





# ***SAYLOR MA 304***

***Topics in Applied Mathematics***

by

Charles K. Chui

Menlo Park, California



Published by The Saylor Foundation  
1000 Wisconsin Ave NW, Suite 220  
Washington, D.C. 20007

<http://www.saylor.org>

Licensed under Creative Commons Attribution – Non-  
Commercial 3.0 (CC BY-NC)  
<http://creativecommons.org/licenses/by-nc/3.0/legalcode>

Copyright 2013 by Charles K. Chui

# Contents

<b>List of Figures</b>	<b>7</b>
<b>Preface</b>	<b>9</b>
<b>1 LINEAR ANALYSIS</b>	<b>1</b>
1.1 Inner Product and Norm Measurements . . . . .	1
1.1.1 Definition of inner product . . . . .	2
1.1.2 Cauchy-Schwarz Inequality . . . . .	3
1.1.3 Norm measurement and angle between vectors . . . . .	5
1.1.4 Gram-Schmidt orthogonalization process . . . . .	6
1.2 Eigenvalue Problems . . . . .	6
1.2.1 Linear Transformations . . . . .	6
1.2.2 Bounded linear functionals and operators . . . . .	7
1.2.3 Eigenvalues and eigenspaces . . . . .	12
1.2.4 Self-adjoint positive definite operators . . . . .	12
1.3 Singular Value Decomposition (SVD) . . . . .	13
1.3.1 Normal operators and spectral decomposition . . . . .	13
1.3.2 Singular values . . . . .	14
1.3.3 Reduced singular value decomposition . . . . .	17
1.3.4 Full singular value decomposition . . . . .	20
1.4 Principal Component Analysis (PCA) and Its Applications . . . . .	23
1.4.1 Frobenius norm measurement . . . . .	23
1.4.2 Principal components for data-dependent basis . . . . .	26
1.4.3 Pseudo-inverses . . . . .	29
1.4.4 Minimum-norm least-squares estimation . . . . .	30
1.5 Applications to Data Dimensionality Reduction . . . . .	37
1.5.1 Representation of matrices by sum of norm-1 matrices . . . . .	37
1.5.2 Approximation by matrices of lower ranks . . . . .	38
1.5.3 Motivation to data-dimensionality reduction . . . . .	40
1.5.4 Principal components as basis for dimension-reduced data . . . . .	50
<b>2 DATA COMPRESSION</b>	<b>55</b>
2.1 Discrete and Fast Fourier Transform (FFT) . . . . .	55
2.1.1 Definition of DFT . . . . .	56
2.1.2 Lanczos matrix factorization . . . . .	56

2.1.3	FFT for fast computation . . . . .	59
2.2	Discrete Cosine Transform (DCT) . . . . .	61
2.2.1	Derivation of DCT from DFT . . . . .	61
2.2.2	8-Point DCT . . . . .	61
2.2.3	2-Dimensional DCT . . . . .	62
2.3	Information Coding . . . . .	64
2.3.1	Probability distributions . . . . .	64
2.3.2	Histogram . . . . .	68
2.3.3	Entropy . . . . .	72
2.3.4	Binary codes . . . . .	73
2.4	Data Compression Schemes . . . . .	76
2.4.1	Lossless and lossy compression . . . . .	77
2.4.2	Kraft inequality . . . . .	78
2.4.3	Huffman coding scheme . . . . .	78
2.4.4	Noiseless coding theorem . . . . .	78
2.5	Image and Video Compression Schemes and Standards . . . . .	83
2.5.1	Image compression scheme . . . . .	84
2.5.2	Quantization . . . . .	85
2.5.3	Huffman, DPCM, and run-length coding . . . . .	86
2.5.4	Encoder - Decoder (Codec) . . . . .	88
2.5.5	I, P, and B video frames . . . . .	91
2.5.6	Macro-blocks . . . . .	91
2.5.7	Motion search and compensation . . . . .	91
<b>3</b>	<b>FOURIER METHODS</b>	<b>95</b>
3.1	Fourier Series . . . . .	95
3.1.1	Fourier series representations . . . . .	97
3.1.2	Orthogonality and computation . . . . .	105
3.2	Orthogonal Projection . . . . .	105
3.2.1	Pythagorean theorem . . . . .	105
3.2.2	Parallelogram law . . . . .	107
3.2.3	Best mean-square approximation . . . . .	108
3.3	Dirichlet's and Fejér's Kernels . . . . .	110
3.3.1	Partial sums as convolution with Dirichlet's kernels . . . . .	110
3.3.2	Césaro means and derivation of Fejér's kernels . . . . .	112
3.3.3	Positive approximate identity . . . . .	114
3.4	Completeness . . . . .	116
3.4.1	Pointwise and uniform convergence . . . . .	117
3.4.2	Trigonometric approximation . . . . .	124
3.5	Parseval's Identity . . . . .	127
3.5.1	Derivation of Parseval's identities . . . . .	127
3.5.2	The Basel problem and Fourier method . . . . .	130
3.5.3	Bernoulli numbers and Euler's formula . . . . .	133

<b>4</b>	<b>TIME-FREQUENCY ANALYSIS</b>	<b>139</b>
4.1	Fourier Transform . . . . .	140
4.1.1	Definition and essence of the Fourier transform . . . .	140
4.1.2	Properties of the Fourier transform . . . . .	141
4.1.3	Sampling Theorem . . . . .	141
4.2	Convolution Filter and Gaussian Kernel . . . . .	141
4.2.1	Convolution filter . . . . .	142
4.2.2	Fourier transform of the Gaussian . . . . .	142
4.2.3	Inverse Fourier transform . . . . .	144
4.3	Localized Fourier Transform . . . . .	148
4.3.1	Short-time Fourier Transform (STFT) . . . . .	149
4.3.2	Gabor transform . . . . .	150
4.3.3	Inverse of localized Fourier transform . . . . .	151
4.4	Uncertainty Principle . . . . .	153
4.4.1	Time-frequency localization window measurement . .	153
4.4.2	Gaussian as optimal time-frequency window . . . . .	154
4.4.3	Derivation of the Uncertainty Principle . . . . .	156
4.5	Time-Frequency Bases . . . . .	158
4.5.1	Balian-Low restriction . . . . .	159
4.5.2	Frames . . . . .	161
4.5.3	Localized cosine basis . . . . .	165
4.5.4	Malvar wavelets . . . . .	165
<b>5</b>	<b>PDE METHODS</b>	<b>169</b>
5.1	From Gaussian Convolution to Diffusion Process . . . . .	170
5.1.1	Gaussian as solution for delta heat source . . . . .	170
5.1.2	Gaussian convolution as solution of heat equation for the real-line . . . . .	171
5.1.3	Gaussian convolution as solution of heat equation in the Euclidean space . . . . .	177
5.2	The method of separation of variables . . . . .	181
5.2.1	Separation of Time and Spatial Variables . . . . .	181
5.2.2	Superposition Solution . . . . .	182
5.2.3	Extension to two spatial variables . . . . .	182
5.3	Fourier series solution . . . . .	184
5.3.1	One-spatial dimension . . . . .	185
5.3.2	Extension to higher dimensional domains . . . . .	189
5.4	Boundary Value Problems . . . . .	192
5.4.1	Neumann boundary value problems . . . . .	193
5.4.2	Anisotropic diffusion . . . . .	197
5.4.3	Solution in terms of eigenvalue problems . . . . .	200
5.5	Application to Image De-Noising . . . . .	203
5.5.1	Diffusion as quantizer for image compression . . . . .	203
5.5.2	Diffusion for noise reduction . . . . .	206
5.5.3	Enhanced JPEG compression . . . . .	209

<b>6</b>	<b>WAVELET METHODS</b>	<b>211</b>
6.1	Time-Scale Analysis . . . . .	211
6.1.1	Wavelet transform . . . . .	212
6.1.2	Frequency versus scale . . . . .	213
6.1.3	Partition into frequency bands . . . . .	214
6.1.4	Parseval's identity for wavelet transform . . . . .	216
6.1.5	Inverse wavelet transform . . . . .	219
6.2	Multiresolution Analysis (MRA) . . . . .	220
6.2.1	Function refinement . . . . .	221
6.2.2	<i>B</i> -spline examples . . . . .	223
6.2.3	The MRA architecture . . . . .	225
6.3	Wavelet Construction . . . . .	227
6.3.1	Quadrature mirror filter . . . . .	228
6.3.2	Matrix extension . . . . .	236
6.3.3	Orthogonal and bi-orthogonal wavelets . . . . .	243
6.4	Wavelet Algorithms . . . . .	256
6.4.1	Wavelet decomposition and reconstruction . . . . .	257
6.4.2	Filter Banks . . . . .	257
6.4.3	The Lifting Scheme . . . . .	257
6.5	Application to Image Coding . . . . .	257
6.5.1	Mapping digital images to the wavelet domain . . . . .	258
6.5.2	Progressive image transmission . . . . .	266
6.5.3	Lossless JPEG-2000 compression . . . . .	268
	<b>Index</b>	<b>273</b>

## *List of Figures*

2.1	<i>Encoder: <math>Q</math> = quantization; <math>E</math> = entropy encoding . . . . .</i>	88
2.2	<i>Decoder: <math>Q^{-1}</math> = de-quantization; <math>E^{-1}</math> = de-coding . . . . .</i>	88
2.3	<i>Quantizers: low compression ratio . . . . .</i>	88
2.4	<i>Quantizers: high compression ratio . . . . .</i>	89
2.5	<i>Zig-zag ordering . . . . .</i>	89
3.1	<i>Dirichlet's kernel <math>D_{16}(x)</math> . . . . .</i>	112
3.2	<i>Fejér's kernel <math>\sigma_{16}(x)</math> . . . . .</i>	113
5.1	<i>Diffusion with delta heat source (top) and arbitrary heat source (bottom) . . . . .</i>	173





## *Preface*

Mathematics was coined the “queen of sciences” by Carl Friedrich Gauss, one of the greatest mathematicians of all time. The name of Gauss is associated with essentially all areas of mathematics. Therefore to him, and most of the great mathematicians before the end of the nineteenth century, there was really no clear boundary between “pure mathematics” and “applied mathematics.” To ensure financial independence, Gauss chose a stable career in astronomy, which is one of the oldest sciences and was perhaps the most popular one during the eighteenth and nineteenth centuries. In his study of celestial motion and orbits and a diversity of disciplines later in his career, including (in chronological order): geodesy, magnetism, dioptrics, and actuarial science, Gauss has developed a vast volume of mathematical methods and tools that are still instrumental to our current study of applied mathematics.

During the twentieth century, with the exciting development of quantum field theory, with the prosperity of the aviation industry, and with the bullish activity in financial market trading, and so forth, the sovereignty of the “queen of sciences” has turned her attention to the theoretical development and numerical solutions of partial differential equations (PDE’s). Indeed, the non-relativistic modeling of quantum mechanics is described by the Schrödinger equation; the fluid flow formulation, as an extension of Newtonian physics by incorporating motion and stress, is modeled by the Navier-Stokes equation; and option stock trading with minimum risk can be modeled by the Black-Scholes equation. All of these equations are PDEs. In general, PDE’s are used to describe a wide variety of phenomena, including: heat diffusion, sound wave propagation, electromagnetic wave radiation, vibration, electrostatics, electrodynamics, fluid flow, and elasticity, just to name a few. For this reason, the theoretical and numerical development of PDEs has been considered the core of applied mathematics, at least in the academic environment.

On the other hand, over the past two decades, we have been facing a rapidly increasing volume of “information” contents to be processed and understood. For instance, the popularity and significant impact of the open education movement (OEM) have contributed to an enormous amount of educational information in the web that are difficult to sort out, due to unavoidable redundancy, occasional contradiction, extreme variation in quality, and even erroneous opinions. This motivated the founding of the “Saylor Foundation courseware” to provide perhaps one of the most valuable, and certainly more reliable, high-quality educational materials, with end-to-end solutions, that

are free to all. With the recent advances of various high-tech fields and the popularity of social networking, the trend of exponential growth of easily accessible information is certainly going to continue well into the twenty-first century, and the bottleneck created by this information explosion will definitely require innovative solutions from the scientific and engineering communities, particularly those technologists with better understanding of and a strong background in applied mathematics. In this regard, mathematics extends its influence and impact by providing innovative theory, methods, and algorithms to virtually every discipline, far beyond sciences and engineering, for processing, transmitting, receiving, understanding, and visualizing data sets, which could be very large or live in some high-dimensional space.

Of course the basic mathematical tools, such as PDE methods and least-squares approximation introduced by Gauss, are always among the core of the mathematical toolbox for applied mathematics. But other innovations and methods must be integrated in this toolbox as well. One of the most essential ideas is the notion of frequency of the data information. Joseph Fourier, a contemporary of Gauss, instilled this important concept to our study of physical phenomena by his innovation of trigonometric series representations, along with powerful mathematical theory and methods, to significantly expand the core of the toolbox for applied mathematics. The frequency content of a given data-set facilitates the processing and understanding of the data information. Another important idea is the “multi-scale” structure of data sets. Less than three decades ago, with the birth of another exciting mathematical subject, called “wavelets,” the data-set of information can be put in the wavelet domain for multi-scale processing as well. On the other hand, it is unfortunate that some essential basic mathematical tools for information processing are not commonly taught in a regular applied mathematics course in the university. Among the commonly missing ones, the topics that are addressed in this Saylor course MA304 include: information coding, data dimensionality reduction, data compression, and image manipulation.

The objective of this course is to study the basic theory and methods in the toolbox of the core of applied mathematics, with a central scheme that addresses “information processing” and with an emphasis on manipulation of digital image data. Linear algebra in the Saylor Foundation’s MA211 and MA212 are extended to “linear analysis” with applications to principal component analysis (PCA) and data dimensionality reduction (DDR). For data compression, the notion of entropy is introduced to quantify coding efficiency as governed by Shannon’s Noiseless Coding theorem. Discrete Fourier transform (DFT) followed by an efficient computational algorithm, called fast Fourier transform (FFT), as well as a real-valued version of the DFT, called discrete cosine transform (DCT) are discussed, with application to extracting frequency content of the given discrete data set that facilitates reduction of the entropy and thus significant improvement of the coding efficiency. DFT can be viewed as a discrete version of the Fourier series, which will be studied in some depth, with emphasis on orthogonal projection, the property of

positive approximate identity of Fejer's kernels, Parseval's identity and the concept of completeness. The integral version of the sequence of Fourier coefficients is called the Fourier transform (FT). Analogous to the Fourier series, the formulation of the inverse Fourier transform (IFT) is derived by applying the Gaussian function as sliding time-window for simultaneous time-frequency localization, with optimality guaranteed by the Uncertainty Principle. Local time-frequency basis functions are also introduced in this course by discretization of the frequency-modulated sliding time-window function at the integer lattice points. Replacing the frequency modulation by modulation with the cosines avoids the Balian-Low stability restriction on the local time-frequency basis functions, with application to elimination of blocky artifact caused by quantization of tiled DCT in image compression. Gaussian convolution filtering also provides the solution of the heat (partial differential) equation with the real-line as the spatial domain. When this spatial domain is replaced by a bounded interval, the method of separation of variables is applied to separate the PDE into two ordinary differential equations (ODEs). Furthermore, when the two end-points of the interval are insulated from heat loss, solution of the spatial ODE is achieved by finding the eigenvalue and eigenvector pairs, with the same eigenvalues to govern the exponential rate of decay of the solution of the time ODE. Superposition of the products of the spatial and time solutions over all eigenvalues solves the heat PDE, when the Fourier coefficients of the initial heat content are used as the coefficients of the terms of the superposition. This method is extended to the two-dimensional rectangular spatial domain, with application to image noise reduction. The method of separation of variables is also applied to solving other typical linear PDEs. Finally, multi-scale data analysis is introduced and compared with the Fourier frequency approach, and the architecture of multiresolution analysis (MRA) is applied to the construction of wavelets and formulation of the multi-scale wavelet decomposition and reconstruction algorithms. The lifting scheme is also introduced to reduce the computational complexity of these algorithms, with applications to digital image manipulation for such tasks as progressive transmission, image edge extraction, and image enhancement.

Portions of this manuscript are revised and modified versions of certain contents extracted from the book, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications," authored by Charles K. Chui and Qingtang Jiang. The book is published by Atlantis Press, and promoted, distributed and sold by Springer, both in print (ISBN 978-94-6239-008-9) and as e-book (ISBN 978-94-6239-009-6), available on Springers internet platform <http://www.springerlink.com>. The author of this text owns the copyright of the book, with signed agreement from his co-author, for non-commercial use and publication.

I would like to take this opportunity to acknowledge several individuals within the Saylor Foundation organization (<http://www.saylor.org>) for their encouragement, generous support, patience, and prompt responses, in the preparation of these contents. First and foremost, I would like to thank

Jennifer Shoop, who left Saylor less than a year ago to join MoneyThink (<http://www.moneythink.org>). It was Jen who suggested and initiated this project, which would not have happened without her enthusiastic encouragement and unfailing support. I am also indebted to Steve Phillips, who helped in finalizing the development contract and oversaw the initiation of this project, and to Tanner Huggins, who took over the responsibility from Steve a year ago in overseeing the progress of the content development. Tanners kind understanding, patience, and prompt responses certainly had a lot to do in keeping me going and finally completing this content development project. To my book publishers and friends, Zeger Karssen and Keith Jones, of Atlantis Press (<http://www.atlantis-press.com>), I would like to express my appreciation to their generous agreement with me to modify some of the contents from our book they published for the content development of MA 304. Finally, I am grateful to my wife, Margaret, for her assistance in typing and formatting the entire manuscript.

Charles Chui  
Menlo Park, California  
September, 2014

# *Unit 1*

## *LINEAR ANALYSIS*

The concepts and basic topics in elementary Linear Algebra and Calculus are reviewed and generalized to Linear Analysis in this first unit. In particular, the dot product, the vector space of  $\mathbb{R}^2$  (that is, the  $x - y$  plane in Plane Geometry), and measurement of lengths of vectors in  $\mathbb{R}^2$ , are extended to the “inner product,” the “inner-product space,” and “norm” measurement, respectively; the notion of eigenvalues and eigenvectors from Linear Algebra is extended to the “eigenvalue problem” of bounded linear operators; and eigenvalues are replaced by singular values for self-adjoint positive definite linear operators. Based on this preparation, the concept of principal component analysis (PCA) is introduced and studied in some detail, along with discussion of its applications to introduce the inverse of a singular matrix or an arbitrary rectangular matrix, minimum-norm least-squares estimation, and above all, to data dimensionality reduction.

### **1.1 Inner Product and Norm Measurements**

The notion of “inner product,” to be introduced in Subunit 1.1.1, is an extension of the “dot product, studied in beginning Vector Calculus. The most important property of the inner product is the Cauchy-Schwartz inequality, to be derived in Subunit 1.1.2. As an immediate application of this inequality, the notion of “norm,” defined in Subunit 1.1.1, is shown to satisfy the “triangle inequality in Subunit 1.1.3, which justifies the use of the norm for measuring the “lengths of elements in an “inner-product space. This allows us to extend beginning Vector Calculus to Linear Analysis of “Function and sequence Spaces. In addition, a combination of the inner-product and norm measurement can be applied to introduce the notion of “angles among elements (such as functions and sequences) of an inner-product space. This is a topic of discussion in Subunit 1.1.3. Furthermore, the Gram-Schmidt orthogonalization process is introduced and studied in Subunit 1.1.4 for changing a linearly independent set of elements (called vectors) in an inner-product space to a mutually orthogonal set of unit vectors.

### 1.1.1 Definition of inner product

The notion of inner product defined on a vector space, as introduced in Subunit 1.1.1, gives a very rich mathematical structure for the vectors in the space. Endowed with this inner product, the vector space will be called an inner-product space, as follows.

**Definition 1.1.1** Let  $\mathbb{V}$  be a vector space over the scalar field of complex numbers  $\mathbb{C}$ . Then  $\mathbb{V}$  is called an inner-product space, if there is a function  $\langle \cdot, \cdot \rangle$  defined on  $\mathbb{V} \times \mathbb{V}$ , with range in  $\mathbb{C}$ , that satisfies the following conditions:

- (a) Conjugate symmetry:  $\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$  for all  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ ;
- (b) Linearity:  $\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle$  for all  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{V}$ ,  $a, b \in \mathbb{C}$ ;
- (c) Positivity:  $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$  for all  $\mathbf{x} \in \mathbb{V}$ , and  $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ .

We remark that the above definition remains valid if the scalar field  $\mathbb{C}$  is replaced by the scalar field  $\mathbb{R}$  of real numbers, as demonstrated in the following example.

**Example 1.1.1** Let  $\mathbb{R}$  denote the set of real numbers and  $n$  any positive integer. Recall that the Euclidean space  $\mathbb{R}^n$  of  $n$ -tuples  $\mathbf{x} = (x_1, \dots, x_n)$ , where  $x_1, \dots, x_n \in \mathbb{R}$ , is a vector space over the scalar field  $\mathbb{R}$ , and that for any  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{y} = (y_1, \dots, y_n)$  in  $\mathbb{R}^n$ , the “dot product” of  $\mathbf{x}$  and  $\mathbf{y}$  is defined by

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \dots + x_n y_n. \quad (1.1.1)$$

Verify that the Euclidean space  $\mathbb{R}^n$  with the inner product  $\langle \cdot, \cdot \rangle$  defined by the dot product, namely,  $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y}$ , as in (1.1.1), is an inner-product space.

**Solution** We verify (a)–(c) of Definition (1.1.1), as follows:

- (a) For  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , since the complex conjugate of a real number is the real number itself, we have

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle &= x_1 \overline{y_1} + \dots + x_n \overline{y_n} \\ &= y_1 x_1 + \dots + y_n x_n \\ &= y_1 \overline{x_1} + \dots + y_n \overline{x_n} = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}. \end{aligned}$$

- (b) For all  $a, b \in \mathbb{R}$  and  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ , with  $\mathbf{z} = (z_1, \dots, z_n)$ , we have

$$\begin{aligned} \langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle &= (ax_1 + by_1)z_1 + \dots + (ax_n + by_n)z_n \\ &= (ax_1 z_1 + \dots + ax_n z_n) + (by_1 z_1 + \dots + by_n z_n) \\ &= a\langle \mathbf{x}, \mathbf{z} \rangle + b\langle \mathbf{y}, \mathbf{z} \rangle. \end{aligned}$$

(c) For any  $\mathbf{x} \in \mathbb{R}^n$ , we have

$$\langle \mathbf{x}, \mathbf{x} \rangle = |x_1|^2 + \cdots + |x_n|^2 \geq 0,$$

and that  $\langle \mathbf{x}, \mathbf{x} \rangle = 0$  if and only if  $|x_1|^2 + \cdots + |x_n|^2 = 0$ , or  $x_1 = 0, \dots, x_n = 0$ , or  $\mathbf{x} = \mathbf{0}$ . ■

Since the inner product of a vector  $\mathbf{x}$  with itself is a non-negative real number, we may use its square-root to define the length of the vector, called the norm of  $\mathbf{x}$ , as follows.

**Definition 1.1.2** For any vector  $\mathbf{x}$  in an inner-product space  $\mathbb{V}$  over the field of real or complex numbers, the norm of  $\mathbf{x}$ , induced by the inner product, is defined by

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}.$$

### 1.1.2 Cauchy-Schwarz Inequality

The key property of an inner-product space is the following Cauchy-Schwarz inequality that governs the size of the inner product of two vectors by the product of their norms.

**Theorem 1.1.1** Let  $\mathbb{V}$  be an inner-product space over the scalar field  $\mathbb{F} = \mathbb{C}$  or its subfield  $\mathbb{R}$ . Then for all  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ ,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|, \quad (1.1.2)$$

where  $\|\cdot\|$  is defined by (1.1.2). Furthermore, equality in (1.1.2) holds, if and only if

$$\mathbf{x} = c\mathbf{y}, \text{ or } \mathbf{y} = c\mathbf{x}$$

for some scalar (also called constant)  $c \in \mathbb{F}$ .

**Proof** We only prove this theorem for  $\mathbb{F} = \mathbb{R}$  and leave the proof for  $\mathbb{F} = \mathbb{C}$  as an exercise. Let  $a \in \mathbb{R}$  be any constant. We compute, according to (1.1.1),

$$\begin{aligned} 0 \leq \|\mathbf{x} - a\mathbf{y}\|^2 &= \langle \mathbf{x} - a\mathbf{y}, \mathbf{x} - a\mathbf{y} \rangle \\ &= \langle \mathbf{x} - a\mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x} - a\mathbf{y}, -a\mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle - a\langle \mathbf{y}, \mathbf{x} \rangle - a\langle \mathbf{x}, \mathbf{y} \rangle + a^2\langle \mathbf{y}, \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle - 2a\langle \mathbf{x}, \mathbf{y} \rangle + a^2\langle \mathbf{y}, \mathbf{y} \rangle \\ &= \|\mathbf{x}\|^2 - 2a\langle \mathbf{x}, \mathbf{y} \rangle + a^2\|\mathbf{y}\|^2. \end{aligned} \quad (1.1.3)$$

If  $\mathbf{y} = \mathbf{0}$ , then the theorem trivially holds. So we may assume  $\mathbf{y} \neq \mathbf{0}$ . Then by setting

$$a = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2},$$

we have

$$\begin{aligned} 0 &\leq \|\mathbf{x}\|^2 - 2 \frac{\langle \mathbf{x}, \mathbf{y} \rangle \langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2} + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \|\mathbf{y}\|^2}{\|\mathbf{y}\|^4} \\ &= \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2} \end{aligned}$$

or

$$0 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 - |\langle \mathbf{x}, \mathbf{y} \rangle|^2,$$

which is the same as (1.1.2). Moreover, the equality in (1.1.2) holds if and only if

$$0 = \|\mathbf{x} - a\mathbf{y}\|^2$$

in (1.1.3) with  $a = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2}$ , or  $\mathbf{x} = a\mathbf{y}$ . ■

Next, we extend the Euclidean space  $\mathbb{R}^n$  to the set  $\ell_2$  of infinite sequences or bi-infinite sequences defined as follows.

**Definition 1.1.3** *The set of infinite sequences,  $\mathbf{x} = \{x_j\}$ , where  $x_j \in \mathbb{C}$  and  $j$  runs from 0 to  $\infty$ , (or bi-infinite sequences, where  $j$  runs from  $-\infty$  to  $\infty$ ), that satisfy*

$$\sum_j |x_j|^2 < \infty,$$

*will be denoted by  $\ell_2$ .*

On some occasions, the same notation  $\ell_2$  will also denote the set of real-valued sequences, with  $\mathbb{C}$  in the above definition replaced by  $\mathbb{R}$ . As an application of the Cauchy-Schwarz inequality (1.1.2) in the above theorem, we will show that  $\ell_2$  is a vector space, and in fact an inner-product space, by extending the dot product of the Euclidean space in Example (1.1.1) to the inner product, as follows.

**Definition 1.1.4** *The set  $\ell_2$  of infinite sequences, endowed with the inner product*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_j x_j \bar{y}_j, \quad (1.1.4)$$

*defined for  $\mathbf{x} = \{x_j\}$  and  $\mathbf{y} = \{y_j\}$  in  $\ell_2$ , is called the  $\ell_2$  inner-product sequence space.*

Observe that since the definition of  $\langle \mathbf{x}, \mathbf{y} \rangle$  in (1.1.4) satisfies the three conditions in Definition 1.1.1, it follows from Theorem 1.1.1 that

$$|\langle \mathbf{x}, \mathbf{y} \rangle| = \left| \sum_j x_j \bar{y}_j \right| \leq \left( \sum_j |x_j|^2 \right)^{1/2} \left( \sum_j |y_j|^2 \right)^{1/2} = \|\mathbf{x}\| \|\mathbf{y}\|, \quad (1.1.5)$$



which is finite. Hence, (1.1.4) is well-defined for all  $\mathbf{x}, \mathbf{y} \in \ell_2$ , so that  $\ell_2$  is indeed an inner-product (vector) space. We will call (1.1.5) the Cauchy-Schwarz inequality for the inner-product (sequence) space  $\ell_2$ .

Analogously, we may introduce the inner-product space of functions defined on an interval  $J$ , which may be bounded or unbounded, such as the entire real-line  $\mathbb{R}$ . In the following, the reader, who might not be familiar with Lebesgue integration, may consider the integral of piecewise continuous functions defined on  $J$ . As in the above discussion of the sequence space  $\ell_2$ , we first introduce the set  $L_2$  of square-integrable functions, as follows.

**Definition 1.1.5** *The set of functions  $f$ , defined on the interval  $J$ , that satisfy*

$$\int_J |f(x)|^2 dx < \infty,$$

*will be denoted by  $L_2$ .*

For the set of square-integrable functions  $f, g$ , we introduce the operation  $\langle f, g \rangle$  to be defined below, and observe that it is finite.

**Definition 1.1.6** *For  $f, g \in L_2$ , set*

$$\langle f, g \rangle = \int_J f(x) \overline{g}(x) dx. \quad (1.1.6)$$

Since the definition of  $\langle f, g \rangle$  in (1.1.6) satisfies the three conditions in Definition 1.1.1, it follows from Theorem 1.1.1 that

$$|\langle f, g \rangle| = \left| \int_J f(x) \overline{g}(x) dx \right| \leq \left( \int_J |f(x)|^2 dx \right)^{1/2} \left( \int_J |g(x)|^2 dx \right)^{1/2} = \|f\| \|g\|, \quad (1.1.7)$$

which is finite. Hence, (1.1.6) is well-defined for all  $f, g \in L_2$ , so that  $L_2$  is indeed an inner-product (vector) space. We will call (1.1.7) the Cauchy-Schwarz inequality for the  $L_2$  inner-product (function) space.

### 1.1.3 Norm measurement and angle between vectors

#### References

- (1) Marcus Pivato, "Linear Partial Differential Equations and Fourier Theory 6A: Inner Products, Cambridge University Press.
- (2) Isaiah Lankham, Bruno Nachtergaele, and Anne Schilling, "Linear Algebra: As an Introduction to Abstract Mathematics, University of California, Davis.

- (3) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 28. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

#### 1.1.4 Gram-Schmidt orthogonalization process

##### References

- (1) Gilbert Strang, “Linear Algebra Lecture 17: Orthogonal Matrices and Gram-Schmidt (YouTube).
- (2) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 48–49. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

## 1.2 Eigenvalue Problems

In Linear Algebra, it is shown that multiplication of an  $m \times n$  matrix  $A$  to an  $n$ -dimensional vector  $\mathbf{x}$  yields an  $m$ -dimensional vector  $\mathbf{y}$ , for arbitrary positive integers  $m$  and  $n$ . This operation has the linearity property, in that

$$A(a_1\mathbf{x}_1 + a_2\mathbf{x}_2) = a_1A\mathbf{x}_1 + a_2A\mathbf{x}_2,$$

for all arbitrary  $n$ -dimensional vectors  $\mathbf{x}_1, \mathbf{x}_2$  and constants  $a_1, a_2$ . The operation of matrix-to-vector multiplication is extended to the notion of linear transformation in this subunit. In particular, the concepts of bounded linear functionals and bounded linear operators are introduced and discussed in some details in Subunit 1.2.2. Furthermore, the notion of “adjoints of linear transformations and that of self-adjoint operators are introduced in this subunit. An important extension of square matrices to linear operators is the eigenvalue problem. In Subunit 1.2.3, the topic of eigenvalues and eigenspace of linear operators is studied; and in Subunit 1.2.4, special properties of the eigenvalues of self-adjoint operators are derived.

### 1.2.1 Linear Transformations

Multiplication of an  $m \times n$  matrix  $A$  to a column vector  $\mathbf{x} \in \mathbb{C}^n$  results in a column vector  $\mathbf{z} \in \mathbb{C}^m$ . Hence, the matrix  $A$  can be considered as a

transformation from the Euclidean space  $\mathbb{C}^n$  to the Euclidean space  $\mathbb{C}^m$ . This transformation has the important property that

$$A(a\mathbf{x} + b\mathbf{y}) = aA\mathbf{x} + bA\mathbf{y}$$

for all vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$  and all scalars  $a, b \in \mathbb{C}$ . To be called the linearity property of the matrix  $A$ , this concept extends to transformations defined on finite or infinite-dimensional vector spaces, including fairly general differential and integral operators on certain appropriate subspaces of the function spaces  $\tilde{L}_2$  as studied in Subunit 1.1.2.

**Definition 1.2.1** *Let  $\mathbb{V}$  and  $\mathbb{W}$  be two vector spaces over the scalar field  $\mathbb{C}$  (or in some cases the field  $\mathbb{R}$  of real numbers). A transformation  $T$  from  $\mathbb{V}$  to  $\mathbb{W}$  is said to be linear, if it satisfies:*

$$T(a\mathbf{x} + b\mathbf{y}) = aT\mathbf{x} + bT\mathbf{y} \quad (1.2.1)$$

for all  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$  and  $a, b \in \mathbb{C}$ .

It follows from (1.2.1) that a linear transformation  $T$  satisfies

$$T(\mathbf{x} + \mathbf{y}) = T\mathbf{x} + T\mathbf{y}; \quad (1.2.2)$$

$$T(a\mathbf{x}) = aT\mathbf{x} \quad (1.2.3)$$

for all  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$  and  $a \in \mathbb{C}$ . Conversely, (1.2.1) follows from (1.2.2) and (1.2.3) as well.

## 1.2.2 Bounded linear functionals and operators

Observe that the scalar field  $\mathbb{C}$  (or  $\mathbb{R}$ ) can be considered as a vector space over the field itself. In Definition 1.2.1, if the vector space  $\mathbb{W}$  is chosen to be  $\mathbb{C}$  (or  $\mathbb{R}$ ), the linear transformation  $T$  from a vector space  $\mathbb{V}$  to  $\mathbb{C}$  (or  $\mathbb{R}$ ) is called a linear functional. Also, if the vector space  $\mathbb{W}$  is chosen to be  $\mathbb{V}$ , then the linear transformation  $T$  from a vector space  $\mathbb{V}$  to  $\mathbb{V}$  itself is called a linear operator. If the vector spaces  $\mathbb{V}$  and  $\mathbb{W}$  are both inner-product spaces, such as the sequence space  $\ell_2$  or the function space  $L_2$ , then a linear transformation  $T$  from  $\mathbb{V}$  to  $\mathbb{W}$  is said to be bounded, provided that the norm of  $T\mathbf{x}$ , induced by the inner product of the inner-product space  $\mathbb{W}$ , is uniformly bounded for all unit vectors  $\mathbf{x} \in \mathbb{V}$ . For bounded linear transformations, we introduce the notion of operator norm, as follows.

**Definition 1.2.2** *Let  $\mathbb{V}$  and  $\mathbb{W}$  be two inner-product spaces (over the scalar field  $\mathbb{C}$  or  $\mathbb{R}$ ) with norms induced by the inner products and denoted by  $\|\cdot\|_{\mathbb{V}}$  and  $\|\cdot\|_{\mathbb{W}}$ , respectively. Then the norm of a linear transformation  $T$  from  $\mathbb{V}$  to  $\mathbb{W}$  is defined by*

$$\|T\|_{\mathbb{V} \rightarrow \mathbb{W}} = \sup \left( \frac{\|T\mathbf{x}\|_{\mathbb{W}}}{\|\mathbf{x}\|_{\mathbb{V}}} : \mathbf{0} \neq \mathbf{x} \in \mathbb{V} \right). \quad (1.2.4)$$

If  $\|T\|_{\mathbb{V} \rightarrow \mathbb{W}}$  is finite, then the transformation  $T$  is said to be a bounded linear transformation.

For convenience, if  $\mathbb{W} = \mathbb{C}$  (that is, for bounded linear functionals  $T$ ), we adopt the abbreviated notation

$$\|T\| = \|T\|_{\mathbb{V} \rightarrow \mathbb{C}}.$$

Also, if  $\mathbb{W} = \mathbb{V}$  (that is, for bounded linear operators  $T$  on  $\mathbb{V}$ ), we adopt the abbreviated notation

$$\|T\|_{\mathbb{V}} = \|T\|_{\mathbb{V} \rightarrow \mathbb{V}}.$$

**Example 1.2.1** Let  $\mathbb{V} = L_2^c(J)$  be the subspace of the inner-product space  $L_2 = L_2(J)$ , consisting only of continuous functions on a compact (that is, closed and bounded) interval  $J$  on the real-line  $\mathbb{R}$ . Then

$$Tf = \int_J f \quad (1.2.5)$$

is a bounded linear functional on  $\mathbb{V}$ .

For linear functionals, the following example is very general and useful in applications.

**Example 1.2.2** Let  $\mathbb{V}$  be an inner-product space with inner product  $\langle \cdot, \cdot \rangle$ , and let  $\mathbf{x}_0 \in \mathbb{V}$  be arbitrarily chosen. Then

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_0 \rangle, \quad \mathbf{x} \in \mathbb{V},$$

is a bounded linear functional from  $\mathbb{V}$  to  $\mathbb{C}$  with  $\|T\| = \|\mathbf{x}_0\|$ , where  $\|\cdot\|$  is the norm induced by the inner product of  $\mathbb{V}$ .

The linearity of  $T$  follows from the second property of the inner product, and the uniform boundedness of  $\|T\mathbf{x}\| = |\langle \mathbf{x}, \mathbf{x}_0 \rangle|$ , for all vectors  $\mathbf{x}$  in  $\mathbb{V}$  with  $\|\mathbf{x}\| = 1$ , is a consequence of the Cauchy-Schwarz inequality. In addition, this inequality assures that  $\|\mathbf{x}_0\|$  is an upper bound of  $\|T\|$ . On the other hand, by choosing  $\mathbf{x} = \mathbf{x}_0$ , it follows from Definition 1.2.2 that

$$\begin{aligned} \|\mathbf{x}_0\|^2 &= |\langle \mathbf{x}_0, \mathbf{x}_0 \rangle| = \|T\mathbf{x}_0\| = \left( \frac{\|T\mathbf{x}_0\|}{\|\mathbf{x}_0\|} \right) \|\mathbf{x}_0\| \\ &\leq \sup \left( \frac{\|T\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{0} \neq \mathbf{x} \in \mathbb{V} \right) \|\mathbf{x}_0\| = (\|T\|) \|\mathbf{x}_0\|. \end{aligned}$$

Hence,  $\|\mathbf{x}_0\|$  is also a lower bound of  $\|T\|$ , so that  $\|T\| = \|\mathbf{x}_0\|$ .

An inner-product space  $\mathbb{V}$  is said to be complete, if there exists a (finite or countably infinite) set  $\{\mathbf{v}_k, k = 1, 2, \dots\}$ , such that every  $\mathbf{x} \in \mathbb{V}$  can be represented as a (finite or countably infinite) linear combination of  $\{\mathbf{v}_k, k =$

$1, 2, \dots\}$ . In addition, if for every  $\mathbf{x} \in \mathbb{V}$ , such a linear combination is a unique representation of  $\mathbf{x}$ , then the set  $\{\mathbf{v}_k, k = 1, 2, \dots\}$  is called a basis of  $\mathbb{V}$ . It is easy to show that for a complete inner-product space, a basis is a linear independent set of vectors.

The converse of Example 1.2.2 is also valid by employing a pair of dual bases (such as an orthonormal basis), defined as follows.

**Definition 1.2.3** *Let  $\mathbb{V}$  be a complete inner-product space. Also let  $\{\mathbf{v}_k, k = 1, 2, \dots\}$  and  $\{\tilde{\mathbf{v}}_k, k = 1, 2, \dots\}$  be two bases of  $\mathbb{V}$ . Then these two bases are said to constitute a dual pair (or a pair of dual bases of  $\mathbb{V}$ ), if*

$$\langle \mathbf{v}_k, \tilde{\mathbf{v}}_j \rangle = \delta_{j-k}, \quad j, k = 1, 2, \dots \quad (1.2.6)$$

*In addition, if  $\{\mathbf{v}_k, k = 1, 2, \dots\}$  is self-dual (that is, the dual pair can be so chosen that  $\{\mathbf{v}_k, k = 1, 2, \dots\} = \{\tilde{\mathbf{v}}_k, k = 1, 2, \dots\}$ ), then  $\{\mathbf{v}_k, k = 1, 2, \dots\}$  is called an orthonormal basis of the inner-product space  $\mathbb{V}$ .*

**Theorem 1.2.1** *Let  $\mathbb{V}$  be a complete inner-product space with dual bases  $\{\mathbf{v}_k\}$  and  $\{\tilde{\mathbf{v}}_k\}$ , and let  $T$  be a bounded linear functional on  $\mathbb{V}$ , such that  $\mathbf{x}_T$ , defined by*

$$\mathbf{x}_T = \sum_j \overline{(T\mathbf{v}_j)} \tilde{\mathbf{v}}_j, \quad (1.2.7)$$

*is in  $\mathbb{V}$ . Then  $T: \mathbb{V} \rightarrow \mathbb{F}$  can be formulated as:*

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_T \rangle \quad \text{for all } \mathbf{x} \in \mathbb{V}. \quad (1.2.8)$$

*Furthermore,  $\mathbf{x}_T$  in (1.2.7), called the representer of the linear functional  $T$ , is unique.*

**Proof** To derive the representation (1.2.8) of  $T\mathbf{x}$  for each  $\mathbf{x} \in \mathbb{V}$ , write

$$\mathbf{x} = \sum_k c_k \mathbf{v}_k.$$

Then by the linearity property, we have

$$T\mathbf{x} = \sum_k c_k T\mathbf{v}_k. \quad (1.2.9)$$

On the other hand, again by linearity and the definition of  $\mathbf{x}_T$  in (1.2.7), we also have

$$\begin{aligned} \langle \mathbf{x}, \mathbf{x}_T \rangle &= \sum_k \left\langle c_k \mathbf{v}_k, \sum_j \overline{T\mathbf{v}_j} \tilde{\mathbf{v}}_j \right\rangle \\ &= \sum_k c_k \sum_j \overline{T\mathbf{v}_j} \langle \mathbf{v}_k, \tilde{\mathbf{v}}_j \rangle \\ &= \sum_k c_k \sum_j T\mathbf{v}_j \delta_{k-j} = \sum_k c_k T\mathbf{v}_k. \end{aligned}$$

Hence, the representation formula in (1.2.8) is established.

To prove that the representer  $\mathbf{x}_T$  of  $T$  in (1.2.7) is unique, let  $\mathbf{y}_0 \in \mathbb{V}$  be another representer of  $T$ . Then

$$T\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_T \rangle = \langle \mathbf{x}, \mathbf{y}_0 \rangle, \text{ for all } \mathbf{x} \in \mathbb{V}$$

so that

$$\langle \mathbf{x}, \mathbf{x}_T - \mathbf{y}_0 \rangle = 0.$$

By choosing  $\mathbf{x} = \mathbf{x}_T - \mathbf{y}_0$ , we have

$$\|\mathbf{x}_T - \mathbf{y}_0\|^2 = 0,$$

or  $\mathbf{y}_0 = \mathbf{x}_T$ . ■

We next turn our attention to the study of bounded linear operators  $T$  that transform vectors  $\mathbf{x}$  in an inner-product space  $\mathbb{V}$  to vectors  $\mathbf{y}$  in the same space  $\mathbb{V}$ . Hence, it should be of great interest (and definitely very useful for applications) to study the existence and uniqueness of the linear operator  $T^*$ , corresponding to a given bounded linear operator  $T$ , such that  $T^*$  takes  $\mathbf{y}$  back to  $\mathbf{x}$ , in such a manner that the value of the inner product of  $T\mathbf{x}$  with  $\mathbf{y}$ , for any  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ , is preserved by that of the inner product of  $\mathbf{x}$  with  $T^*\mathbf{y}$ . This is the concept of adjoints of bounded linear operators. In this regard, it is important to point out that although any square matrix  $A$  is a bounded linear operator on the Euclidean space  $\mathbb{V} = \mathbb{C}^n$ , the notion of the adjoint  $A^*$  of  $A$ , as an operator, is different from the definition of the matrix adjoint of  $A$ , when  $A$  is considered as a matrix, for the formulation of matrix inverses in an elementary course of Matrix Theory.

**Theorem 1.2.2** *Let  $\mathbb{V}$  be a complete inner-product space over the scalar field  $\mathbb{C}$  (or  $\mathbb{R}$ ), such that at least a dual pair of bases for  $\mathbb{V}$  exists. Then corresponding to any bounded linear operator  $T$  on  $\mathbb{V}$ , there exists a linear operator  $T^*$ , also defined on  $\mathbb{V}$ , that satisfies the property*

$$\langle T\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, T^*\mathbf{y} \rangle, \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{V}. \quad (1.2.10)$$

*Furthermore, the linear operator  $T^*$ , called the adjoint of  $T$ , is uniquely determined by (1.2.10).*

**Proof** In view of Theorem 1.2.1, we observe that for any fixed  $\mathbf{y} \in \mathbb{V}$ , the linear functional  $L = L_{\mathbf{y}}$ , defined by

$$L_{\mathbf{y}}\mathbf{x} = \langle T\mathbf{x}, \mathbf{y} \rangle, \quad (1.2.11)$$

is bounded, and hence, has a unique representer  $\mathbf{x}_{L_{\mathbf{y}}}$ , in that  $L_{\mathbf{y}}\mathbf{x} = \langle \mathbf{x}, \mathbf{x}_{L_{\mathbf{y}}} \rangle$  for all  $\mathbf{x} \in \mathbb{V}$ . Since  $\mathbf{x}_{L_{\mathbf{y}}}$  is a function of  $\mathbf{y}$ , we may write  $\mathbf{x}_{L_{\mathbf{y}}} = F(\mathbf{y})$ , so that

$$\langle T\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, F(\mathbf{y}) \rangle \quad (1.2.12)$$

for all  $\mathbf{x} \in \mathbb{V}$ , where (1.2.12) holds for all  $\mathbf{y} \in \mathbb{V}$ . Next, let us verify that  $F$  is a linear operator on  $\mathbb{V}$ . From its definition,  $F$  is a transformation from  $\mathbb{V}$  into itself. That  $F$  is a linear operator on  $\mathbb{V}$  follows from the linearity of the inner product. Indeed, for any fixed  $\mathbf{y}, \mathbf{z} \in \mathbb{V}$  and fixed  $a, b \in \mathbb{C}$  (or  $a, b \in \mathbb{R}$ ), it follows from (1.2.12) that, for all  $\mathbf{x} \in \mathbb{V}$ ,

$$\begin{aligned}\langle \mathbf{x}, F(a\mathbf{y} + b\mathbf{z}) \rangle &= \langle T\mathbf{x}, a\mathbf{y} + b\mathbf{z} \rangle \\ &= \langle T\mathbf{x}, a\mathbf{y} \rangle + \langle T\mathbf{x}, b\mathbf{z} \rangle \\ &= \bar{a}\langle T\mathbf{x}, \mathbf{y} \rangle + \bar{b}\langle T\mathbf{x}, \mathbf{z} \rangle \\ &= \bar{a}\langle \mathbf{x}, F(\mathbf{y}) \rangle + \bar{b}\langle \mathbf{x}, F(\mathbf{z}) \rangle \\ &= \langle \mathbf{x}, aF(\mathbf{y}) \rangle + \langle \mathbf{x}, bF(\mathbf{z}) \rangle,\end{aligned}$$

so that  $\langle \mathbf{x}, F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z})) \rangle = 0$  for all  $\mathbf{x} \in \mathbb{V}$ . By setting  $\mathbf{x} = F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z}))$ , we have

$$\|F(a\mathbf{y} + b\mathbf{z}) - (aF(\mathbf{y}) + bF(\mathbf{z}))\|^2 = 0;$$

that is,  $F(a\mathbf{y} + b\mathbf{z}) = aF(\mathbf{y}) + bF(\mathbf{z})$ . Hence, by setting  $T^* = F$ , we have derived (1.2.10). Furthermore, since the representer in Theorem 1.2.1 is unique, we may conclude that  $T^* = F$  is unique. ■

**Example 1.2.3** Consider an  $n \times n$  matrix  $A \in \mathbb{C}^{n,n}$  as a linear operator on the vector space  $\mathbb{V} = \mathbb{C}^n$ . Determine the adjoint  $A^*$  of  $A$ .

**Solution** For any  $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ , consider  $\mathbf{x}$  and  $\mathbf{y}$  as column vectors:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

so that  $\mathbf{x}^T = [x_1, \dots, x_n]$  and  $\mathbf{y}^T = [y_1, \dots, y_n]$  are row vectors (where the superscript  $T$  denotes, as usual, the transpose of the matrix). Hence, it follows from the definition of the inner product for  $\mathbb{C}^n$  that

$$\begin{aligned}\langle A\mathbf{x}, \mathbf{y} \rangle &= (A\mathbf{x})^T \bar{\mathbf{y}} = \mathbf{x}^T A^T \bar{\mathbf{y}} \\ &= \mathbf{x}^T \overline{(\bar{A})^T \mathbf{y}} = \langle \mathbf{x}, (\bar{A})^T \mathbf{y} \rangle,\end{aligned}$$

so that from the uniqueness of the adjoint  $A^*$  of  $A$ , we have

$$A^* = (\bar{A})^T. \quad (1.2.13)$$

In other words, the adjoint of  $A$ , as an operator, is the transpose-conjugate  $A^*$  of  $A$ . ■

As pointed out above, the operator adjoint is different from the matrix adjoint, for the purpose of formulating matrix inverses, as follows.

**Remark 1.2.1** In Matrix Theory, the adjoint (which we may call matrix adjoint) of a square matrix  $A = [a_{j,k}]_{1 \leq j,k \leq n}$  is the matrix

$$\left( [A_{j,k}]_{1 \leq j,k \leq n} \right)^T,$$

where  $A_{j,k}$  denotes the cofactor of  $a_{j,k}$ . Hence in general, the matrix adjoint of  $A$  is different from the operator adjoint  $A^* = (\overline{A})^T$ , when  $A$  is considered as an operator on the vector space  $\mathbb{C}^n$ . ■

**Definition 1.2.4** A linear operator  $T$  is said to be self-adjoint if  $T^* = T$ .

**Remark 1.2.2** Recall that in an elementary course on Linear Algebra or Matrix Theory, a square matrix  $A \in \mathbb{C}^{n,n}$  is said to be Hermitian, if  $A = (\overline{A})^T$ . Thus, in view of (1.2.13), when considered as a linear operator, a square matrix  $A \in \mathbb{C}^{n,n}$  is self-adjoint if and only if it is Hermitian. Henceforth, for linear operators  $T$ , which may not be square matrices, we say that  $T$  is Hermitian if it is self-adjoint. Clearly, if a matrix  $A \in \mathbb{C}^{n \times n}$  is Hermitian, then all the diagonal entries of  $A$  are real. It will be shown in the next two subunits, 1.2.3 and 1.2.4, that all eigenvalues of self-adjoint operators are real in general. ■

### 1.2.3 Eigenvalues and eigenspaces

#### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 30: Linear Transformations and Their Matrices (YouTube), presented by Gilbert Strang.

### 1.2.4 Self-adjoint positive definite operators

#### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 6: Eigen Values (Part 2) and Positive Definite (Part 1) (YouTube), presented by Gilbert Strang.
- (2) MIT: Department of Computational Science and Engineering's "Lecture 27: Positive Definite Matrices (YouTube), presented by Gilbert Strang.



### 1.3 Singular Value Decomposition (SVD)

In elementary Linear Algebra, a square matrix  $A$  is diagonalizable if there exists a nonsingular matrix  $V$  such that  $V^{-1}AV = D$ , where  $D$  is a diagonal matrix. More precisely, an  $n \times n$  square matrix  $A$  is diagonalizable, if and only if it has  $n$  eigenvalues, counting multiplicities, that constitute a diagonal matrix  $D$  of dimension  $n$ , such that  $A = VDV^{-1}$ , where the  $n$  columns of  $V$  are the eigenvectors of  $A$  associated with the corresponding eigenvalues of  $A$ , in the order as listed in the diagonal of  $D$ . A more useful decomposition  $A = VDV^{-1}$  of the given matrix  $A$  is achieved, when the matrix  $V$  in the decomposition is an orthogonal matrix, for which  $V^{-1}$  is simply the transpose  $V^T$  of the real-valued matrix  $V$ ; and more generally, a unitary matrix, for which  $V^{-1}$  is simply the adjoint (that is, the complex conjugate of  $V^T$ ) of  $V$ , denoted by  $V^*$ . In Subunit 1.3.1, the notion of normal matrices (and in general, normal operators) is introduced and studied. In particular, the spectral decomposition,  $A = V^*DV$ , for normal operators  $A$ , is derived. To generalize this concept to matrices that are not necessarily normal, and to arbitrary  $m \times n$  matrices, where  $m$  and  $n$  are allowed to be different, the notion of singular values is introduced in Subunit 1.3.2. The extension of the spectral decomposition (of normal matrices) to allow the decomposition of arbitrary matrices, including all rectangular ones, is the notion of singular value decomposition (SVD). In Subunit 1.3.3, the reduced SVD (or SVD restricted to only non-zero singular values) is derived, and in Subunit 1.3.4, the SVD computation is extended to full SVD, including all singular values.

#### 1.3.1 Normal operators and spectral decomposition

##### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 30: Linear Transformations and Their Mtrices" (YouTube), presented by Gilbert Strang.
- (2) MIT: Department of Computational Science and Engineering's "Lecture 27: Positive Definite Matrices (YouTube), presented by Gilbert Strang.

### 1.3.2 Singular values

Let  $B \in \mathbb{C}^{m,n}$  be any  $m \times n$  matrix, and consider its corresponding Gram matrix  $A$ , defined by

$$A = BB^*, \quad (1.3.1)$$

where  $B^* = (\overline{B})^T$ . Then  $A$  is self-adjoint and positive semi-definite, and is therefore normal. Hence, according to the study in Subunit 1.3.1,  $A$  admits the following spectral decomposition:

$$A = U\Lambda U^*, \quad (1.3.2)$$

with diagonal matrix  $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_m\}$  and unitary matrix

$$U = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m],$$

where for each  $j = 1, \dots, m$ ,  $(\lambda_j, \mathbf{u}_j)$  is an eigenvalue-eigenvector pair of the matrix  $A$ . Furthermore, since  $A$  is self adjoint and positive semi-definite, we may write  $\lambda_j = \sigma_j^2$ , where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_m = 0 \quad (1.3.3)$$

for some  $r$ , with  $0 \leq r \leq m$ . Hence, the diagonal matrix  $\Lambda$  in the spectral decomposition (1.3.2) has the more precise formulation

$$\Lambda = \text{diag}\{\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0\}, \quad (1.3.4)$$

where we have adopted the standard convention that  $\{\sigma_{r+1}, \dots, \sigma_m\}$  is an empty set if  $r = m$ .

Furthermore, from elementary matrix theory, we have

$$\text{rank}(B) = \text{rank}(BB^*) = \text{rank}(\Lambda) = r,$$

so that  $r \leq \min\{m, n\}$ . Let

$$\Sigma_r = \text{diag}\{\sigma_1, \dots, \sigma_r\} \quad (1.3.5)$$

and consider the  $m \times n$  matrix

$$S = \begin{bmatrix} \Sigma_r & \vdots & O \\ \dots & \dots & \dots \\ O & \vdots & O \end{bmatrix}, \quad (1.3.6)$$

where  $O$  denotes the zero matrix (possibly with different dimensions), so that

$$S = \begin{bmatrix} \Sigma_n \\ \dots \\ O \end{bmatrix} \text{ or } S = \begin{bmatrix} \Sigma_m \vdots O \end{bmatrix} \quad (1.3.7)$$

if  $r = n < m$  or  $r = m < n$ , respectively. Observe that the diagonal matrix  $\Lambda$  in (1.3.4) can be written as

$$\Lambda = SS^T = SS^*, \quad (1.3.8)$$

and the spectral decomposition of  $A$  in (1.3.2) can be re-formulated as

$$A = USS^*U^* = (US)(US)^*. \quad (1.3.9)$$

In the following, we will study the existence of two unitary matrices  $U$  and  $V$ , of dimensions  $m \times m$  and  $n \times n$ , respectively, such that

$$B = USV^*. \quad (1.3.10)$$

To understand the factorization in (1.3.10), let us write

$$U = [\mathbf{u}_1, \dots, \mathbf{u}_m], \quad V = [\mathbf{v}_1, \dots, \mathbf{v}_n], \quad (1.3.11)$$

where  $\{\mathbf{u}_1, \dots, \mathbf{u}_m\}$  and  $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$  are orthonormal bases of  $\mathbb{C}^m$  and  $\mathbb{C}^n$ , respectively. Then it follows from (1.3.10) that  $BV = US$  and  $B^*U = VS^T$ , so that

(i) if  $n < m$ , then

$$\begin{aligned} B\mathbf{v}_j &= \sigma_j \mathbf{u}_j, \quad B^* \mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad \text{for } j = 1, \dots, n, \\ B^* \mathbf{u}_j &= \mathbf{0}, \quad \text{for } j = n+1, \dots, m; \end{aligned} \quad (1.3.12)$$

(ii) if  $n \geq m$ , then

$$\begin{aligned} B\mathbf{v}_j &= \sigma_j \mathbf{u}_j, \quad B^* \mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad \text{for } j = 1, \dots, m, \\ B\mathbf{v}_j &= \mathbf{0}, \quad \text{for } j = m+1, \dots, n. \end{aligned} \quad (1.3.13)$$

**Definition 1.3.1** In (1.3.10), the diagonal entries  $\sigma_1, \dots, \sigma_r$  of  $\Sigma_r$  in (1.3.6) are called the (non-zero) singular values of the matrix  $B$ , and the pair  $(\mathbf{v}_j, \mathbf{u}_j)$  of vectors in (1.3.12) or (1.3.13) is called a singular-vector pair associated with the singular value  $\sigma_j$ .

Clearly, if  $(\mathbf{v}_j, \mathbf{u}_j)$  is a singular-vector pair of  $B$  associated with  $\sigma_j$ , then  $(\mathbf{u}_j, \mathbf{v}_j)$  is a singular-vector pair of  $B^*$  associated with the same  $\sigma_j$ .

**Example 1.3.1** For the matrix

$$B_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

with  $m = 2$  and  $n = 3$ , verify that  $\sigma_1 = \sqrt{2}, \sigma_2 = 1$  are singular values of

$B_1$ , with corresponding singular-vector pairs  $(\mathbf{v}_1, \mathbf{u}_1)$ ,  $(\mathbf{v}_2, \mathbf{u}_2)$ , and  $B_1 \mathbf{v}_3 = \mathbf{0}$ , where

$$\begin{aligned}\mathbf{v}_1 &= \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}^T, \\ \mathbf{v}_2 &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T, \\ \mathbf{v}_3 &= \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}^T, \\ \mathbf{u}_1 &= \begin{bmatrix} 0 & 1 \end{bmatrix}^T \text{ and } \mathbf{u}_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T.\end{aligned}$$

**Solution**

(i) For  $\sigma_1 = \sqrt{2}$ ,

$$\begin{aligned}B_1 \mathbf{v}_1 &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix} \\ &= \sqrt{2} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \sigma_1 \mathbf{u}_1; \\ B_1^* \mathbf{u}_1 &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \\ &= \sqrt{2} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ -\frac{1}{\sqrt{2}} \end{bmatrix} = \sigma_1 \mathbf{v}_1;\end{aligned}$$

(ii) for  $\sigma_2 = 1$ ,

$$\begin{aligned}B_1 \mathbf{v}_2 &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \sigma_2 \mathbf{u}_2; \\ B_1^* \mathbf{u}_2 &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \sigma_2 \mathbf{v}_2;\end{aligned}$$

(iii) for  $\mathbf{v}_3$ ,

$$B_1 \mathbf{v}_3 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}.$$

Observe that  $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$  is an orthonormal basis of  $\mathbb{R}^3$  (and of  $\mathbb{C}^3$ ), and  $\{\mathbf{u}_1, \mathbf{u}_2\}$  is an orthonormal basis of  $\mathbb{R}^2$  (and of  $\mathbb{C}^2$ ). In addition, in this example  $2 = m < n = 3$ , and (1.3.13) holds with  $B = B_1$ . ■

### 1.3.3 Reduced singular value decomposition

In this subunit, we formulate and derive the spectral decomposition formula of any  $m \times n$  matrix  $B$  of complex numbers, only in terms of its non-zero singular values, as follows.

**Theorem 1.3.1** *Let  $B$  be an  $m \times n$  matrix with  $\text{rank}(B) = r$ . Then there exists an  $m \times r$  matrix  $U_1$  and an  $n \times r$  matrix  $V_1$ , with*

$$U_1^* U_1 = I_r, \quad V_1^* V_1 = I_r, \quad (1.3.14)$$

*such that  $B$  has the reduced SVD*

$$B = U_1 \Sigma_r V_1^*, \quad (1.3.15)$$

*where  $\Sigma_r = \text{diag}\{\sigma_1, \dots, \sigma_r\}$  with  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . Furthermore, if  $B$  is a real-valued matrix, then  $U_1$  and  $V_1$  in (1.3.15) can be chosen to be real-valued matrices.*

**Proof** To prove Theorem 1.3.1, let  $A = BB^*$ . Then  $A$  has the spectral decomposition (1.3.2) for some  $m \times m$  unitary matrix  $U$  and  $\Lambda = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}$  with

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_m = 0,$$

where  $0 \leq r \leq \min\{m, n\}$ . Furthermore, since  $\text{rank}(B) = \text{rank}(BB^*)$ , we have  $\text{rank}(B) = \text{rank}(A) = \text{rank}(\Lambda) = r$ .

Write  $U$  as  $U = [U_1 : U_2]$ , with  $U_1$  being the  $m \times r$  matrix consisting of the first  $r$  columns of  $U$ . Then  $U_1^* U_1 = I_r$  and  $U_1^* U_2 = O$ , where  $I_r$  denotes, as usual, the  $r \times r$  identity matrix. Observe from (1.3.2), that

$$U_1^* BB^* U_1 = U_1^* U \Lambda U^* U_1 = [I_r \ O] \Lambda [I_r \ O]^* = (\Sigma_r)^2. \quad (1.3.16)$$

Similarly, it can be shown that  $U_2^* BB^* U_2 = O$ , yielding

$$U_2^* B = O. \quad (1.3.17)$$

Next, we define  $V_1$  by

$$V_1 = B^* U_1 \Sigma_r^{-1}. \quad (1.3.18)$$

Then  $V_1$  satisfies (1.3.15). Indeed, from (1.3.17) and the definition of  $V_1$  in (1.3.18), we have

$$\begin{aligned} U^*(B - U_1 \Sigma_r V_1^*) &= \begin{bmatrix} U_1^* B \\ U_2^* B \end{bmatrix} - \begin{bmatrix} I_r \\ O \end{bmatrix} \Sigma_r V_1^* \\ &= \begin{bmatrix} U_1^* B - \Sigma_r V_1^* \\ O \end{bmatrix} = O. \end{aligned}$$

Thus, since  $U^*$  is nonsingular, we have  $B - U_1 \Sigma_r V_1^* = O$ ; that is,  $B = U_1 \Sigma_r V_1^*$ , as desired.

Furthermore, for real matrices  $B$ , the unitary matrix  $U$  in the spectral decomposition of  $A = BB^* = BB^T$  in (1.3.2) can be chosen to be an  $m \times m$  orthogonal matrix, and hence the matrix  $V_1$  defined by (1.3.18) is also real. ■

**Example 1.3.2** Let  $B = B_1$  in Example 1.3.1; that is,

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

Compute the reduced SVD of  $B$ .

**Solution** Since  $B^*B$  is  $3 \times 3$  and  $BB^*$  is  $2 \times 2$ , we compute the eigenvalues of one with lower dimension, namely:

$$BB^* = BB^T = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix},$$

by taking the determinant of  $\begin{bmatrix} 1 - \lambda & 0 \\ 0 & 2 - \lambda \end{bmatrix}$ , yielding the eigenvalues  $\sigma_1^2 = 2$  and  $\sigma_2^2 = 1$  (since they are arranged in decreasing order). Then the (non-zero) singular values of  $B$  are:

$$\sigma_1 = \sqrt{2}, \sigma_2 = 1.$$

To compute  $\mathbf{u}_1$  and  $\mathbf{u}_2$ , note that

$$BB^* - \sigma_j^2 I_2 = \begin{bmatrix} 1 - \sigma_j^2 & 0 \\ 0 & 2 - \sigma_j^2 \end{bmatrix}, \quad j = 1, 2;$$

that is,  $\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$  and  $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ . Hence, we may select

$$\mathbf{u}_1 = [0 \ 1]^T \quad \text{and} \quad \mathbf{u}_2 = [1 \ 0]^T,$$

yielding

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Since  $r = m = 2$ , in this case  $U_1 = U$ , and  $V_1$  as defined by (1.3.18) is simply

$$V_1 = B^* U_1 \Sigma_2^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}.$$

Hence, the reduced SVD of  $B$  is given by

$$B = U_1 \Sigma_2 V_1^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{bmatrix}.$$

■

**Example 1.3.3** Compute the reduced SVD of  $B$ , where

$$B = \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix}.$$

**Solution** Since  $B$  is  $4 \times 3$  with  $m = 4 > n = 3$ , we consider the SVD of  $B^*$  first and compute the spectral decomposition of  $A = (B^*)(B^*)^* = B^*B$ :

$$A = \begin{bmatrix} 1 & 1 & -1 & 0 \\ 0 & i & 0 & 1 \\ -i & 0 & -1 & -i \end{bmatrix} \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix} = \begin{bmatrix} 3 & -i & 1+i \\ i & 2 & i \\ 1-i & -i & 3 \end{bmatrix}.$$

To compute the eigenvalues of  $A$ , we evaluate the determinant of the matrix  $\lambda I_3 - A$  and factorize the characteristic polynomial, yielding

$$(\lambda - 2)(\lambda^2 - 6\lambda + 5) = (\lambda - 2)(\lambda - 5)(\lambda - 1),$$

so that  $\lambda_1 = 5, \lambda_2 = 2, \lambda_3 = 1$ , when arranged in the decreasing order. Hence, the (non-zero) singular values of  $B$  are

$$\sigma_1 = \sqrt{5}, \sigma_2 = \sqrt{2}, \sigma_3 = 1.$$

To compute the eigenvectors, we simply solve the three homogeneous linear systems  $(A - \lambda_j I_3)\mathbf{x} = \mathbf{0}, j = 1, 2, 3$ . After dividing each solution by its Euclidean norm, we obtain the normalized eigenvectors  $\mathbf{u}_j$  associated with  $\lambda_j$ , for  $j = 1, 2, 3$ , listed as follows:

$$\mathbf{u}_1 = \frac{1}{2\sqrt{6}} \begin{bmatrix} 1-3i \\ 2 \\ -1-3i \end{bmatrix}, \mathbf{u}_2 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \mathbf{u}_3 = \frac{1}{2} \begin{bmatrix} 1 \\ 1-i \\ i \end{bmatrix}.$$

Let  $U = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3]$ . Since  $r = 3$ , we have  $U_1 = U$  and  $\Sigma_3 = \text{diag}(\sqrt{5}, \sqrt{2}, 1)$ .

Applying (1.3.18), we have

$$\begin{aligned}
 V_1 &= (B^*)^* U_1 \Sigma_3^{-1} = BU \Sigma_3^{-1} \\
 &= \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix} \begin{bmatrix} \frac{1-3i}{2\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{2} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{1-i}{2} \\ \frac{-1-3i}{2\sqrt{6}} & -\frac{1}{\sqrt{3}} & \frac{i}{2} \end{bmatrix} \text{diag}\left(\frac{1}{\sqrt{5}}, \frac{1}{\sqrt{2}}, 1\right) \\
 &= \begin{bmatrix} \frac{2-2i}{\sqrt{30}} & \frac{1-i}{\sqrt{6}} & 0 \\ \frac{1-5i}{2\sqrt{30}} & \frac{1+i}{\sqrt{6}} & -\frac{i}{2} \\ \frac{3i}{\sqrt{30}} & 0 & -\frac{1+i}{2} \\ \frac{5-i}{2\sqrt{30}} & -\frac{1+i}{\sqrt{6}} & -\frac{i}{2} \end{bmatrix}.
 \end{aligned}$$

Thus, the reduced SVD for  $B^*$  is given by  $B^* = U \Sigma_3 V_1^*$ , from which we arrive at the following reduced SVD for  $B$ :

$$B = V_1 \Sigma_3 U^*.$$

■

### 1.3.4 Full singular value decomposition

When the zero singular values are also taken into consideration, we have the following full spectral decomposition.

**Theorem 1.3.2** *Let  $B$  be an  $m \times n$  matrix with  $\text{rank}(B) = r$ . Then there exist  $m \times m$  and  $n \times n$  unitary matrices  $U$  and  $V$ , respectively, such that*

$$B = USV^*, \quad (1.3.19)$$

where  $S$  is an  $m \times n$  matrix introduced in (1.3.6), with  $\Sigma_r$  in (1.3.5) given by  $\Sigma_r = \text{diag}\{\sigma_1, \dots, \sigma_r\}$ ,  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$ . Furthermore, if  $B$  is a real matrix, then the unitary matrices  $U$  and  $V$  in (1.3.19) can be chosen to be orthogonal matrices.

**Proof** To prove Theorem 1.3.2, let us again write  $A = BB^*$  and consider the matrix  $V_1$  defined by (1.3.18) so that (1.3.15) holds. Also let  $U = [U_1 : U_2]$ , where  $U_1$  and  $U_2$  are the matrices introduced in the proof of Theorem 1.3.1. Then by (1.3.18) and (1.3.16), we have

$$V_1^* V_1 = \Sigma_r^{-1} U_1^* B B^* U_1 \Sigma_r^{-1} = \Sigma_r^{-1} (\Sigma_r)^2 \Sigma_r^{-1} = I_r.$$



Hence, the columns of  $V_1$  constitute an orthonormal family in  $\mathbb{C}^n$ , and we may extend  $V_1$  to a unitary matrix  $V = [V_1 \ V_2]$  by introducing another matrix  $V_2$  with orthonormal column vectors. Thus, it follows from (1.3.15) that

$$B = U_1 \Sigma_r V_1^* = U_1 [\Sigma_r \ O] [V_1 \ V_2]^* = [U_1 \ U_2] \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} V^* = U S V^*,$$

completing the proof of (1.3.19).

Furthermore, for real matrices  $B$ , the unitary matrix  $U$  in the spectral decomposition (1.3.2) of  $A = BB^T$  can be chosen to be an orthogonal matrix. In addition, as already shown above, the columns of  $V_1$  defined by (1.3.18) constitute an orthonormal family of  $\mathbb{R}^n$ , so that  $V_1$  can be extended to an orthogonal matrix  $V \in \mathbb{R}^{n,n}$  that satisfies  $B = USV^T$ . ■

From a full SVD of  $B$  in (1.3.19), the construction of a reduced SVD (1.3.15) of  $B$  is obvious, simply by keeping only the first  $r$  columns of  $U$  and  $V$  to obtain  $U_1$  and  $V_1$ , respectively. Conversely, from a reduced SVD (1.3.15) of  $B$ , it is also possible to recover a full SVD of  $B$  in (1.3.19) by extending  $\Sigma_r$  to  $S$ , defined in (1.3.5), as well as extending  $U_1$  and  $V_1$  to unitary matrices  $U$  and  $V$ , respectively. In the literature, both the reduced SVD (1.3.15) and full SVD (1.3.19) are called the SVD of  $B$ .

**Remark 1.3.1** To compute the singular value decomposition (SVD) of a rectangular matrix  $B$ , the first step is to compute the eigenvalues of  $BB^*$ . Then the non-zero singular values of  $B$  are the positive square-roots of the non-zero eigenvalues of  $BB^*$ . The unitary matrix  $U$  in the SVD of  $B$  is the unitary matrix in the spectral decomposition of  $BB^*$ . It is important to emphasize that eigenvectors of  $BB^*$  associated with the same eigenvalue must be orthogonalized by applying the Gram-Schmidt process, and that all eigenvectors must be normalized to have unit norm. Of course  $U$  and  $V$  are not unique, although the singular values are unique. ■

**Remark 1.3.2** To compute the singular value decomposition (SVD) of a rectangular matrix  $B$  of dimension  $m \times n$  with  $n < m$ , the computational cost can be reduced by computing the spectral decomposition of the  $n \times n$  matrix  $B^*B$  instead of  $BB^*$ , which has larger dimension. To do so, simply replace  $B$  by  $B^*$  and consider  $A = (B^*)(B^*)^*$ . Hence, the reduced SVD and full SVD are given by  $B^* = U_1 \Sigma_r V_1^*$  and  $B^* = U \Lambda V^*$ , respectively; so that

$$B = V_1 \Sigma_r U_1^* = V \Lambda U^*.$$

■

**Example 1.3.4** As a continuation of Example 1.3.2, compute the full SVD of

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

**Solution** To obtain a full SVD of  $B$ , observe that in the solution of Example 1.3.2,  $V$  can be obtained by extending  $V_1$  to a  $3 \times 3$  orthogonal matrix:

$$V = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Therefore the SVD of  $B$  is given by

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

■

**Example 1.3.5** As a continuation of Example 1.3.3, compute the full SVD of the matrix

$$B = \begin{bmatrix} 1 & 0 & i \\ 1 & -i & 0 \\ -1 & 0 & -1 \\ 0 & 1 & i \end{bmatrix}.$$

**Solution** To obtain the full SVD for  $B$ , we must extend

$$V_1 = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3] \in \mathbb{C}^{4,3}$$

in the solution of Example 1.3.3 to a unitary matrix

$$V = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4] \in \mathbb{C}^{4,4},$$

by filling in the missing column vector  $\mathbf{v}_4$ . To do so, we may select any vector  $\mathbf{w}_4 \in \mathbb{C}^4$  which is not a linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$  and apply the Gram-Schmidt orthogonalization process to the linearly independent set  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{w}_4$ . In this example, we may choose  $\mathbf{w}_4 = [1, 0, 0, 0]^T$ , and compute:

$$\begin{aligned} \tilde{\mathbf{w}}_4 &= \mathbf{w}_4 - \langle \mathbf{w}_4, \mathbf{v}_1 \rangle \mathbf{v}_1 - \langle \mathbf{w}_4, \mathbf{v}_2 \rangle \mathbf{v}_2 - \langle \mathbf{w}_4, \mathbf{v}_3 \rangle \mathbf{v}_3 \\ &= \mathbf{w}_4 - \frac{2+2i}{\sqrt{30}} \mathbf{v}_1 - \frac{1+i}{\sqrt{6}} \mathbf{v}_2 - 0 \mathbf{v}_3 \\ &= \frac{1}{5} [2, -1-i, 1-i, -1+i]^T, \end{aligned}$$

followed by normalization of  $\tilde{\mathbf{w}}_4$ :

$$\mathbf{v}_4 = \frac{\tilde{\mathbf{w}}_4}{\|\tilde{\mathbf{w}}_4\|} = \frac{1}{\sqrt{20}} [2, -1-i, 1-i, -1+i]^T.$$

With this unitary matrix  $V = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4]$ , we obtain the full SVD  $B^* = U [\Sigma_3, O] V^*$  of  $B^*$ , yielding the full SVD

$$B = V \begin{bmatrix} \Sigma_3 \\ O \end{bmatrix} U^*,$$

after taking the complex conjugate of the transpose. ■

## 1.4 Principal Component Analysis (PCA) and Its Applications

The theory and methods of singular value decomposition (SVD) studied in the previous subunit are instrumental to data organization in terms of the degree of significance of the data information. To apply SVD, the concept of principal components is studied in this subunit. In this regard, since data matrices should not be treated as linear transformations, the notion of operator norm is not used in our study. Instead, we will introduce the notions of Frobenius norm and pseudo-inverses of rectangular matrices.

### 1.4.1 Frobenius norm measurement

To define the Frobenius norm of a matrix  $B = [b_{jk}] \in \mathbb{C}^{m,n}$ , we simply consider the matrix  $B$  as an  $mn \times 1$ -vector  $\mathbf{b} \in \mathbb{C}^{mn,1}$ , by arranging all the entries of  $B$  as a finite sequence, such as

$$\mathbf{b} = (b_{11}, \dots, b_{m1}, b_{12}, \dots, b_{m2}, \dots, b_{1n}, \dots, b_{mn}).$$

Then the Frobenius norm of the matrix  $B$  is defined by the Euclidean norm of the vector  $\mathbf{b}$ , as follows.

**Definition 1.4.1** *The Frobenius norm of an  $m \times n$  matrix  $B = [b_{jk}]$  is defined by*

$$\|B\|_F = \left( \sum_{j=1}^m \sum_{k=1}^n |b_{jk}|^2 \right)^{1/2}.$$

An important property of the Frobenius norm for data analysis is that it is governed by the  $\ell_2$  norm of the singular values of the (data) matrix, as follows.

**Theorem 1.4.1** *Let  $B \in \mathbb{C}^{m,n}$ , with  $\text{rank}(B) = r$ . Then*

$$\|B\|_F = \left( \sum_{j=1}^r \sigma_j^2 \right)^{1/2}, \quad (1.4.1)$$

where  $\sigma_1, \dots, \sigma_r$  are the (non-zero) singular values of  $B$ .

**Proof** To derive (1.4.1), let  $A = BB^*$  and observe that the Frobenius norm  $\|B\|_F$  of  $B$  agrees with the trace of  $A = [a_{jk}]$ ,  $j = 1, \dots, n$ . This fact is a

simple consequence of the definition of the trace, namely:

$$\begin{aligned}\text{Tr}(A) &= \sum_{k=1}^n a_{k,k} = \sum_{k=1}^n \left( \sum_{\ell=1}^m b_{k,\ell} \overline{b_{k,\ell}} \right) \\ &= \sum_{k=1}^n \sum_{\ell=1}^m |b_{k,\ell}|^2 = \|B\|_F^2.\end{aligned}$$

On the other hand,  $\text{Tr}(A)$  is the sum of the eigenvalues of  $A$ , and by the definition of the singular values of  $B$ , these eigenvalues

$$\lambda_1 = \sigma_1^2, \lambda_2 = \sigma_2^2, \dots, \lambda_n = \sigma_n^2$$

of  $A$  are squares of the singular values of the matrix  $B$ . Hence, we may conclude that

$$\|B\|_F = (\text{Tr}(A))^{1/2} = \left( \sum_{j=1}^n \sigma_j^2 \right)^{1/2} = \left( \sum_{j=1}^r \sigma_j^2 \right)^{1/2},$$

since  $\sigma_{r+1} = \dots = \sigma_m = 0$ . ■

**Example 1.4.1** Compute the Frobenius norm of the data matrix

$$B = \begin{bmatrix} 0 & 1 & 1 \\ 3 & -1 & 1 \\ 3 & 1 & -1 \end{bmatrix}.$$

**Solution** The Frobenius norm  $\|B\|_F$  can be computed directly by applying the definition, namely:

$$\|B\|_F^2 = \sum_{j=1}^3 \sum_{k=1}^3 |b_{jk}|^2 = 0^2 + 3^2 + 3^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 = 24,$$

so that  $\|B\|_F = \sqrt{24} = 2\sqrt{6}$ . ■

**Example 1.4.2** To demonstrate the result derived in Theorem 1.4.1, apply the formula (1.4.1) to compute the Frobenius norm of the data matrix  $B$  in Example 1.4.1.

**Solution** To compute the singular values of  $B$ , consider the matrix

$$A = BB^T = \begin{bmatrix} 0 & 1 & 1 \\ 3 & -1 & 1 \\ 3 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & 3 & 3 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 11 & 7 \\ 0 & 7 & 11 \end{bmatrix};$$

whose eigenvalues can be computed from the determinant of  $\lambda I_3 - A$ , namely:

$$\begin{aligned}(\lambda - 2) \begin{vmatrix} \lambda - 11 & -7 \\ -7 & \lambda - 11 \end{vmatrix} &= (\lambda - 2)(\lambda^2 - 22\lambda + 72) \\ &= (\lambda - 2)(\lambda - 4)(\lambda - 18),\end{aligned}$$

so that the eigenvalues of  $A$ , or equivalently, the squares of the singular values of  $B$ , are

$$\sigma_1^2 = 18, \sigma_2^2 = 4, \sigma_3^2 = 2.$$

Hence, it follows from Theorem 1.4.1, the Frobenius norm of the data matrix  $B$  is given by

$$\|B\|_F = \left( \sum_{j=1}^3 \sigma_j^2 \right)^{\frac{1}{2}} = (18 + 4 + 2)^{\frac{1}{2}} = 2\sqrt{6},$$

which agrees with the answer obtained in Example 1.4.1, computed directly by applying the definition. ■

Another important property of the Frobenius norm is that in view of Theorem 1.4.1, we may introduce the notion of the Schatten norm, by extending the  $\ell_2$ -norm to the general  $\ell_p$ -norm of the sequence of singular values of  $B$ , as opposed to the sequence of all of the entries of the matrix, as follows.

**Definition 1.4.2** Let  $B \in \mathbb{C}^{m,n}$  with  $\text{rank}(B) = r$ . For  $1 \leq p \leq \infty$ , the Schatten  $p$ -norm of  $B$  is defined by

$$\|B\|_{*,p} = \left( \sum_{j=1}^r \sigma_j^p \right)^{1/p},$$

where  $\sigma_1, \dots, \sigma_r$  are the (non-zero) singular values of  $B$ .

**Remark 1.4.1** The Schatten 1-norm,  $\|B\|_{*,1} = \sum_{j=1}^r \sigma_j$ , for  $p = 1$ , is called the nuclear norm (also called the trace norm or Ky Fan norm). Since this norm is very useful for low-rank matrix approximation and sparse matrix decomposition, the abbreviated notation

$$\|B\|_* = \|B\|_{*,1} = \sum_{j=1}^r \sigma_j \tag{1.4.2}$$

is commonly used for simplicity. ■

In the following, we establish the unitary invariance property of the Schatten  $p$ -norm, that includes the Frobenius norm (with  $p = 2$ ).

**Theorem 1.4.2** For any  $1 \leq p \leq \infty$  and any  $m \times n$  matrix  $B$ , the Schatten  $p$ -norm of an arbitrary unitary transformation of  $B$  remains the same as the Schatten  $p$ -norm of  $B$ . More generally,

$$\|WBR\|_{*,p} = \|B\|_{*,p}$$

for all unitary matrices  $W$  and  $R$  of dimension  $m \times m$  and  $n \times n$ , respectively. In particular, for  $p = 2$ ,

$$\|WBR\|_F = \|B\|_F, \tag{1.4.3}$$

The proof of this theorem can be accomplished by applying the unitary invariant property of singular values. ■

The Frobenius norm will be used to assess the exact error in the approximation of matrices  $B$  with rank  $r$ , by matrices  $C$  with rank  $\leq d$ , for any desired  $d < r$ . Approximation by lower-rank matrices is the first step to the understanding of data dimensionality reduction, a topic of applications to be discussed in Subunit 1.5.

#### 1.4.2 Principal components for data-dependent basis

Let  $B$  denote a data-set of  $m$  vectors  $\mathbf{b}_1, \dots, \mathbf{b}_m$  in  $\mathbb{C}^n$ . For convenience, the notation for the set  $B$  is also used for the matrix  $B \in \mathbb{C}^{m,n}$ , called a data-matrix, with row vectors:

$$\mathbf{b}_j^T = [b_{j,1}, \dots, b_{j,n}],$$

or column vectors  $\mathbf{b}_j$ , where  $j = 1, \dots, m$ ; that is,

$$B = \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_m^T \end{bmatrix} = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_m]^T.$$

Observe that the inner product of the  $j^{\text{th}}$  and  $k^{\text{th}}$  rows of  $B$ , defined by

$$\langle \mathbf{b}_j, \mathbf{b}_k \rangle = \sum_{\ell=1}^n b_{j,\ell} \overline{b_{k,\ell}}, \quad (1.4.4)$$

reveals the “correlation” of the data  $\mathbf{b}_j$  and  $\mathbf{b}_k$  in terms of the ratio of its magnitude with respect to the product of their norms. Recall from the Cauchy-Schwarz inequality from Subunit 1.1.2 that, with  $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$ ,

$$|\langle \mathbf{b}_j, \mathbf{b}_k \rangle| \leq \|\mathbf{b}_j\| \|\mathbf{b}_k\|,$$

and equality holds, if and only if  $\mathbf{b}_k$  is a constant multiple of  $\mathbf{b}_j$ . Hence,  $\mathbf{b}_j$  and  $\mathbf{b}_k$  are a good “match” of each other if the ratio

$$\frac{|\langle \mathbf{b}_j, \mathbf{b}_k \rangle|}{\|\mathbf{b}_j\| \|\mathbf{b}_k\|} \quad (1.4.5)$$

(which is between 0 and 1) is close to 1, and a poor “match”, if the ratio in (1.4.5) is close to 0. Since the inner product  $\langle \mathbf{b}_j, \mathbf{b}_k \rangle$  is the  $(j, k)^{\text{th}}$  entry of the  $m \times m$  square matrix  $A = BB^*$ , called the Gram matrix of  $B$ , the Gram matrix of a data-set is often used to process the data.

On the other hand, even if  $\mathbf{b}_j$  is a perfect match of  $\mathbf{b}_k$ , with ratio in (1.4.5) equal to 1, a minor “shift” of  $\mathbf{b}_j$  may be a bad match of  $\mathbf{b}_k$ , with ratio

in (1.4.5) much smaller than 1. Hence, to provide a better measurement of data correlation, all data vectors are shifted by their average, as follows. For  $\mathbf{b}_1, \dots, \mathbf{b}_m \in \mathbb{C}^n$ , their average is defined by

$$\mathbf{b}^{\text{av}} = \frac{1}{m} \sum_{j=1}^m \mathbf{b}_j, \quad (1.4.6)$$

and the **centered data-matrix**  $\tilde{B}$ , associated with the given data-matrix  $B$ , is defined by

$$\tilde{B} = [\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_m]^T = [\mathbf{b}_1 - \mathbf{b}^{\text{av}} \ \dots \ \mathbf{b}_m - \mathbf{b}^{\text{av}}]^T. \quad (1.4.7)$$

We remark that the matrix

$$\frac{1}{m} \tilde{B}(\tilde{B})^* \quad (1.4.8)$$

is the **covariance matrix** of the data-set  $B$ , if  $\mathbf{b}_1, \dots, \mathbf{b}_m$  are observations of an  $n$ -dimensional random variable. Clearly, both Gram matrices  $\tilde{B}(\tilde{B})^*$  and  $B B^*$  are self-adjoint and positive semi-definite matrices.

