\Lambda}^{-1}$ since the λ_i are real. Let the $M \times r$ matrix $\mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1} = \mathbf{U}_1$, i.e.

$$\mathbf{U}_1 = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r) = \left(\frac{\mathbf{A}\mathbf{v}_1}{\lambda_1}, \frac{\mathbf{A}\mathbf{v}_2}{\lambda_2}, ..., \frac{\mathbf{A}\mathbf{v}_r}{\lambda_r} \right). \tag{2.204}$$

From (2.203), we obtain $\mathbf{U}_1^H\mathbf{U}_1 = \mathbf{I}$ which gives

$$\mathbf{U}_1^H(\mathbf{A}\mathbf{V}_1\mathbf{\Lambda}^{-1}) = \mathbf{I}$$

or

$$\mathbf{A}\mathbf{V}_1 = \mathbf{U}_1\mathbf{\Lambda}. \tag{2.205}$$

Choose an $M \times (M - r)$ matrix $\mathbf{U}_2 = (\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, ..., \mathbf{u}_M)$ such that $\mathbf{U} = (\mathbf{U}_1 \ \mathbf{U}_2)$ is an $M \times M$ unitary matrix, i.e. $\mathbf{U}_2^H\mathbf{U}_2 = \mathbf{I}$, $\mathbf{U}_2^H\mathbf{U}_1 = \mathbf{0}$, and $\mathbf{U}_1^H\mathbf{U}_2 = \mathbf{0}$. Then

$$\begin{aligned}\mathbf{U}^H\mathbf{A}\mathbf{V} &= \begin{pmatrix}\mathbf{U}_1^H\\ \mathbf{U}_2^H\end{pmatrix}\mathbf{A}(\mathbf{V}_1\ \mathbf{V}_2) = \begin{pmatrix}\mathbf{U}_1^H\mathbf{A}\mathbf{V}_1 & \mathbf{U}_1^H\mathbf{A}\mathbf{V}_2\\ \mathbf{U}_2^H\mathbf{A}\mathbf{V}_1 & \mathbf{U}_2^H\mathbf{A}\mathbf{V}_2\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{U}_1^H\mathbf{U}_1\boldsymbol{\Lambda} & \mathbf{0}\\ \mathbf{U}_2^H\mathbf{U}_1\boldsymbol{\Lambda} & \mathbf{0}\end{pmatrix} \qquad \text{(by (2.201) and (2.205))}\\ &= \begin{pmatrix}\boldsymbol{\Lambda} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix} = \boldsymbol{\Sigma}. \end{aligned} \tag{2.206}$$

This, together with the fact that both $\mathbf{U}$ and $\mathbf{V}$ are unitary, therefore proves the theorem.

Approach II: Derivation from the Matrix $\mathbf{A}\mathbf{A}^H$

According to Properties 2.13 and 2.11, the eigenvalues of the $M \times M$ matrix $\mathbf{A}\mathbf{A}^H$ are all real nonnegative. Therefore, let λ_1, λ_2, ..., λ_M be nonnegative real numbers, and λ_1^2, λ_2^2, ..., λ_M^2 be the eigenvalues of $\mathbf{A}\mathbf{A}^H$. Furthermore, let these eigenvalues be arranged in the following order:

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_r^2 > 0$$

and

$$\lambda_{r+1}^2 = \lambda_{r+2}^2 = \cdots = \lambda_M^2 = 0. \tag{2.207}$$

Let $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$ be the orthonormal eigenvectors of $\mathbf{A}\mathbf{A}^H$ corresponding to the eigenvalues λ_1^2, λ_2^2, ..., λ_M^2, respectively. That is,

$$\mathbf{A}\mathbf{A}^H\mathbf{u}_i = \lambda_i^2\mathbf{u}_i, \quad i = 1, 2, ..., M \tag{2.208}$$

and

$$\mathbf{u}_i^H\mathbf{u}_j = \begin{cases} 1, & \text{for } i = j,\\ 0, & \text{for } i \neq j. \end{cases} \tag{2.209}$$

Let $\mathbf{U} = (\mathbf{U}_1\ \mathbf{U}_2)$ be an $M \times M$ matrix where $\mathbf{U}_1 = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_r)$ is an $M \times r$ matrix and $\mathbf{U}_2 = (\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, ..., \mathbf{u}_M)$ is an $M \times (M - r)$ matrix. Then, from (2.209), it follows that

$$\mathbf{U}^H\mathbf{U} = \begin{pmatrix}\mathbf{U}_1^H\\ \mathbf{U}_2^H\end{pmatrix}(\mathbf{U}_1\ \mathbf{U}_2) = \begin{pmatrix}\mathbf{U}_1^H\mathbf{U}_1 & \mathbf{U}_1^H\mathbf{U}_2\\ \mathbf{U}_2^H\mathbf{U}_1 & \mathbf{U}_2^H\mathbf{U}_2\end{pmatrix} = \mathbf{I},$$