When the same notation  $B$  for the data-set is used for the data-matrix  $B = [\mathbf{b}_1 \ \dots \ \mathbf{b}_m]^T$ , the following notion of principal components of  $B$  plays an essential role in the analysis of the data.

**Definition 1.4.3** Let  $B = USV^*$  be the SVD of an  $m \times n$  matrix  $B$ , where  $S$  is given by (1.3.6) with singular values  $\sigma_1, \dots, \sigma_r$  of  $B$  in the sub-block  $\Sigma_r$  of  $S$ , arranged in non-increasing order as in (1.3.5). Then the singular vector pair  $(\mathbf{v}_1, \mathbf{u}_1)$  associated with the largest singular value  $\sigma_1$  is called the *principal component* of  $B$ . Furthermore, the singular vector pairs  $(\mathbf{v}_2, \mathbf{u}_2), \dots, (\mathbf{v}_r, \mathbf{u}_r)$ , associated with the corresponding singular values  $\sigma_2, \dots, \sigma_r$ , are called the *second, \dots,  $r^{\text{th}}$  principal components* of  $B$ , respectively.

**Remark 1.4.2** Let  $B = U_1 \Sigma_r V_1^*$  be the reduced SVD of an  $m \times n$  matrix  $B$  in Theorem 1.3.1, where  $\Sigma_r = \text{diag}\{\sigma_1, \dots, \sigma_r\}$  with  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$  being the singular values of  $B$ . Since  $U_1^* U_1 = I_r, V_1^* V_1 = I_r$ , we have

$$B B^* = U_1 \Sigma_r^2 U_1^*, \quad B^* B = V_1 \Sigma_r^2 V_1^*,$$

and hence,

$$B B^* U_1 = U_1 \Sigma_r^2, \quad B^* B V_1 = V_1 \Sigma_r^2.$$

Thus, for each  $j = 1, \dots, r$ ,  $\mathbf{u}_j$  is an eigenvector of  $B B^*$  associated with eigenvalue  $\sigma_j^2$  and  $\mathbf{v}_j$  is an eigenvector of  $B^* B$  associated with  $\sigma_j^2$ . Therefore, if  $\sigma_1 > \sigma_2 > \dots > \sigma_s$ , with  $2 \leq s \leq r$ , then  $\sigma_j^2, 1 \leq j \leq s$  are simple eigenvalues of  $B B^*$  and  $B^* B$  as well, and hence, the normalized corresponding eigenvectors  $\mathbf{u}_j$  and  $\mathbf{v}_j$  are unique. ■

**Remark 1.4.3** In application to data analysis, when the matrix  $B \in \mathbb{C}^{m,n}$  is a data-matrix, with the data in  $\mathbb{C}^n$  given by the  $m$  rows  $\mathbf{b}_1, \dots, \mathbf{b}_m$  of  $B$ , then the principal components of  $B$  provide a new “coordinate system”, with origin given by the average

$$\mathbf{b}^{\text{av}} = \frac{1}{m} \sum_{j=1}^m \mathbf{b}_j.$$

This coordinate system facilitates the analysis of the data, called **principal component analysis (PCA)**. All linear methods based on this data-dependent coordinate system are collectively called methods of principal component analysis (PCA).

Observe that for PCA, because

$$\sum_{j=1}^m \tilde{\mathbf{b}}_j = \sum_{j=1}^m (\mathbf{b}_j - \mathbf{b}^{\text{av}}) = \mathbf{0},$$

both the centered matrix  $\tilde{B}$  and its Gram  $\tilde{B}(\tilde{B})^*$  are centered at  $\mathbf{0}$ , in the sense that the sum of all row vectors (of  $\tilde{B}$ , and also of  $\tilde{B}\tilde{B}^*$ ) is the zero vector  $\mathbf{0}$ . In addition, the geometry and topology of the data set  $B$  are unchanged, when  $B$  is replaced by  $\tilde{B}$ , since for all  $j, k = 1, \dots, m$ ,

$$\|\tilde{\mathbf{b}}_j - \tilde{\mathbf{b}}_k\| = \|\mathbf{b}_j - \mathbf{b}_k\|. \quad (1.4.9)$$

In view of these nice properties of the centered matrix, when we say that PCA is applied to  $B$ , what we mean is that PCA is applied to the centered matrix  $\tilde{B}$  defined in (1.4.7). Hence, throughout our discussions, all data-matrices are assumed to be centered, and  $\tilde{B}$  is replaced by  $B$ .

**Definition 1.4.4** Let  $1 \leq d \leq q$  be integers. The notation  $\mathcal{O}_{q,d}$  will be used for the collection of all  $q \times d$  (complex or real) matrices  $W = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d]$  with mutually orthonormal column vectors; that is  $W^*W = I_d$  or

$$\langle \mathbf{w}_j, \mathbf{w}_k \rangle = \delta_{j-k}, \quad \text{all } 1 \leq j, k \leq d.$$

For any  $W = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_d] \in \mathcal{O}_{n,d}$ , its column vectors constitute an orthonormal basis of its algebraic span,

$$\text{span } W = \text{span } \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d\},$$

which is a  $d$ -dimensional subspace of  $\mathbb{C}^n$ , so that any  $\mathbf{w} \in \text{span } W$  has a unique representation:

$$\mathbf{w} = \sum_{j=1}^d c_j \mathbf{w}_j = W \begin{bmatrix} c_1 \\ \vdots \\ c_d \end{bmatrix},$$



for some  $c_j \in \mathbb{C}$ . The components of the column vector

$$[c_1 \cdots c_d]^*$$

will be called the “coordinates” of vector  $\mathbf{w}$ , in the “coordinate system” with coordinate axes determined by the unit vectors:  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_d$ . Recall that the superscript “\*” is used to denote the transpose of the complex conjugate of a matrix (or vector). All coordinate systems are restricted to orthogonal systems, with orthonormal vectors as unit vectors for the coordinate axes. This notion will be adopted in the study of data dimensionality reduction based on PCA in Subunit 1.5.

### 1.4.3 Pseudo-inverses

The notion of the inverse of a non-singular square matrix is generalized to the pseudo-inverse of any rectangular matrix  $B$ , by using its SVD, as follows.

**Definition 1.4.5** *Let  $B$  be an  $m \times n$  matrix of real or complex numbers and  $B = USV^*$  be its SVD, with unitary matrices  $U, V$  and  $S$  as given by (1.3.6), with diagonal sub-block  $\Sigma_r = \text{diag} \{\sigma_1, \dots, \sigma_r\}$  defined in (1.3.5). Set  $\Sigma_r^{-1} = \text{diag} \{\sigma_1^{-1}, \dots, \sigma_r^{-1}\}$  and define the  $n \times m$  matrix  $\tilde{S}$  by*

$$\tilde{S} = \begin{bmatrix} \Sigma_r^{-1} & \vdots & O \\ \dots\dots\dots & & \\ O & \vdots & O \end{bmatrix}_{n \times m}.$$

Then the  $n \times m$  matrix

$$B^\dagger = V\tilde{S}U^* \quad (1.4.10)$$

is called the pseudo-inverse of the given matrix  $B$ .

Here and throughout, the subscript  $n \times m$  of a matrix, such as of  $\tilde{S}$ , is used to indicate the matrix dimension. Observe that

$$BB^\dagger = (USV^*)(VS^+U^*) = U \begin{bmatrix} I_r & \vdots & O \\ \dots\dots\dots & & \\ O & \vdots & O \end{bmatrix}_{m \times m} U^*$$

and

$$B^\dagger B = (VS^+U^*)(USV^*) = V \begin{bmatrix} I_r & \vdots & O \\ \dots\dots\dots & & \\ O & \vdots & O \end{bmatrix}_{n \times n} V^*$$

are  $m \times m$  and  $n \times n$  square matrices, respectively, with  $r \times r$  identity matrix sub-block  $I_r$ , where  $r = \text{rank}(B)$ . Hence, for non-singular square matrices, the pseudo-inverse agrees with the inverse matrix. For this reason, the pseudo-inverse is also called the generalized inverse.

#### 1.4.4 Minimum-norm least-squares estimation

In this subunit, we will study the “solution” of the following (possibly over-determined, under-determined, or even inconsistent) system of linear equations

$$B\mathbf{x} = \mathbf{b}, \quad (1.4.11)$$

(to be called a “linear system” for convenience), where  $B$  is an  $m \times n$  coefficient matrix and  $\mathbf{b}$  an  $m$ -dimensional (known) column vector, by applying the pseudo-inverse  $B^\dagger$  of  $B$  to  $\mathbf{b}$  to formulate the vector:

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b}. \quad (1.4.12)$$

Throughout this subunit, the norm  $\|\mathbf{y}\|$  of any vector  $\mathbf{y} \in \mathbb{C}^n$  is the Euclidean (or  $\ell_2$ ) norm of  $\mathbf{y}$ . In the following we will show that taking the pseudo-inverse of the coefficient matrix  $B$  yields the minimum-norm least-squares solution  $\mathbf{x}^\diamond$ , as defined in (1.4.12), of the linear system (1.4.11).

**Theorem 1.4.3** *For the linear system (1.4.11) with coefficient matrix  $B \in \mathbb{C}^{m,n}$  and (known)  $\mathbf{b} \in \mathbb{C}^m$ , the vector  $\mathbf{x}^\diamond$  defined in (1.4.12) has the following properties:*

(i) *for all  $\mathbf{x} \in \mathbb{C}^n$ ,*

$$\|B\mathbf{x}^\diamond - \mathbf{b}\| \leq \|B\mathbf{x} - \mathbf{b}\|;$$

(ii) *the linear system (1.4.11), with unknown  $\mathbf{x}$ , has a solution if and only if the pseudo-inverse  $B^\dagger$  of  $B$  satisfies the condition:  $BB^\dagger \mathbf{b} = \mathbf{b}$ , namely,  $\mathbf{x} = \mathbf{x}^\diamond$  is a solution;*

(iii) *if (1.4.11) has a solution, then the general solution of (1.4.11)  $\mathbf{x} \in \mathbb{C}^n$  is given by*

$$\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w},$$

*for all  $\mathbf{w} \in \mathbb{C}^n$ ;*

(iv) *if (1.4.11) has a solution, then among all solutions,  $\mathbf{x}^\diamond$  is the unique solution*

*with the minimal Euclidean norm, namely:*

$$\|\mathbf{x}^\diamond\| \leq \|\mathbf{x}\|$$

*for any solution  $\mathbf{x}$  of (1.4.11); and*

(v) *if (1.4.11) has a unique solution, then  $\text{rank}(B) = n$ .*

*The above statements remain valid, when  $\mathbb{C}^{m,n}, \mathbb{C}^n, \mathbb{C}^m$  are replaced by  $\mathbb{R}^{m,n}, \mathbb{R}^n, \mathbb{R}^m$ , respectively.*

To prove the above theorem, we need the following properties of the pseudo-inverse.

**Theorem 1.4.4** *Let  $B \in \mathbb{C}^{n,m}$  or  $B \in \mathbb{R}^{n,m}$ . Then*

- (i)  $(BB^\dagger)^* = BB^\dagger$ ;
- (ii)  $(B^\dagger B)^* = B^\dagger B$ ;
- (iii)  $BB^\dagger B = B$ ; and
- (iv)  $B^\dagger BB^\dagger = B^\dagger$ .

Furthermore,  $B^\dagger$  as defined by (1.4.10), is the only  $n \times m$  matrix that satisfies the above conditions (i)–(iv).

**Proof** Derivation of the properties (i)–(iv) is an easy exercise. To show that  $B^\dagger$  is unique, let  $A \in \mathbb{C}^{m,n}$  satisfy (i)–(iv); that is,

- (i)  $(BA)^* = BA$ ;
- (ii)  $(AB)^* = AB$ ;
- (iii)  $BAB = B$ ; and
- (iv)  $ABA = A$ .

In view of the definition  $B^\dagger = V\tilde{S}U^*$  of  $B^\dagger$  in (1.4.10), we introduce four matrix sub-blocks  $A_{11}, A_{12}, A_{21}, A_{22}$  of dimensions  $r \times r, r \times (n-r), (m-r) \times r, (m-r) \times (n-r)$ , respectively, defined by

$$V^*AU = \begin{bmatrix} A_{11} & \vdots & A_{12} \\ \dots\dots\dots & & \\ A_{21} & \vdots & A_{22} \end{bmatrix}.$$

Then by the SVD formulation  $B = USV^*$  of the given matrix  $B$ , it follows from the assumption  $BAB = B$  in (iii) that

$$(U^*BV)(V^*AU)(U^*BV) = U^*(BAB)V = U^*BV.$$

Hence, by the definition (1.3.5) of  $S$ , we have

$$\begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix} = \begin{bmatrix} \Sigma_r & O \\ O & O \end{bmatrix},$$

which is equivalent to  $\Sigma_r A_{11} \Sigma_r = \Sigma_r$ . This yields  $A_{11} = \Sigma_r^{-1}$ . By applying the assumptions (i) and (ii) on  $A$  above, respectively, it is also an easy exercise to show that  $A_{12} = O$  and  $A_{21} = O$ . Finally, by applying these results along

with the assumption  $ABA = A$  in (iv), a similar derivation yields  $A_{22} = O$ . Hence, we have

$$A = V \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} U^* = V \begin{bmatrix} \Sigma_r^{-1} & O \\ O & O \end{bmatrix} U^* = B^\dagger.$$

This completes the proof of the theorem.  $\blacksquare$

We are now ready to prove Theorem 1.4.3. Write  $B\mathbf{x} - \mathbf{b} = (B\mathbf{x} - B\mathbf{x}^\diamond) + (B\mathbf{x}^\diamond - \mathbf{b})$ , and observe that the two vectors  $B\mathbf{x} - B\mathbf{x}^\diamond$  and  $B\mathbf{x}^\diamond - \mathbf{b}$  are orthogonal to each other. The reason is that

$$\begin{aligned} (B\mathbf{x}^\diamond - \mathbf{b})^*(B\mathbf{x} - B\mathbf{x}^\diamond) &= (BB^\dagger\mathbf{b} - \mathbf{b})^*B(\mathbf{x} - \mathbf{x}^\diamond) \\ &= \mathbf{b}^*((BB^\dagger)^* - I)B(\mathbf{x} - \mathbf{x}^\diamond) = \mathbf{b}^*(BB^\dagger B - B)(\mathbf{x} - \mathbf{x}^\diamond) = 0, \end{aligned}$$

where the last two equalities follow from (i) and (iii) of Theorem 1.4.4, respectively. Thus, it follows from the Pythagorean theorem (see Subunit 3.2.1 on the derivation for an arbitrary inner-product space) that

$$\|B\mathbf{x} - \mathbf{b}\|^2 = \|B\mathbf{x} - B\mathbf{x}^\diamond\|^2 + \|B\mathbf{x}^\diamond - \mathbf{b}\|^2 \geq \|B\mathbf{x}^\diamond - \mathbf{b}\|^2,$$

establishing statement (i) in Theorem 1.4.3.

Statement (ii) follows immediately from statement (i), since if the system (1.4.11) has some solution, then  $\mathbf{x}^\diamond$  in (1.4.10) is also a solution of (1.4.11).

To derive statement (iii), we first observe, in view of  $BB^\dagger B = B$  in (iii) of Theorem 1.4.4, that for any  $\mathbf{w} \in \mathbb{C}^n$ , the vector  $\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}$  is a solution of (1.4.11), since (1.4.11) has a solution, namely,  $\mathbf{x}^\diamond$ . On the other hand, suppose that  $\mathbf{x}$  is a solution. Then by setting  $\mathbf{w} = \mathbf{x} - \mathbf{x}^\diamond$ , we have  $B\mathbf{w} = \mathbf{0}$ , so that

$$\mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w} = \mathbf{x}^\diamond + \mathbf{w} - B^\dagger B\mathbf{w} = \mathbf{x}^\diamond + \mathbf{w} = \mathbf{x};$$

which establishes statement (iii).

To prove statement (iv), we apply (ii) and (iv) of Theorem 1.4.4 to show that in statement (iii) of the theorem, the vector  $\mathbf{x}^\diamond$  is orthogonal to  $(I_n - B^\dagger B)\mathbf{w}$ , namely:

$$\left((I_n - B^\dagger B)\mathbf{w}\right)^* \mathbf{x}^\diamond = \mathbf{w}^* \left(I_n - (B^\dagger B)^*\right) B^\dagger \mathbf{b} = \mathbf{w}^* \left(B^\dagger - B^\dagger B B^\dagger\right) \mathbf{b} = \mathbf{0}.$$

Hence, since every solution  $\mathbf{x}$  can be written as  $\mathbf{x} = \mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}$ , we may apply the Pythagorean theorem to conclude that

$$\begin{aligned} \|\mathbf{x}\|^2 &= \|\mathbf{x}^\diamond + (I_n - B^\dagger B)\mathbf{w}\|^2 \\ &= \|\mathbf{x}^\diamond\|^2 + \|(I_n - B^\dagger B)\mathbf{w}\|^2 \geq \|\mathbf{x}^\diamond\|^2. \end{aligned}$$

Thus,  $\mathbf{x}^\diamond$  is the unique solution of (1.4.11) with minimal norm.

Finally, to see that statement (v) holds, we simply observe that for the solution  $\mathbf{x}^\diamond$  of (1.4.11) to be unique, the matrix  $(I_n - B^\dagger B)$  in the general solution (in statement (iii)) must be the zero matrix; that is,  $B^\dagger B = I_n$  or  $B^\dagger$  is the right inverse of  $B$ , so that  $\text{rank}(B) = n$ . ■

**Example 1.4.3** Compute the pseudo-inverse of the matrix

$$B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}.$$

**Solution** Recall from Example 1.3.2 that the SVD of  $B$  is given by

$$B = USV^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Hence, by the definition (1.4.10) of the pseudo-inverse  $B^\dagger$  of  $B$ , we have

$$B^\dagger = V\tilde{S}U^* = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}.$$

■

**Example 1.4.4** Write the system of linear equations

$$\begin{cases} x_2 = 3, \\ x_1 - x_3 = -1, \end{cases}$$

in matrix formulation  $B\mathbf{x} = \mathbf{b}$  with  $\mathbf{x} = [x_1, x_2, x_3]^T$  and  $\mathbf{b} = [3, -1]^T$ . Apply the result from Example 1.4.3 to obtain the solution  $\mathbf{x}^\diamond = B^\dagger \mathbf{b}$  and verify that  $\|\mathbf{x}^\diamond\| \leq \|\mathbf{x}\|$  for all solutions  $\mathbf{x}$  of the linear system.

**Solution** Since the coefficient matrix  $B$  is the matrix in Example 1.4.3, we may apply  $B^\dagger$  computed above to obtain the solution

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b} = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2} \\ 3 \\ \frac{1}{2} \end{bmatrix}.$$

Furthermore, it is clear that the general solution of the system is

$$\mathbf{x} = [a - 1, 3, a]^T, \text{ for any real number } a,$$

so that

$$\begin{aligned}
 \|\mathbf{x}\|^2 &= (a-1)^2 + 3^2 + a^2 = a^2 - 2a + 1 + 9 + a^2 \\
 &= 2(a^2 - a) + 10 = 2\left(a^2 - a + \frac{1}{4}\right) + 10 - \frac{2}{4} \\
 &= 2\left(a - \frac{1}{2}\right)^2 + \left(3^2 + \left(\frac{-1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) \\
 &= 2\left(a - \frac{1}{2}\right)^2 + \|\mathbf{x}^\diamond\|^2 \geq \|\mathbf{x}^\diamond\|^2,
 \end{aligned}$$

with  $\|\mathbf{x}\| = \|\mathbf{x}^\diamond\|$  if and only if  $a = \frac{1}{2}$ , or  $\mathbf{x} = \mathbf{x}^\diamond$ . ■

**Example 1.4.5** Consider the inconsistent system of linear equations

$$\begin{cases} x_2 = 1, \\ x_1 = -1, \\ -x_2 = 1, \end{cases}$$

with matrix formulation  $B\mathbf{x} = \mathbf{b}$ , where

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}.$$

What would be a “reasonable” solution of the system?

**Solution** From Example 1.4.3 (see Example 1.3.2), since the coefficient matrix  $B$  is the transpose of the matrix  $B$  in Example 1.4.3, we have the SVD,  $B = USV^*$  with

$$\begin{aligned}
 U &= \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}, \\
 V &= \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad S = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.
 \end{aligned}$$

Hence, the pseudo-inverse  $B^\dagger$  of  $B$  is given by

$$B^\dagger = V\tilde{S}U^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix};$$

so that

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \\ \sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

That is,  $x_1 = -1$  and  $x_2 = 0$  is the “reasonable” solution of the inconsistent system. Observe that the average of the inconsistency  $x_2 = 1$  and  $-x_2 = 1$  is the “reasonable” solution  $x_2 = 0$ . ■

Next, we apply Theorem 1.4.3 to study the problem of least-squares estimation.

Let  $\mathbb{V}$  be an inner-product space over the scalar field  $\mathbb{C}$  or  $\mathbb{R}$  and  $S_n = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$  be a (possibly linearly dependent) set of vectors in  $\mathbb{V}$  with  $\mathbb{W} = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ . Since the cardinality  $n$  of the set  $S_n$  can be very large, to find a satisfactory representation of an arbitrarily given  $\mathbf{v} \in \mathbb{V}$  from  $\mathbb{W}$ , it is often feasible to acquire only a subset of measurements  $\langle \mathbf{v}, \mathbf{v}_\ell \rangle$  for  $\ell \in \{n_1, \dots, n_m\} \subset \{1, \dots, n\}$ . Let

$$\mathbf{b} = [b_1, \dots, b_m]^T = [\langle \mathbf{v}, \mathbf{v}_{n_1} \rangle, \dots, \langle \mathbf{v}, \mathbf{v}_{n_m} \rangle]^T \quad (1.4.13)$$

be the data vector in  $\mathbb{C}^m$  associated with  $\mathbf{v}$  (where  $m \leq n$ ). The least-squares estimation problem is to identify the “best” approximants

$$\mathbf{w} = \sum_{j=1}^n x_j \mathbf{v}_j \in \mathbb{W}$$

of the vector  $\mathbf{v}$ , based only on the measurement  $\mathbf{b}$  in (1.4.13). Now, since  $\langle \mathbf{w}, \mathbf{v}_{n_\ell} \rangle = \sum_{j=1}^n \langle \mathbf{v}_j, \mathbf{v}_{n_\ell} \rangle x_j$  is supposed to “match” the data component  $\langle \mathbf{v}, \mathbf{v}_{n_\ell} \rangle = b_\ell$  for  $\ell = 1, \dots, m$ , we consider the system of linear equations

$$\sum_{j=1}^n \langle \mathbf{v}_j, \mathbf{v}_{n_\ell} \rangle x_j = b_\ell = \langle \mathbf{v}, \mathbf{v}_{n_\ell} \rangle, \quad \ell = 1, \dots, m, \quad (1.4.14)$$

or in matrix formulation,

$$B\mathbf{x} = \mathbf{b}, \quad (1.4.15)$$

where  $\mathbf{x} = [x_1, \dots, x_n]^T$  and

$$B = [\langle \mathbf{v}_j, \mathbf{v}_{n_\ell} \rangle], \quad (1.4.16)$$

with  $1 \leq \ell \leq m$  and  $1 \leq j \leq n$ , is the  $m \times n$  coefficient matrix. Therefore, by Theorem 1.4.3, the “solution” to (1.4.15) is given by

$$\mathbf{x}^\diamond = B^\dagger \mathbf{b}$$

(where  $B^\dagger$  is the pseudo-inverse of  $B$ ) in that

$$\|B\mathbf{x}^\diamond - \mathbf{b}\| \leq \|B\mathbf{x} - \mathbf{b}\|$$

for all  $\mathbf{x} \in \mathbb{C}^n$  (or  $\mathbf{x} \in \mathbb{R}^n$ ) and that  $\|\mathbf{x}^\diamond\| \leq \|\mathbf{y}\|$  if

$$\|B\mathbf{y} - \mathbf{b}\| = \|B\mathbf{x}^\diamond - \mathbf{b}\|.$$

Of course, by setting  $\mathbf{x}^\diamond = (x_1^\diamond, \dots, x_n^\diamond)$ , the (unique) optimal minimum-norm least-squares representation of  $\mathbf{v} \in \mathbb{V}$  is given by

$$\sum_{j=1}^n x_j^\diamond \mathbf{v}_j. \quad (1.4.17)$$

**Remark 1.4.4** Matching the inner product with the data vector  $\mathbf{v}$  in (1.4.14) is a consequence of the variational method, when  $\sum_j x_j \mathbf{v}_j$  is required to be the best approximation of  $\mathbf{v}$  in the Euclidean (or  $\ell_2$ ) norm. Indeed, for the quantity  $\|\mathbf{v} - \sum_j x_j \mathbf{v}_j\|^2$  to be the smallest for all choices of coefficients  $x_1, \dots, x_n$ , the partial derivatives with respect to each of  $x_1, \dots, x_n$  must be zero. For convenience, we only consider the real-valued setting, so that for each  $\ell = 1, \dots, n$ ,

$$\begin{aligned} 0 &= \frac{\partial}{\partial x_\ell} \|\mathbf{v} - \sum_j x_j \mathbf{v}_j\|^2 \\ &= \frac{\partial}{\partial x_\ell} \left( \|\mathbf{v}\|^2 - 2 \sum_{j=1}^n x_j \langle \mathbf{v}, \mathbf{v}_j \rangle + \sum_{j=1}^n \sum_{k=1}^n x_j x_k \langle \mathbf{v}_j, \mathbf{v}_k \rangle \right) \\ &= -2 \langle \mathbf{v}, \mathbf{v}_\ell \rangle + \sum_{j=1}^n x_j \langle \mathbf{v}_j, \mathbf{v}_\ell \rangle + \sum_{k=1}^n x_k \langle \mathbf{v}_\ell, \mathbf{v}_k \rangle, \end{aligned}$$

or

$$\sum_{j=1}^n x_j \langle \mathbf{v}_j, \mathbf{v}_\ell \rangle = \langle \mathbf{v}, \mathbf{v}_\ell \rangle, \quad (1.4.18)$$

which is (1.4.14), when  $n_\ell$  is replaced by  $\ell$ . ■

**Remark 1.4.5** For computational efficiency and stability, the coefficient matrix  $B$  in (1.4.15) should be “sparse” by choosing locally supported vectors (or functions)  $\mathbf{v}_j$  in  $\mathbb{V}$ . For example, when piecewise polynomials (or splines) are used, it is best to use  $B$ -splines. In particular, when piecewise linear polynomials with equally spaced continuous “turning points” (called simple knots) are considered, then the linear  $B$ -splines are “hat” functions and the full matrix  $B_h = [\langle v_j, v_k \rangle]$ , for  $1 \leq j, k \leq n$ , is the banded square matrix

$$B_h = \frac{1}{6h} \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ & & \dots\dots\dots & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & 2 \end{bmatrix}, \quad (1.4.19)$$

where  $h > 0$  is the distance between two adjacent knots. ■



## 1.5 Applications to Data Dimensionality Reduction

What is data dimension? To understand this concept, let us consider the example of “color images”. Since human vision is “trichromatic”, meaning that our retina contains three types of color cone cells with different absorption spectra, the three primary color components: R (red), G (green) and B (blue), are combined, in various ratios, to yield a wide range of (wonderful) colors that we see. With the rapid technological advancement in high-quality digital image and video display, more accurate color profiles for consistent imaging workflow require significantly more sophisticated color calibration, by using spectro-colorimeters that take narrow-band measurements, even below  $10nm$  ( $nm$  = nano meter) increments. Hence, for visible light; that is, electromagnetic radiation (EMR) with wavelengths that range from  $400nm$  for the primary color “B” to  $700nm$  for the primary color “R”, even a  $10nm$  increment requires 31 readings, a 10-fold increase in data dimension over the 3 “RGB” readings. In other words, the “spectral curve” dimension for every single image pixel goes up from dimension 3 (for RGB) to dimension 31 (for 31 shades of colors), and even higher, if sub  $10nm$  increments are preferred. In Subunit 1.5.3, we will elaborate the discussion on the EMR range, beyond visible light, and mention various important application areas.

### 1.5.1 Representation of matrices by sum of norm-1 matrices

Let  $B \in \mathbb{R}^{m,n}$  denote the data matrix under consideration, with each of the  $m$  rows of  $B$  representing a data vector in the  $n$ -dimensional Euclidean space  $\mathbb{R}^n$ . The objective of the problem of dimensionality reduction is to reduce the dimension  $n$  for the purpose of facilitating data understanding, analysis and visualization, without loss of the essential data information, such as data geometry and topology. Let  $\mathbf{b}_j^T = [b_{j,1}, \dots, b_{j,n}]$  denote the row vectors of  $B$ , or equivalently, the column vectors of  $B$  are given by  $\mathbf{b}_j$ , where  $j = 1, \dots, m$ . For convenience, as already discussed in Subunit 1.4.2, without loss of data geometry and topology information, we may, and will, assume that the data matrix  $B$  has been centered; that is,

$$\sum_{j=1}^m \mathbf{b}_j = \mathbf{0}.$$

Suppose that the rank of the matrix  $B$  is  $r \geq 1$ . Our approach to the study of data dimensionality reduction, from dimension  $n$  to dimension  $d < n$ , is to decompose the data matrix  $B$  into the sum of  $r$  matrices, each with rank equal to 1, called rank-1 matrices in  $\mathbb{R}^{m,n}$ , in an “optimal” way and to

extract  $d$  of these  $r$  components. For this purpose, we apply the singular value decomposition (SVD) of  $B$  to derive the following (rank-1 decomposition) formula:

$$B = U_1 \Sigma_r V_1^* = U S V^* = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^*, \quad (1.5.1)$$

where  $U, V$  are unitary matrices of dimensions  $m, n$  respectively, and  $U_1 = [\mathbf{u}_1, \dots, \mathbf{u}_r]$ ,  $V_1 = [\mathbf{v}_1, \dots, \mathbf{v}_r]$  are obtained from  $U, V$  by keeping only the first  $r$  columns, where  $\sigma_1 \geq \dots \geq \sigma_r > 0$  are (all of) the non-zero singular values of  $B$  (with multiplicities being listed), and  $\mathbf{v}_j^*$  denotes the complex conjugate of the transpose of the  $j^{\text{th}}$  column  $\mathbf{v}_j$  of the matrix  $V_1$  in (1.3.15) or  $V$  in (1.3.19). To derive (1.5.1), we simply apply the reduced SVD of  $B$  in (1.3.15), studied in Subunit 1.3.3, to re-formulate the two matrix components  $U_1 \Sigma_r$  and  $V_1^*$ , and finally multiply the corresponding column and row sub-blocks, as follows:

$$B = U_1 \Sigma_r V_1^* = [\sigma_1 \mathbf{u}_1 \ \dots \ \sigma_r \mathbf{u}_r] \begin{bmatrix} \mathbf{v}_1^* \\ \vdots \\ \mathbf{v}_r^* \end{bmatrix} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^*.$$

Of course for real-valued matrix  $V_1$ , we have  $\mathbf{v}_j^* = \mathbf{v}_j^T$ . In addition, for each  $j = 1, \dots, r$ , observe that  $\mathbf{u}_j \mathbf{v}_j^*$  is an  $m \times n$  matrix with rank = 1. Such matrices are called rank-1 matrices. We remark that an  $m \times n$  matrix  $B$  is a rank-1 matrix, if and only if

$$B = \mathbf{v} \mathbf{w}^*,$$

where  $\mathbf{v}$  and  $\mathbf{w}$  are  $m$ -dimensional and  $n$ -dimensional column vectors, respectively. It is also easy to verify that the rank of the sum of  $r$  rank-1  $m \times n$  matrices does not exceed  $r$ .

## 1.5.2 Approximation by matrices of lower ranks

In this subunit, we will apply the rank-1 decomposition formula (1.5.1), namely:

$$B = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^*$$

of the data-matrix  $B$  into the sum of  $r$  rank-1 matrices, derived in the above subunit, to formulate the following best approximation result by matrices with rank  $d < r$ . The Frobenius norm, introduced and studied in Subunit 1.4.1, is used for the measurement of best approximation.

**Theorem 1.5.1** *Let  $B$  be any  $m \times n$  matrix with  $\text{rank}(B) = r \geq 1$  and with singular values  $\sigma_1 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = 0$ . Then for any integer  $d$ , with*

$1 \leq d < r$ , the  $d^{\text{th}}$  partial sum

$$B(d) = \sum_{j=1}^d \sigma_j \mathbf{u}_j \mathbf{v}_j^*, \quad (1.5.2)$$

of the rank-1 matrix series representation of  $B$  provides the best approximation of  $B$  by all matrices of rank  $\leq d$  under the Frobenius norm, with precise error given by  $\sigma_{d+1}^2 + \cdots + \sigma_r^2$ ; that is,

$$\|B - B(d)\|_F^2 = \sum_{j=d+1}^r \sigma_j^2, \quad (1.5.3)$$

and

$$\|B - B(d)\|_F^2 \leq \|B - C\|_F^2 \quad (1.5.4)$$

for all  $m \times n$  matrices  $C$  with rank  $\leq d$ . Furthermore,  $B(d)$  in (1.5.2) is the unique best approximant of  $B$ , again under the Frobenius norm.

**Proof** Let  $S_d$  denote the matrix obtained from the matrix  $S \in \mathbb{R}^{m,n}$ , introduced in (1.3.6)–(1.3.7), by replacing each of  $\sigma_1, \dots, \sigma_d$  by 0. Then we have

$$B - B(d) = US_d V^*.$$

Hence, it follows from (1.4.3) that

$$\|B - B(d)\|_F^2 = \|US_d V^*\|_F^2 = \|S_d\|_F^2 = \sum_{j=d+1}^r \sigma_j^2,$$

completing the derivation of (1.5.3).

To prove (1.5.4), assume that  $C \in \mathbb{C}^{m,n}$  with rank  $= k \leq d$  provides the best approximation to  $B$  under the Frobenius norm  $\|\cdot\|_F$ , so that  $\|B - C\|_F^2 \leq \|B - B(d)\|_F^2$ . Hence, it follows from (1.5.3) that

$$\|B - C\|_F^2 \leq \sum_{j=d+1}^r \sigma_j^2. \quad (1.5.5)$$

Let  $U, V$  be the unitary matrices in the SVD of  $B$  in (1.5.1) and set  $G = U^* C V$ , so that  $G$  has the same rank  $k$  as  $C$  and that  $C = U G V^*$ . Hence, by applying (1.4.3), we have

$$\begin{aligned} \|B - C\|_F &= \|USV^* - UGV^*\|_F \\ &= \|U(S - G)V^*\|_F = \|S - G\|_F. \end{aligned}$$

Set  $G = [g_{j,\ell}]$  and  $S = [s_{j,\ell}]$ , where in view of (1.3.6)–(1.3.7), we have  $s_{j,\ell} = 0$

for all  $j$  different from  $\ell$  and  $s_{j,j} = 0$  for  $j > r$ . Since the Frobenius norm  $\|S - G\|_F$  is defined by the  $\ell_2$  sequence norm of the sequence consisting of all the entries  $s_{j,\ell} - g_{j,\ell}$  of the matrix  $S - G$ , it should be clear that for  $\|S - G\|_F$  to be minimum among all  $G \in \mathbb{C}^{m,n}$ , this optimal  $G$  is given by

$$G = \begin{bmatrix} \Sigma'_r & 0 \\ 0 & 0 \end{bmatrix},$$

where  $\Sigma'_r = \text{diag}(g_1, g_2, \dots, g_r)$ , with each  $g_j \geq 0$ , so that

$$\|S - G\|_F^2 = \sum_{j=1}^r |s_{j,j} - g_j|^2 = \sum_{j=1}^r |\sigma_j - g_j|^2.$$

Now, since the rank of  $G$  is  $k \leq d \leq r$ , only  $k$  of the  $r$  diagonal entries  $g_1, g_2, \dots, g_r$  of  $\Sigma'_r$  are non-zero, and the minimum of the above sum is achieved only when these  $k$  non-zero entries match the largest  $k$  values of  $\sigma_1, \dots, \sigma_r$ . In other words, we have  $g_1 = \sigma_1, \dots, g_k = \sigma_k$ , and  $g_j = 0$  for  $j > k$ , and that

$$\|B - C\|_F^2 = \|S - G\|_F^2 = \sum_{j=k+1}^r \sigma_j^2. \quad (1.5.6)$$

Hence, by combining (1.5.5) and (1.5.6), we may conclude that  $k = d$ ,

$$\|B - C\|_F^2 = \sum_{j=d+1}^r \sigma_j^2,$$

and  $C = UGV^*$ , with

$$G = \begin{bmatrix} \Sigma'_r & 0 \\ 0 & 0 \end{bmatrix},$$

where  $\Sigma'_r = \text{diag}(\sigma_1, \dots, \sigma_d)$ . This implies that  $C = B(d)$  in (1.5.2), completing the proof of the theorem.  $\blacksquare$

### 1.5.3 Motivation to data-dimensionality reduction

Probably the most interesting examples of high-dimensional data are those generated by the electromagnetic (EM) waves, which can be described by their wavelengths (denoted by  $\lambda$ ), their frequencies (denoted by  $\nu$ ), or by their energies (denoted by  $E$ ). However, though it is important to know that energy is directly proportional to frequency, namely:  $E = h\nu$ , where  $h$  is called Planck's constant, the energies generated by EM waves do not play any role in our discussion of data dimension in this subunit. In other words, we will be only concerned with the study of wavelengths and frequencies, which are, of course, inversely proportional to each other, namely:  $\nu\lambda = v$ , where  $v$  denotes the velocity of the traveling EM wave, measured in meters per second (m/s).

For example, since light travels at the speed  $c = 299,792,458m/s$  in vacuum, we will use the formula

$$\nu\lambda = 299,792,458$$

in the following discussion for convenience. This assumption is quite accurate in the consideration of traveling EM waves emitted by our Sun.

The radiation of EM waves is called electromagnetic radiation (EMR), which has a wide spectrum from very low frequency-range (measured in Hertz, with one Hertz, or  $1Hz$ , equal to 1 cycle per second), to very high frequency-range, or equivalently very small wavelengths (measured in nanometers). The notation for nanometer is  $nm$ , with  $1nm = 10^{-9}$  meter. On the other hand, the tradition in the measurement of very low-frequency EMR, such as radio waves, is in terms of Hertz, such as  $1kHz$  (for one kilo  $Hz$ ) or  $1MHz$  (for one mega  $Hz$ ). For example, in the United States, while typical AM radio frequencies range from  $1,610kHz$  to  $1,710kHz$  (and not exceeding  $5,000kHz$ ), FM radio broadcast is in the frequency range from  $88MHz$  to  $108MHz$ . Observe that at  $100MHz = 10^8Hz$ , the wavelength of the FM radio signal is approximately 3 meters. Hence, since our discussion in this subunit is limited to the “visual” aspect of EMR, of which the very large wavelengths of the EM spectrum has not yet found its place in “visual” applications, we are only concerned with the EMR spectrum well below one-hundredth of a meter, and will use the nanometer,  $nm$ , unit for wavelength measurement.

As mentioned in the introduction of Subunit 1.5, the human vision is “trichromatic”, meaning that our retina contains three types of color cone cells for absorption of the three spectra: blue (B), green (G) and red (R). In other words, visible light to the human eye is in the EMR spectrum range of approximately  $380nm$  to  $750nm$ . More precisely, the rainbow colors, in terms of wavelengths, are given by:

red ( $620-750nm$ ),  
orange ( $590-620nm$ ),  
yellow ( $570-590nm$ ),  
green ( $495-570nm$ ),  
blue ( $450-495nm$ ),  
indigo (between blue and violet, but not specific), and  
violet ( $380-450nm$ ).

Although the human eye is not capable of “seeing” EMR beyond the rainbow spectrum, there are other creatures that can “see” different EMR spectra, though mostly narrower. For example, the majority of the compound eyes of insects are “bichromatic”, with narrow spectral vision, but can see EMR of higher-frequencies or shorter wavelengths in the  $340 - 380nm$  range of the ultraviolet (UV) spectrum. On the other hand, creatures such as snakes

can “see” EMR of the lower-frequencies or longer wavelengths, even in the  $5,000 - 30,000nm$  range of the infrared (IR) spectrum. Of course, since the IR spectrum is completely “dark”, the meaning of “vision” is actually “thermal sensing”, in terms of temperature differencing. The most advanced eyes of the entire animal kingdom are those of the mantis shrimp family, with visual capability up to the  $200 - 800nm$  range of the EMR spectrum, well beyond both ends of the (human) visible light spectrum of  $380 - 750nm$ .

So why is it so important to be able to “see” beyond our visible light spectrum? The truth is that all spectral bands with frequencies higher than the radio-frequency range are currently used for various image acquisition, with important applications to our daily lives. Before going further into this topic, let us first list all the spectral bands of the EMR that are relevant to imaging, as follows:

- (i) Gamma rays (with wavelengths less than  $0.02nm$ ),
- (ii) X-rays (with wavelengths in the  $0.01-10nm$  range),
- (iii) Ultraviolet, UV ( with wavelengths in the  $10-380nm$  range),
- (iv) Visible light (with wavelengths in the  $380-750$  range),
- (v) Infrared, IR (with wavelengths  $750nm-1$  millimeter; or frequencies  $300GHz-400THz$ ),
- (vi) Microwave (with wavelengths  $1$  millimeter– $1$  meter; or frequencies  $300MHz-300GHz$ ).

Since the Gamma ray spectrum is the one with frequencies over  $10^{19}Hz$ , Gamma rays produce the highest energy. For radiation therapy, gamma rays are used to kill cancerous cells. In the area of imaging, since gamma rays can penetrate through thick metal walls, they are used for taking pictures of illegal weapons and other suspicious objects hidden inside steel containers. In addition, gamma rays are used in astronomy research as well as in many manufacturing sectors, of which we can only compile the following short list:

- (i) for the automobile industry – test steel quality in the manufacture of cars and to obtain the proper thickness of tin and aluminum;
- (ii) for the aircraft industry – to check for flaws in jet engines;
- (iii) for road construction – to gauge the density of road surfaces and sub-surfaces;
- (iv) for pipeline companies – to test the strength of welds;
- (v) for the oil, gas, and mining industries – to map the contours of test wells and mine bores; and
- (vi) for cable manufacturers – to check ski lift cables for cracks.

As to medical applications mentioned above, due to their lower frequency (and hence smaller energy), X-rays are much less invasive than Gamma rays. Thanks to the discoverer, Wilhelm Roentgen, X-rays have been essential to various imaging applications in our daily lives, including: medical imaging, crystallography, and security screening. Here, medical X-ray imaging includes but is definitely not limited to: computed tomography (CT), fluoroscopy, radiography, and “conventional X-ray” such as mammography. As to crystallography, since the principle is to characterize atomic structures, crystallography is fundamental to many scientific fields, including: DNA matching and cancer drug research. The most noticeable application of X-ray image screening is probably inspection of airline carry-on bags and checked luggage for airport and air travel security. However, whole-body X-ray screening machines in major U.S. airports have recently been replaced by non-invasive “millimeter scanners”, to be discussed below, under the topic of thermal imaging.

Ultraviolet image acquisition is usually called forensic imaging, with applications that include but are not limited to:

- (i) authentication of original oil paintings by revealing brush strokes;
- (ii) verification of credit cards and important personal identities;
- (iii) detection of counterfeit (paper money);
- (iv) criminal identification by using finger-print image to match and by using forensic pictures to discover removed blood stains; and
- (v) separation of minerals to identify precious metals or gems.

More recently, UV imaging is also applied to enable automated systems to detect scratches and digs on optical surfaces such as lenses or windows. In the semiconductor industry, photolithography applies to inspection of photo-masks with very fine lines and features to locate defects that may be of submicron size. Confocal microscopy, operating at wavelength of  $248nm$  (generated by krypton fluoride) and at wavelength of  $266nm$  (generated by frequency-quadrupled lasers) is used to detect such image features. As an important application, this process can be applied to detect tiny defects in silicon wafers before carrying out semiconductor device (or silicon chip) fabrication in big volumes.

Infrared (IR) imaging, also called “thermal imaging”, is another popular image acquisition area, by using EMR outside the visible spectrum. In fact, IR cameras, also called “thermographic cameras”, have been used to take pictures in the dark many years ago. In general, the basic principle of thermographic cameras is to detect heat radiation in the infrared wavelength range between  $9,000nm$  and  $14,000nm$ , of the EMR spectrum, to produce images of the radiation, called “thermograms”. Since infrared radiation is emitted by all objects with heat content above absolute zero (according to the “black body radiation law” in physics), thermography makes it possible for us to “see” any environment with or without visible illumination. The reason is that the amount of

radiation emitted by an object increases with temperature, so that thermography allows us to see variations in temperature. When viewed through a thermal imaging camera, warm objects stand out well against cooler backgrounds. For example, humans and other warm-blood animals become easily visible against the environment, in bright daylight or pitch darkness. As a result, thermography has been particularly useful to military and other users of surveillance cameras. Other applications include seeing through smoke for the firefighters, seeing through light fog for the enthusiastic out-door athletes, locating overheating joints of power lines for electrical maintenance technicians, finding heat leaks (called thermal signatures) in faulty heat insulation for building construction, monitoring of physiological changes in people at home or hospital care, and monitoring of patients during clinical diagnostics. More recently, IR imaging plays a vital role in the auto industry, due to the increasing large number of electronic and mechanical components, in the areas of product assurance and driving reliability.

As mentioned above, the whole-body X-ray screening machines in the major U.S. airports have recently been replaced by “millimeter scanners”, by using radiation of EM waves with wavelengths that range from 0.1 millimeter (i.e.  $10^5 nm$ ) to 1 millimeter (i.e.  $10^6 nm$ ); or equivalently with frequencies ranging from  $3 \times 10^{11} Hz$  (i.e. 0.3 terahertz) to  $3 \times 10^{12} Hz$  (i.e. 3 terahertz). In physics, radiation of EM waves with frequencies in this range is called “terahertz radiation”. Observe that this “terahertz frequency band” is a narrow band in the EMR spectrum that consists of small portions of the far IR and near microwave spectra. Hence, in view of lower frequencies, millimeter scanning is even less invasive than scanning by thermographic cameras, in terms of radiation energy.

With the technological advancement of image sensors and other image acquisition devices, it was already feasible over three decades ago to capture images of the same scene simultaneously at several different desired wavelengths of the EMR spectrum, selected from visible light to long-wave infrared. The captured images constitute an image stack, called a “multispectral image”. Hence, a multispectral image can reveal the same scene both in bright daylight and in pitch darkness, when both the visible and infrared spectra are used in taking the picture. In this regard, multispectral imaging only deals with discrete and somewhat narrow (spectral) bands. Being “discrete and narrow” is what distinguishes multispectral imaging from color photography (which is accomplished by analyzing the spectrum of colors into three channels of information, one dominated by red, one by green, and the third by blue, in imitation of the way the normal human eye senses color).

While multispectral images have narrow bands, “hyperspectral” imaging is the popular choice for such applications that can benefit from using a vast portion of the electromagnetic spectrum, anywhere between ultraviolet and the terahertz band (if desired). Just like multispectral imaging, hyperspectral sensors simultaneously collect image data as a stack of images (of the same scene), but usually at equally spaced wavelengths called “spectral resolution”



(with small wavelength spacing for finer spectral resolution). These images are then combined to form a three-dimensional image stack, called an “HSI cube” (though it is only a right parallelepiped, rather than a cube, in general), for processing, analysis, and visualization. However, since a wide spectral band is used, if equal spacing is preferred, as in most applications, the spectral resolution cannot be too fine, with common choices of only  $5nm$  and  $10nm$ . More recently, spectral imaging with finer spectral resolution (i.e. smaller wavelength spacing between two consecutive images) is used for various specific applications. Such spectral images are called “ultraspectral” images, captured by using interferometer-type imaging sensors designed for very fine spectral resolution imaging. Unfortunately, these sensors have fairly low spatial resolution, due to high data-rate requirement for the desired spectral resolution.

For convenience, the totality of multispectral imaging, hyperspectral imaging, and ultraspectral imaging, will be called “spectral imaging” in our discussion. For each of these three types of imaging, a spectral image is a stack of digital images (of the same scene), captured at various desired wavelengths, to be called spectral bands, labeled by  $\ell^{\text{th}}$  band, for  $\ell = 1, \dots, n$ , where  $n$  is the number of (spectral) bands (or number of images in the stack). Hence, corresponding to each spatial pixel location  $(ih, jh)$ , where  $h > 0$  denotes the spatial distance between adjacent pixels, we may consider a pixel value  $b_{ih,jh}^\ell$  in the  $\ell^{\text{th}}$  band. For the sake of mathematical development and analysis, we use the roster ordering of the pixel locations; that is, horizontally from left to right, and line by line, starting from the first row. In other words, we consider the one-one map of the sequence

$$\{(h, h), \dots, (h, m_2h), (2h, h), \dots, (2h, m_2h), \dots, \dots, (m_1h, h), \dots, (m_1h, m_2h)\}$$

to the sequence  $\{1, \dots, m\}$ , where  $m = m_1 \times m_2$ . Hence, the pixel values  $b_{ih,jh}^\ell$  are re-labeled as  $b_k^\ell$ , where  $k = 1, \dots, m$  and  $\ell = 1, \dots, n$ . Now consider the row-vector

$$\mathbf{b}_k^T = [b_k^1 \ \dots \ b_k^n]$$

as a data-vector in  $\mathbb{R}^n$ . Then the matrix  $B \in \mathbb{R}^{m,n}$ , defined by

$$B = \begin{bmatrix} \mathbf{b}_1^T \\ \vdots \\ \mathbf{b}_m^T \end{bmatrix} = [\mathbf{b}_1 \ \dots \ \mathbf{b}_m]^T,$$

is the corresponding data-matrix of the spectral image under consideration. Therefore, we have formulated a spectral image as a data-set  $B$  of  $m$  points in the  $n$ -dimensional Euclidean space  $\mathbb{R}^n$ . Recall that the data-set notation is also used as the data-matrix notation.

Returning to our discussion of multispectral imaging, we remark that the most well-known and most accomplished application of multispectral image

acquisition is the on-going spectral image capture by remote sensing (RS) radiometers from Landsat satellites. For over forty years since 1972, the Landsat Program of NASA is a series of Earth-observing satellite missions of taking multispectral images of Earth's continents and surrounding coastal regions to facilitate and enable scientists to study various aspects of our planet and to evaluate the dynamic changes caused by both natural processes and human practices, with immediate applications to monitoring water quality, glacier recession, sea ice movement, invasive species encroachment, coral reef health, land use change, deforestation rates and population growth, and so forth. In addition, Landsat has helped in assessing damages from natural disasters such as fires, floods, and tsunamis, and subsequently, planning disaster relief and flood control programs. Even some fast-food restaurants have used population information to estimate community growth sufficient to warrant a new franchise. There were a total of eight Landsat satellites. The pioneering Landsat #1 was launched in 1972 and lasted till 1978, followed by Landsat #2 (1975-1982), Landsat #3 (1978-1983), and Landsat #4 (1982-2001). Landsat #5 (1984 - ) and Landsat #7 (1999 - ) are still busily sending imagery data to the ground stations for processing and storage. Unfortunately, Landsat #6 was lost at launch in 1993, but the good news is that Landsat #8, launched on February 11, 2013, is most successful. Orbiting the Earth every 99 minutes, Landsat #8 takes pictures of the entire Earth every 16 days, with 400 gigabytes (or 400 billion bytes) of most valuable imagery data downlinked to ground stations everyday for processing and archived in the U.S. Geological Survey (USGS). Since Landsat #8 includes additional bands, the combinations used to create RGB composites differ from Landsat #7 and Landsat #5. For instance, while bands #4, #3, #2 are used to create a color infrared (CIR) image by using data from Landsat #7 or Landsat #5, CIR composites are created by using Landsat #8 bands #5, #4, #3 instead. All Landsat data in the USGS archive are free and can be ordered from the USGS website ([www.usgs.gov](http://www.usgs.gov)). In the following, we list the spectral bands being used by Landsat #8:

- Band 1 - Coastal aerosol 430nm–450nm (visible light)
- Band 2 - Blue 450nm–510nm (visible light)
- Band 3 - Green 530nm–590nm (visible light)
- Band 4 - Red 640nm–670nm (visible light)
- Band 5 - Infrared 850nm–880nm (near Infrared)
- Band 6 - Infrared 1, 570nm–1, 650nm (short-wavelength IR)
- Band 7 - Infrared 2, 1100nm–1, 290nm (short-wavelength IR)
- Band 8 - Panchromatic 500nm–680nm (visible light)
- Band 9 - Cirrus 1, 360nm–1, 380nm (short-wavelength Infrared)

Band 10 - Infrared 10,600nm–11,190nm (long-wavelength-Infrared)

Band 11 - Infrared 11,500nm–12,510nm (long-wavelength-Infrared)

While multispectral sensors record discrete wavelengths of light, essentially sampling sections of the electromagnetic spectrum; a hyperspectral instrument, such as the Hyperion system, records many adjacent wavelengths to image most of the spectrum within a set range. In other words, in contrast to multispectral imaging that uses less than 10 (spectral) bands in general, the hyperspectral sensors look for objects by using hundreds (or occasionally even over a thousand) spectral bands to find and detect objects with traces of unique “fingerprints” across the electromagnetic spectrum. These “fingerprints” are known as spectral signatures and enable identification of the materials that make up a scanned object. Again for remote sensing (RS), the Hyperion sensor on Earth Observing #1 resolves 220 bands from 400nm to 2,500nm, with a spectral resolution (i.e. distance between adjacent bands) of 10 nm. Hyperspectral remote sensing is used in a wide array of applications. Although originally developed for mining and geology (the ability of hyperspectral imaging to identify various minerals makes it ideal for the mining and oil industries, where it can be used to look for ore and oil), it has now spread into fields as widespread as ecology and surveillance, as well as historical manuscript research, such as the imaging of the Archimedes Palimpsest. This technology is continually becoming more available to the public. Organizations such as NASA and the USGS have catalogues of various minerals and their spectral signatures, and have posted them online to make them readily available for researchers. According to a NASA Deputy Director, Bryant Cramer, “Hyperion is probably the future of remote sensing”. “Hyperion is a hyperspectral instrument, a change in technology that is like going from black-and-white to color television”, Mandel adds. Chemists have long used spectroscopy to identify substances because everything reflects electromagnetic energy (including light) at specific wavelengths and in ways that are as unique as a fingerprint. By measuring the energy that comes from a material, scientists can figure out what the material is. Hyperion measures reflected light like many other satellite imagers, but since it is recording more than 200 wavelengths, it can detect the fingerprints of the materials on Earth’s surface. Just as iron and copper look different in visible light, iron- and copper-rich minerals reflect varying amounts of light in the infrared spectrum.

“Hyperion has really opened up a whole new avenue of analysis that we hadn’t even explored before, and I can tell you where in the area the ore is coming from, which parts of the site were used for smelting and which were not; and that different parts of the site were drawing ore from different regions,” says NASA Contractor Sabrina Savage. Such information would be prohibitively expensive to gather in field research, but Hyperion provides an UCSD distinguished professor, Thomas Evan Levy, with the most valuable data for his Archaeology research, including better target excavation at likely smelting sites and mines.

HSI data have also assisted in the interpretation of ancient papyri, such as those found at Herculaneum, by imaging the fragments in the infrared range ( $1000\text{nm}$ ). Often, the text on the documents appears to be as black ink on black paper to the naked eye. At  $1000\text{ nm}$ , the difference in light reflectivity makes the text clearly readable. It has also been used to image the Archimedes palimpsest by imaging the parchment leaves in bandwidths from  $365 - 870\text{nm}$ , and then using advanced digital image processing techniques to reveal the under-text of Archimedes' work. Besides NASA's satellite image Hyperion, HSI cubes are also generated from airborne sensors like the NASA's Airborne Visible/Infrared Imaging Spectrometer (AVIRIS). In fact, over the past decade, HSI images are taken by ground troops. The list of applications of HSI is too long to list. Let us compile a short list as follows.

- (i) Agriculture: Although the cost of acquiring hyperspectral images is typically high, yet for specific crops and in specific climates, hyperspectral remote sensing use is increasing for monitoring the development and health of crops. In Australia, work is under way to use imaging spectrometers to detect grape variety and develop an early warning system for disease outbreaks. Furthermore, work is underway to use hyperspectral data to detect the chemical composition of plants, which can be used to detect the nutrient and water status of wheat in irrigated systems. Another application in agriculture is the detection of animal proteins in compound feeds to avoid bovine spongiform encephalopathy (BSE), also known as mad-cow disease. Different studies have been done to propose alternative tools to the reference method of detection (classical microscopy). One of the first alternatives is near infrared microscopy (NIR), which combines the advantages of microscopy and NIR. In 2004, the first study relating this problem with hyperspectral imaging was published. Hyperspectral libraries that are representative of the diversity of ingredients usually present in the preparation of compound feeds were constructed. These libraries can be used together with chemometric tools to investigate the limit of detection, specificity and reproducibility of the NIR hyperspectral imaging method for the detection and quantification of animal ingredients in feed.
- (ii) Mineralogy: A set of stones is scanned with a Specim LWIR-C imager in the thermal infrared range from  $7.7\text{ m}$  to  $12.4\text{ m}$ . The quartz and feldspar spectra are clearly recognizable. Geological samples, such as drill cores, can be rapidly mapped for nearly all minerals of commercial interest with hyperspectral imaging. Fusion of SWIR and LWIR spectral imaging is standard for the detection of minerals in the feldspar, silica, calcite, garnet, and olivine groups, as these minerals have their most distinctive and strongest spectral signature in the LWIR regions. Hyperspectral remote sensing of minerals is well developed. Many minerals can be identified from airborne images, and their relation to the presence of valuable minerals, such as gold and diamonds, is well un-

derstood. Currently, progress is towards understanding the relationship between oil and gas leakages from pipelines and natural wells, and their effects on the vegetation and the spectral signatures.

- (iii) Surveillance: In hyperspectral thermal infrared emission measurement, an outdoor scan in winter conditions, with ambient temperature of 15, relative radiance spectra from various targets in the image are shown with arrows. The infrared spectra of the different objects such as the watch glass have clearly distinctive characteristics. The contrast level indicates the temperature of the object. This image was produced with a Specim LWIR hyperspectral imager. Hyperspectral surveillance is the implementation of hyperspectral scanning technology for surveillance purposes. Hyperspectral imaging is particularly useful in military surveillance because of countermeasures that military entities now take to avoid airborne surveillance. Aerial surveillance was used by French soldiers using tethered balloons to spy on troop movements during the French Revolutionary Wars, and since that time, soldiers have learned not only to hide from the naked eye, but also to mask their heat signatures to blend into the surroundings and avoid infrared scanning. The idea that drives hyperspectral surveillance is that hyperspectral scanning draws information from such a large portion of the light spectrum that any given object should have a unique spectral signature in at least a few of the many bands that are scanned. The SEALs from DEVGRU who killed Osama bin Laden in May 2011 used this technology while conducting the raid (Operation Neptune's Spear) on Osama bin Laden's compound in Abbottabad, Pakistan.

On the other hand, for many applications, such as medical imaging and homeland security screening, it is highly recommended to use the EMR in the range between  $1nm$  (long X-ray) and  $10^6nm$  (for long-wave infrared), with spectral resolution of  $10nm$ . The purpose is that for each pixel location, there is an almost continuous curve  $\mathbf{b}$  (called spectral curve) consisting of  $10^5$  points (i.e.  $\mathbf{b} \in \mathbb{R}^{10^5}$ ), with one point from each band of the HSI cube. Therefore, to detect abnormal growth or calcification, the radiologist can compare the spectral curves of a patient with a library of compilations from previous cancer patients; and for homeland security screening, the spectral curves can be compared with a library of curves of suspicious illegal substances, such as chemicals and biological poison, for screening carry-on bags and checked luggage for airports security and for entering other public buildings, etc. To meet this specification, the high data-dimension of  $10^5$  cannot be processed without a super computer. Therefore, to facilitate computational efficiency, memory usage, data understanding and visualization, it is necessary to reduce the data dimension, while preserving data similarities (and dis-similarities), data geometry, and data topology.

### 1.5.4 Principal components as basis for dimension-reduced data

Recall from the discussion in Subunit 1.4.2 that to provide a better measurement of data correlation, the data vectors  $\mathbf{b}_1, \dots, \mathbf{b}_m \in \mathbb{C}^n$ , should be shifted by their average; that is, by replacing  $\mathbf{b}_j$  with  $\tilde{\mathbf{b}}_j = \mathbf{b}_j - \mathbf{b}^{\text{av}}$ , for each  $j = 1, \dots, m$ , where

$$\mathbf{b}^{\text{av}} = \frac{1}{m} \sum_{k=1}^m \mathbf{b}_k. \quad (1.5.7)$$

In other words, the data-matrix

$$B = [\mathbf{b}_1, \dots, \mathbf{b}_m]^T \quad (1.5.8)$$

is replaced by the centered data-matrix

$$\tilde{B} = [\tilde{\mathbf{b}}_1, \dots, \tilde{\mathbf{b}}_m]^T = [\mathbf{b}_1 - \mathbf{b}^{\text{av}} \dots \mathbf{b}_m - \mathbf{b}^{\text{av}}]^T. \quad (1.5.9)$$

For convenience, the data-matrix  $B$  in the following discussion is assumed to have been centered; that is, we will assume that  $B = \tilde{B}$ . Now, consider the rank-1 decomposition formula

$$B = U_1 \Sigma_r V_1^* = U S V^* = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^*$$

of the centered data-matrix  $B$  in (1.5.1). Recall that the  $m$  rows of  $B$  (before subtracting the average) are the  $m$  data-vectors. By applying the above rank-1 decomposition formula, the data-dimension  $n$  is already reduced to dimension  $r$  (which is the rank of the data matrix), provided that  $r < n$ . To achieve steeper dimension reduction,  $r$  is further reduced to any desirable dimension  $d$ , with  $1 \leq d \leq r$ . To do so, we first recall the notation  $\mathcal{O}_{n,d}$ , in Definition 1.4.4, for the collection of all  $n \times d$  matrices  $W = [\mathbf{w}_1 \dots \mathbf{w}_d]$  with mutually orthonormal column vectors (that is,  $W^* W = I_d$ ), and we will consider the matrix  $W = W_d$ , defined by

$$W_d = [\bar{\mathbf{v}}_1 \dots \bar{\mathbf{v}}_d],$$

which is in  $\mathcal{O}_{n,d}$ . Hence, as mentioned in Subunit 1.4.2, the coordinate system, with the orthonormal unit vectors  $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_d$  as coordinate axes, is the optimal coordinate system, in the sense of PCA, for reducing the data dimension from  $n$  to  $d$ . To be more precise, we return to the SVD of the data matrix  $B$  (that has been centered), namely

$$B = [\mathbf{b}_1 \dots \mathbf{b}_m]^T = U S V^*, \quad (1.5.10)$$

where

$$U = [\mathbf{u}_1 \dots \mathbf{u}_m], \quad V = [\mathbf{v}_1 \dots \mathbf{v}_n],$$

are unitary matrices, with orthonormal column vectors  $\mathbf{u}_j$  and  $\mathbf{v}_j$ , respectively. The first idea in the PCA dimensionality reduction approach is to replace the matrices  $U, V, S$  in the SVD of the data-matrix  $B$ , by the truncated matrices

$$U_d = [\mathbf{u}_1 \cdots \mathbf{u}_d], V_d = [\mathbf{v}_1 \cdots \mathbf{v}_d], \Sigma_d = \text{diag}\{\sigma_1, \dots, \sigma_d\}, \quad (1.5.11)$$

respectively. Then, in view of the SVD of  $B$  in (1.5.10), or equivalently  $BV = US$ , we may consider the  $m \times d$  matrix representation:

$$BV_d = U_d \Sigma_d,$$

and apply this  $m \times d$  matrix  $Y_d = BV_d = U_d \Sigma_d$  to introduce the notion of dimension-reduced data, as follows.

**Definition 1.5.1** *Let  $0 \leq d < r$ . The column vectors  $\mathbf{y}_1, \dots, \mathbf{y}_m$  of the matrix  $Y_d^T = (BV_d)^T = (U_d \Sigma_d)^T$ ; or equivalently,*

$$\begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_m^T \end{bmatrix} = Y_d = BV_d = U_d \Sigma_d, \quad (1.5.12)$$

*are said to constitute the dimension-reduced data of the given data-set  $B = \{\mathbf{b}_1, \dots, \mathbf{b}_m\} \subset \mathbb{C}^n$ .*

The main result on dimensionality reduction is the following theorem which states that the dimension-reduced data-set  $\{\mathbf{y}_1, \dots, \mathbf{y}_m\}$  of the given data  $B = \{\mathbf{b}_1, \dots, \mathbf{b}_m\} \subset \mathbb{C}^n$  is the optimal choice, in the sense that when measured in the “coordinate system”, with coordinate axes determined by the unit vectors:  $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_d \in \mathbb{C}^n$ , this set provides the best  $\ell_2$ -approximation of the given data-set  $B$  among all possible choices  $\mathbf{q}_1, \dots, \mathbf{q}_m \in \mathbb{C}^d$  and all possible “coordinate systems”  $W_d$  of  $\mathbb{C}^d$ .

**Theorem 1.5.2** *Let  $B = [\mathbf{b}_1 \cdots \mathbf{b}_m]^T$  be an arbitrary  $m \times n$  data-matrix with SVD representation given by (1.5.10). Then the dimension-reduced data  $\mathbf{y}_1, \dots, \mathbf{y}_m$  of  $B$ , defined by (1.5.12), lie in a  $d$ -dimensional subspace of  $\mathbb{C}^n$  and satisfy the following best approximation property:*

$$\sum_{j=1}^m \|\bar{V}_d \mathbf{y}_j - \mathbf{b}_j\|^2 \leq \sum_{j=1}^m \|W_d \mathbf{q}_j - \mathbf{b}_j\|^2, \quad (1.5.13)$$

*for all  $W_d = [\mathbf{w}_1 \cdots \mathbf{w}_d] \in \mathcal{O}_{n,d}$  and all  $\{\mathbf{q}_1, \dots, \mathbf{q}_m\} \subset \mathbb{C}^d$ , where  $V_d, \Sigma_d$  are defined by (1.5.11) with  $\bar{V}_d \in \mathcal{O}_{n,d}$ .*

In (1.5.13), for each  $j = 1, \dots, m$ , the vector  $\bar{V}_d \mathbf{y}_j$  lies in a  $d$ -dimensional subspace of  $\mathbb{C}^n$  with basis  $\{\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_d\}$ , and the set  $\{\bar{V}_d \mathbf{y}_1, \dots, \bar{V}_d \mathbf{y}_m\}$  is an approximation of the given dataset  $B = \{\mathbf{b}_1, \dots, \mathbf{b}_m\}$ . Observe that

$\mathbf{y}_1, \dots, \mathbf{y}_m$  are the coordinates of the approximants  $\bar{V}_d \mathbf{y}_1, \dots, \bar{V}_d \mathbf{y}_m$  in the coordinate system  $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_d$ . Hence, the inequality (1.5.13) guarantees that the first  $d$  principal components  $\mathbf{v}_1, \dots, \mathbf{v}_d$  of  $B$  provide the best coordinate system (in hierarchal order), for optimal dimensionality reduction of the dataset  $B \subset \mathbb{C}^n$  to a  $d$ -dimensional subspace, for any choice of dimension  $d < n$ . In particular, to reduce  $B$  to a 1-dimensional subspace, the first principal component  $\mathbf{v}_1$  of  $B$  should be used to give the generator of this subspace for computing and representing the best 1-dimensional reduced data; to reduce  $B$  to a 2-dimensional subspace, the first and second principal components  $\mathbf{v}_1, \mathbf{v}_2$  of  $B$  should be used for computing and representing the best 2-dimensional reduced data; and so forth. In general, to reduce the dimension of a given dataset  $B = \{\mathbf{b}_1, \dots, \mathbf{b}_m\} \subset \mathbb{C}^n$  to a  $d$ -dimensional subspace of  $\mathbb{C}^n$ , for any  $d < n$ , the best replacement of  $B$  is given by:

$$\begin{bmatrix} (\bar{V}_d \mathbf{y}_1)^T \\ \vdots \\ (\bar{V}_d \mathbf{y}_m)^T \end{bmatrix} = Y_d V_d^* = U_d \Sigma_d V_d^* = B V_d V_d^*. \quad (1.5.14)$$

To prove the above theorem, we first observe that in view of (1.5.12) and the above dimension-reduced data representation formula (1.5.14), we may apply the formulation of  $B(d)$  in (1.5.2) to write

$$\begin{aligned} \begin{bmatrix} (\bar{V}_d \mathbf{y}_1)^T \\ \vdots \\ (\bar{V}_d \mathbf{y}_m)^T \end{bmatrix} - B &= Y_d V_d^* - B = [\mathbf{u}_1 \ \dots \ \mathbf{u}_d] \Sigma_d [\mathbf{v}_1 \ \dots \ \mathbf{v}_d]^* - B \\ &= B(d) - B. \end{aligned}$$

Hence, since the left-hand side of (1.5.13) can be written as

$$\sum_{j=1}^m \|\mathbf{y}_j^T V_d^* - \mathbf{b}_j^T\|^2,$$

which is precisely the square of the Frobenius norm of  $Y_d V_d^* - B$  (see (1.4.1)), we have

$$\sum_{j=1}^m \|\bar{V}_d \mathbf{y}_j - \mathbf{b}_j\|^2 = \|B(d) - B\|_F^2.$$

Let  $Q$  be the  $m \times d$  matrix with the  $j^{\text{th}}$  row given by  $\mathbf{q}_j^T$  for  $1 \leq j \leq m$ . Then the right-hand side of (1.5.13) can also be written as

$$\sum_{j=1}^m \|W_d \mathbf{q}_j - \mathbf{b}_j\|^2 = \sum_{j=1}^m \|\mathbf{q}_j^T W_d^T - \mathbf{b}_j^T\|^2 = \|R - B\|_F^2,$$

where

$$R = Q W_d^T. \quad (1.5.15)$$



Since  $W_d \in \mathcal{O}_{n,d}$ , the rank of the matrix  $R$  in (1.5.15) does not exceed  $d$ . Therefore, the desired inequality (1.5.13) follows from (1.5.4) in Theorem 1.5.1 on p.38. Furthermore, again by this theorem, (the square of) the error of the dimensionality reduction from  $B \subset \mathbb{C}^n$  to  $\mathbf{y}_1, \dots, \mathbf{y}_m \subset \mathbb{C}^n$  is given by

$$\sum_{j=1}^m \|\bar{V}_d \mathbf{y}_j - \mathbf{b}_j\|^2 = \|B(d) - B\|_F^2 = \sum_{j=d+1}^r \sigma_j^2. \quad (1.5.16)$$

■

**Example 1.5.1** Let  $B = [\mathbf{b}_1 \ \mathbf{b}_2 \ \mathbf{b}_3]^T \subset \mathbb{R}^2$ , where

$$\mathbf{b}_1 = \begin{bmatrix} 2.5 \\ 4.5 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 9 \\ 4 \end{bmatrix}, \quad \mathbf{b}_3 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}.$$

Compute the dimension-reduced data  $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$  of  $B$ , which lie in a 1-dimensional subspace of  $\mathbb{R}^2$ , by considering the dimension-reduced data  $\{\tilde{y}_1, \tilde{y}_2, \tilde{y}_3\}$  of the corresponding centered dataset  $\tilde{B} = [\tilde{\mathbf{b}}_1; \tilde{\mathbf{b}}_2; \tilde{\mathbf{b}}_3]^T$  of  $B$ . In addition, compute the  $2 \times 1$  (real) matrix transformation  $V_1 = [\mathbf{v}_1]$  in (1.5.11), for which

$$\sum_{j=1}^3 \|\tilde{y}_j \mathbf{v}_1 - \tilde{\mathbf{b}}_j\|^2 \leq \sum_{j=1}^3 \|q_j \mathbf{w}_1 - \tilde{\mathbf{b}}_j\|^2$$

for any unit vector  $\mathbf{w}_1 \in \mathbb{R}^2$  and all  $q_1, q_2, q_3 \in \mathbb{R}$ .

**Solution** Since the average  $\mathbf{b}^{\text{av}}$  of  $B$  is

$$\mathbf{b}^{\text{av}} = \begin{bmatrix} 3.5 \\ 2.5 \end{bmatrix},$$

we have

$$\begin{aligned} \tilde{\mathbf{b}}_1 &= \mathbf{b}_1 - \mathbf{b}^{\text{av}} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \\ \tilde{\mathbf{b}}_2 &= \mathbf{b}_2 - \mathbf{b}^{\text{av}} = \begin{bmatrix} 5.5 \\ 1.5 \end{bmatrix}, \\ \tilde{\mathbf{b}}_3 &= \mathbf{b}_3 - \mathbf{b}^{\text{av}} = \begin{bmatrix} -4.5 \\ -3.5 \end{bmatrix}; \end{aligned}$$

so that the centered dataset is given by

$$\tilde{B} = \begin{bmatrix} \tilde{\mathbf{b}}_1^T \\ \tilde{\mathbf{b}}_2^T \\ \tilde{\mathbf{b}}_3^T \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 5.5 & 1.5 \\ -4.5 & -3.5 \end{bmatrix}.$$

Since  $\tilde{B}$  has more rows than columns, we may consider the spectral decomposition of  $\tilde{B}^*\tilde{B}$  (instead of  $\tilde{B}\tilde{B}^*$ ) to obtain  $V_1$  in the reduced SVD of  $\tilde{B}$  and apply  $\tilde{Y}_1 = \tilde{B}V_1$  to obtain the reduced data for  $\tilde{B}$ , as follows. By direct calculation, we have

$$\tilde{B}^*\tilde{B} = \begin{bmatrix} 51.5 & 22 \\ 22 & 18.5 \end{bmatrix},$$

with eigenvalues  $125/2$  and  $15/2$  and corresponding eigenvectors given by

$$\mathbf{v}_1 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}.$$

Hence, it follows from (1.5.12) that

$$Y_1 = \tilde{B}V_1 = \tilde{B}\mathbf{v}_1 = \begin{bmatrix} 0 \\ \frac{5\sqrt{5}}{2} \\ -\frac{5\sqrt{5}}{2} \end{bmatrix}.$$

That is, the dimension-reduced dataset for  $\tilde{B}$  is

$$\{\tilde{y}_1, \tilde{y}_2, \tilde{y}_3\} = \{0, \frac{5\sqrt{5}}{2}, -\frac{5\sqrt{5}}{2}\}.$$

Hence, the dimension-reduced dataset  $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$  of  $B$  can be obtained by adding  $\mathbf{b}^{\text{av}}$  to  $\tilde{y}_1\mathbf{v}_1, \tilde{y}_2\mathbf{v}_1, \tilde{y}_3\mathbf{v}_1$ , namely:

$$\mathbf{y}_1 = \begin{bmatrix} 3.5 \\ 2.5 \end{bmatrix}, \quad \mathbf{y}_2 = \begin{bmatrix} 8.5 \\ 5 \end{bmatrix}, \quad \mathbf{y}_3 = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix}.$$

In the above discussion, the matrix  $Y_d = \tilde{B}V_d$  is applied to obtain the dimension-reduced data. If the reduced SVD of  $\tilde{B}$  is applied, we may use  $Y_d = U_d\Sigma_d$  to obtain the reduced data instead. ■

# Unit 2

## DATA COMPRESSION

This unit is a comprehensive study of data compression, with emphasis on the compression of digital images. The discrete Fourier transform (DFT) and a fast implementation, called FFT, of the DFT for  $\mathbb{R}^n$  or  $\mathbb{C}^n$ , with  $n = 2^m$ , are studied in some detail, and applied to introduce the discrete cosine transform (DCT) and a fast computation of the DCT. The basic topics of information representation and information coding, including the notion of histograms, entropy of probability distributions, and binary codes, are discussed. For data compression, Shannon's Noiseless Coding Theorem is derived, based on Kraft's inequality and in terms of the entropy; and the Huffman coding scheme is presented. When applied to digital image compression, the methods for lossless (or reversible) compression are briefly discussed, but an indepth study of lossy compression is presented, with quantization of the DCT coefficients as the key component of the lossy compression scheme. In addition, the study of digital image compression is extended to digital video compression, and the current image and video compression standards are also discussed in this unit.

### 2.1 Discrete and Fast Fourier Transform (FFT)

Fourier series representation of functions on a bounded interval is a basic tool in applied mathematics. The study of this important topic requires some depth of mathematical analysis and will be delayed to the next unit, with the concept to be introduced in Subunit 3.1. On the other hand, the discrete version of the Fourier coefficients of a Fourier series, to be called the discrete Fourier transform (DFT) of some finite dimensional vector  $\mathbf{v}_n$ , is simply a matrix-to-vector multiplication operation  $F_n \mathbf{v}_n$ , where  $F_n$  is some square matrix of dimension  $n$ , and the operation is considered as a linear transformation from  $\mathbb{C}^n$  to  $\mathbb{C}^n$ . This notion of DFT is introduced and studied in Subunit 2.1.1. For even integers  $n$ , the Lanczos matrix factorization of  $F_n$  is derived in Subunit 2.1.2. This matrix factorization result will be applied in Subunit 2.1.3 to decompose a DFT matrix  $F_n$  for  $n = 2^m$ , yielding the fast Fourier transform (FFT) computational scheme.

### 2.1.1 Definition of DFT

#### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 8: Discrete Time Fourier Transform (YouTube), presented by Gilbert Strang.
- (2) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications," pages 171–179. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 2.1.2 Lanczos matrix factorization

To study the fast Fourier transform (FFT) computational scheme, we need the following result due to Cornelius Lanczos to reduce a  $2n$ -point discrete Fourier transform (DFT) to an  $n$ -point DFT via multiplication by certain sparse matrices.

Denote

$$D_n = \text{diag}\{1, e^{-i\pi/n}, \dots, e^{-i(n-1)\pi/n}\},$$

and let

$$P_n^e = [\delta_{2j-k}] = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix};$$

$$P_n^o = [\delta_{2j-k+1}] = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \cdots & 0 \\ & & & \cdots & & \\ 0 & 0 & & \cdots & 0 & 1 \end{bmatrix},$$

where  $0 \leq j \leq n-1$  and  $0 \leq k \leq 2n-1$ , be  $n \times (2n)$  matrices. Also, let  $I_n$  denote the  $n \times n$  identity matrix and

$$E_{2n} = \begin{bmatrix} I_n & \vdots & D_n \\ \cdots & \cdots & \cdots \\ I_n & \vdots & -D_n \end{bmatrix}. \quad (2.1.1)$$

For example, with the above notations, we have

$$\begin{aligned} D_1 &= [1], P_1^e = [1 \ 0], P_1^o = [0 \ 1], E_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \\ D_2 &= \begin{bmatrix} 1 & 0 \\ 0 & -i \end{bmatrix}, P_2^e = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, P_2^o = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\ E_4 &= \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -i \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & i \end{bmatrix}. \end{aligned}$$

**Theorem 2.1.1** *The  $2n$ -point DFT  $F_{2n}$  can be factored out in terms of two diagonal blocks of the  $n$ -point DFT  $F_n$  as follows:*

$$F_{2n} = E_{2n} \begin{bmatrix} F_n & \vdots & O \\ \dots\dots\dots & & \\ O & \vdots & F_n \end{bmatrix} \begin{bmatrix} P_n^e \\ \dots \\ P_n^o \end{bmatrix}, \quad (2.1.2)$$

where  $O$  denotes the  $n \times n$  zero matrix.

**Proof** To derive (2.1.2), consider  $\mathbf{x} = [x_0, \dots, x_{2n-1}]^T \in \mathbb{C}^{2n}$ , with DFT given by  $\hat{\mathbf{x}} = F_{2n}\mathbf{x} = [\hat{x}_0, \dots, \hat{x}_{2n-1}]^T$ , where

$$\begin{aligned} \hat{x}_\ell &= \sum_{k=0}^{2n-1} x_k e^{-i2\pi k\ell/(2n)} \\ &= \sum_{j=0}^{n-1} x_{2j} e^{-i2\pi(2j)\ell/(2n)} + \sum_{j=0}^{n-1} x_{2j+1} e^{-i2\pi(2j+1)\ell/(2n)} \\ &= \sum_{j=0}^{n-1} x_{2j} e^{-i2\pi j\ell/n} + e^{-i\pi\ell/n} \sum_{j=0}^{n-1} x_{2j+1} e^{-i2\pi j\ell/n}. \end{aligned} \quad (2.1.3)$$

Set

$$\mathbf{x}_e = [x_0, x_2, \dots, x_{2n-2}]^T, \mathbf{x}_o = [x_1, x_3, \dots, x_{2n-1}]^T.$$

Then

$$\mathbf{x}_e = P_n^e \mathbf{x}, \mathbf{x}_o = P_n^o \mathbf{x}.$$

Upon using the notation  $(\mathbf{v})_\ell$  for the  $\ell^{\text{th}}$  component of the vector  $\mathbf{v}$ , the result in (2.1.3) is simply

$$\begin{aligned} \hat{x}_\ell &= (F_n \mathbf{x}_e)_\ell + e^{-i\pi\ell/n} (F_n \mathbf{x}_o)_\ell \\ &= (F_n P_n^e \mathbf{x})_\ell + e^{-i\pi\ell/n} (F_n P_n^o \mathbf{x})_\ell \\ &= \begin{cases} (F_n P_n^e \mathbf{x})_\ell + e^{-i\pi\ell/n} (F_n P_n^o \mathbf{x})_\ell, & \text{for } 0 \leq \ell \leq n-1, \\ (F_n P_n^e \mathbf{x})_\ell - e^{-i\pi(\ell-n)/n} (F_n P_n^o \mathbf{x})_\ell, & \text{for } n \leq \ell \leq 2n-1. \end{cases} \end{aligned} \quad (2.1.4)$$

In (2.1.4), we have used the fact that  $e^{-i\pi\ell/n} = -e^{-i\pi(\ell-n)/n}$ . Thus,

$$(F_{2n}\mathbf{x})_\ell = \begin{cases} ((F_n P_n^e + D_n F_n P_n^o)\mathbf{x})_\ell, & \text{for } 0 \leq \ell \leq n-1, \\ ((F_n P_n^e - D_n F_n P_n^o)\mathbf{x})_\ell, & \text{for } n \leq \ell \leq 2n-1, \end{cases}$$

which yields (2.1.2). ■

Let  $n = 2^m$  with  $m \geq 1$ . For each  $k$ ,  $0 \leq k \leq m-1$ , let  $G_k^m$  be the  $2^m \times 2^m$  matrix defined by

$$G_k^m = \text{diag}\{\underbrace{E_{2^{m-k}}, \dots, E_{2^{m-k}}}_{2^k \text{ copies}}\} = \begin{bmatrix} E_{2^{m-k}} & & O \\ & \ddots & \\ O & & E_{2^{m-k}} \end{bmatrix}, \quad (2.1.5)$$

where, according to (2.1.1),  $E_{2^{m-k}}$  is a  $2^{m-k} \times 2^{m-k}$  matrix:

$$E_{2^{m-k}} = \begin{bmatrix} I_{2^{m-k-1}} & \vdots & D_{2^{m-k-1}} \\ \dots\dots\dots & & \\ I_{2^{m-k-1}} & \vdots & -D_{2^{m-k-1}} \end{bmatrix}.$$

Furthermore, let

$$P_{2^m} = \begin{bmatrix} P_{2^{m-1}}^e \\ \vdots \\ P_{2^{m-1}}^o \end{bmatrix}$$

denote the  $2^m \times 2^m$  “permutation matrix” in (2.1.2) with  $n = 2^m$ ; and define, inductively, the permutation matrices  $\tilde{P}_1 = \tilde{P}_{2^0}$ ,  $\tilde{P}_2 = \tilde{P}_{2^1}$ ,  $\tilde{P}_4 = \tilde{P}_{2^2}$ ,  $\dots$ ,  $\tilde{P}_{2^m}$  by

$$\tilde{P}_1 = [1], \dots, \tilde{P}_{2^\ell} = \begin{bmatrix} \tilde{P}_{2^{\ell-1}} & O \\ O & \tilde{P}_{2^{\ell-1}} \end{bmatrix} P_{2^\ell},$$

where  $\ell = 1, 2, \dots, m$ .

**Example 2.1.1** Consider  $n = 4, m = 2$ . We have

$$G_0^2 = E_{2^2} = E_4,$$

$$G_1^2 = \text{diag}\{E_2, E_2\} = \begin{bmatrix} E_2 & O \\ O & E_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

and

$$\begin{aligned}
 P_2 &= \begin{bmatrix} P_1^e \\ P_1^o \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\
 \tilde{P}_2 &= \begin{bmatrix} \tilde{P}_1 & 0 \\ 0 & \tilde{P}_1 \end{bmatrix} P_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} P_2 = P_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\
 P_4 &= \begin{bmatrix} P_2^e \\ P_2^o \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \\
 \tilde{P}_4 &= \begin{bmatrix} \tilde{P}_2 & O \\ O & \tilde{P}_2 \end{bmatrix} P_4 = \begin{bmatrix} I_2 & O \\ O & I_2 \end{bmatrix} P_4 = P_4.
 \end{aligned}$$

■

### 2.1.3 FFT for fast computation

In view of the Lanczos matrix factorization result established in Subunit 2.1.2, we can now derive the following fast Fourier transform (FFT) computational scheme.

**Theorem 2.1.2** *Let  $n = 2^m$ , where  $m \geq 1$  is an integer. Then the  $n$ -point DFT has the formulation*

$$F_n = F_{2^m} = G_0^m G_1^m \cdots G_{m-1}^m \tilde{P}_{2^m}. \quad (2.1.6)$$

**Proof** Proof of the FFT scheme (2.1.6) can be carried out by mathematical induction on  $m = 1, 2, \dots$

For  $m = 1$ , since

$$G_0^1 = E_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

and  $\tilde{P}_2 = I_2$  as shown in Example 2.1.1, it follows that

$$G_0^1 \tilde{P}_2 = G_0^1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = F_2,$$

which is (2.1.6) for  $m = 1$ .

In general, by Theorem 2.1.1 and the induction hypothesis, we have

$$\begin{aligned}
F_{2^m} &= E_{2^m} \begin{bmatrix} F_{2^{m-1}} & O \\ O & F_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} \begin{bmatrix} G_0^{m-1} \cdots G_{m-2}^{m-1} \tilde{P}_{2^{m-1}} & O \\ O & G_0^{m-1} \cdots G_{m-2}^{m-1} \tilde{P}_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} \begin{bmatrix} G_0^{m-1} & O \\ O & G_0^{m-1} \end{bmatrix} \cdots \begin{bmatrix} G_{m-2}^{m-1} & O \\ O & G_{m-2}^{m-1} \end{bmatrix} \begin{bmatrix} \tilde{P}_{2^{m-1}} & O \\ O & \tilde{P}_{2^{m-1}} \end{bmatrix} P_{2^m} \\
&= E_{2^m} G_1^m G_2^m \cdots G_{m-1}^m \tilde{P}_{2^m} \\
&= G_0^m G_1^m \cdots G_{m-1}^m \tilde{P}_{2^m},
\end{aligned}$$

since  $G_0^m = E_{2^m}$  by (2.1.5), and

$$\begin{bmatrix} G_{k-1}^{m-1} & O \\ O & G_{k-1}^{m-1} \end{bmatrix} = G_k^m, \quad (2.1.7)$$

for  $1 \leq k \leq m-1$ . ■

**Example 2.1.2** Verify (2.1.6) for  $n = 4, n = 8$ .

**Solution** For  $n = 4$ , we have, from Theorem 2.1.1, that

$$F_4 = E_4 \begin{bmatrix} F_2 & O \\ O & F_2 \end{bmatrix} \begin{bmatrix} P_2^e \\ P_2^o \end{bmatrix} = E_4 \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} P_4 = G_0^2 G_1^2 \tilde{P}_4,$$

where the last equality follows from  $G_0^2 = E_4$ , the expression of  $G_1^2$  and  $P_4 = \tilde{P}_4$  as shown in Example 2.1.1. This is (2.1.6) for  $n = 4$ .

Similarly, for  $n = 8$ , we have

$$\begin{aligned}
F_8 &= E_8 \begin{bmatrix} F_4 & O \\ O & F_4 \end{bmatrix} \begin{bmatrix} P_4^e \\ P_4^o \end{bmatrix} = E_8 \begin{bmatrix} G_0^2 G_1^2 \tilde{P}_4 & O \\ O & G_0^2 G_1^2 \tilde{P}_4 \end{bmatrix} P_8 \\
&= E_8 \begin{bmatrix} G_0^2 & O \\ O & G_0^2 \end{bmatrix} \begin{bmatrix} G_1^2 & O \\ O & G_1^2 \end{bmatrix} \begin{bmatrix} \tilde{P}_4 & O \\ O & \tilde{P}_4 \end{bmatrix} P_8 \\
&= G_0^3 G_1^3 G_2^3 \tilde{P}_8,
\end{aligned}$$

where the last equality follows from (2.1.7). ■



## 2.2 Discrete Cosine Transform (DCT)

The discrete cosine transform (DCT) of a vector is the discrete version of the coefficients of the Fourier cosine series of a function on a bounded interval (to be studied in Subunit 3.1.1, with formula given in (3.1.8)). In Subunit 2.2.1, the DCT is derived from the DFT, and in Subunit 2.2.2, the example of 8-point DCT is given. The DCT is then extended to two dimensions in Subunit 2.2.3 to define the DCT of a matrix. This is important for applications to image compression, since a digital image block is nothing but a rectangular matrix, and the 2-dimensional DCT of this image block reveals its frequency content (for applying quantization to decrease its entropy) to be studied in Subunit 2.2.5.

### 2.2.1 Derivation of DCT from DFT

#### References

- (1) Stanford University: Department of Electrical Engineering's "Lecture 1: The Fourier Transforms and Its Applications (YouTube).
- (2) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 179–189. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com)

### 2.2.2 8-Point DCT

#### References

- (1) National Program on Technology Enhanced Learning's "Lecture 17: Lossy Image Compression: DCT (YouTube).
- (2) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 197–199 and pages 203–204. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 2.2.3 2-Dimensional DCT

To apply the DCT to data sets in higher dimensions, we may consider one dimension at a time. For example, to apply an  $n$ -point DCT to 2-dimensional data sets, such as digital images, we may first apply the transform in the horizontal direction, followed by the same transform in the vertical direction, as follows.

**Definition 2.2.1** *Let  $A$  be an  $m \times n$  data matrix. The 2-dimensional DCT of  $A$  is defined by the  $n$ -point DCT  $C_n$  of the transpose  $A^T$  of  $A$ , followed by the  $m$ -point DCT  $C_m$  of the transpose of  $C_n A^T$ ; so that the DCT of the data matrix  $A$  is defined by*

$$\hat{A} = C_m(C_n A^T)^T = C_m A C_n^T. \quad (2.2.1)$$

Furthermore, the corresponding inverse 2D-DCT is given by

$$A = C_m^T \hat{A} C_n. \quad (2.2.2)$$

The reason for the need of taking matrix transposes in (2.2.1) is that in implementation, the 1-dimensional DCT is operated on rows of the data matrix.

**Example 2.2.1** Compute the 2-dimensional DCT of the  $2 \times 2$  matrix

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}.$$

**Solution** First, it can be easily shown that the 2-point DCT is given by

$$C_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Hence, it follows from (2.2.1) of Definition 2.2.1 that the 2-dimensional DCT of  $A$  is given by

$$\begin{aligned} \hat{A} &= C_2 A C_2^T = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ -1 & -1 \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 2 & -2 \\ 4 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 2 & 0 \end{bmatrix}. \end{aligned}$$

■

**Example 2.2.2** Compute the 2-dimensional DCT of the rectangular matrix

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

**Solution** From the definition of the  $n$ -point DCT for  $n = 3$ , it can be shown by direct computation that

$$C_3 = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & 0 & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & -\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Hence, it follows from (2.2.1) of Definition 2.2.1 that the 2-dimensional DCT of  $A$  is given by

$$\begin{aligned} \hat{A} &= C_2 A C_3^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} C_3^T \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 1 & 1 \\ -2 & -1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\sqrt{\frac{2}{3}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix} \\ &= \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{2}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{-2}{\sqrt{3}} & \frac{-3}{\sqrt{2}} & \frac{1}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} \sqrt{\frac{2}{3}} & \frac{-1}{2} & \frac{-1}{2\sqrt{3}} \\ -\sqrt{\frac{2}{3}} & \frac{-3}{2} & \frac{1}{2\sqrt{3}} \end{bmatrix}. \end{aligned}$$

■

To formulate the 2-dimensional DCT and inverse DCT in terms of sums of products (without matrix-to-matrix multiplications), we may write out the  $j^{\text{th}}$  row of the DCT, namely:

$$\mathbf{c}_j^T = d_j \sqrt{\frac{2}{n}} \left[ \cos \frac{j\pi}{2n} \cos \frac{j3\pi}{2n} \cdots \cos \frac{j(2n-1)\pi}{2n} \right]$$

for  $j = 0, \dots, n-1$ , where

$$d_0 = \frac{1}{\sqrt{2}}; \quad d_1 = \cdots = d_{n-1} = 1.$$

Then for a given  $n \times n$  square matrix  $A$ , with 2-dimensional DCT denoted by

$$\hat{A} = C_n A C_n^T,$$

as defined in (2.2.1), we have, by using the notation:

$$A = [a_{j,k}]_{0 \leq j, k \leq n-1}; \quad \hat{A} = [\hat{a}_{\ell,s}]_{0 \leq \ell, s \leq n-1},$$

that

$$\begin{aligned}\hat{a}_{j,k} &= \frac{2}{n} \sum_{\ell=0}^{n-1} \sum_{s=0}^{n-1} \left( d_j \cos \frac{j(2\ell-1)\pi}{2n} \right) a_{\ell,s} \left( d_k \cos \frac{k(2s-1)\pi}{2n} \right) \\ &= \frac{2}{n} d_j d_k \sum_{\ell=0}^{n-1} \sum_{s=0}^{n-1} \left( \cos \frac{j(2\ell-1)\pi}{2n} \cos \frac{k(2s-1)\pi}{2n} \right) a_{\ell,s},\end{aligned}\quad (2.2.3)$$

for  $j, k = 0, 1, \dots, n-1$ ; and

$$\begin{aligned}a_{\ell,s} &= \frac{2}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \left( d_j \cos \frac{j(2\ell-1)\pi}{2n} \right) \hat{a}_{j,k} \left( d_k \cos \frac{k(2s-1)\pi}{2n} \right) \\ &= \frac{2}{n} \sum_{j=0}^{n-1} \sum_{k=0}^{n-1} \left( d_j d_k \cos \frac{j(2\ell-1)\pi}{2n} \cos \frac{k(2s-1)\pi}{2n} \right) \hat{a}_{j,k},\end{aligned}\quad (2.2.4)$$

for  $\ell, s = 0, \dots, n-1$ .

## 2.3 Information Coding

The difference between data reduction, including data dimensionality reduction studied in Subunit 1.5, and the topic of data compression studied in the present Unit 2 is that compressed data must be recoverable, at least approximately. The most commonly used representation of data (particularly compressed data) for communication and storage is a string of numbers consisting only of zeros, 0's, and ones, 1's, without using any punctuation. This string of 0's and 1's is called a "binary code" of the data. For the recovery of the data, the binary code must be decipherable by referring to the corresponding codetable. The length of the binary code depends on coding efficiency, which is governed by the "entropy" of the source data, a notion to be introduced and studied in Subunit 2.3.3, in terms of the probability distribution of the data.

### 2.3.1 Probability distributions

Probability is a measure of the chance of success or failure of an "outcome" (such as a bet) from empirical evidence, resulting from inductive reasoning and statistical inference. It is an estimation of how likely (or unlikely) it is for the outcome to happen. For some situations, this estimation could be measured quantitatively by some real number between 0 and 1, called the "probability"

(or probability value) of occurrence of the outcome. If the probability (value) is 0, the chance for the outcome to happen is 0%. On the other hand, if the probability is 1, then there is a 100% chance for the outcome to take place. The larger the probability value (between 0 and 1), the more likely the outcome is to happen.

In some situations, probability values can be computed, at least under the “fairness” condition. For example, when a fair coin is tossed twice, the probability value for each of the four different outcomes: HH, HT, TH and TT (where H stands for head, and T stands for tail), is  $1/4$ , because the chance for each of the four (and only four) outcomes to happen is the same, or 25%. More generally, again under the “fairness” assumption, the probability value can be computed by dividing the number of desired outcomes with the total number of all possible outcomes.

As another example, let us consider the outcomes of rolling fair dice. To be specific, we assume that each of the fair dice is a rounded cube with six faces, engraved with different numbers of dots, that range from 1 dot to 6 dots. For instance, when two dice are rolled and come to rest, a pair of random numbers is generated, as given by the number of dots on each of the two top faces. This pair is called an “outcome”, and there are precisely 21 possible different outcomes in total, namely:

$$(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 2), (2, 3), (2, 4), (2, 5), \\ (2, 6), (3, 3), (3, 4), (3, 5), (3, 6), (4, 4), (4, 5), (4, 6), (5, 5), (5, 6), (6, 6).$$

For a gambler to bet on a certain number (of the dots), he/she places a bet on the sum of the pair of two numbers randomly generated by rolling the dice. This is a positive integer  $n$ , with  $n = 2, \dots, 12$ . It is not difficult to tabulate the number,  $c_n$ , of outcomes that generate the number  $n$ . Precisely, we have

$$c_2 = c_3 = 1, c_4 = c_5 = 2, c_6 = c_7 = c_8 = 3, c_9 = c_{10} = 2, c_{11} = c_{12} = 1.$$

The reason is that the only possibilities to achieve  $n$  by adding two numbers between 1 and 6 are:

$$\begin{aligned} 2 &= 1 + 1, & 3 &= 1 + 2, & 4 &= 1 + 3 = 2 + 2, & 5 &= 1 + 4 = 2 + 3, \\ 6 &= 1 + 5 = 2 + 4 = 3 + 3, & 7 &= 1 + 6 = 2 + 5 = 3 + 4, \\ 8 &= 2 + 6 = 3 + 5 = 4 + 4, & 9 &= 3 + 6 = 4 + 5, \\ 10 &= 4 + 6 = 5 + 5, & 11 &= 5 + 6, & 12 &= 6 + 6, \end{aligned}$$

and there are no other combinations. Hence, the probability (value),  $p_n$ , of winning the bet on the number  $n$  is given by

$$p_n = \frac{\text{number of desired outcomes}}{\text{total number of possible outcomes}} = \frac{c_n}{21}; \quad (2.3.1)$$

so that

$$\begin{aligned} p_1 &= 0, & p_2 &= p_3 = p_{11} = p_{12} = \frac{1}{21}, & p_4 &= p_5 = p_9 = p_{10} = \frac{2}{21}, \\ p_6 &= p_7 = p_8 = \frac{3}{21}. \end{aligned}$$

Hence, if the payoff would be the same for the bet of any random number  $n$ , the gambler is advised to bet on one of the three numbers: 6, 7, 8. Only a fool would bet on the number 1, with probability  $p_1 = 0$ .

In the above example, we have listed 12 probability values,

$$p_1, \dots, p_{12}.$$

Observe that each value is between 0 and 1, and the sum of these values is precisely equal to 1. The set  $\{p_1, \dots, p_{12}\}$  is called a probability distribution, and more precisely, a discrete probability distribution, as follows.

**Definition 2.3.1** *Let  $n$  be any positive integer, and  $p_1, \dots, p_n$  be real numbers that satisfy both  $0 \leq p_1, \dots, p_n \leq 1$  and*

$$p_1 + \dots + p_n = 1. \quad (2.3.2)$$

*Then the set  $S_n = \{p_1, \dots, p_n\}$  is called a (discrete) probability distribution.*

We remark that although the index set  $I_n = 1, \dots, n$  can be replaced by any set  $x_1, \dots, x_n$  of real numbers, we prefer to simplify notations by only using  $I_n$ . After all, the main purpose of this index set  $I_n$  is to represent the set of  $n$  outcomes, with probability distribution  $S_n$ . For example, in the above discussion of tossing a fair coin twice, the integers 1, 2, 3, 4 of the index set  $I_4$  represent the outcomes HH, HT, TH, TT, respectively, with probability distribution  $S_4 = \{p_1, p_2, p_3, p_4\}$ , where  $p_1 = p_2 = p_3 = p_4 = 1/4$ . Also, in our discussion of rolling two fair dice at the same time to arrive at the outcomes  $c_1, \dots, c_{12}$ , the integer  $j \in I_{12}$  represents the outcome  $c_j$  for each  $j = 1, \dots, 12$ . The importance of the discrete probability distribution is that it is used to quantify how likely (or unlikely) any outcome is to take place. This is described by using the notion of a random variable  $X$  governed by the discrete probability distribution, with range in  $I_n$ , as follows.

**Definition 2.3.2** *Let  $n$  be any positive integer, and  $S_n = \{p_1, \dots, p_n\}$  be a probability distribution with the index set  $I_n = \{1, \dots, n\}$ . A non-deterministic function  $X$  is called a random variable governed by  $S_n$ , if the range of  $X$  is the index set  $I_n$ , so defined that the probability of  $X$  equal to  $j$  is  $p_j$ , for all  $j \in I_n$ , namely:*

$$\mathcal{P}\{X = j\} = p_j, \quad (2.3.3)$$

*for  $j = 1, \dots, n$ .*

We emphasize that the random variable  $X$  has no specific function values. There is some uncertainty for  $X$  to be equal to  $j \in I_n$ , in that the chance for  $X = j$  to hold is  $100 \times p_j\%$ . There are many useful discrete probability distributions dictated by certain rules. For example, tossing a fair coin  $k$  times yields a sequence of H and T of length equal to  $k$ . Each sequence is an outcome, and there are  $n = 2^k$  different outcomes, with the same probability for each outcome. In other words, the discrete probability distribution is the following notion of uniform distribution.

**Definition 2.3.3** *Discrete uniform probability distribution:*

$$S_n = \{p_1, \dots, p_n\}, \quad p_1 = \dots = p_n = \frac{1}{n}. \quad (2.3.4)$$

Note that the ordering of H and T is taken into consideration in the definition of each outcome in arriving at the discrete uniform probability distribution. On the other hand, if only the number of heads ( $H$ ) and number of tails ( $T$ ) are counted without concern of the order of occurrence (e.g.  $HHT = HTH = THH$ ), then the number of outcomes reduces to  $n = k + 1$ , with the so-called

**Definition 2.3.4** *Binomial probability distribution:*

$$S_n = \{p_1, \dots, p_n\}, \quad p_{j+1} = \frac{1}{2^{n-1}} \binom{n-1}{j}, \quad (2.3.5)$$

for  $j = 0, \dots, n-1$ .

In general, if the coin is not necessarily fair, with probability  $s$  for the occurrence of  $H$ , where  $0 < s < 1$ , then we have the following

**Definition 2.3.5** *Bernoulli probability distribution:*

$$S_2 = \{p_1 = s, p_2 = 1 - s\}, \quad (2.3.6)$$

with  $0 < s < 1$ .

When this unfair coin is tossed  $k = n-1$  times and each outcome is defined by the number of heads ( $H$ ) without consideration of the order of  $H$  and  $T$  in the sequence; that is, when the so-called Bernoulli trial is performed, we have the following

**Definition 2.3.6** *General binomial probability distribution:*

$$S_n = \{p_1, \dots, p_n\}, \quad p_{j+1} = \binom{n-1}{j} s^j (1-s)^{n-j-1}, \quad (2.3.7)$$

where  $j = 0, \dots, n-1$  and  $0 < s < 1$ .

However, in many applications there are no rules to govern the discrete probability distribution, which could be tabulated by performing many tireless experiments. For instance, for lossless (or reversible) compression of 8-bit digital gray-scale images, each pixel of an image, with resolution  $m \times n$ , is an integer  $j \in \{0, \dots, 255\}$ . The  $k^{\text{th}}$  row of the image, with pixel values given by

$$x_{k,0}, \dots, x_{k,n},$$

is mapped to the sequence

$$y_{k,0}, \dots, y_{k,n},$$

with

$$y_{k,0} = x_{k,0}, y_{k,1} = x_{k,1} - x_{k,0}, \dots, y_{k,n} = x_{k,n} - x_{k,n-1}.$$

This mapping is called DPCM (differential pulse code modulation). The key properties of DPCM are: firstly, it is reversible, and secondly, the discrete probability distribution of the outcomes  $y_{k,0}, \dots, y_{k,n}$  is much less uniform than that of the original sequence  $x_{k,0}, \dots, x_{k,n}$  and with smaller integers (though some are negative). Such probability distributions can be encoded with shorted average code-lengths, to be studied in Subunits 2.3, 2.3.4, and 2.4.4. The probability distributions can be tabulated by using the histogram of the DPCM encoded sequences, to be discussed in the next subunit.

We conclude this subunit by introducing the notions of the expected value and variance of a random variable governed by a discrete probability distribution.

**Definition 2.3.7** Let  $S_n = \{p_1, \dots, p_n\}$  be a discrete probability distribution with the index set  $I_n = \{1, \dots, n\}$ . The expected value of the random variable  $X$  defined in (2.3.3) is defined by

$$\mu = E[X] = \sum_{j=1}^n j p_j, \quad (2.3.8)$$

and the variance of  $X$  by

$$\text{Var}[X] = E[(X - \mu)^2]. \quad (2.3.9)$$

In view of (2.3.8), the definition of variance can be re-formulated as

$$\text{Var}[X] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - \mu^2.$$

For example, the expected value of the random variable for the probability distribution in (2.3.4) is  $\mu = \frac{n+1}{2}$ , while that for the probability distribution in (2.3.7) is  $\mu = s(n-1)$ . In particular, by setting  $s = \frac{1}{2}$ , it follows that the expected value of the random variable for the binomial distribution in (2.3.5) is  $\mu = \frac{n-1}{2}$ .

In addition, the variance of  $X$  for the probability distributions in (2.3.7) is given by

$$\text{Var}[X] = (n-1)s(1-s) \quad (2.3.10)$$

and hence, by setting  $s = \frac{1}{2}$ , the variance of  $X$  for the probability distributions in (2.3.5) is given by  $\text{Var}[X] = \frac{n-1}{4}$ . The above results can be verified by applying the binomial expansion of  $(1+x)^{n-1}$  and taking appropriate derivatives.

## 2.3.2 Histogram

Information Theory is an area of Applied Mathematics which is concerned



with such topics as the quantification, coding, communication, and storage of information. Here, the term “information” should not be confused with the word “meaning”, since in Information Theory, “pure nonsense” could be considered as an information source. In other words, “information” should not be measured in terms of “what is to be said”, but only of “what could be said”. As Claude Shannon, the father of Information Theory, emphasized in his 1948 pioneering paper, the semantic aspects of the communication of information are irrelevant to mathematical and engineering considerations.

On the other hand, coding of an information source is essential for communication and storage of the information. Here, the term “coding” consists of two operations: “encoding” of the information source to facilitate effective communication and storage, and “decoding” to recover the original information source. In other words, decoding is the inverse operation of encoding. In Subunits 2.3.4 and 2.4.4, we will consider “binary coding”, meaning that the encoded information, called a binary code, is a (finite) “sequence” of numbers consisting only of zeros, 0’s, and ones, 1’s; and that all “code-words” and “instructions” that constitute this so-called sequence can be identified from some “code-table”, even though no punctuation are used in this “sequence” to separate them. Furthermore, all elements of the set of the information source are represented as code-words in the code-table. Here again, the term “sequence” is in quotation, since no “commas” are allowed, as opposed to the usual mathematical definition of sequences.

Hence, if each element of the given set of information source is assigned some non-negative integer, all of such integers can be, in turn, assigned some “code-words” contained in a “code-table”, then decoding to receive and recover the original information source from the binary code is possible by using the code-table. To quantify the integer representation of (the elements of the set of) the information source, the unit “bit”, coined by John Tukey, as the abbreviation of “binary digits”, is used. For example, if the largest (non-negative) integer used in the integer representation of the information source is 255, we say that the source is an 8-bit information source (since the binary representation of 255 is 11111111, a string of 8 ones). In other words, one 0 or one 1 is one bit (or 1 b). In addition, the measurement in the unit of “bytes” is also used. One byte (1B) is equal to 8 bits (1B = 8b). In practice, the unit “bit” is used for transmission of the encoded data, and the unit “byte” is used for storage of the encoded data.

When a binary code is stored in some memory device (such as hard disk, flash memory, or server), the length of the “sequence” of 0’s and 1’s is called the “file size” of the encoded data. When the binary code is transmitted (such as in broadcasting or video streaming), the length of the sequence of 0’s and 1’s is called the length of the bit-stream, and the speed of transmission is called the “bit-rate”. While file sizes are measured in kilo-bytes, mega-bytes, giga-bytes, and tera-bytes (or kB, MB, GB, and TB, respectively), bit-rates are measured in kilo-bits per second, mega-bits per second, and giga-bits per second (or kb/s, Mb/s, Gb/s, respectively).

A typical example is an 8-bit gray-scale digital image as discussed in the previous subunit, with the “intensity” of each pixel (considered as an element of the set of the information source, which is the image) being calibrated in increasing integer steps from 0 to 255, with 0 representing “black” (or no light) and 255 representing “white”, while for  $j = 1, \dots, 254$ , the increase in intensity yields increasingly lighter shades (of gray). Another example is a 24-bit color digital image. But instead of using 24 bits in the binary expression of the largest integer used in the integer representation of the information source (which is a color image), the convention is to assign 8-bits to each of the three primary color components, “red” (R), “green” (G), and “blue” (B), in that a triple  $(i, j, k)$  of integers, with each of  $i, j$ , and  $k$ , ranging from 0 to 255, for (R, G, B) represents increasing intensities of the red, green, and blue (visible) lights, respectively. Recall that as primary additive colors, addition of different intensities for R, G, B (that is, different values of  $i, j, k$ ) yield  $2^8 \times 2^8 \times 2^8 = 2^{24}$  colors of various shades, as discussed in Subunit 1.5.3. In particular, a gray-scale digital image is obtained by setting  $i = j = k$ , where  $0 \leq i \leq 255$ . Therefore, the term “24-bit color” is justified.

It must be understood that the meaning of a 12-bit novel only indicates an upper bound of the number of different words being used in the novel. The actual encoded file size is significantly larger, and often quantified in megabytes. A typical compressed file size of a novel is about 1 MB. Similarly, by a 24-bit color picture, we only mean that the quality (in terms of shades of color) is limited to 24 bits. The file size of a JPEG compressed image usually exceeds several kilo-bytes and occasionally even over 1 MB, depending on the image resolution (that is, the number of pixels).

In the above examples, it should be clear that the notion of probability, as studied in the previous subunit (in the sense of percentages of equal pixel values for a digital image and percentages of the same word being used in the novel), plays an important role in information coding. For instance, in a picture with blue sky and blue water in the background, the percentages of RGB pixels with values of  $(0, 0, k)$ , where  $50 \leq k \leq 150$ , are much higher than those with  $(i, j, 0)$  for all  $0 \leq i, j \leq 255$ . Also, for the novel example mentioned above, the frequency of occurrence of the word “the” is much higher than just about all of the other words in the novel. In fact, according to some study, less than half of the vocabulary used in a typical novel constitutes over 80% of all the words in the book.

Furthermore, it should be emphasized that the notion of “entropy” of the information source (defined in terms of the probabilities of occurrence as discussed above) often decreases when some suitable mathematical transformation is applied to the integer representation of the information source. Typical transforms include RLE (run-length encoding) and DPCM (differential pulse code modulation, as already mentioned in the previous subunit, and to be discussed briefly in Example 2.3.1 in this sub-section. Since the encoded file size and length of an encoded bit-stream are governed by the entropy, it is important to understand this concept well.

**Definition 2.3.8** Let  $X_n = \{x_1, \dots, x_n\}$  be an information source (or some mathematical transformation of a given information source), and let  $Z = \{z_1, \dots, z_m\}$  be the subset of all distinct elements of the set  $X_n$ . Corresponding to each  $z_j \in Z, j = 1, \dots, m$ , let  $p_j$  denote some probability value associated with  $z_j \in Z_n$ , such that  $S_m = \{p_1, \dots, p_m\}$  is a discrete probability distribution, as defined in (2.3.2) of the previous subunit. Then the function

$$H(S_m) = H(p_1, \dots, p_m) = \sum_{j=1}^m p_j \log_2 \frac{1}{p_j} \quad (2.3.11)$$

is called the entropy of the probability distribution  $S_m$  for the information source  $X_n$ .

The theory of entropy will be studied in Subunit 2.3.3.

**Remark 2.3.1** For most applications, since the  $\#X_n = n$  is very large and since many (or even most) information sources of the same application are quite similar, the same discrete probability distribution  $S_m$  is often used for all such information sources. This  $S_m$  is usually obtained from large volumes of previous experiments.

Of course, the precise values of  $p_1, \dots, p_m$  that constitute  $S_m$  can be defined and computed in terms of the “histogram” of  $X_n$ , to be defined below. This specific  $S_m$  will be called the optimal discrete probability distribution for  $X_n$ . For example, consider a digital gray-scale image with pixels  $x_j, j = 1, \dots, n$ , where  $n$  is the resolution of the image. Hence, for a 10-Mega pixel digital image,  $n = 10^7$ . On the other hand, most of these pixels are the same, since each  $x_j \in \{0, \dots, 255\}$ . Hence, the subset  $Z_m$  of distinct pixel values of  $X_n$  has at most 256 elements; or  $0 < m \leq 256$ . We will formulate the “optimal” probability  $S_m = \{p_1, \dots, p_m\}$  for the information source  $X_n$  in terms of its histogram, defined as follows.

**Definition 2.3.9** Let  $X_n = \{x_1, \dots, x_n\}$  be an information source, and  $Z = \{z_1, \dots, z_m\}$  be the subset of all distinct elements of the set  $X_n$ . For each  $j = 1, \dots, m$ , let  $g_j$  denote the cardinality (that is, the count of number of elements) of the set  $\{x_i \in X_n : x_i = z_j\}$ . Then the histogram of the data-set  $X_n$  is defined by

$$G_m = \{g_1, \dots, g_m\}. \quad (2.3.12)$$

Hence, the histogram of  $X_n$  is a set of positive integers. To formulate the optimal probability distribution for the information source  $X_n$ , we simply set

$$p_j = \frac{g_j}{n}, \quad (2.3.13)$$

for each  $j = 1, \dots, m$ . It is clear that  $S_m = \{p_1, \dots, p_m\}$ , with  $p_1, \dots, p_m$  defined in (2.3.13) satisfies (2.3.2), and is therefore a probability distribution.

**Example 2.3.1** Consider the information source  $X = \{0, 1, \dots, 255\}$ . Discuss its histogram and the corresponding probability distribution. Also, introduce a reversible transformation of  $X$  to another set  $Y$ , with a much more compact histogram, that facilitates significant reduction of the encoded bit-stream.

**Solution** The set  $X$  can be written as  $X = \{x_0, \dots, x_{255}\}$  with  $x_j = j$  for  $j = 0, \dots, 255$ . Observe that since the elements of  $X$  are distinct, the histogram is simply  $\{1, \dots, 1\}$ , so that the discrete probability distribution is uniform, in that

$$p_1 = \dots = p_n = \frac{1}{n}.$$

This is extremely costly to code. The reason is that if no commas are used to separate the code-words, then we need to use all 8 bits to encode each  $x_j$ , namely:  $x_0 = 00000000, x_1 = 00000001, \dots, x_{255} = 11111111$ . Therefore the encoded bit-stream is 000000000000000010...011111111, which has a length of  $256 \times 8 = 2048$  bits.

On the other hand, if we set  $y_0 = 0$  and

$$y_j = x_j - x_{j-1} - 1, \text{ for } j = 1, \dots, 255, \quad (2.3.14)$$

then we have  $y_j = 0$ , for all  $j = 0, \dots, 255$ . Therefore, the histogram becomes a singleton  $G_1 = \{256\}$ , with corresponding discrete probability distribution,  $S_1 = \{p_1\} = \{1\}$ . Observe that the histogram is most compact, yielding the one-value probability distribution. The encoded bit-stream is a string of 256 zeros. This code can be further shortened by coding the size of the block of 256 zeros, by applying “run-length encoding (RLE)”, to be discussed later in Subunit 2.5.3, in the study of digital image compression. Also, observe that the transformation from  $X$  to  $Y$  is reversible, since

$$x_j = y_j + x_{j-1} + 1, \text{ } j = 1, \dots, 255, \quad (2.3.15)$$

for  $y_j = 0, j = 0, \dots, 255$ , with initial condition  $x_0 = 0$ . Of course the formula (2.3.15) along with the initial value  $x_0 = 0$  must be coded, but this requires only a few bits. We remark that the code of the transformation in (2.3.14) is called DPCM (differential pulse code modulation), as already mentioned in the previous subunit. Both RLE and DPCM are commonly used, and are part of the JPEG coding scheme for image compression, to be studied in Subunit 2.5.3. ■

### 2.3.3 Entropy

#### References

- (1) Web Media: CSLearning101’s “Huffman Coding Tutorial (YouTube).

- (2) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 208–217.. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 2.3.4 Binary codes

Before the study of binary codes, we give a brief review of binary representations of non-negative integers. In other words, we will use base 2, instead of using base 10, to represent the natural numbers:  $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, \dots, n, \dots$ . For any natural number  $n$ , the subscript 10 is used to indicate that it is the customary base-ten representation, namely:  $n = n_{10}$ . Similarly, the subscript 2 will be used to indicate it is the binary representation  $n$ . For example,

$$\begin{aligned} 1 &= 1_{10} = 1_2, & 2 &= 2_{10} = 10_2, & 3 &= 3_{10} = 11_2, & 4 &= 4_{10} = 100_2, \\ 5 &= 5_{10} = 101_2, & 6 &= 6_{10} = 110_2, & 7 &= 7_{10} = 111_2, & 8 &= 8_{10} = 1000_2, \\ 9 &= 9_{10} = 1001_2, \dots, & 255 &= 255_{10} = 11111111_2, \\ 256 &= 256_{10} = 100000000_2, \dots \end{aligned}$$

**Remark 2.3.2** When it is clear that binary representation is used for all natural numbers, the subscript 2 is dropped for convenience. Observe firstly, that the first bit of the binary representation of any natural number is always 1; and secondly, in adding two binary representations, there is a carry of 1 to the next column (i.e., the column to the left), when adding two 1s on the same column. For example,  $11 + 10 = 101$ , since  $0 + 1 = 1$  on the first column (i.e. the column on the right), and  $1 + 1 = 10$ , with a carry of 1, when adding the two 1s in the second column.

In the following, we present a method for converting any positive number  $n$  to its binary representation  $n_2$ . The method is an instruction for step-by-step computations. Such instructions are called algorithms.

**Remark 2.3.3** Let  $k$  be any natural number. Suppose that  $k$  is even.

Step 1. Write  $k = a_0 2^{j_0}$ , where  $a_0$  is an odd number.

Step 2 . If  $a_0 = 1$ , stop. If  $a_0 > 1$ , since  $a_0 - 1$  is even, we repeat Step 1 to write

$$a_0 - 1 = a_1 2^{j_1},$$

where  $a_1 \in \mathbb{N}$  is odd.

Step 3. If  $a_1 = 1$ , stop. For  $\ell = 1, 2, \dots$ , if  $a_\ell > 1$ , repeat Step 2 by writing

$$a_\ell - 1 = a_{\ell+1} 2^{j_{\ell+1}},$$

where  $a_{\ell+1}$  is odd, and stop if  $a_{\ell+1} = 1$ .

Suppose that after carrying out Step 2 for  $n$  times, we have  $a_n = 1$ . Then  $k$  can be written as

$$\begin{aligned} k &= 2^{j_0}[1 + (a_0 - 1)] = 2^{j_0}[1 + a_1 2^{j_1}] = 2^{j_0} + a_1 2^{j_0+j_1} = \dots \\ &= 2^{j_0} + 2^{j_0+j_1} + \dots + 2^{j_0+\dots+j_{n-1}} + a_n 2^{j_0+\dots+j_n} \\ &= 2^{j_0} + 2^{j_0+j_1} + \dots + 2^{j_0+\dots+j_{n-1}} + 2^{j_0+\dots+j_n}. \end{aligned}$$

Therefore the binary representation of the even integer  $k$  is given by

$$k = (k)_{10} = (\underbrace{10\dots0}_{j_n-1}1\dots\underbrace{10\dots0}_{j_\ell-1}10\dots\underbrace{010\dots0}_{j_0})_2.$$

On the other hand, suppose that  $k > 1$  is odd. Then the same procedure as described above applies to  $k = a_0$  in Step 2. Hence, the representation in Step 2 holds for  $j_0 = 0$ ; that is, we have

$$k = 1 + 2^{j_1} + \dots + 2^{j_1+\dots+j_n},$$

so that the binary representation of the odd integer  $k$  is given by

$$k = (k)_{10} = (\underbrace{10\dots0}_{j_n-1}1\dots\underbrace{10\dots0}_{j_\ell-1}10\dots\underbrace{010\dots0}_{j_1-1})_2.$$

**Remark 2.3.4** In practice, we emphasize again that if it is clear that the representation is binary, then the subscript 2 is omitted for convenience. For example,

$$254 = 11111110_2 = 11111110.$$

Let us now turn to the study of binary codes. The term codes (or coding) may mean different things, including: encryption, channel coding, and source coding. In this unit, we will be concerned only with source coding, meaning: encoding a given information source, that can be de-coded, uniquely, without any ambiguity, and unnecessary delay. An information source may be represented by a finite set  $X_n = \{x_1, x_2, \dots, x_n\}$  of symbols, which may be a message, a novel, a digital image, a digital video, and so forth. As discussed in Subunit 2.3.2, we may consider the subset  $Z_m = \{z_1, \dots, z_m\}$  of all distinct elements of  $X_n$  and some probability value  $p_j$  associated with  $z_j$  for all  $j = 1, \dots, m$ , such that  $S_m = \{p_1, \dots, p_m\}$  is a discrete probability distribution, as defined in (2.3.2). This set  $Z_m$  will be called the “source alphabet” of the source information  $X_n$ . Recall from Subunit 2.3.1 that the source alphabet  $Z_m = \{z_1, \dots, z_m\}$  can be represented by the index set  $I_m = \{1, \dots, m\}$ , and each  $j \in I_m$  may be considered as an outcome. Hence, when  $x_k \in X_n$ , where  $1 \leq k \leq n$ , is considered as a random variable, then the probability for  $x_k$  to be equal to  $z_j \in Z_m$  is given by

$$\mathcal{P}\{x_k = j\} = p_j,$$

as defined in (2.3.3). To construct a binary code for the source information  $X_n$ , we will use the “code alphabet”  $\{0, 1\}$ . Before doing so, let us first motivate our discussion by considering the “code alphabet”  $\{0, 1, 2\}$ . For this alphabet, observe that the set of “words”

$$\mathbb{C}_1 = \{01, 12, 010, 012\}$$

constructed by using this alphabet does not provide a code-table. The reason is that there could be different messages that can be represented by the same sentence. For example, 01012 could mean 01, 012 or 010, 12. So if the source alphabet is  $Z_4 = \{\text{go, no, way, ahead}\}$ , then by using the words 01, 12, 010, 012 to represent “go”, “no”, “way”, “ahead”, respectively, the sentence 01012 conveys two contradictory messages, namely: 01, 012 = “Go ahead” and 010, 12 = “No way”. Observe that the above table  $\mathbb{C}_1$  can be easily modified to yield a code-table, simply by deleting the “word” 01. That is, the set

$$\mathbb{C}_2 = \{12, 010, 012\}$$

is now a code-table, in the sense that every sentence made up by using the words in  $\mathbb{C}_2$  is uniquely decodable. For example,

$$0100100121201201012$$

represents one and only one sentence:

$$010, 010, 012, 12, 012, 010, 12,$$

even without using commas to separate the words. So if the source alphabets are “that” = 010, “is” = 012, and “good” = 12, then the sentence 0100100121201201012 reads “That that is good is that good”.

Also, observe that the “word” 01, in the table  $\mathbb{C}_1$  is a “prefix” of the “word” 010. By deleting all prefix words, we obtain a code-table, called a “prefix code”. On the other hand, not all code-tables are prefix codes. For example,

$$\mathbb{C}_3 = \{0, 01, 11\}$$

is a code-table, but not a prefix code, since the codeword 0 is a prefix of the codeword 01. It is easy to see that  $\mathbb{C}_3$  is a code-table. However, this is not a desirable code. Indeed, the sentence

$$0111 \dots 1$$

cannot be decoded till the entire sentence is read. The reason is that if there is an even number of 1’s, then the sentence reads:

$$0, 11, 11, 11, \dots, 11,$$

but if there is an odd number of 1’s, then the sentence reads:

$$01, 11, 11, 11, \dots, 11.$$

On the other hand, by replacing the codeword 01 with 10, the new codetable

$$\mathbb{C}_4 = \{0, 10, 11\}$$

is “instantaneous”, in that every sentence can be read without any delay. For example, the same sentence 0111...1 must be 0, 11, 11, 11, ..., 11 and the number of 1’s in this sentence must be even. Also, the sentence 1...101...1 must be one of the two sentences

$$11, \dots, 11, 0, 11, \dots, 11$$

and

$$11, \dots, 11, 10, 11, \dots, 11,$$

depending on whether the number of 1’s before the 0 is even or odd. In any case, the number of 1’s following the 0 must be even, without any ambiguity.

Observe that the code alphabet for the codes  $\mathbb{C}_3$  and  $\mathbb{C}_4$  is  $\{0, 1\}$ . Codes constructed by using the code alphabet  $\{0, 1\}$  are called “binary codes”, while codes, such as  $\mathbb{C}_2$ , constructed by using the code alphabet  $\{0, 1, 2\}$  are called “ternary codes”. Since binary codes are much more commonly used in applications, we only study binary codes in this course.

**Definition 2.3.10** *Let  $\mathbb{C}_n$  be a code-table, with index set  $I_n = \{1, \dots, n\}$  and with code-words  $c_1, \dots, c_n$ . Then the code-table  $\mathbb{C}_n$  is called a prefix-code, if for each (fixed)  $j \in I_n$ , any code-word  $c_k \in \mathbb{C}_n$  is not a prefix of  $c_j$ , for  $k \neq j$ . It is called an instantaneous code-table, if every code-word  $c_{\ell_k}$ ,  $1 \leq k \leq M$ , in an arbitrary bit-stream*

$$c_{\ell_1} c_{\ell_2} c_{\ell_3} c_{\ell_4} \cdots c_{\ell_M} \tag{2.3.16}$$

*can be decoded, as soon as this code-word  $c_{\ell_k}$  is read, even before the first bit of the next code-word  $c_{\ell_{k+1}}$  of the bit-stream (2.3.16) arrives.*

The following result, which assures that every prefix code is instantaneous and vice versa, can be easily established.

**Theorem 2.3.1** *A code-table  $\mathbb{C}$  is a prefix code, if and only if it is instantaneous.*

## 2.4 Data Compression Schemes

When an information volume is too large, the most sensible way is to compress it before sending it to a receiver or a storage device. Of course the compressed



information has to be recovered by the receiver or from the storage. The most common way of representing an information source is to use a string of 0's and 1's, called a bit-stream. Therefore, a bit-stream is a binary code, consisting of a sequence of code words, along with certain punctuations, to constitute such information as phases, sentences, paragraphs, chapters, etc.. A code table is needed to represent the actual words and punctuations in a unique way. Hence, both encoding the information into a binary code and decoding from the binary code to recover the original information requires the same code table. Data compression can be classified into two types: lossless compression and lossy compression. When a lossless compression scheme is applied to encode an information source, the same information can be recovered perfectly from the compressed bit-stream. In other words, a lossless compression scheme is reversible. On the other hand, when certain portions of the information source are less important or even irrelevant, they can be ignored by applying a lossy compression scheme to achieve a much shorter compressed bit-stream. For example, since noise is irrelevant, it should be separated and removed before encoding. This subunit is devoted to a discussion of data compression, with formulation of Kraft's inequality in Subunit 2.4.2 that governs the minimum lengths of admissible code words and presentation of the Huffman coding scheme in Subunit 2.4.3, with average code word length not exceeding 1 plus the entropy of the information source, as governed by the Noiseless coding theorem (to be discussed in Subunit 2.4.4), provided that the probability distribution  $p_j$  of the information source satisfies  $p_j = 2^{-m_j}$  for some positive integers  $m_j$ .

### 2.4.1 Lossless and lossy compression

#### References

- (1) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications," page 232. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).
- (2) National Program on Technology Enhanced Learning's "Lecture 19: Data Compression (YouTube).
- (3) National Program on Technology Enhanced Learning's "Lecture 17: Lossy Image Compression: DCT (YouTube).
- (4) National Program on Technology Enhanced Learning's "Lecture 18: DCT Quantization and Limitations (YouTube).

### 2.4.2 Kraft inequality

#### References

- (1) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, page 219. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 2.4.3 Huffman coding scheme

#### References

- (1) Web Media: CSLearning101’s “Huffman Coding Tutorial (YouTube).
- (2) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, page 222–230. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 2.4.4 Noiseless coding theorem

As already studied in Subunit 2.4.3, a necessary condition for a binary table (or dictionary)  $\mathbb{C}_n = \{c_1, \dots, c_n\}$  to be qualified as a binary code-table is that the lengths of the code-words are governed by Kraft’s inequality.

**Definition 2.4.1** *The length of a code-word  $c_j$  in a code-table  $\mathbb{C}_1 = \{c_1, \dots, c_n\}$  is the number of code alphabets; that is, the number of digits 0 and/or 1 in  $c_j$ . The notation for the length of the code-word  $c_j$  to be used in our discussions is:*

$$\ell_j = \text{length}(c_j). \quad (2.4.1)$$

A restriction for a given table  $\mathbb{C}_n$ , consisting of  $n$  words  $c_1, \dots, c_n$ , to be qualified as a code-table is that the lengths  $\ell_1, \dots, \ell_n$  of  $c_1, \dots, c_n$ , respectively, satisfy the following Kraft’s inequality (or more precisely, Kraft-McMillan’s inequality):

$$\sum_{j=1}^n 2^{-\ell_j} \leq 1. \quad (2.4.2)$$

Observe that if we choose  $c_1 = 0$  and  $c_2 = 1$  and include both of them in a table  $\mathbb{C}_n$ , then for the table  $\mathbb{C}_n$  to be a code-table, the size of  $\mathbb{C}_n$  must be  $n = 2$ . The reason is that

$$2^{-\ell_1} + 2^{-\ell_2} = \frac{1}{2} + \frac{1}{2} = 1,$$

which is already the upper bound in (2.4.2).

**Remark 2.4.1** Kraft only proved (2.4.2) for instantaneous (or prefix) codes in his 1949 MIT Master's Thesis in Electrical Engineering. Several years later, in 1956, McMillan removed the “instantaneous” restriction, by showing that all (decodable or decipherable) codes must satisfy (2.4.2).

**Remark 2.4.2** To assess the suitability of code-tables for encoding a given information source  $X = \{x_1, \dots, x_N\}$  with a relatively shorter bit-stream (or smaller file size), recall the notion of discrete probability distributions  $S_n = \{p_1, \dots, p_n\}$  associated with  $X$ . Assuming that a code-table  $\mathbb{C}_n$  is constructed for encoding  $X$ , then since each code-word  $c_j \in \mathbb{C}_n$  is constructed according to the frequency of occurrence of the source data  $x_k \in X$ , the value of each corresponding  $p_j \in S_n$  is positive and increases according to the increase of frequency of the occurrence of the same  $x_k \in X$  (see, for example, (2.3.13)). Therefore, to achieve a shorter encoded bit-stream, the length  $\ell_j$  of the code-word  $c_j \in \mathbb{C}_n$  should be chosen to be relatively shorter for larger values of  $p_j$ . For this reason, the values  $p_1, \dots, p_n \in S_n$  are chosen to be the weights to define the following weighted average of the lengths of code-words:

$$\text{avlength}(\mathbb{C}_n) = \text{avlength}\{c_1, \dots, c_n\} = \sum_{j=1}^n p_j \ell_j, \quad (2.4.3)$$

called “**average code-word length**” of  $\mathbb{C}_n$ , where  $\ell_j = \text{length}(c_j)$  as introduced in (2.4.1).

**Remark 2.4.3** In applications, the same probability distribution  $S_n$ , and hence, the same code-table  $\mathbb{C}_n$ , is constructed for a large class of information sources  $X$ , with different cardinalities  $N = \#X$ . In particular,  $N$  is usually much, much larger than  $n = \#S_n$ , where  $S_n$  is used as a discrete probability distribution associated with all such  $X$ . For example, the same code-table  $\mathbb{C}_n$ , called the “Huffman table”, to be discussed in Subunit 2.4.3, is used for most (if not all) JPEG compressed digital images and MPEG compressed videos, a topic of investigation in Subunit 2.5.

Let us return to Kraft's inequality (2.4.2) and observe that it governs the necessity of fairly long code-word lengths, when the number of code-words must be sufficiently large for many practical applications. However, it does not give a quantitative measurement of the code-word lengths. In the following, we first show that the entropy  $H(S_n)$  of a given discrete probability distribution  $S_n = \{p_1, \dots, p_n\}$ , with  $n$  being the size of the desired code-table  $\mathbb{C}_n = \{c_1, \dots, c_n\}$ , provides a useful measurement stick.

**Theorem 2.4.1** *Let  $\mathbb{C}_n = \{c_1, \dots, c_n\}$  be a code-table with code-word lengths  $\ell_j = \text{length}\{c_j\}$ ,  $j = 1, \dots, n$ . Then the average code-word length of  $\mathbb{C}_n$  is*

bounded below by the entropy of the desired discrete probability distribution that governs the construction of the code-table; namely,

$$H(S_n) \leq \text{avlength}(\mathbb{C}_n), \quad (2.4.4)$$

where  $\text{avlength}(\mathbb{C}_n)$  is defined in (2.4.3). Furthermore, equality in (2.4.4) is achieved if and only if both of the following conditions are satisfied:

$$p_j = \frac{1}{2^{k_j}}, \quad j = 1, \dots, n, \quad (2.4.5)$$

for some positive integers  $k_1, \dots, k_n$ ; and

$$\ell_j = k_j, \quad j = 1, \dots, n, \quad (2.4.6)$$

so that Kraft's inequality becomes equality.

**Proof** The proof of this theorem is an application of Kraft's inequality (2.4.2). Indeed, by setting

$$q_j = \frac{1}{C} 2^{-\ell_j}, \quad j = 1, \dots, n,$$

where

$$C = \sum_{j=1}^n 2^{-\ell_j},$$

it is clear that  $q_1 + \dots + q_n = 1$  and  $0 \leq q_j \leq 1$  for all  $j$ . Now, by an application of the variational method of Lagrange multipliers, it is not difficult to show that for all discrete probability distributions, including  $Q_n = \{q_1, \dots, q_n\}$ , the quantity  $G(Q_n)$ , defined by

$$G(Q_n) = \sum_{j=1}^n p_j \log_2 \frac{1}{q_j},$$

satisfies

$$G(Q_n) \geq G(S_n) = H(S_n), \quad (2.4.7)$$

and that equality holds, if and only if  $q_j = p_j$  for all  $j = 1, \dots, n$ . Hence, it follows from (2.4.7) that

$$\begin{aligned} H(S_n) &\leq G(Q_n) = \sum_{j=1}^n p_j \log_2 (C 2^{\ell_j}) \\ &= \sum_{j=1}^n p_j (\ell_j + \log_2 C) \\ &= \text{avlength}(\mathbb{C}_n) + \log_2 C, \end{aligned}$$

since  $\sum_{j=1}^n p_j = 1$ . In view of Kraft's inequality, we have  $\log_2 C \leq 0$ , completing the derivation of (2.4.4). Furthermore, equality in (2.4.7) holds if and only if  $\log_2 C = 0$ , or  $C = 1$  and

$$H(S_n) = G(Q_n),$$

which is equivalent to  $q_1 = p_1, \dots, q_n = p_n$ , or

$$p_1 = \frac{1}{2^{\ell_1}}, \dots, p_n = \frac{1}{2^{\ell_n}}.$$

This completes the proof of (2.4.5)–(2.4.6), with  $k_j = \ell_j$ , which is an integer, being the length of the code-word  $c_j, j = 1, \dots, n$ . ■

**Remark 2.4.4** Since discrete probability distributions are seldom positive integer powers of  $\frac{1}{2}$ , one cannot expect to achieve a code-table with minimum average code-word lengths in general. On the other hand, there are fairly complicated coding schemes, such as “arithmetic coding”, that could reach the entropy lower bound as close as desired.

In introducing the notion of entropy, Claude Shannon also showed that the entropy can be used as a measuring stick in that there exist instantaneous code-tables with average code-word lengths bounded above by the entropy plus 1. In other words, we have the following result, called “noiseless coding” by Shannon.

**Theorem 2.4.2** *For any discrete probability distribution  $S_n = \{p_1, \dots, p_n\}$ ,*

$$H(S_n) \leq \min_{\mathbb{C}_n} \text{avlength}(\mathbb{C}_n) < H(S_n) + 1,$$

*where the minimum is taken over all instantaneous code-tables  $\mathbb{C}_n$ .*

**Proof** Let  $S_n = \{p_1, \dots, p_n\}$  be an arbitrarily given discrete probability distribution. In view of Theorem 2.4.1, it is sufficient to construct a certain instantaneous code  $\mathbb{C}_n = \{c_1, \dots, c_n\}$ , with length  $(c_j) = \ell_j$ , as defined in (2.4.1), for  $j = 1, \dots, n$ , that satisfies:

$$\sum_{j=1}^n p_j \ell_j < H(S_n) + 1.$$

Without loss of generality, we may, and do, assume that

$$1 > p_1 \geq p_2 \geq \dots \geq p_n > 0.$$

Indeed, while the zero probability value can be omitted in our discussion, the value  $p_1 = 1$  implies that  $p_2 = \dots = p_n = 0$ , or  $n = 1$ , so that the required code-table is simply  $\mathbb{C}_n = \mathbb{C}_1 = \{0\}$ , which is trivial. Now consider the set of positive integers  $\ell_1, \dots, \ell_n$ , with  $\ell_j$  defined by the smallest integer which

is greater than or equal to  $-\log_2 p_j$ , for each  $j = 1, \dots, n$ . Hence, under the assumption on the given discrete probability distribution  $S_n$ , we have

$$1 \leq \ell_1 \leq \ell_2 \leq \dots \leq \ell_n.$$

It is sufficient to prove that there exists an instantaneous code  $\mathbb{C}_n = \{c_1, \dots, c_n\}$ , with  $\text{length}(c_j) = \ell_j$ , for  $j = 1, \dots, n$ .

Indeed, on one hand, we have

$$\sum_{j=1}^n 2^{-\ell_j} \leq \sum_{j=1}^n p_j = 1,$$

which assures that  $\{\ell_1, \dots, \ell_n\}$  satisfies (2.4.2). On the other hand, we also have

$$\sum_{j=1}^n p_j \ell_j < \sum_{j=1}^n p_j (-\log_2 p_j) + 1 = \sum_{j=1}^n p_j (-\log_2 p_j) + \sum_{j=1}^n p_j = H(S_n) + 1,$$

which satisfies (2.4.4) as desired. To construct the desired instantaneous code-table (or prefix code)  $\mathbb{C}_n$ , we first consider the set of integers,  $\{w_1, \dots, w_n\}$ , defined by

$$w_1 = 0, w_2 = 2^{\ell_2 - \ell_1}, \dots, w_n = 2^{\ell_n - \ell_1} + 2^{\ell_n - \ell_2} + \dots + 2^{\ell_n - \ell_{n-1}}. \quad (2.4.8)$$

Also, let  $a_j = (a_j)_2$  denote the binary representation of  $w_j$  for each  $j = 1, \dots, n$ ; that is,  $a_j$  is a string of 0's and/or 1's, with  $w_j = (a_j)_2$ . For  $j \geq 1$ , the algorithm for computing  $a_j$  from  $w_j = (w_j)_{10}$  is given in Remark 2.3.3; and as mentioned in Remark 2.3.4, the subscript 2 of the binary representation  $(a_j)_2$  is omitted for convenience. Let us first observe that the length of  $a_j$  is less than or equal to  $\ell_j$  for all  $j = 1, \dots, n$ . Indeed, while the length of  $a_1$  is 1 (since  $a_1 = 0$ ), we have, for  $2 \leq j \leq n$ ,

$$\log_2(w_j) = \ell_j + \log_2\left(\sum_{k=1}^{j-1} 2^{-\ell_k}\right) \leq \ell_j + \log_2\left(\sum_{k=1}^{j-1} p_k\right) < \ell_j.$$

Hence, we can introduce the code

$$c_1 = 0 \dots 0, c_2 = (a_2)_2 0 \dots 0, \dots, c_n = (a_n)_2 0 \dots 0,$$

where the number of 0's attached to the right of each  $a_j$  is to ensure that the length of the code-word  $c_j$ , as defined in (2.4.1), is precisely  $\ell_j$ , namely:

$$\text{length}(c_j) = \ell_j,$$

for each  $j = 1, \dots, n$ . What is left in the proof of the Noiseless Coding Theorem, or Theorem 2.4.2, is to verify that  $\mathbb{C}_n = \{c_1, \dots, c_n\}$ , as defined in (2.4.4), is a prefix code. To do so, we first recall from Remark 2.3.2 that for

each  $j \geq 2$ ,  $w_j$  is a positive integer, and hence, the first bit of its binary representation  $a_j$  is always equal to 1. This implies that  $c_1 = 0 \cdots 0$  is not a prefix of all  $c_k$  for  $k \geq 2$ . To show that  $c_j$  is not a prefix of  $c_k$  for all  $k \neq j$  and  $j, k \geq 2$ , let us assume, on the contrary, that it is. Then the length  $\ell_j$  of  $c_j$  is less than the length  $\ell_k$ , and the first  $\ell_j$  bits of  $c_k$  is precisely  $(c_j)_2$ . Hence, in terms of the integers  $w_k$  and  $w_j$ , we may conclude that  $w_j$  is equal to the largest integer, not exceeding  $\frac{w_k}{2^{\ell_k - \ell_j}}$ , so that

$$w_j \leq \frac{w_k}{2^{\ell_k - \ell_j}} \quad (2.4.9)$$

On the other hand, from its definition (2.4.8), we have

$$\begin{aligned} \frac{w_k}{2^{\ell_k - \ell_j}} &= \sum_{i=1}^{k-1} 2^{(\ell_k - \ell_i) - (\ell_k - \ell_j)} \\ &= \sum_{i=1}^{k-1} 2^{\ell_j - \ell_i} = \sum_{i=1}^{j-1} 2^{\ell_j - \ell_i} + \sum_{i=j}^{k-1} 2^{\ell_j - \ell_i} \\ &= w_j + \sum_{i=j}^{k-1} 2^{\ell_j - \ell_i} = w_j + 1 + \sum_{i=j+1}^{k-1} 2^{\ell_j - \ell_i} > w_j + 1. \end{aligned} \quad (2.4.10)$$

Hence, combining the results in (2.4.9) and (2.4.10), we arrive at the absurd conclusion,  $w_j + 1 < w_j$ . This contradiction completes the proof of the theorem. ■

## 2.5 Image and Video Compression Schemes and Standards

In this current era of “information revolution”, data compression is a necessity rather than convenience or luxury. Without the rapid advancement of the state-of-the-art compression technology, not only the internet highway is unbearably over-crowded, but data management would be a big challenge also. In addition, data transmission would be most competitive and data storage would be very costly.

On the other hand, for various reasons, including strict regulations (such as storage of medical images), industry standards (for the capability of data retrieval by non-proprietary hardware and software installed in PC's or hand-held devices), and the need for a user-friendly environment, the more complex and/or proprietary solutions are not used by the general public, though they usually have better performance. For instance, ASCII is the preferred text format for word processors and Huffman coding is the most popular binary coding scheme, even though arithmetic coding is more optimal.

In general, there are two strategies in data compression, namely: lossless

compression and lossy compression. Lossless compression is reversible, meaning that the restored data file is identical to the original data information. There are many applications that require lossless compression. Examples include compression of executable codes, word processing files, and to some extent, medical records, and medical images. On the other hand, since lossless compression does not satisfactorily reduce file size in general, lossy compression is most widely used.

Lossy compression is non-reversible, but allows insignificant loss of the original data source, in exchange for significantly smaller compressed file size. Typical applications are image, video, and audio compressions. For compression of digital images and videos, for instance, compressed imagery is often more visually pleasing than the imagery source, which inevitably consists of (additive) random noise due to perhaps insufficient lighting and non-existence of “perfect sensors” for image capture. Since random noise increases entropy, which in turn governs the lower bound of the compressed file size according to the Noiseless Coding Theorem (or Theorem 2.4.2), lossy compression via removing a fair amount of such additive noise is a preferred approach. The topic of noise removal will be studied in Unit 5, and in more detail in Subunit 5.5.2.

### 2.5.1 Image compression scheme

There are three popular methods for lossless data compression: (i) run-length encoding (RLE), (ii) delta encoding, and (iii) entropy coding. RLE simply involves encoding the number of the same source word that appears repeatedly in a row. For instance, to encode a one-bit line drawing along each scan-line, only very few dots of the line drawing are on the scan-line. Hence, if “1” is used for the background and “0” for the line drawing, there are long rows of repeating “1” before a “0” is encountered. Therefore, important applications of RLE include graphic images (cartoons) and animation movies. It is also used to compress Windows 3.x bitmap for the computer startup screen. In addition, RLE is incorporated with Huffman encoding for the JPEG compression standard to be discussed in some details later in this unit.

Delta encoding is a simple idea for encoding data in the form of differences. In Example 2.3.1, the notion of differential pulse code modulation (DPCM) is introduced to create long rows of the same source word for applying RLE. In general, delta encoding is used only as an additional coding step to further reduce the encoded file size. In video compression, “delta frames” can be used to reduce frame size and is therefore used in every video compression standard. We will also mention DPCM in JPEG compression in the next section.

Perhaps the most powerful stand-alone lossless compression scheme is LZW compression, named after the developers A. Lempel, J. Ziv, and T. Welch. Typical applications are compression of executable codes, source codes, tabulated numbers, and data files with extreme redundancy. In addition, LZW is



used in GIF image files and as an option in TIFF and PostScript. However, LZW is a proprietary encoding scheme owned by Unisys Corporation.

### 2.5.2 Quantization

Let us now focus on the topic of lossy compression, to be applied to the compression of digital images and video, the topic of discussion in this unit. The general overall lossy compression scheme consists of three steps:

- (i) Transformation of source data,
- (ii) Quantization,
- (iii) Entropy coding.

Both (i) and (iii) are reversible at least in theory, but (ii) is not. To recover the information source, the three steps are:

- (i) De-coding by applying the code-table,
- (ii) De-quantization,
- (iii) Inverse transformation.

We have briefly discussed, in Example 2.3.1 in Subunit 2.3.1, that an appropriate transformation could be introduced to significantly reduce the entropy of certain source data, without loss of any data information. This type of specific transformations is totally data-dependent and therefore not very practical. Fortunately, there are many transformations that can be applied for sorting data information without specific knowledge of the data content, though their sole purpose is not for reduction of the data entropy. When data information is properly sorted out, the less significant content can be suppressed and the most insignificant content can be eliminated, if desired, so as to reduce the entropy. As a result, shortened binary codes, as governed by the Noiseless Coding Theorem studied in Subunit 2.4.4, can be constructed.

The most common insignificant data content is (additive) random noise, with probability values densely distributed on the (open) unit interval  $(0, 1)$ . Hence, embedded with such noise, the source data has undesirably large entropy. Fortunately, partially due to the dense distribution, random noise lives in the high-frequency range. Therefore, all transformations that have the capability of extracting high-frequency contents could facilitate suppressing the noise content by means of “quantization”, to be discussed in the next paragraph. Such transformations include DCT and DFT studied in Chapter 4, DST (discrete sine transform), Hardamard transform, and DWT (discrete wavelet transform), which will be introduced and studied in some depth in Unit 6. Among all of these transformations, DCT, and particularly DCT-II,

is the most popular for compression of digital images, videos, and digitized music.

The key to the feasibility of significant file size reduction for lossy compression is the “quantization” process that maps a “fine” set of real numbers to a “coarse” set of integers. Of course such mappings are irreversible. For any real numbers  $x$ ,  $\text{sgn } x$  (called the sign of  $x$ ) is defined by

$$\text{sgn } x = \begin{cases} 1, & \text{for } x > 0, \\ 0, & \text{for } x = 0, \\ -1, & \text{for } x < 0. \end{cases}$$

Also, for any non-negative real number  $x$ ,  $\lfloor x \rfloor$  will denote the largest integer not exceeding  $x$ . Then the most basic quantization process is the mapping of  $x \in \mathbb{R}$  to an integer  $\tilde{x}$ , defined by

$$\tilde{x} = \text{round} \left( \frac{x}{Q} \right) = (\text{sgn } x) \left\lfloor \frac{|x|}{Q} \right\rfloor, \quad (2.5.1)$$

where  $Q$  is a positive integer, called the “quantizer” of the “round-off” function defined in (2.5.1). Hence,  $\tilde{x}$  is an integer and  $Q\tilde{x}$  is an approximation of  $x$ , in that

$$|x - Q\tilde{x}| < 1.$$

A better approximation of the given real number  $x$  could be achieved by  $Q\hat{x}$ , with  $\hat{x}$  defined by

$$\hat{x} = (\text{sgn } x) \left\lfloor \frac{|x \pm \lfloor \frac{Q}{2} \rfloor|}{Q} \right\rfloor, \quad (2.5.2)$$

where the “+” sign or “−” sign is determined by whichever choice yields the smaller  $|x - Q\hat{x}|$ . In any case, it is clear that the binary representation of  $\tilde{x}$  (or of  $\hat{x}$ ) requires fewer bits for larger integer values of the quantizer  $Q$ . In applications to audio and image compressions, since the human ear and human eye are less sensitive to higher frequencies, larger values of the quantizer  $Q$  can be applied to higher-frequency DCT terms to save more bits. Moreover, since additive random noise lives in the higher frequency range, such noise could be suppressed, often resulting in more pleasing audio and imagery quality.

In summary, lossy compression is achieved by binary encoding (such as Huffman coding) of the quantized values of the transformed data; and recovery of the source data is accomplished by applying the inverse transformation to the de-quantized values of the decoded data. Since the quantization process is irreversible, the compression scheme is lossy.

### 2.5.3 Huffman, DPCM, and run-length coding

For compression of digital images, the industry standard is called JPEG, which is the acronym for “Joint Photographic Experts Group”. The compression scheme has been described above, with the transformation being the

two-dimensional DCT-II (studied in Subunit 2.2.3) applied to  $8 \times 8$  tiles of the digital image. In other words, we apply (2.2.3)–(2.2.4) to  $8 \times 8$  data matrices, as follows.

**Theorem 2.5.1** *For each  $8 \times 8$  sub-block*

$$A = \begin{bmatrix} a_{0,0} & \cdots & a_{0,7} \\ \vdots & & \\ \cdots & & \\ a_{7,0} & \cdots & a_{7,7} \end{bmatrix} \quad (2.5.3)$$

*of a digital image, the DCT*

$$\hat{A} = \begin{bmatrix} \hat{a}_{0,0} & \cdots & \hat{a}_{0,7} \\ \vdots & & \\ \cdots & & \\ \hat{a}_{7,0} & \cdots & \hat{a}_{7,7} \end{bmatrix}$$

*of  $A$  is given by*

$$\hat{a}_{j,k} = \frac{d_j d_k}{4} \sum_{\ell=0}^7 \sum_{s=0}^7 \left( \cos \frac{j(2\ell-1)\pi}{16} \cos \frac{k(2s-1)\pi}{16} \right) a_{\ell,s}, \quad (2.5.4)$$

*for  $j, k = 0, \dots, 7$ ; and the IDCT of  $\hat{A}$  is given by*

$$a_{\ell,s} = \frac{1}{4} \sum_{j=0}^7 \sum_{k=0}^7 \left( d_j d_k \cos \frac{j(2\ell-1)\pi}{16} \cos \frac{k(2s-1)\pi}{16} \right) \hat{a}_{j,k}, \quad (2.5.5)$$

*for  $\ell, s = 0, \dots, 7$ . In (2.5.4) and (2.5.5),  $d_0 = \frac{1}{\sqrt{2}}$  and  $d_1 = \dots = d_7 = 1$ .*

**Remark 2.5.1** For 8-bit images, the entries  $a_{\ell,s}$  of the  $8 \times 8$  image block  $A$  in (2.5.3) are integers that range from 0 to 255 (that is,  $0 \leq a_{\ell,s} \leq 255$ ). Hence, it follows from (2.5.4) that the dc (direct current) term  $\hat{a}_{0,0}$  of the DCT of  $A$  is given by

$$\hat{a}_{0,0} = \frac{1}{8} \sum_{\ell=0}^7 \sum_{s=0}^7 a_{\ell,s},$$

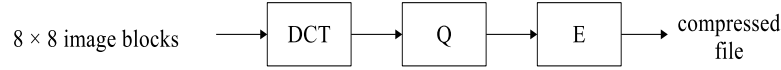
which is some integer between 0 and 2040 and therefore is an 11-bit integer. In addition, all the other DCT coefficients  $\hat{a}_{j,k}$  may oscillate in signs and are called ac (alternate current) terms.

**Remark 2.5.2** To reduce the size of the dc term  $\hat{a}_{0,0}$ , DPCM is used in JPEG by taking the difference with the dc term of the previous  $8 \times 8$  block. On the other hand, to achieve much higher compression ratio, the ac terms  $\hat{a}_{j,k}$  (for  $(j, k) \neq (0, 0)$ ) are to be “quantized”. This is the lossy (or non-reversible)

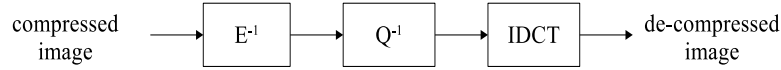
encoding component of the JPEG encoder. The quantization process is to divide each  $\hat{a}_{j,k}$  by some positive integer  $q_{j,k}$ , followed by rounding off the quotient to the nearest integer, to be denoted by  $(\hat{a}_{j,k}/q_{j,k})$ . For instance, for large values of  $q_{j,k}$ , the round-off integers  $(\hat{a}_{j,k}/q_{j,k})$  are equal to 0, yielding a long row of consecutive zeros (0's) for high-frequency ac terms (that is, relatively larger values of  $j + k$ ). The division by positive integers  $q_{j,k}$  also yield smaller numbers (and hence, less bits) to encode. When decoding the (encoded) compressed image file, the integers  $q_{j,k}$  are multiplied to the round-off integers  $(\hat{a}_{j,k}/q_{j,k})$  before the inverse DCT (denoted by IDCT) is applied.

#### 2.5.4 Encoder - Decoder (Codec)

The schematic diagrams of the JPEG encoder and decoder are shown in Fig. 2.1 and Fig. 2.2.



**FIGURE 2.1:** *Encoder:  $Q$  = quantization;  $E$  = entropy encoding*



**FIGURE 2.2:** *Decoder:  $Q^{-1}$  = de-quantization;  $E^{-1}$  = de-coding*

In Figs.2.3–2.4, we give two examples of the  $8 \times 8$  quantizers, with one for achieving low compression ratio and one for achieving high compression ratio.

1	1	1	1	1	2	2	4
1	1	1	1	1	2	2	4
1	1	1	1	2	2	2	4
1	1	1	1	1	2	4	8
1	1	2	2	2	2	4	8
2	2	2	2	2	4	8	8
2	2	2	4	4	8	8	16
4	4	4	4	8	8	16	16

**FIGURE 2.3:** *Quantizers: low compression ratio*

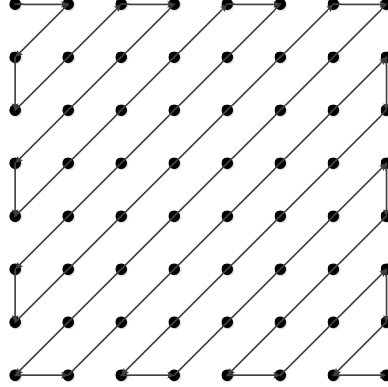
1	2	4	8	16	32	64	128
2	4	4	8	16	32	64	128
4	4	8	16	32	64	128	128
8	8	16	32	64	128	128	256
16	16	32	64	128	128	256	256
32	32	64	128	128	256	256	256
64	64	128	128	256	256	256	256
128	128	128	256	256	256	256	256

**FIGURE 2.4:** *Quantizers: high compression ratio*

The round-off integers, denoted by

$$b_{j,k} = (a_{j,k}/q_{j,k}),$$

for  $j, k = 0, \dots, 7$ , are arranged as a sequence of 64 integers by following the zig-zag ordering as shown in Fig. 2.5 and written out precisely as in (2.5.6).

**FIGURE 2.5:** *Zig-zag ordering*

$$\begin{aligned}
B = \{ & b_{0,0}, b_{0,1}, b_{1,0}, b_{2,0}, b_{1,1}, b_{0,2}, b_{0,3}, b_{1,2}, b_{2,1}, b_{3,0}, b_{4,0}, \\
& b_{3,1}, b_{2,2}, b_{1,3}, b_{0,4}, b_{0,5}, b_{1,4}, b_{2,3}, b_{3,2}, b_{4,1}, b_{5,0}, b_{6,0}, \\
& b_{5,1}, b_{4,2}, b_{3,3}, b_{2,4}, b_{1,5}, b_{0,6}, b_{0,7}, b_{1,6}, b_{2,5}, b_{3,4}, b_{4,3}, \\
& b_{5,2}, b_{6,1}, b_{7,0}, b_{7,1}, b_{6,2}, b_{5,3}, b_{4,4}, b_{3,5}, b_{2,6}, b_{1,7}, b_{2,7}, \\
& b_{3,6}, b_{4,5}, b_{5,4}, b_{6,3}, b_{7,2}, b_{7,3}, b_{6,4}, b_{5,5}, b_{4,6}, b_{3,7}, b_{4,7}, \\
& b_{5,6}, b_{6,5}, b_{7,4}, b_{7,5}, b_{6,6}, b_{5,7}, b_{6,7}, b_{7,6}, b_{7,7} \}.
\end{aligned} \tag{2.5.6}$$

Then  $B$  is considered as an information source to be encoded. The following modification is specified when the Huffman table, provided by the JPEG standard, is applied.

- (i) The dc term  $b_{0,0}$  in (2.5.6) is subtracted from the dc term of the previous  $8 \times 8$  (DCT-quantized) block, and the result  $\tilde{b}_{0,0}$  replaces  $b_{0,0}$  of (2.5.6) as the modified information source. (Recall that this differencing step is called DPCM.)
- (ii) Two code-words from the Huffman table are reserved for encoding rows of 0's in the sequence  $B$  in (2.5.6):
  - (a) EOB (end-of-block) with code-word 00000000 (of 8 zeros) means that the remaining source words in  $B$  are zeros and there is no need to encode them. For example, in (2.5.6), if  $b_{7,1} \neq 0$  but all the 27 source words in the sequence  $B$  after  $b_{7,1}$  are equal to 0, then after encoding  $b_{7,1}$  by applying the Huffman table, tack eight 0's to indicate  $b_{7,1}$  is the last word to be encoded in this  $8 \times 8$  block;
  - (b) ZRL (zero run length) with code-word 11110000 (of four 1's followed by four 0's) means that there are words of consecutive zeros in  $B$  before the last non-zero source word. We will not go into details of this process.

In the above discussion, we only considered compression of 8-bit gray-scale images, for convenience. For 24-bit  $RGB$  color images, with 8-bit  $R$  (red), 8-bit  $G$  (green), and 8-bit  $B$  (blue), it is recommended to apply the color transform from  $RGB$  to  $YC_bC_r$ , where  $Y$  stands for luminance (that is, light intensity, which by itself can be used to display the gray-scale image). The other two color components  $C_b$  and  $C_r$ , called chrominance (more precisely, chrominance blue and chrominance red, respectively), convey the color information by carrying the difference in intensities from the intensity of the luminance.

During the composite (analog) TV era, the  $YIQ$  color coordinates were introduced by RCA in the 1950's for broadcast bandwidth saving by "chroma-subsampling" of the  $IQ$  color components. It was later adopted by the NTSC standard, with the luminance  $Y$  specified to be

$$Y = 0.299R + 0.587G + 0.114B, \quad (2.5.7)$$

and chrominance  $I$  and  $Q$  to be

$$\begin{aligned} I &= 0.736(R - Y) - 0.268(B - Y), \\ Q &= 0.478(R - Y) + 0.413(B - Y). \end{aligned} \quad (2.5.8)$$

To overcome certain shortcomings, particularly in up-sampling, Germany introduced the PAL standard in the 1960's with  $YUV$  color transform given by

$$Y = 0.3R + 0.6G + 0.1B, \quad (2.5.9)$$

and the chrominance components  $U$  and  $V$  given simply by

$$U = B - Y, \quad V = R - Y. \quad (2.5.10)$$

Meanwhile, France also introduced another standard called SECAM. The importance of the luminance - chrominance formats (as opposed to the  $RGB$  color coordinates) is that human vision is much more sensitive to light intensity (or brightness) than color differences, particularly for scenes in motion. Observe that both  $I, Q$  in (2.5.8) and particularly  $U, V$  in (2.5.10) are defined by taking color differences. Consequently, chroma-subsampling by down-sampling the chrominance components is hardly noticeable in digital TV broadcasting. Currently, the so-called 4:1:1 and 4:2:0 formats allow an additional 33% increase in compression ratio. Furthermore, if noise removal is applied appropriately to the chrominance components, the  $Y$  component maintains the sharpness of the video imagery.

To adopt the  $YIQ$  and  $YUV$  color coordinates for color image compression, the  $YC_bC_r$  format was developed by the JPEG image compression standard, by specifying

$$\begin{aligned} C_b &= \frac{1}{2} U + 0.5; \\ C_r &= \frac{1}{1.6} V + 0.5, \end{aligned} \quad (2.5.11)$$

for the chrominance components, where the values of the colors  $R, G, B$  are expressed by a relative scale from 0 to 1, with 0 indicating no phosphor excitation and 1 indicating maximum phosphor excitation. Hence, the additive factor of 0.5, with the decrease in the color differences  $U = B - Y$  and  $V = R - Y$ , facilitates a better preservation of the blue and red colors, even after down-sampling of  $C_b$  and  $C_r$  to achieve higher compression ratio.

### 2.5.5 I, P, and B video frames

### 2.5.6 Macro-blocks

### 2.5.7 Motion search and compensation

We end this unit by giving a very brief introduction to (digital) video compression.

While the image compression standard JPEG was developed by the “Joint Photographic Experts Group” under the auspices of the three major international standard organizations ISO, CCITT, and IEC, the video compression standard MPEG was developed by the “Moving Pictures Expert Group” of ISO. The first standard completed in 1991 is known as MPEG-1 for the first-generation digital video compression. The second-generation digital video standard, known as MPEG-2, is currently adopted for HDTV broadcasting. In addition, MPEG-4 was developed in the late 1990’s for low bit-rate video

compression. More recently, the new video standard H.264, also called MPEG-4 Part 10 and AVC (for Advanced Video Coding), was successfully developed in 2003 for up to an additional 50% compression saving over MPEG-2 and MPEG-4 Part 1, while maintaining comparable video quality. H.264 is the video compression standard for Blu-ray discs, and is widely used for internet video streaming by such giants as YouTube (of Google) and iTunes stores (of Apple). In addition, it is embedded in the web software Adobe Flash Player, and is the preferred video format for most cable and satellite television service providers, including Direct TV.

In any case, all effective video compression schemes are similar, with *I*-pictures (or *I*-frames), *P*-pictures (or *P*-frames, and *B*-pictures (or *B*-frames). With the exception of H.264, all MPEG and other H.26x standards adopt the JPEG image compression standard for *I*-picture (or intra-frame) compression. To meet the mandate of  $2\times$  compression efficiency over MPEG-2, *I*-frame image compression for H.264 departs from the JPEG standard by introducing “*I*-slices” that include  $4 \times 4$  DCT blocks, with applications to frame-by-frame video such as “iFrame video”, developed by Apple in 2009 to facilitate video editing and high-quality video camcorder recording, particularly in iMovie’09 and iMovie’11. It is also adopted by camera manufacturers to capture HD video in  $1920 \times 1080$  resolution, such as the AVC-Intra video codec (that is, encoding and decoding), developed by Panasonic in 2007 for HD video broadcasting.

On the other hand, to facilitate video frame prediction (for *P*-frames and *B*-frames), the *I*-frame format for video encoding is slightly different from JPEG in that the *I*-frames are partitioned into macroblocks of sizes  $8 \times 8$  or  $16 \times 16$ . In other words, DPCM encoding of the dc coefficients is limited to at most four  $8 \times 8$  blocks. To encode a *P*-frame (or prediction frame), adjacent macroblocks of the current *P*-frame (called intra-macroblocks) are compared with macroblocks of previous *I* or *P* frames (called inter-macroblocks) by “motion search”. If an inter-macroblock (from some previous frame) is suitable to replace an adjacent intra-macroblock, then compression of this adjacent macroblock is eliminated simply by coding the “motion vector” that tells the decoder which inter-macroblock is used to replace the adjacent macroblock of the current frame. Bi-directional frame prediction is an extension of the *P*-frame prediction to allow searching of inter-macroblock replacement (for replacing adjacent macroblocks of the current frame) from both previous video frames and future video frames. Such prediction frames are called *B*-frames (or bi-directional prediction frames).

## References

- (1) John Loomis’s “JPEG Tutorial (HTML).
- (2) National Program on Technology Enhanced Learning (NPTEL), “Lecture 16: Introduction to Image and Video Compression, (YouTube).



- (3) National Program on Technology Enhanced Learning (NPTEL), “Lecture 23: Video Compression Basic Building Blocks, (YouTube).
- (4) National Program on Technology Enhanced Learning (NPTEL), “Lecture 24: Motion Estimation Techniques, (YouTube).
- (5) National Program on Technology Enhanced Learning (NPTEL), “Lecture 26: Video Coding Standards (YouTube).



# Unit 3

## FOURIER METHODS

The subject of Fourier series is one of the most important topics in Applied Mathematics. For example, the matrix transformation DFT studied in Unit 2 is only a discrete version of the Fourier coefficients of the Fourier series. The theory of Fourier series is very rich and well documented in the mathematics literature, with numerous existing textbooks and research monographs. The objective of this unit is to study the most basic topics of this subject and to prepare for its applications to solving partial differential equations (PDEs) in Unit 5. In Subunit 3.2, it is shown that the partial sums of a Fourier series of some function  $f$  are the orthogonal projections of  $f$  to the corresponding subspaces of trigonometric polynomials. In addition, these partial sums can be formulated as convolution of the function with the Dirichlet kernels, to be introduced in Subunit 3.3. Since averaging of the Dirichlet kernels yields the Fejér kernels (introduced in Subunit 3.3) that constitute a positive approximate identity (to be shown in Subunit 3.4), it follows that convergence of the sequence of trigonometric polynomials, resulting from convolution of the function  $f$  with the Fejér kernels, to the function  $f$  itself is assured in the mean-square sense. Consequently, being orthogonal projections, the partial sums of the Fourier series also converge to the function represented by the Fourier series, again in the mean-square sense. The topic of point-wise and uniform convergence, under a differentiability assumption on the function  $f$  is studied in Subunit 3.4, where the concept of completeness is also discussed. An interesting observation is that  $L_2$ -completeness is equivalent to Parseval's identity for Fourier series, with such interesting applications as solving the famous Basel problem. Based on the Bernoulli polynomials, we will also derive Euler's formula (in terms of the Bernoulli numbers) as solution of the general Basel problem (for all even powers) in Subunit 3.5. The completeness property of Fourier series will also be applied to solving boundary value problems of PDE in Unit 5.

### 3.1 Fourier Series

The continuous version of the notions of DFT and DCT (for finite sequences,

such as digital signals, studied in Subunits 2.1 and 2.2, respectively) is the concept of “Fourier-coefficients”, defined for piecewise continuous functions on bounded intervals  $J$ , such as analog signals, and more generally, for functions in  $L_2(J)$ . To facilitate the theoretical development and computational simplicity, we will first study the complex-valued setting, before deriving the real-valued formulation, in terms of the cosine and/or sine basis functions. The analogy of the inverse DFT (IDFT) and inverse DCT (IDCT) for finite sequences is the “Fourier series” representation (or expansion) of the given piecewise continuous functions (or more generally, functions in  $L_2(J)$  for infinite sequences), in that the Fourier series can be applied to recover the given function from the infinite sequence of its Fourier coefficients.

For this study, a piecewise continuous function  $f(x)$  defined on some bounded interval  $J = [a, b]$  is extended to a periodic function on  $\mathbb{R} = (-\infty, \infty)$ , with period  $= (b - a)$ , by setting

$$f(x + \ell(b - a)) = f(x), \quad x \in [a, b],$$

for all  $\ell = \pm 1, \pm 2, \dots$ , after replacing the values  $f(a)$  and  $f(b)$  by their average value  $(f(a) + f(b))/2$ ; and the sequence of its Fourier coefficients  $c_k = c_k(f)$  is defined by

$$c_k = c_k(f) = \frac{1}{b - a} \int_a^b f(x) e^{-i2\pi k(x-a)/(b-a)} dx, \quad k \in \mathbb{Z}.$$

Observe that the sequence  $\{c_k\} = \{c_k(f)\}$  is an infinite (and more precisely, a bi-infinite) sequence. Furthermore, in view of Euler’s formula,

$$e^{-i2\pi k(x-a)/(b-a)} = \cos \frac{2\pi k(x-a)}{b-a} - i \sin \frac{2\pi k(x-a)}{b-a},$$

the sequence  $\{c_k\}$  reveals the frequency content of the given analog signal  $f \in PC[a, b]$ , in a similar manner as the DFT and DCT reveal the frequency content of a digital signal.

As mentioned above, the Fourier series of the given function  $f \in PC[a, b]$ , defined by

$$(Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f) e^{i2\pi k(x-a)/(b-a)},$$

is used to recover  $f(x)$  from the sequence of its Fourier coefficients. In this regard, the study of Fourier series is much more sophisticated than that of IDFT and IDCT, since it requires the study of the convergence of an infinite series  $Sf$ ; or equivalently, the convergence of the sequence  $\{S_n f\}$  of its “partial sums”, defined by

$$(S_n f)(x) = \sum_{k=-n}^n c_k(f) e^{i2\pi k(x-a)/(b-a)}$$

for  $n = 0, 1, \dots$

The topic of convergence of Fourier series will be discussed in some depth in Subunit 3.4.1. To develop the mathematical tools for the study of pointwise and uniform convergence, the notions of Dirichlet's and Fejér's kernels will be introduced in Subunits 3.3.1 and 3.3.2, respectively. In particular, Fejér's kernels, defined by averaging Dirichlet's kernels, provide the important property of "Positive Approximate Identity" to be investigated in Subunit 3.3.3. We remark that the property of positive approximate identity is the key ingredient in the proof of uniform convergence, to be discussed in Subunit 3.4.1.