i.e. $\mathbf{U}$ is a unitary matrix. By (2.207) and (2.208), we have

$$\mathbf{A}\mathbf{A}^H\mathbf{U}_2 = (\lambda_{r+1}^2\mathbf{u}_{r+1}, \lambda_{r+2}^2\mathbf{u}_{r+2}, ..., \lambda_M^2\mathbf{u}_M) = \mathbf{0},$$

implying that

$$(\mathbf{U}_2^H\mathbf{A})(\mathbf{U}_2^H\mathbf{A})^H = \mathbf{U}_2^H(\mathbf{A}\mathbf{A}^H\mathbf{U}_2) = \mathbf{0}$$

or

$$\mathbf{U}_2^H\mathbf{A} = \mathbf{0}. \tag{2.210}$$

In the same way, by (2.208), we have

$$\mathbf{A}\mathbf{A}^H\mathbf{U}_1 = (\lambda_1^2\mathbf{u}_1, \lambda_2^2\mathbf{u}_2, ..., \lambda_r^2\mathbf{u}_r) = \mathbf{U}_1\mathbf{\Lambda}^2 \tag{2.211}$$

or

$$\mathbf{U}_1^H\mathbf{A}\mathbf{A}^H\mathbf{U}_1 = \mathbf{U}_1^H\mathbf{U}_1\mathbf{\Lambda}^2 = \mathbf{\Lambda}^2,$$

implying that

$$(\mathbf{A}^H\mathbf{U}_1\mathbf{\Lambda}^{-1})^H(\mathbf{A}^H\mathbf{U}_1\mathbf{\Lambda}^{-1}) = \mathbf{\Lambda}^{-1}(\mathbf{U}_1^H\mathbf{A}\mathbf{A}^H\mathbf{U}_1)\mathbf{\Lambda}^{-1} = \mathbf{I}. \tag{2.212}$$

Let the $K \times r$ matrix $\mathbf{A}^H\mathbf{U}_1\mathbf{\Lambda}^{-1} = \mathbf{V}_1$, i.e.

$$\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r) = \left(\frac{\mathbf{A}^H\mathbf{u}_1}{\lambda_1}, \frac{\mathbf{A}^H\mathbf{u}_2}{\lambda_2}, ..., \frac{\mathbf{A}^H\mathbf{u}_r}{\lambda_r}\right). \tag{2.213}$$

From (2.212), we obtain $\mathbf{V}_1^H\mathbf{V}_1 = \mathbf{I}$, which gives

$$(\mathbf{A}^H\mathbf{U}_1\mathbf{\Lambda}^{-1})^H\mathbf{V}_1 = \mathbf{I}$$

or

$$\mathbf{U}_1^H\mathbf{A} = \mathbf{\Lambda}\mathbf{V}_1^H. \tag{2.214}$$

Choose a $K \times (K-r)$ matrix $\mathbf{V}_2 = (\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, ..., \mathbf{v}_K)$ such that $\mathbf{V} = (\mathbf{V}_1\ \mathbf{V}_2)$ is a $K \times K$ unitary matrix, i.e. $\mathbf{V}_2^H\mathbf{V}_2 = \mathbf{I}$, $\mathbf{V}_2^H\mathbf{V}_1 = \mathbf{0}$, and $\mathbf{V}_1^H\mathbf{V}_2 = \mathbf{0}$. Then

$$\begin{aligned}
\mathbf{U}^H\mathbf{A}\mathbf{V} &= \begin{pmatrix} \mathbf{U}_1^H \\ \mathbf{U}_2^H \end{pmatrix}\mathbf{A}(\mathbf{V}_1\ \mathbf{V}_2) = \begin{pmatrix} \mathbf{U}_1^H\mathbf{A}\mathbf{V}_1 & \mathbf{U}_1^H\mathbf{A}\mathbf{V}_2 \\ \mathbf{U}_2^H\mathbf{A}\mathbf{V}_1 & \mathbf{U}_2^H\mathbf{A}\mathbf{V}_2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{\Lambda}\mathbf{V}_1^H\mathbf{V}_1 & \mathbf{\Lambda}\mathbf{V}_1^H\mathbf{V}_2 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \text{(by (2.210) and (2.214))} \\
&= \begin{pmatrix} \mathbf{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \mathbf{\Sigma}.
\end{aligned} \tag{2.215}$$

This, together with the fact that both $\mathbf{U}$ and $\mathbf{V}$ are unitary, proves this theorem, too.

Q.E.D.

Appendix 2B Some Terminologies of Functions

A *function* written as $y = f(x)$ is a rule that assigns to each element x in a set A one and only one element y in a set B. The set A is called the *domain* of $f(x)$, while the set B is called the *range* of $f(x)$. The symbol x representing an arbitrary element in A is called an *independent variable.* Some terminologies for $f(x)$ defined on an interval $[x_\mathrm{L}, x_\mathrm{U}]$ (the domain of $f(x)$) are as follows.