### 3.1.1 Fourier series representations

For convenience, we only consider the interval  $J = [a, b] = [-\pi, \pi]$ . There is certainly no loss of generality for such restriction, in view of the simple change of variables:

$$x \longleftrightarrow \frac{2\pi}{b-a} (x - a) - \pi;$$

namely, the functions  $\tilde{f} \in PC[a, b]$  and  $f \in PC[-\pi, \pi]$  are interchangeable by considering

$$f(x) = \tilde{f}\left(a + \frac{b-a}{2\pi} (x + \pi)\right)$$

and

$$\tilde{f}(x) = f\left(\frac{2\pi}{b-a}(x - a) - \pi\right).$$

As mentioned above, each function  $f \in PC[-\pi, \pi]$  can be extended to a  $2\pi$ -periodic function, as follows:

**Definition 3.1.1** *Let  $f(x)$  be a function defined on an interval  $[-\pi, \pi]$ . The  $2\pi$ -periodic extension  $F(x)$  of  $f(x)$  is a function defined on  $\mathbb{R}$ , by*

$$\begin{cases} F(x) = f(x), & x \in (-\pi, \pi), \\ F(-\pi) = F(\pi) = \frac{1}{2}(f(\pi) + f(-\pi)), \end{cases}$$

and  $F(x + 2\ell\pi) = F(x)$  for all  $\ell \in \mathbb{Z}$ . For convenience,  $F(x)$  is re-named by the given function  $f(x)$ , to avoid unnecessarily additional notation. In addition, the notation  $PC_{2\pi}^* = PC^*[-\pi, \pi]$  is used to denote the inner-product space of such  $2\pi$ -periodic piecewise continuous functions, with inner product  $\langle\langle \cdot, \cdot \rangle\rangle$ , defined by

$$\langle\langle f, g \rangle\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{g(x)} \, dx. \quad (3.1.1)$$

Observe that the only difference between the inner product  $\langle\langle \cdot, \cdot \rangle\rangle$  and the

inner product  $\langle \cdot, \cdot \rangle$ , as introduced in Subunit 1.1, is the additional normalization constant of  $(2\pi)^{-1}$  in (3.1.1), namely:

$$\langle\langle f, g \rangle\rangle = \frac{1}{2\pi} \langle f, g \rangle.$$

Let us first consider the infinite sequence  $\{c_k\} = \{c_k(f)\}$  (of Fourier coefficients) for a given function  $f(x)$  in the vector space  $\mathbb{V} = PC_{2\pi}^*$ , defined by

$$c_k = c_k(f) = \langle\langle f, e_k \rangle\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \overline{e_k(x)} \, dx,$$

where  $e_k(x) = e^{ikx} = \cos kx + i \sin kx$ , or

$$c_k = c_k(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} \, dx; \quad (3.1.2)$$

and the corresponding infinite (Fourier) series:

$$(Sf)(x) = \sum_{k=-\infty}^{\infty} c_k(f) e_k(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}. \quad (3.1.3)$$

**Definition 3.1.2** Let  $f \in PC_{2\pi}^*$ . The infinite series in (3.1.3), with  $c_k = c_k(f)$ , defined by (3.1.2), is called the Fourier series representation (or expansion) of  $f(x)$ , and  $c_k = c_k(f)$  is called the  $k^{\text{th}}$  Fourier coefficient of  $f(x)$ .

**Example 3.1.1** Let  $f_1(x)$  be defined by

$$f_1(x) = \begin{cases} 1, & \text{for } 0 \leq x \leq \pi, \\ -1, & \text{for } -\pi \leq x < 0, \end{cases}$$

and extended  $2\pi$ -periodically to all  $x \in \mathbb{R}$  (see Definition 3.1.1). Compute the Fourier coefficients and Fourier series expansion of  $f_1(x)$ .

**Solution** By (3.1.2), we have

$$\begin{aligned} c_k &= c_k(f_1) = \frac{1}{2\pi} \left\{ \int_{-\pi}^0 (-1) e^{-ikx} \, dx + \int_0^{\pi} e^{-ikx} \, dx \right\} \\ &= \frac{1}{2\pi} \int_0^{\pi} (-e^{ikx} + e^{-ikx}) \, dx \\ &= \frac{-i}{\pi} \int_0^{\pi} \sin kx \, dx. \end{aligned}$$

Hence,  $c_0 = 0$  and for  $k \neq 0$ ,

$$c_k = \left[ \frac{i \cos kx}{\pi k} \right]_0^{\pi} = \frac{i}{\pi k} ((-1)^k - 1).$$

In other words, the Fourier coefficients of  $f_1(x)$  are given by

$$c_{2\ell} = 0, \quad c_{2\ell+1} = \frac{-2i}{\pi(2\ell+1)}, \quad \text{for all } \ell = 0, \pm 1, \pm 2, \dots \quad (3.1.4)$$

(where we consider  $k = 2\ell$  and  $k = 2\ell + 1$  separately), and the Fourier series expansion of  $f_1(x)$  is given by

$$\begin{aligned} (Sf_1)(x) &= \frac{2}{i\pi} \sum_{\ell=-\infty}^{\infty} \frac{e^{i(2\ell+1)x}}{2\ell+1} \\ &= \frac{2}{i\pi} \left( \sum_{\ell=0}^{\infty} \frac{e^{i(2\ell+1)x}}{2\ell+1} + \sum_{\ell=-\infty}^{-1} \frac{e^{i(2\ell+1)x}}{2\ell+1} \right) \\ &= \frac{2}{i\pi} \left\{ (e^{ix} - e^{-ix}) + \frac{e^{i3x} - e^{-i3x}}{3} + \frac{e^{i5x} - e^{-i5x}}{5} + \dots \right\} \\ &= \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1}. \end{aligned}$$

■

**Example 3.1.2** Let  $g_1(x)$  be defined on  $[-\pi, \pi]$  by

$$g_1(x) = \begin{cases} 1, & \text{for } |x| \leq \frac{\pi}{2}, \\ 0, & \text{for } \frac{\pi}{2} < |x| \leq \pi, \end{cases}$$

and extended periodically to  $\mathbb{R}$ . Compute the Fourier coefficients and Fourier series representation of  $g_1(x)$ .

**Solution** By (3.1.2), we have

$$c_k = c_k(g_1) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g_1(x) e^{-ikx} dx = \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} e^{-ikx} dx,$$

so that  $c_0 = \frac{1}{2}$  and for  $k \neq 0$ ,

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} (\cos kx - i \sin kx) dx \\ &= \frac{1}{2\pi} \int_{-\pi/2}^{\pi/2} \cos kx dx = \frac{1}{\pi} \int_0^{\pi/2} \cos kx dx \\ &= \left[ \frac{1}{\pi} \frac{\sin kx}{k} \right]_0^{\pi/2} = \frac{1}{\pi} \frac{\sin k\pi/2}{k}. \end{aligned}$$

Again, since  $c_k = 0$  for all even  $k \neq 0$ , we only consider odd  $k = 2\ell + 1$ , for which

$$\sin \frac{(2\ell+1)\pi}{2} = \sin \left( \ell\pi + \frac{\pi}{2} \right) = (-1)^\ell.$$

Therefore, we have  $c_0 = \frac{1}{2}$ ,  $c_{2\ell} = 0$  for all  $\ell = \pm 1, \pm 2, \dots$ , and

$$c_{2\ell+1} = \frac{(-1)^\ell}{(2\ell+1)\pi}, \quad \ell = 0, \pm 1, \pm 2, \dots;$$

and the Fourier series representation of  $g_1(x)$  is given by

$$\begin{aligned} (Sg_1)(x) &= \frac{1}{2} + \sum_{\ell=-\infty}^{\infty} \frac{(-1)^\ell}{(2\ell+1)\pi} e^{i(2\ell+1)x} \\ &= \frac{1}{2} + \frac{1}{\pi} \left\{ (e^{ix} + e^{-ix}) - \frac{e^{i3x} + e^{-i3x}}{3} + \dots \right\} \\ &= \frac{1}{2} + \frac{2}{\pi} \sum_{k=0}^{\infty} (-1)^k \frac{\cos(2k+1)x}{2k+1}. \end{aligned}$$

■

For various important reasons, particularly for more effective implementation in applications, it is necessary to convert the Fourier series expansion of functions  $f \in PC_{2\pi}^*$  to a cosine/sine series (that is, in terms of cosines and sines) by getting rid of the imaginary unit,  $i$ , in Euler's formula, as follows.

**Theorem 3.1.1** *The Fourier series expansion  $Sf$  of  $f \in PC_{2\pi}^*$  in (3.1.3) can be re-formulated as*

$$(Sf)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx), \quad (3.1.5)$$

called the (Fourier) trigonometric (or cosine and sine) series expansion, where

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

and

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

**Proof** Observe that

$$\frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, dx = c_0,$$

where  $c_0$  is defined in (3.1.2) for  $k = 0$ . For  $k = 1, 2, \dots$ ,

$$\begin{aligned} c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} \, dx = \frac{1}{2} (a_k - ib_k); \\ c_{-k} &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{ikx} \, dx = \frac{1}{2} (a_k + ib_k), \end{aligned}$$



so that

$$\begin{aligned}
 (Sf)(x) &= c_0 + \sum_{k=1}^{\infty} c_k e^{ikx} + \sum_{k=-\infty}^{-1} c_k e^{ikx} \\
 &= c_0 + \sum_{k=1}^{\infty} c_k e^{ikx} + \sum_{k=1}^{\infty} c_{-k} e^{-ikx} \\
 &= \frac{a_0}{2} + \sum_{k=1}^{\infty} \frac{1}{2} (a_k - ib_k) e^{ikx} + \frac{1}{2} (a_k + ib_k) e^{-ikx} \\
 &= \frac{a_0}{2} + \sum_{k=1}^{\infty} \frac{1}{2} a_k (e^{ikx} + e^{-ikx}) + \frac{1}{2} (-ib_k) (e^{ikx} - e^{-ikx}) \\
 &= \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx.
 \end{aligned}$$

This proves (3.1.5). ■

In general, the interval  $[-\pi, \pi]$  can be replaced by  $[-d, d]$  for any  $d > 0$ , as follows.

**Theorem 3.1.2** *Let  $d > 0$ . Then the Fourier series expansion of  $Sf$  of any function  $f \in PC[-d, d]$ , extended periodically to  $\mathbb{R}$ , is given by*

$$(Sf)(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left( a_k \cos \frac{k\pi x}{d} + b_k \sin \frac{k\pi x}{d} \right), \quad (3.1.6)$$

*called the (Fourier) trigonometric (or cosine and sine) series expansion, where*

$$a_k = \frac{1}{d} \int_{-d}^d f(x) \cos \frac{k\pi x}{d} dx, \quad k = 0, 1, 2, \dots,$$

and

$$b_k = \frac{1}{d} \int_{-d}^d f(x) \sin \frac{k\pi x}{d} dx, \quad \text{for } k = 1, 2, \dots \quad (3.1.7)$$

**Example 3.1.3** Let  $f(x) = 1$  for  $-d \leq x \leq 0$ , and  $f(x) = 0$  for  $0 < x \leq d$ . Compute the Fourier cosine and sine series representation  $Sf$  of  $f$  in (3.1.6).

**Solution** For  $k = 0$ ,

$$a_0 = \frac{1}{d} \int_{-d}^d f(x) dx = \frac{1}{d} \int_{-d}^0 1 dx = \frac{d}{d} = 1,$$

and for  $k \geq 1$ ,

$$\begin{aligned}
 a_k &= \frac{1}{d} \int_{-d}^d f(x) \cos \frac{k\pi x}{d} dx = \frac{1}{d} \int_{-d}^0 1 \cdot \cos \frac{k\pi x}{d} dx \\
 &= \left[ \frac{1}{d} \frac{\sin \frac{k\pi x}{d}}{\frac{k\pi}{d}} \right]_{-d}^0 = \frac{1}{k\pi} (\sin 0 - \sin(-k\pi)) = 0; \\
 b_k &= \frac{1}{d} \int_{-d}^d f(x) \sin \frac{k\pi x}{d} dx = \frac{1}{d} \int_{-d}^0 1 \cdot \sin \frac{k\pi x}{d} dx \\
 &= \left[ -\frac{1}{d} \frac{\cos \frac{k\pi x}{d}}{\frac{k\pi}{d}} \right]_{-d}^0 = -\frac{1}{k\pi} (\cos 0 - \cos(-k\pi)) \\
 &= -\frac{1}{k\pi} (1 - (-1)^k).
 \end{aligned}$$

Thus, the Fourier cosine and sine series expansion of  $f$  is given by

$$\begin{aligned}
 (Sf)(x) &= \frac{1}{2} - \sum_{k=1}^{\infty} \frac{1}{k\pi} (1 - (-1)^k) \sin \frac{k\pi x}{d} \\
 &= \frac{1}{2} - \sum_{n=0}^{\infty} \frac{2}{(2n+1)\pi} \sin \frac{(2n+1)\pi x}{d}.
 \end{aligned}$$

■

For the Fourier series expansion  $(Sf)(x)$  in terms of both cosines and sines in (3.1.5) of Theorem 3.1.1 or in (3.1.6) in Theorem 3.1.2, observe that the computational complexity can be reduced by eliminating either the sine series component, or the cosine series component. To this end, recall that any function  $f(x)$ , defined on an interval  $[a, b]$ , can be treated as a function defined on the interval  $[0, d]$ , by the change of variables  $t = \frac{d}{b-a}(x-a)$ . In other words, we may consider the Fourier trigonometric series of  $\tilde{f}(t)$  (with  $\tilde{f}(t) = f(x)$ ) on  $[0, d]$  (instead of  $f(x)$  on  $[a, b]$ ), while recovering both  $f(x)$  and its Fourier cosine and sine series by the change of variables  $x = \frac{b-a}{d}t + a$ .

Without loss of generality, let us consider functions  $f(x)$  in  $L_2[0, d]$ ,  $d > 0$ . For such functions  $f$ , we extend  $f$  to an even function  $f_e$  on  $[-d, d]$  by setting

$$f_e(x) = \begin{cases} f(x), & \text{for } 0 \leq x \leq d, \\ f(-x), & \text{for } -d \leq x < 0. \end{cases}$$

Then, since  $\sin(\frac{k\pi x}{d})$  is an odd function on  $[-d, d]$ , we have  $b_k(f_e) = 0$  in (3.1.7), so that

$$(Sf_e)(x) = \frac{a_0(f_e)}{2} + \sum_{k=1}^{\infty} a_k(f_e) \cos \frac{k\pi x}{d}, \quad (3.1.8)$$

where

$$a_k(f_e) = \frac{1}{d} \int_{-d}^d f_e(x) \cos \frac{k\pi x}{d} dx = \frac{2}{d} \int_0^d f(x) \cos \frac{k\pi x}{d} dx$$

for  $k = 0, 1, 2, \dots$ . But since the given function  $f(x)$  is the same as  $f_e(x)$  for  $x \in [0, d]$ , we have  $Sf = Sf_e$ , and therefore have eliminated the sine series component of the Fourier series representation  $Sf$  of  $f$  on  $[0, d]$ . For the Fourier cosine series representation, we will use the notation  $S^c f$  for  $Sf$ , for clarity.

**Example 3.1.4** Find the Fourier cosine series representation  $Sf$  (to be relabeled by  $(S^c f)$  to emphasize a Fourier series without the sine terms) of  $f(x) = 1 + 2 \cos^2 x, 0 \leq x \leq \pi$ .

**Solution** By the trigonometric identity  $\cos^2 x = \frac{1}{2}(1 + \cos 2x)$ , we may write  $f(x)$  as

$$f(x) = 1 + 2 \cdot \frac{1}{2}(1 + \cos 2x) = 2 + \cos 2x.$$

Hence, by direct application of the definition, we can also compute

$$\begin{aligned} a_0 &= \frac{2}{\pi} \int_0^\pi (2 + \cos 2x) \, dx = 4; \\ a_2 &= \frac{2}{\pi} \int_0^\pi (2 + \cos 2x) \cos 2x \, dx \\ &= \frac{2}{\pi} \int_0^\pi \cos^2 2x \, dx = \frac{2}{\pi} \frac{\pi}{2} = 1, \end{aligned}$$

and for  $k = 3, 4, \dots$ ,  $a_k = 0$ , since  $\cos kx$  is orthogonal to 1 and  $\cos 2x$ . Thus, the cosine series expansion  $S^c f = Sf$  of  $f(x) = 1 + 2 \cos^2 x$  is

$$(S^c f)(x) = \frac{a_0}{2} + a_2 \cos 2x = 2 + \cos 2x.$$

Observe that there is really no need to compute  $a_k$ , since the cosine polynomial  $2 + \cos 2x$  is already a Fourier cosine series in terms of  $\cos kx, k = 0, 1, \dots$  ■

In general, for integer powers of the cosine and sine functions, we may apply Euler's formula to avoid the task of integration.

**Example 3.1.5** Find the Fourier cosine series expansion  $(S^c f)(x) = (Sf)(x)$  of  $f(x) = \cos^5 x, 0 \leq x \leq \pi$ .

**Solution** By Euler's formula, we have

$$\begin{aligned} \cos^5 x &= \left( \frac{e^{ix} + e^{-ix}}{2} \right)^5 \\ &= \frac{1}{32} (e^{i5x} + 5e^{i3x} + 10e^{ix} + 10e^{-ix} + 5e^{-i3x} + e^{-i5x}) \\ &= \frac{1}{32} (e^{i5x} + e^{-i5x} + 5(e^{i3x} + e^{-i3x}) + 10(e^{ix} + e^{-ix})) \\ &= \frac{1}{16} (\cos 5x + 5 \cos 3x + 10 \cos x). \end{aligned}$$

Since the above trigonometric polynomial is already a “series” in terms of the cosine basis functions  $\cos kx$ ,  $k = 0, 1, \dots$ , we know that its Fourier cosine series expansion is itself. Thus, the cosine series expansion  $S^c f$  of  $f(x) = \cos^5 x$  is given by

$$(S^c f)(x) = \frac{5}{8} \cos x + \frac{5}{16} \cos 3x + \frac{1}{16} \cos 5x,$$

where the terms are arranged in increasing order of “frequencies”. ■

Similarly, a function  $f \in L_2[0, d]$  can be expanded as a (Fourier) sine series. The trick is to consider the odd extension  $f_o$  of  $f$  to  $[-d, d]$ , namely:

$$f_o(x) = \begin{cases} f(x), & \text{for } 0 < x \leq d, \\ -f(-x), & \text{for } -d \leq x < 0, \\ 0, & \text{for } x = 0. \end{cases}$$

Then, since  $\cos \frac{k\pi x}{d}$  is an even function on  $[-d, d]$ , for each  $k = 0, 1, \dots$ , we have  $a_k(f_o) = 0$  and

$$b_k(f_o) = \frac{1}{d} \int_{-d}^d f_o(x) \sin \frac{k\pi x}{d} dx = \frac{2}{d} \int_0^d f(x) \sin \frac{k\pi x}{d} dx$$

for  $k = 1, 2, \dots$ . In other words, we have eliminated the cosine series component of the Fourier series representation  $Sf$  of  $f$  on  $[0, d]$ . For clarity, we will use the notation  $S^s f$  for  $Sf$  for the Fourier sine series representation.

**Example 3.1.6** Compute the Fourier sine series representation  $(S^s f)(x)$  of  $f(x) = x$ ,  $0 \leq x \leq 1$ .

**Solution** For  $d = 1$ , we have

$$\begin{aligned} b_k &= 2 \int_0^1 f(x) \sin k\pi x dx = 2 \int_0^1 x \sin k\pi x dx \\ &= 2 \int_0^1 x \left( \frac{-\cos k\pi x}{k\pi} \right)' dx = \frac{-2}{k\pi} \int_0^1 x (\cos k\pi x)' dx \\ &= \frac{-2}{k\pi} \left\{ \left[ x \cos k\pi x \right]_0^1 - \int_0^1 \cos k\pi x dx \right\} \\ &= \frac{-2}{k\pi} \left\{ \cos k\pi - 0 - \left[ \frac{\sin k\pi x}{k\pi} \right]_0^1 \right\} \\ &= \frac{-2}{k\pi} \left\{ (-1)^k - \frac{\sin k\pi}{k\pi} + \frac{\sin 0}{k\pi} \right\} = \frac{-2}{k\pi} \{ (-1)^k - 0 + 0 \} \\ &= \frac{2}{k\pi} (-1)^{k+1} = \frac{2}{\pi} \frac{(-1)^{k+1}}{k}. \end{aligned}$$

Thus, the Fourier sine series representation of  $f(x) = x$  is given by

$$\frac{2}{\pi} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \sin k\pi x. \quad \blacksquare$$

### 3.1.2 Orthogonality and computation

#### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 28: Fourier Series (Part 1) (YouTube)", presented by Gilbert Strang.
- (2) MIT: Department of Computational Science and Engineering's "Lecture 29: Fourier Series (Part 2) (YouTube)", presented by Gilbert Strang.

## 3.2 Orthogonal Projection

In Subunit 1.1, the extension of the dot product to the inner product, and that of the Euclidean space  $\mathbb{R}^2$  (or  $x - y$  plane) to the general inner-product (vector) spaces, such as the  $\ell_2$  sequence space and  $L_2$  function space, allow us to generalize the well-known geometric properties of the Pythagorean theorem and the Parallelogram law, studied in pre-college Plane Geometry, to an arbitrary inner-product space. The derivations of these two properties are presented in Subunits 3.2.1 and 3.2.2, respectively, simply by applying the Cauchy-Schwarz inequality, studied in Subunit 1.1.2. In addition, the method of orthogonal projection, discussed in Subunit 1.1.4, is applied to the function space  $L_2$  to show that for every function  $f \in PC_{2\pi}^*$ , the  $n$ -th partial sum  $S_n f$  of the Fourier series of  $f$  provides the best mean-square approximation of  $f$  from the subspace  $\mathbb{V}_{2n+1}$ , with basis functions  $e^{ikx} = \cos(kx) + i\sin(kx)$ , where  $-n \leq k \leq n$ . This will be proved, with illustrative examples, in Subunit 3.2.3.

### 3.2.1 Pythagorean theorem

Recall, from Subunit 1.1, the definition of the inner product and the notion (1.1.2) of the corresponding norm measurement

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle},$$

for any vector  $\mathbf{x}$  in an inner-product space  $\mathbb{V}$ . Also recall that two vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$  are said to be orthogonal to each other, if  $\langle \mathbf{x}, \mathbf{y} \rangle = 0$ . Hence, for any two vectors  $\mathbf{x}$  and  $\mathbf{y}$  that are orthogonal to each other, we have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2. \end{aligned} \tag{3.2.1}$$

Observe that the three vectors  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z} = \mathbf{x} + \mathbf{y}$  constitute a right triangle, with  $\mathbf{x}$  and  $\mathbf{y}$  as the two legs, and  $\mathbf{z}$  as the hypotenuse. Hence, the identity (3.2.1) says that the square of the length (i.e. norm) of the hypotenuse is equal to the sum of squares on the two legs of a right triangle. For the special case where the vector space  $\mathbb{V}$  is the Euclidean space  $\mathbb{R}^2$ , this is called the Pythagorean Theorem. We have therefore extended the statement of the Pythagorean Theorem from the Euclidean space  $\mathbb{R}^2$  to any inner-product space  $\mathbb{V}$ , which includes both sequence and function spaces, that could be infinite dimensional spaces. For convenience, the identity (3.2.1) is also called the Pythagorean Theorem.

To apply the Pythagorean Theorem to our study of Fourier series, let us again use the notation

$$e_k(x) = e^{ikx} = \cos kx + i \sin kx, \quad (3.2.2)$$

and consider the subspace

$$\mathbb{V}_{2n+1} = \text{span}\{e_k(x) : -n \leq k \leq n\} \quad (3.2.3)$$

of the inner-product space  $PC_{2\pi}^*$ , with inner product  $\langle\langle \cdot, \cdot \rangle\rangle$  introduced in (3.1.1). Since

$$\langle\langle e_j, e_k \rangle\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} e_j(x) \overline{e_k(x)} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i(j-k)x} dx = \delta_{j-k}$$

for all  $j, k = -n, \dots, n$ ,  $\mathbb{V}_{2n+1}$  is a  $(2n+1)$ -dimensional vector space with orthonormal basis  $\{e_{-n}(x), \dots, e_n(x)\}$ .

Now for any function  $f \in PC_{2\pi}^*$ , recall from Subunit 1.1.4 that the orthogonal projection  $g_{2n+1} = P_{2n+1}f$  of  $f \in PC_{2\pi}^*$  onto the subspace  $\mathbb{V}_{2n+1}$  is the function  $g_{2n+1} \in \mathbb{V}_{2n+1}$ , determined by the orthogonality of  $f - g_{2n+1}$  to the entire space  $\mathbb{V}_{2n+1}$ ; or equivalently, to its basis  $\{e_{-n}(x), \dots, e_n(x)\}$  of  $\mathbb{V}_{2n+1}$ . This, in turn, is equivalent to the condition:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} [f(x) - g_{2n+1}(x)] e^{-ikx} dx = 0;$$

or equivalently,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} g_{2n+1}(x) e^{-ikx} dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx = c_k(f), \quad (3.2.4)$$

for all  $k = -n, \dots, n$ . To draw the conclusion of the above derivation, we need the following notation of the partial sums of a Fourier series.

**Definition 3.2.1** Let  $f \in PC_{2\pi}^*$ . Then for  $n = 0, 1, \dots$ , the  $n^{\text{th}}$  partial sum of its Fourier series  $Sf$  is defined by

$$(S_n f)(x) = \sum_{k=-n}^n c_k e^{ikx}. \quad (3.2.5)$$

Hence, in view of the fact that both functions,  $g_{2n+1}$  and  $S_n f$ , are in the same vector space  $\mathbb{V}_{2n+1}$  with basis  $\{e_{-n}(x), \dots, e_n(x)\}$ , it follows from (3.2.4) that  $g_{2n+1} = S_n f$ , as follows.

**Theorem 3.2.1** *Let  $f \in PC_{2\pi}^*$ . Then for each  $n = 0, 1, \dots$ , the orthogonal projection  $P_{2n+1}f$  of  $f$  from  $PC_{2\pi}^*$  to its subspace  $\mathbb{V}_{2n+1}$  is the  $n^{\text{th}}$  partial sum of the Fourier series  $Sf$  of  $f$ , namely:*

$$(P_{2n+1}f)(x) = (S_n f)(x). \quad (3.2.6)$$

### 3.2.2 Parallelogram law

The Pythagorean Theorem studied in Subunit 3.2.1 has a natural generalization from a right triangle to a parallelogram. Indeed, if a rectangle is partitioned into two right triangles by using one of the two diagonals, then applying the Pythagorean Theorem to each of the two right triangles, we may conclude that the sum of the squares on the two diagonals is the same as the sum of the squares on the four sides of the rectangle. The reason is that the two diagonals of a rectangle have the same length. The generalization from a rectangle to an arbitrary parallelogram is to allow the two diagonals to have different lengths. In the following, we will show, by using the inner product, that the above statement (that the sum of the squares on the two diagonals is equal to the sum of the squares on the four sides) holds for all parallelograms.

As in Subunit 3.2.1, let  $\mathbb{V}$  be any inner-product space with inner product  $\langle \mathbf{x}, \mathbf{y} \rangle$  and norm  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ , for any vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ . Then assuming that  $\mathbf{x}$  and  $\mathbf{y}$  are not parallel, we can formulate a parallelogram, with two opposite sides given by  $\mathbf{x}$  and the other two opposite sides given by  $\mathbf{y}$ . The two diagonal vectors are then given by  $\mathbf{x} + \mathbf{y}$  and  $\mathbf{x} - \mathbf{y}$ . Therefore, the sum of the squares on the two diagonals is precisely  $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2$ , which can be expanded and simplified as follows:

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle + \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \\ &= (\langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle) \\ &\quad + (\langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle + \langle -\mathbf{y}, -\mathbf{y} \rangle) \quad (3.2.7) \\ &= 2\langle \mathbf{x}, \mathbf{x} \rangle + 2\langle \mathbf{y}, \mathbf{y} \rangle = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2, \end{aligned}$$

where the right-hand side is the sum of the squares on the four sides of the parallelogram. Therefore, the identity in (3.2.7) is called the Parallelogram Law. Of course, the identity (3.2.7) is an extension from the Euclidean space  $\mathbb{R}^2$  to an arbitrary inner-product space  $\mathbb{V}$ , which includes both sequence and function spaces, that may be infinite dimensional.

In the following example, we use the inner product  $\langle \cdot, \cdot \rangle$  introduced in Subunit 1.1, instead of the normalized inner product  $\langle\langle \cdot, \cdot \rangle\rangle$  defined in (3.1.1).

**Example 3.2.1** Consider the inner-product space  $L_2[0, 1]$ , with inner product defined in Subunit 1.1. Verify the validity of the Parallelogram Law by using the function  $f(x) = x + 1$  as two opposite sides, and the function  $g(x) = x - 1$  as the other two opposite sides, of a parallelogram.

**Solution** The areas of the squares on two adjacent sides of the parallelogram are given by

$$\|f\|^2 = \int_0^1 (f(x))^2 dx = 1 + 1 + \frac{1}{3};$$

and

$$\|g\|^2 = \int_0^1 (g(x))^2 dx = 1 - 1 + \frac{1}{3},$$

respectively. Hence, the sum of the squares on the four sides of the parallelogram is given by  $2[(1 + 1 + \frac{1}{3}) + (1 - 1 + \frac{1}{3})] = 4 + \frac{4}{3}$ . On the other hand, since the two diagonals are represented by  $f(x) + g(x) = 2x$  and  $f(x) - g(x) = 2$ , the sum of the squares on the two diagonals is given by

$$\|f + g\|^2 + \|f - g\|^2 = \int_0^1 (2x)^2 dx + \int_0^1 (2)^2 dx = 4 \times \frac{1}{3} + 4 = 4 + \frac{4}{3};$$

which agrees with the sum of the squares on the four sides of the parallelogram. ■

### 3.2.3 Best mean-square approximation

To prepare for our study of convergence of Fourier series in Subunit 3.4.1, we recall the notion of the  $n^{\text{th}}$  partial sums  $(S_n f)(x)$ , introduced in (3.2.13), of the Fourier series  $(Sf)(x)$  in (3.1.3), and apply the Pythagorean Theorem in Subunit 3.2.1 to Theorem 3.2.1 to derive the following result on best mean-square approximation.

**Theorem 3.2.2** Let  $f \in PC_{2\pi}^*$ . Then for each  $n = 0, 1, \dots$ , the best  $L^2$ -approximation of  $f$  from the subspace  $\mathbb{V}_{2n+1}$  is achieved by the  $n^{\text{th}}$  partial sum  $S_n f \in \mathbb{V}_{2n+1}$ , namely:

$$\|f - S_n f\|_2 \leq \|f - g\|_2, \text{ for all } g \in \mathbb{V}_{2n+1}. \quad (3.2.8)$$

**Proof;** Let  $g \in \mathbb{V}_{2n+1}$  be arbitrarily chosen, and observe that the function  $h(x)$ , defined by

$$h = S_n f - g$$

is also in the space  $\mathbb{V}_{2n+1}$ . Hence, according to (3.2.6) in Theorem 3.2.1, the function  $f - S_n f$  is orthogonal to the function  $h = (S_n f - g)$ . This allows us to apply the Pythagorean Theorem to the right triangle, with the two legs



$(f - S_n f)$  and  $h$ , and hypotenuse  $(f - S_n f) + h = (f - S_n f) + (S_n f - g) = f - g$ , to conclude that

$$\|f - S_n f\|_2^2 + \|h\|_2^2 = \|f - g\|_2^2,$$

which implies (3.2.8) and completes the proof of Theorem 3.2.2. ■

**Example 3.2.2** As an application of Theorem 3.2.2 to Example 3.1.1 in Subunit 3.1, let  $f_1(x)$  be defined by

$$f_1(x) = \begin{cases} 1, & \text{for } 0 \leq x \leq \pi, \\ -1, & \text{for } -\pi \leq x < 0. \end{cases}$$

Determine the values of  $a$  and  $b$ , for which the norm measurement

$$\|f_1(x) - (a \cos x + b \sin x)\|_2$$

is minimum.

**Solution** According to the solution of Example 3.1.1, the Fourier series expansion  $Sf_1$  of  $f_1$  is given by

$$(Sf_1)(x) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{\sin(2k+1)x}{2k+1}.$$

Hence, the first-order partial sum  $S_1 f_1$  of the Fourier series  $Sf_1$  is

$$(S_1 f_1)(x) = \frac{4}{\pi} \sin x,$$

by considering only the term with  $k = 0$  in the above series expansion. Therefore it follows from Theorem 3.2.2 that the smallest value of  $\|f_1(x) - (a \cos x + b \sin x)\|_2$  is achieved by choosing  $a = 0$  and  $b = \frac{4}{\pi}$ . ■

**Example 3.2.3** As another application of Theorem 3.2.2 to Example 3.1.2 in Subunit 3.1, let  $g_1(x)$  be defined on  $[-\pi, \pi]$  by

$$g_1(x) = \begin{cases} 1, & \text{for } |x| \leq \frac{\pi}{2}, \\ 0, & \text{for } \frac{\pi}{2} < |x| \leq \pi. \end{cases}$$

Determine the values of  $a$ ,  $b$ , and  $c$ , for which the norm measurement

$$\|g_1(x) - (a + b \cos x + c \sin x)\|_2$$

is minimum.

**Solution** According to the solution of Example 3.1.2, the Fourier series expansion  $Sg_1$  of  $g_1$  is given by

$$(Sg_1)(x) = \frac{1}{2} + \frac{2}{\pi} \sum_{k=0}^{\infty} (-1)^k \frac{\cos(2k+1)x}{2k+1}.$$

Hence, the partial sum  $S_1 g_1$  of  $S g_1$  is  $(S_1 g_1)(x) = \frac{1}{2}$  by considering only the term with  $k = 0$  in the above series expansion. Therefore it follows from Theorem 3.2.2 that the smallest value of  $\|g_1(x) - (a + b \cos x + c \sin x)\|_2$  is achieved by choosing  $a = \frac{1}{2}$ ,  $b = 0$ , and  $c = 0$ . ■

### 3.3 Dirichlet's and Fejér's Kernels

Two important families of (integral) convolution kernels are introduced and studied in this subunit. Firstly, the family of Dirichlet's kernels, derived in Subunit 3.3.1, is used as convolution kernels (with a given function  $f$  in  $PC_{2\pi}^*$ ) to yield the  $n$ -th partial sums  $S_n f$  of the Fourier series of  $f$ , for all  $n = 0, 1, \dots$ . In Subunit 3.3.2, the family of Fejér's kernels is introduced as averages of Dirichlet's kernels. An elegant expression of Fejér's kernels is also derived in this subunit, showing that Fejér's kernels are non-negative. Furthermore, it is shown in Subunit 3.3.3 that the family of Fejér's kernels constitutes a positive approximate identity. This important property is also applied in Subunit 3.3.3 to show that all continuous functions on the closed interval  $[-\pi, \pi]$  can be uniformly approximated by trigonometric polynomials on the interval.

#### 3.3.1 Partial sums as convolution with Dirichlet's kernels

For any  $f \in PC_{2\pi}^*$ , the  $n^{\text{th}}$  partial sum  $(S_n f)(x)$  of the Fourier series expansion (3.1.3) can be formulated as

$$\begin{aligned} (S_n f)(x) &= \sum_{k=-n}^n c_k e^{ikx} = \sum_{k=-n}^n \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-ikt} dt \right) e^{ikx} \\ &= \sum_{k=-n}^n \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{ik(x-t)} dt \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) D_n(x-t) dt, \end{aligned} \quad (3.3.1)$$

where

$$D_n(x) = \sum_{k=-n}^n e^{ikx} \quad (3.3.2)$$

is called Dirichlet's kernel of order  $n$ .

In other words, the  $n^{\text{th}}$  partial sum  $S_n f$  of the Fourier series expansion of any function  $f \in PC_{2\pi}^*$  can be obtained by applying the (integral) convolution

operation with the kernel  $D_n$  in (3.3.2), where the convolution operation is defined as follows.

**Definition 3.3.1** For the inner-product space  $PC_{2\pi}^*$  with inner product defined in (3.1.1), the operation

$$(f * g)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)g(x-t) dt \quad (3.3.3)$$

is called the (integral) convolution of  $f, g \in PC_{2\pi}^*$ .

**Remark 3.3.1** We remark that the convolution operation defined by (3.3.3) is commutative, since

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} f(t)g(x-t) dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t)g(t) dt,$$

which can be easily verified by applying the  $2\pi$ -periodicity property of  $f(x)$  and  $g(x)$  in the change of the variable from  $(x-t)$  to  $t$  in the integration. ■

**Theorem 3.3.1** For  $n = 1, 2, 3, \dots$ , the Dirichlet kernel  $D_n(x)$ , as defined in (3.3.2), is given by

$$D_n(x) = \frac{\sin(n + \frac{1}{2})x}{\sin \frac{x}{2}}. \quad (3.3.4)$$

To derive (3.3.4), we may apply the well-known summation formula for finite geometric series, as follows:

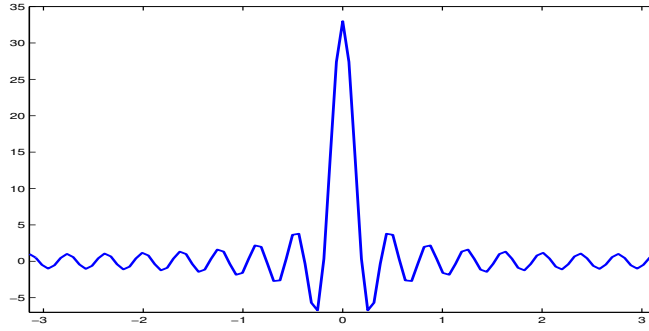
$$\begin{aligned} D_n(x) &= e^{-inx} \sum_{k=0}^{2n} e^{ikx} = e^{-inx} \frac{1 - e^{i(2n+1)x}}{1 - e^{ix}} \\ &= \frac{e^{-inx} - e^{i(n+1)x}}{e^{i\frac{x}{2}}(e^{-i\frac{x}{2}} - e^{i\frac{x}{2}})} = \frac{e^{-i(n+\frac{1}{2})x} - e^{i(n+\frac{1}{2})x}}{e^{-i\frac{x}{2}} - e^{i\frac{x}{2}}} \\ &= \frac{(-2i) \sin(n + \frac{1}{2})x}{(-2i) \sin \frac{x}{2}} = \frac{\sin(n + \frac{1}{2})x}{\sin \frac{x}{2}}. \quad \blacksquare \end{aligned}$$

**Remark 3.3.2** From the definition in (3.3.2), it is clear that  $D_n(0) = 2n+1$ . Hence, the formula in (3.3.4) of  $D_n(x)$  for  $x \neq 0$  also applies to  $x = 0$  by applying L'Hospital's rule, namely:

$$\begin{aligned} \lim_{x \rightarrow 0} \frac{\sin(n + \frac{1}{2})x}{\sin \frac{x}{2}} &= \lim_{x \rightarrow 0} \frac{(n + \frac{1}{2}) \cos(n + \frac{1}{2})x}{\frac{1}{2} \cos \frac{x}{2}} \\ &= (2n + 1) \frac{\cos 0}{\cos 0} = 2n + 1. \end{aligned}$$



The graph of  $y = D_n(x)$ , for  $n = 16$ , is displayed in Fig. 3.1. Observe the sign changes of this graph at  $x = \frac{2j\pi}{33}$  for  $j = 1, \dots, 32$  and  $j = -32, \dots, -1$ .



**FIGURE 3.1:** *Dirichlet's kernel  $D_{16}(x)$*

### 3.3.2 Césaro means and derivation of Fejér's kernels

In this subunit, we introduce the notion of Césaro means (that is, arithmetic averages) of Dirichlet's kernels  $D_j(x)$  to formulate the following notion of Fejér's kernels  $\sigma_n$  of order  $n = 0, 1, 2, \dots$ , namely:

$$\sigma_n(x) = \frac{1}{n+1} \sum_{j=0}^n D_j(x). \quad (3.3.5)$$

To derive a useful formula for  $\sigma_n(x)$ , we set  $z = e^{ix}$  and observe that

$$D_j(x) = \frac{1 - z^{2j+1}}{z^j(1 - z)} = \frac{1}{z - 1} \left( z^{j+1} - \frac{1}{z^j} \right),$$

so that

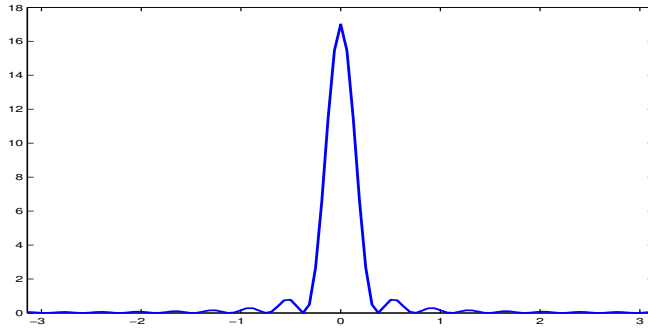
$$\begin{aligned}
 (n+1)\sigma_n(x) &= D_0(x) + D_1(x) + \cdots + D_n(x) \\
 &= \frac{1}{z-1}(z-1) + \frac{1}{z-1}\left(z^2 - \frac{1}{z}\right) + \cdots + \frac{1}{z-1}\left(z^{n+1} - \frac{1}{z^n}\right) \\
 &= \frac{1}{z-1}\left(z + z^2 + \cdots + z^{n+1} - 1 - \frac{1}{z} - \cdots - \frac{1}{z^n}\right) \\
 &= \frac{1}{z-1}\left(z - \frac{1}{z^n}\right)(1 + z + z^2 + \cdots + z^n) \\
 &= \frac{1}{(z-1)^2}\left(z - \frac{1}{z^n}\right)(z^{n+1} - 1) \\
 &= \frac{(z^{n+1} - 1)^2}{z^n(z-1)^2} = \frac{(z^{\frac{n+1}{2}} - z^{-\frac{n+1}{2}})^2}{(z^{\frac{1}{2}} - z^{-\frac{1}{2}})^2} \\
 &= \left(\frac{\sin \frac{(n+1)x}{2}}{\sin \frac{x}{2}}\right)^2.
 \end{aligned}$$

Hence, Fejér's kernels can be formulated as follows.

**Theorem 3.3.2** *For  $n = 1, 2, \dots$ , the  $n^{\text{th}}$  order Fejér kernel  $\sigma_n(x)$ , as defined in (3.3.5), can be written as*

$$\sigma_n(x) = \frac{1}{n+1} \left( \frac{\sin \frac{(n+1)x}{2}}{\sin \frac{x}{2}} \right)^2. \quad (3.3.6)$$

The graph of  $y = \sigma_n(x)$ , for  $n = 16$ , is displayed in Fig. 3.2. Observe that  $\sigma_n(x) \geq 0$ , for all  $x \in \mathbb{R}$ . This positivity property of Fejér's kernels is the key ingredient for them to constitute a positive approximate identity. This property will be applied to prove that the set of trigonometric polynomials is dense in the space of periodic continuous functions. This study is delayed to the next subunit.



**FIGURE 3.2:** Fejér's kernel  $\sigma_{16}(x)$

Next observe that since the  $n^{\text{th}}$  partial sum  $S_n f$  of the Fourier series expansion of any function  $f \in PC_{2\pi}^*$  is precisely the (integral) convolution of  $f$  with Dirichlet's kernel  $D_n(x)$ , we should be motivated to study the  $n^{\text{th}}$  Césaro means of the partial sums  $\{(S_j f)(x) : j = 0, 1, \dots\}$  and expect the result to be the (integral) convolution of  $f$  with Fejér's kernel  $\sigma_n(x)$ , as in (3.3.8) to be formulated below. But first let us write out the Césaro means  $(C_n f)(x)$ , as follows.

**Definition 3.3.2** *The  $n^{\text{th}}$ -order Césaro means  $(C_n f)(x)$  of any function  $f \in PC_{2\pi}^*$  is defined by*

$$(C_n f)(x) = \frac{(S_0 f)(x) + \dots + (S_n f)(x)}{n+1}. \quad (3.3.7)$$

Recall that  $\mathbb{V}_{2n+1}$ , as defined by (3.2.3), is a subspace of  $PC_{2\pi}^*$ . Hence, since  $(C_n f)(x)$  is a linear combination of  $\{(S_j f)(x) : 0 \leq j \leq n\}$  which is a subset of  $\mathbb{V}_{2n+1}$ , the Césaro means  $(C_n f)(x)$  is a function in  $\mathbb{V}_{2n+1}$ .

**Remark 3.3.3** Let  $\sigma_n(x)$  be Fejér's kernel of order  $n$ . It follows from (3.3.1) and (3.3.5) that the Césaro means of the partial sums of the Fourier series representation of any function  $f \in PC_{2\pi}^*$  is the following  $n^{\text{th}}$ -degree trigonometric polynomial

$$(C_n f)(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) \sigma_n(x-t) dt = (f * \sigma_n)(x). \quad (3.3.8)$$

■

### 3.3.3 Positive approximate identity

In this subunit, we introduce the concept of “positive approximate identity” and show that the sequence  $\{\sigma_n(x)\}$  of Fejér's kernels has this important property.

**Theorem 3.3.3** *The sequence  $\{\sigma_n(x)\}$  of Fejér's kernels constitutes a “positive approximate identity”, meaning that it has the following properties:*

- (a)  $\sigma_n(x) \geq 0$ , for all  $x$ ;
- (b)  $\frac{1}{2\pi} \int_{-\pi}^{\pi} \sigma_n(x) dx = 1$ ;
- (c) for any positive number  $\delta$  with  $0 < \delta < \pi$ ,

$$\sup_{\delta \leq |x| \leq \pi} \sigma_n(x) \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (3.3.9)$$

The first property is evident from (3.3.6), and the second property follows from the definitions of  $D_k(x)$  and  $\sigma_n(x)$ , while the third property follows from the fact that  $|\sin y| \leq 1$  for all  $y \in \mathbb{R}$  and  $|\sin \frac{x}{2}| \geq |\sin \frac{\delta}{2}| > 0$  for  $\delta \leq |x| \leq \pi$ . ■

The importance of the property of positive approximate identity of the sequence  $\{\sigma_n(x)\}$  is the following theorem on uniform approximation of continuous periodic functions by trigonometric polynomials

**Theorem 3.3.4** *Let  $f$  be a continuous function on the closed interval  $[-\pi, \pi]$  that satisfies  $f(-\pi) = f(\pi)$ . Then the sequence of trigonometric polynomials  $(C_n f)(x) = (f * \sigma_n)(x) \in \mathbb{V}_{2n+1}$ , as defined in (3.3.8), converges uniformly to  $f(x)$  for all  $x \in \mathbb{R}$ ; that is,*

$$\|f - C_n f\|_\infty \rightarrow 0, \text{ as } n \rightarrow \infty,$$

where

$$\|f - C_n f\|_\infty = \max\{|f(x) - (C_n f)(x)| : x \in \mathbb{R}\} \quad (3.3.10)$$

is called the uniform error of approximation of  $f$  by  $C_n f \in \mathbb{V}_{2n+1}$ .

**Proof** Let  $\epsilon > 0$  be arbitrarily given, and set  $M = \|f\|_\infty$ . Then since  $f$  is continuous on the closed and bounded interval  $[-\pi, \pi]$ , it is uniformly continuous on  $[-\pi, \pi]$ , and therefore uniformly continuous on the real line  $\mathbb{R}$ , by the assumption  $f(-\pi) = f(\pi)$ . Let  $\delta > 0$  be so chosen, that

$$|f(x) - f(t)| < \frac{\epsilon}{2} \quad (3.3.11)$$

for all  $x, t$  with  $|x - t| < \delta$ . On the other hand, by the third property of Fejér's kernels  $\sigma_n(x)$  in Theorem 3.3.3, there exists a positive integer  $N$ , such that

$$\sup_{\delta \leq |x-t| \leq \pi} \sigma_n(x-t) \leq \frac{\epsilon}{4M} \quad (3.3.12)$$

for all  $n \geq N$ . In addition, by consecutive application of the properties (b) and (a) of  $\sigma_n(x)$  in the same theorem, we have

$$\begin{aligned} |f(x) - (C_n f)(x)| &= \left| \frac{1}{2\pi} \int_{-\pi}^{\pi} (f(x) - f(t)) \sigma_n(x-t) dt \right| \\ &\leq \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x) - f(t)| \sigma_n(x-t) dt. \end{aligned}$$

Hence, when the integral over  $[-\pi, \pi]$  is partitioned into two integrals, with one over  $0 \leq |t - x| < \delta$ , and the other over  $\delta \leq |t - x| \leq \pi$ , it follows from

(3.3.11) and (3.3.12) that

$$\begin{aligned}
 |f(x) - (C_n f)(x)| &\leq \frac{1}{2\pi} \int_{|x-t| < \delta} |f(x) - f(t)| \sigma_n(x-t) dt \\
 &\quad + \frac{1}{2\pi} \int_{|x-t| \geq \delta} |f(x) - f(t)| \sigma_n(x-t) dt \\
 &< \frac{\epsilon}{2} \left( \frac{1}{2\pi} \int_{-\pi}^{\pi} \sigma_n(x-t) dt \right) + \frac{1}{2\pi} \int_{-\pi}^{\pi} 2M \cdot \frac{\epsilon}{4M} dt \\
 &= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
 \end{aligned}$$

That is,  $\|f - C_n f\|_{\infty} < \epsilon$  for all  $n \geq N$ , or equivalently,

$$\|f - C_n f\|_{\infty} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

This completes the proof of (3.3.10). ■

**Remark 3.3.4** Of course the interval  $[-\pi, \pi]$  can be replaced by any closed and bounded interval  $[a, b]$  and the above theorem assures that any continuous periodic function can be uniformly approximated as close as desired by trigonometric polynomials with sufficiently high degrees. This result will be applied in Subunit 3.4.1 to derive the result on the convergence of Fourier series in the  $L_2$ -norm.

### 3.4 Completeness

The main objective of this subunit is to establish both the pointwise and uniform convergence results of Fourier series, as well as the result on the density of trigonometric polynomials in the  $L_2$  space of square-integrable functions. Since there exist continuous  $2\pi$ -periodic functions whose Fourier series diverge everywhere, the objective of Subunit 3.4.1 is to show that under the condition of existence of one-sided derivatives, the Fourier series do converge. On the other hand, we will prove in Subunit 3.4.2 that trigonometric polynomials are dense in  $L_2(-\pi, \pi)$ . Hence, since the partial sums  $S_n f$  of the Fourier series of  $f$  are orthogonal projections of  $f$ , and thus provide best trigonometric polynomial approximation of  $f$  in  $L_2(-\pi, \pi)$ , it follows that Fourier series do converge in the  $L_2(-\pi, \pi)$ -norm without any restriction at all.



### 3.4.1 Pointwise and uniform convergence

Although the partial sums of the Fourier series expansion  $(Sf)(x)$  of any square-integrable function  $f(x)$  provide the best approximation of  $f(x)$  in the  $L_2$  norm, among all trigonometric polynomials of the same degree, as proved in Theorem 3.2.2 in Subunit 3.2.3, it is not clear whether the Fourier series  $(Sf)(x)$  would converge pointwise to  $f(x)$  (that is, at each fixed  $x \in (-\pi, \pi]$ ). The fact is that pointwise convergence is not assured, unless some smoothness condition is imposed on the given function  $f(x)$ . For example, there exists a continuous  $2\pi$ -periodic function whose Fourier series diverges at every  $x = r\pi$ , where  $r$  is any rational number. In this subunit, we will show that under the assumption that if both one-sided derivatives of  $f(x) \in L_2(-\pi, \pi]$  exist for all  $x \in (-\pi, \pi]$ , then the Fourier series  $Sf$  of  $f$  indeed converges pointwise. Furthermore, we will also show that if the given  $2\pi$ -periodic function  $f(x)$  is continuous on  $(-\pi, \pi]$  and is almost everywhere differentiable, with derivative  $f'(x) \in L_2(-\pi, \pi)$ , then its Fourier series  $(Sf)(x)$  converges absolutely and uniformly to  $f(x)$  for all  $x \in \mathbb{R}$ . To facilitate the proof of the result on pointwise convergence of Fourier series, we first derive the following inequality for any orthonormal family, called Bessel's inequality.

**Theorem 3.4.1** *Let  $\{\phi_k\}$ , where  $k = 1, 2, \dots, n$ , be an orthonormal family in an inner-product space  $\mathbb{V}$ . Then for any  $f \in \mathbb{V}$  and  $c_k = \langle f, \phi_k \rangle$ ,*

$$\sum_{k=1}^n |c_k|^2 = \sum_{k=1}^n |\langle f, \phi_k \rangle|^2 \leq \|f\|^2.$$

The proof of the above theorem follows from the identity (3.2.1) in Subunit 3.2.1 for the derivation of the Pythagorean Theorem, namely:

$$\begin{aligned} 0 \leq \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle + \langle -\mathbf{y}, -\mathbf{y} \rangle \\ &= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2, \end{aligned} \tag{3.4.1}$$

simply by setting  $\mathbf{x} = f(x)$  and

$$\mathbf{y} = \sum_{k=1}^n \langle f, \phi_k \rangle \phi_k(x),$$

while observing that for such  $\mathbf{x}$  and  $\mathbf{y}$ , we have

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle.$$

Indeed, while

$$\begin{aligned}
 \langle \mathbf{x}, \mathbf{y} \rangle &= \sum_{k=1}^n \langle f, \langle f, \phi_k \rangle \phi_k \rangle \\
 &= \sum_{k=1}^n \overline{\langle f, \phi_k \rangle} \langle f, \phi_k \rangle \\
 &= \sum_{k=1}^n \left| \langle f, \phi_k \rangle \right|^2 = \langle \mathbf{y}, \mathbf{x} \rangle
 \end{aligned}$$

it follows from the property:

$$\langle \phi_k, \phi_j \rangle = \delta_{k-j}$$

of the orthonormal family  $\{\phi_k\}$  that

$$\begin{aligned}
 \langle \mathbf{y}, \mathbf{y} \rangle &= \sum_{k=1}^n \sum_{j=1}^n \left\langle \langle f, \phi_k \rangle \phi_k, \langle f, \phi_j \rangle \phi_j \right\rangle \\
 &= \sum_{k=1}^n \sum_{j=1}^n \langle f, \phi_k \rangle \overline{\langle f, \phi_j \rangle} \langle \phi_k, \phi_j \rangle \\
 &= \sum_{k=1}^n \sum_{j=1}^n \langle f, \phi_k \rangle \overline{\langle f, \phi_j \rangle} \delta_{k-j} \\
 &= \sum_{k=1}^n \left| \langle f, \phi_k \rangle \right|^2
 \end{aligned}$$

as well. ■

**Example 3.4.1** Verify that the family of functions

$$\phi_k(x) = e^{i2\pi kx}, \quad -n \leq k \leq n,$$

constitute an orthonormal family of their algebraic span  $\mathbb{V}_{2n+1}$  in  $\mathbb{V} = L_2(0, 1]$ . Then compute  $\langle f, \phi_k \rangle$  and  $\|f\|^2$ , where  $f(x) = x$ , and determine the error:

$$E_n = \|f\|^2 - \sum_{k=-n}^n |\langle f, \phi_k \rangle|^2.$$

**Solution** For  $k \neq j$ ,

$$\begin{aligned}
 \langle \phi_k, \phi_j \rangle &= \int_0^1 e^{i2\pi kx} e^{-i2\pi jx} dx \\
 &= \int_0^1 e^{i2\pi(k-j)x} dx \\
 &= \frac{1}{i2\pi(k-j)} e^{i2\pi(k-j)x} \Big|_0^1 \\
 &= \frac{1}{i2\pi(k-j)} (1 - 1) = 0.
 \end{aligned}$$

For  $k = j$ ,

$$\langle \phi_k, \phi_j \rangle = \langle \phi_k, \phi_k \rangle = \int_0^1 e^{i2\pi(k-k)x} dx = \int_0^1 dx = 1.$$

Hence, the family of functions  $\phi_k$ ,  $k = -n, \dots, n$ , is orthonormal.

Next,

$$\|f\|^2 = \int_0^1 x^2 dx = \frac{1}{3};$$

and for  $k \neq 0$ ,

$$\begin{aligned}
 \langle f, \phi_k \rangle &= \int_0^1 x e^{-i2\pi kx} dx \\
 &= x \frac{e^{-i2\pi kx}}{-i2\pi k} \Big|_0^1 - \int_0^1 \frac{1}{-i2\pi k} e^{-i2\pi kx} dx \\
 &= \frac{i}{2\pi k} - \frac{i}{2\pi k} \cdot 0 = \frac{i}{2\pi k};
 \end{aligned}$$

while for  $k = 0$ ,

$$\langle f, \phi_k \rangle = \int_0^1 x dx = \frac{1}{2}.$$

Hence, the error is given by:

$$\begin{aligned}
 E_n &= \|f\|^2 - \sum_{k=-n}^n \left| \langle f, \phi_k \rangle \right|^2 = \frac{1}{3} - \left( 2 \sum_{k=1}^n \left( \frac{1}{2\pi k} \right)^2 + \left( \frac{1}{2} \right)^2 \right) \\
 &= \frac{1}{3} - \frac{1}{2\pi^2} \sum_{k=1}^n \frac{1}{k^2} - \frac{1}{4} \\
 &= \frac{1}{12} - \frac{1}{2\pi^2} \sum_{k=1}^n \frac{1}{k^2}.
 \end{aligned}$$

■

We remark that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

is the solution of the Basel problem to be discussed in Subunit 3.5.2. Hence,

$$E_n > \frac{1}{12} - \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{1}{12} - \frac{1}{2\pi^2} \frac{\pi^2}{6} = 0.$$

We are now ready to establish the following pointwise convergence result.

**Theorem 3.4.2** *Let  $f \in L_2(-\pi, \pi]$  be extended periodically to  $\mathbb{R}$ . Then for any  $x_0 \in \mathbb{R}$ , if both of the one-sided derivatives of  $f(x)$  exist and are finite at  $x = x_0$ ; that is, if*

$$f'(x_0^+) = \lim_{h \rightarrow 0^+} \frac{f(x_0 + h) - f(x_0^+)}{h},$$

$$f'(x_0^-) = \lim_{h \rightarrow 0^+} \frac{f(x_0 - h) - f(x_0^-)}{-h}$$

*exist, then the Fourier series  $(Sf)(x)$  of  $f(x)$  converges to  $\frac{1}{2}(f(x_0^-) + f(x_0^+))$  at  $x = x_0$ . Here,  $f(x_0^+)$  and  $f(x_0^-)$  denote the right-hand and left-hand limits of  $f$  at  $x_0$ .*

To prove this theorem, we first observe that in view of the  $2\pi$ -periodic extension of  $f(x)$ ,

$$\begin{cases} f(\pi^+) = f(-\pi^+), & f(-\pi^-) = f(\pi^-); \\ f'(\pi^+) = f'(-\pi^+), & f'(-\pi^-) = f'(\pi^-). \end{cases} \quad (3.4.2)$$

Secondly, recall from (3.3.1) of Subunit 3.3.1 that the partial sums  $S_n f$  of the Fourier series  $Sf$  of the  $2\pi$ -periodic extension of  $f$  can be formulated as follows:

$$\begin{aligned}(S_n f)(x) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x-t) D_n(t) dt \\ &= \frac{1}{2\pi} \left( \int_{-\pi}^0 + \int_0^{\pi} \right) (f(x-t) D_n(t)) dt \\ &= \frac{1}{2\pi} \int_0^{\pi} (f(x+t) + f(x-t)) D_n(t) dt,\end{aligned}$$

where the change of the variables of integration  $u = -t$  for the integral over  $[-\pi, 0]$  and the fact that  $D_n(-t) = D_n(t)$  have been applied. Since

$$\frac{1}{2\pi} \int_0^{\pi} D_n(t) dt = \frac{1}{2},$$

by (3.3.2), we have, for any  $x = x_0$ ,

$$\begin{aligned}(S_n f)(x_0) &- \frac{1}{2} (f(x_0^-) + f(x_0^+)) \\ &= \frac{1}{2\pi} \int_0^{\pi} (f(x_0+t) + f(x_0-t)) - (f(x_0^-) + f(x_0^+)) D_n(t) dt \\ &= \int_0^{\pi} h_1(t) \sin nt dt + \int_0^{\pi} h_2(t) \cos nt dt.\end{aligned}$$

Here, in view of the formula of  $D_n(x)$  in (3.3.4), we have introduced the two functions:

$$h_1(t) = \frac{\cos \frac{t}{2}}{2\pi \sin \frac{t}{2}} (f(x_0+t) - f(x_0^+) + f(x_0-t) - f(x_0^-));$$

$$h_2(t) = \frac{\sin \frac{t}{2}}{2\pi} (f(x_0+t) - f(x_0^+) + f(x_0-t) - f(x_0^-)).$$

It is clear that since  $|\sin \frac{t}{2}| \leq 1$  and  $f \in L_2[0, \pi]$ , we have  $h_2 \in L_2[0, \pi]$ . Also, by the assumption that  $f'(x_0^+)$  and  $f'(x_0^-)$  exist and the fact that  $\frac{t}{\pi} \leq \sin \frac{t}{2}$  for all  $0 \leq t \leq \pi$ , we may conclude that  $h_1(t)$  is bounded on  $[0, \pi]$ , so that  $h_1 \in L_2[0, \pi]$  as well.

Next, observe that the two families

$$\left\{ \frac{1}{\sqrt{\pi}} \sin t, \dots, \frac{1}{\sqrt{\pi}} \sin nt \right\}$$

and

$$\left\{ \frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}} \cos t, \dots, \frac{1}{\sqrt{\pi}} \cos nt \right\}$$

are orthonormal families in  $L_2[-\pi, \pi]$ . So, if we extend  $h_1 \in L_2[0, \pi]$  to an odd function in  $L_2[-\pi, \pi]$  and  $h_2 \in L_2[0, \pi]$  to an even function in  $L_2[-\pi, \pi]$ , we may simplify the Fourier coefficients of the extended functions as follows:

$$\begin{cases} a_0(h_1) = \int_{-\pi}^{\pi} h_1(x) \frac{1}{\sqrt{2\pi}} dx = 0, \\ a_k(h_1) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} h_1(x) \cos kx dx = 0, \quad k = 1, \dots, n; \\ b_k(h_1) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} h_1(x) \sin kx dx = \frac{2}{\sqrt{\pi}} \int_0^{\pi} h_1(x) \sin kx dx, \end{cases}$$

and

$$\begin{cases} a_0(h_2) = \int_{-\pi}^{\pi} h_2(x) \frac{1}{\sqrt{2\pi}} dx = \sqrt{\frac{2}{\pi}} \int_0^{\pi} h_2(x) dx, \\ a_k(h_2) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} h_2(x) \cos kx dx = \frac{2}{\sqrt{\pi}} \int_0^{\pi} h_2(x) \cos kx dx, \quad k = 1, \dots, n; \\ b_k(h_2) = \frac{1}{\sqrt{\pi}} \int_{-\pi}^{\pi} h_2(x) \sin kx dx = 0. \end{cases}$$

Therefore, it follows from the above Theorem 3.4.1 that the sequences  $\{a_n(h_2)\}$  and  $\{b_n(h_1)\}$  converge to zero. The reason is that for the infinite series  $\sum |a_n(h_2)|^2$  to converge, the sequence  $\{|a_n(h_2)|^2\}$  must tend to zero. This implies that  $\{a_n(h_2)\}$  converges to zero. Similarly,  $\{b_n(h_1)\}$  also converges to zero. That is, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} (S_n f)(x_0) - \frac{1}{2}(f(x_0^-) + f(x_0^+)) \\ = \lim_{n \rightarrow \infty} \left( \int_0^{\pi} h_1(t) \sin nt dt + \int_0^{\pi} h_2(t) \cos nt dt \right) = 0. \end{aligned} \quad (3.4.3)$$

■

As a continuation of the pointwise convergence result in Theorem 3.4.2, we will show that if  $f'(x)$  is also in  $L_2(-\pi, \pi]$ , then the  $2\pi$ -periodic extension of  $f(x)$ , also denoted by  $f(x)$ , must be a continuous function in  $\mathbb{R}$ , and that its Fourier series  $(Sf)(x)$  converges uniformly in  $\mathbb{R}$ .

**Theorem 3.4.3** *Let  $f(x)$  be a continuous function in  $L_2[-\pi, \pi]$  with  $f(-\pi) = f(\pi)$ , such that its derivative  $f'(x)$ , defined almost everywhere, is in  $L_2(-\pi, \pi]$ . Then the Fourier series  $(Sf)(x)$  converges absolutely and uniformly to  $f(x)$ .*

For the proof of Theorem 3.4.3, we will show that the sequence of partial sums  $(S_n f)(x)$  of  $(Sf)(x)$  is a uniformly convergent Cauchy sequence, in that for an arbitrarily given  $\epsilon > 0$ , there exists some natural number  $N$ , such that for all  $n > m \geq N$ ,

$$|(S_n f)(x) - (S_m f)(x)| \leq \sum_{m \leq |k| < n} |c_k(f)| < \epsilon \quad (3.4.4)$$

for all  $x \in \mathbb{R}$ . Indeed, if (3.4.4) holds for all  $x$ , then the convergence of  $\{(S_n f)(x)\}$  is uniform; and since each  $(S_n f)(x)$  is a  $2\pi$ -periodic continuous function, the uniform limit is also  $2\pi$ -periodic continuous. Hence, in view of the point-wise convergence result from Theorem 3.4.2, the limit function is  $f(x)$ . Note that the last inequality in (3.4.4) implies absolute convergence as well.

In the following, since we will consider the Fourier series of both  $f(x)$  and its derivative  $f'(x)$ , we must use the notations  $c_k(f)$  and  $c_k(f')$  to distinguish the Fourier coefficients of the two functions. Now, applying integration by parts, we have, for  $k \neq 0$ ,

$$\begin{aligned} c_k(f) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-ikx} dx \\ &= \frac{i}{2\pi k} \left( e^{-ik\pi} f(\pi) - e^{ik\pi} f(-\pi) \right) - \frac{i}{2\pi k} \int_{-\pi}^{\pi} f'(x) e^{-ikx} dx \\ &= \frac{-i}{2\pi k} c_k(f'), \end{aligned}$$

where we have assumed that  $f(\pi) = f(-\pi)$  by the  $2\pi$ -periodic extension of  $f(x)$ . Therefore by the Cauchy-Schwarz inequality from Subunit 1.1.2, we have

$$\sum_{k \neq 0} |c_k(f)| \leq \frac{1}{2\pi} \left( \sum_{k \neq 0} \frac{1}{k^2} \right)^{\frac{1}{2}} \left( \sum_{k \neq 0} |c_k(f')|^2 \right)^{\frac{1}{2}}.$$

On the other hand, since  $f' \in L_2(-\pi, \pi]$ , it follows from the Bessel inequality from Theorem 3.4.1 that

$$\sum_{k \neq 0} |c_k(f')|^2 \leq \|f'\|^2 < \infty.$$

Hence, since

$$\sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,$$

the infinite series  $\sum_{k \neq 0} |c_k(f)|$  converges, so that the sequence  $\{\sum_{|k| < n} c_k(f)\}$  is a Cauchy sequence. That is, given  $\epsilon > 0$ , there exists a natural number  $N$ ,

such that

$$\sum_{m \leq |k| < n} |c_k(f)| < \epsilon,$$

for all integers  $m$  and  $n$  with  $n > m \geq N$ . This yields

$$\begin{aligned} |(S_n f)(x) - (S_m f)(x)| &= |\sum_{m < |k| \leq n} c_k(f) e^{ikx}| \\ &\leq \sum_{m < |k| \leq n} |c_k(f)| < \epsilon, \end{aligned}$$

as desired. ■

**Example 3.4.2** Let  $f(x)$  be a continuous function on the closed interval  $[-\pi, \pi]$ . Construct a linear polynomial  $p(x)$ , such that the function

$$\tilde{f}(x) = f(x) - p(x)$$

has continuous  $2\pi$ -periodic extension to  $\mathbb{R}$ . Show that if  $f'(x)$  exists almost everywhere with  $f' \in L_2(-\pi, \pi]$ , then the Fourier series  $(S\tilde{f})(x)$  converges uniformly to  $\tilde{f}(x)$  on  $[-\pi, \pi]$ .

**Solution** The linear polynomial

$$p(x) = \frac{f(\pi) - f(-\pi)}{2\pi} (x + \pi) + f(-\pi)$$

clearly satisfies  $p(-\pi) = f(-\pi)$  and  $p(\pi) = f(\pi)$ . Hence,

$$\tilde{f}(-\pi) = f(-\pi) - p(-\pi) = 0,$$

and

$$\tilde{f}(\pi) = f(\pi) - p(\pi) = 0;$$

so that  $\tilde{f}(-\pi) = \tilde{f}(\pi) = 0$ , and  $\tilde{f}$  has continuous  $2\pi$ -periodic extension to  $\mathbb{R}$ . Furthermore, since

$$\tilde{f}'(x) = f'(x) + c$$

where  $c = -((f(\pi) - f(-\pi))/2\pi)$  is a constant,  $\tilde{f}' \in L_2(-\pi, \pi]$ . By Theorem 3.4.3, the Fourier series  $(S\tilde{f})(x)$  of  $\tilde{f}(x)$  converges uniformly to  $\tilde{f}(x)$  on  $[-\pi, \pi]$ . ■

### 3.4.2 Trigonometric approximation

In this subunit, we will apply Theorem 3.3.4 from Subunit 3.3.3 to show that the family of trigonometric polynomials is dense in the space  $L_2[-\pi, \pi]$ , meaning that for any function  $f \in L_2[-\pi, \pi]$ , there exists a sequence of trigonometric polynomials  $p_n$ , such that

$$\|f - p_n\|_2^2 = \int_{-\pi}^{\pi} |f(x) - p_n(x)|^2 dx$$



converges to 0, as  $n$  tends to infinity.

Without requiring any knowledge of the Lebesgue integration theory, we only consider piecewise continuous functions in  $L_2[-\pi, \pi]$ . Hence, to apply Theorem 3.3.4 from Subunit 3.3.3, it is sufficient to show that every piecewise continuous function can be approximated as closely as we wish by continuous functions in the  $L_2[-\pi, \pi]$  norm.

Since a piecewise continuous  $2\pi$ -periodic function has at most finitely many jump discontinuities, we may assume that  $f(x)$  has  $m$  jump discontinuities at  $x_1, \dots, x_m$  in  $(-\pi, \pi]$ , where  $-\pi < x_1 < \dots < x_m \leq \pi$ , and set

$$x_{m+1} = x_1 + 2\pi.$$

Let  $\eta > 0$  be so chosen that the intervals

$$I_k = [x_k - \eta, x_k + \eta], \quad k = 1, \dots, m$$

do not overlap. This allows us to introduce a  $2\pi$ -periodic continuous function  $\tilde{f}(x)$ , defined by:

$$\tilde{f}(x) = f(x), \text{ for } x \notin \cup_{k=1}^m I_k, \quad (3.4.5)$$

and

$$\tilde{f}(x) = \frac{f(x_k + \eta) - f(x_k - \eta)}{2\eta} (x - x_k + \eta) + f(x_k - \eta), \quad (3.4.6)$$

for  $x \in I_k$  and  $k = 1, \dots, m$ .

For each  $k = 1, \dots, m$ , extend the function  $f(x)$  from the open interval  $(x_k, x_{k+1})$  to a continuous function on the closed interval  $[x_k, x_{k+1}]$ , by taking the one-sided limits; that is,

$$f(x_k) = f(x_k^+) = \lim_{0 < x - x_k \rightarrow 0} f(x),$$

and

$$f(x_{k+1}) = f(x_{k+1}^-) = \lim_{0 < x_{k+1} - x \rightarrow 0} f(x).$$

Then the extended function, being continuous on a compact interval, is bounded. Set

$$M_k = \max_{x_k \leq x \leq x_{k+1}} |f(x)|. \quad (3.4.7)$$

Hence, in view of (3.4.5)–(3.4.7), we have

$$\begin{aligned} \|f - \tilde{f}\|_2^2 &= \int_{-\pi}^{\pi} |f(x) - \tilde{f}(x)|^2 dx \\ &= \sum_{k=1}^m \int_{I_k} |f(x) - \tilde{f}(x)|^2 dx \\ &\leq 4\eta \sum_{k=1}^m M_k^2. \end{aligned}$$

Given  $\eta > 0$ . By choosing  $\eta$  so small that

$$0 < \eta < \epsilon^2 / \left( 4 \sum_{k=1}^m M_k^2 \right),$$

we have

$$\|f - \tilde{f}\|_2 < \epsilon.$$

Next, since  $\tilde{f} \in C[-\pi, \pi]$  with  $\tilde{f}(-\pi) = \tilde{f}(\pi)$ , we may apply Theorem 3.3.4 of Subunit 3.3.3 to conclude that the trigonometric polynomials

$$\tilde{p}_n(x) = (C_n \tilde{f})(x),$$

defined by the Césaro means of the  $n^{\text{th}}$  partial sums  $S_n \tilde{f}$  of the Fourier series of  $\tilde{f}$ , satisfy

$$\|\tilde{f} - \tilde{p}_n\|_\infty \rightarrow 0,$$

for  $n \rightarrow \infty$ . Hence, by introducing the trigonometric polynomials

$$p_n(x) = (C_n f)(x),$$

we have

$$\begin{aligned} \|f - p_n\|_2 &= \|(f - \tilde{f}) + (\tilde{f} - C_n \tilde{f}) + (C_n \tilde{f} - C_n f)\|_2 \\ &\leq \|f - \tilde{f}\|_2 + \|\tilde{f} - \tilde{p}_n\|_2 + \|C_n(\tilde{f} - f)\|_2 \\ &\leq 2 \|f - \tilde{f}\|_2 + \|\tilde{f} - \tilde{p}_n\|_2 \\ &\leq \sqrt{2\pi} \left( 2\|f - \tilde{f}\|_\infty + \|\tilde{f} - \tilde{p}_n\|_\infty \right) \\ &< 6\sqrt{2\pi} \epsilon + \sqrt{2\pi} \|\tilde{f} - \tilde{p}_n\|_\infty, \end{aligned}$$

where the result (3.3.8) of Subunit 3.3.2 has been applied to conclude that

$$\begin{aligned} \|C_n(\tilde{f} - f)\|_2 &= \|\sigma_n \star (\tilde{f} - f)\|_2 \\ &\leq \|\tilde{f} - f\|_2. \end{aligned}$$

Hence, since  $\|\tilde{f} - \tilde{p}_n\|_\infty \rightarrow 0$  and  $\epsilon > 0$  is arbitrary, we have established the following result.

**Theorem 3.4.4** *For any  $f \in L_2(-\pi, \pi]$ , there exist trigonometric polynomials  $p_n(x)$ , such that*

$$\|f - p_n\|_2 \rightarrow 0.$$

The result in Theorem 3.4.4 implies that the family  $\{e^{ikx}\}$ ,  $k = 0, \pm 1, \pm 2, \dots$ , is complete in  $L_2(-\pi, \pi]$ .

As an immediate application of Theorem 3.4.4, we may conclude from the best mean-square approximation result for partial sums of Fourier series in Theorem 3.2.2 of Subunit 3.2.3 the following.

**Theorem 3.4.5** *Let  $f \in L_2(-\pi, \pi]$  and  $S_n f$  be the  $n^{\text{th}}$  partial sum of its Fourier series  $Sf$ . Then*

$$\|f - S_n f\|_2 \rightarrow 0$$

*as  $n \rightarrow \infty$ . That is, the Fourier series  $Sf$  of  $f$  converges to  $f$  in the  $L_2$ -norm.*

### 3.5 Parseval's Identity

As a consequence of the “completeness” property, studied in Subunit 3.4, we will show that Bessel's inequality, as derived in Theorem 3.4.1 of Subunit 3.4.1, becomes an equality, when the orthonormal family is complete. In other words, if  $\{\phi_k\}$  is a complete orthonormal family in an inner-product space  $\mathbb{V}$ , then we have the following equality, called Parseval's identity:

$$\sum_k |\langle f, \phi_k \rangle|^2 = \|f\|^2$$

for all  $f \in \mathbb{V}$ . A complete orthonormal family will be called an orthonormal basis.

The objective of this subunit is to derive such identities for various orthonormal bases in Subunit 3.5.1, and to apply Parseval's identity to solving the Basel problem in Subunit 3.5.2. Furthermore, we will introduce the notions of Bernoulli numbers and Bernoulli polynomials to study the extension of Basel problem by L. Euler to all even powers, and to derive Euler's extension by applying the Fourier series of Bernoulli polynomials. Euler's solution of the general Basel problem will be called Euler's formula, shown in (3.5.5).

#### 3.5.1 Derivation of Parseval's identities

Let  $\{\phi_k\}$  be an orthonormal family in an inner-product space  $\mathbb{V}$ , such that  $\{\phi_k\}$  is complete in  $\mathbb{V}$ , in the sense that every  $f \in \mathbb{V}$  can be approximated as closely as we wish, by finite linear combinations of  $\{\phi_k\}$ . More precisely, for

an given  $\epsilon > 0$ , there exist constants  $d_1, \dots, d_n$  such that

$$\|f - \sum_{k=1}^n d_k \phi_k\| < \epsilon.$$

Then by the same derivation of Theorem 3.2.2 of Subunit 3.2.3, we may conclude that

$$\|f - \sum_{k=1}^n \langle f, \phi_k \rangle \phi_k\| \leq \|f - \sum_{k=1}^n d_k \phi_k\|$$

for all choices of  $\{d_1, \dots, d_n\}$ , so that

$$\lim_{n \rightarrow \infty} \|f - \sum_{k=1}^n \langle f, \phi_k \rangle \phi_k\| = 0. \quad (3.5.1)$$

For this reason, a complete orthonormal family  $\{\phi_k\}$  of  $\mathbb{V}$  is called an orthonormal basis of  $\mathbb{V}$ . In Theorem 3.4.1 of Subunit 3.4.1, if the orthonormal family  $\{\phi_k\}$  is an orthonormal basis, then Bessel's inequality becomes Parseval's identity as follow.

**Theorem 3.5.1** *Let  $\{\phi_k\}$  be an orthonormal basis of an inner-product space  $\mathbb{V}$ . Then for every  $f \in \mathbb{V}$ ,*

$$\sum_k |\langle f, \phi_k \rangle|^2 = \|f\|^2. \quad (3.5.2)$$

*This is called Parseval's identity.*

The proof of (3.5.2) is an extension of that of Theorem 3.4.1 in Subunit 3.4.1. Indeed, for  $\mathbf{x} = f(x)$  and

$$\mathbf{y}_n = - \sum_{|k| \leq n} \langle f, \phi_k \rangle \phi_k,$$

recall that

$$\langle f, \mathbf{y}_n \rangle = \langle \mathbf{y}_n, f \rangle = -\|\mathbf{y}_n\|^2,$$

so that it follows from Bessel's inequality in Theorem 3.4.1 that

$$\begin{aligned} 0 &\leq \|f\|^2 - \sum_{|k| \leq n} |\langle f, \phi_k \rangle|^2 \\ &= \|f\|^2 - \|\mathbf{y}_n\|^2 \\ &= \|f\|^2 + \langle f, \mathbf{y}_n \rangle + \langle \mathbf{y}_n, f \rangle + \|\mathbf{y}_n\|^2 \\ &= \|f + \mathbf{y}_n\|^2 = \left\| f - \sum_{|k| \leq n} \langle f, \phi_k \rangle \phi_k \right\|^2 \rightarrow 0, \end{aligned}$$

by applying (3.5.1). ■

In the following, we apply Theorem 3.5.1 to various orthonormal bases of the  $L_2$  spaces, with the first being the Fourier expansion of functions in  $L_2(\pi, \pi]$ .

**Theorem 3.5.2** *Let  $f \in L_2(-\pi, \pi]$  and  $\{c_k\}$  be the sequence of Fourier coefficients of  $f$ , namely:  $c_k = \frac{1}{2\pi} \langle f, c_k \rangle$  where  $e_k(x) = e^{ikx}$ . Then*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx = \sum_{k=-\infty}^{\infty} |c_k|^2. \quad (3.5.3)$$

Parseval's identity in (3.5.3) assures that the "energy" of the sequence  $\{c_k\}$  of the Fourier coefficients of  $f \in L_2(-\pi, \pi]$  preserves the energy of  $f$ .

When the interval  $[-\pi, \pi]$  is replaced by  $[-d, d]$  for any value  $d > 0$ , then Parseval's identity can be formulated as follows:

$$\frac{1}{2d} \int_{-d}^d |f(x)|^2 dx = \sum_{k=-\infty}^{\infty} |c_k|^2, \quad \text{for all } f \in L_2[-d, d].$$

For the cosine and/or sine series, we have the following versions of Parseval's identity:

- (1) For the interval  $[-\pi, \pi]$ :

The Fourier cosine and sine coefficients in (3.1.5) of Subunit 3.1.1 satisfy Parseval's identity:

$$\frac{|a_0|^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} (|a_k|^2 + |b_k|^2) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |f(x)|^2 dx,$$

for all  $f \in L_2(-\pi, \pi]$ .

- (2) In general, for any  $d > 0$ , the Fourier cosine and sine coefficients in (3.1.7) of Subunit 3.1.1 satisfy Parseval's identity:

$$\frac{|a_0|^2}{4} + \frac{1}{2} \sum_{k=1}^{\infty} (|a_k|^2 + |b_k|^2) = \frac{1}{2d} \int_{-d}^d |f(x)|^2 dx.$$

To derive Parseval's identity for cosine/sine series, we observe that since  $c_0 = a_0/2$  and

$$\begin{aligned} |c_k|^2 + |c_{-k}|^2 &= \frac{1}{4} |a_k - ib_k|^2 + \frac{1}{4} |a_k + ib_k|^2 \\ &= \frac{1}{2} (|a_k|^2 + |b_k|^2), \quad k = 1, 2, \dots, \end{aligned}$$

where we have applied the identity that

$$|a - ib|^2 + |a + ib|^2 = 2(|a|^2 + |b|^2),$$

which is valid for all complex numbers  $a$  and  $b$ . ■

### 3.5.2 The Basel problem and Fourier method

Before the birth of Calculus in the 17<sup>th</sup> century, P. Mengoli proposed the problem of finding the exact value of the infinite sum

$$\sum_{k=1}^{\infty} \frac{1}{k^2}$$

in 1644. This so-called **Basel Problem** became one of the most popular problems since Calculus was introduced, and such outstanding mathematicians, as the Bernoulli brothers, Jacob and Johan, tried but failed in meeting the challenge. It was not till the year 1735, when Johan Bernoulli's student, L. Euler solved the problem by using Taylor's series expansion. In this subunit, we will compute the infinite series

$$\sum_{k=1}^{\infty} \frac{1}{k^{2n}}$$

for  $n = 1, 2, 3$ , by applying Parseval's identity to certain appropriate function. It must be pointed out that the notion of Fourier series was introduced many years later by J. Fourier (1768-1830).

**First Solution** Let  $f_1(x)$  be the function introduced in Example 3.1.1 of Subunit 3.1.1. Then it is clear that  $\|f_1\|_2^2 = 2\pi$ . Hence, it follows from (3.1.4) in the same example that

$$\begin{aligned} 1 &= \sum_{\ell=-\infty}^{\infty} \left| \frac{-2i}{\pi(2\ell+1)} \right|^2 \\ &= \frac{4}{\pi^2} \left( \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=-\infty}^{-1} \frac{1}{(2\ell+1)^2} \right) \\ &= \frac{4}{\pi^2} \left( \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=1}^{\infty} \frac{1}{(2\ell-1)^2} \right) \\ &= \frac{4}{\pi^2} \left( \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} + \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} \right) \\ &= \frac{8}{\pi^2} \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2}, \end{aligned}$$

or

$$\sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} = \frac{\pi^2}{8}. \quad (3.5.4)$$

To complete the solution, we simply partition the sum over  $k$  into two sums, with one over even  $k$ 's and the other over odd  $k$ 's, namely:

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{k^2} &= \sum_{\ell=1}^{\infty} \frac{1}{(2\ell)^2} + \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^2} \\ &= \frac{1}{4} \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} + \frac{\pi^2}{8}, \end{aligned}$$

by applying (3.5.4), so that

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \left(1 - \frac{1}{4}\right)^{-1} \frac{\pi^2}{8} = \frac{4}{3} \frac{\pi^2}{8} = \frac{\pi^2}{6}.$$

■

**Second Solution** We may also select the function  $f(x) = x$  on the interval  $[0, 1]$  in Example 3.4.1 of Subunit 3.4.1 to solve the Basel problem. Indeed, from the error formula in Example 3.4.1, we have

$$0 < E_n = \frac{1}{12} - \frac{1}{2\pi^2} \sum_{k=1}^n \frac{1}{k^2} \rightarrow 0$$

as  $n \rightarrow \infty$ , yielding

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{2\pi^2}{12} = \frac{\pi^2}{6},$$

where  $E_n \rightarrow 0$  in view of Parseval's identity. ■

In the following two examples, we further explore the application of Parseval's identity for Fourier series to computing higher-order infinite sums.

**Example 3.5.1** Apply Parseval's identity to compute  $\sum_{k=1}^{\infty} \frac{1}{k^4}$  by selecting an appropriate function  $f_2 \in L_2(-\pi, \pi]$ .

**Solution** We choose  $f_2(x) = |x|$  for  $x \in (-\pi, \pi]$  and extend this function from  $(-\pi, \pi]$  to  $\mathbb{R}$  by setting  $f_2(x) = f_2(x + 2\pi)$ , so that  $f_2 \in PC_{2\pi}^*$ . Observe that  $f_2'(x) = f_1(x)$  in Example 3.1.1 of Subunit 3.1.1. Hence, by integration by parts, we have, for  $k \neq 0$ ,

$$\begin{aligned} c_k(f_2) &= \frac{1}{2\pi} \left[ f_2(x) \frac{e^{-ikx}}{-ik} \right]_{-\pi}^{\pi} - \frac{1}{2\pi} \int_{-\pi}^{\pi} f_1(x) \frac{e^{-ikx}}{-ik} dx \\ &= \frac{-i}{k} c_k(f_1), \end{aligned}$$

where the first term vanishes due to the fact that  $f_2(-\pi) = f_2(\pi)$ . Since  $c_0(f_2) = \frac{\pi}{2}$  and  $c_{2\ell}(f_1) = 0$  for all  $\ell \neq 0$ , we have

$$\begin{aligned} \sum_{k=-\infty}^{\infty} |c_k(f_2)|^2 &= \frac{\pi^2}{4} + \sum_{\ell=-\infty}^{\infty} \frac{4}{\pi^2(2\ell+1)^4} \\ &= \frac{\pi^2}{4} + \frac{8}{\pi^2} \sum_{\ell=0}^{\infty} \frac{1}{(2\ell+1)^4} \\ &= \frac{\pi^2}{4} + \left(1 - \frac{1}{2^4}\right)^{-1} \frac{8}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{k^4}. \end{aligned}$$

Therefore, by Parseval's identity, we have

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{k^4} &= \left( \frac{1}{2\pi} \|f_2\|_2^2 - \frac{\pi^2}{4} \right) \frac{2^4}{2^4-1} \frac{\pi^2}{8} \\ &= \left( \frac{\pi^2}{3} - \frac{\pi^2}{4} \right) \frac{2^4}{2^4-1} \frac{\pi^2}{2^3} = \frac{\pi^4}{90}. \end{aligned}$$

■

**Example 3.5.2** Select an appropriate function  $f_3 \in PC_{2\pi}^*$  to compute the exact value of  $\sum_{k=1}^{\infty} \frac{1}{k^6}$ .

**Solution** We choose the odd function extension of

$$f_3(x) = \frac{\pi^2}{8} - \frac{1}{2} \left(x - \frac{\pi}{2}\right)^2, \quad 0 \leq x \leq \pi,$$

by setting  $f_3(x) = -f_3(-x)$  for  $-\pi \leq x < 0$ . Then extend  $f_3(x)$  from  $(-\pi, \pi]$  periodically to  $f_3 \in PC_{2\pi}^*$ . Observe that

$$f_3'(x) = \frac{\pi}{2} - f_2(x),$$

$f_3(-\pi) = f_3(\pi) = 0$ , and that

$$\frac{1}{2\pi} \|f_3\|_2^2 = \frac{\pi^4}{120}.$$

Hence, by applying Parseval's identity and by following the same method of derivation in Example 3.5.1, we have

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{1}{k^6} &= \frac{2^6}{2^6-1} \frac{\pi^2}{8} \frac{1}{2\pi} \|f_3\|_2^2 \\ &= \frac{8\pi^2}{63} \cdot \frac{\pi^4}{120} = \frac{\pi^6}{945}. \end{aligned}$$

■



### 3.5.3 Bernoulli numbers and Euler's formula

As mentioned in Subunit 3.5.2, the Basel problem posed by P. Mengoli in the year 1644, which created a lot of excitement among mathematicians, was solved by L. Euler in 1735, before the introduction of Fourier series by J. Fourier in the early 1800's.

In fact, Euler not only solved the Basel problem, but also derived the exact formula

$$\sum_{k=1}^{\infty} \frac{1}{k^{2n}} = \frac{(-1)^{n-1} 2^{2n-1} \pi^{2n}}{(2n)!} b_{2n}, \quad (3.5.5)$$

for all  $n = 1, 2, \dots$ , where the  $b_{2n}$ 's are the Bernoulli numbers, whose exact values can be easily computed recursively. We will call (3.5.5) Euler's formula. The first three Bernoulli numbers  $b_{2n}$  are:

$$b_2 = \frac{1}{6}, \quad b_4 = -\frac{1}{30}, \quad b_6 = \frac{1}{42}.$$

On the other hand, the problem of finding the exact values of

$$\sum_{k=1}^{\infty} \frac{1}{k^{2n+1}}, \quad n = 1, 2, \dots$$

remains unsolved, although Euler was able to compute the infinite sums for at least  $n = 1$ , and  $n = 2$  fairly accurately, of course without the computer (in the 18<sup>th</sup> century).

In general, the Bernoulli numbers  $b_0, b_1, b_2, \dots$  were introduced by Jacob Bernoulli by using the Taylor series expansion of the function  $t/(e^t - 1)$ , as follows:

$$\frac{t}{e^t - 1} = \sum_{j=0}^{\infty} \frac{b_j}{j!} t^j. \quad (3.5.6)$$

Hence, it is clear that

$$b_0 = 1. \quad (3.5.7)$$

Now, apply the Taylor series expansion of  $e^t$  to write

$$\begin{aligned} t &= \left( \sum_{j=0}^{\infty} \frac{b_j}{j!} t^j \right) (e^t - 1) \\ &= \left( \sum_{j=0}^{\infty} \frac{b_j}{j!} t^j \right) \left( \sum_{k=1}^{\infty} \frac{1}{k!} t^k \right) \\ &= \sum_{\ell=1}^{\infty} \sum_{k=1}^{\ell} \frac{b_{\ell-k}}{(\ell-k)! k!} t^{\ell}. \end{aligned}$$

Hence, by applying (3.5.7), we have

$$\sum_{\ell=2}^{\infty} \sum_{k=1}^{\ell} \frac{b_{\ell-k}}{(\ell-k)! k!} t^{\ell} = 0,$$

or equivalently,

$$\sum_{k=1}^{\ell} \frac{b_{\ell-k}}{(\ell-k)! k!} = 0, \quad \ell = 2, 3, \dots$$

By a change of indices, we may conclude that

$$\sum_{k=0}^j \frac{b_{j-k}}{(j-k)! (k+1)!} = 0, \quad j = 1, 2, \dots$$

Hence,  $b_1, b_2, \dots$  can be computed recursively, by applying the formula

$$b_j = - \sum_{k=1}^j \frac{b_{j-k}}{(j-k)! (k+1)!} \quad (3.5.8)$$

for  $j = 1, 2, \dots$ , with initial value  $b_0 = 1$  as given in (3.5.7).

The function  $t/(e^t - 1)$  in (3.5.6) is called the generating function of the Bernoulli numbers, which are the values  $b_j = B_j(0)$  of the Bernoulli polynomials  $B_n(x)$  evaluated at  $x = 0$ . Here, the Bernoulli polynomials are defined by using the generating function:

$$\frac{te^{xt}}{e^t - 1} = \sum_{n=0}^{\infty} B_n(x) \frac{t^n}{n!}, \quad (3.5.9)$$

again by applying the Taylor series expansion. From (3.5.9), it is not difficult to show that  $B_j(x)$  is a monic polynomial (that is, with 1 as its leading coefficient), given by

$$B_n(x) = \sum_{k=0}^n \binom{n}{k} b_k x^{n-k}. \quad (3.5.10)$$

To derive the formula (3.5.10), we apply (3.5.6) and the Taylor series ex-

pansion of  $e^{xt}$  to (3.5.9), yielding:

$$\begin{aligned}
 \sum_{n=0}^{\infty} B_n(x) \frac{t^n}{n!} &= (e^{xt}) \frac{t}{e^t - 1} \\
 &= \left( \sum_{k=0}^{\infty} \frac{x^k t^k}{k!} \right) \left( \sum_{j=0}^{\infty} \frac{b_j}{j!} t^j \right) \\
 &= \sum_{n=0}^{\infty} \left[ \sum_{j=0}^n \frac{b_j}{(n-j)! j!} x^{n-j} \right] t^n \\
 &= \sum_{n=0}^{\infty} \left[ n! \sum_{j=0}^n \frac{b_j}{(n-j)! j!} x^{n-j} \right] \frac{t^n}{n!}.
 \end{aligned}$$

Hence, equating the coefficients of  $t^n/n!$ , we obtain

$$\begin{aligned}
 B_n(x) &= \sum_{j=0}^n \frac{n! b_j}{(n-j)! j!} x^{n-j} \\
 &= \sum_{j=0}^n \binom{n}{j} b_j x^{n-j}
 \end{aligned}$$

Among the other important properties of the Bernoulli numbers and Bernoulli polynomials, we only mention (but without proof) that

$$b_1 = -\frac{1}{2} \quad \text{and} \quad b_{2k+1} = 0, \quad k = 1, 2, \dots \quad (3.5.11)$$

and

$$\begin{cases} B_0(x) = 1 \\ B_1(x) = x - \frac{1}{2} \\ B'_n(x) = n B_{n-1}(x) \\ B_n(1-x) = (-1)^n B_n(x). \end{cases} \quad (3.5.12)$$

for  $n = 1, 2, \dots$

Let  $n$  be fixed, and consider the Bernoulli polynomial  $B_n(x)$  as a function in  $L_2[0, 1]$  to compute its Fourier coefficients  $c_k = c_k(B_n)$ ,  $k = 0, \pm 1, \dots$ , defined by

$$c_k = \int_0^1 B_n(x) e^{-i2\pi kx} dx. \quad (3.5.13)$$

To do so, we simply apply (3.5.12) to integrate (3.5.13) by parts, yielding

$$c_k = \frac{(-1)^{n-1} n!}{(-2\pi i k)^n}, \quad \text{for } k \neq 0,$$

while

$$\begin{aligned} c_0 &= \int_0^1 B_n(x) dx = \frac{1}{n+1} \int_0^1 B'_{n+1}(x) dx \\ &= \frac{1}{n+1} (B_{n+1}(1) - B_{n+1}(0)) \\ &= 0, \end{aligned}$$

because  $B_{2k+1}(1) = -b_{2k+1} = 0$  and  $B_{2k}(1) = B_{2k}(0)$  for all  $k = 1, 2, \dots$ . Since  $\{e^{i2\pi k}\}, k = 0, \pm 1, \pm 2, \dots$ , is an orthonormal basis of  $L_2[0, 1]$ , we may apply Parseval's identity to conclude that

$$\int_0^1 (B_n(x))^2 dx = \frac{2(n!)^2}{(2\pi)^{2n}} \sum_{k=1}^{\infty} \frac{1}{k^{2n}}. \quad (3.5.14)$$

To compute the integral on the left of the equality (3.5.14), we again apply (3.5.12) in the following computation of integration by parts, namely:

$$\begin{aligned} \int_0^1 (B_n(x))^2 dx &= B_n \frac{B_{n+1}(x)}{n+1} \Big|_0^1 - \int_0^1 \frac{n}{n+1} B_{n-1}(x) B_{n+1}(x) dx = \dots \\ &= (-1)^{n-1} \frac{n!}{(n+1)\dots(2n-1)} \int_0^1 B_1(x) B_{2n-1}(x) dx \\ &= (-1)^{n-1} \frac{n!}{(n+1)\dots(2n)} B_1(x) B_{2n}(x) \Big|_0^1 \\ &\quad - (-1)^{n-1} \frac{n!}{(n+1)\dots(2n)} \int_0^1 B_{2n}(x) dx \\ &= (-1)^{n-1} \frac{n!}{(n+1)\dots(2n)} \left( \frac{1}{2} B_{2n}(1) - \left(\frac{-1}{2}\right) B_{2n}(0) \right) \\ &= (-1)^{n-1} \frac{(n!)^2}{(2n)!} b_{2n}, \end{aligned}$$

where we have applied the properties in (3.5.12), specifically  $B_1(x) = x - \frac{1}{2}$  and  $B_{2n+1}(1) = -B_{2n+1}(0) = -b_{2n+1} = 0$ . Hence, by putting this into (3.5.14),

we obtain

$$\begin{aligned}\sum_{k=1}^{\infty} \frac{1}{k^{2n}} &= (-1)^{n-1} \frac{(n!)^2}{(2n)!} b_{2n} \frac{(2\pi)^{2n}}{2(n!)^2} \\ &= \frac{(-1)^{n-1} 2^{2n-1} \pi^{2n}}{(2n)!} b_{2n}\end{aligned}$$

which agrees with Euler's formula in (3.5.5).



# Unit 4

## TIME-FREQUENCY ANALYSIS

The Fourier transform (FT) introduced in this unit is an analogue of the sequence of Fourier coefficients of the Fourier series discussed in Unit 3, in that the normalized integral over the circle in the definition of Fourier coefficients is replaced by the integral over the real line to define the FT. While the Fourier series is used to recover the given function it represents from the sequence of Fourier coefficients, it is non-trivial to justify the validity of the seemingly obvious formulation of the inverse Fourier transform (IFT) for the recovery of a function from its FT. In this unit, the notions of localized FT (LFT) and localized inverse FT (LIFT) will be introduced, and an identity that governs the relationship between LFT and LIFT is also established, with the Gabor transform, with certain Gaussian time window, as an example. The importance of the Gabor transform is due to the fact that the Fourier transform of a Gaussian function remains to be a Gaussian function. The identity that governs the relation between LFT and LIFT is also applied to verify the validity of the IFT formulation, by using the Gaussian function with variance  $\sigma^2$  as the time localization window, and taking the limit, with  $\sigma^2$  approaching to zero. Another important consequence of this identity is the Uncertainty Principle, which states that the Gaussian is the only window function that provides optimal simultaneous time-frequency localization with area of the time-frequency window equal to 2. Discretization of any frequency-modulated sliding time-window of the LFT at the integer lattice yields a family of local time-frequency basis functions. Unfortunately, the Balian-Low restriction excludes any of such sliding time-window functions, including the Gaussian, to attain finite area of the time-frequency window, while providing stability for the family of local time-frequency basis functions, called a frame. This unit ends with a discussion of a way for avoiding the Balian-Low restriction by replacing the frequency-modulation of the sliding time-window function with modulation by certain cosine functions. More precisely, a family of stable local cosine basis functions, sometimes called Malvar “wavelets”, is introduced to achieve good time-frequency localization.

## 4.1 Fourier Transform

In this first subunit, after introducing the definition of the Fourier transform and the motivation of the transform in Subunit 4.1.1, the basic properties are derived in Subunit 4.1.2. Application to deriving the Sampling Theorem is given in Subunit 4.1.3, and other applications in Subunit 4.1.4.

### 4.1.1 Definition and essence of the Fourier transform

The notion of Fourier transform is introduced in this subunit. When a non-periodic function  $f$  is considered as an analog signal with time-domain  $(-\infty, \infty)$ , the Fourier transform of  $f$ , defined by

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-ix\omega} dx \quad (4.1.1)$$

is used to reveal the frequency content of  $f$ . An important application of the Fourier transform is that it takes the convolution filtering of the analog signal  $f(t)$  to multiplication of its Fourier transform  $\hat{f}(\omega)$  by the Fourier transform of the convolution filter. Precisely, if  $h$  denotes the filter with Fourier transform  $\hat{h}$ , then while the filtering output is given by the convolution operation:

$$(f \star h)(x) = \int_{-\infty}^{\infty} f(t) h(x-t) dt, \quad (4.1.2)$$

the Fourier transform of this output is simply the product of  $\hat{f}(\omega)$  and  $\hat{h}(\omega)$ ; that is,

$$(\widehat{f \star h})(\omega) = \hat{h}(\omega) \hat{f}(\omega). \quad (4.1.3)$$

Hence, in applications to signal processing, the filtering objective can be met by properly designing the filter characteristic,  $|\hat{h}(\omega)|$ . For example, to suppress or remove low-frequency contents, the filter function  $h$  should be so chosen that the magnitude of its Fourier transform,  $|\hat{h}(\omega)|$ , is small for the frequency variable  $\omega$  in some neighborhood of the zero frequency 0, called the stop-band. For the choice of such high-pass (or more generally, band-pass) filters  $h$ , the convolution operation of the input signal  $f$  with  $h$  reduces (or even removes) the low-frequency content of  $f$ . On the other hand, if the low-frequency content is to be retained and the high-frequency content to be suppressed, then the low-pass filter  $h(t)$  should have the property  $|\hat{h}(\omega)| \doteq 1$  for small values of  $|\omega|$ , while  $|\hat{h}(\omega)| \doteq 0$  for larger values of  $|\omega|$ .



### 4.1.2 Properties of the Fourier transform

#### References

- (1) Stanford University: Department of Electrical Engineering's "Lecture 1: The Fourier Transforms and Its Applications (YouTube)", "Lecture 6: The Fourier Transforms and Its Applications (YouTube)", and "Lecture 8: The Fourier Transforms and Its Applications (YouTube)".
- (2) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 319–329. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 4.1.3 Sampling Theorem

#### References

- (1) MIT: Department of Computational Science and Engineering's "Lecture 36: Sampling Theorem (YouTube)", presented by Gilbert Strang.
- (2) Charles K. Chui and Qingtang Jiang, "Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 336–338. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

## 4.2 Convolution Filter and Gaussian Kernel

In this subunit, convolution filtering with a filter function  $h(t)$ , as defined in (4.1.2) of Subunit 4.1.1, is elaborated. Since the Gaussian function provides

a lowpass filter with the optimal time-frequency localization property, we will compute its Fourier transform in Subunit 4.2.2. By applying the Gaussian, the inverse Fourier transform is introduced and studied in Subunit 4.2.3.

### 4.2.1 Convolution filter

#### References

- (1) Gilbert Strang's Computational Science and Engineering: "Lecture 32: Convolution (Part 2), Filtering (YouTube).

### 4.2.2 Fourier transform of the Gaussian

The Gaussian function is the only function whose Fourier transform remains to be a Gaussian function, or more precisely, another Gaussian function. Hence, since the Gaussian is a low-pass filter function, it is theoretically the optimal filter for simultaneous time-frequency localization. The Gaussian function is defined by

$$g_\sigma(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-(\frac{x}{2\sigma})^2}, \quad (4.2.1)$$

where  $\sigma > 0$ . The division by  $2\sigma\sqrt{\pi}$  in (4.2.1) is to assure the integral to be equal to 1, namely:

$$\int_{-\infty}^{\infty} g_\sigma(x) dx = 1, \text{ all } \sigma > 0. \quad (4.2.2)$$

To prove (4.2.2), observe that by changing the Cartesian coordinates to polar coordinates, we have

$$\begin{aligned} \left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy \\ &= \int_0^{\infty} \int_{-\pi}^{\pi} e^{-r^2} r dr d\theta = 2\pi \int_0^{\infty} e^{-r^2} r dr = \pi, \end{aligned}$$

which yields

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \quad (4.2.3)$$

after taking the square-root. Hence, for any parameter  $\alpha > 0$ , by the change of variables of integration from  $\sqrt{\alpha} x$  to  $x$ , we have

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}}, \quad (4.2.4)$$

which implies that (4.2.2) holds, by choosing  $\alpha = 1/(4\sigma^2)$ .

To compute the Fourier transform of  $g_\sigma(x)$ , we first consider  $v(x) = e^{-x^2}$  and formulate its Fourier transform as

$$\begin{aligned} G(\omega) &= \widehat{v}(\omega) = \int_{-\infty}^{\infty} e^{-x^2} e^{-i\omega x} dx \\ &= \int_{-\infty}^{\infty} e^{-(x^2+i\omega x)} dx. \end{aligned}$$

Then for  $y \in \mathbb{R}$ , we have

$$\begin{aligned} G(-iy) &= \int_{-\infty}^{\infty} e^{-(x^2+yx)} dx \\ &= e^{y^2/4} \int_{-\infty}^{\infty} e^{-(x+y/2)^2} dx = \sqrt{\pi} e^{y^2/4}. \end{aligned} \quad (4.2.5)$$

Hence, when the function

$$H(z) = G(z) - \sqrt{\pi} e^{-z^2/4} \quad (4.2.6)$$

is considered as a function of a complex variable  $z$ ,  $H(z)$  is analytic for all  $z \in \mathbb{C}$  (or  $H(z)$  is called an entire function) and  $H(-iy) = 0$  for all  $y \in \mathbb{R}$ . Recall that if an analytic function vanishes on a set with at least a finite accumulation point in the domain of analyticity, then the function vanishes identically in this domain. Hence,  $H(z) = 0$  for all  $z \in \mathbb{C}$ , so that

$$G(\omega) - \sqrt{\pi} e^{-\omega^2/4} = 0, \quad \omega \in \mathbb{R};$$

or

$$\widehat{v}(\omega) = \int_{-\infty}^{\infty} e^{-x^2} e^{-i\omega x} dx = \sqrt{\pi} e^{-\omega^2/4}. \quad (4.2.7)$$

This enables us to compute the Fourier transform  $\widehat{g}_\sigma(\omega)$  of  $g_\sigma(x)$ ; namely, in view of (4.2.1) and (4.2.7), we have

$$\begin{aligned} \widehat{g}_\sigma(\omega) &= \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-(\frac{x}{2\sigma})^2} e^{-i\omega x} dx \\ &= \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2} e^{-i(2\sigma\omega)y} (2\sigma) dy \\ &= \frac{1}{\sqrt{\pi}} \sqrt{\pi} e^{-(2\sigma\omega)^2/4} = e^{-(\sigma\omega)^2}. \end{aligned}$$

In other words, we have established the following result.

**Theorem 4.2.1** *The Fourier transform of the Gaussian function  $g_\sigma(x)$ , where  $\sigma > 0$ , defined in (4.2.1), is given by*

$$\widehat{g}_\sigma(\omega) = e^{-\sigma^2 \omega^2}. \quad (4.2.8)$$

**Remark 4.2.1** By (4.2.7) and the change of variables of integration, we also have

$$\int_{-\infty}^{\infty} e^{-\sigma^2 \omega^2} e^{ix\omega} d\omega = \frac{\sqrt{\pi}}{\sigma} e^{-\left(\frac{x}{2\sigma}\right)^2} = 2\pi g_\sigma(x). \quad (4.2.9)$$

This formula will be used to the study of the inversion of the Fourier transform in the next subunit. ■

### 4.2.3 Inverse Fourier transform

An immediate application of the Gaussian function is to introduce the inverse Fourier transform. Let us first consider the companion transform  $\mathbb{F}^\#$  of the Fourier transform  $\mathbb{F}$ , defined by

$$(\mathbb{F}^\# g)(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} g(\omega) e^{ix\omega} d\omega \quad (4.2.10)$$

for  $g \in L_1(\mathbb{R})$ .

**Remark 4.2.2** Observe that  $\mathbb{F}^\#$  introduced in (4.2.10) is related to the Fourier transform  $\mathbb{F}$  by

$$(\mathbb{F}^\# g)(x) = \frac{1}{2\pi} (\mathbb{F}g)(-x) = \frac{1}{2\pi} \overline{(\mathbb{F}\overline{g})(x)}. \quad (4.2.11)$$

■

Let  $\widehat{f}$  denote the Fourier transform of a given function  $f \in L_1(\mathbb{R})$ . Then under the additional assumption that  $\widehat{f} \in L_1(\mathbb{R})$ , we can recover  $f$  from  $\widehat{f}(\omega)$  by applying  $\mathbb{F}^\#$ , as in the following theorem.

**Theorem 4.2.2** *Let  $f$  be in  $L_1(\mathbb{R})$  or  $L_2(\mathbb{R})$ , and let  $\widehat{f}(\omega)$  denote its Fourier transform. Then under the assumption that  $\widehat{f} \in L_1(\mathbb{R})$ , the given function  $f$  can be recovered by applying the formula:*

$$f(x) = (\mathbb{F}^\# \widehat{f})(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\omega) e^{ix\omega} d\omega. \quad (4.2.12)$$

That is,  $\mathbb{F}^\# = \mathbb{F}^{-1}$  is the inverse Fourier transform (or IFT).

**Remark 4.2.3** The assumption  $f \in L_1(\mathbb{R})$  in Theorem 4.2.2 is not necessary and can be replaced by  $f \in L_2(\mathbb{R})$ . To remove this assumption, let us consider the following truncation of  $f$

$$f_N(x) = \begin{cases} f(x), & \text{for } |x| \leq N, \\ 0, & \text{for } |x| > N, \end{cases} \quad (4.2.13)$$

for  $N = 1, 2, \dots$ . Observe that each  $f_N$  is compactly supported. Thus, from the assumption that  $f \in L_2(\mathbb{R})$ , we have  $f_N \in (L_2 \cap L_1)(\mathbb{R})$ , so that  $\widehat{f}_N$  is well-defined. In addition, since  $\{f_N\}$  converges to  $f$  in  $L_2(\mathbb{R})$ ,  $\{f_N\}$  is a Cauchy sequence in  $L_2(\mathbb{R})$ . In the following, by applying the Gaussian function and its Fourier transform, it will be shown that for  $N = 1, 2, \dots$ ,

$$\|\widehat{f}_N\|_2^2 = 2\pi \|f_N\|_2^2. \quad (4.2.14)$$

It then follows from (4.2.14) and the fact that  $\{f_N\}$  is a Cauchy sequence in  $L_2(\mathbb{R})$ , that the sequence  $\{\widehat{f}_N(\omega)\}_N$  is also a Cauchy sequence in  $L_2(\mathbb{R})$ , and its limit, being a function in  $L_2(\mathbb{R})$ , can be used as the definition of the Fourier transform of  $f$ .

Let us now give a precise definition of the Fourier transform of  $L_2$  functions.

**Definition 4.2.1** Let  $f \in L_2(\mathbb{R})$  and  $f_N(x)$  be the truncations of  $f$  defined by (4.2.13). Then the Fourier transform  $\widehat{f}(\omega)$  of  $f$  is defined as the limit of  $\{\widehat{f}_N(\omega)\}_N$  in  $L_2(\mathbb{R})$ .

**Remark 4.2.4** The Fourier transform  $\widehat{f}$  for  $f \in L_2(\mathbb{R})$  defined in Definition 4.2.1 is independent of the choice of  $\{f_N\}$  in the sense that if  $\{g_N\}$  with  $g_N \in (L_2 \cap L_1)(\mathbb{R})$  converges to  $f$  in  $L_2(\mathbb{R})$ , then  $\{\widehat{g}_N\}$  has the same limit as  $\{\widehat{f}_N\}$ . Indeed, by (4.2.14), we have

$$\begin{aligned} \|\widehat{g}_N - \widehat{f}_N\|_2 &= 2\pi \|g_N - f_N\|_2 \\ &\leq 2\pi \|g_N - f\|_2 + 2\pi \|f - f_N\|_2 \rightarrow 0 \end{aligned}$$

as  $N \rightarrow \infty$ . Thus,  $\{\widehat{g}_N\}$  and  $\{\widehat{f}_N\}$  have the same limit. ■

Before we are ready to prove the above theorem on the Inverse Fourier transform, we need to establish the following two results.

**Theorem 4.2.3** Let  $\widehat{f}(\omega)$  be the Fourier transform of  $f \in L_2(\mathbb{R})$ . Then

$$\|\widehat{f}\|_2^2 = 2\pi \|f\|_2^2. \quad (4.2.15)$$

The identity (4.2.15) is called Plancherel's formula.

To prove the theorem, we observe that by the definition of the Fourier

transform for functions in  $L_2(\mathbb{R})$ , it is sufficient to derive the identity (4.2.15) for the truncated functions; and in view of Remark 4.2.4, we may assume that  $f \in (L_1 \cap L_2)(\mathbb{R})$  and consider its corresponding “autocorrelation function”  $F(x)$ , defined by

$$F(x) = \int_{-\infty}^{\infty} f(t) \overline{f(t-x)} dt. \quad (4.2.16)$$

By setting

$$f^-(x) = f(-x),$$

the autocorrelation can be viewed as the convolution of  $f$  and  $\overline{f^-}$ , namely:

$$F(x) = (f * \overline{f^-})(x),$$

so that it can be seen that  $F(x)$  is in both  $L_\infty$  and  $L_1(\mathbb{R})$ . Furthermore, it is also easy to see that

$$\widehat{F}(\omega) = \widehat{f}(\omega) \overline{\widehat{f}(\omega)} = |\widehat{f}(\omega)|^2 \quad (4.2.17)$$

Therefore, by applying (4.2.17) together with (4.2.8), we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} |\widehat{f}(\omega)|^2 e^{-\sigma^2 \omega^2} d\omega &= \int_{-\infty}^{\infty} \widehat{f}(\omega) \overline{\widehat{f}(\omega)} \widehat{g}_\sigma(\omega) d\omega \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(y) e^{-i\omega y} \int_{-\infty}^{\infty} \overline{f(x)} e^{i\omega x} \widehat{g}_\sigma(\omega) dx dy \right\} d\omega \\ &= \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} \overline{f(x)} \left\{ \int_{-\infty}^{\infty} \widehat{g}_\sigma(\omega) e^{i(x-y)\omega} d\omega \right\} dx dy \\ &= 2\pi \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} \overline{f(x)} g_\sigma(x-y) dx dy \\ &= 2\pi \int_{-\infty}^{\infty} f(y) \int_{-\infty}^{\infty} \overline{f(y+t)} g_\sigma(t) dt dy \\ &= 2\pi \int_{-\infty}^{\infty} F(-t) g_\sigma(t) dt = 2\pi (F * g_\sigma)(0), \end{aligned}$$

where the 4<sup>th</sup> equality follows from (4.2.9) in Remark 4.2.1 and the change of variables of integration  $t = x - y$  is applied to derive the last second line. In addition, since  $F(x)$  is a continuous function and  $\{g_\sigma\}$  constitutes a convolution identity, meaning that  $\{g_\sigma\}$  converges to the delta distribution, it follows from Theorem 4.2.1 with  $x_0 = 0$  that by taking the limit as  $\sigma \rightarrow 0$ ,

$$\|\widehat{f}\|_2^2 = 2\pi F(0) = 2\pi \|f\|_2^2. \quad (4.2.18)$$

This completes the proof of Theorem 4.2.3. ■

As a consequence of Theorem 4.2.3, we have the following result, also called Plancherel’s formula.

**Theorem 4.2.4** *Let  $f, g \in L_2(\mathbb{R})$ . Then*

$$\langle f, g \rangle = \frac{1}{2\pi} \langle \widehat{f}, \widehat{g} \rangle. \quad (4.2.19)$$

We only derive (4.2.19) for real-valued functions, since the complex-valued setting is similar, though requires more calculation. For real-valued  $f(x)$  and  $g(x)$ , since

$$\|f \pm g\|_2^2 = \|f\|_2^2 \pm 2\langle f, g \rangle + \|g\|_2^2,$$

we have

$$\langle f, g \rangle = \frac{1}{4} \left( \|f + g\|_2^2 - \|f - g\|_2^2 \right).$$

Hence, it follows from (4.2.15) (with  $f$  replaced by  $(f + g)$  and  $(f - g)$ , respectively) that

$$\begin{aligned} \langle f, g \rangle &= \frac{1}{4} \frac{1}{2\pi} \left( \|\widehat{f} + \widehat{g}\|_2^2 - \|\widehat{f} - \widehat{g}\|_2^2 \right) \\ &= \frac{1}{2\pi} \langle \widehat{f}, \widehat{g} \rangle. \end{aligned}$$

■

The following result, though similar to the formulation of (4.2.19), is more elementary and can be proved by a straight-forward change of order of integrations.

**Theorem 4.2.5** *Let  $f, g \in L_2(\mathbb{R})$ . Then*

$$\langle \widehat{f}, \overline{g} \rangle = \langle \widehat{g}, \overline{f} \rangle. \quad (4.2.20)$$

To prove this theorem, it is sufficient to derive (4.2.20) for  $f, g \in (L_1 \cap L_2)(\mathbb{R})$ , since the completion of  $(L_1 \cap L_2)(\mathbb{R})$  in  $L_2(\mathbb{R})$  is  $L_2(\mathbb{R})$ . Under this additional assumption, we may apply Fubini's theorem to interchange the order of integrations. Thus, we have

$$\begin{aligned} \langle \widehat{f}, \overline{g} \rangle &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x) e^{-iyx} dx \right\} g(y) dy \\ &= \int_{-\infty}^{\infty} f(x) \left\{ \int_{-\infty}^{\infty} g(y) e^{-ixy} dy \right\} dx \\ &= \int_{-\infty}^{\infty} f(x) \widehat{g}(x) dx = \langle \widehat{g}, \overline{f} \rangle. \end{aligned}$$

■

We are now ready to prove the result on the Inverse Fourier transform.

**Proof of Theorem 4.2.2** To prove Theorem 4.2.2, we set  $g = \widehat{f}$  and apply (4.2.11), Theorem 4.2.5, and Theorem 4.2.4, consecutively, to compute

$$\begin{aligned}
\|f - \mathbb{F}^\# g\|_2^2 &= \|f\|_2^2 - \langle f, \mathbb{F}^\# g \rangle - \langle \mathbb{F}^\# g, f \rangle + \|\mathbb{F}^\# g\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \langle f, \overline{\mathbb{F} g} \rangle - \frac{1}{2\pi} \langle \overline{\mathbb{F} g}, f \rangle + \left(\frac{1}{2\pi}\right)^2 \|\mathbb{F} g\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \langle \mathbb{F} g, \overline{f} \rangle - \frac{1}{2\pi} \overline{\langle \mathbb{F} g, f \rangle} + \left(\frac{1}{2\pi}\right)^2 (2\pi) \|\overline{g}\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \langle \widehat{f}, g \rangle - \frac{1}{2\pi} \overline{\langle \widehat{f}, g \rangle} + \frac{1}{2\pi} \|g\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \langle \widehat{f}, \widehat{f} \rangle - \frac{1}{2\pi} \overline{\langle \widehat{f}, \widehat{f} \rangle} + \frac{1}{2\pi} \|\widehat{f}\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 + \frac{1}{2\pi} \|\widehat{f}\|_2^2 \\
&= \|f\|_2^2 - \frac{1}{2\pi} \|\widehat{f}\|_2^2 = \|f\|_2^2 - \|f\|_2^2 = 0.
\end{aligned}$$

Here, we have used the fact that  $\|\bar{h}\|_2 = \|h\|_2$  and  $g = \widehat{f}$ . Hence,

$$(f - \mathbb{F}^\# g)(x) = 0$$

for almost all  $x$ , which implies the validity of (4.2.12). ■

**Remark 4.2.5** Since the operator  $\mathbb{F}^\#$  can be used to recover  $f$  from its Fourier transform  $\widehat{f}$ , as long as  $\widehat{f} \in L_1(\mathbb{R})$ , we will replace this notation by  $\mathbb{F}^{-1}$ , or

$$\mathbb{F}^{-1} = \mathbb{F}^\#, \quad (4.2.21)$$

called the inverse Fourier transform, IFT.

### 4.3 Localized Fourier Transform

In this subunit, the Fourier transform studied in Subunit 4.1 is localized by introducing a (sliding) time-window function. As an example, the Gabor transform, formulated by selecting a certain Gaussian function as the time-window, is discussed in some detail. In addition, it will be shown that the original function can be recovered by using a suitable corresponding (sliding) frequency-window function to localize the inverse Fourier transform. In Subunit 4.3.1,



the notion of short-time Fourier transform (STFT) is introduced as an example of the general localized Fourier transform (LFT) with time-window  $u(t)$ , along with its corresponding localized inverse Fourier transform (LIFT) with frequency window  $v(\xi)$ . The relationship between the time and frequency window functions  $u(t)$  and  $v(\xi)$  is derived in Subunit 4.3.3 to assure perfect recovery by the LIFT from the LFT. This result is an extension to the general setting from the Gabor transform studied in Subunit 4.3.2.

#### 4.3.1 Short-time Fourier Transform (STFT)

To compute the Fourier transform of a given function  $f(x)$ , if it is truncated by some characteristic function  $\chi_{(a,b)}(x)$ , then computation of

$$(f\chi_{(a,b)})^\wedge(\omega) = \int_a^b f(x)e^{-i\omega x} dx$$

is certainly simpler than that of  $\hat{f}(\omega)$ . In general, a more desirable window function  $u(x)$  could be used in place of the characteristic function  $\chi_{(a,b)}(x)$ , and this window should be allowed to slide (continuously) along the  $x$ -axis, instead of partitioning the  $x$ -axis into disjoint intervals. This is the key idea of the so-called “short-time” Fourier transform (STFT). Since this transform localizes the function  $f(x)$  before the Fourier transform is applied, we will also call it localized Fourier transform (LFT), as follows.

**Definition 4.3.1** Let  $u \in (L_1 \cap L_2)(\mathbb{R})$  and  $x \in \mathbb{R}$ . Then for any  $f \in L_2(\mathbb{R})$ , the integral transform

$$(\mathbb{F}_u f)(x, \omega) = \int_{-\infty}^{\infty} f(t)u(t-x)e^{-i\omega t} dt \quad (4.3.1)$$

is called the localized Fourier transform (LFT) or short-time Fourier transform (STFT) of the function  $f(x)$  at the time-frequency (or space-frequency) point  $(x, \omega) \in \mathbb{R}^2$ .

**Remark 4.3.1** In contrast with the Fourier transformation  $\mathbb{F}$  that takes a function  $f(x)$  from the time (or spatial) domain  $\mathbb{R}$  to  $\hat{f}(\omega)$  in the frequency domain  $\mathbb{R}$ , the LFT  $\mathbb{F}_u$ , with “time-window” function  $u(x)$ , takes  $f(x)$  from the time (or spatial) domain  $\mathbb{R}$  to the time-frequency domain  $\mathbb{R}^2$ . For this reason, we use  $t$  (instead of  $x$ ) as the dummy variable for the integration in (4.3.1), while reserving the variable  $x$  as the time component of the time-frequency coordinate  $(x, \omega) \in \mathbb{R}^2$ . ■

To localize the inverse Fourier transform, a topic to be studied in Subunit

4.3.3 later, we will use another window function  $v$  to define the localized inverse Fourier transform of the function  $\hat{f}(\omega)$ , as follows.

**Definition 4.3.2** Let  $v \in (L_1 \cap L_2)(\mathbb{R})$  and  $\omega \in \mathbb{R}$ . Then for any  $f(x) \in L_2(\mathbb{R})$  with Fourier transform  $\hat{f}(\omega)$ , the integral transform

$$(\mathbb{F}_v^\# \hat{f})(x, \omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) v(\xi - \omega) e^{ix\xi} d\xi \quad (4.3.2)$$

is called the localized inverse Fourier transform (LIFT) of  $\hat{f}(\omega)$  at the time-frequency (or spatial-frequency) point  $(x, \omega) \in \mathbb{R}^2$ .

Here, the notation  $\mathbb{F}^\#$  from (4.2.12) is adopted, since we will be interested in recovering the given function  $f$  from its LFT or STFT. This result will be derived in Subunit 4.3.3, by choosing a suitable frequency-window function  $v$  associated with the time-window function  $u$ . In the next Subunit 4.3.2, we will use the Gaussian function  $g_\sigma(x)$  defined in (4.2.1), with certain specific choice of  $\sigma > 0$ , as the window function  $u$  to demonstrate the power of the LFT. More precisely, by selecting the value  $\sigma = 1/(2\sqrt{\pi})$ , we will introduce the so-called Gabor transform which has the important property that the given function  $f$  can be recovered from its Gabor transform, simply by taking the inverse Fourier transform.

### 4.3.2 Gabor transform

If the Gaussian function  $g_\sigma(x)$  defined in (4.2.1) of Subunit 4.2.2, with  $\sigma = 1/(2\sqrt{\pi})$ , is used as the function  $u(x)$  in (4.3.1), then since  $g_\sigma$  is certainly in  $(L_1 \cap L_2)(\mathbb{R})$ , it can be used as a time-window for the LFT (or STFT). This particular LFT is called the Gabor transform, defined as follows.

**Definition 4.3.3** The integral transform

$$(\mathbb{G}f)(x, \omega) = \int_{-\infty}^{\infty} f(t) e^{-\pi(t-x)^2} e^{-it\omega} dt \quad (4.3.3)$$

of functions  $f \in L_1(\mathbb{R})$  is called the Gabor transform of  $f(x)$  at the  $(x, \omega)$  position of the time-frequency domain  $\mathbb{R}^2$ .

The reason for the choice of  $\sigma = 1/(2\sqrt{\pi})$  is that to recover  $f(x)$  from its Gabor transform  $(\mathbb{G}f)(x, \omega)$ , we may simply apply the inverse Fourier transformation  $\mathbb{F}^{-1} = \mathbb{F}^\#$  as defined in (4.2.10) and (4.2.20), with the integral over the frequency domain, namely:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{G}f)(x, \omega) e^{ix\omega} d\omega. \quad (4.3.4)$$

This formula holds even for measurable functions  $f(x)$  with at most polynomial growth in  $\mathbb{R}$ , since by using the Gaussian as the window function, both of the integrals in LFT and the inverse Fourier transform always exist and are finite. The formula is an application of Theorem 4.3.2 to be established in the next Subunit 4.3.3, by using the following three properties of the window function  $u(x) = e^{-\pi(x)^2}$ :

(a)  $u(0) = 1$  ;

(b)  $\widehat{u}(\omega) = e^{-\frac{1}{4\pi}(\omega)^2}$  ;

(b)  $\widehat{u}(-\omega) = \widehat{u}(\omega)$ .

The first two properties hold for the window function  $u(x) = g_\sigma(x)$  with  $\sigma = 1/(2\sqrt{\pi})$ , since  $u(x) = e^{-\sigma^2 x^2}$  and  $\widehat{u}(\sigma) = \widehat{g}_\sigma(\omega) = e^{-\sigma^2 \omega^2}$  for  $\sigma = 1/(2\sqrt{\pi})$  as well, by applying Theorem 4.2.1 of Subunit 4.2.2.

### 4.3.3 Inverse of localized Fourier transform

In this subunit, we will first show that if  $u \in (L_1 \cap L_2)(\mathbb{R})$  with Fourier transform  $\widehat{u}(\omega)$  also in  $(L_1 \cap L_2)(\mathbb{R})$ , then by choosing  $v(\omega) = (\overline{\mathbb{F}u})(\omega) = \widehat{u}(-\omega)$  for the LIFT  $\mathbb{F}_v^\#$  in (4.3.2), we have simultaneous time and frequency localization, as follows.

**Theorem 4.3.1** *Let  $u \in (L_1 \cap L_2)(\mathbb{R})$  with Fourier transform  $\mathbb{F}u = \widehat{u} \in (L_1 \cap L_2)(\mathbb{R})$ . Then for any  $f \in L_1(\mathbb{R})$ ,*

$$(\mathbb{F}_u f)(x, \omega) = e^{-ix\omega} \left( \mathbb{F}_{u^\star}^\# \widehat{f} \right)(x, \omega), \quad (4.3.5)$$

for  $(x, \omega) \in \mathbb{R}^2$ , when  $u^\star$  is defined by

$$u^\star(\xi) = \widehat{u}(-\xi), \quad \text{for } \xi \in \mathbb{R}.$$

**Proof** The proof of (4.3.5) follows by applying Plancherel's formula (4.2.19) in Theorem 4.2.4 of Subunit 4.2.3. Indeed, by considering the function

$$g(t) = \overline{u(t-x)} e^{i\omega t},$$

it follows from (4.2.19) that

$$\widehat{g}(\xi) = \widehat{u}(\xi - \omega) e^{-ix(\xi - \omega)} = \overline{\widehat{u}(\omega - \xi)} e^{-ix(\xi - \omega)},$$

so that for fixed  $(x, \omega)$ ,

$$\begin{aligned}
 (\mathbb{F}_u f)(x, \omega) &= \langle f, g \rangle = \frac{1}{2\pi} \langle \hat{f}, \hat{g} \rangle \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \hat{u}(\omega - \xi) e^{ix(\xi - \omega)} d\xi \\
 &= e^{-ix\omega} \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) u^*(\xi - \omega) e^{ix\xi} d\xi \\
 &= e^{-ix\omega} \left( \mathbb{F}_{u^*}^{\#} \hat{f} \right)(x, \omega).
 \end{aligned}$$

■

As an application of the above theorem, we derive the following formula for recovering the given function  $f$  from its LFT.

**Theorem 4.3.2** *Let  $u \in (L_1 \cap L_2)(\mathbb{R})$  with Fourier transform  $\hat{u} \in (L_1 \cap L_2)(\mathbb{R})$ , such that  $u(0) \neq 0$ . Then for any  $f \in (L_1 \cap L_2)(\mathbb{R})$  with  $\hat{f} \in L_1(\mathbb{R})$ ,*

$$f(x) = \frac{1}{u(0)} \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_u f)(x, \omega) e^{ix\omega} d\omega. \quad (4.3.6)$$

The derivation of (4.3.6) follows from Theorem 4.3.1 by multiplying both sides of (4.3.5) with  $e^{ix\omega}$  and then taking the integral; namely,

$$\begin{aligned}
 \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_u f)(x, \omega) e^{ix\omega} d\omega &= \frac{1}{2\pi} \int_{-\infty}^{\infty} (\mathbb{F}_{u^*}^{\#} \hat{f})(x, \omega) d\omega \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(-(\xi - \omega)) d\omega \right) d\xi \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(y) dy \right) d\xi \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(y) e^{i0y} dy \right) d\xi \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} u(0) d\xi \\
 &= u(0) \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{ix\xi} d\xi \right) = u(0) f(x),
 \end{aligned}$$

where Theorem 4.2.2 is applied to both  $\hat{u}$  and  $\hat{f}$  which are assumed to be in  $L_1(\mathbb{R})$ . ■

## 4.4 Uncertainty Principle

The main objective of this subunit is to introduce the notions of center and width of window functions for the purpose of quantifying the localization property of any given function and its Fourier transform. As already shown in the previous subunit, for simultaneous time and frequency localization, and hence for function recovery, the frequency localization window function is essentially the Fourier transform of the time localization window function. Hence, there is always a balance between localization in the time and frequency domains, in that to achieve better localization in the time-domain, localization in the frequency-domain must be sacrificed by a corresponding amount; and vice versa. In this regard, it is natural to consider the product of the widths of the time-window and frequency-window. We will derive the lower bound of this product for all choices of window functions and show that the Gaussian function is the only window function that achieves this lower bound. This result, called the uncertainty principle, will also be derived in this subunit.

### 4.4.1 Time-frequency localization window measurement

To quantify the localization properties of the LFT and LIFT, we introduce the notion of “window center” and “window width” in the following.

**Definition 4.4.1** *Let  $u \in (L_1 \cap L_2)(\mathbb{R})$  be a nonzero function such that  $xu(x) \in L_2(\mathbb{R})$ . Then*

$$x^* = \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} x|u(x)|^2 dx \quad (4.4.1)$$

*is called the center of the window function  $u(x)$ , and*

$$\Delta_u = \left\{ \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} (x - x^*)^2 |u(x)|^2 dx \right\}^{1/2} \quad (4.4.2)$$

*is called the radius of  $u(x)$ . In addition, the window width of  $u(x)$  is defined by  $2\Delta_u$ .*

Observe that for  $u \in L_2(\mathbb{R})$ , if  $xu(x) \in L_2(\mathbb{R})$ , then  $xu(x)^2 \in L_1(\mathbb{R})$ . Thus, the center  $x^*$  is well-defined. In view of the simultaneous time-frequency localization identity (4.3.5) of Subunit 4.3.3, it is imperative to come up with window functions  $u(x)$  such that both  $xu(x)$  and  $\omega\hat{u}(\omega)$  are in  $L_2(\mathbb{R})$  in order to achieve finite window widths  $2\Delta_u$  and  $2\Delta_{u^*}$ , as defined in (4.4.3)–(4.4.4).

**Example 4.4.1** The window function

$$u(x) = \chi_{(-\frac{1}{2}, \frac{1}{2})}(x)$$

with

$$\int_{-\infty}^{\infty} u(x) dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} 1 dx = 1$$

and center  $x^* = 0$  has finite  $\Delta_u$ , but  $\Delta_{u^*} = \infty$ .

**Solution** Clearly, with

$$\|u\|_2^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} 1^2 dx = 1,$$

we have

$$x^* = \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} xu(x)^2 dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x dx = 0$$

and

$$\Delta_u^2 = \frac{1}{\|u\|_2^2} \int_{-\infty}^{\infty} x^2 u(x)^2 dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x^2 dx = \frac{1}{12}$$

is finite. On the other hand,

$$\begin{aligned} \widehat{u}(\omega) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{-i\omega x} dx = \frac{e^{-i\omega/2} - e^{i\omega/2}}{-i\omega} \\ &= \frac{\sin(\omega/2)}{\omega/2}. \end{aligned}$$

Hence,  $u^*(\omega) = \widehat{u}(-\omega) = \widehat{u}(\omega)$ , the center of the frequency window function  $u^*$  is  $\omega^* = 0$ , and the window radius is  $\Delta_{\omega^*} = \infty$ , since

$$\int_{-\infty}^{\infty} |\omega \widehat{u}(\omega)|^2 d\omega = 4 \int_{-\infty}^{\infty} \sin^2\left(\frac{\omega}{2}\right) d\omega = \infty. \quad \blacksquare$$

#### 4.4.2 Gaussian as optimal time-frequency window

The most commonly used time-window function is the Gaussian function

$$g_\alpha(x) = e^{-\alpha x^2}$$

with parameter  $\alpha > 0$ . Recall from (4.2.4) of Subunit 4.2.2 that

$$\int_{-\infty}^{\infty} g_\alpha(x) dx = \sqrt{\frac{\pi}{\alpha}}.$$

Hence, to compute its window width, we differentiate both sides of the above equation with respect to  $\alpha$ , and then set  $\alpha = 1/(2\sigma^2)$ , to yield

$$\int_{-\infty}^{\infty} x^2 e^{-2(\frac{x}{2\sigma})^2} dx = \frac{1}{2} \sqrt{\pi} \left( \frac{1}{2\sigma^2} \right)^{-3/2},$$

so that

$$\int_{-\infty}^{\infty} x^2 g_{\sigma}^2(x) dx = \frac{1}{(2\sigma\sqrt{\pi})^2} \frac{\sqrt{\pi}}{2} 2^{3/2} \sigma^3,$$

where  $g_{\sigma}(x)$  is the normalized Gaussian defined in (4.2.1). Since  $g_{\sigma}$  is an even function, the center  $x^*$  of  $g_{\sigma}$  is 0. In addition, since

$$\begin{aligned} \int_{-\infty}^{\infty} g_{\sigma}^2(x) dx &= \frac{1}{(2\sigma\sqrt{\pi})^2} \int_{-\infty}^{\infty} e^{-2(\frac{x}{2\sigma})^2} dx \\ &= \frac{1}{(2\sigma\sqrt{\pi})^2} (\sqrt{2}\sigma) \sqrt{\pi}, \end{aligned}$$

we also have

$$\begin{aligned} (\Delta_{g_{\sigma}})^2 &= \int_{-\infty}^{\infty} x^2 g_{\sigma}^2(x) dx / \int_{-\infty}^{\infty} g_{\sigma}^2(x) dx \\ &= \frac{\sqrt{\pi}}{2} 2^{3/2} \sigma^3 / (\sqrt{2}\sigma\sqrt{\pi}) = \sigma^2, \end{aligned}$$

so that the radius of the window function  $g_{\sigma}$  is

$$\Delta_{g_{\sigma}} = \sigma. \quad (4.4.3)$$

Hence, the window width of  $g_{\sigma}$  is  $2\sigma$ .

To compute the window width of the Fourier transform  $\hat{g}_{\sigma}$  of  $g_{\sigma}$ , we rewrite the Fourier transform  $\hat{g}_{\alpha}(\omega)$  in (4.2.8) as

$$\hat{g}_{\sigma}(\omega) = e^{-(\omega/2\eta)^2} \text{ with } \eta = \frac{1}{2\sigma},$$

so that  $\hat{g}_{\sigma}(\omega) = c g_{\eta}(\omega)$  for some suitable normalization constant  $c$ . This allows us to conclude, by applying the result  $\Delta_{g_{\eta}} = \eta$  in (4.4.3), that  $\Delta_{\hat{g}_{\sigma}} = \eta = \frac{1}{2\sigma}$ ;

so that the width of the window function  $\hat{g}_{\sigma}$  is  $2\Delta_{\hat{g}_{\sigma}} = \frac{1}{\sigma}$ . We summarize the above results in the following theorem.

**Theorem 4.4.1** *The radii of the window functions of the Gaussian function  $g_{\sigma}(x)$  and its Fourier transform  $\hat{g}_{\sigma}(\omega)$  are given by*

$$\Delta_{g_{\sigma}} = \sigma, \quad \Delta_{\hat{g}_{\sigma}} = \frac{1}{2\sigma}; \quad (4.4.4)$$

so that the area of the time-frequency localization window in  $\mathbb{R}^2$ , defined by

$$[-\Delta_{g\sigma}, \Delta_{g\sigma}] \times [-\Delta_{\hat{g}\sigma}, \Delta_{\hat{g}\sigma}] \quad (4.4.5)$$

is always the same value 2, independent of  $\sigma$ , namely:

$$(2\Delta_{g\sigma})(2\Delta_{\hat{g}\sigma}) = 2. \quad (4.4.6)$$

It turns out that the area = 2 is the smallest among all time-frequency localization windows  $[-\Delta_u, \Delta_u] \times [-\Delta_{\hat{u}}, \Delta_{\hat{u}}]$ , where  $u \in (L_1 \cap L_2)(\mathbb{R})$ , as asserted by the following so-called “uncertainty principle”.

**Theorem 4.4.2** *Let  $u \in (L_1 \cap L_2)(\mathbb{R})$  with Fourier transform  $\hat{u}(\omega)$ . Then*

$$\Delta_u \Delta_{\hat{u}} \geq \frac{1}{2}, \quad (4.4.7)$$

where  $\Delta_u$  or  $\Delta_{\hat{u}}$  may be infinite. Furthermore, equality in (4.4.7) holds if and only if

$$u(x) = cg_\sigma(x - b) \quad (4.4.8)$$

for any  $\sigma > 0, b \in \mathbb{R}$ , and  $c \neq 0$ .

In other words, the Gaussian function is the only time-window function that provides optimal time-frequency localization. The proof of Theorem 4.4.2 will be given in the next subunit.

#### 4.4.3 Derivation of the Uncertainty Principle

To prove Theorem 4.4.2 in the previous subunit, we may assume, without loss of generality, that the centers of  $u(x)$  and  $\hat{u}(\omega)$  are  $x^* = 0$  and  $\omega^* = 0$ , respectively. Hence,

$$(\Delta_u \Delta_{\hat{u}})^2 = \frac{1}{\|u\|_2^2 \|\hat{u}\|_2^2} \left( \int_{-\infty}^{\infty} x^2 |u(x)|^2 dx \right) \left( \int_{-\infty}^{\infty} \omega^2 |\hat{u}(\omega)|^2 d\omega \right). \quad (4.4.9)$$

In (4.4.9), we may apply Plancherel’s formula to conclude that

$$\int_{-\infty}^{\infty} \omega^2 |\hat{u}(\omega)|^2 d\omega = \|\hat{u}'\|_2^2 = 2\pi \|u'\|_2^2. \quad (4.4.10)$$

Here, we remark that if the integral on the left of (4.4.10) is finite, then both  $\hat{u}'$  and  $u'$  exist almost everywhere. In addition, in the denominator of (4.4.9), we have  $\|\hat{u}\|_2^2 = 2\pi \|u\|_2^2$ , again by the Plancherel formula. Therefore, applying



the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
 (\Delta_u \Delta_{\hat{u}})^2 &= \|u\|_2^{-4} \int_{-\infty}^{\infty} |xu(x)|^2 dx \int_{-\infty}^{\infty} |u'(x)|^2 dx \\
 &\geq \|u\|_2^{-4} \left( \int_{-\infty}^{\infty} |xu(x) \overline{u'(x)}| dx \right)^2 \\
 &\geq \|u\|_2^{-4} \left| \int_{-\infty}^{\infty} \operatorname{Re} \{xu(x) \overline{u'(x)}\} dx \right|^2.
 \end{aligned} \tag{4.4.11}$$

But since

$$\begin{aligned}
 x \frac{d}{dx} |u(x)|^2 &= x \frac{d}{dx} u(x) \overline{u(x)} \\
 &= x (u(x) \overline{u'(x)} + u'(x) \overline{u(x)}) \\
 &= 2 \operatorname{Re} \{xu(x) \overline{u'(x)}\},
 \end{aligned}$$

the right-hand side of (4.4.11) can be written as

$$\begin{aligned}
 &\|u\|_2^{-4} \left\{ \frac{1}{2} \int_{-\infty}^{\infty} \left( x \frac{d}{dx} |u(x)|^2 \right) dx \right\}^2 \\
 &= \frac{1}{4} \|u\|_2^{-4} \left\{ \left[ x |u(x)|^2 \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} |u(x)|^2 dx \right\}^2 \\
 &= \frac{1}{4} \|u\|_2^{-4} \|u\|_2^4 = \frac{1}{4},
 \end{aligned} \tag{4.4.12}$$

and hence,  $\Delta_u \Delta_{\hat{u}} \geq \frac{1}{2}$ . In both (4.4.11) and (4.4.12), we have assumed that  $u \in PC(\mathbb{R})$ . In addition, since  $u \in L_2(\mathbb{R})$  or  $|u|^2 \in L_1(\mathbb{R})$ , the function  $|u(x)|^2$  must decay to 0 faster than  $\frac{1}{x}$ , when  $|x| \rightarrow \infty$ . That (4.4.7) and (4.4.9) are valid for any  $u \in (L_1 \cap L_2)(\mathbb{R})$  follows from a standard “density” argument of

$$\operatorname{closure}_{L_2}(PC(\mathbb{R})) = L_2(\mathbb{R}).$$

Finally, if the inequality in (4.4.7) becomes equality, we recall from the derivation of the Cauchy-Schwarz inequality that

$$|xu(x)| = r|u'(x)| \tag{4.4.13}$$

for some constant  $r > 0$  and

$$\pm \operatorname{Re} xu(x) \overline{u'(x)} = |xu(x)u'(x)|. \tag{4.4.14}$$

From (4.4.13), we have

$$xu(x) = ru'(x)e^{i\theta(x)} \tag{4.4.15}$$

for some real-valued function  $\theta(x)$ . Hence, by (4.4.14) together with (4.4.15), we may conclude that

$$\pm \operatorname{Re} r |u'(x)|^2 e^{i\theta(x)} = r |u'(x)|^2.$$

Thus  $\pm \operatorname{Re}(e^{i\theta(x)}) = 1$ , which implies that  $\pm e^{i\theta(x)}$  is the constant function 1. Therefore, (4.4.15) becomes

$$\frac{u'(x)}{u(x)} = \frac{1}{r}x \quad \text{or} \quad \frac{u'(x)}{u(x)} = \frac{-1}{r}x,$$

or equivalently,

$$u(x) = \tilde{c}e^{x^2/2r} \quad \text{or} \quad u(x) = \tilde{c}e^{-x^2/2r}.$$

But since  $r > 0$  and  $u(x) \in L_1(\mathbb{R})$ ,  $u(x)$  cannot be  $\tilde{c}e^{x^2/2r}$  and must be the Gaussian function

$$u(x) = \tilde{c}e^{-\alpha^2 x^2}$$

with  $\alpha^2 = \frac{1}{2r} > 0$ . In the above argument, we have assumed that the center  $x^*$  of the time-window function  $u(x)$  is  $x^* = 0$ . Therefore, in general,  $u(x)$  can be formulated as

$$u(x) = cg_\sigma(x - b)$$

for  $\sigma = \frac{1}{2\alpha}$  and some  $c \neq 0, x^* = b \in \mathbb{R}$ . This completes the proof of the uncertainty principle.  $\blacksquare$

## 4.5 Time-Frequency Bases

Since computation of the LFT (or STFT) and its corresponding inverse, or LIFT, is very costly, particularly when the Gaussian is used as the window function, this subunit is devoted to the study of time-frequency analysis by considering sampling of the LFT  $(\mathbb{F}_{\bar{u}}f)(x, \omega)$  of  $f \in (L_1 \cap L_2)(\mathbb{R})$ , where the complex conjugate  $\overline{u(x)}$  of  $u \in (L_1 \cap L_2)(\mathbb{R})$  is used as the time-window function. Again, for the sake of low computational cost, only uniform discretization is considered. Let  $a > 0$  be the spacing between two neighboring sample points in the time-domain, and let  $b > 0$  denote the sampling scale in frequency space; that is, by sampling the frequency domain  $\mathbb{R}$  at the discrete set of points  $2\pi kb$ , for integers  $k$ . Analogous to the uncertainty principle studied in Subunit 4.4, we will study the Balian-Low restriction on the product of sampling rate  $a$  and sampling scale  $b$ , namely:  $0 < ab \leq 1$ . This is the restriction for the sampled basis-functions to constitute a complete family of functions in the space  $L_2(\mathbb{R})$ , regardless of the choice of the window functions. A detailed discussion is given in Subunit 4.5.2, in terms of the “frame” condition. We will also demonstrate

a way of getting around the Balian-Low restriction by replacing the LFT with the localized cosine transform, by introducing the notion of localized cosine basis in Subunit 4.5.3, and introduce a class of admissible window functions in Subunit 5.5.4 with the so-called Malvar wavelets as a class of demonstrative examples. .

#### 4.5.1 Balian-Low restriction

Let  $\mathbb{Z}$  denote the set of all integers. By sampling the LFT  $(\mathbb{F}_{\bar{u}}f)(x, \omega)$  of  $f$  at  $(x, \omega) = (m, 2\pi k)$ , with  $m, k \in \mathbb{Z}$ , we may formulate the discrete LFT as the inner product of the given function  $f$  with a family of basis-functions  $h_{m,k}(x)$ , namely:

$$\begin{aligned} (\mathbb{F}_{\bar{u}}f)(m, 2\pi k) &= \int_{-\infty}^{\infty} f(x) \overline{u(x-m)} e^{-i2\pi kx} dx \\ &= \int_{-\infty}^{\infty} f(x) \overline{h_{m,k}(x)} dx = \langle f, h_{m,k} \rangle \end{aligned}$$

where the basis functions  $h_{m,k}(x)$  are defined by

$$h_{m,k}(x) = u(x-m) e^{i2\pi kx}. \quad (4.5.1)$$

**Remark 4.5.1** Since  $e^{-i2\pi kx} = e^{-i2\pi k(x-m)}$ , the functions  $h_{m,k}(x)$  in (4.5.1) can be formulated as

$$h_{m,k}(x) = H_k(x-m) \quad (4.5.2)$$

with  $H_k(x)$  defined by

$$H_k(x) = u(x) e^{i2\pi kx}.$$

While  $H_k(x)$  localizes the frequency  $k \in \mathbb{Z}$  of  $f(x)$  only at the time sample point  $m = 0$ ,  $h_{m,k}(x) = H_k(x-m)$  localizes the same frequency of  $f(x)$  at any time instant  $m \in \mathbb{Z}$ . ■

**Remark 4.5.2** On the other hand, if the frequency  $k \in \mathbb{Z}$  in the definition (4.5.1) is not an integer, such as

$$h_{m,kb} = u(x-m) e^{i2\pi kb x},$$

where  $b \notin \mathbb{Z}$ , then (4.5.2) does not apply. Later, we will also consider

$$(\mathbb{F}_{\bar{u}}f)(ma, 2\pi kb) = \langle f, h_{ma,kb} \rangle$$

with  $a, b > 0$  and

$$h_{ma,kb}(x) = u(x-ma) e^{i2\pi kb x}. \quad (4.5.3)$$

■

**Example 4.5.1** Let  $u(x)$  be the characteristic function of  $[-\frac{1}{2}, \frac{1}{2})$ ; that is,

$$u(x) = \chi_{[-\frac{1}{2}, \frac{1}{2})}(x) = \begin{cases} 1, & \text{for } -\frac{1}{2} \leq x < \frac{1}{2}, \\ 0, & \text{otherwise.} \end{cases}$$

Then the family  $\{h_{m,k}(x)\}$ ,  $m, k \in \mathbb{Z}$  defined by (4.5.1) with this particular time-window function  $u(x)$ , constitutes an orthonormal basis of  $L_2(\mathbb{R})$ .

**Solution** For each  $m \in \mathbb{Z}$ ,

$$\langle h_{m,k}, h_{m,\ell} \rangle = \int_{m-\frac{1}{2}}^{m+\frac{1}{2}} e^{i2\pi(k-\ell)x} dx = \delta_{k-\ell}.$$

For all  $k, \ell \in \mathbb{Z}$  and  $m \neq n$ ,

$$\langle h_{m,k}, h_{n,\ell} \rangle = 0,$$

since the supports of  $h_{m,k}$  and  $h_{n,\ell}$  do not overlap. Hence,

$$\langle h_{m,k}, h_{n,\ell} \rangle = \delta_{m-n} \delta_{k-\ell};$$

or  $\{h_{m,k}(x)\}$ ,  $m, k \in \mathbb{Z}$ , is an orthonormal family.

In addition, for any  $f \in L_2(\mathbb{R})$ , by setting  $f_m(x) = f(x+m)$ , where  $-\frac{1}{2} \leq x \leq \frac{1}{2}$ , namely:

$$f_m(x) = u(x)f(x+m) = \chi_{[-\frac{1}{2}, \frac{1}{2})}(x)f(x+m), \quad -\frac{1}{2} \leq x \leq \frac{1}{2},$$

and extending it to  $\mathbb{R}$  periodically such that  $f_m(x+1) = f_m(x)$ , then for each fixed  $m \in \mathbb{Z}$ , the Fourier series

$$(Sf_m)(x) = \sum_{k=-\infty}^{\infty} a_k(f_m) e^{i2\pi kx}$$

of  $f_m(x)$  converges to  $f_m(x)$  in  $L_2[-\frac{1}{2}, \frac{1}{2}]$ , where

$$\begin{aligned} a_k(f_m) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} f_m(t) e^{-i2\pi kt} dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(t+m) e^{-i2\pi kt} dt \\ &= \int_{m-\frac{1}{2}}^{m+\frac{1}{2}} f(y) e^{-i2\pi k(y-m)} dy \\ &= \int_{-\infty}^{\infty} f(t)u(t-m) e^{-i2\pi kt} dt = \langle f, h_{m,k} \rangle. \end{aligned}$$

Thus, for each  $m \in \mathbb{Z}$ , we have

$$\begin{aligned} f(x)u(x-m) &= f_m(x-m)u(x-m) = ((Sf_m)(x-m))u(x-m) \\ &= \sum_{k=-\infty}^{\infty} a_k(f_m) e^{i2\pi k(x-m)} u(x-m) = \sum_{k=-\infty}^{\infty} \langle f, h_{m,k} \rangle h_{m,k}(x). \end{aligned}$$

Hence, summing both sides over all  $m \in \mathbb{Z}$  yields

$$\sum_{m=-\infty}^{\infty} f(x)u(x-m) = \sum_{m=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \langle f, h_{m,k} \rangle h_{m,k}(x),$$

so that

$$\begin{aligned} f(x) &= \sum_{m=-\infty}^{\infty} f(x)\chi_{[m-\frac{1}{2}, m+\frac{1}{2})}(x) = \sum_{m=-\infty}^{\infty} f(x)\chi_{[-\frac{1}{2}, \frac{1}{2})}(x-m) \\ &= \sum_{m=-\infty}^{\infty} f(x)u(x-m) = \sum_{m=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \langle f, h_{m,k} \rangle h_{m,k}(x), \end{aligned}$$

where the convergence is in the  $L_2(\mathbb{R})$ -norm.  $\blacksquare$

**Remark 4.5.3** The limitation of the window function  $u(x) = \chi_{[-\frac{1}{2}, \frac{1}{2})}(x)$  in the above example is that it provides very poor frequency localization. Unfortunately, the formulation of  $h_{m,k}(x)$  in (4.5.1) cannot be improved by too much, as dictated by the so-called “Balian-Low” restriction to be stated in Theorem 4.5.1 in the next Subunit 4.5.2  $\blacksquare$

However, before we could fully understand the Balian-Low restriction, we need to recall the notion of completeness, as studied in Subunit 3.4. In the next subunit, we will see that the completeness property is guaranteed by any family of functions in  $L_2(\mathbb{R})$  that constitute a frame of  $L_2(\mathbb{R})$ .

## 4.5.2 Frames

**Definition 4.5.1** A family of functions  $\{h_\alpha(x)\}$  in  $L_2(\mathbb{R})$ ,  $\alpha \in J$ , where  $J$  denotes an infinite index set, such as  $\mathbb{Z}$  and  $\mathbb{Z}^2$ , is called a frame of  $L_2(\mathbb{R})$ , if there exist some constants  $A$  and  $B$ , with  $0 < A \leq B < \infty$ , such that

$$A\|f\|_2^2 \leq \sum_{\alpha \in J} |\langle f, h_\alpha \rangle|^2 \leq B\|f\|_2^2, \quad (4.5.4)$$

for all  $f \in L_2(\mathbb{R})$ . Here,  $A$  and  $B$  are called frame bounds. Furthermore, if  $A$  and  $B$  can be so chosen that  $A = B$  in (4.5.4), then  $\{h_\alpha\}$  is called a tight frame.

This definition of a complete family extends the notion of complete orthonormal family studied in Subunit 3.4.2 and Subunit 3.5.1.

**Remark 4.5.4** If  $\{h_\alpha\}, \alpha \in J$ , is a tight frame with frame bound  $A = B$ , then the family  $\{\tilde{h}_\alpha(x)\}$ , defined by

$$\tilde{h}_\alpha(x) = \frac{1}{\sqrt{A}} h_\alpha(x), \quad \alpha \in J,$$

satisfies Parseval's identity:

$$\|f\|_2^2 = \sum_{\alpha \in J} |\langle f, \tilde{h}_\alpha \rangle|^2, \quad f \in L_2(\mathbb{R}). \quad (4.5.5)$$

■

Recall that an orthonormal basis of  $L_2(\mathbb{R})$  also satisfies Parseval's identity (4.5.5). To understand the identity (4.5.5) for the tight frame  $\{\tilde{h}_\alpha\}$ , let us consider the function  $f(x) = \tilde{h}_{\alpha_0}(x)$  for any fixed index  $\alpha_0 \in J$ , so that

$$\begin{aligned} \|\tilde{h}_{\alpha_0}\|_2^2 &= \sum_{\alpha \in J} |\langle \tilde{h}_{\alpha_0}, \tilde{h}_\alpha \rangle|^2 \\ &= \|\tilde{h}_{\alpha_0}\|_2^4 + \sum_{\alpha \neq \alpha_0} |\langle \tilde{h}_{\alpha_0}, \tilde{h}_\alpha \rangle|^2 \end{aligned}$$

or

$$\|\tilde{h}_{\alpha_0}\|_2^2 \left(1 - \|\tilde{h}_{\alpha_0}\|_2^2\right) = \sum_{\alpha \neq \alpha_0} |\langle \tilde{h}_{\alpha_0}, \tilde{h}_\alpha \rangle|^2.$$

Since the right-hand side is non-negative, we see that

$$\|\tilde{h}_{\alpha_0}\|_2^2 \leq 1.$$

In addition, if  $\|\tilde{h}_{\alpha_0}\|_2 = 1$ , then the right-hand side vanishes, or  $\langle \tilde{h}_{\alpha_0}, \tilde{h}_\alpha \rangle = 0$  for all  $\alpha \neq \alpha_0$ . Thus if  $\|\tilde{h}_\alpha\|_2 = 1$  for each  $\alpha \in J$ , then  $\{\tilde{h}_\alpha\}_{\alpha \in J}$  is an orthonormal family.

Furthermore, observe that any frame  $\{h_\alpha\}$  of  $L_2(\mathbb{R})$  is a complete family in  $L_2(\mathbb{R})$ . To prove this claim, assume, on the contrary, that  $\{h_\alpha\}$  is not complete in  $L_2(\mathbb{R})$ . Then there would exist some non-trivial  $f \in L_2(\mathbb{R})$  which is orthogonal to all  $h_\alpha$ . This violates the lower-bound frame condition, in that

$$0 < A\|f\|_2^2 \leq \sum_{\alpha \in J} |\langle f, h_\alpha \rangle|^2 = 0.$$

Therefore, we have shown that any frame of  $L_2(\mathbb{R})$  is a complete family in  $L_2(\mathbb{R})$ .

Let us summarize the above derivations, applied to tight frames, in the following theorem.

**Theorem 4.5.1** *Let  $\{h_\alpha\}, \alpha \in J$ , be a tight frame of  $L_2(\mathbb{R})$  with frame bound  $A = B = 1$ . Then*

- (a)  $\|h_\alpha\|_2 \leq 1$  for all  $\alpha \in J$ ;
- (b) the  $L_2(\mathbb{R})$  closure of  $\text{span}\{h_\alpha : \alpha \in J\}$  is the entire  $L_2(\mathbb{R})$  space;
- (c) if  $\|h_\alpha\|_2 = 1$  for all  $\alpha \in J$ , then  $\{h_\alpha\}$  is an orthonormal basis of  $L_2(\mathbb{R})$ .

We now return to the study of time-frequency analysis by stating the following Balian-Low restriction, a concept introduced in the Subunit 4.5.1.

**Theorem 4.5.2** *Let  $\{h_{m,k}(x)\}, (m,k) \in \mathbb{Z}^2$ , be defined by (4.5.1) with window function  $u(x) \in (L_1 \cap L_2)(\mathbb{R})$ . Then a necessary condition for  $\{h_{m,k}(x)\}$  to be a frame of  $L_2(\mathbb{R})$  is that at least one of the two integrals*

$$\int_{-\infty}^{\infty} |xu(x)|^2 dx \quad \text{and} \quad \int_{-\infty}^{\infty} |\omega \hat{u}(\omega)|^2 d\omega$$

*is equal to  $\infty$ .*

**Remark 4.5.5** Since the Fourier transform of the derivative of a function is  $i\omega$  multiple of the Fourier transform of the function, it follows from the Plancherel formula that

$$\begin{aligned} \int_{-\infty}^{\infty} |\omega \hat{f}(\omega)|^2 d\omega &= \int_{-\infty}^{\infty} |(\mathbb{F}f')(\omega)|^2 d\omega \\ &= 2\pi \int_{-\infty}^{\infty} |f'(x)|^2 dx. \end{aligned}$$

Hence, if  $u(x)$  is the window function in  $\{h_{m,k}(x)\}$  with finite window width (that is,  $\Delta_u < \infty$ ), then for  $\{h_{m,k}(x)\}$  to be a frame of  $L_2(\mathbb{R})$ , it is necessary that

$$\int_{-\infty}^{\infty} |u'(x)|^2 dx = \infty$$

according to the Balian-Low restriction in Theorem 4.5.2. Consequently, any continuous differentiable function  $u(x)$ , that vanishes outside a finite interval, cannot be used as the window function  $u(x)$  in (4.5.1) to achieve good time-frequency localization. ■

Observe that the family in (4.5.1) is obtained from the family in (4.5.3) by setting  $a = b = 1$ , so that  $ab = 1$ . Hence, in view of the above remark, in order to use a smooth function with finite support as the time-window to achieve good time-frequency localization, the only chance is to choose  $a$  and  $b$  in (4.5.3) with  $ab < 1$ , and this is indeed assured by the following theorem.

**Theorem 4.5.3** *Let  $a, b > 0$  and  $\{h_{ma,kb}(x)\}, (m, k) \in \mathbb{Z}^2$ , be defined by (4.5.3) with window function  $u \in (L_1 \cap L_2)(\mathbb{R})$ . Then the following statements hold.*

- (a) *For  $ab > 1$ , there does not exist any window function  $u(x)$  for which the family  $\{h_{ma,kb}(x)\}, (m, k) \in \mathbb{Z}^2$ , is complete in  $L_2(\mathbb{R})$ .*
- (b) *For  $ab = 1$  (such as the family of functions in (4.5.1)), there exists  $u(x) \in (L_1 \cap L_2)(\mathbb{R})$  such that  $\{h_{ma,kb}(x)\}$  is a frame (such as an orthonormal basis in Example 4.5.1), but the time-frequency window*

$$[-\Delta_u, \Delta_u] \times [-\Delta_{\hat{u}}, \Delta_{\hat{u}}]$$

*necessarily has infinite area, namely:*

$$\Delta_u \Delta_{\hat{u}} = \infty.$$

- (c) *For  $0 < ab < 1$ , there exists  $u \in (L_1 \cap L_2)(\mathbb{R})$  such that  $\Delta_u \Delta_{\hat{u}} < \infty$  and the corresponding family  $\{h_{ma,kb}(x)\}, (m, k) \in \mathbb{Z}^2$ , is a tight frame of  $L_2(\mathbb{R})$ .*

**Remark 4.5.6** Let  $a$  and  $b$  in (4.5.3) be restricted by  $0 < ab < 1$  to achieve good time-frequency localization, as guaranteed by Theorem 4.5.3(c). Then the frame  $\{h_{ma,kb}(x)\}$  of  $L_2(\mathbb{R})$ , with window function  $u(x)$  (that satisfies  $\Delta_u \Delta_{\hat{u}} < \infty$ ) cannot be formulated as the translation (by  $ma, m \in \mathbb{Z}$ ) of some localized function

$$H_{kb}(x) = u(x)e^{i2\pi kbx}$$

as in (4.5.2) for  $h_{m,k}(x)$ , where  $a = b = 1$ . Indeed, for  $0 < ab < 1$ , computation of  $H_{kb}(x - ma)$  requires additional computation of the phase modulation:

$$A_{ab}(km) = e^{-i(2\pi km)ab},$$

which is no longer equal to 1 in general (for  $k, m \in \mathbb{Z}$ ). The reason is that

$$H_{kb}(x - ma) = \left( u(x - ma)e^{i2\pi kbx} \right) A_{ab}(km).$$

Consequently, the computational aspect of time-frequency analysis is much less effective. ■



### 4.5.3 Localized cosine basis

The good news is that by replacing  $e^{i2\pi kx}$  with the sine and/or cosine functions to formulate a frequency basis, such as

$$c_k(x) = \sqrt{2} \cos(k + \frac{1}{2})\pi x, \quad k = 0, 1, \dots, \quad (4.5.6)$$

as the orthonormal basis of  $L_2[0, 1]$ , then localization by window functions  $u(x)$  with finite time-frequency windows (that is,  $\Delta_u \Delta_{\hat{u}} < \infty$ ) is feasible, even by formulating the local basis functions  $h_{mn}^c(x)$  as integer translates, such as

$$h_{m,k}^c(x) = H_k^c(x - m) = u(x - m) c_k(x - m), \quad m, k \in \mathbb{Z},$$

of the localized frequency basis

$$H_k^c(x) = u(x) c_k(x), \quad k \in \mathbb{Z}.$$

In other words, it is possible to get around the Balian-Low restriction by replacing  $e^{i2\pi kx}$  by sine and/or cosine functions. However, to accomplish this goal, we must spend some effort to construct the localization window functions  $u$ . This topic will be studied in the next subunit, where we give a precise formulation of the commonly used localized cosine basis functions, the so-called Malvar wavelets.

### 4.5.4 Malvar wavelets

In this subunit, we are only concerned with window functions  $u(x)$  that satisfy the following admissible conditions.

**Definition 4.5.2** *A function  $u(x)$  is said to be an admissible window function, if it satisfies the following:*

- (i) *there exists some positive number  $\delta$ , with  $0 < \delta < \frac{1}{2}$ , such that*

$$u(x) = 1, \text{ for } x \in [\delta, 1 - \delta],$$

*and*

$$u(x) = 0, \text{ for } x \notin [-\delta, 1 + \delta];$$

- (ii)  $0 \leq u(x) \leq 1$ ;

- (iii)  $u(x)$  is symmetric with respect to  $x = \frac{1}{2}$ ; that is,:

$$u(\frac{1}{2} - x) = u(\frac{1}{2} + x), \text{ for all } x;$$

- (iv) both  $u(x)$  and  $u'(x)$  are at least piecewise continuous on  $[-\delta, 1 + \delta]$ ; and
- (v)  $u^2(x) + u^2(-x) = 1$ , for  $x \in [-\delta, \delta]$ .

**Remark 4.5.7** It follows from (i), (ii), and (iv) that any admissible window function  $u$  satisfies:

$$\int_{-\infty}^{\infty} x^2 u^2(x) dx < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} |u'(x)|^2 dx < \infty,$$

so that by the Plancherel formula, we have

$$\int_{-\infty}^{\infty} \omega^2 |\widehat{u}(\omega)|^2 d\omega = \int_{-\infty}^{\infty} |(\mathbb{F}u')|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} |u'(x)|^2 dx < \infty.$$

Hence,  $u(x)$  provides good time-frequency localization, meaning that  $\Delta_u \Delta_{\widehat{u}} < \infty$ . Furthermore, conditions (i), (iii) and (v) for an admissible window function also imply that

$$\sum_{m=-\infty}^{\infty} u^2(x - m) = 1, \quad \text{for all } x \in \mathbb{R}. \quad (4.5.7)$$

■

**Example 4.5.2** Let  $0 < \delta < \frac{1}{2}$ . Then the function  $u(x)$  defined by

$$u(x) = \begin{cases} 0, & x < -\delta \text{ or } x > 1 + \delta; \\ \frac{1}{\sqrt{2}} \left(1 + \sin \frac{\pi}{2\delta} x\right), & -\delta \leq x < 0; \\ \sqrt{1 - \frac{1}{2} \left(1 - \sin \frac{\pi}{2\delta} x\right)^2}, & 0 \leq x < \delta; \\ 1, & \delta \leq x \leq 1 - \delta; \\ \sqrt{1 - \frac{1}{2} \left(1 - \sin \frac{\pi}{2\delta} (1 - x)\right)^2}, & 1 - \delta < x \leq 1; \\ \frac{1}{\sqrt{2}} \left(1 + \sin \frac{\pi}{2\delta} (1 - x)\right), & 1 < x \leq 1 + \delta, \end{cases} \quad (4.5.8)$$

is an admissible window function.

**Solution** Since verification of (i), (ii), (iii) and (v) for  $u(x)$  is straightforward, we only verify that  $u'(x)$  exists for any  $x \in \mathbb{R}$ , and this is reduced to the points  $x_0 = -\delta, 0, \delta, 1 - \delta, 1, 1 + \delta$ . Since it is also clear that  $u(x)$  is continuous, to verify that  $u'(x_0)$  exists, it is enough to show that the left-hand and right-hand limits of  $u'(x)$  at  $x_0$  exist and are equal.

For  $x_0 = -\delta$ , clearly

$$\lim_{x \rightarrow -\delta^-} u'(x) = \lim_{x \rightarrow -\delta^-} 0 = 0.$$

On the other hand,

$$\lim_{x \rightarrow -\delta^+} u'(x) = \lim_{x \rightarrow -\delta^+} \frac{1}{\sqrt{2}} \cos\left(\frac{\pi}{2\delta}x\right) \frac{\pi}{2\delta} = \frac{1}{\sqrt{2}} \frac{\pi}{2\delta} \cos\left(-\frac{\pi}{2}\right) = 0.$$

Thus  $u'(-\delta)$  exists.

For  $x_0 = 0$ , we have

$$\begin{aligned} \lim_{x \rightarrow 0^-} u'(x) &= \lim_{x \rightarrow 0^-} \frac{1}{\sqrt{2}} \cos\left(\frac{\pi}{2\delta}x\right) \frac{\pi}{2\delta} = \frac{1}{\sqrt{2}} \frac{\pi}{2\delta} \cos 0 = \frac{\pi}{2\sqrt{2}\delta}; \\ \lim_{x \rightarrow 0^+} u'(x) &= \lim_{x \rightarrow 0^+} \frac{\pi}{4\delta} \left(1 - \frac{1}{2} \left(1 - \sin \frac{\pi}{2\delta}x\right)^2\right)^{-\frac{1}{2}} \left(1 - \sin \frac{\pi}{2\delta}x\right) \cos \frac{\pi}{2\delta}x \\ &= \frac{\pi}{4\delta} \left(\frac{1}{2}\right)^{-\frac{1}{2}} = \frac{\pi}{2\sqrt{2}\delta}. \end{aligned}$$

Therefore  $u'(0)$  exists. Verification of the existence of  $u'(x)$  at  $x_0 = \delta$  is similar. Finally, the existence of  $u'(x)$  at  $x_0 = 1 - \delta, 1, 1 + \delta$  follows from the symmetry of  $u(x)$ . In fact, both  $u(x)$  and  $u'(x)$  are continuous for all  $x$ . ■

We end this subunit by formulating the so-called Malvar wavelets, as follows.

**Theorem 4.5.4** *Let  $u(x)$  be an admissible window function that satisfies the conditions (i)–(v) in Definition 4.5.3. Then  $u(x)$  has the property*

$$\Delta_u \Delta_{\hat{u}} < \infty, \quad (4.5.9)$$

and the family  $\{\psi_{m,k}(x)\}$  of functions defined by

$$\begin{aligned} \psi_{m,k}(x) &= u(x - m) c_k(x - m) \\ &= \sqrt{2} u(x - m) \cos\left((k + \tfrac{1}{2})\pi(x - m)\right), \quad m \in \mathbb{Z}, k \geq 0, \end{aligned}$$

where  $c_k(x)$  is defined in (4.5.6), constitutes an orthonormal basis of  $L_2(\mathbb{R})$ .