- A function $f(x)$ is said to be *even* (*odd*) if $f(-x) = f(x)$ $(f(-x) = -f(x))$ for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$.
- A function $f(x)$ is said to be *periodic with period* T if $f(x+kT) = f(x)$ for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$ and any nonzero integer k.
- A function $f(x)$ is said to be *bounded* if $|f(x)| \le M$ for all $x \in [x_\mathrm{L}, x_\mathrm{U}]$ where M is a positive constant.
- A function $f(x)$ is said to be *increasing* (*decreasing*) or, briefly, *monotonic* if $f(x_0) \le f(x_1)$ $(f(x_0) \ge f(x_1))$ for all x_0, $x_1 \in [x_\mathrm{L}, x_\mathrm{U}]$ and $x_0 < x_1$.
- A function $f(x)$ is said to be *strictly increasing* (*strictly decreasing*) if $f(x_0) < f(x_1)$ $(f(x_0) > f(x_1))$ for all x_0, $x_1 \in [x_\mathrm{L}, x_\mathrm{U}]$ and $x_0 < x_1$.

Continuity of Functions

A function $f(x)$ is said to be *continuous at a point* x_0 if $\lim_{x\to x_0} f(x) = f(x_0)$, i.e. for every real number $\varepsilon > 0$ there exists a real number $\Delta x > 0$ such that

$$|f(x) - f(x_0)| < \varepsilon \quad \text{whenever } 0 < |x - x_0| < \Delta x \tag{2.216}$$

where Δx is, in general, dependent on ε and x_0. Furthermore, define the *left-hand limit* of $f(x)$ at a point x_0 as

$$f(x_0^-) = \lim_{x\to x_0^-} f(x) = \lim_{\substack{x\to x_0\\ x<x_0}} f(x) \tag{2.217}$$

and the *right-hand limit* of $f(x)$ at x_0 as

$$f(x_0^+) = \lim_{x\to x_0^+} f(x) = \lim_{\substack{x\to x_0\\ x>x_0}} f(x). \tag{2.218}$$

Then $f(x)$ is continuous at x_0 if and only if [29, 33]

$$f(x_0^-) = f(x_0^+) = f(x_0). \tag{2.219}$$

On the other hand, a function $f(x)$ is said to be *discontinuous at a point* x_0 if it fails to be continuous at x_0.

A function $f(x)$ is said to be *continuous on an open interval* $(x_\mathrm{L}, x_\mathrm{U})$ if it is continuous at every point $x \in (x_\mathrm{L}, x_\mathrm{U})$. A function $f(x)$ is said to be *continuous on a closed interval* $[x_\mathrm{L}, x_\mathrm{U}]$ if it is continuous on $(x_\mathrm{L}, x_\mathrm{U})$ and, meanwhile, $f(x_{\rm L}^+) = f(x_{\rm L})$ and $f(x_{\rm U}^-) = f(x_{\rm U})$. Furthermore, as illustrated in Fig. 2.10, a function $f(x)$ is said to be *piecewise continuous on an interval* $[x_{\rm L}, x_{\rm U}]$ if there are at most a finite number of points $x_{\rm L} = x_1 < x_2 < \cdots < x_n = x_{\rm U}$ such that (i) $f(x)$ is continuous on each subinterval (x_k, x_{k+1}) for $k = 1, 2, ..., n-1$ and (ii) both $f(x_k^-)$ and $f(x_k^+)$ exist for $k = 1, 2, ..., n$ [13, 18, 34]. In a word, a piecewise continuous function has a finite number of finite discontinuities. Moreover, a continuous function is merely a special case of a piecewise continuous function.

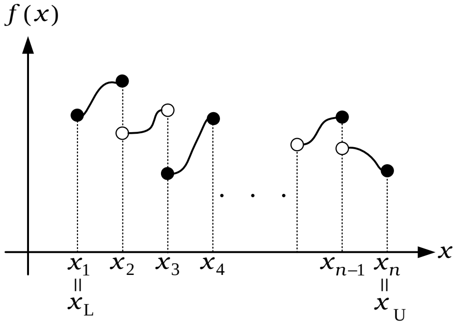

Fig. 2.10 A piecewise continuous function $f(x)$ on an interval $[x_{\rm L}, x_{\rm U}]$

Continuity of Derivatives

The *derivative* of a function $f(x)$ at a point x_0 is defined as

$$f'(x_0) = \left.\frac{df(x)}{dx}\right|_{x=x_0} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \tag{2.220}$$

provided that the limit exists. Define the *left-hand derivative* of $f(x)$ at x_0 as

$$f'(x_0^-) = \lim_{\Delta x \to 0^-} \frac{f(x_0 + \Delta x) - f(x_0^-)}{\Delta x} \tag{2.221}$$

and the *right-hand derivative* of $f(x)$ at x_0 as

$$f'(x_0^+) = \lim_{\Delta x \to 0^+} \frac{f(x_0 + \Delta x) - f(x_0^+)}{\Delta x}. \tag{2.222}$$

Then the derivative $f'(x)$ is said to be *piecewise continuous on an interval* $[x_{\rm L}, x_{\rm U}]$ if $f(x)$ is piecewise continuous on $[x_{\rm L}, x_{\rm U}]$ and, meanwhile, there are at most a finite number of points $x_{\rm L} = x_1 < x_2 < \cdots < x_n = x_{\rm U}$ such that (i) $f'(x)$ exists and is continuous on each subinterval (x_k, x_{k+1}) for $k = 1, 2, ..., n-1$ and (ii) both $f'(x_k^-)$ and $f'(x_k^+)$ exist for $k = 1, 2, ..., n$ [13, 18]. Note that if $f'(x)$ exists at a point x_0, then $f(x)$ is continuous at x_0.

Appendix 2C Proof of Theorem 2.33

From the assumption that $\sum_{n=-\infty}^{\infty} |a_n|^s < \infty$, it follows that $|a_n|$ is bounded above. Let $\beta = \max\{|a_n|, n = -N \sim N\}$. Since $\max\{|a_n|/\beta, n = -N \sim N\} = 1$ and $l > s \geq 1$, one can easily infer that

$$1 \leq \sum_{n=-N}^{N} \left(\frac{|a_n|}{\beta}\right)^l \leq \sum_{n=-N}^{N} \left(\frac{|a_n|}{\beta}\right)^s , \tag{2.223}$$

which further leads to

$$\left\{\sum_{n=-N}^{N} \left(\frac{|a_n|}{\beta}\right)^l\right\}^{1/l} \leq \left\{\sum_{n=-N}^{N} \left(\frac{|a_n|}{\beta}\right)^s\right\}^{1/s} . \tag{2.224}$$

Canceling the common term β on both sides of (2.224) yields

$$\left\{\sum_{n=-N}^{N} |a_n|^l\right\}^{1/l} \leq \left\{\sum_{n=-N}^{N} |a_n|^s\right\}^{1/s} . \tag{2.225}$$

Since (2.225) holds for any N and $\left\{\sum_{n=-\infty}^{\infty} |a_n|^s\right\}^{1/s} < \infty$, letting $N \to \infty$ therefore gives (2.59). Thus, what remains to prove is the equality condition of (2.59).

Suppose that there are M terms of the sequence $\{a_n\}$ corresponding to $|a_n| = \beta$, and that $|a_n| < \beta$ for $n \in \Omega$ where Ω is a set of indices. It can be seen that the equality of (2.59) requires the equality of (2.223) and the equality of (2.224) for $N \to \infty$. From the equality of (2.223) for $N \to \infty$, we have

$$M + \sum_{n\in\Omega} \left(\frac{|a_n|}{\beta}\right)^l = M + \sum_{n\in\Omega} \left(\frac{|a_n|}{\beta}\right)^s ,$$

implying that $|a_n| = 0$ for $n \in \Omega$. From this result and the equality of (2.224) for $N \to \infty$, we have

$$M^{1/l} = M^{1/s},$$

implying that $M = 1$. This therefore completes the proof.

Q.E.D.

Appendix 2D Proof of Theorem 2.36

Since $s_n(x)$ is periodic with period 2π, substituting (2.76) into (2.77) yields

$$\begin{aligned} s_n(x) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-t)D_n(t)dt = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x+t)D_n(t)dt \\ &= g_n(x) + \widetilde{g}_n(x) \end{aligned} \tag{2.226}$$

where

$$D_n(t) = \sum_{k=-n}^{n} e^{jkt} = \frac{\sin\left[(2n+1)t/2\right]}{\sin(t/2)} \tag{2.227}$$

is the so-called *nth Dirichlet kernel* [14] and

$$g_n(x) = \frac{1}{2\pi}\int_{0}^{\pi} f(x+t)D_n(t)dt, \tag{2.228}$$

$$\widetilde{g}_n(x) = \frac{1}{2\pi}\int_{-\pi}^{0} f(x+t)D_n(t)dt. \tag{2.229}$$

By expressing $D_n(t) = 1 + 2\sum_{k=1}^{n}\cos kt$, we obtain the integrations

$$\frac{1}{2\pi}\int_{0}^{\pi} D_n(t)dt = \frac{1}{2} \quad \text{and} \quad \frac{1}{2\pi}\int_{-\pi}^{0} D_n(t)dt = \frac{1}{2}. \tag{2.230}$$

Further express $g_n(x)$ given by (2.228) as

$$g_n(x) = \frac{1}{2\pi}\int_{0}^{\pi} f(x^+)D_n(t)dt + \frac{1}{2\pi}\int_{0}^{\pi} [f(x+t) - f(x^+)]D_n(t)dt$$

which, together with (2.227) and (2.230), gives

$$\begin{aligned} g_n(x) - \frac{f(x^+)}{2} &= \frac{1}{2\pi}\int_{-\pi}^{\pi} h(t)\sin\frac{(2n+1)t}{2}dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} h_1(t)\sin(nt)dt + \frac{1}{2\pi}\int_{-\pi}^{\pi} h_2(t)\cos(nt)dt \end{aligned} \tag{2.231}$$

where

$$h(t) = \begin{cases} \dfrac{f(x+t) - f(x^+)}{\sin(t/2)}, & 0 \le t < \pi, \\ 0, & -\pi \le t < 0, \end{cases}$$

and $h_1(t) = h(t)\cos(t/2)$, $h_2(t) = h(t)\sin(t/2)$. By definition, the left-hand limit $h(0^-) = \lim_{t\to 0^-} h(t) = 0$, and the right-hand limit