# Unit 5

## PDE METHODS

When the variance  $\sigma^2$  of the Gaussian convolution filter is replaced by  $ct$ , where  $c$  is a fixed positive constant and  $t$  is used as the time parameter, then the convolution filtering of any input function  $f$  describes the heat diffusion process with initial temperature given by  $f(x)$  at the spatial position  $x \in \mathbb{R}$ . More precisely, if the function  $u(x, t)$  of two variables is used to represent the heat content (or temperature) at the position  $x$  and time  $t > 0$ , then  $u(x, t)$ , obtained by the Gaussian convolution of the given function  $f$ , is the solution of the heat diffusion PDE with initial condition  $u(x, 0) = f(x)$ , where the positive constant  $c$  is called the heat conductivity constant. However, this elegant example has little practical value, because the spatial domain is the entire  $x$ -axis, but it serves the purpose as a convincing motivation for the study of linear PDE methods, to be studied in this unit. To solve the same heat diffusion PDE as described in this example, but with initial heat source  $f$  given on a bounded interval instead, and with perfect insulation at the two end-points to avoid any heat loss, the method of “separation of variables is introduced in this unit. This method is applied to separate the PDE into two ordinary differential equations (ODE’s) that can be easily solved by appealing to the eigenvalue problem, studied in Subunit 1.2, for linear differential operators, with eigenfunctions given by the cosine function in  $x$ , and with frequency governed by the eigenvalues, which also dictate the rate of exponential decay in the time variable  $t$ . Superposition of the product of these corresponding eigenfunctions with coefficients given by the Fourier coefficients of the Fourier series representation of the initial heat content, studied in Unit 3, solves this heat equation. Extension of the method of separation of variables to the study of boundary value problems on a bounded rectangular domain in  $\mathbb{R}^s$  for any  $s > 2$  as well as the solution of other popular linear PDE’s, is also studied in this unit. In addition, the linear diffusion PDE, called isotropic diffusion, is generalized to a non-linear PDE that describes anisotropic diffusion; and the solution in terms of eigenvalue problems is studied, by introducing the notion of lagged anisotropic diffusion. Application of the diffusion process is applied, in Subunit 5.5, to image noise reduction that facilitates the efficiency of digital image and video compression, as studied in Unit 2.

## 5.1 From Gaussian Convolution to Diffusion Process

As an application of the Gaussian function, the solution of the (heat) diffusion partial differential equation (PDE)

$$\begin{cases} \frac{\partial}{\partial t} u(\mathbf{x}, t) = c \nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^s. \end{cases} \quad (5.1.1)$$

in the Euclidean space  $\mathbb{R}^s$ , for any dimension  $s \geq 1$ , is studied this subunit. Here and throughout,  $c$  is a positive constant, called heat diffusion conductivity.

Our mathematical tool is the convolution operation with the Gaussian function

$$g_\sigma(\mathbf{x}) = g_\sigma(x_1) \cdots g_\sigma(x_s), \quad (5.1.2)$$

where  $\mathbf{x} = (x_1, \dots, x_s) \in \mathbb{R}^s$ . We will first consider the one spatial-dimension with initial heat source given by the delta function, and then apply it to consider any initial heat source. Finally, we extend the one-dimensional result to an arbitrarily high spatial dimensional space  $\mathbb{R}^s$ .

### 5.1.1 Gaussian as solution for delta heat source

For the one spatial-dimensional space, the Laplacian  $\nabla^2$  in (5.1.1) becomes the second partial derivative with respect to the spatial variable  $x$ . Recall that for any constant  $\sigma > 0$ , the Gaussian function  $g_\sigma(x)$ , defined by

$$g_\sigma(x) = \frac{1}{2\sigma\sqrt{\pi}} e^{-(\frac{x}{2\sigma})^2}$$

as in (4.2.1) of Subunit 4.2.2, satisfies

$$\int_{-\infty}^{\infty} g_\sigma(x) dx = 1, \text{ all } \sigma > 0$$

(see (4.2.2)), and its Fourier transform is given by

$$\widehat{g}_\sigma(\omega) = e^{-\sigma^2 \omega^2}, \quad (5.1.3)$$

as shown in (4.2.8) of Theorem 4.2.1. Again, let  $c > 0$  denote the heat diffusion conductivity constant in the PDE (5.1.1). We introduce the “time” parameter

$$t = \frac{\sigma^2}{c} \text{ or } \sigma^2 = ct \quad (5.1.4)$$

to define the time-spatial Gaussian function  $G(x, t)$  (of two variables) as follows:

$$G(x, t) = g_\sigma(x) = g_{\sqrt{ct}}(x) = \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}}, \quad (5.1.5)$$

with  $x \in \mathbb{R}$  to be called the spatial variable, and  $t \geq 0$  to be called the time variable. Then  $G(x, t)$  satisfies the PDE (5.1.1) in that

$$\frac{\partial}{\partial t}G(x, t) = c \frac{\partial^2}{\partial x^2}G(x, t), \quad x \in \mathbb{R}, t > 0. \quad (5.1.6)$$

Indeed, by taking the first partial derivatives of  $G(x, t)$  in (5.1.5), we have

$$\begin{aligned} \frac{\partial}{\partial t}G(x, t) &= \frac{1}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}} \left\{ -\frac{1}{2}t^{-\frac{3}{2}} + t^{-\frac{1}{2}}\left(\frac{x^2}{4c}\right)t^{-2} \right\}; \\ \frac{\partial}{\partial x}G(x, t) &= \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}} \left\{ -\frac{t^{-1}}{2c}x \right\}, \end{aligned}$$

so that the second spatial partial derivative is given by

$$\begin{aligned} \frac{\partial^2}{\partial x^2}G(x, t) &= \frac{t^{-\frac{1}{2}}}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}} \left\{ -\frac{t^{-1}}{2c} + \left(-\frac{t^{-1}}{2c}x\right)^2 \right\} \\ &= \frac{1}{c} \frac{1}{2\sqrt{\pi c}} e^{-\frac{x^2}{4c}t^{-1}} \left\{ -\frac{1}{2}t^{-\frac{3}{2}} + \left(\frac{x^2}{4c}\right)t^{-\frac{1}{2}} \cdot t^{-2} \right\} \\ &= \frac{1}{c} \frac{\partial}{\partial t}G(x, t), \end{aligned}$$

which proves that  $G(x, t)$  satisfies the PDE in (5.1.6). To study the initial condition  $u(x, 0)$  in (5.1.1), observe that  $g_\sigma(x) \rightarrow 0$  as  $\sigma \rightarrow 0$ , for all  $x \neq 0$ . But since the integral of  $g_\sigma(x)$  over  $(-\infty, \infty)$  is 1 for all  $\sigma$ ,  $g_\sigma(x)$  does not converge to the zero function, but instead to the delta distribution. Hence,

$$G(x, 0) = \delta(x), \quad x \in \mathbb{R},$$

where  $\delta(x)$  denotes the “Dirac delta” distribution (also commonly called the “delta function”). In other words, the time-spatial Gaussian function  $G(x, t) = g_\sigma(x) = g_{\sqrt{ct}}(x)$  is the solution of the heat diffusion partial differential equation with unit point-heat source at  $x = 0$ .

### 5.1.2 Gaussian convolution as solution of heat equation for the real-line

Next, we consider the one-dimensional heat diffusion process again, but with arbitrary initial heat content  $u_0(x)$ ,  $-\infty < x < \infty$ .

Recall that if  $f \in L_2(\mathbb{R})$  is continuous at  $x$ , then  $f(x)$  is “reproduced” by convolution with the delta function, meaning that

$$f * \delta(x) = f(x),$$

or more precisely,

$$\lim_{t \rightarrow 0^+} \int_{-\infty}^{\infty} f(y)G(x-y, t) dy = f(x), \quad x \in \mathbb{R}. \quad (5.1.7)$$

**Remark 5.1.1** Although we usually assume that  $f \in L_\infty(\mathbb{R})$  for (5.1.7) to hold, yet it is only for simplicity. In fact, since for any fixed  $t > 0$ ,  $f(y)G(x-y, t)$  is integrable for all  $f \in PC(\mathbb{R})$  with “at most polynomial growth”, meaning that

$$f(x)x^{-n} \in L_\infty(\mathbb{R})$$

for some integer  $n > 0$ , (5.1.7) is valid for all  $f \in PC(\mathbb{R})$  with at most polynomial growth. ■

In other words, we have the following result.

**Theorem 5.1.1** *Let  $u_0 \in PC(\mathbb{R})$  with at most polynomial growth. Then the solution of the initial value PDE*

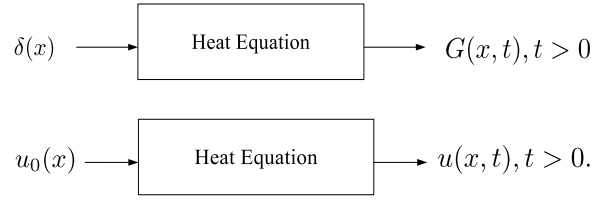
$$\begin{cases} \frac{\partial}{\partial t} u(x, t) = c \frac{\partial^2}{\partial x^2} u(x, t), & x \in \mathbb{R}, t \geq 0, \\ u(x, 0) = u_0(x), & x \in \mathbb{R}, \end{cases} \quad (5.1.8)$$

*is given by*

$$u(x, t) = \int_{-\infty}^{\infty} u_0(y)G(x-y, t) dy. \quad (5.1.9)$$

The proof of the claim that  $u(x, t)$  in (5.1.9) satisfies the initial condition in (5.1.8) has already been discussed in (5.1.7) with  $f(x) = u_0(x)$  and Remark 5.1.1. The idea is that if the heat source (i.e. initial heat content) is the delta function  $\delta(x)$ , then the heat distribution for  $t > 0$  is the Gaussian function  $G(x, t)$ , as shown at the top of Fig. 5.1.





**FIGURE 5.1:** *Diffusion with delta heat source (top) and arbitrary heat source (bottom)*

To show that the function  $u(x, t)$  defined in (5.1.9) is the solution of the initial PDE in (5.1.8), we simply apply (5.1.6) to obtain

$$\begin{aligned}
 \frac{\partial}{\partial t} u(x, t) &= \frac{\partial}{\partial t} \int_{-\infty}^{\infty} u_0(y) G(x - y, t) dy \\
 &= \int_{-\infty}^{\infty} u_0(y) \frac{\partial}{\partial t} G(x - y, t) dy \\
 &= \int_{-\infty}^{\infty} u_0(y) \left\{ c \frac{\partial^2}{\partial x^2} G(x - y, t) \right\} dy \\
 &= c \frac{\partial^2}{\partial x^2} \int_{-\infty}^{\infty} u_0(y) G(x - y, t) dy \\
 &= c \frac{\partial^2}{\partial x^2} u(x, t).
 \end{aligned}$$

Hence,  $u(x, t)$  as defined by (5.1.9) is the solution of the heat diffusion PDE (5.1.8) with heat source  $u_0(x)$ . That is,  $u(x, t)$  is the output as shown in the bottom of Fig. 5.1, with input  $u_0(x)$ . ■

**Example 5.1.1** Compute the solution  $u(x, t)$  of the initial value (heat diffusion) PDE in (5.1.8), with initial (or input) function

$$u_0(x) = a_\alpha \cos \alpha x + b_\alpha \sin \alpha x,$$

where  $\alpha$  is any real number and  $a_\alpha, b_\alpha$  are arbitrary constants.

**Solution** By Theorem 5.1.1, the solution is obtained by computing the convolution of  $u_0(x)$  with the Gaussian function  $G(x, t)$  with respect to the spatial variable  $x$ , where  $t \geq 0$  is fixed while the convolution operation is performed. Although there is no need to consider the following three separate cases, we will do so in this first example to show the computational steps more transparently.

**Case 1.** For  $\alpha = 0$ , the input function  $u_0(x) = a_0$  is a constant. Hence,

$u_0(x - y) = a_0$  and

$$u(x, t) = (u_0 * G(\cdot, t))(x) = \int_{-\infty}^{\infty} a_0 g_{\sigma}(y) dy = a_0,$$

since the Gaussian  $g_{\sigma}(x)$  is normalized with integral over  $(-\infty, \infty)$  equal to 1, as shown in (4.2.2) of Subunit 4.2.

**Case 2.** For  $b_{\alpha} = 0$ , the initial function is

$$u_0(x) = a_{\alpha} \cos \alpha x = \frac{a_{\alpha}}{2} (e^{i\alpha x} + e^{-i\alpha x}).$$

Hence, for fixed  $\sigma^2 = ct$ , the convolution becomes

$$\begin{aligned} u(x, t) &= (u_0 * g_{\sigma})(x) \\ &= \frac{a_{\alpha}}{2} \left( \int_{-\infty}^{\infty} e^{i\alpha(x-y)} g_{\sigma}(y) dy + \int_{-\infty}^{\infty} e^{-i\alpha(x-y)} g_{\sigma}(y) dy \right) \\ &= \frac{a_{\alpha}}{2} \left( e^{i\alpha x} \widehat{g_{\sigma}}(\alpha) + e^{-i\alpha x} \widehat{g_{\sigma}}(-\alpha) \right) \\ &= \frac{a_{\alpha}}{2} \left( e^{i\alpha x} e^{-\sigma^2 \alpha^2} + e^{-i\alpha x} e^{-\sigma^2 (-\alpha)^2} \right) \\ &= \frac{a_{\alpha}}{2} \left( e^{i\alpha x} + e^{-i\alpha x} \right) e^{-\sigma^2 \alpha^2} \end{aligned}$$

by (5.1.3). Since  $\sigma^2 = ct$ , we have

$$u(x, t) = a_{\alpha} e^{-c\alpha^2 t} \cos \alpha x.$$

**Case 3.** For  $a_{\alpha} = 0$ , the initial function is

$$u_0(x) = b_{\alpha} \sin \alpha x = \frac{b_{\alpha}}{2i} (e^{i\alpha x} - e^{-i\alpha x}).$$

Therefore, the same computation as above yields

$$\begin{aligned} u(x, t) &= \frac{b_{\alpha}}{2i} (e^{i\alpha x} - e^{-i\alpha x}) e^{-\sigma^2 \alpha^2} \\ &= b_{\alpha} e^{-c\alpha^2 t} \sin \alpha x, \end{aligned}$$

since  $\sigma^2 = ct$ .

Combining the above computational results, we obtain the solution of the initial value PDE (5.1.8):

$$u(x, t) = e^{-c\alpha^2 t} (a_{\alpha} \cos \alpha x + b_{\alpha} \sin \alpha x) = e^{-c\alpha^2 t} u_0(x)$$

for all  $x \in \mathbb{R}$  and  $t \geq 0$ . ■

In general, if the initial (input) function  $u_0(x)$  is a  $2\pi$ -periodic piecewise continuous function, the same computational steps apply to yield the solution  $u(x, t)$  of the initial value PDE (5.1.8), as follows.

**Example 5.1.2** Let  $u_0 \in PC[-M, M]$ ,  $M > 0$ , such that the partial sums  $(S_n u_0)(x)$  of the Fourier series of  $u_0(x)$  are uniformly bounded. Show that the solution  $u(x, t)$  of the initial value PDE (5.1.8) with initial heat content  $u_0(x)$  is given by

$$u(x, t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-c(\frac{k\pi}{M})^2 t} \left( a_k \cos \frac{k\pi}{M} x + b_k \sin \frac{k\pi}{M} x \right), \quad (5.1.10)$$

where

$$a_k = \frac{1}{M} \int_{-M}^M u_0(x) \cos \frac{k\pi}{M} x \, dx, \quad k = 0, 1, 2, \dots$$

$$b_k = \frac{1}{M} \int_{-M}^M u_0(x) \sin \frac{k\pi}{M} x \, dx, \quad k = 1, 2, \dots$$

**Solution** Consider the Fourier cosine and sine series expansion of  $u_0(x)$  in Theorem 3.1.1 of Subunit 3.1.1, with  $d = M$ . Since  $(S_n u_0)(x)$  are uniformly bounded, we may apply Lebesgue's dominated convergence theorem to interchange summation and integration, namely:

$$\begin{aligned} u(x, t) &= (u_0 * G(\cdot, t))(x) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left( \frac{a_k - ib_k}{2} \int_{-\infty}^{\infty} e^{i \frac{k\pi}{M}(x-y)} G(y, t) \, dy \right. \\ &\quad \left. + \frac{a_k + ib_k}{2} \int_{-\infty}^{\infty} e^{-i \frac{k\pi}{M}(x-y)} G(y, t) \, dt \right) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left( \frac{a_k - ib_k}{2} e^{i \frac{k\pi}{M} x} \hat{g}_{\sigma}(k) + \frac{a_k + ib_k}{2} e^{-i \frac{k\pi}{M} x} \hat{g}_{\sigma}(-k) \right) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} \left( a_k \frac{e^{i \frac{k\pi}{M} x} + e^{-i \frac{k\pi}{M} x}}{2} + b_k \frac{e^{i \frac{k\pi}{M} x} - e^{-i \frac{k\pi}{M} x}}{-2i} \right) \hat{g}_{\sigma}(k) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-\sigma^2 (\frac{k\pi}{M})^2} \left( a_k \cos \frac{k\pi}{M} x + b_k \sin \frac{k\pi}{M} x \right) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty} e^{-(\frac{k\pi}{M})^2 ct} \left( a_k \cos \frac{k\pi}{M} x + b_k \sin \frac{k\pi}{M} x \right), \end{aligned}$$

since  $\sigma^2 = ct$ , where again the formula  $\widehat{g}_\sigma(\omega) = e^{-\sigma^2 \omega^2}$  in (5.1.3) is applied. ■

**Example 5.1.3** Find the solution  $u(x, t)$  of the initial value (heat diffusion) PDE (5.1.8) with initial heat content  $u_0(x) = x^n$ , for  $n = 1$  and 2.

**Solution** Since the PDE (5.1.8) describes the heat diffusion process, let us consider  $u_0(x)$  as the initial temperature at  $x \in \mathbb{R}$ . Hence, the solution  $u(x, t)$  is the temperature at the time instant  $t > 0$ , at the same position  $x \in \mathbb{R}$ .

For  $n = 1$ , the temperature at  $t > 0$  and location  $x \in \mathbb{R}$  is given by

$$\begin{aligned} u(x, t) &= \int_{-\infty}^{\infty} (x - y) g_\sigma(y) \, dy \\ &= x \int_{-\infty}^{\infty} g_\sigma(y) \, dy - \int_{-\infty}^{\infty} y g_\sigma(y) \, dy, \end{aligned}$$

where  $\sigma^2 = ct$ . Since the first integral is equal to 1 and the second integral is 0 (with odd function  $y g_\sigma(y)$ ), we have

$$u(x, t) = x, \text{ for all } t \geq 0.$$

That is, the temperature does not change with time; or there is no diffusion at all.

For  $n = 2$ , since  $(x - y)^2 = x^2 - 2xy + y^2$ , the same argument as above yields

$$u(x, t) = \int_{-\infty}^{\infty} (x - y)^2 g_\sigma(y) \, dy = x^2 + \int_{-\infty}^{\infty} y^2 g_\sigma(y) \, dy,$$

where  $\sigma^2 = ct$ . Observe that because

$$\begin{aligned} \int_{-\infty}^{\infty} y^2 e^{-\alpha y^2} \, dy &= -\frac{\partial}{\partial \alpha} \int_{-\infty}^{\infty} e^{-\alpha y^2} \, dy \\ &= -\frac{\partial}{\partial \alpha} \sqrt{\frac{\pi}{\alpha}} = \frac{1}{2} \sqrt{\pi} \alpha^{-3/2}, \end{aligned}$$

we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} y^2 g_\sigma(y) \, dy &= \frac{1}{2\sigma\sqrt{\pi}} \int_{-\infty}^{\infty} y^2 e^{-(\frac{y}{2\sigma})^2} \, dy \\ &= \frac{1}{2\sigma\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} (2\sigma)^3 = 2\sigma^2 = 2ct, \end{aligned}$$

and this yields:

$$u(x, t) = x^2 + 2ct, \text{ for all } t > 0.$$

Observe that the temperature at  $x$  increases from  $u(x, 0) = x^2$  to  $u(x, c) = x^2 + 2ct$  as  $t > 0$  increases. This is truly “global warming” everywhere! ■

### 5.1.3 Gaussian convolution as solution of heat equation in the Euclidean space

We now extend our discussion from one spatial variable  $x \in \mathbb{R}$  to  $s$  spatial variables  $\mathbf{x} = (x_1, \dots, x_s) \in \mathbb{R}^s$ . Let  $|\mathbf{x}|$  denote the Euclidean norm of  $\mathbf{x} \in \mathbb{R}^s$ ; that is

$$|\mathbf{x}|^2 = x_1^2 + \dots + x_s^2,$$

and observe that the Gaussian function  $g_\sigma(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{R}^s$ , defined by

$$g_\sigma(\mathbf{x}) = \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4\sigma^2}}, \quad (5.1.11)$$

can be written as the product of the 1-dimensional Gaussian functions; namely

$$g_\sigma(\mathbf{x}) = g_\sigma(x_1, \dots, x_s) = g_\sigma(x_1) \cdots g_\sigma(x_s).$$

Hence, when the time variable  $t$  is defined by (5.1.4); that is  $\sigma^2 = ct$ , the extension of  $G(x, t)$  in (5.1.5) to  $\mathbb{R}^s$ ,  $s \geq 2$ , is given by

$$G(\mathbf{x}, t) = G(x_1, \dots, x_s, t) = g_\sigma(\mathbf{x}) = \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}}. \quad (5.1.12)$$

Indeed, by (5.1.11) and (5.1.5), we have

$$\begin{aligned} G(\mathbf{x}, t) &= g_\sigma(x_1) \cdots g_\sigma(x_s) \\ &= \left( \frac{1}{\sqrt{4\pi\sigma^2}} e^{-x_1^2/4\sigma^2} \right) \cdots \left( \frac{1}{\sqrt{4\pi\sigma^2}} e^{-x_s^2/4\sigma^2} \right) \\ &= \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-x_1^2/4\sigma^2} \cdots e^{-x_s^2/4\sigma^2} \\ &= \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-(x_1^2 + \dots + x_s^2)/4\sigma^2} = \frac{1}{(4\pi\sigma^2)^{s/2}} e^{-|\mathbf{x}|^2/4\sigma^2} \\ &= \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \end{aligned}$$

as desired.

Next, we extend the convolution operation from 1-dimension to  $s$ -dimension by

$$(u_0 * h)(\mathbf{x}) = \int_{\mathbb{R}^s} u_0(\mathbf{y}) h(\mathbf{x} - \mathbf{y}) d\mathbf{y}. \quad (5.1.13)$$

Then for fixed  $t \geq 0$ , the  $s$ -dimensional convolution with the spatial variables  $\mathbf{x} = (x_1, \dots, x_s)$  of  $h(\mathbf{x}) = G_c(\mathbf{x}, t)$  can be written as consecutive 1-

dimensional convolutions; namely,

$$\begin{aligned}
 (u_0 * G(\cdot, t))(\mathbf{x}) &= \int_{\mathbb{R}^s} u_0(\mathbf{y}) G(\mathbf{x} - \mathbf{y}, t) d\mathbf{y} \\
 &= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{s \text{ integrals}} u_0(y_1, \dots, y_s) G(x_1 - y_1, t) \cdots G(x_s - y_s, t) dy_1 \cdots dy_s \\
 &= (u_0(\underbrace{\cdot, \cdot, \dots, \cdot}_{s \text{ components}}) *_1 g_\sigma *_2 \cdots *_s g_\sigma)(x_1, \dots, x_s), \tag{5.1.14}
 \end{aligned}$$

where “ $*_k$ ” denotes the 1-dimensional convolution with respect to the  $k^{\text{th}}$  component (say  $y_k$  in (5.1.13)).

Hence, as in the 1-dimensional case, it is not difficult to verify that

$$G(\mathbf{x}, t) = G(x_1, \dots, x_s, t)$$

is the solution of the  $s$ -dimensional heat equation with  $\delta(\mathbf{x}) = \delta(x_1) \cdots \delta(x_s)$  as the initial heat source; namely,

$$\begin{cases} \frac{\partial}{\partial t} G(\mathbf{x}, t) = c \nabla^2 G(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\ G(\mathbf{x}, 0) = \delta(\mathbf{x}) = \delta(x_1) \cdots \delta(x_s), & \mathbf{x} \in \mathbb{R}^s, \end{cases} \tag{5.1.15}$$

where  $\nabla^2$  denotes the Laplace operator, defined by

$$\nabla^2 G(\mathbf{x}, t) = \frac{\partial^2}{\partial x_1^2} G(\mathbf{x}, t) + \cdots + \frac{\partial^2}{\partial x_s^2} G(\mathbf{x}, t).$$

To verify that  $G(\mathbf{x}, t)$  satisfies the heat diffusion equation, we simply follow the same computations in the derivation of (5.1.6), as follows:

$$\begin{aligned}
 \frac{\partial}{\partial t} G(\mathbf{x}, t) &= \frac{1}{(4\pi c)^{s/2}} e^{-(\frac{|\mathbf{x}|^2}{4c})t^{-1}} \left\{ -\frac{s}{2} t^{-\frac{s+2}{2}} + t^{-\frac{s}{2}} \left( \frac{|\mathbf{x}|^2}{4c} \right) t^{-2} \right\}; \\
 \frac{\partial}{\partial x_k} G_c(x_1, \dots, x_s, t) &= \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{t^{-1}}{2c} x_k \right\},
 \end{aligned}$$

so that

$$\frac{\partial^2}{\partial x_k^2} G(\mathbf{x}, t) = \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{t^{-1}}{2c} + \left( -\frac{t^{-1}}{2c} x_k \right)^2 \right\}.$$

Hence, it follows that

$$\begin{aligned}
 c\nabla G(\mathbf{x}, t) &= c \sum_{k=1}^s \frac{\partial^2}{\partial x_k^2} G(x_1, \dots, x_s, t) \\
 &= \frac{t^{-\frac{s}{2}}}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{st^{-1}}{2c} + \frac{t^{-2}}{4c^2} \sum_{k=1}^s x_k^2 \right\} \\
 &= \frac{1}{(4\pi c)^{s/2}} e^{-\frac{|\mathbf{x}|^2}{4c} t^{-1}} \left\{ -\frac{s}{2} t^{-\frac{s}{2}-1} + t^{-\frac{s}{2}} \left( \frac{|\mathbf{x}|^2}{4c} \right)^2 t^{-2} \right\} \\
 &= \frac{\partial}{\partial t} G(\mathbf{x}, t).
 \end{aligned}$$

To apply the above result to an arbitrary heat source, we simply follow the same argument for the 1-dimensional setting and apply (5.1.14) and (5.1.15) to obtain the solution of the initial value PDE:

$$\begin{cases} \frac{\partial}{\partial t} u(\mathbf{x}, t) = c\nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in \mathbb{R}^s, t \geq 0, \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^s, \end{cases} \quad (5.1.16)$$

where the initial condition is any integrable function  $u_0(\mathbf{x}) = u_0(x_1, \dots, x_s)$  in  $\mathbb{R}^s$ .

We summarize the above discussion in the following theorem.

**Theorem 5.1.2** *Let  $u_0(\mathbf{x})$  be a measurable function in  $\mathbb{R}^s$ ,  $s \geq 1$ , with at most polynomial growth such that the set of  $\mathbf{x} \in \mathbb{R}^s$  on which  $u_0(\mathbf{x})$  is discontinuous has (Lebesgue) measure zero. Then the solution of the initial value PDE (5.1.16) is given by*

$$u(\mathbf{x}, t) = (u_0 * G(\cdot, t))(\mathbf{x}), \quad (5.1.17)$$

as defined in (5.1.14).

**Example 5.1.4** Let  $c > 0$  and  $\nabla^2$  be the Laplace operator defined by

$$\nabla^2 f(x, y) = \frac{\partial^2}{\partial x^2} f(x, y) + \frac{\partial^2}{\partial y^2} f(x, y).$$

Re-formulate the solution  $u(x, y, t)$  of the 2-dimensional (heat diffusion) initial value PDE

$$\frac{\partial}{\partial t} u(x, y, t) = c\nabla^2 u(x, y, t)$$

in (5.1.16) with initial (input) function  $u_0(x, y)$ , according to Theorem 5.1.2 explicitly in terms of the 1-dimensional Gaussian function.

**Solution** According to Theorem 5.1.2 and the commutative property of the convolution operation, we have

$$\begin{aligned} u(x, y, t) &= \int_{-\infty}^{\infty} g_{\sigma}(y_1) \left\{ \int_{-\infty}^{\infty} u_0(x - x_1, y - y_1) g_{\sigma}(x_1) dx_1 \right\} dy_1 \\ &= \frac{t^{-1}}{4\pi c} \int_{-\infty}^{\infty} e^{-\frac{t-1}{4c} y_1^2} \left\{ \int_{-\infty}^{\infty} u_0(x - x_1, y - y_1) e^{-\frac{t-1}{4c} x_1^2} dx_1 \right\} dy_1, \end{aligned} \quad (5.1.18)$$

where  $\sigma^2 = ct$ . ■

**Definition 5.1.1** A function  $u_0(x, y)$  of two variables is said to be separable, if there exist functions  $f_j$  and  $h_k$  of one variable, such that

$$u_0(x, y) = \sum_{j,k} a_{j,k} f_j(x) h_k(y) \quad (5.1.19)$$

for some constants  $a_{j,k}$ , where  $\sum_{j,k}$  denotes a finite double sum, such as:

$$\sum_{j=0}^m \sum_{k=0}^n, \sum_{\ell=0}^n \sum_{j+k=\ell}, \sum_{j=0}^m \sum_{k=0}^{n-j}, \text{ etc.}$$

**Example 5.1.5** Let  $u_0(x, y)$  be a separable function as defined by (5.1.19). Write out the solution  $u(x, y, t)$  of the 2-dimensional (heat diffusion) PDE in Example 5.1.4 with initial condition  $u_0(x, y)$  in the most useful form for computation.

**Solution** Let  $\sigma^2 = ct$  and apply the formulation (5.1.18) to obtain

$$\begin{aligned} u(x, y, t) &= \int_{-\infty}^{\infty} g_{\sigma}(y_1) \left\{ \int_{-\infty}^{\infty} \sum_{j,k} a_{j,k} f_j(x - x_1) h_k(y - y_1) g_{\sigma}(x_1) dx_1 \right\} dy_1 \\ &= \sum_{j,k} a_{j,k} \left( \int_{-\infty}^{\infty} f_j(x - x_1) g_{\sigma}(x_1) dx_1 \right) \\ &\quad \times \left( \int_{-\infty}^{\infty} h_k(y - y_1) g_{\sigma}(y_1) dy_1 \right), \end{aligned} \quad (5.1.20)$$

with  $\sigma = \sqrt{ct}$ . After computing (5.1.20), we may write out the solution  $u(x, y, t)$  by replacing  $\sigma$  with  $\sqrt{ct}^{1/2}$ , as follows:

$$\begin{aligned} u(x, y, t) &= \frac{t^{-1}}{4\pi c} \sum_{j,k} a_{j,k} \left( \int_{-\infty}^{\infty} f_j(x - x_1) e^{-\frac{t-1}{4c} x_1^2} dx_1 \right) \\ &\quad \times \left( \int_{-\infty}^{\infty} h_k(y - y_1) e^{-\frac{t-1}{4c} y_1^2} dy_1 \right). \end{aligned} \quad \blacksquare$$



**Example 5.1.6** Apply the solution in Example 5.1.5 to compute the solution  $u(x, y, t)$  of the 2-dimensional (heat diffusion) initial value PDE in Example 5.1.4 with initial condition  $u_0(x, y)$  given by (5.1.19), where  $f_j(x) = \cos \frac{j\pi}{M}x$  and  $h_k(y) = \cos \frac{k\pi}{N}y$  for  $M, N > 0$ .

**Solution** For  $f_j(x) = \cos \frac{j\pi}{M}x$  and  $h_k(y) = \cos \frac{k\pi}{N}y$ , application of the solution in Example 5.1.1 yields,

$$\int_{-\infty}^{\infty} f_j(x - x_1) g_{\sigma}(x_1) dx_1 = e^{-\sigma^2 (\frac{j\pi}{M})^2} \cos \frac{j\pi}{M}x = e^{-(\frac{j\pi}{M})^2 ct} \cos \frac{j\pi}{M}x.$$

Hence, it follows from (5.1.20) in the previous example that

$$u(x, y, t) = \sum_{j,k} a_{j,k} e^{-\left((\frac{j\pi}{M})^2 + (\frac{k\pi}{N})^2\right) ct} \cos \frac{j\pi}{M}x \cos \frac{k\pi}{N}y.$$

■

## 5.2 The method of separation of variables

The most elementary and commonly used approach to solving a linear partial differential equation (PDE) is to treat the solution as an infinite sum of simple-minded solutions that the variables are separated. In this subunit, the method of separation of variables is introduced to change a partial differential equation (PDE) of  $n$  independent variables to two differential equations, with one ordinary differential equation (ODE) and one PDE of  $n - 1$  independent variables.

### 5.2.1 Separation of Time and Spatial Variables

First, let us consider  $n = 2$  and separate a PDE to two ODE's. If a function  $u(x, t)$  of two independent variables  $x$  and  $t$  is written as the product of two dependent variables  $X = X(x)$  and  $T = T(t)$ , with  $X(x)$  independent of  $t$  and  $T(t)$  is independent of  $x$ , we say that  $u(x, t)$  is separated, namely

$$u(x, t) = X(x)T(t).$$

Of course, this artificial way of writing the original function  $u(x, t)$  makes little sense. But it helps in changing a linear partial differential equation (PDE) of  $u$  into two ordinary differential equations (ODE). For example, the one-dimensional heat equation

$$\frac{\partial}{\partial t} u(x, t) = c \frac{\partial^2}{\partial x^2} u(x, t) \quad (5.2.1)$$

becomes

$$X(x)T'(t) = cX''(x)T(t). \quad (5.2.2)$$

Hence, dividing both sides by  $X(x)T(t)$ , we obtain

$$\frac{T'(t)}{cT(t)} = \frac{X''(x)}{X(x)}. \quad (5.2.3)$$

Observe that since the right-hand side  $T'(t)/cT(t)$  is independent of  $x$  and the left-hand side  $X''(x)/X(x)$  is independent of  $t$ , they must be the same constant, say  $\lambda$ . Hence, the PDE (5.2.1) becomes two ODE's (ordinary differential equations):

$$T'(t) = \lambda cT(t) \quad (5.2.4)$$

$$X''(x) = \lambda X(x). \quad (5.2.5)$$

The constant  $\lambda$  is instrumental to the solution of a PDE with boundary conditions and/or initial values.

### 5.2.2 Superposition Solution

Indeed, if  $T_j(t)$ ,  $j = 0, 1, \dots$  are solutions of (5.2.4) and  $X_j(x)$ ,  $j = 0, \dots$  are solutions of (5.2.5), where the "eigenvalues"  $\lambda = \lambda_j$  are chosen appropriately to satisfy the given "boundary conditions", then since each

$$u_j(x, t) = X_j(x)T_j(t), \quad j = 0, 1, \dots,$$

is a solution of the original PDE (5.2.1), the formal (possibly infinite) linear combination

$$\sum_{j=0}^{\infty} b_j u_j(x, t) = \sum_{j=0}^{\infty} b_j X_j(x)T_j(t) \quad (5.2.6)$$

is also a solution of (5.2.1). If the infinite series indeed converges, then by applying the given boundary or initial conditions to determine the coefficients  $b_0, b_1, b_2, \dots$ , we obtain the solution of the given boundary-value and/or initial-value problem with the PDE model (5.2.1). The consideration of the formal linear combination (5.2.6) is called the "Principle of superposition."

### 5.2.3 Extension to two spatial variables

The method of separation of (dependent) variables and the principle of superposition can be extended to linear PDE models for higher dimensions. To demonstrate the feasibility of this extension, let us consider the two-dimensional heat equation

$$\frac{\partial}{\partial t} = c\nabla^2 u(x, y, t) \quad (5.2.7)$$

where  $\nabla^2$  denotes the two-dimensional Laplacian operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$$

with  $(x, y)$  denoting the two-dimensional spatial coordinates. Then we may first write

$$u(x, y, t) = U(x, y)T(t)$$

to separate  $u$  to two dependent variables  $U = U(x, y)$  and  $T = T(t)$ . Then the PDE (5.2.7) becomes

$$U(x, y)T'(t) = c\nabla U(x, y)T(t). \quad (5.2.8)$$

Dividing both sides of (5.2.8) by  $cU(x, y)T(t)$ , we rewrite (5.2.8) as

$$\frac{T'(t)}{cT(t)} = \frac{\nabla U(x, y)}{U(x, y)}. \quad (5.2.9)$$

Again, since  $T'(t)/cT(t)$  is independent of  $x, y$  and  $\nabla U(x, y)/U(x, y)$  is independent of  $t$ , they must be the same constant, say  $-\lambda$  again. Thus, the PDE (5.2.7) becomes an ODE (for  $T(t)$ ) and a PDE (for  $U(x, y)$ ), namely:

$$\begin{aligned} T'(t) &= -\lambda cT(t) \\ \nabla U(x, y) + \lambda U(x, y) &= 0. \end{aligned} \quad (5.2.10)$$

To separate the PDE (5.2.11), we may write

$$U(x, y) = X(x)Y(y).$$

Then (5.2.11) yields

$$X''(x)Y(y) + X(x)Y''(y) + \lambda X(x)Y(y) = 0$$

or

$$\begin{aligned} \frac{X''(x)}{X(x)} + \frac{Y''(y)}{Y(y)} + \lambda &= 0 \\ -\frac{X''(x)}{X(x)} &= \frac{Y''(y)}{Y(y)} + \lambda. \end{aligned}$$

The left-hand side of the above equation is a function of  $x$  only, while the right-hand side is a function of  $y$  only. Thus both sides are the same constant, say,  $-\tilde{\lambda}$ . Therefore, we have two ODE's:

$$X''(x) - \tilde{\lambda}X(x) = 0, \quad (5.2.11)$$

$$Y''(y) + (\lambda - \tilde{\lambda})Y(y) = 0. \quad (5.2.12)$$

To summarize, we have to solve three ODE's by solving three eigenvalue

problems (5.2.10), (5.2.12), and (5.2.13). The eigenvalues  $\lambda = \lambda_j$  and  $\tilde{\lambda} = \tilde{\lambda}_{k,\ell}$  must be so chosen that the boundary and/or initial values are satisfied.

Finally, the principle of superposition allows us to formulate the general solution

$$u(x, y, t) = \sum_{j=0}^{\infty} \sum_{k,\ell=0}^{\infty} b_{j,k,\ell} T_j(t) X_k(x) Y_\ell(y), \quad (5.2.13)$$

where the coefficients  $b_{j,k,\ell}$  are to be determined by the boundary and/or initial values.

### References

- (1) Peter Olver, “Introduction to Partial Differential Equations, pages 103–109, University of Minnesota link.
- (2) Marcus Pivato, “Linear Partial Differential Equations and Fourier Theory, Cambridge University Press link.

## 5.3 Fourier series solution

In the history of mathematics, the most fruitful research collaboration, before the famous team of Godfrey Hardy and John Littlewood of Cambridge University almost two centuries later, was probably the collaboration between Daniel Bernoulli (1700–1782) and Leonhard Euler (1707–1783), during the five years (1727–1732), when they were together at the St. Petersburg Academy of Science. Together, they accomplished important work in hydrodynamics, probability, and the theory of oscillations. It is interesting to learn, however, that even these two giants could not agree on the solution of the two-point boundary-valued problem:

$$\begin{cases} \frac{\partial^2}{\partial t^2} u(x, t) = c^2 \frac{\partial^2}{\partial x^2} u(x, t), & 0 < x < L; \\ u(0, t) = u(L, t) = 0, & t \geq 0, \end{cases} \quad (5.3.1)$$

which is the partial differential equation that describes the vibrating string. Bernoulli proposed that the solution should be the “infinite series”

$$u(x, t) = \sum_{k=1}^{\infty} c_k \sin \frac{k\pi}{L} \cos \frac{ck\pi}{L} t \quad (5.3.2)$$

which clearly satisfies

$$\frac{\partial^2}{\partial t^2} u(x, t) = c^2 \frac{\partial^2}{\partial x^2} u(x, t)$$

with  $u(0, t) = u(L, t) = 0$  for all  $t \geq 0$ . Furthermore, for  $t = 0$ , before the vibration starts, the initial displacement of the string is given by

$$u(x, 0) = \sum_{k=1}^{\infty} c_k \sin \frac{k\pi}{L} x. \quad (5.3.3)$$

Unfortunately, Bernoulli could not relate the coefficients  $c_k, k = 1, 2, \dots$ , with the function that represents the initial displacement of the string. The reader should be reminded that this discovery by Bernoulli is more than 50 years before the introduction of Fourier series by Joseph Fourier (1768–1830).

Euler thought that Bernoulli's proposed solution was absurd, pointing out that if (5.3.2) would be the general solution of (5.3.1), then the initial displacement given by (5.3.3) must be an odd function, and gave a counter-example  $f(x) = x(L - x)$ , which certainly does not satisfy  $f(-x) = -f(x)$ . Euler then found his own general solution

$$u(x, t) = \frac{1}{2} (f(x + ct) + f(x - ct)) \quad (5.3.4)$$

of the initial-valued/boundary-valued problem for the vibrating string PDE

$$\begin{cases} \frac{\partial^2}{\partial t^2} u(x, t) = c^2 \frac{\partial^2}{\partial x^2} u(x, t), & 0 < x < L; \\ u(0, t) = u(L, t) = 0, & t \geq 0; \\ u(x, 0) = f(x), & 0 \leq x \leq L, \end{cases} \quad (5.3.5)$$

where  $f(x)$  is any given function in  $C^2[0, L]$ .

The reader is reminded of the study of Fourier series in Unit 3, and particularly the formulation of the Fourier coefficients of the sine series on  $[0, L]$  in Subunit 3.1 and the convergence of the Fourier series in Subunit 3.4. Indeed, Bernoulli's proposed solution (5.3.2)–(5.3.3) is correct, but he did not realize the coefficients  $c_k, k = 1, 2, \dots$ , in (5.3.3) can be computed, as being done by Fourier some 50 years later, from the given initial function  $u(x, 0) = f(x)$ . In the following, we will apply the method of separation of variables studied in Subunit 5.2 to derive Fourier series solutions of linear PDE governed by boundary and/or initial values.

### 5.3.1 One-spatial dimension

Let us first consider the one-spatial variable setting. In Subunit 5.1.2, we have

seen that for the heat equation on the entire real-line, if the initial heat content (or temperature) is a periodic function represented by a (Fourier) cosine series, then the Gaussian convolution of it yields the solution of the heat (diffusion) PDE with the given initial heat content as the initial-value function.

Hence, to unify our presentation, we will again study the heat equation, but now on a bounded interval of the real-line with perfect insulation at the two (boundary) endpoints. In other words, we will discuss the solution of the following initial-valued Neumann PDE, where the Neumann condition at the two end-points is described by the partial derivative with respect to the spatial variable being zero for all  $t \geq 0$ .

$$\begin{cases} \frac{\partial}{\partial t} u(x, t) = c \frac{\partial^2}{\partial x^2} u(x, t), & x \in (0, M), t > 0; \\ \frac{\partial}{\partial x} u(0, t) = \frac{\partial}{\partial x} u(M, t) = 0, & t > 0; \\ u(x, 0) = f(x), & 0 \leq x \leq M. \end{cases} \quad (5.3.6)$$

Observe that the heat equation (5.3.6) is different from the vibrating string PDE (5.3.5), in that the partial derivative with respect to the time variable  $t$  in (5.3.6) is of the first order, while that in (5.3.5) is of the second order. In addition, for the vibrating string PDE (5.3.5), the two endpoints of the string are fixed for all  $t \geq 0$ , so that the boundary condition is the so-called Dirichlet condition, while for the heat equation (5.3.6), the two endpoints are perfectly insulated to avoid heat diffusion across the boundary, so that the first derivative in the spatial variable  $x$  is zero, called a Neumann condition.

Recall from Subunit 5.2.1, by writing

$$U(x, t) = X(x)T(t),$$

where  $X$  is independent of  $t$  and  $T$  is independent of  $x$ , the PDE

$$\frac{\partial}{\partial t} U(x, t) = c \frac{\partial^2}{\partial x^2} U(x, t) \quad (5.3.7)$$

becomes

$$XT' = cX''T$$

or

$$\frac{T'}{cT} = \frac{X''}{X} = \lambda$$

for some constant  $\lambda$ . This separates the PDE (5.3.7) into two ODE's:

$$\begin{cases} T' &= \lambda cT \\ X'' &= \lambda X \end{cases}$$

where  $X = X(x)$  must satisfy the Neumann condition  $X'(0) = X'(M) = 0$ .

The general solution of the first order ODE  $T' = \lambda cT$  is simply

$$T(t) = a_0 e^{\lambda cT}.$$

Since the constant  $c$  is positive, being the heat conductivity constant, it is clear that  $\lambda$  cannot be positive, since heat (or temperature) cannot increase with increasing time  $t > 0$  (without further additional heat source for  $t > 0$ ). For this reason, we may write

$$\lambda = -\mu^2, \quad \mu \geq 0.$$

Of course, we may also deduce this conclusion from the ODE with zero Neumann condition:

$$\begin{cases} X''(x) &= \lambda X(x), \quad 0 \leq x \leq M; \\ X'(0) &= X'(M) = 0. \end{cases}$$

Indeed, if  $\lambda > 0$ , then

$$X(x) = a_1 e^{\sqrt{\lambda}x} + a_2 e^{-\sqrt{\lambda}x},$$

so that  $X'(x) = a_1 \sqrt{\lambda} e^{\sqrt{\lambda}x} - a_2 \sqrt{\lambda} e^{-\sqrt{\lambda}x}$ , and hence,

$$\begin{cases} 0 &= X'(0) = a_1 \sqrt{\lambda} - a_2 \sqrt{\lambda} = \sqrt{\lambda} (a_1 - a_2); \\ 0 &= X'(M) = (a_1 e^{\sqrt{\lambda}M} - a_2 e^{-\sqrt{\lambda}M}) \sqrt{\lambda}; \end{cases}$$

from which it follows that

$$a_1 = a_2 = 0.$$

Hence, for a non-trivial solution  $X(x)$ , we must choose  $\lambda \leq 0$ .

Now, for  $\lambda = -\mu^2$ , the general solution of  $X'' + \mu^2 X = 0$  is given by

$$X(x) = a_1 \cos \mu x + a_2 \sin \mu x,$$

which yields

$$X'(x) = -a_1 \mu \sin \mu x + a_2 \mu \cos \mu x.$$

To satisfy the Neumann condition  $X'(0) = 0$ , we have

$$0 = a_2 \mu.$$

Therefore, for  $\mu > 0$ ,  $a_2$  must be zero. To satisfy the other Neumann condition  $X'(M) = 0$ , we have

$$0 = -a_1 \mu \sin \mu M,$$

so that for  $\mu > 0$ ,  $\mu$  must be

$$\mu = \frac{k\pi}{M} \quad \text{for integers } k;$$

or

$$\lambda = -\mu^2 = -\left(\frac{k\pi}{M}\right)^2, \quad k = 0, 1, 2, \dots \quad (5.3.8)$$

Here, we have included  $k = 0$  to allow  $\mu = 0$  and arbitrary  $a_1$ . In addition, since we have  $k^2$  in general, we may select positive values of  $k$  without changing the values of  $\mu$ . Putting (5.3.8) into  $U = XT$ , we obtain

$$U(t) = b_k e^{-(\frac{k\pi}{M})^2 t} \cos\left(\frac{k\pi}{M} t\right),$$

where we have replaced the constant  $a_0 a_1$  by  $b_k$ . Therefore, by applying the principle of superposition as introduced in Subunit 5.2, we may formulate the general solution of the heat diffusion PDE (5.3.6) as

$$U(x, t) = \frac{b_0}{2} + \sum_{k=1}^{\infty} b_k e^{-(\frac{k\pi}{M})^2 t} \cos\left(\frac{k\pi}{M} x\right) \quad (5.3.9)$$

with  $u(x, 0) = f(x)$  for  $0 < x < L$ . Here, we replace  $b_0/2$  for the convenience of formulating the (Fourier) cosine coefficients:

$$b_k = \frac{2}{M} \int_0^M f(x) \cos \frac{k\pi x}{M} dx. \quad (5.3.10)$$

In summary, we have proved the following theorem.

**Theorem 5.3.1** *Let  $f(x)$  be a square integrable function on  $[0, M]$  with (Fourier) cosine coefficients  $b_0, b_1, b_2, \dots$  in (5.3.10). Then the solution of the heat equation (5.3.6) is given by  $u(x, t)$  in (5.3.9).*

**Example 5.3.1** Solve the heat equation (5.3.9) on the interval  $[0, \pi]$  with initial heat content  $f(x) = 2 - \cos 3x$ .

**Solution** Since  $f(x) = 2 - \cos 3x$  is already a (Fourier) cosine series on the interval  $[0, \pi]$ , the solution of (5.3.9) with  $M = \pi$  is simply

$$\begin{aligned} u(x, t) &= 2 - e^{-c(\frac{3\pi}{\pi})^2 t} \cos 3x \\ &= 2 - e^{-9t} \cos 3x. \end{aligned}$$

■

**Example 5.3.2** Solve the heat equation (5.3.9) on the interval  $[0, \pi]$  with initial heat content  $f(x) = \frac{x}{\pi}$ .

**Solution** We first compute the (Fourier) cosine coefficients, with  $[0, M] =$



$[0, \pi]$ . For  $k > 0$ , we have

$$\begin{aligned}
 b_k &= \frac{2}{\pi} \int_0^\pi \frac{x}{\pi} \cos kx dx \\
 &= \frac{2}{\pi^2} \left[ \left( \frac{x \sin kx}{k} \right) \Big|_0^\pi - \left( \int_0^\pi \frac{\sin kx}{k} dx \right) \right] \\
 &= \frac{2}{\pi^2 k^2} \left[ (0 - 0) - \left( -\frac{\cos kx}{k^2} \right) \Big|_0^\pi \right] \\
 &= \frac{2}{\pi^2 k^2} ((-1)^k - 1).
 \end{aligned}$$

On the other hand, for  $k = 0$ , we have

$$b_0 = \frac{2}{\pi} \int_0^\pi \frac{x}{\pi} dx = \frac{2}{\pi^2} \cdot \frac{\pi^2}{2} = 1.$$

Hence,  $b_0 = 1$ ,  $b_{2k} = 0$ , and  $b_{2k-1} = \frac{-4}{\pi^2(2k-1)^2}$ . Therefore, putting these values in (5.3.9), we have

$$u(x, t) = \frac{1}{2} - \frac{4}{\pi^2} \sum_{k=1}^{\infty} \frac{1}{(2k-1)^2} e^{-c(2k-1)^2 t} \cos(2k-1)x.$$

■

### 5.3.2 Extension to higher dimensional domains

In this subunit, we extend the discussion in Subunit 5.3.1 to higher spatial dimensions  $\mathbb{R}^s$ , for  $s \geq 2$ . We remark that the key contents of Subunit 5.3.1 include the supplementary reading materials: Markus Pivoto, “Linear PDE and Fourier Theory” 1A, and Peter Olver, “Introduction to PDE” Chapter 8.

Let  $\nabla^2$  denote the Laplacian operator defined by

$$\nabla^2 U(\mathbf{x}) = \sum_{k=1}^s \frac{\partial^2}{\partial x_k^2} U(\mathbf{x}), \quad (5.3.11)$$

where  $\mathbf{x} := (x_1, \dots, x_s) \in D$ , and  $D$  is any bounded and connected region (i.e. the closure of a domain) in  $\mathbb{R}^s$ ,  $s > 1$ . The extension of the initial-valued Neumann PDE (5.3.5) from a bounded interval  $(0, M) \subset \mathbb{R}^1$  to a bounded open set  $D$  in  $\mathbb{R}^s$  is simply

$$\begin{cases} \frac{\partial}{\partial t} u(\mathbf{x}, t) = c \nabla^2 u(\mathbf{x}, t), & \mathbf{x} \in D, \quad t \geq 0; \\ \frac{\partial}{\partial \mathbf{n}} u(\mathbf{x}, t) = 0, & \mathbf{x} \in \partial D, \quad t > 0; \\ u(\mathbf{x}, 0) = f(\mathbf{x}), & \mathbf{x} \in D, \end{cases} \quad (5.3.12)$$

where  $\partial D$  denotes the boundary curve of  $D$ , which is assumed to have continuous turning tangent almost everywhere,  $\mathbf{n}$  denotes the outer unit normal vector on  $\partial D$ , and

$$\frac{\partial}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla$$

with  $\nabla$  denoting the vector-valued gradient operator. Of course, the initial heat content  $f(\mathbf{x})$  is assumed to be square-integrable on  $D$ . Another way to formulate the above equation is

$$\frac{\partial}{\partial \mathbf{n}} u(\mathbf{x}, t) = \sum_{k=1}^s n_k \frac{\partial}{\partial x_k} u(\mathbf{x}, t)$$

where  $\mathbf{n} = (n_1, \dots, n_s)$ .

In this subunit, we only consider the rectangular region

$$D = (0, M_1) \times \dots \times (0, M_s),$$

and in particular,  $D = (0, M) \times (0, N)$  for the two-dimensional spatial space with  $s = 2$ . Recall from Subunit 5.2 that by the method of separation of variables (5.2.8)–(5.2.13) restricts the dependent variable  $X = X(x)$  to satisfy

$$X'(0) = X'(M) = 0.$$

Since the solution of O.D.E. (5.2.12), for negative  $\tilde{\lambda}$ , is

$$X(x) = B \cos \left( \sqrt{-\tilde{\lambda}} x \right) + C \sin \left( \sqrt{-\tilde{\lambda}} x \right),$$

that

$$X'(x) = -B \sqrt{-\tilde{\lambda}} \sin \left( \sqrt{-\tilde{\lambda}} x \right) + \sqrt{-\tilde{\lambda}} C \cos \left( \sqrt{-\tilde{\lambda}} x \right),$$

the condition  $X'(0) = 0$  implies  $C = 0$  and the condition  $X'(M) = 0$  implies  $\sqrt{-\tilde{\lambda}} = \frac{\pi k}{M}$  for any integer  $k$ . Hence, under these two zero Neumann conditions, the general solution of the O.D.E. (5.2.12) is

$$X(x) = X_k(x) = b_k \cos \frac{k\pi x}{M}$$

for  $k = 0, 1, 2, \dots$

Similarly, under the two Neumann boundary conditions

$$Y'(0) = Y'(N) = 0,$$

the general solution of the O.D.E. (5.2.13) restricts the values of  $\lambda - \tilde{\lambda}$  to be

$$\lambda - \tilde{\lambda} = -\left(\frac{\ell\pi}{N}\right)^2,$$

for  $\ell = 0, 1, 2, \dots$ . Since  $\lambda = -\left(\frac{\pi k}{M}\right)^2$ , we have

$$\lambda - \tilde{\lambda} = -\left(\left(\frac{\pi k}{M}\right)^2 + \left(\frac{\pi \ell}{N}\right)^2\right)$$

for  $k, \ell = 0, 1, 2, \dots$ . Hence, by the principle of superposition, the general solution of the Neumann PDE in (5.3.12), with  $D = [0, M] \times [0, N]$  is given by

$$u(x, y, t) = \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} d(k, \ell) b_{k, \ell} e^{-c\left(\left(\frac{k\pi}{M}\right)^2 + \left(\frac{\ell\pi}{N}\right)^2\right)t} \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right), \quad (5.3.13)$$

where we have introduced the notation

$$d(k, \ell) = 2^{-(\delta_k + \delta_\ell)}, \quad k, \ell = 0, 1, 2, \dots, \quad (5.3.14)$$

by using the Kronecker delta symbol  $\delta_j$  defined by  $\delta_0 = 1$  and  $\delta_j = 0$  for  $j \neq 0$  (applied to  $j = k$  and  $j = \ell$ ), and where  $b_{k, \ell}$  are certain constants. To impose the initial condition, we have

$$u(x, y, 0) = f(x, y) = \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} d(k, \ell) b_{k, \ell} \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right), \quad (5.3.15)$$

which is the Fourier cosine series representation of  $f(\mathbf{x}) = f(x, y)$  in (5.3.12). Thus,  $b_{k, \ell}, k = 0, 1, \dots, \ell = 0, 1, \dots$  in (5.3.13) are the coefficients of the cosine series of the initial function  $u_0(x, y)$ , namely:

$$b_{k, \ell} = \frac{4}{MN} \int_0^M \int_0^N f(x, y) \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right) dx dy, \quad (5.3.16)$$

for  $k, \ell = 0, 1, 2, \dots$ . ■

For  $s \geq 3$ , we may consider separation of one variable at a time, by writing

$$U(\mathbf{x}) = U(x_1, \dots, x_{s-1}, x_s) = V(x_1, \dots, x_{s-1})X(x_s),$$

so that

$$\nabla_s^2 U = X \nabla_{s-1}^2 V + X'' V$$

(where the subscript of  $\nabla^2$  denotes the dimension of the Laplace operator). Hence, by an induction argument, we have the following result, which extends Theorem 5.3.1 to any dimension  $s \geq 2$ .

For convenience, the notation  $d(k, \ell)$  in (5.3.14) is extended to an arbitrary  $s \geq 2$ ; namely

$$d(k_1, \dots, k_s) = 2^{-\sum_{j=1}^s \delta_{k_j}}.$$

**Theorem 5.3.2** Let  $u_0 \in L_1((0, M_1) \times \cdots \times (0, M_s))$ , where  $M_1, \dots, M_s > 0$ . Then the solution of the Neumann diffusion PDE (5.3.12) for  $D = (0, M_1) \times \cdots \times (0, M_s)$  with initial heat content  $u(\mathbf{x}, 0) = u_0(\mathbf{x})$ ,  $\mathbf{x} \in D$ , is given by

$$u(\mathbf{x}, t) = \sum_{k_1=0}^{\infty} \cdots \sum_{k_s=0}^{\infty} d(k_1, \dots, k_s) b_{k_1 \dots k_s} e^{-c \sum_{j=1}^s \left(\frac{k_j \pi}{M_j}\right)^2 t} \times \cos\left(\frac{k_1 \pi}{M_1} x_1\right) \cdots \cos\left(\frac{k_s \pi}{M_s} x_s\right), \quad (5.3.17)$$

with

$$b_{k_1 \dots k_s} = \frac{2^s}{M_1 \cdots M_s} \int_D f(x_1, \dots, x_s) \cos\left(\frac{k_1 \pi}{M_1} x_1\right) \cdots \cos\left(\frac{k_s \pi}{M_s} x_s\right) d\mathbf{x},$$

where  $u(\mathbf{x}, 0) = f(\mathbf{x}) = f(x_1, \dots, x_s)$  and the convergence of the series in (5.3.17) is in the  $L_2(D)$ -norm.

**Example 5.3.3** Solve the Neumann diffusion PDE (5.3.12) for  $M = N = \pi$ , with initial heat content

$$u(\mathbf{x}, 0) = f(\mathbf{x}) = 1 + 2 \cos x \cos y + \cos x \cos 3y.$$

**Solution** Since the representation of  $u_0(x, y)$  is already its Fourier cosine series representation, it follows from Theorem 5.3.2 that the solution is given by

$$\begin{aligned} u(x, y, t) &= 1 + 2e^{-c(1^2+1^2)t} \cos x \cos y + e^{-c(1^2+3^2)t} \cos x \cos 3y \\ &= 1 + 2e^{-2ct} \cos x \cos y + e^{-10ct} \cos x \cos 3y. \end{aligned}$$

■

## 5.4 Boundary Value Problems

In this subunit, we extend the linear PDE model (5.3.12), that describes the isotropic heat diffusion process, with positive constant heat conductivity, to anisotropic heat diffusion, with the conductivity constant  $c$  replaced by a function that dictates the direction of heat diffusion. We first recall that

$$\nabla^2 = \nabla \cdot \nabla$$

where  $\nabla$  is the vector-valued **gradient operator**

$$\nabla f(x_1, \dots, x_s) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_s} \right),$$

(with scalar-valued function  $f$ ), and the scalar-valued **divergence operator**

$$\begin{aligned}\nabla \cdot \mathbf{F}(x_1, \dots, x_s) &= \nabla \cdot (f_1(x_1, \dots, x_s), \dots, f_s(x_1, \dots, x_s)) \\ &= \left( \frac{\partial}{\partial x_1} f_1 + \dots + \frac{\partial}{\partial x_s} f_s \right)(x_1, \dots, x_s),\end{aligned}$$

(with vector-valued function  $\mathbf{F} = (f_1, \dots, f_s)$ ). Therefore, for any constant  $c$  and scalar-valued function  $v(\mathbf{x}) = v(x_1, \dots, x_s)$ , we have

$$c \nabla^2 u(\mathbf{x}) = c \nabla \cdot \nabla v(\mathbf{x}) = \nabla \cdot (c \nabla v(\mathbf{x})).$$

This formulation allows us to replace the constant  $c$  by any differentiable function  $w(\mathbf{x})$ . With  $u(\mathbf{x}, t)$  denoting the heat content at the spatial location  $\mathbf{x}$  and time  $t$ , then the isotropic diffusion PDE

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) = c \nabla^2 u(\mathbf{x}, t)$$

with constant heat conductivity  $c > 0$ , can be extended to the anisotropic diffusion PDE

$$\frac{\partial}{\partial t} u(\mathbf{x}, t) = \nabla \cdot (c(|\nabla u(\mathbf{x}, t)|) \nabla u(\mathbf{x}, t)), \quad (5.4.1)$$

where  $c(p)$  is a function of the variable  $p$  defined for  $p \geq 0$ , and we will consider

$$p = |\nabla u(\mathbf{x}, t)| = \sqrt{u_{x_1}^2(\mathbf{x}, t) + \dots + u_{x_s}^2(\mathbf{x}, t)}.$$

Observe that since the PDE (5.4.1) is highly non-linear, we first describe how heat diffuses with various choices of the conductivity function  $c(p)$  in terms of the “geometry” of “arcs” or “edges.” This is usually called “diffusion geometry.” To compute reasonably good approximate solutions of an anisotropic PDE, we introduce the notion of “lagged anisotropic diffusion” in Subunit 5.4.2, converting the non-linear PDE to a system of linear PDE’s. To facilitate our discussion and solution of the PDE system, we replace  $c(p)$  in (5.4.1) by a general differentiable function  $w(\mathbf{x}) \geq 0$  and study the Neumann boundary value problem (in the spatial domain) obtained from (5.4.1) with  $c(p)$  replaced by  $w(\mathbf{x})$ , by application of the method of separation of variables to isolate the spatial PDE model.

### 5.4.1 Neumann boundary value problems

Let  $D$  be a bounded and simply connected domain in  $\mathbb{R}^2$ , such that the boundary  $\partial D$  is a piecewise smooth curve  $C$ , so that, with the exception of possibly a finite number of “corners,”  $C$  has continuous turning tangent  $\boldsymbol{\tau} = \boldsymbol{\tau}(x, y)$ , for  $(x, y) \in C = \partial D$ . To unify the notations with the previous subunit, we set the closure of  $D$  to be

$$\text{clos } D = D \cup C = D \cup \partial D.$$

In particular, for the rectangular region

$$D = (0, M) \times (0, N)$$

considered in Subunit 5.3.2, the boundary curve  $C$  of  $D$  consists of 4 straight edges and 4 corners.

We will use the notation

$$\mathbf{n} = \mathbf{n}(x, y)$$

for the unit normal of any smooth arc  $\gamma$  in  $D$ , and in particular, the outer normal of the boundary curve  $C$  of  $D$ . For the smooth arc  $\gamma$ , corresponding to  $\mathbf{n} = \mathbf{n}(x, y)$ ,  $(x, y) \in \gamma$ , we will introduce the unit tangent

$$\boldsymbol{\tau} = \boldsymbol{\tau}(x, y)$$

for the same  $(x, y) \in \gamma$ . Hence,  $(\boldsymbol{\tau}, \mathbf{n})$  constitutes a pair of local coordinates at  $(x, y)$ .

The Neumann boundary value problem to be studied in this subunit is the “anisotropic” PDE:

$$\begin{cases} \frac{\partial}{\partial t} u(x, y, t) = \nabla \cdot (w(x, y) \nabla u(x, y, t)), & (x, y) \in D; \\ \frac{\partial}{\partial \mathbf{n}} u(x, y, t)|_{(x, y) \in \partial D} = 0, \end{cases} \quad (5.4.2)$$

for all  $t > 0$ . As mentioned previously,  $w(x, y) > 0$  is a differentiable function. Observe that since (5.4.2) is a linear PDE, we may apply the method of separation of variables studied in Subunit 5.2 to write

$$u(x, y, t) = U(x, y)T(t),$$

so that

$$\frac{\partial u}{\partial t} = U T' \quad \text{and} \quad \nabla \cdot (w \nabla u) = (\nabla \cdot (w \nabla U))T,$$

which yields

$$\frac{T'}{T} = \frac{\nabla \cdot (w \nabla U)}{U},$$

where the left-hand side is independent of  $(x, y)$  and the right-hand side is independent of  $t$ . Thus, they must be equal to the same constant, say  $-\lambda$ . Also, since  $\frac{T'}{T} = -\lambda$ ,  $T(t) = e^{-\lambda t}$  (with any multiplication constant). It is intuitively clear that  $\lambda \geq 0$  for “diffusion” for increasing value of  $t$ . This will be shown rigorously later in this subunit. As to the spatial consideration, we have an eigenvalue problem

$$\nabla \cdot (w \nabla U) = -\lambda U.$$

Indeed, in view of the zero Neumann condition in (5.4.2), this eigenvalue problem can be formulated, more precisely, as

$$\begin{cases} \nabla \cdot (w \nabla v) = -\lambda v; \\ \frac{\partial}{\partial \mathbf{n}} v|_{\partial D} = 0, \end{cases} \quad (5.4.3)$$

where  $v = U$ . For any fixed  $\lambda$ , the set of solutions of the Neumann boundary value problem (5.4.3) is the same as the eigenspace:

$$S(\lambda) := \left\{ v : \frac{\partial}{\partial \mathbf{n}} v \Big|_{\partial D} = 0, \nabla \cdot (w \nabla v) = -\lambda v \right\} \quad (5.4.4)$$

corresponding to the eigenvalue  $\lambda$ . For  $\lambda_0 = 0$ , it is clear that  $S(\lambda_0)$  is non-trivial, since the constant functions lie in  $S(\lambda_0)$ . By applying the Gram-Schmidt orthonormalization procedure, we may assume that  $\{v_{0,k}\}_k$  is an orthonormal basis of the eigen-space  $S(\lambda_0)$ .

Consider two different eigenvalues  $\lambda_i$  and  $\lambda_j$  (with e.g.  $\lambda_i = \lambda_0 = 0$ ). Let  $(\lambda_i, v_i)$  and  $(\lambda_j, v_j)$  be two such eigenvalue-function pairs. In the following, we will prove that the eigenfunctions  $v_i$  and  $v_j$  must be orthogonal, namely,

$$\int_D v_i(x, y) v_j(x, y) dx dy = 0. \quad (5.4.5)$$

**Theorem 5.4.1** *Eigenfunctions corresponding to different eigenvalues of the operator defined in (5.4.4) are orthogonal.*

Our proof will depend on the following “Divergence Theorem,” namely: For vector-valued functions  $\vec{F}$  on  $D$ ,

$$\iint_D \nabla \cdot \vec{F}(x, y) dx dy = \oint_C \vec{F} \cdot \mathbf{n} ds, \quad (5.4.6)$$

where  $\mathbf{n}(x, y)$  is the unit outer normal at  $(x, y)$ , and the line integral over the close curve  $C = \partial D$  is considered to be in the counter-clockwise direction.

Now for the eigenfunction  $v_i$  and  $v_j$  discussed above, observe that

$$\begin{aligned} \nabla \cdot (v_i w \nabla v_j) &= w(\nabla v_i) \cdot (\nabla v_j) + v_i \nabla \cdot (w \nabla v_j); \\ \nabla \cdot (v_j w \nabla v_i) &= w(\nabla v_j) \cdot (\nabla v_i) + v_j \nabla \cdot (w \nabla v_i), \end{aligned}$$

so that by taking the difference, we have

$$\nabla \cdot (v_i w \nabla v_j) - \nabla \cdot (v_j w \nabla v_i) = v_i \nabla \cdot (w \nabla v_j) - v_j \nabla \cdot (w \nabla v_i). \quad (5.4.7)$$

The formula (5.4.6) of the Divergence Theorem can be applied to both  $\vec{F} = v_i w \nabla v_j$  and  $\vec{F} = v_j w \nabla v_i$  to show that both integrals  $\int \int_D$  on the left-hand side of (5.4.7) vanish, since

$$(\nabla u_j) \cdot \mathbf{n}|_{\partial D} = \frac{\partial}{\partial \mathbf{n}} u_j|_{\partial D} = 0 \quad \text{and} \quad (\nabla u_i) \cdot \mathbf{n}|_{\partial D} = \frac{\partial}{\partial \mathbf{n}} u_i|_{\partial D} = 0.$$

Therefore, from (5.4.7), we have

$$\iint_D v_i \nabla \cdot (w \nabla v_j) dx dy = \iint_D v_j \nabla \cdot (w \nabla v_i) dx dy,$$

or equivalently,

$$-\lambda_j \iint_D v_i v_j dx dy = -\lambda_i \iint_D v_j v_i dx dy,$$

since  $v_j \in S(\lambda_j)$  and  $v_i \in S(\lambda_i)$ . Hence, if  $\lambda_j \neq \lambda_i$ , (5.4.5) follows, completing the proof of the theorem. ■

To apply the above theorem to investigate if any eigenvalue  $\lambda \neq \lambda_0 (= 0)$  exists, we now know that if  $v$  is an eigenfunction corresponding to  $\lambda$ , then  $v$  must be orthogonal to the eigenspace  $S(\lambda_0)$ . (We'll use the notation  $v \perp S(\lambda_0)$  or  $v \in S^\perp(\lambda_0)$ .) Now, for  $v \in S(\lambda)$ ,

$$-\lambda v = \nabla \cdot (w \nabla v)$$

so that

$$-\lambda v^2 = v \nabla \cdot (w \nabla v).$$

On the other hand, since

$$\begin{aligned} \nabla \cdot (v w \nabla v) &= v \nabla \cdot (w \nabla v) + (\nabla v) \cdot (w \nabla v) \\ &= v \nabla \cdot (w \nabla v) + w |\nabla v|^2, \end{aligned}$$

it follows that

$$-\lambda v^2 = \nabla \cdot (v w \nabla v) - w |\nabla v|^2.$$

Hence, by applying the Divergence Theorem (5.4.6) again, but with  $\vec{F} = v w \nabla v$  in (5.4.6), and using the Neumann boundary condition

$$(\nabla v) \cdot \mathbf{n}|_{\partial D} = \frac{\partial}{\partial \mathbf{n}} v|_{\partial D} = 0,$$

(which implies  $\iint_D \nabla \cdot (v w \nabla v) dx dy = 0$ ), we have

$$\iint_D \lambda v^2 dx dy = \iint_D w |\nabla v|^2 dx dy.$$

That  $\lambda \geq 0$  is in agreement with the “diffusion” process  $T(t) = e^{-\lambda t}$ . In addition, the eigenvalue  $\lambda$  is given by the “Rayleigh quotient”:

$$\lambda = \frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2(x, y) dx dy}. \quad (5.4.8)$$

We will continue this discussion in Subunit 5.4.3, when we study the solution of the system of linear PDE's obtained by “linearization” of the non-linear anisotropic PDE model.



### 5.4.2 Anisotropic diffusion

In this subunit, we discuss the “geometry” of the diffusion process governed by various conductivity functions  $c(p)$ ,  $p \geq 0$ , for the anisotropic heat diffusion PDE, (5.4.9) below, with initial heat content  $u(x, y, 0) = u_0(x, y) = f(x, y)$ , for a bounded and simply connected region  $\text{clos } D = \Omega \cup C \subset \mathbb{R}^2$ , as introduced in the previous subunit:

$$\begin{cases} \frac{\partial}{\partial t} u(x, y, t) = \nabla \cdot (c(|\nabla u(x, y, t)|) \nabla u(x, y, t)), & (x, y) \in D; \\ \frac{\partial}{\partial \mathbf{n}} u(x, y, t)|_{\partial D} = 0, & (x, y) \in \partial D; \\ u(x, y, 0) = u_0(x, y), & (x, y) \in D, \end{cases} \quad (5.4.9)$$

for  $t \geq 0$ . For notational convenience, we will write  $u = u(x, y, t)$ , when it is understood that  $(x, y) \in D$  and  $t > 0$  (see (5.4.2) for  $\omega(x, y)$  in place of  $c(|\nabla u|)$  and the Neumann boundary condition, where  $\mathbf{n} = (x, y)$  denotes the outer unit normal at  $(x, y) \in \partial D$ ). Observe that if  $\alpha$  denotes the angle of inclination of the unit tangent vector  $\boldsymbol{\tau}$  of  $C$  with the  $x$ -axis, we have

$$\begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{n} \end{bmatrix} = \begin{bmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \quad (5.4.10)$$

where  $\mathbf{e}_1 = (1, 0)$  and  $\mathbf{e}_2 = (0, 1)$ . Here, we consider  $(\boldsymbol{\tau}, \mathbf{n})$  as the unit tangent-normal pair at any point  $(x, y) \in \partial D$ .

To change the derivative from the  $(x, y)$  coordinates to the local coordinates  $(\boldsymbol{\tau}, \mathbf{n})$ , we may apply (5.4.10) to write

$$\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\tau}} \\ \frac{\partial}{\partial \mathbf{n}} \end{bmatrix} = \begin{bmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix}, \quad (5.4.11)$$

or

$$\begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\tau}} \\ \frac{\partial}{\partial \mathbf{n}} \end{bmatrix}. \quad (5.4.12)$$

As to the second-order partial derivatives, we present the following computations:

$$\begin{aligned} \text{(i)} \quad \frac{\partial^2}{\partial x^2} &= \frac{\partial}{\partial x} \left( \cos \alpha \frac{\partial}{\partial \boldsymbol{\tau}} - \sin \alpha \frac{\partial}{\partial \mathbf{n}} \right) \\ &= \cos \alpha \frac{\partial}{\partial \boldsymbol{\tau}} \left( \cos \alpha \frac{\partial}{\partial \boldsymbol{\tau}} - \sin \alpha \frac{\partial}{\partial \mathbf{n}} \right) - \sin \alpha \frac{\partial}{\partial \mathbf{n}} \left( \cos \alpha \frac{\partial}{\partial \boldsymbol{\tau}} - \sin \alpha \frac{\partial}{\partial \mathbf{n}} \right) \\ &= \cos^2 \alpha \frac{\partial^2}{\partial \boldsymbol{\tau}^2} - 2 \cos \alpha \sin \alpha \frac{\partial^2}{\partial \boldsymbol{\tau} \partial \mathbf{n}} + \sin^2 \alpha \frac{\partial^2}{\partial \mathbf{n}^2}. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \text{(ii)} \quad \frac{\partial^2}{\partial x \partial y} &= \sin \alpha \cos \alpha \frac{\partial^2}{\partial \boldsymbol{\tau}^2} + (\cos^2 \alpha - \sin^2 \alpha) \frac{\partial^2}{\partial \boldsymbol{\tau} \partial \mathbf{n}} - \sin \alpha \cos \alpha \frac{\partial^2}{\partial \mathbf{n}^2}; \\ \text{(iii)} \quad \frac{\partial^2}{\partial y^2} &= \sin^2 \alpha \frac{\partial^2}{\partial \boldsymbol{\tau}^2} + 2 \sin \alpha \cos \alpha \frac{\partial^2}{\partial \boldsymbol{\tau} \partial \mathbf{n}} + \cos^2 \alpha \frac{\partial^2}{\partial \mathbf{n}^2}. \end{aligned}$$

In matrix formulation, we have:

$$\begin{bmatrix} \frac{\partial^2}{\partial x^2} \\ \frac{\partial^2}{\partial x \partial y} \\ \frac{\partial^2}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \cos^2 \alpha & -2 \sin \alpha \cos \alpha & \sin^2 \alpha \\ \sin \alpha \cos \alpha & \cos^2 \alpha - \sin^2 \alpha & -\sin \alpha \cos \alpha \\ \sin^2 \alpha & 2 \sin \alpha \cos \alpha & \cos^2 \alpha \end{bmatrix} \begin{bmatrix} \frac{\partial^2}{\partial \tau^2} \\ \frac{\partial^2}{\partial \tau \partial \mathbf{n}} \\ \frac{\partial^2}{\partial \mathbf{n}^2} \end{bmatrix}. \quad (5.4.13)$$

In particular, we remark that the Laplace operator  $\nabla^2$  is rotationally invariant, namely:

$$\begin{aligned} \nabla^2 &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \\ &= (\cos^2 \alpha + \sin^2 \alpha) \frac{\partial^2}{\partial \tau^2} + (-2 \cos \alpha \sin \alpha + 2 \cos \alpha \sin \alpha) \frac{\partial^2}{\partial \tau \partial \mathbf{n}} \\ &\quad + (\sin^2 \alpha + \cos^2 \alpha) \frac{\partial^2}{\partial \mathbf{n}^2} \\ &= \frac{\partial^2}{\partial \tau^2} + \frac{\partial^2}{\partial \mathbf{n}^2}, \end{aligned}$$

or we may write

$$\nabla_{(x,y)}^2 = \nabla_{(\tau,\mathbf{n})}^2, \quad (5.4.14)$$

In addition, by taking matrix inverse in (5.4.13), we have

$$\begin{bmatrix} \frac{\partial^2}{\partial \tau^2} \\ \frac{\partial^2}{\partial \tau \partial \mathbf{n}} \\ \frac{\partial^2}{\partial \mathbf{n}^2} \end{bmatrix} = \begin{bmatrix} \cos^2 \alpha & 2 \sin \alpha \cos \alpha & \sin^2 \alpha \\ -\sin \alpha \cos \alpha & \cos^2 \alpha - \sin^2 \alpha & \sin \alpha \cos \alpha \\ \sin^2 \alpha & -2 \sin \alpha \cos \alpha & \cos^2 \alpha \end{bmatrix} \begin{bmatrix} \frac{\partial^2}{\partial x^2} \\ \frac{\partial^2}{\partial x \partial y} \\ \frac{\partial^2}{\partial y^2} \end{bmatrix}. \quad (5.4.15)$$

Now, returning to PDE (5.4.9), by setting

$$p = |\nabla u| = \sqrt{u_x^2 + u_y^2},$$

we have

$$\begin{cases} \frac{u_x}{p} = \frac{u_x}{|\nabla u|} = -\sin \alpha, \\ \frac{u_y}{p} = \frac{u_y}{|\nabla u|} = \cos \alpha, \end{cases} \quad (5.4.16)$$

Hence, it follows that

$$\begin{cases} \frac{\partial p}{\partial x} = \frac{u_{xx}u_x + u_{xy}u_y}{\sqrt{u_x^2 + u_y^2}}, \\ \frac{\partial p}{\partial y} = \frac{u_{xy}u_x + u_{yy}u_y}{\sqrt{u_x^2 + u_y^2}}. \end{cases} \quad (5.4.17)$$

This implies, by direct computations, that

$$\begin{aligned}
\nabla \cdot (c(p) \nabla u) &= (\nabla c(p)) \cdot (\nabla u) + c(p) \nabla u \\
&= c'(p) \left( \frac{\partial p}{\partial x}, \frac{\partial p}{\partial y} \right) \cdot (u_x, u_y) + c(p) \nabla u \\
&= c'(p) \frac{u_{xx} u_x^2 + 2u_{xy} u_x u_y + u_{yy} u_y^2}{\sqrt{u_x^2 + u_y^2}} + c(p) \nabla u \\
&= c'(p) \left[ u_{xx} \frac{u_x^2}{p} + 2u_{xy} \frac{u_x u_y}{p} + u_{yy} \frac{u_y^2}{p} \right] + c(p) \nabla u \\
&= p c'(p) \left[ u_{xx} \left( \frac{u_x}{p} \right)^2 + 2u_{xy} \left( \frac{u_x}{p} \right) \left( \frac{u_y}{p} \right) + u_{yy} \left( \frac{u_y}{p} \right)^2 \right] + c(p) \nabla u \\
&= p c'(p) [(\sin^2 \alpha) u_{xx} - 2(\sin \alpha \cos \alpha) u_{xy} + (\cos^2 \alpha) u_{yy}] + c(p) \nabla u.
\end{aligned} \tag{5.4.18}$$

Therefore, by applying the change of variables from  $(x, y)$ -coordinates to the local coordinates  $(\boldsymbol{\tau}, \mathbf{n})$  in (5.4.14), we have, from (5.4.18),

$$\nabla \cdot (c(p) \nabla u) = p c'(p) u_{\mathbf{nn}} + c(p) (u_{\mathbf{nn}} + u_{\boldsymbol{\tau}\boldsymbol{\tau}}). \tag{5.4.19}$$

Hence, by introducing the function:

$$F(p) := p c(p), \tag{5.4.20}$$

so that

$$F'(p) := p c'(p) + c(p), \tag{5.4.21}$$

it follows from (5.4.19) that

$$\nabla \cdot (c(p) \nabla u) = F'(|\nabla u|) u_{\mathbf{nn}} + c(|\nabla u|) u_{\boldsymbol{\tau}\boldsymbol{\tau}}. \tag{5.4.22}$$

In other words, the anisotropic diffusion PDE in (5.4.9) can be written, for each  $(x, y) \in D$ , as

$$\frac{\partial u(x, y; t)}{\partial t} = F'(|\nabla u(x, y; t)|) u_{\mathbf{nn}} + c(|\nabla u(x, y; t)|) u_{\boldsymbol{\tau}\boldsymbol{\tau}}, \tag{5.4.23}$$

where  $\mathbf{n} = \mathbf{n}(x, y)$ ,  $\boldsymbol{\tau} = \boldsymbol{\tau}(x, y)$  depend on  $(x, y)$ ,  $F'(|\nabla u|)$  is the conductivity of the diffusion of  $u = u(x, y; t)$  at  $(x, y)$  in the **normal** direction  $\mathbf{n} = \frac{\nabla u}{|\nabla u|}$ , and  $c(|\nabla u|)$  is the conductivity of the diffusion of in the **tangential** direction  $\boldsymbol{\tau} = \boldsymbol{\tau}(x, y)$ .

This observation of the “diffusion geometry” has important applications to digital image noise removal (called “denoising”), while preserving sharpness of image edges. We will elaborate on this application in Subunit 5.5.2, with various examples of the heat conductivity function  $c(p)$ ,  $p \geq 0$ .

### 5.4.3 Solution in terms of eigenvalue problems

We emphasize that it is not feasible, in general, for the non-linear PDE (5.4.9) to have an exact solution. In this subunit, we introduce the notion of lagged anisotropic transform by considering the system of linear PDE's

$$\begin{cases} \frac{\partial}{\partial t} u^n(x, y, t) = \nabla \cdot (c(|\nabla u^{n-1}(x, y, t)|) \nabla u^n(x, y, t)), & (x, y) \in D \\ \frac{\partial}{\partial \mathbf{n}} u^n(x, y, t)|_{\partial D} = 0, & (x, y) \in D; \\ u^n(x, y, 0) = u_0(x, y), & (x, y) \in D, \end{cases} \quad (5.4.24)$$

for  $n = 1, 2, \dots$ , with

$$u^0(x, y, 0) := u_0(x, y), \quad t > 0,$$

or if necessary,  $u_0(x, y)$  should be convolved with some lowpass filter to define the initial function  $u^0(x, y, t)$ . The reason for this extra step is that  $c(|\nabla u^0|)$  must be differentiable for the PDE in (5.4.24), for  $n = 1$ , to be well posed. To solve the linear PDE for each  $n \geq 1$ , we rely on the results derived in Subunit 5.4.1, with  $w = c(|\nabla u^{n-1}|)$  in the Rayleigh quotient (5.4.8). More precisely, while the eigenspace  $S(0) = S(\lambda_0)$  in (5.4.4) is easy to determine (see Example 5.4.1 to follow), the positive eigenvalues  $\lambda_k, k = 1, 2, \dots$ , namely:

$$0 = \lambda_0 < \lambda_1 < \lambda_2 < \dots,$$

can be computed, by solving the minimization problems with

$$\lambda_1 = \min \left\{ \frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2 dx dy} : v \in S^\perp(\lambda_0) \text{ and } \frac{\partial v}{\partial \mathbf{n}}|_{\partial D} = 0 \right\}. \quad (5.4.25)$$

Then by applying the Gram-Schmidt orthonormalization procedure, we have an orthonormal basis  $\{v_{1,k}\}_k$  of  $S(\lambda_1)$ . Hence, by the Theorem 5.4.1 of Subunit 5.4.1, the set of functions  $\{v_{0,k}\} \cup \{v_{1,k}\}$  constitute an orthonormal basis of  $S(\lambda_0) \oplus S(\lambda_1)$ .

In general, for  $j > 1$ , compute

$$\lambda_j = \min \left\{ \frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2 dx dy} : v \in \left( \oplus_{\ell=0}^{j-1} S(\lambda_\ell) \right)^\perp, \quad \frac{\partial v}{\partial \mathbf{n}}|_{\partial D} = 0 \right\}, \quad (5.4.26)$$

with  $S(\lambda_j)$  being its corresponding eigen-space with  $\lambda = \lambda_j$ , with orthonormal basis  $\{v_{j,k}\}_k$ . Then

$$\cup_{j=0}^{\infty} \{v_{j,k}\}_k \quad (5.4.27)$$

is an orthonormal basis of the space its spans, namely

$$\text{clos}_{L^2} \left( \oplus_{j=0}^{\infty} S(\lambda_j) \right) = L_0^2(D),$$

where

$$L_0^2(D) := \left\{ f \in L^2(D) : \frac{\partial f}{\partial \mathbf{n}}|_{\partial D} = 0 \right\}. \quad (5.4.28)$$

That is, the family of  $v_{j,k}$  in (5.4.27) is an orthonormal basis of  $L_0^2(D)$  defined in (5.4.28).