$$h(0^+) = \lim_{t\to 0^+} h(t) = \lim_{t\to 0^+} \left[\frac{f(x+t) - f(x^+)}{t}\right] \cdot \left[\frac{t}{\sin(t/2)}\right] = 2f'(x^+)$$

exists since $f'(x^+)$ exists. From this and the condition that $f(x)$ is piecewise continuous on $[-\pi, \pi)$, it follows that $h(t)$ is piecewise continuous on $[-\pi, \pi)$ and, thus, square integrable on $[-\pi, \pi)$. In other words, $h(t)$ is in $\mathcal{L}^2[-\pi, \pi)$ and so are $h_1(t)$ and $h_2(t)$. Accordingly, the two terms in the second line of (2.231) are identical to zero as $n \to \infty$ (by Problem 2.16) and therefore

$$\lim_{n\to\infty} g_n(x) = \frac{f(x^+)}{2}. \tag{2.232}$$

In a similar way, by expressing $\widetilde{g}_n(x)$ given by (2.229) as

$$\widetilde{g}_n(x) = \frac{1}{2\pi}\int_{-\pi}^{0} f(x^-)D_n(t)dt + \frac{1}{2\pi}\int_{-\pi}^{0} [f(x+t) - f(x^-)]D_n(t)dt$$

and with the condition that $f'(x^-)$ exists and $f(x)$ is piecewise continuous on $[-\pi, \pi)$, we also have

$$\lim_{n\to\infty} \widetilde{g}_n(x) = \frac{f(x^-)}{2}. \tag{2.233}$$

Equation (2.78) then follows from (2.226), (2.232) and (2.233).

Q.E.D.

Appendix 2E Proof of Theorem 2.38

Since $f'(x)$ is piecewise continuous on $[-\pi, \pi)$, it is integrable on $[-\pi, \pi)$ and has the Fourier series

$$f'(x) \sim \sum_{k=-\infty}^{\infty} \widetilde{c}_k e^{jkx}$$

where

$$\widetilde{c}_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f'(x)dx = \frac{f(\pi) - f(-\pi)}{2\pi} = 0, \tag{2.234}$$

$$\begin{aligned} \widetilde{c}_k &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f'(x)e^{-jkx}dx = \left.\frac{f(x)e^{-jkx}}{2\pi}\right|_{-\pi}^{\pi} + \frac{jk}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-jkx}dx \\ &= jkc_k, \quad \text{for } |k| = 1 \sim \infty. \end{aligned} \tag{2.235}$$

By (2.235) and the Cauchy–Schwartz inequality (Theorem 2.32), we have

$$\begin{aligned}\sum_{k=-n}^{n}|c_k| &= |c_0| + \sum_{k=-n,k\neq 0}^{n} |k|^{-1}\cdot|\widetilde{c}_k| \\ &\leq |c_0| + \left\{\sum_{k=-n,k\neq 0}^{n}|k|^{-2}\right\}^{1/2}\left\{\sum_{k=-n,k\neq 0}^{n}|\widetilde{c}_k|^2\right\}^{1/2} \\ &= |c_0| + \sqrt{2}\left\{\sum_{k=1}^{n}k^{-2}\right\}^{1/2}\left\{\sum_{k=1}^{n}\left(|\widetilde{c}_k|^2+|\widetilde{c}_{-k}|^2\right)\right\}^{1/2}. \qquad (2.236)\end{aligned}$$

As shown in Example 2.26, the series $\sum_{k=1}^{\infty} k^{-2}$ converges, indicating that

$$\sum_{k=1}^{\infty} k^{-2} < \infty. \qquad (2.237)$$

Moreover, since $f'(x)$ is piecewise continuous on $[-\pi,\pi)$, it is square integrable on $[-\pi,\pi)$ and therefore is in $\mathcal{L}^2[-\pi,\pi)$. Accordingly, by Bessel's inequality (2.69) and (2.234),

$$\sum_{k=-\infty,k\neq 0}^{\infty}|\widetilde{c}_k|^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi}|f'(x)|^2\,dx < \infty. \qquad (2.238)$$

As a consequence of (2.236), (2.237) and (2.238), $\sum_{k=-\infty}^{\infty}|c_k| < \infty$ and, by Theorem 2.37, the Fourier series given by (2.75) is uniformly and absolutely convergent on $[-\pi,\pi)$. From this and by the pointwise convergence theorem, it then follows that the Fourier series given by (2.75) converges uniformly and absolutely to $f(x)$ since $f(x)$ is continuous on $[-\pi,\pi)$.

Q.E.D.

Appendix 2F Proof of Theorem 2.46

According to Property 2.12, the proof is equivalent to showing that $\mathbf{P}^{[i+1]}$ is positive definite under the conditions that (i) both $\mathbf{P}^{[i]}$ and $\mathbf{Q}^{[i]}$ are positive definite, (ii) $\mathbf{r}_{i+1} \neq \mathbf{0}$, (iii) $\mathbf{s}_{i+1} \neq \mathbf{0}$, and (iv) $\mu^{[i]}$ is optimum for iteration i.