**Example 5.4.1** Let  $D = (0, \pi)^2 = (0, \pi) \times (0, \pi)$  and  $w = c_0$  be a constant. Define

$$h_{m,n}(x, y) = \begin{cases} \frac{1}{\pi}, & \text{for } m = n = 0, \\ \frac{\sqrt{2}}{\pi} \cos mx, & \text{for } m = 1, 2, \dots, \text{ and } n = 0, \\ \frac{\sqrt{2}}{\pi} \cos ny, & \text{for } m = 0 \text{ and } n = 1, 2, \dots, \\ \frac{2}{\pi^2} \cos mx \cos ny, & \text{for } m, n = 1, 2, \dots. \end{cases} \quad (5.4.29)$$

Then  $\{h_{m,n}\}$  is an orthonormal basis of  $L_0^2(0, \pi)^2$ . Let  $\lambda_0 = 0$ . For  $\lambda_1$ , since  $S(\lambda_0)$  is a one-dimensional vector space generated by the constant basis function  $h_{0,0}$ , we must consider

$$\frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2 dx dy}, \quad \text{with } \iint_D v dx dy = 0, \quad (5.4.30)$$

(i.e.  $v \in S^\perp(\lambda_0)$ ) to find the smallest eigenvalue  $\lambda_1 > \lambda_0 (= 0)$ . Let us first observe that

$$\int_0^\pi \cos^2 mx dx = \int_0^\pi \sin^2 mx dx = \frac{\pi}{2}, \quad m \geq 1, \quad (5.4.31)$$

so that for all  $m = 1, 2, \dots$ ,

$$\iint_D w |\nabla h_{m,0}|^2 dx dy = c_0 \int_0^\pi \int_0^\pi \frac{2}{\pi^2} m^2 \sin^2 mx dx dy = c_0 m^2.$$

Similarly,  $\iint_D w |\nabla h_{0,m}|^2 dx dy = c_0 m^2$ . Now, write  $v = \sum_{m,n=0}^\infty a_{m,n} h_{m,n}$ . Since  $v \in S^\perp(\lambda_0)$ ,  $a_{0,0} = 0$ . Therefore, by applying the Parseval's identity to  $v = \sum_{m+n \geq 1}^\infty a_{m,n} h_{m,n}$ , we have

$$\iint_D v^2 dx dy = \sum_{j+k \geq 1}^\infty (a_{j,k})^2.$$

Furthermore, applying Parseval's identity again to  $\nabla v = \sum_{m+n \geq 1}^\infty a_{m,n} \nabla h_{m,n}$  with the orthonormal family  $\{\sin mx \cos ny, \cos mx \sin ny\}$ , we have

$$\iint_D w v^2 dx dy = \sum_{j+k \geq 1}^\infty c_0 (m^2 + n^2) (a_{j,k})^2. \quad (5.4.32)$$

In the following, consider the weights

$$b_{m,n} := \frac{a_{m,n}^2}{\sum_{j+k \geq 1} a_{j,k}^2}, \quad m+n \geq 1$$

which satisfy:

$$0 \leq b_{m,n} \leq 1, \quad \sum_{m+n \geq 1} b_{m,n} = 1. \quad (5.4.33)$$

Therefore, from (5.4.32), we may write

$$\frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2 dx dy} = \sum_{m+n \geq 1} b_{m,n} (c_0(m^2 + n^2)). \quad (5.4.34)$$

In view of (5.4.33), we see that the Rayleigh quotient in (5.4.34) is a **convex combination** of the sequence  $\{c_0(m+n)^2\}$ ,  $m+n \geq 1$ , with weights  $b_{m,n}$ . Thus

$$\lambda_1 = \min_v \frac{\iint_D w |\nabla v|^2 dx dy}{\iint_D v^2 dx dy} = \sum_{b_{m,n}} c_0(m^2 + n^2) = c_0, \quad (5.4.35)$$

which is attained at  $b_{1,0} = 1$  or  $b_{0,1} = 1$ . Therefore, the eigen-space  $S(\lambda_1)$  has dimension = 2, with orthonormal basis  $\{h_{1,0}, h_{0,1}\}$ . This procedure obviously extends to  $\lambda_1 < \lambda_2 < \lambda_3 < \dots$ . In the following, we compile a table of eigenvalues  $0 = \lambda_0 < \lambda_1 < \lambda_2 < \dots$  with dimensions of the eigen-spaces  $S(\lambda_0), S(\lambda_1), S(\lambda_2), \dots$ , and orthonormal bases. See the following table for the first 11 eigenvalues with corresponding orthonormal bases of eigenfunctions.

eigenvalues ( $\lambda$ )	dimension of $S(\lambda)$	orthonormal bases
$\lambda_0 = 0$	$\dim S(\lambda_0) = 1$	$\{h_{0,0}\}$
$\lambda_1 = c_0$	$\dim S(\lambda_1) = 2$	$\{h_{1,0}, h_{0,1}\}$
$\lambda_2 = 2c_0$	$\dim S(\lambda_2) = 1$	$\{h_{1,1}\}$
$\lambda_3 = 4c_0$	$\dim S(\lambda_3) = 2$	$\{h_{2,0}, h_{0,2}\}$
$\lambda_4 = 5c_0$	$\dim S(\lambda_4) = 2$	$\{h_{2,1}, h_{1,2}\}$
$\lambda_5 = 8c_0$	$\dim S(\lambda_5) = 1$	$\{h_{2,2}\}$
$\lambda_6 = 9c_0$	$\dim S(\lambda_6) = 2$	$\{h_{3,0}, h_{0,3}\}$
$\lambda_7 = 10c_0$	$\dim S(\lambda_7) = 2$	$\{h_{3,1}, h_{1,3}\}$
$\lambda_8 = 13c_0$	$\dim S(\lambda_8) = 2$	$\{h_{3,2}, h_{2,3}\}$
$\lambda_9 = 17c_0$	$\dim S(\lambda_9) = 2$	$\{h_{4,1}, h_{1,4}\}$
$\lambda_{10} = 18c_0$	$\dim S(\lambda_{10}) = 1$	$\{h_{3,3}\}$

■  
 We remark that the computation of the eigenvalues (and corresponding eigenfunctions) by means of the minimization criteria (5.4.25)–(5.4.26) applies to an arbitrary simply connected bounded region with piecewise smooth boundary curve, while the Fourier series solution studied in Subunit 5.3.2 only applies to bounded rectangular regions. The above simple example is included here only for the demonstrative purpose. In general,  $w(x, y)$  is not a constant function  $c_0$ , which reduces to the isotropic diffusion PDE studied in Subunit 5.3.

## 5.5 Application to Image De-Noising

Observe that the entropy (of the probability distribution  $\mathbb{P}_n$ ) of an information source increases, often quite significantly, when the information source is contaminated with additive noise. The reason is that the noise random behavior causes the distribution of  $\mathbb{P}_n =: p_1, \dots, p_n$  to be more uniform, in that the contaminated values of  $p_1, \dots, p_n$  are more likely to be different. Since the entropy directly governs coding efficiency according to the Noiseless Coding Theorem, as studied in Subunit 2.4.2, noise reduction certainly facilitates data compression efficiency. When the information source is some (noisy) digital image, quantization reduces the entropy, due to the sparseness of the quantized DCT blocks. From a theoretical point of view, if each image block is represented by a Fourier series, then the Fourier coefficients are divided by some exponentially increasing terms, induced by the diffusion process (as the time parameter increases). In other words, isotropic diffusion performs just like quantization of the DCT coefficients. This concept will be discussed in Subunit 5.5.1. On the other hand, since isotropic diffusion causes uniform image blurring, it would be preferable to replace the quantization process by anisotropic diffusion. In Subunit 5.5.2, four anisotropic diffusion models are introduced and the implementation issue is also discussed. Our study of the application of image de-noising ends with a brief discussion in Subunit 5.5.3 of the pros and cons of such enhancement of the JPEG image compression standard.

### 5.5.1 Diffusion as quantizer for image compression

Recall from Subunit 2.5 that quantization of the discrete cosine transform (DCT) of a digital image is the critical step to achieve high compression ratios at the cost of removing certain high-frequency image content. The quantizers “ $Q$ ”, however, were artificially chosen by means of experimentation and visual

judgment. In the continuous-time setting, if the image to be compressed is represented by a Fourier cosine series, as given by (5.3.15) of Subunit 5.3.2, namely

$$u(x, y, 0) = f(x, y) = \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} d(k, \ell) b_{k, \ell} \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right),$$

then the solution of the isotropic heat diffusion PDE (5.3.12) is given by

$$u(x, y, t) = \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} d(k, \ell) b_{k, \ell} e^{-c\left(\left(\frac{k\pi}{M}\right)^2 + \left(\frac{\ell\pi}{N}\right)^2\right)t} \cos\left(\frac{k\pi}{M}x\right) \cos\left(\frac{\ell\pi}{N}y\right).$$

Hence, it is natural to apply

$$Q = Q(t) := e^{c\left(\left(\frac{k\pi}{M}\right)^2 + \left(\frac{\ell\pi}{N}\right)^2\right)t} \quad (5.5.1)$$

as the quantizer by choosing desirable values of  $t > 0$ .

To apply (5.5.1) to the discrete-time and DCT setting for digital image compression, we may “sample” the solution  $u(x, y, t)$  in (5.3.13) in the spatial domain  $(x, y) \in D$ . Let us first discuss this aspect before considering discretization of the time parameter  $t$  (since this should be done with input from experimentation and visual judgment).

Let  $\Delta x$  and  $\Delta y$  denote the uniform spacings; that is

$$v_k := u_0(a + k\Delta x); \quad v_{k, \ell} := u_0(a + k\Delta x, b + \ell\Delta y)$$

for the domain  $D = (a, c) \times (b, d)$ . Then the A/D (analog-to-digital) converter may be described by

$$\begin{aligned} u_0(x, y) &\longrightarrow \boxed{\text{A/D}} \text{ (with } \Delta x, \Delta y) \longrightarrow \\ &\longrightarrow \begin{bmatrix} v_{0,0} & v_{1,0} & \cdots & v_{M,0} \\ v_{0,1} & v_{1,1} & \cdots & v_{M,1} \\ \vdots & \vdots & \ddots & \vdots \\ v_{0,N} & v_{1,N} & \cdots & v_{M,N} \end{bmatrix} = \mathbf{v}^{M,N}, \end{aligned}$$

where  $\mathbf{v}^{M,N}$  is an  $(N+1) \times (M+1)$  matrix. Recall that to apply the DCT to  $\mathbf{v}^{M,N}$ , the data matrix  $\mathbf{v}^{M,N}$  must be first transposed; that is, we will apply DCT to

$$\mathbf{v}_0 := (\mathbf{v}^{M,N})^T.$$

When the matrix dimension is not clear, we then write  $(\mathbf{v}_0)_{(M+1) \times (N+1)}$  by tacking on the dimension indices. Let us replace the (Fourier) cosine coefficients

$$\begin{aligned} \frac{b_{0,0}}{4}, \frac{b_{0,k}}{2}, \quad k = 1, 2, \dots; \\ \frac{b_{\ell,0}}{2}, \quad \ell = 1, 2, \dots; \end{aligned}$$



and  $b_{\ell,k}, \ell, k = 1, 2, \dots$ , in the continuous-spatial setting (5.3.15) by the discrete-spatial setting:

$$C := \begin{bmatrix} c_{0,0} & c_{0,1} & \cdots & c_{0,N} \\ c_{1,0} & c_{1,1} & \cdots & c_{1,N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M,0} & c_{M,1} & \cdots & c_{M,N} \end{bmatrix}.$$

Then by introducing the DCT matrix:

$$\mathcal{E}_{M+1} = \frac{\sqrt{2}}{\sqrt{M+1}} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & \cdots & \frac{\sqrt{2}}{2} \\ \cos(\frac{\pi}{M+1} \cdot \frac{1}{2}) & \cos(\frac{\pi}{M+1} \cdot \frac{3}{2}) & \cdots & \cos(\frac{\pi}{M+1} \cdot \frac{2M+1}{2}) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(\frac{M\pi}{M+1} \cdot \frac{1}{2}) & \cos(\frac{M\pi}{M+1} \cdot \frac{3}{2}) & \cdots & \cos(\frac{M\pi}{M+1} \cdot \frac{2M+1}{2}) \end{bmatrix}, \quad (5.5.2)$$

we see that the 2-D discrete cosine transform of the discretized version  $\mathbf{v}$  of  $u_0(x, y)$  is given by

$$C = \mathcal{E}_{M+1} \mathbf{v}_0 \mathcal{E}_{N+1}^T, \quad (5.5.3)$$

where

$$\mathbf{v}_0 = (\mathbf{v}_0)_{(M+1) \times (N+1)}.$$

Next, let us consider the “diffusion terms”  $e^{-c(\frac{m\pi}{M+1})^2 t}$  and  $e^{-c((\frac{m\pi}{M+1})^2 + (\frac{n\pi}{N+1})^2)t}$  in (5.3.13) or  $Q^{-1}(t)$  in (5.5.1), and consider the diagonal matrices

$$D_{M+1} := \begin{bmatrix} 1 & & & \\ & e^{-c(\frac{\pi}{M+1})^2 t} & & \\ & & e^{-c(\frac{2\pi}{M+1})^2 t} & \\ & & & \ddots \\ & & & & e^{-c(\frac{M\pi}{M+1})^2 t} \end{bmatrix}$$

and

$$E_{N+1} := \begin{bmatrix} 1 & & & \\ & e^{-c(\frac{\pi}{N+1})^2 t} & & \\ & & e^{-c(\frac{2\pi}{N+1})^2 t} & \\ & & & \ddots \\ & & & & e^{-c(\frac{N\pi}{N+1})^2 t} \end{bmatrix}.$$

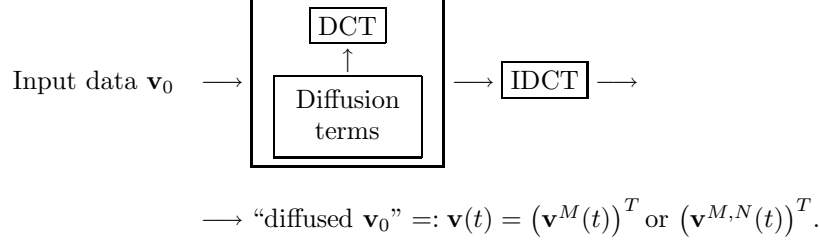
Then quantization of  $C$  in (5.5.3) can be described by

$$\mathbf{C} = \mathcal{E}_{M+1} \mathbf{v}_0 \mathcal{E}_{N+1}^T \longrightarrow D_{M+1} \mathbf{C} E_{N+1}^T = (D_{M+1} \mathcal{E}_{M+1}) \mathbf{v}_0 (E_{N+1} \mathcal{E}_{N+1})^T.$$

We remark that the diffusion process can be pre-computed by

$$\mathcal{E}_{M+1} \longrightarrow D_{M+1} \mathcal{E}_{M+1}, \quad \mathcal{E}_{N+1} \longrightarrow D_{N+1} \mathcal{E}_{N+1}. \quad (5.5.4)$$

In other words, the same “diffusion system” handles **all** (arbitrary) data.



Next, returning to the solution  $u(x, y; t)$  in (5.3.15), we observe that the cosine terms in the continuous-space setting must be replaced by IDCT in the discrete-space setting. Hence, by introducing the notation

$$\mathbf{v}^{M,N}(t) = \begin{bmatrix} v_{0,0}(t) & v_{1,0}(t) & \cdots & v_{M,0}(t) \\ v_{0,1}(t) & v_{1,1}(t) & \cdots & v_{M,1}(t) \\ \vdots & \vdots & \ddots & \vdots \\ v_{0,N}(t) & v_{1,N}(t) & \cdots & v_{M,N}(t) \end{bmatrix}, \quad t \geq 0, \quad (5.5.5)$$

with  $\mathbf{v}^{M,N}(0) = \mathbf{v}^{M,N}$ , we have

$$(\mathbf{v}^{M,N}(t))^T = (\mathcal{E}_{M+1}^T D_{M+1} \mathcal{E}_{M+1}) \mathbf{v}_0 (\mathcal{E}_{M+1}^T D_{M+1} \mathcal{E}_{M+1})^T.$$

Of course, uniform time discretization is to replace  $t$  by  $t_k := t_0 + k\Delta T$ ,  $\Delta T > 0$ , where  $t_0 = 0$ . But non-uniform time discretization should be more desirable, depending on experimentation and human visual judgment.

### 5.5.2 Diffusion for noise reduction

It is well-known that convolution with a lowpass filter is commonly used for (high-frequency) noise reduction, and that the most popular lowpass filter is the Gaussian function. Recall from Subunit 5.1.3 that Gaussian convolution also solves the (isotropic) heat diffusion PDE with spatial domain  $\mathbb{R}^2$ . Hence, for image noise reduction, with image domain  $D = (0, M) \times (0, N) \subset \mathbb{R}^2$ , the solution  $u(x, y, t)$  of the initial-value Neumann problem (5.3.12) provides an image de-noising tool for the given image  $u_0(x, y) = u(x, y, 0)$ . However, removal of high-frequency noise also removes high-frequency image content. To preserve high-frequency content, such as image edges and certain visible textures, anisotropic diffusion should be used. In Subunit 5.4.2, we have derived a general theory of anisotropic heat diffusion PDE with conductivity function

$c(p)$ , where the magnitude of the image gradient  $|\nabla u(x, y, t)|$  is used for  $p$ . The key property is the re-formation of the anisotropic PDE (5.4.9) as

$$\frac{\partial}{\partial t} u = F'(p) u_{\mathbf{n}\mathbf{n}} + c(p) u_{\boldsymbol{\tau}\boldsymbol{\tau}}$$

in (5.4.23), with  $u = u(x, y, t)$  and  $p = |\nabla u|$ ; where the coefficient  $F'(p) = pc(p)$  in (5.4.20) of  $u_{\mathbf{n}\mathbf{n}}$  quantifies the amount of diffusion in the normal direction  $\mathbf{n}$  of any image edge or visible image texture, and  $c(p)$  quantifies the amount of diffusion in the tangential direction  $\boldsymbol{\tau}$ . Hence, since diffusion in the normal direction blurs image edges and removes certain image textures, while diffusion in the tangential direction is much less visible, it is usually wise to select the conductivity function  $c(p) > 0$  such that

$$F'(p) := pc'(p) + c(p)$$

is small (or even negative). Here, negative  $F'(p)$  implies backward diffusion that sharpens image edges for application to image enhancement. In the following, we compile a list of popular choices of the conductivity function  $c(p)$ .

**Example 5.5.1** (TV model)  $c(p)$  is defined by

$$c(p) := \frac{1}{p}. \quad (5.5.6)$$

Thus,  $F(p) = 1$  and hence,  $F'(p) = 0$ . Therefore, the geometry of diffusion for the PDE

$$\frac{\partial}{\partial t} u = \nabla \cdot \left( \frac{\nabla u}{|\nabla u|} \right)$$

is

$$\frac{\partial}{\partial t} u = \frac{1}{|\nabla u|} u_{\boldsymbol{\tau}\boldsymbol{\tau}}, \quad (5.5.7)$$

which implies that the image diffuses only along image edges and no visibly blurring occurs, but the price to pay is that: large  $\frac{1}{|\nabla u|}$  (for small  $|\nabla u|$ , i.e. low-intensity edges and texture) means enormous diffusion, though in the tangential direction, which results in loss of texture. In addition, modification at  $|\nabla u| \approx 0$  (in terms of truncation) is needed for computation and implementation. ■

**Example 5.5.2** (Gaussian model)  $c(p)$  is defined by

$$c(p) := c_0 e^{-p^2/2p_0^2}, \quad (5.5.8)$$

where  $c_0$  and  $p_0$  are positive constants. Then  $F(p) = c_0 p e^{-p^2/2p_0^2}$ . Thus

$$\begin{aligned} F'(p) &= c_0 e^{-p^2/2p_0^2} (1 + p(-p/p_0^2)) \\ &= c_0 e^{-p^2/2p_0^2} (1 - (p/p_0)^2). \end{aligned}$$

From the geometry of diffusion for the PDE given in (5.4.23), if  $|\nabla u(x, y; t)| > p_0$ , then  $F'(p) < 0$ . Thus, the image edges are enhanced (due to backward diffusion). Note: In contrast to the TV model, low-variation content (i. e.  $0 < |\nabla u| \approx 0$ ) is preserved (though slightly diffused in the normal direction, which results in blurring). ■

**Example 5.5.3** (Perona-Malik model)  $c(p)$  is defined by

$$c(p) := \frac{c_0}{1 + (p^2/p_0)^2}, \quad (5.5.9)$$

where  $c_0$  and  $p_0$  are positive constants. Thus,  $F(p) = \frac{c_0 p}{1 + (p^2/p_0)^2}$ , and

$$F'(p) = \frac{c_0(1 - (p^2/p_0)^2)}{(1 + (p^2/p_0)^2)^2}.$$

Note that image edges may be over-sharpened for  $|\nabla u| > p_0$ . ■

**Example 5.5.4** In this example, we consider the conductivity function  $c(p)$  proposed by You, Xu, Tannenbaum and Kavech in IEEE Trans. Image Proc., vol.5, pp.1539-1553, 1996. The design of  $c(p)$  takes full advantage of backward diffusion for edge sharpening, while preserving “isotropic diffusion” for  $|\nabla u| \leq p_0$ .

Let  $\epsilon > 0$  and  $0 < \alpha < 1$  be constants and let

$$c_0 = \frac{1}{p_0} \left( 1 + \frac{\alpha}{(p_0 + \epsilon)^{1-\alpha}} \right)$$

be the conductivity constant for isotropic diffusion.

The conductivity function  $c(p)$  is defined by

$$c(p) := \begin{cases} c_0, & \text{for } 0 \leq p \leq p_0 \\ \frac{1}{p} \left( 1 + \frac{\alpha}{(p+\epsilon)^{1-\alpha}} \right), & \text{if } p_0 < p. \end{cases}$$

where  $p_0 > 0$ . Then

$$F'(p) = \begin{cases} c_0, & \text{for } 0 \leq p \leq p_0 \\ -\frac{\alpha(1-\alpha)}{(p+\epsilon)^{2-\alpha}}, & \text{if } p_0 < p. \end{cases}$$

■

As to the issue of implementation, since images to be de-noised are digital, we are only interested in discretization of the anisotropic PDE (5.4.9). In view of the popularity of the TV model in Example 5.5.1, we only consider the formulation (5.5.7), namely:

$$\frac{\partial}{\partial t} u(x, y, t) = \frac{1}{\sqrt{u_x^2 + u_y^2}} \frac{\partial^2}{\partial \tau^2} u(x, y, t),$$

where  $\boldsymbol{\tau} = \boldsymbol{\tau}(x, y)$  denotes the unit tangent vector at  $(x, y)$ . By (5.4.15), we have (with  $\cos \alpha = \frac{u_y}{|\nabla u|}$  and  $\sin \alpha = \frac{u_x}{|\nabla u|}$  from (5.4.16)),

$$\begin{aligned} \frac{\partial^2}{\partial \boldsymbol{\tau}^2} &= \cos^2 \alpha \frac{\partial^2}{\partial x^2} + 2 \sin \alpha \cos \alpha \frac{\partial^2}{\partial x \partial y} + \sin^2 \alpha \frac{\partial^2}{\partial y^2} \\ &= \frac{1}{u_x^2 + u_y^2} \left( u_y^2 \frac{\partial^2}{\partial x^2} - 2 u_x u_y \frac{\partial^2}{\partial x \partial y} + u_x^2 \frac{\partial^2}{\partial y^2} \right). \end{aligned}$$

Hence, it follows that the anisotropic PDE (5.4.9) (for the TV model) is:

$$\frac{\partial u}{\partial t} u_t = \frac{u_y^2 u_{xx} - 2 u_x u_y u_{xy} + u_x^2 u_{yy}}{(u_x^2 + u_y^2)^{3/2}}. \quad (5.5.10)$$

To discretize the PDE (5.5.10), let  $\Delta t, \Delta x, \Delta y > 0$ . Then for  $\ell = 0, 1, 2, \dots$  and  $j, k \in \mathbb{Z}$  such that  $(j\Delta x, k\Delta y) \in D \cup \partial D$ , set

$$u_\ell(j, k) := u(j\Delta x, k\Delta y, \ell\Delta), \quad (5.5.11)$$

and replace the partial derivatives  $u_t, u_x, u_y, u_{xx}, u_{xy}, u_{yy}$  by partial divided differences  $\delta_t u, \delta_x u, \delta_y u, \delta_{xx} u, \delta_{xy} u, \delta_{yy} u$  respectively, defined by

$$\begin{aligned} (\delta_t u)_\ell(j, k) &:= (u_{\ell+1}(j, k) - u_\ell(j, k)) / \Delta t, \\ (\delta_x u)_\ell(j, k) &:= (u_\ell(j+1, k) - u_\ell(j, k)) / \Delta x, \\ (\delta_y u)_\ell(j, k) &:= (u_\ell(j, k+1) - u_\ell(j, k)) / \Delta y, \\ (\delta_{xx} u)_\ell(j, k) &:= (u_\ell(j+1, k) - 2u_\ell(j, k) + u_\ell(j-1, k)) / (\Delta x)^2, \\ (\delta_{xx} u)_\ell(j, k) &:= (u_\ell(j+1, k+1) - u_\ell(j+1, k) - u_\ell(j, k+1) + \\ &\quad + u_\ell(j, k)) / \Delta x \Delta y, \\ (\delta_{yy} u)_\ell(j, k) &:= (u_\ell(j, k+1) - 2u_\ell(j, k) + u_\ell(j, k-1)) / (\Delta y)^2. \end{aligned} \quad (5.5.12)$$

Therefore, discretization of (5.5.10) becomes

$$\begin{aligned} \delta_t u &= (\delta_t u)_\ell(j, k) \\ &= \left[ \frac{(\delta_y u)^2 (\delta_{xx} u) - 2(\delta_x u) (\delta_y u) (\delta_{xy} u) + (\delta_x u)^2 (\delta_{yy} u)}{((\delta_x u)^2 + (\delta_y u)^2)^{3/2}} \right]_\ell(j, k), \end{aligned} \quad (5.5.13)$$

for  $\ell = 0, 1, \dots$ , and for  $(j\Delta x, k\Delta y) \in D \cup \partial D$ .

### 5.5.3 Enhanced JPEG compression

We conclude this unit (Unit 5) by proposing an improvement of the JPEG compression standard studied in Subsection 2.5. Recall that other than the  $8 \times 8$  DCT blocks, the quantization step discussed in (2.5.1)–(2.5.2) of Subunit 2.5 could benefit from the theoretical approach of isotropic diffusion studied

in Subunit 5.5.1. Unfortunately, as pointed out in Subunit 5.5.2, isotropic diffusion necessarily blurs the original images. On the other hand, since random noise could be significantly reduced from the diffusion process, the entropy of the diffused DCT is much smaller in general. This ensures smaller compressed file size, since the size of the Huffman code, as studied in Subunit 2.5.3 is governed by the “noiseless coding” theorem of Shannon (see Theorem 2.4.2).

Of course, as a digital image compressing standard, all JPEG compressed images must be recoverable (i.e. decompressed and open) by the existing devices; that is, by using the existing software or hardware devices. This restriction significantly limits the freedom in modifying the JPEG compression scheme. One improvement which is not restricted by the JPEG standard is image pre-processing (before JPEG compression is to be applied). To reduce the entropy, image enhancement by reducing noise and sharpening image edges is feasible by applying anisotropic diffusion as studied in Subunit 5.5.2. If further reduction of compressed file size is desired by the user, as improved quantizer, as studied in Subunit 5.5.1, can be applied to replace the limited JPEG quantization tables. Of course, since the modified quantizer must be sent to the receiver, the application of this recommendation is somewhat limited.

# *Unit 6*

## *WAVELET METHODS*

This final unit is concerned with the study of multi-scale data analysis and wavelet transform, with emphasis on construction of compactly supported wavelets, development of wavelet algorithms, and application to image coding. The continuous wavelet transform (CWT) is introduced and the relation of scale and frequency is described by using a high-pass filter. For the CWT, an inner product on the time-scale space is introduced, and Parseval's identity for this inner-product space is derived, with application to introducing the inverse continuous transform (ICWT). For dyadic discretization of the CWT, resulting in the discrete wavelet transform (DWT), the notion of multiresolution analysis (MRA) is introduced to give an effective architecture, both for wavelet construction and for DWT algorithm development. Construction of compactly supported wavelets is achieved by the method of matrix extension, to be studied in some depth in Subunit 6.3. Wavelet algorithms to be studied include derivation of the wavelet decomposition and reconstruction algorithms, extension of these algorithms to tensor products, and the lifting scheme. This unit ends with a study of embedding a digital image in the wavelet-domain for image manipulation, such as progressive transmission, and the lossless JPEG-2000 digital image compression standard.

### **6.1 Time-Scale Analysis**

The continuous wavelet transform (CWT), defined by translation and dilation (also called scaling) of some convolution wavelet kernel, is introduced in Subunit 6.1.1. The importance of the (positive) scaling parameter is that it can be viewed as the reciprocal of the frequency of the signal (represented by a function) being analyzed by the CWT. This concept is illustrated by some ideal high-pass filter in Subunit 6.1.2. Since a (finite) frequency band can be partitioned into a disjoint union of ideal high-pass bands, the discussion in Subsection 6.1.2 is extended in Subunit 6.1.3 to introduce the notion of the discrete wavelet transform (DWT). A class of admissible wavelets is introduced in Subunit 6.1.4. In terms of each admissible wavelet, an inner product operation is introduced and Parseval's identity for the correspond-

ing inner-product space is established. Furthermore, as an application of this identity, the inverse continuous wavelet transform (ICWT) is derived and the reproducing kernel (again in term of the given admissible wavelet) for this inner-product space is formulated in Subunit 6.1.4.

### 6.1.1 Wavelet transform

Let  $\psi \in L_2(\mathbb{R})$  be a continuous function such that  $\psi(t) \rightarrow 0$  for  $t \rightarrow \pm\infty$ , for which the Cauchy principal value of the integral of  $\psi$  on  $\mathbb{R}$  exists and vanishes; that is,

$$\text{PV} \int_{-\infty}^{\infty} \psi(t) dt = \lim_{A \rightarrow \infty} \int_{-A}^A \psi(t) dt = 0. \quad (6.1.1)$$

Then  $\psi$  is called a “wavelet.” For any given wavelet  $\psi$ , by introducing two real numbers  $b \in \mathbb{R}$  and  $a > 0$ , we have a two-parameter family of functions

$$\psi_{b,a}(t) = \frac{1}{a} \psi\left(\frac{t-b}{a}\right), \quad (6.1.2)$$

called “wavelets generated by  $\psi$ .” The word “wavelets” means “small waves.” In view of (6.1.1), the graph of  $\psi(t)$  is oscillating (and hence, “wavy”); and this wave dies down to 0, since  $\psi(t) \rightarrow 0$  as  $|t| \rightarrow \infty$ . Moreover, observe that  $\psi_{b,a}(t)$  zooms in to a smaller region near  $t = b$  as the positive parameter  $a$  tends to 0. Therefore the graphs of  $\psi_{b,a}(t)$  are indeed small or large waves, depending on small values or large values of the parameter  $a > 0$ ; and this family of wavelets covers the entire “time-domain”  $\mathbb{R} = (-\infty, \infty)$  as  $b$  runs over  $\mathbb{R}$ .

The wavelets  $\psi_{b,a}(t)$  defined in (6.1.2) have both the localization and oscillation features for the analysis of (input) functions  $f \in L_2(\mathbb{R})$ , when used as the “integration kernel” for the (continuous) **wavelet transform** (CWT) of  $f(t)$ , defined by

$$(W_\psi f)(b, a) = \langle f, \psi_{b,a} \rangle = \frac{1}{a} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt \quad (6.1.3)$$

for analyzing the oscillation behavior of  $f(t)$ . The localization feature is easy to understand, since  $\psi(t)$  is a window function already. However, it is important to point out that the window size (as defined in terms of the width  $2\Delta_\psi$  defined in (4.4.2) of Subunit 4.4.1) varies since

$$\Delta_{\psi_{b,a}} = a\Delta_\psi \text{ and } \Delta_{\widehat{\psi_{b,a}}} = \frac{1}{a}\Delta_{\widehat{\psi}}.$$

Hence, the wavelet transform  $(W_\psi f)(b, a)$  of  $f(t)$  zooms in, as the time-window width  $2\Delta_{\psi_{b,a}} = 2a\Delta_\psi$  narrows (for smaller values of  $a > 0$ , with wider frequency-window) and zooms out as the window width widens (for larger



values of  $a > 0$ , with narrower frequency-window for analyzing high-frequency contents).

Next, we must understand the feature of oscillation, or frequency. Since the translation parameter  $b$  has nothing to do with oscillation, the frequency must be governed by the scale parameter  $a > 0$  as well.

For this purpose, let us consider the single frequency signal  $f_\omega(t) = e^{i\omega t}$  with frequency  $= \omega/2\pi$  Hz (where  $\omega$  is fixed). Although  $f_\omega$  is not in  $L_2(\mathbb{R})$ , the inner product in (6.1.3) is still well-defined (for any wavelet  $\psi \in L_2(\mathbb{R})$ , since

$$\overline{\langle f_\omega, \psi_{b,a} \rangle} = \int_{-\infty}^{\infty} \psi_{b,a}(t) e^{-it\omega} dt$$

is the Fourier transform of the function  $\psi_{b,a} \in L_2(\mathbb{R})$ . Hence, we have, from (6.1.3), that

$$\begin{aligned} (W_\psi f_\omega)(b, a) &= \langle f_\omega, \psi_{b,a} \rangle = \overline{(\mathbb{F}\psi_{b,a})}(\omega) \\ &= \frac{1}{a} \int_{-\infty}^{\infty} \bar{\psi}\left(\frac{t-b}{a}\right) e^{i\omega t} dt \\ &= e^{i\omega b} \widehat{\bar{\psi}}(a\omega). \end{aligned} \tag{6.1.4}$$

### 6.1.2 Frequency versus scale

Let us apply (6.1.4) to explore the relationship between the notions of “frequency”  $\omega$  and “scale”  $a$ , by considering an appropriate wavelet  $\psi$  whose Fourier transform is formulated as the difference of two ideal lowpass filters.

First consider a pure-tone signal

$$g_{\omega_0}(t) = d_0 \cos \omega_0 t,$$

with amplitude  $= d_0 \neq 0$  and frequency  $= \frac{1}{2\pi}\omega_0$  Hz. Let  $h_\eta$ , defined by

$$h_\eta(t) = \frac{\sin \eta t}{\pi t}, \quad \eta > 0,$$

be the ideal lowpass filter with Fourier transform given by

$$\widehat{h}_\eta(\omega) = \chi_{[-\eta, \eta]}(\omega), \tag{6.1.5}$$

and consider the function  $\psi_\epsilon(t)$  defined by

$$\psi_\epsilon(t) = h_{1+\epsilon}(t) - h_{1-\epsilon}(t),$$

with Fourier transform given by

$$\widehat{\psi}_\epsilon(\omega) = \chi_{[-1-\epsilon, -1+\epsilon]}(\omega) + \chi_{[1-\epsilon, 1+\epsilon]}(\omega)$$

by applying (6.1.5). Since  $0 < \epsilon < 1$ , we have  $\widehat{\psi}_\epsilon(0) = 0$ , or equivalently,

$$\int_{-\infty}^{\infty} \psi_\epsilon(t) dt = \widehat{\psi}_\epsilon(0) = 0,$$

so that  $\psi_\epsilon(t)$  is a wavelet. Now, applying (6.1.4), we have

$$\begin{aligned} (W_{\psi_\epsilon} g_{\omega_0})(b, a) &= \frac{1}{2} d_0 \left( W_{\psi_\epsilon} e^{i\omega_0 t} \right)(b, a) + \frac{1}{2} d_0 \left( W_{\psi_\epsilon} e^{-i\omega_0 t} \right)(b, a) \\ &= \frac{1}{2} d_0 \left( e^{i\omega_0 b} + e^{-i\omega_0 b} \right) \overline{\widehat{\psi}_\epsilon(a\omega_0)} \\ &= d_0 (\cos \omega_0 b) \left( \chi_{[-1-\epsilon, -1+\epsilon]} + \chi_{[1-\epsilon, 1+\epsilon]} \right)(a\omega_0) \\ &= d_0 (\cos \omega_0 b) \chi_{[1-\epsilon, 1+\epsilon]}(a\omega_0), \end{aligned} \quad (6.1.6)$$

since  $a\omega_0 > 0$  for positive values of  $\omega_0$ . An equivalent formulation of (6.1.6) is that

$$(W_{\psi_\epsilon} g_{\omega_0})(b, a) = 0, \text{ for } |a\omega_0 - 1| > \epsilon$$

and

$$(W_{\psi_\epsilon} g_{\omega_0})(b, a) = d_0 (\cos \omega_0 b) = g_{\omega_0}(b), \text{ for } |a\omega_0 - 1| < \epsilon$$

(where we have ignored the consideration of  $|a\omega_0 - 1| = \epsilon$  for convenience). Hence, for small values of  $\epsilon > 0$ , we see that the relation of the scale  $a$  and frequency  $\omega_0$  is

$$\frac{1}{a} \approx \omega_0, \quad (6.1.7)$$

and that the wavelet transform  $\mathbb{W}_{\psi_\epsilon}$  is an ideal band-pass filter which preserves the signal content with frequencies in some  $\epsilon$ -neighborhood of  $\frac{1}{a}$ . In view of (6.1.7), it is customary to say that the scale  $a$  is inversely proportional to the frequency, although this statement is somewhat misleading and only applies to the wavelet  $\psi_\epsilon(t)$ .

### 6.1.3 Partition into frequency bands

The relationship as illustrated in (6.1.7) between the scale  $a > 0$  and frequency can be extended from a single frequency to a range of frequencies, called “frequency band”, with bandwidth determined by  $\Delta_{\widehat{\psi}}$ . (Observe that  $\Delta_{\widehat{\psi}_\epsilon} \rightarrow 0$  as  $\epsilon \rightarrow 0$ .) To illustrate this concept, let us again consider some wavelet  $\psi$  as the difference of two ideal lowpass filters in the following example.

**Example 6.1.1** Apply (6.1.4) to explore the relationship between the scale  $a > 0$  and “frequency bands”  $[d^j, d^{j+1})$  for any  $d > 1$  and all integers,  $j = 0, 1, 2, \dots$ , by considering some wavelet  $\psi$  as the difference of two appropriate ideal lowpass filters.

**Solution** Let us again consider the ideal lowpass filter  $h_\eta(t)$  with Fourier transform  $\hat{h}_\eta(\omega) = \chi_{[-\eta, \eta]}(\omega)$  as in (6.1.5), but consider the wavelet

$$\psi_I(t) = h_d(t) - h_1(t), \quad (6.1.8)$$

where  $d > 1$ , so that

$$\hat{\psi}_I(\omega) = \chi_{[-d, -1]}(\omega) + \chi_{(1, d]}(\omega) \quad (6.1.9)$$

is an “ideal” bandpass filter, with pass-band  $=[-d, -1] \cup (1, d]$ . Then for a multi-frequency signal

$$g(t) = \sum_{k=0}^n c_k \cos kt \quad (6.1.10)$$

with DC (direct current) term  $c_0$ , and AC (alternating current) terms with amplitudes  $c_k$ , for the frequency components of  $k/2\pi$  Hz, respectively, for  $k = 1, \dots, n$ , it follows from the same computation as in Subunit 6.1.2 that

$$(W_{\psi_I} g)(b, a) = \sum_{k=0}^n c_k (\cos kb) \chi_{[1, d]}(ak).$$

In particular, for each  $j = 0, \dots, \lfloor \log_d n \rfloor$ ,

$$\begin{aligned} (W_{\psi_I} g)\left(b, \frac{1}{d^j}\right) &= \sum_{k=0}^n c_k (\cos kb) \chi_{[1, d]}(ak) \left(\frac{k}{d^j}\right) \\ &= \sum_{d^j \leq k < d^{j+1}} c_k \cos kb. \end{aligned} \quad (6.1.11)$$

That is,  $(W_{\psi_I} g)(b, d^{-j})$  is precisely the restriction of the given signal  $g(t)$  on the “frequency band”  $[d^j, d^{j+1})$ , where the time variable  $t$  of  $g(t)$  is replaced by the translation parameter  $b$ .

To capture the dc term  $c_0$ , we simply use the lowpass filter  $h_1(t)$  that generates the wavelet  $\psi(t)$ , again by taking the inner product, namely,

$$\begin{aligned} \langle g, h_1 \rangle &= \int_{-\infty}^{\infty} g(t) \overline{h_1(t)} dt = \int_{-\infty}^{\infty} g(t) h_1(t) dt \\ &= \frac{1}{2} \sum_{k=0}^n c_k \left( \int_{-\infty}^{\infty} h_1(t) e^{ikt} dt + \int_{-\infty}^{\infty} h_1(t) e^{-ikt} dt \right) \\ &= \frac{1}{2} \sum_{k=0}^n c_k \left( \chi_{(-1, 1)}(-k) + \chi_{(-1, 1)}(k) \right) \\ &= \frac{1}{2} c_0 (1 + 1) = c_0. \end{aligned} \quad (6.1.12)$$

This result, together with (6.1.11), gives rise to the following decomposition of

the given signal  $g(t)$  into the frequency bands  $[0, 1), [1, d), [d, d^2), \dots$ , namely:

$$\begin{aligned} g(t) &= \langle g, h_1 \rangle + \sum_{j=0}^{\lfloor \log_d n \rfloor} (W_{\psi_I} g)\left(t, \frac{1}{d^j}\right) \\ &= \langle g, h_1 \rangle + \sum_{j=0}^{\lfloor \log_d n \rfloor} \langle g, \psi_{t, d^{-j}}^I \rangle, \end{aligned} \quad (6.1.13)$$

where

$$\psi_{t, d^{-j}}^I(x) := d^j \psi_I(d^j(x - t)),$$

and

$$\langle g, \psi_{t, d^{-j}}^I \rangle = d^j (W_{\psi_I} g)(t, d^{-j})$$

as introduced in (6.1.2) and (6.1.3), respectively, for  $b = t$  and  $a = d^{-j}$ . ■

We remark that the decomposition formula (6.1.13) derived in Example 6.1.1 should be considered only as an illustration of the concept of wavelet decomposition of signals into frequency sub-bands. For computational efficiency, the translation parameter  $b$  is also discretized, namely:  $b = k/d^j$ , so that for  $a = d^{-j}$ , we have

$$\psi_{b, a}^I(x) = d^j \psi_I\left(d^j\left(x - \frac{k}{d^j}\right)\right) = d^j \psi_I(d^j x - k) \quad (6.1.14)$$

and

$$(W_{\psi_I} f)(b, a) = (W_{\psi_I} f)\left(\frac{k}{d^j}, \frac{1}{d^j}\right) = d^j \int_{-\infty}^{\infty} f(t) \overline{\psi_I(d^j t - k)} dt. \quad (6.1.15)$$

#### 6.1.4 Parseval's identity for wavelet transform

In this subunit, we introduce “Parseval's formula” that will be used to derive the inverse wavelet transform later in Subunit 6.1.5.

**Definition 6.1.1** Let  $\mathbb{R}_+^2$  denote the upper-half plane  $(-\infty, \infty) \times (0, \infty)$ . Then for  $F(b, a)$  and  $G(b, a)$  with  $\frac{1}{a}F(b, a), \frac{1}{a}G(b, a) \in L_2(\mathbb{R}_+^2)$ , the inner product  $\langle F, G \rangle_W$  is defined by

$$\langle F, G \rangle_W = \int_0^\infty \left\{ \int_{-\infty}^\infty F(b, a) \overline{G(b, a)} db \right\} \frac{da}{a}. \quad (6.1.16)$$

Furthermore, the vector space with inner product defined by (6.1.16) will be denoted by  $L_2(\mathbb{R}_+^2, \frac{db da}{a})$  and

$$\|F\|_W = \sqrt{\langle F, F \rangle_W}.$$

**Definition 6.1.2** A wavelet  $\psi \in L_2(\mathbb{R})$  is said to be admissible, if its Fourier transform  $\widehat{\psi}$  satisfies

$$C_\psi = \int_0^\infty \frac{|\widehat{\psi}(\omega)|^2}{\omega} d\omega < \infty. \quad (6.1.17)$$

**Theorem 6.1.1** Let  $\psi \in L_2(\mathbb{R})$  be an admissible wavelet as introduced in Definition 6.1.2. Then for any  $f \in L_2(\mathbb{R})$ , the wavelet transform  $(W_\psi f)(b, a)$  is in  $L_2\left(\mathbb{R}_+^2, \frac{db da}{a}\right)$ .

**Proof** Since both  $f$  and  $\psi$  are in  $L_2(\mathbb{R})$  and

$$\widehat{\psi}_{b,a}(\omega) = \widehat{\psi}(a\omega) e^{-ib\omega},$$

it follows from Parseval's formula for Fourier transform that

$$(W_\psi f)(b, a) = \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega) \overline{\widehat{\psi}(a\omega)} e^{ib\omega} d\omega. \quad (6.1.18)$$

Hence, by introducing the notation

$$\widehat{F}_a(\omega) = \widehat{f}(\omega) \overline{\widehat{\psi}(a\omega)}, \quad (6.1.19)$$

which is an  $L_1(\mathbb{R})$  function in view of the Cauchy-Schwarz inequality, we may conclude that

$$F_a(b) = \left(\mathbb{F}^{-1} \widehat{F}_a\right)(b) = (\mathbb{F}^\# \widehat{F}_a)(b)$$

is well defined, with

$$F_a(b) = (W_\psi f)(b, a)$$

almost everywhere by (6.1.19). Hence, we have

$$\begin{aligned}
& \int_0^\infty \left\{ \int_{-\infty}^\infty |(W_\psi f)(b, a)|^2 db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \int_{-\infty}^\infty |F_a(b)|^2 db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty |\widehat{F}_a(\omega)|^2 d\omega \right\} \frac{da}{a} \\
&= \frac{1}{2\pi} \int_0^\infty \left\{ \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 |\widehat{\psi}(a\omega)|^2 d\omega \right\} \frac{da}{a} \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 \left\{ \int_0^\infty |\widehat{\psi}(a\omega)|^2 \frac{da}{a} \right\} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty |\widehat{f}(\omega)|^2 \left\{ \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi \right\} d\omega \\
&= C_\psi \frac{1}{2\pi} \|\widehat{f}\|_2^2 = C_\psi \|f\|_2^2 < \infty.
\end{aligned}$$

■

We are now ready to derive Parseval's formula for the wavelet transform, as follows.

**Theorem 6.1.2** *Let  $\psi \in L_2(\mathbb{R})$  be an admissible wavelet as defined by (6.1.17). Then*

$$\langle W_\psi f, W_\psi g \rangle_W = C_\psi \langle f, g \rangle, \quad (6.1.20)$$

for all  $f, g \in L_2(\mathbb{R})$ , where the inner product  $\langle \cdot, \cdot \rangle_W$  is defined in (6.1.16) and the constant  $C_\psi$  is defined in (6.1.17).

**Proof** To prove this theorem, we first observe that the left-hand side of (6.1.20) is well defined and finite by applying Theorem 6.1.1 and the Cauchy-Schwarz inequality for the inner product  $\langle \cdot, \cdot \rangle_W$ . Hence, by introducing the notation  $\widehat{F}_a(\omega) = \widehat{f}(\omega) \widehat{\psi}(a\omega)$  and  $\widehat{G}_a(\omega) = \widehat{g}(\omega) \widehat{\psi}(a\omega)$  as in (6.1.19) and observing that they are  $L_2(\mathbb{R})$  functions with inverse Fourier transform given by

$$F_a(b) = (W_\psi f)(b, a), \quad G_a(b) = (W_\psi g)(b, a)$$

almost everywhere, respectively, we may apply Fubini's theorem to compute

$$\begin{aligned}
\langle W_\psi f, W_\psi g \rangle_W &= \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)(b, a) \overline{(W_\psi g)(b, a)} db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \int_{-\infty}^\infty F_a(b) \overline{G_a(b)} db \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{F}_a(\omega) \overline{\widehat{G}_a(\omega)} d\omega \right\} \frac{da}{a} \\
&= \int_0^\infty \left\{ \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega) \overline{\widehat{g}(\omega)} |\widehat{\psi}(a\omega)|^2 d\omega \right\} \frac{da}{a} \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega) \overline{\widehat{g}(\omega)} \left\{ \int_0^\infty |\widehat{\psi}(a\omega)|^2 \frac{da}{a} \right\} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega) \overline{\widehat{g}(\omega)} \left\{ \int_0^\infty \frac{|\widehat{\psi}(\xi)|^2}{\xi} d\xi \right\} d\omega \\
&= C_\psi \frac{1}{2\pi} \int_{-\infty}^\infty \widehat{f}(\omega) \overline{\widehat{g}(\omega)} d\omega \\
&= C_\psi \langle f, g \rangle.
\end{aligned}$$

■

### 6.1.5 Inverse wavelet transform

The objective of this subunit is to derive the formula for recovering  $f(x)$  from its wavelet transform  $(W_\psi f)(b, a)$ .

Let  $g_\sigma(x)$  be the (normalized) Gaussian function defined in (4.2.1) of Subunit 4.2.2 and recall that for any  $f \in L_\infty(\mathbb{R})$ ,

$$(f * g_\sigma)(x) \rightarrow f(x) \quad (6.1.21)$$

as  $0 < \sigma \rightarrow 0$  at each  $x \in \mathbb{R}$  where  $f$  is continuous. For each  $x \in \mathbb{R}$  where both  $f(t)$  and  $\psi(\frac{t-b}{a})$  are continuous at  $t = x$ , in addition to (6.1.21), we also have

$$\begin{aligned}
(W_\psi g_\sigma(x - \cdot))(b, a) &= \frac{1}{a} \int_{-\infty}^\infty g_\sigma(x - t) \overline{\psi\left(\frac{t-b}{a}\right)} dt \\
&\rightarrow \frac{1}{a} \overline{\psi\left(\frac{x-b}{a}\right)} = \bar{\psi}_{b,a}(x) \quad (6.1.22)
\end{aligned}$$

as  $0 < \sigma \rightarrow 0$ , which yields

$$\begin{aligned}
 & \langle W_\psi f, W_\psi g_\sigma(x - \cdot) \rangle_W \\
 &= \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)(b, a) \overline{(W_\psi g_\sigma(x - \cdot))}(b, a) db \right\} \frac{da}{a} \\
 &\rightarrow \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)(b, a) \psi_{b,a}(x) db \right\} \frac{da}{a} \\
 &= \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)(b, a) \psi\left(\frac{x-b}{a}\right) db \right\} \frac{da}{a^2}.
 \end{aligned}$$

This, together with (6.1.21), yields the following result on the inverse wavelet transform (IWT).

**Theorem 6.1.3** *Let  $\psi \in L_2(\mathbb{R})$  satisfy (6.1.17). Then for all  $f \in (L_2 \cap L_\infty)(\mathbb{R})$ ,*

$$\begin{aligned}
 f(x) &= \frac{1}{C_\psi} \int_0^\infty \left\{ \int_{-\infty}^\infty (W_\psi f)(b, a) \psi\left(\frac{x-b}{a}\right) db \right\} \frac{da}{a^2} \\
 &= \frac{1}{C_\psi} \int_0^\infty \left\{ \int_{-\infty}^\infty \langle f, \psi_{b,a} \rangle \psi_{b,a}(x) db \right\} \frac{da}{a}
 \end{aligned} \tag{6.1.23}$$

almost everywhere, where  $C_\psi$  is given by (6.1.17).

To write out (6.1.23) without using the notation  $\psi_{b,a}$ , we may set

$$K(x, t, b, a) = \frac{1}{a^3} \psi\left(\frac{x-b}{a}\right) \overline{\psi\left(\frac{t-b}{a}\right)} \tag{6.1.24}$$

and apply Fubini's theorem to re-formulate (6.1.23) as

$$f(x) = \int_0^\infty \left\{ \int_{-\infty}^\infty \int_{-\infty}^\infty f(t) K(x, t, b, a) dt db \right\} da, \tag{6.1.25}$$

which is a reconstruction (or reproduction) formula for  $(L_2 \cap L_\infty)(\mathbb{R})$ , with wavelet kernel  $K(x, t, b, a)$  given in (6.1.24), that may be considered as a “reproducing kernel”.

## 6.2 Multiresolution Analysis (MRA)

In this subunit we introduce the notion of multiresolution analysis (MRA) and show how MRA leads to the study of wavelet decomposition. This subunit



also serves as an introduction of the MRA method for the construction of compactly supported wavelets, which is studied in some depth in Subunit 6.3. To be more specific, in Subunit 6.2.1, the notion of function refinement, along with refinement equation, scaling function, two-scale relation and two-scale symbol, is introduced, with the *sinc* function (from the Sampling theorem studied in Subunit 4.1.3) as a demonstrative example. To derive compactly supported refinable (or more precisely, scaling) functions, the characteristic function of the unit interval is used to introduce (polynomial) *B*-splines of any positive order in Subunit 6.2.2. The multiresolution (MRA) architecture, based on an arbitrary scaling function, is formally described in Subunit 6.2.3, along with justifications in terms of certain low-pass and high-pass filters that are generated by using the *sinc* function from Shannon's Sampling theorem.

### 6.2.1 Function refinement

In most applications, particularly in signal and image processing, only band-limited functions are of interest. Hence, although the general theory and method of MRA will be studied in this section and the next two chapters under the framework of  $L_2 = L_2(\mathbb{R})$ , we will first introduce the concept of MRA and that of the corresponding wavelet bandpass filters, by considering ideal lowpass and ideal bandpass filters, to apply to all band-limited functions.

Let  $\widehat{\phi}_S(\omega) = \chi_{[-\pi, \pi]}(\omega)$  denote the ideal lowpass filter. Here, the subscript  $S$  of  $\phi_S$  stands for both Claude Shannon and "sinc," since  $\phi_S$  is the sinc function in Shannon's Sampling theorem studied in Subunit 4.1.3. Then for any band-limited function  $f(x)$ , there exists a (sufficiently large) positive integer  $J$  such that  $\widehat{f}(\omega)$  vanishes outside the interval

$$[-2^J\pi, 2^J\pi].$$

Hence, the ideal lowpass filter  $\widehat{\phi}_S(2^{-J}\omega)$  becomes an allpass filter of all functions including  $f(x)$ , with bandwidth  $\leq 2^{J+1}\pi$ ; that is,

$$\widehat{f}(\omega)\widehat{\phi}_S(2^{-J}\omega) = \widehat{f}(\omega)$$

for all  $\omega$ , or equivalently

$$(f * \phi_{S,J})(x) = f(x)$$

for all  $x$ , where

$$\phi_{S,J}(x) = 2^J \phi(2^J x)$$

and the function  $\phi_S$ , given by

$$\phi_S(x) = \text{sinc } x := \frac{\sin \pi x}{\pi x}, \quad (6.2.1)$$

is the inverse Fourier transform of  $\widehat{\phi}(\omega) = \chi_{[-\pi, \pi]}(\omega)$ . On the other hand,

it follows from the Sampling Theorem discussed in Subunit 4.1.3 that such functions  $f(x)$  (with bandwidth not exceeding  $2^{J+1}\pi$ ) can be recovered from its discrete samples  $f(\frac{k}{2^J})$ ,  $k \in \mathbb{Z}$ , via the formula

$$\begin{aligned} f(x) &= \sum_{k=-\infty}^{\infty} f\left(\frac{k}{2^J}\right) \frac{\sin \pi(2^J x - k)}{\pi(2^J x - k)} \\ &= \sum_{k=-\infty}^{\infty} c_k^J \phi(2^J x - k), \end{aligned} \quad (6.2.2)$$

by applying (6.2.1), where

$$c_k^J = f\left(\frac{k}{2^J}\right).$$

Since the function  $f(x) = \phi_S(2^{J-1}x)$  has bandwidth  $= 2^J\pi$ , it follows from (6.2.2) that

$$\phi_S(2^{J-1}x) = \sum_{k=-\infty}^{\infty} c_k^J \phi_S(2^J x - k), \quad (6.2.3)$$

where, in view of (6.2.1),

$$c_k^J = \phi_S\left(\frac{2^{J-1}k}{2^J}\right) = \phi_S\left(\frac{k}{2}\right) = \frac{\sin(\pi k/2)}{k\pi/2},$$

which is independent of  $J$ . Hence, we may introduce the sequence  $\{p_{S,k}\}$ , defined by

$$p_{S,k} = \frac{\sin(\pi k/2)}{\pi k/2} = \begin{cases} \delta_j, & \text{for } k = 2j, \\ \frac{(-1)^j 2}{(2j+1)\pi}, & \text{for } k = 2j+1, \end{cases} \quad (6.2.4)$$

for all  $j \in \mathbb{Z}$ , and replace  $2^{J-1}x$  by  $x$  in (6.2.3) to obtain the identity

$$\phi_S(x) = \sum_{k=-\infty}^{\infty} p_{S,k} \phi_S(2x - k), \quad (6.2.5)$$

to be called the **two-scale relation** or **refinement equation**, with the governing sequence  $\{p_{S,k}\}$  in (6.2.4) called the corresponding **two-scale sequence** or **refinement sequence** of  $\phi_S(x)$ . We will also say that  $\phi_S(x)$  is a **refinable function**.

For each  $j \in \mathbb{Z}$ , let

$$\mathbb{V}_j = \overline{\text{span}} \left\{ \phi_S(2^j x - k) : k \in \mathbb{Z} \right\} \quad (6.2.6)$$

be the  $L_2$ -closure of the (linear) algebraic span of  $\{\phi(2^j x - k)\}$ . Then it follows from the two-scale relation (6.2.5) that the family  $\{\mathbb{V}_j\}$  of vector spaces is a nested sequence

$$\cdots \subseteq \mathbb{V}_{-1} \subseteq \mathbb{V}_0 \subseteq \mathbb{V}_1 \subseteq \mathbb{V}_2 \subseteq \cdots, \quad (6.2.7)$$

of  $L_2(\mathbb{R})$ .

In general, if  $\phi$  is a refinable function with refinement sequence  $\{p_k\}$ ; that is,

$$\phi(x) = \sum_k p_k \phi(2x - k) \quad (6.2.8)$$

for all  $x$ , where the sequence may be finite or infinite, then by taking the Fourier transform of both sides of (6.2.8), we have

$$\widehat{\phi}(\omega) = P(e^{-i\frac{\omega}{2}}) \widehat{\phi}\left(\frac{\omega}{2}\right), \quad (6.2.9)$$

where  $P(x)$  is called the **two-scale symbol** of  $\{p_k\}$ , defined by

$$P(z) = \frac{1}{2} \sum_k p_k z^k, \quad (6.2.10)$$

with  $z = e^{-i\omega/2}$ .

### 6.2.2 B-spline examples

The ideal lowpass filter function  $\phi_S(x)$  in (6.2.1) is a refinable function, as described by the refinement relation (6.2.5), but its refinement sequence  $\{p_{S,k}\}$  (6.2.5) is infinite. In this subunit, we introduce a family of refinable functions with finite refinement sequences.

The first is the obvious example  $\varphi_1(x) := \chi_{[0,1)}(x)$ , the characteristic function of the unit interval  $[0, 1)$ , as follows.

**Example 6.2.1** Let

$$\varphi_1(x) = \begin{cases} 1 & \text{for } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (6.2.11)$$

Then

$$\varphi_1(2x) = \begin{cases} 1 & \text{for } 0 \leq x < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\varphi_1(2x - 1) = \begin{cases} 1 & \text{for } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (6.2.12)$$

Hence,

$$\varphi_1(x) = \varphi_1(2x) + \varphi_1(2x - 1) \quad (6.2.13)$$

for all  $x$ . That is,  $\varphi_1(x)$  is a refinable function, with refinement sequence  $\{p_k\}$  given by

$$p_0 = p_1 = 1$$

and  $p_k = 0$  for all  $k \neq 0, 1$ . ■

Following (6.2.6), consider

$$\mathbb{V}_j = \overline{\text{span}} \left\{ \varphi_1(2^j x - k) : k \in \mathbb{Z} \right\}.$$

Then the family  $\{\mathbb{V}_j\}$  of vector spaces is also a nested sequence of subspaces of  $L_2(\mathbb{R})$  as in (6.2.7).

**Example 6.2.2** Let  $\varphi_1(x) = \chi_{[0,1)}(x)$  as introduced in Example 6.2.1. Then the convolution of  $\varphi_1$  with itself, denoted by  $\varphi_2 = \varphi_1 * \varphi_1$ , is the piecewise linear B-spline, also called the **hat function**, given by

$$\varphi_2(x) = \min\{x, 2-x\} \chi_{[0,2)}(x) = \begin{cases} x, & \text{for } 0 \leq x < 1, \\ 2-x, & \text{for } 1 \leq x < 2, \\ 0, & \text{elsewhere.} \end{cases} \quad (6.2.14)$$

Since the Fourier transform of  $\varphi_1$  is given by

$$\widehat{\varphi}_1(\omega) = \frac{1 - e^{-i\omega}}{i\omega},$$

it follows that

$$\begin{aligned} \widehat{\varphi}_1(\omega) &= (\widehat{\varphi}_1(\omega))^2 = \left( \frac{1 - e^{-i\omega}}{i\omega} \right)^2 \\ &= \left( \frac{1 + e^{-i\omega/2}}{2} \right)^2 \left( \frac{1 - e^{-i\omega/2}}{i\omega/2} \right)^2 \\ &= \left( \frac{1 + e^{-i\omega/2}}{2} \right)^2 \widehat{\varphi}_2\left(\frac{\omega}{2}\right). \end{aligned}$$

Thus,  $\varphi_2$  is refinable with two-scale symbol  $P_2(z)$  given by

$$P_2(z) = \left( \frac{1+z}{2} \right)^2,$$

or the refinement sequence is given by

$$p_0 = \frac{1}{2}, p_1 = 1, p_2 = \frac{1}{2}, p_k = 0, k \neq 0, 1, 2;$$

that is,

$$\varphi_2(x) = \frac{1}{2}\varphi_2(2x) + \varphi_2(2x-1) + \frac{1}{2}\varphi_2(2x-2).$$

■

**Example 6.2.3** More generally, let  $\varphi_m$  be the Cardinal B-spline of order  $m \geq 1$  defined by the  $m$ -fold convolution:

$$\varphi_m(x) = \underbrace{(\varphi_0 * \varphi_0 * \cdots * \varphi_0)}_{m \text{ copies of } \varphi_0}(x). \quad (6.2.15)$$

Then  $\varphi_m$  is refinable with two-scale symbol given by

$$P_m(z) = \left( \frac{1+z}{2} \right)^m, \quad (6.2.16)$$

or equivalently, with refinement sequence

$$p_k = 2^{m-1} \binom{m}{k}$$

for  $0 \leq k \leq m$ ;  $p_k = 0$  for  $k < 0$  or  $k > m$ .

■

### 6.2.3 The MRA architecture

In view of the examples  $\phi_S$  introduced in Subunit 6.2.1 and the  $B$ -spline examples  $\varphi_m, m = 0, 1, 2, \dots$ , studied in Subunit 6.2.2, we are now ready to introduce the notion of multiresolution analysis (MRA) generated by any refinable function  $\phi \in L_2(\mathbb{R})$ .

**Definition 6.2.1** *Let  $\phi \in L_2(\mathbb{R})$  be a refinable function and*

$$\mathbb{V}_j = \overline{\text{span}} \{ \phi(2^j x - k) : k \in \mathbb{Z} \}$$

*be the  $L_2$ -closure of the algebraic span of  $\{ \phi(2^j x - k) \}$ . Then the sequence  $\{ \mathbb{V}_j \}$  of subspaces of  $L_2(\mathbb{R})$  is called a multiresolution analysis (MRA) of  $L_2(\mathbb{R})$  if the following conditions are satisfied*

- (i)  $\mathbb{V}_j \subset \mathbb{V}_{j+1}, \quad j \in \mathbb{Z};$
- (ii)  $\cap_{j \in \mathbb{Z}} \mathbb{V}_j = \{0\};$
- (iii)  $\cup_{j \in \mathbb{Z}} \mathbb{V}_j$  is dense in  $L_2(\mathbb{R})$ ;
- (iv) any function  $f(x) \in \mathbb{V}_j$  if and only if  $f(2x) \in \mathbb{V}_{j+1}$ ;
- (v)  $\{ \phi(x-k) : k \in \mathbb{Z} \}$  is a Riesz basis of  $\mathbb{V}_0$ ; that is, there exist some positive constants  $0 < c \leq C < \infty$ , such that

$$c \sum_k |c_k|^2 \leq \left\| \sum_{k \in \mathbb{Z}} c_k \phi(x-k) \right\|_{L_2(\mathbb{R})}^2 \leq C \sum_{k \in \mathbb{Z}} |c_k|^2 \quad (6.2.17)$$

for all square-summable sequences  $\{c_k\}$ .

We will also say that the refinable function  $\phi$  generates the MRA. The main objective of introducing the notion of MRA is to provide an architecture for the construction of wavelets.

Let us first return to the sinc function example  $\phi_S$  considered in Subunit 6.2.1, defined in terms of its Fourier transform by  $\widehat{\phi}_S(\omega) = \chi_{[\pi, \pi]}(\omega)$ , so that

$$\widehat{\phi}_S\left(\frac{\omega}{2^j}\right) = \chi_{[-2^j\pi, 2^j\pi]}(\omega)$$

is an “ideal lowpass filter” with pass-band  $[-2^j\pi, 2^j\pi]$ . Observe that for each integer  $j \in \mathbb{Z}$ , the difference

$$\widehat{\phi}_S(\omega)\left(\frac{\omega}{2^j}\right) - \widehat{\phi}(\omega)\left(\frac{\omega}{2^{j-1}}\right)$$

is the ideal bandpass filter with pass-band

$$[-2^j\pi, -2^{j-1}\pi) \cup (2^{j-1}\pi, 2^j\pi]. \quad (6.2.18)$$

Therefore, to separate any function  $f_J(x)$ , with bandwidth  $\leq 2^{J+1}\pi$ , into components:

$$f_J(x) = f_0(x) + g_0(x) + \cdots + g_{J-1}(x) \quad (6.2.19)$$

on non-overlapping (ideal) frequency bands; that is,

$$\begin{aligned} \widehat{f}_0(\omega) &= \widehat{f}_J(\omega)\chi_{[-\pi, \pi]}(\omega) \\ \widehat{g}_0(\omega) &= \widehat{f}_J(\omega)\chi_{[-2\pi, -\pi) \cup (\pi, 2\pi]}(\omega) \\ \widehat{g}_1(\omega) &= \widehat{f}_J(\omega)\chi_{[-2^2\pi, -2\pi) \cup (2\pi, 2^2\pi]}(\omega) \\ &\vdots \\ \widehat{g}_{J-1}(\omega) &= \widehat{f}_J(\omega)\chi_{[-2^J\pi, -2^{J-1}\pi) \cup (2^{J-1}\pi, 2^J\pi]}(\omega), \end{aligned}$$

it is required to find an ideal bandpass filter  $\psi_I(x)$  with Fourier transform

$$\widehat{\psi}_{S,I}(\omega) = \chi_{[-2\pi, -\pi) \cup (\pi, 2\pi]}(\omega) = \widehat{\phi}_S\left(\frac{\omega}{2}\right) - \widehat{\phi}_S(\omega), \quad (6.2.20)$$

which yields

$$\widehat{\psi}_{S,I}\left(\frac{\omega}{2^j}\right) = \widehat{\phi}_S\left(\frac{\omega}{2^{j+1}}\right) - \widehat{\phi}_S\left(\frac{\omega}{2^j}\right) = \chi_{[-2^{j+1}\pi, -2^j\pi) \cup (2^j\pi, 2^{j+1}\pi]}(\omega)$$

for  $j = 0, \dots, J-1$ . However, for computational and other reasons, we introduce a phase shift of  $\psi_S(x)$  to define the “wavelet”

$$\psi_S(x) := -2\phi_S(2x-1) + \phi_S\left(x - \frac{1}{2}\right), \quad (6.2.21)$$

so that

$$\widehat{\psi}_S(\omega) = -e^{-i\frac{\omega}{2}}\left(\widehat{\phi}_S\left(\frac{\omega}{2}\right) - \widehat{\phi}_S(\omega)\right). \quad (6.2.22)$$

Observe that  $|\widehat{\psi}_S(\omega)| = |\widehat{\psi}_{S,I}(\omega)|$  by comparing (6.2.20) with (6.2.18), and hence the separation of  $f_J(x)$  into components on ideal frequency bands in (6.2.17) remains valid with only a phase shift of  $g_j(x)$  by  $-(\pi + \omega/2^j)$ ,  $j = 0, \dots, J-1$ .

In the definition of  $\psi_S(x)$  in (6.2.21), we remark that  $\psi_S \in \mathbb{V}_1$ , with

$$\begin{aligned} \psi_S(x) &= \sum_{k=-\infty}^{\infty} p_{S,k} \phi_S(2x - (k+1)) - 2\phi(2x-1) \\ &= \sum_{k=-\infty}^{\infty} (p_{S,k-1} - 2\delta_{k-1}) \phi(2x - k), \end{aligned} \quad (6.2.23)$$

by applying (6.2.5). Furthermore, since we have, from (6.2.4), that  $p_{S,2j} = \delta_{2j}$  and

$$p_{S,1-2j} = \frac{2 \sin \frac{(1-2j)\pi}{2}}{(1-2j)\pi} = \frac{-2 \sin \frac{(2j-1)\pi}{2}}{-(2j-1)\pi} = p_{S,2j-1},$$

so that

$$(-1)^k p_{S,1-k} = \begin{cases} p_{S,2j-1}, & \text{for } k = 2j, \\ -\delta_{2j}, & \text{for } k = 2j+1, \end{cases}$$

for all  $k$ , it follows from (6.2.21) that  $\psi_S$  as defined in (6.2.21) satisfies the two-scale relation

$$\psi_S(x) = \sum_{k=-\infty}^{\infty} q_{S,k} \phi_S(2x - k), \quad (6.2.24)$$

with

$$q_{S,k} = (-1)^k p_{S,1-k}, \quad k \in \mathbb{Z}. \quad (6.2.25)$$

For any function  $\psi$  defined on  $\mathbb{R}$ , the notation

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi(2^j x - k), \quad j, k \in \mathbb{Z}, \quad (6.2.26)$$

is commonly used for the reason that  $\|\psi_{j,k}\|_2 = \|\psi\|_2$  for all  $j, k \in \mathbb{Z}$ . We will call  $\psi \in L_2(\mathbb{R})$  an **orthogonal wavelet** provided that the family  $\{\psi_{j,k} : j, k \in \mathbb{Z}\}$  is an orthonormal basis of  $L_2(\mathbb{R})$ .

The construction of compactly supported orthogonal and biorthogonal wavelets based on the MRA architecture will be studied in Subunit 6.3.3.

### 6.3 Wavelet Construction

This is a very elaborate subunit, with detailed discussions of the theoretical development, computational schemes, and algorithms for implementation.

In Subunit 6.3.1, to prepare for the construction of compactly supported orthogonal wavelets, the notion of quadratic mirror filter (QMF) is introduced and studied. The QMF is generalized, in Subunit 6.3.2, to the formulation of the matrix extension specification for the construction of compactly supported bi-orthogonal wavelets. In Subunit 6.3.3, the notion of direct-sum and orthogonal-sum decomposition (of the  $L_2(\mathbb{R})$  space as a sum of multi-scale wavelet subspaces) is introduced, and the corresponding wavelet decomposition and wavelet reconstruction algorithms are formulated. This subunit ends with the study of analysis and synthesis wavelets, and the method of construction of both orthonormal wavelets (that is, orthogonal wavelets with unit-norm) and bi-orthogonal wavelets.

### 6.3.1 Quadrature mirror filter

Recall from (6.2.8)–(6.2.11) that for any refinable function  $\phi$  with refinement sequence  $\{p_k\}$ , the two-scale symbol of  $\{p_k\}$  is defined by the Laurent series

$$P(z) = \frac{1}{2} \sum_k p_k z^k. \quad (6.3.1)$$

On the other hand, for any sequence  $\{c_k\}$ , its symbol will be defined by the formal Laurent series

$$C(z) = \sum_k c_k z^k$$

(without the  $\frac{1}{2}$  normalization constant). In general, following the example of the wavelet  $\psi_S$  in (6.2.21)–(6.2.23), we introduce, for any refinable function  $\phi$ , its corresponding “wavelet”

$$\psi(x) = \sum_k q_k \phi(2x - k) \quad (6.3.2)$$

where  $\{q_k\}$  has yet to be determined. But analogous to the two-scale relation

$$\widehat{\phi}(\omega) = P(e^{-i\frac{\omega}{2}}) \widehat{\phi}\left(\frac{\omega}{2}\right)$$

in (6.2.9) for the refinement function  $\phi$ , we may also re-write (6.3.2) as

$$\widehat{\psi}(\omega) = Q(e^{-i\frac{\omega}{2}}) \widehat{\phi}\left(\frac{\omega}{2}\right) \quad (6.3.3)$$

where

$$Q(z) := \frac{1}{2} \sum_k q_k z^k. \quad (6.3.4)$$

Inspired by the simple formula of  $\{q_{S,k}\}$  in (6.2.22), we may be tempted to consider

$$q_k = (-1)^k \bar{p}_{1-k}, \quad k \in \mathbb{Z}. \quad (6.3.5)$$



This choice, however, is not necessarily a good one, unless the refinable function is “orthonormal” (i.e. orthogonal, with unit norm), in that

$$\langle \phi(x-k), \phi(x-j) \rangle := \int_{-\infty}^{\infty} \phi(x-k) \overline{\phi(x-j)} dx = \delta_{k-j} \quad (6.3.6)$$

where  $\delta_\ell$  is the Kronecker symbol. Recall that the sinc function example  $\phi_S(x)$  in Subunit 6.2.1 does satisfy (6.3.6), since the Fourier transform of  $\phi_S(x-j)$  is  $\chi_{[\pi, \pi]}(\omega) e^{-ij\omega}$  and

$$\begin{aligned} \langle \phi_S(x-k), \phi_S(x-j) \rangle &= \frac{1}{2\pi} \langle \widehat{\phi}_S(\omega) e^{-ik\omega}, \widehat{\phi}_S(\omega) e^{-ij\omega} \rangle \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-i(k-j)\omega} d\omega = \delta_{k-j}, \end{aligned}$$

by applying Plancherel’s identity. Indeed, for any refinable function  $\phi$  with refinement sequence  $\{p_k\}$ , if  $\phi$  is orthonormal, then the choice of  $\{q_k\}$  in (6.3.5) assures that the function  $\psi$  defined in (6.3.2) is an orthonormal wavelet provided that the refinement sequence  $\{p_k\}$  satisfies the so-called “sum rule”:

$$\sum_j p_{2j} = \sum_j p_{2j-1} = 1, \quad (6.3.7)$$

as stated in the following theorem.

**Theorem 6.3.1** *Let  $\phi \in (L_1 \cap L_2)(\mathbb{R})$  be an orthonormal refinable function with refinement sequences  $\{p_k\}$  that satisfies the sum rule (6.3.7). Then the function  $\psi$  in (6.3.2), with  $\{q_k\}$  defined by (6.3.5), is an orthonormal wavelet, in that it satisfies both (6.1.1) and (6.3.6) (with  $\phi$  replaced by  $\psi$ ). Furthermore,*

$$\langle \psi(x-k), \phi(x-j) \rangle = 0, \quad j, k \in \mathbb{Z}. \quad (6.3.8)$$

We remark that the orthogonality property (6.3.8) implies the orthogonality of the two sub-spaces

$$\mathbb{V}_j = \overline{\text{span}}\{\phi(2^j x - k), \quad k \in \mathbb{Z}\}$$

and

$$\mathbb{W}_j = \overline{\text{span}}\{\psi(2^j x - k), \quad k \in \mathbb{Z}\}$$

of  $\mathbb{V}_{j+1}$ , for all  $j \in \mathbb{Z}$ . Since  $\mathbb{V}_j + \mathbb{W}_j = \mathbb{V}_{j+1}$  (to be shown in Subunit 6.3.2), we will write

$$\mathbb{V}_{j+1} = \mathbb{V}_j \oplus^\perp \mathbb{W}_j, \quad j \in \mathbb{Z}, \quad (6.3.9)$$

called an orthogonal decomposition of the MRA  $\{\mathbb{V}_j\}$  of  $L_2(\mathbb{R})$ . Indeed, in view of the density condition (iii) in the definition of an MRA, it follows from (6.3.9) that

$$L_2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} \mathbb{W}_j, \quad (6.3.10)$$

which implies that every function  $f \in L_2(\mathbb{R})$  has an orthogonal decomposition

$$f(x) = \sum_{j \in \mathbb{Z}} g_j(x),$$

where  $g_j \in \mathbb{W}_j$ .

To prove the above theorem, we first note that  $\psi$  is indeed a wavelet, since

$$\begin{aligned} \int_{-\infty}^{\infty} \psi(x) dx &= \sum_k q_k \int_{-\infty}^{\infty} \phi(x) dx \\ &= \left( \int_{-\infty}^{\infty} \phi(x) dx \right) \left( \sum_k (-1)^k \bar{p}_{1-k} \right) \\ &= \left( \int_{-\infty}^{\infty} \phi(x) dx \right) \left( \sum_k \bar{p}_{2j} - \sum_j \bar{p}_{2j-1} \right) \\ &= \left( \int_{-\infty}^{\infty} \phi(x) dx \right) (1 - 1) = 0 \end{aligned}$$

by applying the sum rule (6.3.7). To derive the property (6.3.6) for the wavelet  $\psi$ , we first observe that the property (6.3.6) for the refinable function  $\phi$  implies that

$$\sum_k p_k \bar{p}_{k-2n} = 2\delta_n, \quad n \in \mathbb{Z}. \quad (6.3.11)$$

Indeed, for any  $n \in \mathbb{Z}$ , applying the refinement relation for  $\phi$  to (6.3.6) yields

$$\begin{aligned} \delta_n &= \langle \phi(x), \phi(x-n) \rangle \\ &= \sum_k \sum_j p_k \bar{p}_j \langle \phi(2x-k), \phi(2x-(j+2n)) \rangle \\ &= \sum_k \sum_j p_k \bar{p}_{j-2n} \langle \phi(2x-k), \phi(2x-j) \rangle \\ &= \frac{1}{2} \sum_j p_j \bar{p}_{j-2n} \end{aligned}$$

which gives (6.3.10). Hence, for any  $n \in \mathbb{Z}$ , the same derivation also yields

$$\begin{aligned} \langle \psi(x), \psi(x-n) \rangle &= \sum_k \sum_j q_k \bar{q}_j \langle \phi(2x-k), \phi(2x-(j+2n)) \rangle \\ &= \frac{1}{2} \sum_j q_j \bar{q}_{j-2n} \end{aligned}$$

Therefore, from the selection of  $\{q_k\}$  in (6.3.5), we have

$$\begin{aligned}
 \langle \psi(x), \psi(x-n) \rangle &= \frac{1}{2} \sum_j (-1)^j \bar{p}_{1-j} (-1)^{j-2n} p_{1-j+2n} \\
 &= \frac{1}{2} \sum_j q_j \bar{p}_{1-j} p_{1-j+2n} = \frac{1}{2} \sum_k q \bar{p}_k \bar{p}_{k+2n} \\
 &= \frac{1}{2} 2\delta_n = \delta_n
 \end{aligned}$$

by applying (6.3.11). That is,

$$\langle \psi(x-k), \psi(x-j) \rangle = \langle \psi(x), \psi(x-(j-k)) \rangle = \delta_{j-k},$$

which is the property (6.3.6) for the wavelet  $\psi$ . Finally, to establish (6.3.8), we consider

$$\begin{aligned}
 d_n &:= \langle \psi(x), \phi(x-n) \rangle \\
 &= \frac{1}{2} \sum_j q_j \bar{p}_{j-2n} = \frac{1}{2} \sum_j (-1)^j \bar{p}_{1-j} \bar{p}_{j-2n} \\
 &= \frac{1}{2} \sum_k (-1)^{1-k+2n} \bar{p}_{k-2n} \bar{p}_{1-k} = -d_n,
 \end{aligned}$$

where we have changed the index of summation from  $1-j$  to  $k-2n$ . Hence,  $2d_n = 0$  implies  $d_n = 0$ , completing the derivation of (6.3.8), and hence, the proof of the theorem.  $\blacksquare$

Next we prove that the property of normalized orthogonality (6.3.6) is equivalent to the identity

$$\sum_{k=-\infty}^{\infty} |\widehat{\phi}(\omega + 2\pi k)|^2 = 1, \quad \omega \in \mathbb{R}. \quad (6.3.12)$$

The reason is that Plancherel's identity can be applied to (6.3.6) to yield

$$\begin{aligned}
 \delta_j &= \langle \phi(x), \phi(x-j) \rangle \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\phi}(\omega) \overline{\widehat{\phi}(\omega)} e^{-ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{\phi}(\omega)|^2 e^{-ij\omega} d\omega \\
 &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{2\pi k}^{2\pi(k+1)} |\widehat{\phi}(\omega)|^2 e^{-ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=-\infty}^{\infty} |\widehat{\phi}(\omega + 2\pi k)|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=-\infty}^{\infty} |\widehat{\phi}(\omega + 2\pi k)|^2 e^{ij\omega} d\omega;
 \end{aligned}$$

that is, the Fourier coefficients of the  $2\pi$ -periodic function  $\sum_{k=-\infty}^{\infty} |\widehat{\phi}(\omega + 2\pi k)|^2$  are the same as those of the constant function 1, so that the uniqueness of Fourier series representations yields (6.3.12). ■

As an application of (6.3.12), we apply the two-scale relation (6.2.9) to the

above derivation to obtain

$$\begin{aligned}
 \delta_j &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{4\pi k}^{4\pi(k+1)} \left| P(e^{-i\frac{\omega}{2}}) \widehat{\phi}\left(\frac{\omega}{2}\right) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{4\pi} \sum_{k=-\infty}^{\infty} \left| P(e^{-i\frac{\omega}{2}}) \widehat{\phi}\left(\frac{\omega}{2} + 2\pi k\right) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{4\pi} \left| P(e^{-i\frac{\omega}{2}}) \right|^2 \sum_{k=-\infty}^{\infty} \left| \widehat{\phi}\left(\frac{\omega}{2} + 2\pi k\right) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{4\pi} \left| P(e^{-i\frac{\omega}{2}}) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \left\{ \int_0^{2\pi} \left| P(e^{-i\frac{\omega}{2}}) \right|^2 e^{ij\omega} d\omega + \int_{2\pi}^{4\pi} \left| P(e^{-i\frac{\omega}{2}}) \right|^2 e^{ij\omega} d\omega \right\} \\
 &= \frac{1}{2\pi} \left\{ \int_0^{2\pi} \left| P(e^{-i\frac{\omega}{2}}) \right|^2 e^{ij\omega} d\omega + \int_0^{2\pi} \left| P(e^{-i(\frac{\omega}{2}+\pi)}) \right|^2 e^{ij\omega} d\omega \right\} \\
 &= \frac{1}{2\pi} \int_0^{2\pi} \left( \left| P(e^{-i\frac{\omega}{2}}) \right|^2 + \left| P(-e^{-i\frac{\omega}{2}}) \right|^2 \right) e^{ij\omega} d\omega,
 \end{aligned}$$

where the identity (6.3.23) was applied. Hence, the uniqueness of Fourier series representations can be applied again to acquire the identity

$$|P(z)|^2 + |P(-z)|^2 = 1, \quad |z| = 1. \quad (6.3.13)$$

The same proof also yields

$$\begin{aligned}
 0 &= \langle \psi(x), \phi(x-j) \rangle \\
 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\psi}(\omega) \overline{\widehat{\phi}(\omega)} e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{4\pi k}^{4\pi(k+1)} Q(e^{-i\frac{\omega}{2}}) \overline{P(e^{-i\frac{\omega}{2}})} \left| \widehat{\phi}\left(\frac{\omega}{2}\right) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{4\pi} Q(e^{-i\frac{\omega}{2}}) \overline{P(e^{-i\frac{\omega}{2}})} \sum_{k=-\infty}^{\infty} \left| \widehat{\phi}\left(\frac{\omega}{2} + 2\pi k\right) \right|^2 e^{ij\omega} d\omega \\
 &= \frac{1}{2\pi} \int_0^{2\pi} \left( Q(e^{-i\frac{\omega}{2}}) \overline{P(e^{-i\frac{\omega}{2}})} + Q(-e^{-i\frac{\omega}{2}}) \overline{P(-e^{-i\frac{\omega}{2}})} \right) e^{ij\omega} d\omega
 \end{aligned}$$

for all  $j \in \mathbb{Z}$ ; so that we have the identity

$$Q(z)\overline{P(z)} + Q(-z)\overline{P(-z)} = 0, \quad |z| = 1. \quad (6.3.14)$$

On the other hand, it follows from (6.3.4)–(6.3.5) that

$$\begin{aligned} Q(z) &= \frac{1}{2} \sum_k (-1)^k \bar{p}_{1-k} z^k \\ &= \frac{1}{2} \sum_k (-1)^{1-j} \bar{p}_j z^{1-j} \\ &= -z \frac{1}{2} \sum_k \bar{p}_j (-z)^j = -z \overline{P(-z)} \end{aligned} \quad (6.3.15)$$

Therefore, as a consequence of (6.3.13), we have the identity

$$|Q(z)|^2 + |Q(-z)|^2 = 1, \quad |z| = 1. \quad (6.3.16)$$

In matrix formulation, the totality of (6.3.13), (6.3.14), and (6.3.16) can be re-written as

$$\begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix} \begin{bmatrix} \overline{P(z)} & \overline{Q(z)} \\ \overline{P(-z)} & \overline{Q(-z)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad |z| = 1. \quad (6.3.17)$$

**Definition 6.3.1** If  $P(z)$  and  $Q(z)$  satisfy (6.3.17) (that is, (6.3.13), (6.3.14), and (6.3.16)) for all  $z$  on the unit circle  $|z| = 1$  of the complex plane, then the pair  $(P(z), Q(z))$  is said to provide a quadrature mirror filter (QMF).