By (2.152), we can express the Hermitian form of $\mathbf{P}^{[i+1]}$ as follows:

$$\begin{aligned}\mathbf{x}^H\mathbf{P}^{[i+1]}\mathbf{x} &= \mathbf{x}^H\mathbf{P}^{[i]}\mathbf{x} + \frac{\left|\mathbf{s}_{i+1}^H\mathbf{x}\right|^2}{\alpha\mathbf{s}_{i+1}^H\mathbf{r}_{i+1}} - \frac{\left|\mathbf{r}_{i+1}^H\mathbf{P}^{[i]}\mathbf{x}\right|^2}{\mathbf{r}_{i+1}^H\mathbf{P}^{[i]}\mathbf{r}_{i+1}} \\ &= \frac{\left(\mathbf{x}^H\mathbf{P}^{[i]}\mathbf{x}\right)\left(\mathbf{r}_{i+1}^H\mathbf{P}^{[i]}\mathbf{r}_{i+1}\right) - \left|\mathbf{r}_{i+1}^H\mathbf{P}^{[i]}\mathbf{x}\right|^2}{\mathbf{r}_{i+1}^H\mathbf{P}^{[i]}\mathbf{r}_{i+1}} + \frac{\left|\mathbf{s}_{i+1}^H\mathbf{x}\right|^2}{\alpha\mathbf{s}_{i+1}^H\mathbf{r}_{i+1}} \qquad (2.239)\end{aligned}$$

for any $\mathbf{x} \neq \mathbf{0}$. Let the SVD of the Hermitian matrix $\mathbf{P}^{[i]}$ be written as

$$\mathbf{P}^{[i]} = \sum_{k=1}^{\widetilde{L}} \lambda_k \mathbf{u}_k \mathbf{u}_k^H \quad \text{(see (2.42))} \tag{2.240}$$

where

$$\widetilde{L} = \begin{cases} L & \text{for real } \boldsymbol{\theta}, \\ 2L & \text{for complex } \boldsymbol{\theta} \end{cases} \tag{2.241}$$

and $\mathbf{u}_k$ is the orthonormal eigenvector of $\mathbf{P}^{[i]}$ associated with the eigenvalue λ_k. Since $\mathbf{P}^{[i]}$ is positive definite, all the eigenvalues λ_k are (real) positive. Then, by the Cauchy–Schwartz inequality (Theorem 2.1), we have

$$\begin{aligned} \left|\mathbf{r}_{i+1}^H \mathbf{P}^{[i]} \mathbf{x}\right|^2 &= \left| \sum_{k=1}^{\widetilde{L}} \left(\sqrt{\lambda_k} \mathbf{r}_{i+1}^H \mathbf{u}_k\right) \left(\sqrt{\lambda_k} \mathbf{u}_k^H \mathbf{x}\right) \right|^2 \\ &\leq \left\{ \sum_{k=1}^{\widetilde{L}} \lambda_k \left|\mathbf{r}_{i+1}^H \mathbf{u}_k\right|^2 \right\} \left\{ \sum_{k=1}^{\widetilde{L}} \lambda_k \left|\mathbf{u}_k^H \mathbf{x}\right|^2 \right\} \\ &= \left\{ \sum_{k=1}^{\widetilde{L}} \lambda_k \mathbf{r}_{i+1}^H \mathbf{u}_k \mathbf{u}_k^H \mathbf{r}_{i+1} \right\} \left\{ \sum_{k=1}^{\widetilde{L}} \lambda_k \mathbf{x}^H \mathbf{u}_k \mathbf{u}_k^H \mathbf{x} \right\} \\ &= \left(\mathbf{r}_{i+1}^H \mathbf{P}^{[i]} \mathbf{r}_{i+1}\right) \left(\mathbf{x}^H \mathbf{P}^{[i]} \mathbf{x}\right). \end{aligned} \tag{2.242}$$

From (2.239), (2.242), and the fact that $\mathbf{r}_{i+1}^H \mathbf{P}^{[i]} \mathbf{r}_{i+1} > 0$ (since $\mathbf{P}^{[i]}$ is positive definite and $\mathbf{r}_{i+1} \neq \mathbf{0}$), it follows that

$$\mathbf{x}^H \mathbf{P}^{[i+1]} \mathbf{x} \geq \frac{\left|\mathbf{s}_{i+1}^H \mathbf{x}\right|^2}{\alpha \mathbf{s}_{i+1}^H \mathbf{r}_{i+1}} \quad \text{for any } \mathbf{x} \neq \mathbf{0} \tag{2.243}$$

and the equality holds only when $\mathbf{x} = c \cdot \mathbf{r}_{i+1}$ for any nonzero scalar c.

On the other hand, since

$$\boldsymbol{\vartheta}^{[i+1]} = \boldsymbol{\vartheta}^{[i]} - \mu^{[i]} \mathbf{d}^{[i]} \tag{2.244}$$

where $\mathbf{d}^{[i]} = \mathbf{Q}^{[i]} \boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i]})$, the necessary condition for the optimum $\mu^{[i]}$ can be derived by minimizing the objective function $f(\mu^{[i]}) \triangleq J(\boldsymbol{\vartheta}^{[i]} - \mu^{[i]} \mathbf{d}^{[i]})$. More specifically, by using the chain rule, we obtain

$$\frac{df(\mu^{[i]})}{d\mu^{[i]}} = \left[\frac{\partial J(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}}\right]^H \Bigg|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{[i+1]}} \cdot \frac{d\boldsymbol{\vartheta}^{[i+1]}}{d\mu^{[i]}} = -\left[\boldsymbol{\nabla} J(\boldsymbol{\vartheta}^{[i+1]})\right]^H \mathbf{d}^{[i]} = 0. \tag{2.245}$$

This result, together with (2.141), (2.142) and (2.244), therefore leads to

$$\begin{aligned}\mathbf{s}_{i+1}^H\mathbf{r}_{i+1} &= \mu^{[i]}\left[\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})\right]^H\mathbf{d}^{[i]}\\ &= \mu^{[i]}\left[\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]})\right]^H\mathbf{Q}^{[i]}\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}) > 0 \qquad (2.246)\end{aligned}$$

since $\mu^{[i]} > 0$, $\mathbf{Q}^{[i]}$ is positive definite and $\boldsymbol{\nabla}J(\boldsymbol{\vartheta}^{[i]}) \neq \mathbf{0}$ before convergence.

As a consequence of (2.243) and (2.246),

$$\mathbf{x}^H\mathbf{P}^{[i+1]}\mathbf{x} \geq \frac{\left|\mathbf{s}_{i+1}^H\mathbf{x}\right|^2}{\alpha\mathbf{s}_{i+1}^H\mathbf{r}_{i+1}} \geq 0 \quad \text{for any } \mathbf{x} \neq \mathbf{0}. \qquad (2.247)$$

The first equality of (2.247) holds only when $\mathbf{x} = c \cdot \mathbf{r}_{i+1}$ for any $c \neq 0$, while the second equality of (2.247) holds only when $\mathbf{s}_{i+1}^H\mathbf{x} = 0$. In other words, for any $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^H\mathbf{P}^{[i+1]}\mathbf{x} = 0$ happens only when $\mathbf{s}_{i+1}^H(c \cdot \mathbf{r}_{i+1}) = 0$, that contradicts (2.246). As a result, the Hermitian form $\mathbf{x}^H\mathbf{P}^{[i+1]}\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$ and accordingly $\mathbf{P}^{[i+1]}$ is positive definite.

Q.E.D.

Problems

2.1. Prove Theorem 2.1.

2.2. Prove Theorem 2.2. (*Hint*: Use the Cauchy–Schwartz inequality.)

2.3. Prove Theorem 2.5.

2.4. Prove Theorem 2.7. (*Hint*: Express **A** as a multiplication of a lower triangular matrix and an upper triangular matrix.)

2.5. Prove Theorem 2.8.

2.6. Prove Properties 2.9 to 2.12.

2.7. Prove Property 2.13.

2.8. Prove Property 2.14. (*Hint*: Use Property 2.10.)

2.9. (a) Find the eigenvalues and the normalized eigenvectors of the matrix

$$\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}.$$

(b) Use part (a) to find the eigendecomposition of the matrix **A**.

2.10. Prove Corollary 2.16.

2.11. Prove Theorem 2.17.

2.12. Prove Theorem 2.18.

2.13. Prove Theorem 2.21.

2.14. Prove Theorem 2.25.

2.15. Prove Theorem 2.30.

2.16. Suppose $\{\phi_n(x), n = -\infty \sim \infty\}$ is a set of real or complex orthogonal functions in $\mathcal{L}^2[x_\mathrm{L}, x_\mathrm{U}]$. Show that if $f(x)$ is a real or complex function in $\mathcal{L}^2[x_\mathrm{L}, x_\mathrm{U}]$, then

$$\lim_{|n|\to\infty} \int_{x_\mathrm{L}}^{x_\mathrm{U}} f(x)\phi_n^*(x)dx = 0.$$

(*Hint*: Use Bessel's inequality.)

2.17. Prove Theorem 2.37. (*Hint*: Use the Weierstrass M-test and Theorem 2.27.)

2.18. Prove Theorem 2.39. (*Hint*: Use Theorem 2.30 to show that the sequence $\{\sum_{k=-n}^{n} c_k e^{jkx}\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathcal{L}^2[-\pi, \pi)$.)

2.19. Prove Theorem 2.42.

2.20. Prove Theorem 2.44.

2.21. Show that if $\mathbf{A}$ is a full-rank $M \times K$ matrix and $M \geq K$, then its pseudoinverse $\mathbf{A}^+ = (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H$.

2.22. Find the LS solution to the set of linear equations $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & -1 \\ 5 & 2 \\ 3 & -4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 2 \\ -1 \\ 1 \\ 3 \end{pmatrix}.$$

2.23. Consider the set of linear equations $\mathbf{A}\boldsymbol{\theta} = \mathbf{b}$ where

$$\mathbf{A} = \begin{pmatrix} 0.6 & -1.6 & 0 \\ 0.8 & 1.2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -0.5 \\ 2 \\ 0 \\ 0 \end{pmatrix}.$$

(a) Find the SVD of $\mathbf{A}$.
(b) Find the LS solution for $\boldsymbol{\theta}$.

Computer Assignments

2.1. Suppose $f(x)$ is a periodic function of period 2π and

$$f(x) = \begin{cases} 1, & |x| \leq \pi/2, \\ 0, & \pi/2 < |x| \leq \pi. \end{cases}$$

(a) Let $s_n(x)$ denote the nth partial sum of the Fourier series of $f(x)$. Find the Fourier series of $f(x)$ and the partial sum $\lim_{n\to\infty} s_n(x)$.

(b) Plot the partial sums $s_1(x)$, $s_3(x)$ and $s_{23}(x)$, and specify what phenomenon you observe.

References

1. R. A. Horn and C. R. Johnson, *Matrix Analysis.* New York: Cambridge University Press, 1990.
2. G. Strang, *Linear Algebra and Its Applications.* New York: Harcourt Brace Jovanovich, 1988.
3. S. H. Friedberg, A. J. Insel, and L. E. Spence, *Linear Algebra.* New Jersey: Prentice-Hall, 1989.
4. E. Kreyszig, *Advanced Engineering Mathematics.* New York: John Wiley & Sons, 1999.
5. H. Stark and Y. Yang, *Vector Space Projections: A Numerical Approach to Signal and Image Processing, Neural Nets, and Optics.* New York: John Wiley & Sons, 1998.
6. G. H. Golub and C. F. van Loan, *Matrix Computations.* London: The Johns Hopkins University Press, 1989.
7. S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* New Jersey: Prentice-Hall, 1993.
8. S. Haykin, *Adaptive Filter Theory.* New Jersey: Prentice-Hall, 1996.
9. J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control.* New Jersey: Prentice-Hall, 1995.
10. V. C. Klema and A. J. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Trans. Automatic Control*, vol. AC-25, no. 2, pp. 164–176, Apr. 1980.
11. D. Kincaid and W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing.* California: Brooks/Cole, 1996.
12. K. G. Binmore, *Mathematical Analysis: A Straightforward Approach.* New York: Cambridge University Press, 1982.
13. M. Stoll, *Introduction to Real Analysis.* Boston: Addison Wesley Longman, 2001.
14. E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction.* New Jersey: Princeton University Press, 2003.
15. R. G. Bartle, *The Elements of Real Analysis.* New York: John Wiley & Sons, 1970.

16. W. R. Wade, *An Introduction to Analysis.* New Jersey: Prentice-Hall, 1995.
17. H. L. Royden, *Real Analysis.* New Jersey: Prentice-Hall, 1988.
18. E. Haug and K. K. Choi, *Methods of Engineering Mathematics.* New Jersey: Prentice-Hall, 1993.
19. C.-Y. Chi and M.-C. Wu, "Inverse filter criteria for blind deconvolution and equalization using two cumulants," *Signal Processing*, vol. 43, no. 1, pp. 55–63, Apr. 1995.
20. G. H. Hardy, J. E. Littlewood, and G. Polya, *Inequalities.* London: Cambridge University Press, 1934.
21. A. V. Oppenheim and A. S. Willsky with I. T. Young, *Signals and Systems.* New Jersey: Prentice-Hall, 1983.
22. O. K. Ersoy, *Fourier-related Transforms, Fast Algorithms and Applications.* New Jersey: Prentice-Hall, 1997.
23. R. N. Bracewell, *The Fourier Transform and Its Applications.* Boston: McGraw-Hill, 2000.
24. M. J. Lighthill, *Introduction to Fourier Analysis and Generalised Functions.* Cambridge: Cambridge University Press, 1958.
25. D. G. Luenberger, *Linear and Nonlinear Programming.* California: Addison-Wesley Publishing, 1984.
26. S. S. Rao, *Engineering Optimization: Theory and Practice.* New York: John Wiley & Sons, 1996.
27. D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.*, vol. 130, pts. F and H, no. 1, pp. 11–16, Feb. 1983.
28. E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis for Mathematics, Science, and Engineering.* New Jersey: Prentice-Hall, 1993.
29. E. W. Swokowski, *Calculus with Analytic Geometry.* Boston: Prindle, Weber & Schmidt, 1983.
30. J. W. Bandler, "Optimization methods for computer-aided design," *IEEE Trans. Microwave Theory and Techniques*, vol. MTT-17, no. 8, pp. 533–552, Aug. 1969.
31. B. S. Gottfried and J. Weisman, *Introduction to Optimization Theory.* New Jersey: Prentice-Hall, 1973.
32. H. Y. Huang, "Unified approach to quadratically convergent algorithms for function minimization," *Journal of Optimization Theory and Applications*, vol. 5, pp. 405–423, 1970.
33. J. Stewart, *Calculus: Early Transcendentals.* New York: International Thomson, 1999.
34. M. R. Spiegel, *Theory and Problems of Fourier Analysis with Applications to Boundary Value Problems.* New York: McGraw-Hill, 1974.