The reason for the terminology of QMF will be clear when we discuss “wavelet decomposition and reconstruction” in Subunit 6.3.3. As already discussed, if  $P(z)$  satisfies (6.3.13), the  $Q(z)$  can be chosen by using (6.3.15), or more generally, by setting

$$Q(z) = -z^{2k+1} \overline{P(-z)} \quad (6.3.18)$$

for any integer  $k$ , for  $(P(z), Q(z))$  to be a QMF pair. In other words, to construct a QMF, the only task is to construct a two-scale symbol  $P(z)$  defined in (6.3.1) that satisfies (6.3.13).

**Example 6.3.1** Let  $\{p_{S,k}\}$  be the refinement sequence, given in (6.2.5), of the refinement function  $\phi_S(x) = \text{sinc } x = \sin \pi x / \pi x$  in (6.2.1). Then the corresponding two-scale symbol,

$$P_S(z) = \frac{1}{2} \sum_{k=-\infty}^{\infty} p_{S,k} z^k, \quad (6.3.19)$$

satisfies the QMF condition (6.3.13).

**Solution** By taking the Fourier transform of  $\phi_S(x)$  in (6.2.5), we have

$$\widehat{\phi}_S(\omega) = P_S(e^{-i\omega/2}) \widehat{\phi}_S\left(\frac{\omega}{2}\right).$$

Since  $\widehat{\phi}_S(\omega) = \psi_{[-\pi, \pi)}(\omega)$ , it is clear that

$$|P_S(z)|^2 = P_S(z) = P_S(e^{-i\omega/2}) = \begin{cases} 1 & \text{if } -\frac{\pi}{2} \leq \omega < \frac{\pi}{2} \\ 0 & \text{if } \frac{\pi}{2} \leq \omega < \frac{3\pi}{2}, \end{cases}$$

where  $z = e^{-i\omega/2}$ . Hence,  $|P_S(-z)|^2 = P_S(-z) = 1 - P_S(z) = 1 - |P_S(z)|^2$ , yielding (6.3.13). ■

We remark that the QMF in Example 6.3.1 is given only as a proof of concept, since the decay of the infinite filter sequence  $\{p_{S,k}\}$  is too slow for any wavelet application.

**Example 6.3.2** Let  $\varphi_1(x) = \chi_{[0,1)}(x)$  be the refinable function discussed in Example 6.3.1 of Subunit 6.2.2 with refinement sequence  $\{p_k\}$  given by  $p_0 = p_1 = 1$  and  $p_k \neq 0$  for all  $k \neq 0, 1$ . Show that the two-scale symbol  $P_1(z) = (1+z)/2$  of  $\varphi_1(x)$  satisfies (6.3.13). Also, construct the (Haar) wavelet  $\psi_0(x)$  associated with  $\varphi_1(x)$  and the corresponding QMF  $(P_1(z), Q_1(z))$ .

**Solution** For  $|z| = 1$ , we have

$$\begin{aligned} |P_1(z)|^2 + |P_1(-z)|^2 &= \frac{1}{2^2}|1+z|^2 + \frac{1}{2^2}|1-z|^2 \\ &= \frac{1}{4}((1+z)\overline{(1+z)} + (1-z)\overline{(1-z)}) \\ &= \frac{1}{4}((1+\bar{z}+z+|z|^2) + ((1-\bar{z}-z+|z|^2))) \\ &= \frac{1}{4}(1+1+1+1) = 1. \end{aligned}$$

This proves that  $P_1(z)$  satisfies (6.3.13). Next, by (6.3.5), we have

$$q_k = (-1)^k \overline{p_{1-k}} = \begin{cases} 1 & \text{if } k = 0 \\ -1 & \text{if } k = 1 \\ 0 & \text{otherwise,} \end{cases}$$

so that the corresponding (Haar) wavelet  $\psi_1(x)$  defined in (6.3.2) is given by

$$\begin{aligned} \psi_1(x) &= \sum_{k=-\infty}^{\infty} q_k \varphi_1(2x - k) \\ &= \varphi_1(2x) - \varphi_1(2x - 1) \\ &= \psi_{[0, \frac{1}{2})}(x) - \psi_{[\frac{1}{2}, 1)}(x). \end{aligned} \tag{6.3.20}$$

Moreover, the two-scale symbol  $Q_1(z)$  defined in (6.3.4) is given by

$$Q_1(x) = \frac{1}{2} \sum_{k=-\infty}^{\infty} q_k q_k z^k = \frac{1}{2}(1 - z),$$

which agrees with  $-z \overline{P_1(-z)}$ , for  $|z| = 1$ , as derived in (6.3.15), yielding the Haar QMF pair

$$(P_1(z), Q_1(z)) = \left( \frac{1+z}{2}, \frac{1-z}{2} \right).$$

■

We conclude this subunit by pointing out that while the QMF for the “Haar wavelet” in Example 6.3.2 is most economical (with the Laurent series being linear polynomials), the QMF for the “Shannon wavelet” in Example 6.3.1 is not too useful, being a bi-infinite Laurent series. In Subunit 6.3.3, we will introduce the notion of Daubechies wavelets whose QMF are Laurent polynomials of degree greater than 1.

### 6.3.2 Matrix extension

A QMF pair  $(P(z), Q(z))$  of Laurent series, studied in the previous subunit, generate a  $2 \times 2$  (Laurent) matrix

$$\mathcal{M}_{P,Q}(z) := \begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix}, \quad |z| = 1, \quad (6.3.21)$$

which satisfies the QMF condition (6.3.17). In this subunit, we re-formulate (6.3.17) as

$$\mathcal{M}_{P,Q}(z) \mathcal{M}_{P,Q}^*(z) = I_2, \quad |z| = 1, \quad (6.3.22)$$

where  $I_2$  is the  $2 \times 2$  identity matrix and  $\mathcal{M}_{P,Q}^*$  denotes the complex conjugate of the transpose of  $\mathcal{M}_{P,Q}$ . Observe that for  $|z| = 1$ , the complex conjugation implies that  $\bar{z} = \frac{1}{z}$ . This is instrumental to the algebraic manipulation in (5.3.15) that leads to the identity (6.3.16). In this regard, we remark that the adjoint of a matrix is also formulated as the complex conjugation of its transpose, as studied in Subunit 1.2.2 (see (1.2.27)). The importance is that the inverse of a unitary matrix is its adjoint. Hence, as a consequence of (6.3.22), for a QMF pair  $(P(z), Q(z))$ , the matrix  $\mathcal{M}_{P,Q}(z)$  is invertible, with inverse given by  $\mathcal{M}_{P,Q}^*(z)$ , for  $|z| = 1$ . In the previous subunit, we have also studied two examples of QMF pairs; but such examples are rare. For instance, let us consider the  $m$ -th order Cardinal  $B$ -spline  $\varphi_m(x)$  defined by the  $m$ -fold convolution of the characteristic function  $\varphi_1(x) = \chi_{[0,1)}(x)$  in (6.2.13), with two-scale symbol

$$P_m(z) = \left( \frac{1+z}{2} \right)^m \quad (6.3.23)$$

in (6.2.14). We will show, in the following example, that for any integer



$m > 1$ ,  $P_m(z)$  does not satisfy the necessary condition (6.3.17) or equivalently (6.3.22), for a QMF.

**Example 6.3.3** Let  $m \geq 1$  and  $P_m(z)$  be the polynomial in (6.3.23) (i.e. (6.2.13)). Show that for  $z = e^{i\theta}$ ,

$$|P_m(z)|^2 + |P_m(-z)|^2 = 1, \quad \theta \in \mathbb{R}$$

holds if and only if  $m = 1$ .

**Solution** Observe that for  $z = e^{i\theta}$ ,

$$|1 \pm z|^2 = (1 \pm z)(1 \pm \bar{z}) = 2(1 \pm \cos \theta).$$

Hence, it follows from (6.3.23) that

$$|P_m(\pm z)|^2 = \left( \frac{1 \pm \cos \theta}{2} \right)^m,$$

so that

$$|P_m(z)|^2 + |P_m(-z)|^2 = \left( \frac{1 + \cos \theta}{2} \right)^m + \left( \frac{1 - \cos \theta}{2} \right)^m.$$

It is clear that the quantity on the right-hand side is equal to 1 for  $m = 1$ . But for  $m \geq 2$ ,

$$\begin{aligned} |P_m(z)|^2 + |P_m(-z)|^2 &= 2^{-m} \sum_{k=0}^m \left( 1 + (-1)^k \right) \binom{m}{k} \cos^k \theta \\ &= 2^{-m+1} \sum_{j=0}^{\lfloor m/2 \rfloor} \binom{m}{2j} \cos^{2j} \theta \\ &< 2^{-m+1} \sum_{j=0}^{\lfloor m/2 \rfloor} \binom{m}{2j} = 1, \end{aligned}$$

for  $0 < |\theta| \leq \pi/2$ . The reason is that while

$$\sum_{k=0}^m \binom{m}{k} = 2^m,$$

or

$$2^{-m+1} \sum_{k=0}^m \binom{m}{k} = 2,$$

we have

$$\sum_{\text{odd } j} \binom{m}{j} = \sum_{\text{even } j} \binom{m}{j}.$$

■

In the next subunit, namely Subunit 6.3.3, we will use  $P_m(z)$ ,  $m \geq 2$ , as a multiplicative factor of some Laurent polynomial  $P_{D,2m}(z)$  of degree  $2m-1$ , that satisfies the necessary condition (6.3.13), so that together with the Laurent polynomial  $Q_{D,2m}(z) = -z \overline{P_{D,2m}(-z)}$  as suggested by (6.3.15), the (Laurent) matrix  $\mathcal{M}_{P_{D,2m}, Q_{D,2m}}(z)$  does satisfy the QMF condition (6.3.22) for all  $|z| = 1$ . Such Laurent polynomials  $P_{D,2m}(z)$  are two-scale symbols of the orthonormal Daubechies scaling function  $\varphi_{D,2m}(x)$ , with  $Q_{D,2m}(z)$  being the two-scale symbols of the corresponding orthonormal Daubechies wavelets  $\psi_{D,2m}(x)$ .

But for now, we are interested in the more general setting of arbitrarily given refinable functions  $\phi(x)$  with refinement sequences  $\{p_k\}$  that satisfy the sum rule condition

$$\sum_k p_{2k} = \sum_k p_{2k-1} = 1,$$

and corresponding two-scale symbols  $Q(z)$ , for which three other Laurent polynomials  $Q(z)$ ,  $A(z)$ ,  $B(z)$  exist, so that  $\mathcal{M}_{P,Q}(z)$  is invertible for  $|z| = 1$ , with inverse given by  $\mathcal{M}_{A,B}^*(z)$ . We will again consider

$$P(z) = \frac{1}{2} \sum_k p_k z^k;$$

$$Q(z) = \frac{1}{2} \sum_k q_k z^k;$$

$$A(z) = \frac{1}{2} \sum_k a_k z^k;$$

$$B(z) = \frac{1}{2} \sum_k b_k z^k,$$

where  $\{p_k\}$ ,  $\{q_k\}$ ,  $\{a_k\}$ , and  $\{b_k\}$  are preferably finite sequences. The general problem of finding Laurent symbols  $Q(z)$ ,  $A(z)$ ,  $B(z)$  that converge absolutely and uniformly on  $|z| = 1$ , such that

$$\mathcal{M}_{P,Q}(z) \mathcal{M}_{A,B}^*(z) = \begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix} \begin{bmatrix} \overline{A(z)} & \overline{B(z)} \\ \overline{A(-z)} & \overline{B(-z)} \end{bmatrix} = I_2 \quad (6.3.24)$$

for  $|z| = 1$ , is called the problem of “matrix extension.” This problem is a generalization of the QMF condition, for which

$$Q(z) = -z \overline{P(-z)}, \quad A(z) = \overline{P(z)},$$

and

$$B(z) = \overline{Q(z)} = z^{-1} P(z)$$

for  $|z| = 1$ .

**Theorem 6.3.2** *Let  $\phi$  be a refinable function with refinement sequence  $\{p_k\} \in \ell^1$  that satisfies the sum rule condition  $\sum_k p_{2k} = \sum_k p_{2k-1} = 1$ . Suppose that corresponding to the two-scale symbol  $P(z)$  of  $\{p_k\}$ , there exist sequences  $\{q_k\}, \{a_k\}, \{b_k\} \in \ell^1$  such that the symbols*

$$Q(z) = \frac{1}{2} \sum_k q_k z^k, A(z) = \frac{1}{2} \sum_k a_k z^k, B(z) = \frac{1}{2} \sum_k b_k z^k,$$

*satisfy the matrix extension condition (6.3.24). Then by setting*

$$\psi(x) = \sum_k q_k \phi(2x - k), \quad (6.3.25)$$

*the given refinable function  $\phi$  also has the decomposition property:*

$$\phi(2x - \ell) = \frac{1}{2} \sum_k \{ \bar{a}_{\ell-2k} \phi(x - k) + \bar{b}_{\ell-2k} \psi(x - k) \} \quad (6.3.26)$$

*for all  $x \in \mathbb{R}$  and all  $\ell \in \mathbb{Z}$ .*

**Proof** We first observe that for the invertible matrix  $\mathcal{M}_{P,Q}(z), |z| = 1$ , its right inverse  $\mathcal{M}_{A,B}^*(z)$  is also its left inverse. This yields

$$\begin{aligned} P(z) \overline{A(z)} + Q(z) \overline{B(z)} &= 1; \\ P(z) \overline{A(-z)} + Q(z) \overline{B(-z)} &= 0, \end{aligned} \quad (6.3.27)$$

by multiplying out the first and second rows of  $\mathcal{M}_{A,B}^*(z)$ , respectively, to the first column of  $\mathcal{M}_{P,Q}(z)$ . Hence, by adding and subtracting the two equations in (6.3.27), we have

$$P(z) (\overline{A(z)} + \overline{A(-z)}) + Q(z) (\overline{B(z)} + \overline{B(-z)}) = 1;$$

$$P(z) (\overline{A(z)} - \overline{A(-z)}) + Q(z) (\overline{B(z)} - \overline{B(-z)}) = 1,$$

for  $|z| = 1$ ; that is,

$$P(z) \left( \sum_k \bar{a}_{-2k} z^{2k} \right) + Q(z) \left( \sum_k \bar{b}_{-2k} z^{2k} \right) = 1;$$

$$P(z) \left( \sum_k \bar{a}_{-2k+1} z^{2k-1} \right) + Q(z) \left( \sum_k \bar{b}_{-2k+1} z^{2k-1} \right) = 1$$

for  $|z| = e^{-i\omega/2}, \omega \in \mathbb{R}$ . Therefore, with both sides of the above two equations multiplied by  $\hat{\phi}(\frac{\omega}{2})$  and  $z\hat{\phi}(\frac{\omega}{2})$ , respectively, it follows from (the Fourier transform of) the two-scale relations that

$$\hat{\phi}(\omega) = P(e^{-\omega/2}) \hat{\phi}(\frac{\omega}{2});$$

$$\hat{\psi}(\omega) = Q(e^{-\omega/2}) \hat{\phi}(\frac{\omega}{2}),$$

that

$$\begin{aligned}\widehat{\phi}\left(\frac{\omega}{2}\right) &= \sum_k \left( \bar{a}_{-2k} e^{-ik\omega} \widehat{\phi}(\omega) + \bar{b}_{-2k} e^{-ik\omega} \widehat{\psi}(\omega) \right); \\ \widehat{\phi}\left(\frac{\omega}{2}\right) e^{-i\omega/2} &= \sum_k \left( \bar{a}_{-2k+1} e^{-ik\omega} \widehat{\phi}(\omega) + \bar{b}_{-2k+1} e^{-ik\omega} \widehat{\psi}(\omega) \right),\end{aligned}$$

or equivalently, by taking the inverse Fourier transform,

$$\begin{aligned}2\phi(2x) &= \sum_k \left( \bar{a}_{-2k} \phi(x-k) + \bar{b}_{-2k} \psi(x-k) \right); \\ 2\phi(2x-1) &= \sum_k \left( \bar{a}_{-2k+1} \phi(x-k) + \bar{b}_{-2k+1} \psi(x-k) \right).\end{aligned}$$

Finally, by changing  $x$  to  $x-j$  in the above two equations, we have

$$\begin{aligned}\phi(2x-2j) &= \sum_k \frac{1}{2} \left( \bar{a}_{2j-2k} \phi(x-k) + \bar{b}_{2j-2k} \psi(x-k) \right); \\ \phi(2x-(2j+1)) &= \frac{1}{2} \sum_k \left( \bar{a}_{2j+1-2k} \phi(x-k) + \bar{b}_{2j+1-2k} \psi(x-k) \right).\end{aligned}$$

The totality of these two equations is the same as (6.3.26), by setting  $\ell = 2j$  to the first equation and  $\ell = 2j+1$  in the second equation. ■

**Remark 6.3.1** In the next subunit, we will apply (6.3.26) of Theorem 6.3.2 to decompose digital signals into low-frequency and high-frequency bands and will also apply the refinement relation (6.3.2) and the MRA wavelet definition (6.3.25) to reconstruct the original digital signals. Theorem 6.3.2 will also be applied to prove that the MRA architecture introduced in Subunit 6.2.3 can be expanded to include decomposition of the  $L_2(\mathbb{R})$  space into a direct sum of multi-scale wavelet subspaces. This topic, along with the construction of orthogonal and bi-orthogonal MRA wavelets will be studied in the next subunit, Subunit 6.3.3. Extension to the two-dimensional setting, with application to digital image decomposition and compression will be discussed in Subunit 6.5.

Let us now return to the matrix extension problem (6.3.24) of finding the Laurent symbols  $Q(z), A(z), B(z)$ , from a given two-scale symbol  $P(z)$ . As an extension of the QMF discussion in Subunit 6.3.1, where the QMF condition (6.3.17) is equivalent to the totality of (6.3.13), (6.3.14), and (6.3.16), the matrix extension problem (6.3.24) is equivalent to finding desirable symbols  $Q(z), A(z), B(z)$  that satisfy the totality of four conditions:

$$P(z)\overline{A(z)} + P(-z)\overline{A(-z)} = 1, \quad |z| = 1; \quad (6.3.28)$$

$$P(z)\overline{B(z)} + P(-z)\overline{B(-z)} = 0, \quad |z| = 1; \quad (6.3.29)$$

$$A(z)\overline{Q(z)} + A(-z)\overline{Q(-z)} = 0, \quad |z| = 1; \quad (6.3.30)$$

$$B(z)\overline{Q(z)} + B(-z)\overline{Q(-z)} = 1, \quad |z| = 1. \quad (6.3.31)$$

A general procedure to solving the matrix extension problem is first to find the desirable Laurent symbol  $A(z)$  that satisfies the “2-duality” condition (6.3.28) corresponding to the given two-scale symbol  $A(z)$ . Concurrently, depending on the order  $m$  of polynomial preservation by the given refinable function  $\phi(x)$  (from which  $P(z)$  is determined), the symbol  $B(z)$ , with polynomial factor  $(1-z)^n$  for any  $n, 1 \leq n \leq m$ , is to be constructed, such that  $B(z)$  satisfies the “2-orthogonal” property (6.3.29). Then the two-scale symbol  $Q(z)$  is unique and satisfies the identities (6.3.30) and (6.3.31).

We remark that the choice of  $B(z)$  with polynomial factor  $(1-z)^n$  assures the  $n^{\text{th}}$ -order vanishing moment of the “analysis wavelet,” associated with the “dual refinable function,” with  $A(z)$  as its refinement symbol. To be more specific, let us consider the  $m^{\text{th}}$ -order Cardinal  $B$ -spline  $\varphi_m(x)$ , defined in (6.2.14) by  $m$ -fold convolution of the characteristic function of the unit interval  $[0, 1)$ . Hence, by (6.2.15), the given two-scale symbol of the centered  $B$ -spline  $\phi_m(x) := \varphi_m(x + \frac{m}{2})$ , for even  $m$ , is given by

$$\tilde{P}_m(z) := z^{-m/2} P_m(z) = z^{-m/2} \left( \frac{1+z}{2} \right)^m. \quad (6.3.32)$$

**Remark 6.3.2** As discussed in Subunits 6.1.2 and 6.1.3, the wavelet transform defined in (6.1.3) of Subunit 6.1.1, with wavelet kernel  $\psi_{b,a}$  introduced in (6.1.2), is a band-pass filter that annihilates the “low-frequency” content of a given signal (or function)  $f(t)$ , while separating the “high-frequency” content of  $f(t)$  for analysis, by adjusting the scale  $a > 0$ . Here, the notion of “low-frequency” means the “slowly oscillating” component of  $f(t)$ . In applications, an algebraic polynomial is used to describe slow oscillation. Hence, if  $f(t)$  has a Taylor representation

$$f(t) = \sum_{k=0}^{n-1} \frac{1}{k!} f^{(k)}(t_0) (t - t_0)^k + R_n(t),$$

with remainder  $R_n(t)$ , in some  $\epsilon$ -neighborhood  $N(\epsilon, t_0)$  of  $t_0$ , then by using a wavelet  $\psi(t)$  with  $n^{\text{th}}$ -order vanishing moment for the wavelet transform  $(W_\psi f)(b, a)$  in (6.1.3), we have

$$(W_\psi f)(b, a) = (W_\psi R_n)(b, a)$$

by sliding the translation parameter  $b$  and adjusting the scale parameter  $a > 0$  to match  $\psi_{b,a}(t)$  with the neighborhood  $N(\epsilon, t_0)$  of  $t_0$ . This enables the wavelet transform to annihilate the low-frequency content of  $f(t)$  and analyze the high-frequency content  $R_n(t)$  in  $N(\epsilon, t_0)$ . Observe that a higher order of vanishing moments improves the quality of high-frequency analysis.

**Theorem 6.3.3** *Let  $m \geq 2$  be an arbitrarily chosen even integer. Then the Laurent polynomial  $A(z)$  with smallest degree that satisfies the “2-duality” condition (6.3.28), with  $\tilde{P}(z)$  given by (6.3.32), is*

$$A_m(z) = \sum_{j=0}^{m/2-1} \binom{\frac{m}{2} + j - 1}{j} \left[ \frac{1}{2} \left( 1 - \frac{(z + \frac{1}{z})}{2} \right) \right]^j. \quad (6.3.33)$$

Since the most commonly used  $B$ -spline is the centered cubic spline  $\phi_4(x) = \varphi_4(x+2)$ , we only consider the 2-duality property (6.3.28) for  $m = 4$  in (6.3.33) as follows.

**Example 6.3.4** For  $\tilde{P}_4(z) = z^{-2} \left( \frac{1+z}{2} \right)^4$  in (6.3.32), verify that  $A(z) = A_4(z)$  in (6.3.33) satisfies the condition (6.3.28).

**Solution** For  $|z| = 1$ , we have  $\frac{1}{z} = \bar{z}$ , so that

$$\overline{A_4(z)} = \left( 1 + \left( 1 - \frac{z + z^{-1}}{2} \right) \right) = \left( 2 - \frac{z + z^{-1}}{2} \right).$$

Hence, we have

$$\begin{aligned} & \tilde{P}_4(z) \overline{A_4(z)} + \tilde{P}_4(-z) \overline{A_4(z)} \\ &= \frac{1}{2z^2} \left( \frac{1+z}{2} \right)^4 \left( 4 - \left( z + \frac{1}{z} \right) \right) + \frac{1}{2z^2} \left( \frac{1-z}{2} \right)^4 \left( 4 + \left( z + \frac{1}{z} \right) \right) \\ &= \frac{1}{32} \frac{1}{z^2} \left\{ (1 + 4z + 6z^2 + 4z^3 + z^4) \left( 4 - \left( z + \frac{1}{z} \right) \right) + \right. \\ & \quad \left. (1 - 4z + 6z^2 - 4z^3 + z^4) \left( 4 + \left( z + \frac{1}{z} \right) \right) \right\} \\ &= \frac{1}{32} \frac{1}{z^2} \left\{ 4(2 + 12z^2 + 2z^4) - \left( z + \frac{1}{z} \right) (8z + 8z^3) \right\} \\ &= \frac{1}{32} \frac{1}{z^2} \left\{ 8 + 48z^2 + 8z^4 - (8 + 16z^2 + 8z^4) \right\} \\ &= \frac{1}{32} \frac{1}{z^2} (32z^2) = 1. \quad \blacksquare \end{aligned}$$

**Remark 6.3.3** The condition (6.3.28) is equivalent to the sequence (time-domain) duality condition

$$\sum_k p_k \bar{a}_{k-2j} = 2\delta_j, \quad j \in \mathbb{Z}, \quad (6.3.34)$$

where  $\delta_j$  denotes the Kronecker symbol. In view of the “down-sample” by 2 (i.e.  $j \rightarrow 2j$ ) in the convolution, we call (6.3.28), or equivalently (6.3.34),

the 2-duality condition. Similarly, the time-domain orthogonality condition, which is equivalent to (6.3.29), is given by

$$\sum_k p_k \bar{b}_{k-2j} = 0, \quad j \in \mathbb{Z}, \quad (6.3.35)$$

and called the 2-orthogonality condition. Of course, (6.3.30) is equivalent to

$$\sum_k \bar{a}_k q_{k-2j} = 0, \quad j \in \mathbb{Z}, \quad (6.3.36)$$

and (6.3.31) is equivalent to

$$\sum_k \bar{b}_k q_{k-2j} = 2\delta_j, \quad j \in \mathbb{Z}. \quad (6.3.37)$$

**Remark 6.3.4** The notion of 2-duality and of 2-orthogonality will be clear from our study of bi-orthogonal wavelets in the next subunit, particularly Theorem 6.3.5.

**Theorem 6.3.4** *For any even integer  $m \geq 2$ , let  $\phi_m(x) = \varphi_m(x + \frac{m}{2})$  be the  $m^{\text{th}}$  order centered Cardinal B-spline with two-scale symbol  $\tilde{P}_m(x)$  given by (6.3.32) and corresponding 2-dual Laurent polynomial  $A_m(z)$  given by (6.3.33). Then the Laurent symbol  $B(z)$  with smallest Laurent polynomial degree and  $n^{\text{th}}$  order vanishing moment, for any  $n, 1 \leq n \leq m$ , for the matrix extension (6.3.24) is given by*

$$B(z) = B_n(z) = z^{-m/2+1}(1-z)^n. \quad (6.3.38)$$

Furthermore, the Laurent polynomial

$$Q_m(z) := \frac{(-1)^{m/2}}{2^m} z^{-1} C(x) A_m(-z), \quad (6.3.39)$$

for some appropriate polynomial  $C(z)$  (see Example 6.3.7, and in particular, Theorem 6.5.2 of Subunit 6.5.3), can be used to complete the solution of the matrix extension problem (6.3.24).

### 6.3.3 Orthogonal and bi-orthogonal wavelets

The notion of multiresolution analysis (MRA) introduced in Subunit 6.2.3 can now be extended to the decomposition of functions in  $L_2(\mathbb{R})$  by applying Theorem 6.3.2 in Subunit 6.3.2. Precisely, let  $\phi$  be a refinable function that satisfies the conditions as stated in Theorem 6.3.2 of Subunit 6.3.2. Also, recall

$$\mathbb{V}_j = \overline{\text{span}} \{ \phi(2^j x - k) : k \in \mathbb{Z} \}$$

where  $\overline{\text{span}}$  denotes the  $L_2$ -closure of the algebraic span. Then  $\phi$  is said to generate an MRA if  $\{\mathbb{V}_j\}$  is a nested sequence of subspaces of  $L_2(\mathbb{R})$ , with the  $L_2$ -closure of the union of all  $\mathbb{V}_j, j \in \mathbb{Z}$ , being the entire  $L_2(\mathbb{R})$  space. Hence, for any fixed integer  $j$ , by defining

$$\mathbb{W}_j = \overline{\text{span}} \{ \psi(2^j x - k) : k \in \mathbb{Z} \}, \quad (6.3.40)$$

where  $\psi$  is called a “wavelet” as defined in (6.3.25), we observe that both  $\mathbb{V}_j$  and  $\mathbb{W}_j$  are subspaces of  $\mathbb{V}_{j+1}$ , namely:

$$\mathbb{V}_j, \mathbb{W}_j \subset \mathbb{V}_{j+1}.$$

On the other hand, by changing  $x$  to  $2^j x$  in (6.3.26) of Theorem 6.3.2, it is clear that

$$\phi(2^{j+1} x - \ell) \in \mathbb{V}_j + \mathbb{W}_j, \ell \in \mathbb{Z}.$$

Since the  $L_2$ -closure of the linear span of the functions  $\phi(2^{j+1} x - \ell), \ell \in \mathbb{Z}$ , is  $\mathbb{V}_{j+1}$ , we also have

$$\mathbb{V}_{j+1} \subseteq \mathbb{V}_j + \mathbb{W}_j;$$

and thus,

$$\mathbb{V}_{j+1} = \mathbb{V}_j + \mathbb{W}_j, \quad j \in \mathbb{Z}. \quad (6.3.41)$$

Now, by applying the properties (ii) and (iii) in the MRA definition, we have the decomposition property:

$$L_2(\mathbb{R}^2) = \sum_{j=-\infty}^{\infty} \mathbb{W}_j. \quad (6.3.42)$$

In addition, by imposing very mild assumption of “linearly independent integer shifts” the refinable function  $\phi$ ; (that is, the assumption that for  $\{c_k\} \in \ell_1$ ,

$$\sum_{k=-\infty}^{\infty} c_k \phi(x - k) = 0, \quad x \in \mathbb{R},$$

implies  $c_k = 0$  for all  $k$ ), it follows that

$$\mathbb{V}_j \cap \mathbb{W}_j = \{0\}, \quad j \in \mathbb{Z}.$$

Hence, for all  $\ell, k \in \mathbb{Z}$  with  $\ell \neq k$ , say  $\ell < k$ , since  $\mathbb{W}_\ell \subset \mathbb{V}_{\ell+1} \subseteq \mathbb{V}_k$  and  $\mathbb{V}_k \cap \mathbb{W}_k = \{0\}$ , it follows that

$$\mathbb{W}_j \cap \mathbb{W}_k = \{0\}, \quad (6.3.43)$$

for all  $j \neq k$ . We remark that the linearly “independent integer shift” condition is satisfied by all orthogonal refinable functions (see (6.3.8)) and all Cardinal  $B$ -splines  $\varphi_m(x)$  (see (6.2.13)). In view of the above discussion, let us introduce the following two notations:

$$L_2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} \mathbb{W}_j \quad (6.3.44)$$



$$L_2(\mathbb{R}) = \bigoplus_{j=-\infty}^{\infty} \mathbb{W}_j, \quad (6.3.45)$$

where  $\bigoplus$  is called “direct sum” and  $\bigoplus^\perp$  is called “orthogonal sum.” For (6.3.44), it follows from (6.3.43) that every  $f \in L_2(\mathbb{R})$  has a **unique** decomposition

$$f(x) = \sum_{j=-\infty}^{\infty} g_j(x), \quad g_j \in \mathbb{W}_j;$$

and for (6.3.45), this unique decomposition has the additional property that

$$\langle g_j, g_k \rangle = 0, \quad \text{for } j \neq k.$$

Furthermore, the function  $\psi(x)$  in (6.3.25) that generates the subspaces  $\mathbb{W}_j$  in (6.3.40) will be called an “MRA wavelet,” and the subspaces  $\mathbb{W}_j$  are called “wavelet subspaces.”

In the following, we apply the “2-dual sequence”  $\{a_k\}$  and the “2-orthogonal sequence”  $\{b_k\}$  from the matrix extension (6.3.24), as described by (6.3.34) and (6.3.35) respectively, to decompose functions  $f_{j+1} \in \mathbb{V}_{j+1}$  as the sum of a function  $f_j \in \mathbb{V}_j$  and a function  $g_j \in \mathbb{W}_j$ , for any integer  $j \in \mathbb{Z}$ . Here, since

$$\mathbb{V}_{j+1} = \mathbb{V}_j \oplus \mathbb{W}_j, \quad j \in \mathbb{Z},$$

the decomposition  $f_{j+1} = f_j + g_j$  is unique and should be called **direct sum decomposition**. For  $f_{j+1} \in \mathbb{V}_{j+1}$ ,  $f_j \in \mathbb{V}_j$  and  $g_j \in \mathbb{W}_j$ , we may write

$$\begin{aligned} f_{j+1}(x) &= \sum_{\ell} c_{\ell}^{j+1} \phi(2^{j+1}x - \ell); \\ f_j(x) &= \sum_k c_k^j \phi(2^jx - k); \\ g_j(x) &= \sum_k d_k^j \psi(2^jx - k). \end{aligned} \quad (6.3.46)$$

Therefore, the task of decomposition of the function  $f_{j+1}$  into a direct sum of two functions  $f_j$  and  $g_j$  is equivalent to computing the coefficient sequences  $\{c_k^j\}$  and  $\{d_k^j\}$  in terms of the coefficient sequence  $\{c_{\ell}^{j+1}\}$  of  $f_{j+1}$ .

By replacing  $x$  in the decomposition relation (6.3.26) with  $2^j x$ , we have

$$\phi(2^{j+1}x - \ell) = \frac{1}{2} \sum_k \{ \bar{a}_{\ell-2k} \phi(2^jx - k) + \bar{b}_{\ell-2k} \psi(2^jx - k) \}.$$

Hence, from the definition of the coefficient sequence  $\{c_{\ell}^j\}, \{d_{\ell}^j\}$ , it follows that

$$\begin{aligned} f_{j+1}(x) &= \sum_{\ell} c_{\ell}^{j+1} \left[ \sum_k \left\{ \frac{1}{2} \bar{a}_{\ell-2k} \phi(2^jx - k) + \frac{1}{2} \bar{b}_{\ell-2k} \psi(2^jx - k) \right\} \right] \\ &= \sum_k \left( \sum_{\ell} \frac{1}{2} \bar{a}_{\ell-2k} c_{\ell}^{j+1} \right) \phi(2^jx - k) + \\ &\quad + \sum_k \left( \sum_{\ell} \frac{1}{2} \bar{b}_{\ell-2k} c_{\ell}^{j+1} \right) \psi(2^jx - k). \end{aligned}$$

Observe that the function represented by first sum on the right-hand side is in  $\mathbb{V}_j$ , while the function represented by the second sum is in  $\mathbb{W}_j$ . Since the decomposition  $\mathbb{V}_{j+1} = \mathbb{V}_j \oplus \mathbb{W}_j$  is a direct sum, we have

$$f_j(x) = \sum_k c_k^j \phi(2^j x - k) = \sum_k \left( \sum_\ell \frac{1}{2} \bar{a}_{\ell-2k} c_\ell^{j+1} \right) \phi(2^j x - k),$$

and

$$g_j(x) = \sum_k d_k^j \psi(2^j x - k) = \sum_k \left( \sum_\ell \frac{1}{2} \bar{b}_{\ell-2k} c_\ell^{j+1} \right) \psi(2^j x - k).$$

Therefore, assuming that the wavelet  $\psi$  also has the property of “linear independent integer shifts” as the refinable function  $\phi$ , we have derived the following “wavelet decomposition” algorithm:

$$\begin{aligned} c_k^j &= \sum_\ell \frac{1}{2} \bar{a}_{\ell-2k} c_\ell^{j+1}; \\ d_k^j &= \sum_\ell \frac{1}{2} \bar{b}_{\ell-2k} c_\ell^{j+1}, \quad k \in \mathbb{Z}, \end{aligned} \tag{6.3.47}$$

for the decomposition of any  $f_{j+1} \in \mathbb{V}_{j+1}$  into the sum of  $f_j \in \mathbb{V}_j$  and  $g_j \in \mathbb{W}_j$ . This, of course, is valid for any integer  $j \in \mathbb{Z}$ . Hence, by applying the decomposition algorithm (6.3.47) to  $f_j, f_{j-1}, \dots, f_{j-L+1}$  for any arbitrarily chosen non-negative integer  $L$ , we have

$$f_{j+1}(x) = f_{j-L}(x) + g_j(x) + \dots + g_{j-L}(x) \tag{6.3.48}$$

where  $f_{j-L} \in \mathbb{V}_{j-L}$  and  $g_\ell \in \mathbb{W}_\ell$  for  $\ell = j-L, \dots, j$  and the sum in (6.3.48) is unique, being a direct sum.

**Remark 6.3.5** By setting  $g_\ell := \frac{1}{2} \bar{a}_{-\ell}$  and  $h_\ell := \bar{b}_{-\ell}$ , the decomposition algorithm (6.3.47) is precisely the convolution of the given sequence  $\{c_{j+1}^\ell\}$  with the “filters”  $\{g_\ell\}$  and  $\{h_\ell\}$ , respectively, followed by “downsampling,” meaning that the terms with odd indices of the output sequences are dropped. The notation for downsampling is  $2 \downarrow$ . That is, (6.3.47) can be described as follows:

$$\begin{aligned} \{c_k^{j+1}\} &\longrightarrow \star \left\{ \frac{1}{2} \bar{a}_{-\ell} \right\} \longrightarrow \boxed{2 \downarrow} \longrightarrow \{c_k^j\} \\ \{c_k^{j+1}\} &\longrightarrow \star \left\{ \frac{1}{2} \bar{b}_{-\ell} \right\} \longrightarrow \boxed{2 \downarrow} \longrightarrow \{d_k^j\}. \end{aligned}$$

We will call  $\{\frac{1}{2} \bar{a}_\ell\}$  a lowpass filter and  $\{\frac{1}{2} \bar{b}_\ell\}$  a highpass filter.

To reconstruct the function  $f_{j+1}$  from the “lowpass” component  $f_j \in \mathbb{V}_j$  and “highpass” component  $g_j \in \mathbb{W}_j$ , we apply the refinement and two-scale relations:

$$\begin{aligned} \phi(x) &= \sum_\ell p_\ell \phi(2x - \ell); \\ \psi(x) &= \sum_\ell q_\ell \phi(2x - \ell), \end{aligned}$$

in (6.2.8) and (6.3.2), respectively. Indeed, by replacing  $x$  with  $2^j x - k$  in the above two relations, we have

$$\begin{aligned}\phi(2^j x - k) &= \sum_{\ell} p_{\ell} \phi(2^{j+1} x - (2k + \ell)) \\ &= \sum_{\ell} p_{\ell-2k} \phi(2^{j+1} x - \ell),\end{aligned}$$

and similarly,

$$\psi(2^j x - k) = \sum_{\ell} q_{\ell-2k} \phi(2^{j+1} x - \ell).$$

Therefore,

$$\begin{aligned}f_j(x) + g_j(x) &= \sum_k c_k^j \phi(2^j x - k) + \sum_k d_k^j \psi(2^j x - k) \\ &= \sum_k \left[ c_k^j \left( \sum_{\ell} p_{\ell-2k} \phi(2^{j+1} x - \ell) \right) + \right. \\ &\quad \left. + d_k^j \left( \sum_{\ell} q_{\ell-2k} \phi(2^{j+1} x - \ell) \right) \right] \\ &= \sum_{\ell} \left[ \sum_k \left( p_{\ell-2k} c_k^j + q_{\ell-2k} d_k^j \right) \right] \phi(2^{j+1} x - \ell).\end{aligned}$$

Since  $f_{j+1}(x) = \sum_{\ell} c_{\ell}^{j+1} \phi(2^{j+1} x - \ell)$  and the refinable function  $\phi$  satisfies the linearly independent integer shifts condition, we may conclude that

$$c_{\ell}^{j+1} = \sum_k \left( p_{\ell-2k} c_k^j + q_{\ell-2k} d_k^j \right). \quad (6.3.49)$$

We call (6.3.48) the “wavelet reconstruction” algorithm.

**Remark 6.3.6** For reconstruction, the input sequences  $\{c_k^j\}$  and  $\{d_k^j\}$  are first upsampled, by inserting one zero in-between every two consecutive terms; that is, by setting

$$\tilde{c}_n^j = \begin{cases} c_k^j, & \text{for } n = 2k, \\ 0, & \text{for } n = 2k + 1, \end{cases}$$

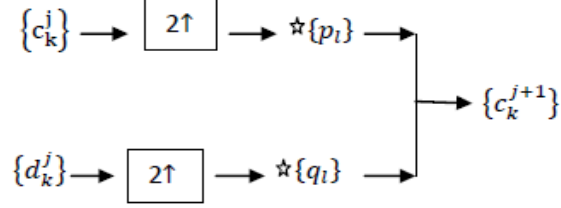
and

$$\tilde{d}_n^j = \begin{cases} d_k^j, & \text{for } n = 2k, \\ 0, & \text{for } n = 2k + 1. \end{cases}$$

Therefore,

$$\sum_k \left( p_{\ell-2k} c_k^j + q_{\ell-2k} d_k^j \right) = \sum_k \left( p_{\ell-n} \tilde{c}_n^j + q_{\ell-n} \tilde{d}_n^j \right)$$

which is the convolution operation. The symbol for upsampling is  $2 \uparrow$ . By applying the upsampling symbol, the wavelet reconstruction algorithm (6.3.49) can be described as follows.



To introduce the notion of bi-orthogonal wavelets, we return to the matrix extension problem studied in the previous subunit and assume that the sequence  $\{a_k\}$ , with two-scale symbol  $A(z) = \frac{1}{2} \sum a_k z^k$  in the matrix extension (6.3.24), is the refinement sequence of some refinable function  $\tilde{\phi} \in L_2(\mathbb{R})$ ; that is,

$$\tilde{\phi}(x) = \sum_k a_k \tilde{\phi}(2x - k). \quad (6.3.50)$$

**Theorem 6.3.5** *Let  $\phi, \tilde{\phi} \in L_2(\mathbb{R})$  be refinable functions with refinement sequences  $\{p_k\}, \{a_k\}$  as in (6.2.8), (6.3.50), respectively, that satisfy the sum rule condition. Suppose that the function pair  $(\phi, \tilde{\phi})$  is a dual pair, namely:*

$$\langle \phi, \tilde{\phi}(\cdot - j) \rangle = \delta_j, \quad j \in \mathbb{Z}, \quad (6.3.51)$$

*then the sequence pair  $(\{p_k\}, \{a_k\})$  is 2-dual, as introduced in (6.3.34), namely:*

$$\sum_k p_k \bar{a}_{k-2j} = 2\delta_j, \quad j \in \mathbb{Z}.$$

**Remark 6.3.7** Under certain appropriate conditions on the symbol  $A(z)$ , the converse of this theorem also holds, in that  $\tilde{\phi} \in L_2(\mathbb{R})$  exists and is dual to  $\phi$ . But the proof is somewhat technical to be included in this writing. In any case, the 2-duality condition (6.3.51) is a necessary condition for the duality of  $\tilde{\phi}$  and  $\phi$ .

To prove the theorem, we simply apply the refinement relations (6.2.8), (6.3.50) and the duality assumption (6.3.51) to compute

$$\begin{aligned} \delta_j &= \langle \phi, \tilde{\phi}(\cdot - j) \rangle = \sum_k \sum_\ell p_k \bar{a}_\ell \langle \phi(2x - k), \tilde{\phi}(2x - (\ell + 2j)) \rangle \\ &= \frac{1}{2} \sum_k \sum_\ell p_k \bar{a}_\ell \langle \phi, \tilde{\phi}(\cdot - (\ell + 2j - k)) \rangle \\ &= \frac{1}{2} \sum_k \sum_n p_k \bar{a}_{n+k-2j} \delta_n \\ &= \frac{1}{2} \sum_k p_k \bar{a}_{k-2j}. \end{aligned}$$

Now, under the existence and duality assumption on  $\tilde{\phi}$ , the other two symbols

$$B(z) = \frac{1}{2} \sum_k b_k z^k, \quad Q_k(z) = \frac{1}{2} \sum_k q_k z^k$$

of the matrix extension (6.3.24) can be applied to introduce two wavelets

$$\tilde{\psi}(x) := \sum_k b_k \tilde{\phi}(2x - k), \quad (6.3.52)$$

and of course,

$$\psi(x) := \sum_k q_k \phi(2x - k),$$

as in (6.3.25). These two wavelets also constitute a dual pair, and the wavelet  $\tilde{\psi}$  in (6.3.52) is orthogonal to  $V_0 := \overline{\text{span}} \langle \phi(x - k) : k \in \mathbb{Z} \rangle$ , as stated in the following.

**Theorem 6.3.6** *Under the assumption that the matrix extension*

$$\mathcal{M}_{P,Q}(z) M_{A,B}^*(z) = I_2$$

*holds for  $|z| = 1$ , the existence of  $\tilde{\phi} \in L_2(\mathbb{R})$ , and the duality (6.3.51) of the pair  $(\phi, \tilde{\phi})$ , then the two wavelets  $\psi$  and  $\tilde{\psi}$  constitute a dual pair, in that*

$$\langle \psi, \tilde{\psi}(\cdot - j) \rangle = \delta_j, \quad j \in \mathbb{Z}, \quad (6.3.53)$$

*and  $\tilde{\psi}(\cdot - j)$  is orthogonal to  $W_0$ , in that*

$$\langle \phi(\cdot - j), \tilde{\psi}(\cdot - k) \rangle = 0, \quad j, k \in \mathbb{Z}, \quad (6.3.54)$$

**Proof** The same computation as in the proof of the above theorem yields:

$$\begin{aligned} \langle \psi, \tilde{\psi}(\cdot - j) \rangle &= \sum_k \sum_\ell q_k \bar{b}_\ell \langle \phi(2x - k), \tilde{\phi}(2x - (\ell + 2j)) \rangle \\ &= \frac{1}{2} \sum_k q_k \bar{b}_{k-2j} = \delta_j, \quad j \in \mathbb{Z}, \end{aligned}$$

by (6.3.37), and

$$\begin{aligned} \langle \phi, \tilde{\psi}(\cdot - j) \rangle &= \sum_k \sum_\ell p_k \bar{b}_\ell \langle \phi(2x - k), \tilde{\phi}(2x - (\ell + 2j)) \rangle \\ &= \frac{1}{2} \sum_k p_k \bar{b}_{k-2j} = 0, \quad j \in \mathbb{Z}, \end{aligned}$$

by (6.3.35). ■

To study the relevance of wavelet decomposition and the wavelet transform, let us apply (6.3.53)–(6.3.54) of Theorem 6.3.6 to the wavelet decomposition

$$f_{j+1}(x) = f_j(x) + g_j(x) \quad (6.3.55)$$

in (6.3.46)–(6.3.48). For  $L = 1$  in (6.3.48), we have (6.3.55), where the representations of the functions  $f_{j+1}, f_j, g_j$  are given in (6.3.46), with corresponding coefficient sequences  $\{c_k^{j+1}\}, \{c_k^j\}, \{d_k^j\}$  governed by the relations in (6.3.47). For any fixed  $j \in \mathbb{Z}$ , it follows from (6.3.54) that

$$\langle f_j, \tilde{\psi}(2^j x - k) \rangle = 0, k \in \mathbb{Z};$$

and from (6.3.53) that

$$\langle g_j, \tilde{\psi}(2^j x - k) \rangle = \sum_{\ell} d_{\ell}^j \langle \psi(2^j x - \ell), \tilde{\psi}(2^j x - k) \rangle = 2^{-k} d_k^j.$$

Hence, the coefficient  $d_k^j$  of  $g_j(x)$  is precisely:

$$d_k^j = 2^j \langle f_{j+1}, \tilde{\psi}(2^j x - k) \rangle = \left( W_{\tilde{\psi}} f_{j+1} \right) \left( \frac{k}{2^j}, \frac{1}{2^j} \right), \quad (6.3.56)$$

as defined in (6.3.3) with

$$b = \frac{k}{2^j}, \quad a = \frac{1}{2^j}$$

for  $f = f_{j+1}$  and wavelet  $\tilde{\psi}$ . In other words, the wavelet decomposition algorithm (6.3.47) provides a very efficient way (simply by discrete convolution followed by downsampling) for computing the wavelet transform

$$(W_{\tilde{\psi}} f)(b, a) = \frac{1}{a} \int_{-\infty}^{\infty} f(t) \overline{\tilde{\psi}\left(\frac{t-b}{a}\right)} dt \quad (6.3.57)$$

of  $f(x) = f_{j+1}(x)$  at the time-scale position

$$(b, a) = \left( \frac{k}{2^j}, \frac{1}{2^j} \right), \quad k \in \mathbb{Z}; \quad (6.3.58)$$

and this applies to all scale levels  $j \in \mathbb{Z}$ .

**Definition 6.3.2** *The integral transform (6.3.57) with wavelet kernel*

$$\tilde{\psi}_{b,a}(t) = \frac{1}{a} \tilde{\psi}\left(\frac{t-b}{a}\right)$$

as previously defined in (6.1.3) is called the **continuous wavelet transform (CWT)** of any function  $f \in L_2(\mathbb{R})$ ; and when  $(b, a)$  is evaluated at the discrete time-scale positions in (6.3.58), then the CWT (6.3.57) becomes

$$\left( W_{\tilde{\psi}} f \right) \left( \frac{k}{2^j}, \frac{1}{2^j} \right) = 2^j \int_{-\infty}^{\infty} f(t) \overline{\tilde{\psi}(2^j t - k)} dt, \quad (6.3.59)$$

which is called the **discrete wavelet transform** (DWT). Furthermore, the wavelet  $\tilde{\psi}$  in (6.3.57) and (6.3.59) is called an “analysis wavelet”. In addition the MRA wavelet  $\psi$  corresponding to the scaling function  $\phi$  is called the “synthesis wavelet” of the MRA architecture.

**Remark 6.3.8** In the wavelet literature, it is more common to define the CWT and DWT by

$$(W_{\tilde{\psi}} f)(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \overline{\tilde{\psi}\left(\frac{t-b}{a}\right)} dt$$

with  $a^{-1/2}$  normalization.

Observe that the wavelet kernel  $\tilde{\psi}_{b,a}$  in the above definition (6.3.57) (see also (6.1.2) in Subunit 6.1) preserves  $L_1(\mathbb{R})$  energy in that

$$\|\tilde{\psi}_{b,a}\|_{L_1(\mathbb{R})} = \int_{-\infty}^{\infty} |\tilde{\psi}_{b,a}(t)| dt = \|\tilde{\psi}\|_{L_1(\mathbb{R})},$$

for all  $a, b \in \mathbb{R}$ , with  $a > 0$ ; while the kernel

$$\frac{1}{\sqrt{a}} \tilde{\psi}\left(\frac{t-b}{a}\right)$$

preserves  $L_2(\mathbb{R})$  energy, in that

$$\begin{aligned} \left\| \frac{1}{\sqrt{a}} \tilde{\psi}\left(\frac{t-b}{a}\right) \right\|_{L_2(\mathbb{R})}^2 &= \int_{-\infty}^{\infty} \left| \frac{1}{\sqrt{a}} \tilde{\psi}\left(\frac{t-b}{a}\right) \right|^2 dt \\ &= \left\| \tilde{\psi} \right\|_{L_2(\mathbb{R})}^2, \end{aligned}$$

for all  $a, b \in \mathbb{R}$ , with  $a > 0$ .

Let us summarize the results obtained in (6.3.46)–(6.3.48) and (6.3.56), as well as (6.3.49) in the following.

**Theorem 6.3.7** Let  $\phi \in (L_1 \cap L_2)(\mathbb{R})$  generate an MRA with corresponding MRA wavelet synthesis wavelet  $\psi$ , in that

$$V_j = \overline{\text{span}}\{\phi(2^j x - k) : k \in \mathbb{Z}\};$$

$$W_j = \overline{\text{span}}\{\psi(2^j x - k) : k \in \mathbb{Z}\}$$

and

$$V_{J+1} = V_{J-L} \oplus W_J \oplus \cdots \oplus W_{J-L}$$

for arbitrarily chosen integers  $J$  and  $L$  with  $L \geq 0$ . Also let  $\tilde{\phi} \in L_2(\mathbb{R})$  be the refinable function generated by some matrix extension (6.3.25), such that  $(\phi, \tilde{\phi})$  satisfies the duality condition (6.3.51), and  $\tilde{\psi}$  be the analysis wavelet

corresponding to  $\tilde{\phi}$ , again from the same matrix extension (6.3.25). Then for any function  $f_{J+1} \in \mathbb{V}_{J+1}$ , the wavelet decomposition algorithm (6.3.47) can be applied to compute the DWT

$$d_k^j = \left( W_{\tilde{\psi}} f_{J+1} \right) \left( \frac{k}{2^j}, \frac{1}{2^j} \right)$$

of  $f_{J+1}$ , with analysis wavelet  $\tilde{\psi}$ , for any  $j = J - L, \dots, J$  and all  $k \in \mathbb{Z}$ . In addition, the wavelet reconstruction algorithm (6.3.49) can be applied to perfectly recover  $f_{J+1}$  from the DWT  $\{d_k^j\}$ ,  $j = J - L, \dots, J$  and the lowpass sequence  $\{c_k^{J-L}\}$ .

We remark that since  $V_{J+1} \approx L_2(\mathbb{R})$  for sufficiently large  $J$ , any function  $f \in L_2(\mathbb{R})$  can be approximated as closely as desired by some  $f_{J+1} \in V_{J+1}$ . Then the DWT  $\{d_k^j\}$ ,  $j = J - L, \dots, J$ , can be considered as the DWT of the given function  $f \in L_2(\mathbb{R})$ . In addition, since  $\mathbb{V}_{J-L} \approx \{0\}$  for sufficiently large  $L$ , the lowpass sequence  $\{c_k^{J-L}\}$  can be ignored. However, in most applications,  $\{c_k^{J-L}\}$  is used as a “thumb-nail” of  $f$ , while the DWT sequences  $\{d_k^j\}$  are used for analyzing the “details” of  $f$ . In Subunit 6.5.1–6.5.2, we will elaborate on this concept in our study of digital image analysis and compression.

We next turn to the discussion of the construction of refinable functions  $\phi, \tilde{\phi}$  and their corresponding bi-orthogonal wavelets. For the special case when  $\mathcal{M}_{P,Q}$  is a QMF, recall that the only task is to construct the two-scale symbol

$$P(z) = \frac{1}{2} \sum_k p_k z^k,$$

or equivalently the refinement sequence  $\{p_k\}$ , since  $A(z), B(z)$ , and  $Q(z)$  are simply  $A(z) = P(z)$ ,  $B(z) = Q(z)$ , and  $Q(z) = -z^{2k+1} \overline{P(-z)}$  for an arbitrary integer  $k$  (see (6.3.28)). On the other hand, for the general matrix extension, the most common choice of  $P(z)$  is

$$P(z) = P_m(z) = \left( \frac{1+z}{2} \right)^m,$$

$m \geq 2$ , or any phase shift of  $P_m(z)$  such as the centered  $z^{-\lfloor m/2 \rfloor} P_m(z)$ . That is, the refinable function  $\phi$  of choice is

$$\phi(x) = \varphi_m(x + \lfloor m/2 \rfloor),$$

where  $\varphi_m$  is the  $m^{\text{th}}$  order Cardinal  $B$ -spline defined by  $m$ -fold convolution of the characteristic function of the unit interval.

In any case, for practical applications, since only finite filters are used for wavelet decomposition (6.3.47) and wavelet reconstruction (6.3.49), we are only interested in Laurent polynomials  $P(z), A(z), B(z), Q(z)$ .

We begin with considering the QMF and the construction of Laurent polynomials  $P(z)$  that satisfied

$$|P(z)|^2 + |P(-z)|^2 = 1, \quad |z| = 1. \quad (6.3.60)$$



Let  $m \geq 1$  be any integer, and consider

$$P(z) = \left( \frac{1+z}{2} \right)^m S(z), \quad (6.3.61)$$

where  $S(z)$  is a Laurent polynomial to be constructed according to the specification (6.3.60). Setting  $z = e^{-i\theta}$ ,  $\theta \in \mathbb{R}$ , we have

$$|P(z)|^2 + |P(-z)|^2 = |S(z)|^2 \cos^{2m} \left( \frac{\theta}{2} \right) + |S(-z)|^2 \sin^{2m} \left( \frac{\theta}{2} \right).$$

Next, replace  $|S(e^{-i\theta})|^2$  by  $p_\ell(\sin^2(\frac{\theta}{2}))$  where  $p_\ell$  is an algebraic polynomial of degree  $\ell$  (to be determined). Then since  $|S(-e^{-i\theta})| = |S(e^{-i(\theta-\pi)})|$ , we have  $|S(-z)|^2 = p_\ell(\cos^2(\frac{\theta}{2}))$ . This translates the condition to

$$p_\ell \left( \sin^2 \frac{\theta}{2} \right) \cos^{2m} \frac{\theta}{2} + p_\ell \left( \cos^2 \frac{\theta}{2} \right) \sin^{2m} \frac{\theta}{2} = 1$$

or equivalently

$$p_\ell(x)(1-x)^m + p_\ell(1-x)x^m = 1, \quad (6.3.62)$$

where  $x := \sin^2 \frac{\theta}{2}$ . To determine the algebraic polynomial  $p_\ell(x)$  in (6.3.62), observe that

$$\begin{aligned} p_\ell(x) &= (1-x)^{-m} (1-x^m p_\ell(1-x)) \\ &= \sum_{k=0}^{\infty} \binom{m+k-1}{k} x^k (1-x^m p_\ell(1-x)) \end{aligned}$$

where the Taylor expansion of  $(1-x)^{-m}$  at  $x=0$  is used. Hence, the polynomial  $p_\ell(x)$  is given by

$$p_\ell(x) = \sum_{k=0}^{m-1} \binom{m+k-1}{k} x^k + x^m r_0(x)$$

where  $r_0(x)$  is a power series with non-negative powers of  $x$ , so that the lowest power of  $x$  of the power series  $x^m r_0(x)$  is at least  $m$ . This leads to the choice of  $\ell = m-1$  to arrive at

$$p_{m-1}(x) = \sum_{k=0}^{m-1} \binom{m+k-1}{k} x^k \quad (6.3.63)$$

for the polynomial  $p_\ell(x)$  in (6.3.62). Returning to  $|S(z)|^2 = p_{m-1}(\sin^2 \frac{\theta}{2})$ ,

since  $z = e^{-i\theta}$ , we have, by (6.3.63),

$$\begin{aligned} |S(z)|^2 &= \sum_{k=0}^{m-1} \binom{m+k-1}{k} \left(1 - \left(\frac{e^{i\theta/2} + e^{-i\theta/2}}{2}\right)^2\right)^k \\ &= \sum_{k=0}^{m-1} \binom{m+k-1}{k} \left(1 - \frac{z + 2 + z^{-1}}{4}\right)^k \\ &= \sum_{k=0}^{m-1} \binom{m+k-1}{k} \left(\frac{2 - z - z^{-1}}{4}\right)^k, \end{aligned}$$

and therefore, the desired two-scale polynomial  $P(z) =: P_{D,2m}(z)$  of the QMF can be computed by the taking “square-root” of

$$|P_{D,2m}(z)|^2 = \left(\frac{2 + z + z^{-1}}{4}\right)^m \sum_{k=0}^{m-1} \binom{m+k-1}{k} \left(\frac{2 - z - z^{-1}}{4}\right)^k. \quad (6.3.64)$$

**Example 6.3.5** For  $m = 1$  in (6.3.64), we have

$$|P_{D,2}(z)|^2 = \frac{2 + z + z^{-1}}{4} = \frac{1+z}{2} \cdot \frac{1+z^{-1}}{2} = \frac{1+z}{2} \left(\frac{1+\overline{z}}{2}\right)$$

where  $|z| = 1$ . Hence,  $P_{D,2}(z) = \frac{1}{2}(1+z)$ , so that the refinement sequence is

$$p_k = \begin{cases} 1 & \text{for } k = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

That is, the refinable function is simply

$$\phi_{D,2}(x) = \chi_{[0,1)}(x) = \varphi_1(x),$$

the first order Cardinal  $B$ -spline. ■

We remark that although it is fairly easy to find the square root for  $m = 1$ , the difficulty increases dramatically for larger integers  $m$ . In the next example, we will see that finding the square root for  $m = 2$  is already somewhat tricky. In general, a systematic method by applying the “Riesz Lemma” allows us to compute the square root for any integer  $m \geq 2$ . This topic will not be discussed in this writing.

**Example 6.3.6** For  $m = 2$  in (6.3.64), we have

$$\begin{aligned}
 |P_{D,4}(z)|^2 &= \left( \frac{2+z+z^{-1}}{4} \right)^2 \left( 1 + 2 \cdot \frac{2-z-z^{-1}}{4} \right) \\
 &= \left( \frac{1+z}{2} \right)^2 \left( \frac{1+z^{-1}}{2} \right)^2 \frac{4-z-z^{-1}}{2} \\
 &= \left( \frac{1+z}{2} \right)^2 \left( \frac{1+z^{-1}}{2} \right)^2 \frac{[(2+\sqrt{3})-z][(2+\sqrt{3})-z^{-1}]}{2(2+\sqrt{3})} \\
 &= \left( \frac{1+2z+z^2}{4} \cdot \frac{(2+\sqrt{3})-z}{\sqrt{4+2\sqrt{3}}} \right) \left( \frac{1+2z+z^2}{4} \cdot \frac{(2+\sqrt{3})-z^{-1}}{\sqrt{4+2\sqrt{3}}} \right).
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 P_{D,4}(z) &= \frac{(1+2z+z^2)(2+\sqrt{3}-z)}{4\sqrt{4+2\sqrt{3}}} \\
 &= \frac{(1+2z+z^2)(2+\sqrt{3}-z)(\sqrt{3}-1)}{8},
 \end{aligned}$$

since  $\frac{1}{\sqrt{4+2\sqrt{3}}} = \frac{\sqrt{3}-1}{2}$ . That is, by writing

$$P_{D,4}(z) = \frac{1}{2} \sum_{k=0}^3 p_{D,k} z^k,$$

it follows that the refinement sequence is given by

$$p_{D,0} = \frac{1+\sqrt{3}}{4}, \quad p_{D,1} = \frac{3+\sqrt{3}}{4}, \quad p_{D,2} = \frac{3-\sqrt{3}}{4}, \quad p_{D,3} = \frac{1-\sqrt{3}}{4},$$

and  $p_{D,k} = 0$  for  $k \neq 0, 1, 2, 3$ . Observe that

$$p_{D,0} + p_{D,2} = \frac{1+\sqrt{3}}{4} + \frac{3-\sqrt{3}}{4} = 1$$

and

$$p_{D,1} + p_{D,3} = \frac{3+\sqrt{3}}{4} + \frac{1-\sqrt{3}}{4} = 1.$$

Thus, the refinement sequence satisfies the sum rule condition. ■

**Remark 6.3.9** The sequences  $\{p_{D,k}\} = \{p_{D,2m,k}\}$ ,  $m = 1, 2, \dots$ , are refinement sequences of the Daubechies orthonormal scaling functions  $\phi_{D,2m}(x)$ . By setting  $q_{D,2m,k} = (-1)^k p_{D,2m,1-k}$ , the orthonormal wavelets

$$\psi_{D,2m}(x) := \sum_k q_{D,2m,k} \phi_{D,2m}(2x - k)$$

are called Daubechies wavelets. These are the first compactly supported refinable functions and wavelets that are orthonormal and reasonably smooth.

We now give examples of the matrix extension problem by considering the most popular (spline) two-scale symbol

$$P(z) = P_m(z) = \left(\frac{1+z}{2}\right)^m.$$

**Example 6.3.7** For  $m = 2$ , the refinement function  $\phi$  is the linear Cardinal  $B$ -spline  $\varphi_2(x)$ . The 2-dual symbol corresponding to  $P_2(z)$  is the Laurent polynomial

$$A_2(z) = \frac{1}{8}(1+z)^2(4-z-z^{-1}).$$

It is easy to verify that

$$P_2(z) \overline{A_2(z)} + P_2(-z) \overline{A_2(-z)} = 1$$

for  $|z| = 1$ . The choice of  $A_2(z)$  follows the recipe (6.3.33). By choosing  $n = m = 2$  in (6.3.38) we achieve the maximum order of vanishing moments for the dual wavelet  $\psi$ , namely:

$$B_2(z) = -\frac{1}{4}z\left(1 - \frac{1}{z}\right)^2,$$

where a different constant multiple and shift are applied. Finally, by choosing  $C(z) = z^2 + 4z + 1$  in (6.3.39), we have

$$Q_2(z) = -\frac{1}{8}\left(z^2 + \frac{1}{z^2}\right) - \frac{1}{4}\left(z + \frac{1}{z}\right) + \frac{3}{4}$$

where a constant multiple and appropriate shift are also applied. The reason for the above minor adjustment of (6.3.38) and (6.3.39) is to give the so-called “5/3 biorthogonal filters” adopted by the JPEG-2000 standard for wavelet image compression. This topic will be discussed in Subunit 6.5.3.

It is easy to verify that

$$\begin{bmatrix} P_2(z) & P_2(-z) \\ Q_2(z) & Q_2(-z) \end{bmatrix} \begin{bmatrix} A_2(z^{-1}) & B_2(z^{-1}) \\ A_2(-z^{-1}) & B_2(-z^{-1}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

for  $|z| = 1$ . ■

More examples will be given in Subunit 6.5.3 in the construction of wavelets with dyadic filter taps for wavelet decomposition and reconstruction.

## 6.4 Wavelet Algorithms

This subunit is a compilation of the wavelet decomposition and reconstruction algorithms, their generalization to filter banks, and their implementation by applying the lifting scheme.

### 6.4.1 Wavelet decomposition and reconstruction

#### References

- (1) This MA 304 text, Subunit 6.3.3: Equations (6.3.47)–(6.3.49) and the two diagrams.
- (2) This MA 304 text, Subunit 6.5.2: Diagrams for “Two-dimensional wavelet decomposition and “Two-dimensional wavelet reconstruction.

### 6.4.2 Filter Banks

#### Reference

- (1) Gilbert Strang, “Lecture Notes: Handouts 116, MIT open courseware.
- (2) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 419–432. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

### 6.4.3 The Lifting Scheme

#### Reference

- (1) Gilbert Strang, “Lecture Notes: Handouts 1-16, MIT open courseware.
- (2) Charles K. Chui and Qingtang Jiang, “Applied Mathematics: Data Compression, Spectral Methods, Fourier Analysis, Wavelets, and Applications, pages 479–498. Atlantis Press, ISBN 978-94-6239-009-6, available on Springer internet platform: [www.springerlink.com](http://www.springerlink.com).

## 6.5 Application to Image Coding

In this subunit, the time variable  $t$  of a “signal”  $f(t)$ , studied previously, is replaced by two spatial variables  $x$  and  $y$  of an “image”  $f(x, y)$ , for  $(x, y)$  in an image domain  $D$ , which is usually a bounded rectangular region, say  $D = [0, b] \times [0, c] \subset \mathbb{R}^2$ , where  $b, c > 0$ . We will discuss the extension of the wavelet transform for signals to the wavelet transform for images, by generalizing the one-dimensional domain to two dimensions. In Subunit 6.5.1, the

theory developed in Subunit 6.3 is applied to map an image defined on  $D$  to a hierarchy of sub-images, by applying a combination of smoothing and wavelet operators, to reveal “low-frequency” and “high-frequency” contents of the images. In practice, low-frequency sub-images can be used as image thumbnails, while the image details in the high-frequency sub-images facilitate such applications as image compression and image edge extraction. The wavelet decomposition and reconstruction algorithms, (6.3.47) and (6.3.49), respectively, derived in Subunit 6.3.3, will be formulated in two-dimensions and applied in Subunit 6.5.2 to decompose a given image into its (wavelet) image hierarchy, and to reconstruct the given image from its wavelet image hierarchy, with application to progressive image transmission and acquisition. Application to image compression will be studied in Subunit 6.5.3, where the image compression industry standard, JPEG-2000, is also discussed.

### 6.5.1 Mapping digital images to the wavelet domain

In this subunit, the theory developed in Subunits 6.3.1–6.3.3 is applied to map an (image) function to the wavelet domain, in the form of an image hierarchy. Let  $D = [0, b] \times [0, c]$ , where  $b, c > 0$ . A function  $f(x, y)$  defined on  $D$  will be used to represent an image.

Let  $\phi$  be a refinable (also called scaling) function introduced and discussed in Subunit 6.2.3. The extension of this MRA architecture from the one-dimensional space  $L_2(\mathbb{R})$  to the two-dimensional space  $L_2(\mathbb{R}^2)$  is easily accomplished by considering the tensor product

$$\Phi(x, y) := \phi(x) \phi(y), \quad (x, y) \in \mathbb{R}^2, \quad (6.5.1)$$

and setting

$$\mathbf{V}_j := \overline{\text{span}} \{ \Phi(2^j x - k, 2^j y - \ell) : k, \ell \in \mathbb{Z} \} \quad (6.5.2)$$

where  $\overline{\text{span}}$  denotes the closure in  $L_2(\mathbb{R}^2)$  of the (linear) algebraic span. It is easy to verify that the properties (i)–(v) in the definition of MRA in Subunit 6.2.3 for  $\phi$  and the nested subspaces  $\{V_j\}$  remain valid for  $\Phi(x, y)$  and the nested subspaces  $\{\mathbf{V}_j\}$  in (6.5.2). In particular, in view of property (iii), while every function  $f(x) \in L_2(\mathbb{R})$  can be approximated as closely as desired by

$$\sum_k c_k^j \phi(2^j x - k),$$

for some sequence  $\{c_k^j\} \in \ell_2 = \ell_2(\mathbb{Z})$ . Every function  $f(x, y)$  in  $L_2(\mathbb{R}^2)$  can also be approximated as closely as desired by

$$\sum_k \sum_\ell c_{k,\ell}^j \Phi(2^j x - k, 2^j y - \ell),$$

for some sequence  $\{c_{k,\ell}^j\} \in \ell_2 = \ell_2(\mathbb{Z}^2)$ , for sufficiently large positive integers

$j \in \mathbb{Z}$ . For this reason, we will always consider functions in  $V_{J+1} \subset L_2(\mathbb{R})$  or  $\mathbf{V}_{J+1} \subset L_2(\mathbb{R}^2)$ , respectively, where  $0 < J \in \mathbb{Z}$  is considered to be sufficiently large.

Next, returning to the matrix extension of the two-scale symbol  $P(z) = \frac{1}{2} \sum_k p_k z^k$ , where  $\{p_k\}$  is the refinement sequence of  $\phi(x)$ , namely:

$$\begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix} \begin{bmatrix} \overline{A(z)} & \overline{B(z)} \\ \overline{A(-z)} & \overline{B(-z)} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (6.5.3)$$

where  $P(z)$ ,  $A(z)$ ,  $B(z)$ , and  $Q(z)$  are assumed to be Laurent polynomials from now on, and assuming that  $A(z) = \frac{1}{2} \sum_k a_k z^k$  is also the two-scale symbol of a refinable function  $\tilde{\phi}(x) \in L_2(\mathbb{R})$ , as in (6.3.50) and Theorem 6.3.3 of Subunit 5.3.3, it follows that

$$\begin{aligned} \tilde{\psi}(x) &:= \sum_k b_k \tilde{\phi}(2x - k); \\ \psi(x) &:= \sum_k q_k \phi(2x - k) \end{aligned} \quad (6.5.4)$$

are also in  $L_2(\mathbb{R})$ . Furthermore, by taking the Fourier transform, we have

$$\begin{aligned} \hat{\phi}(\omega) &= P(e^{-i\omega/2}) \hat{\phi}\left(\frac{\omega}{2}\right); \\ \hat{\psi}(\omega) &= Q(e^{-i\omega/2}) \hat{\phi}\left(\frac{\omega}{2}\right); \\ \hat{\tilde{\phi}}(\omega) &= A(e^{-i\omega/2}) \hat{\tilde{\phi}}\left(\frac{\omega}{2}\right); \\ \hat{\tilde{\psi}}(\omega) &= B(e^{-i\omega/2}) \hat{\tilde{\phi}}\left(\frac{\omega}{2}\right). \end{aligned} \quad (6.5.5)$$

Now, along with the refinable (or scaling) function  $\Phi(x, y) \in L_2(\mathbb{R}^2)$  introduced in (6.5.1), we introduce the following three synthesis wavelets:

$$\begin{aligned} \Psi^1(x, y) &:= \phi(x) \psi(y); \\ \Psi^2(x, y) &:= \psi(x) \phi(y); \\ \Psi^3(x, y) &:= \psi(x) \psi(y). \end{aligned} \quad (6.5.6)$$

In the next subunit (Subunit 6.5.2), we will apply the decomposition and reconstruction algorithms, (6.3.47) and (6.3.49) respectively, derived in Subunit 6.3.3, to decompose each function

$$f_{j+1}(x, y) := \sum_{\ell} \sum_m c_{\ell, m}^{j+1} \Phi(2^{j+1}x - \ell, 2^{j+1}y - m) \quad (6.5.7)$$

into a (direct) sum of four components:

$$\begin{aligned}
 f_j(x, y) &:= \sum_{\ell} \sum_m c_{\ell, m}^j \Phi(2^j x - \ell, 2^j y - m); \\
 g_{1, j}(x, y) &:= \sum_{\ell} \sum_m d_{1; \ell, m}^j \Psi^1(2^j x - \ell, 2^j y - m); \\
 g_{2, j}(x, y) &:= \sum_{\ell} \sum_m d_{2; \ell, m}^j \Psi^2(2^j x - \ell, 2^j y - m); \\
 g_{3, j}(x, y) &:= \sum_{\ell} \sum_m d_{3; \ell, m}^j \Psi^3(2^j x - \ell, 2^j y - m),
 \end{aligned} \tag{6.5.8}$$

namely:

$$f_{j+1}(x, y) = f_j(x, y) + \sum_{n=1}^3 g_{n, j}(x, y), \tag{6.5.9}$$

for  $j = J - L + 1, \dots, J$ , for any desired number of  $L$  levels of decomposition of any given function  $f(x, y) := f_{J+1}(x, y)$ , where  $L \geq 1$ .

**Definition 6.5.1** For  $f(x, y) = f_{J+1}(x, y)$ , the decomposition (6.5.9) maps the “image”  $f(x, y)$  to the hierarchy of sub-images

$$\{f_{j+1-L}(x, y), g_{1, j}(x, y), g_{2, j}(x, y), g_{3, j}(x, y) : j = J + 1 - L, \dots, J\} \tag{6.5.10}$$

for any desirable number  $L \geq 1$  of the hierarchy. In addition, the corresponding set of coefficients

$$\{\{c_{\ell, m}^{j+1-L}\}, \{d_{n; \ell, m}^j\} : n = 1, 2, 3; j = J + 1 - L, \dots, J\} \tag{6.5.11}$$

is called the **representation** of the given image  $f(x, y)$  in the wavelet domain.

**Remark 6.5.1** In (6.5.10), if  $L = 1$  is selected, then  $f(x, y) = f_{J+1}(x, y)$  is decomposed as a direct-sum of four sub-images  $f_J(x, y), g_{1, J}(x, y), g_{2, J}(x, y), g_{3, J}(x, y)$ ; to be called the  $LL, LH, HL, HH$  bands, respectively, of  $f(x, y)$ . Here, “L” stands for “low-frequency” and “H” stands for “high-frequency.” Furthermore, again for one level decomposition, the wavelet-domain representation of  $f(x, y)$  consists of four sequences:  $\{c_{\ell, m}^J\}, \{d_{1; \ell, m}^J\}, \{d_{2; \ell, m}^J\}, \{d_{3; \ell, m}^J\}$ , which reveal the  $LL, LH, HL, HH$  contents, respectively, of  $f(x, y)$ . In applications, the number  $L$  of decomposed levels is chosen to be larger than 1, so that  $f_{j+1-L}(x, y)$  is used as the “thumb-nail” of the given image, and the family of sequences  $\{d_{n; \ell, m}^j\}, n = 1, 2, 3$  and  $j = J + 1 - L, \dots, J$ , is used to reveal the multi-level wavelet details of the image. Observe that the size of the thumb-nail is  $4^{-L}$  of that of the original image  $f(x, y)$ .

To understand the reason for the terminology of thumb-nails and wavelet details, let us first introduce the “smoothing transform”, defined by

$$(S_{\tilde{\phi}} f)(b, a) := \frac{1}{a} \int_{-\infty}^{\infty} f(x) \overline{\tilde{\phi}\left(\frac{x-b}{a}\right)} dx, \tag{6.5.12}$$



for all functions  $f \in L_2(\mathbb{R})$ , where  $\tilde{\phi}$  is a refinable (or scaling) function, such as the  $\tilde{\phi}$  in (6.3.50) and Theorem 6.3.3. of Subunit 5.3.3,  $b \in \mathbb{R}$ , and  $a > 0$ . Of course, the formulation of the smoothing transform  $S_{\tilde{\phi}}$  is the same as the wavelet transform  $W_{\tilde{\psi}}$ , defined by

$$(W_{\tilde{\psi}}f)(b, a) := \frac{1}{a} \int_{-\infty}^{\infty} f(x) \overline{\tilde{\psi}\left(\frac{x-b}{a}\right)} dx,$$

where  $\tilde{\psi}$  is an analysis wavelet, such as the wavelet associated with  $\tilde{\phi}$  as in (6.5.4). However,  $S_{\tilde{\phi}}$  and  $W_{\tilde{\psi}}$  serve two different, yet complementary, purposes. For example, one might think of  $W_{\tilde{\psi}}$  as a “high-pass filter” and  $S_{\tilde{\phi}}$  as a “low-pass filter”. Indeed, in the frequency domain, it follows from Plancherel’s identity that

$$\begin{aligned} (W_{\tilde{\psi}}f)(b, a) &= \langle f, \tilde{\psi}_{b,a} \rangle = \frac{1}{2\pi} \langle \hat{f}, \widehat{\tilde{\psi}_{b,a}} \rangle \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \widehat{\tilde{\psi}}(a\omega) e^{ib\omega} d\omega, \end{aligned}$$

and

$$(S_{\tilde{\phi}}f)(b, a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) \widehat{\tilde{\phi}}(a\omega) e^{ib\omega} d\omega,$$

where  $\widehat{\tilde{\psi}}(a\omega)$  and  $\widehat{\tilde{\phi}}(a\omega)$  are window functions of the above two inverse short-time Fourier transforms (STFT). Since

$$\widehat{\tilde{\psi}}(0) = \int_{-\infty}^{\infty} \tilde{\psi}(x) dx = 0$$

and

$$\widehat{\tilde{\phi}}(0) = \int_{-\infty}^{\infty} \tilde{\phi}(x) dx = 1,$$

the inverse STFT, with window function  $\widehat{\tilde{\psi}}(a\omega)$ , “ignores” the low-frequency content of  $f$ , particularly for large scale  $a > 0$ ; while the inverse STFT, with window function  $\widehat{\tilde{\phi}}(a\omega)$ , “retains” the low-frequency content of  $f$ , particularly for large scale  $a > 0$ .

When the two transforms  $S_{\tilde{\phi}}$  and  $W_{\tilde{\psi}}$  are applied to functions  $f(x, y)$  of two variables  $x$  and  $y$ , we consider one variable at a time, by introducing the superscripts “1” and “2”, when they are applied to the first variable  $x$  and second variable  $y$ , respectively; namely, for fixed values of  $y$ ,

$$\begin{aligned} (S_{\tilde{\phi}}^1 f)(b, y; a) &:= \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\phi}\left(\frac{x-b}{a}\right)} dx; \\ (W_{\tilde{\psi}}^1 f)(b, y; a) &:= \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\psi}\left(\frac{x-b}{a}\right)} dx, \end{aligned} \quad (6.5.13)$$

while for fixed values of  $x$ ,

$$\begin{aligned} (S_{\phi}^2 f)(x, b; a) &:= \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\phi}\left(\frac{y-b}{a}\right)} dy; \\ (W_{\psi}^2 f)(x, b; a) &:= \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\psi}\left(\frac{y-b}{a}\right)} dy. \end{aligned} \quad (6.5.14)$$

Next, analogous to the (tensor-product) synthesis wavelets  $\Psi^1(x, y)$ ,  $\Psi^2(x, y)$  and  $\Psi^3(x, y)$  introduced in (6.5.6), we also need the (tensor product) analysis wavelets:

$$\begin{aligned} \tilde{\Psi}^1(x, y) &:= \tilde{\phi}(x) \tilde{\psi}(y); \\ \tilde{\Psi}^2(x, y) &:= \tilde{\psi}(x) \tilde{\phi}(y); \\ \tilde{\Psi}^3(x, y) &:= \tilde{\psi}(x) \tilde{\psi}(y). \end{aligned} \quad (6.5.15)$$

We now apply (6.5.13)–(6.5.14) to introduce the two-dimensional wavelet transforms  $W_{\tilde{\Psi}^1}$ ,  $W_{\tilde{\Psi}^2}$ , and  $W_{\tilde{\Psi}^3}$ , with analysis wavelets  $\tilde{\Psi}^1$ ,  $\tilde{\Psi}^2$  and  $\tilde{\Psi}^3$ , respectively, as follows: For  $f(x, y) \in L_2(\mathbb{R}^2)$ ,

$$\begin{aligned} (W_{\tilde{\Psi}^1} f)(b_1, b_2; a) &:= (W_{\psi}^2 S_{\phi}^1 f)(b_1, b_2; a) \\ &= \frac{1}{a} \int_{-\infty}^{\infty} \left( \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\phi}\left(\frac{x-b_1}{a}\right)} dx \right) \overline{\tilde{\psi}\left(\frac{y-b_2}{a}\right)} dy; \\ (W_{\tilde{\Psi}^2} f)(b_1, b_2; a) &:= (S_{\psi}^2 W_{\phi}^1 f)(b_1, b_2; a) \\ &= \frac{1}{a} \int_{-\infty}^{\infty} \left( \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\psi}\left(\frac{x-b_1}{a}\right)} dx \right) \overline{\tilde{\phi}\left(\frac{y-b_2}{a}\right)} dy; \\ (W_{\tilde{\Psi}^3} f)(b_1, b_2; a) &:= (W_{\psi}^2 W_{\phi}^1 f)(b_1, b_2; a) \\ &= \frac{1}{a} \int_{-\infty}^{\infty} \left( \frac{1}{a} \int_{-\infty}^{\infty} f(x, y) \overline{\tilde{\psi}\left(\frac{x-b_1}{a}\right)} dx \right) \overline{\tilde{\psi}\left(\frac{y-b_2}{a}\right)} dy, \end{aligned} \quad (6.5.16)$$

where  $b_1, b_2 \in \mathbb{R}$  and  $a > 0$ .

**Theorem 6.5.1** *Let  $(\phi, \tilde{\phi})$  be a dual pair of refinable (or scaling)  $L_2(\mathbb{R})$  functions with 2-dual two-scaling Laurent polynomial symbols  $(P(z), A(z))$  and that together with Laurent polynomial symbols  $B(z), Q(z)$ , constitute the matrix extension (6.5.3). Then by adopting the notations in (6.5.4)–(6.5.8), the coefficients  $d_{1;\ell,m}^j, d_{2;\ell,m}^j, d_{3;\ell,m}^j$ , in (6.5.8) reveal the wavelet details of  $f_{j+1}(x, y)$ ,*

for each  $j = J + 1 - L, \dots, J$ , as follows:

$$\begin{aligned} d_{1;\ell,m}^j &= \left( W_{\tilde{\Psi}_1} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}; \frac{1}{2^j} \right); \\ d_{2;\ell,m}^j &= \left( W_{\tilde{\Psi}_2} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}; \frac{1}{2^j} \right); \\ d_{3;\ell,m}^j &= \left( W_{\tilde{\Psi}_3} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}; \frac{1}{2^j} \right). \end{aligned} \quad (6.5.17)$$

Before attempting to prove the above theorem, we remark that the duality of  $\phi$  and  $\tilde{\phi}$ , defined by

$$\int_{-\infty}^{\infty} \phi(x-j) \overline{\tilde{\phi}(x-k)} = \delta_{j-k}$$

is equivalent to

$$\sum_{k=-\infty}^{\infty} \widehat{\phi}(\omega + 2\pi k) \overline{\widehat{\tilde{\phi}}(\omega + 2\pi k)} = 1. \quad (6.5.18)$$

Since the derivation of (6.5.18) is the same as that of (6.3.12) where  $\tilde{\phi} = \phi$ , it is safe not to provide the proof here.

To prove Theorem 6.5.1, we first apply (6.5.18) to derive the following duality and orthogonality properties, where the inner product notation for  $L_2(\mathbb{R}^2)$  is used:

$$\langle \Phi^p(2^j x - k, 2^j y - \ell), \tilde{\Psi}^q(2^j x - m, 2^j y - n) \rangle = 0 \quad (6.5.19)$$

and

$$\langle \Psi^p(2^j x - k, 2^j y - \ell), \tilde{\Psi}^q(2^j x - m, 2^j y - n) \rangle = 2^{-2j} \delta_{k-m} \delta_{\ell-n} \delta_{p-q} \quad (6.5.20)$$

for all  $p, q = 1, 2, 3$  and all  $k, \ell, m, n \in \mathbb{Z}$ . To prove (6.5.19), we apply

Plancherel's identity to write

$$\begin{aligned}
& \int_{-\infty}^{\infty} \phi(x-k) \overline{\psi(x-m)} dx = \\
& = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\phi}(\omega) \overline{\widehat{\psi}(\omega)} e^{-i(k-m)\omega} d\omega \\
& = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{\phi}\left(\frac{\omega}{2}\right) \overline{\widehat{\phi}\left(\frac{\omega}{2}\right)} P(e^{-i\frac{\omega}{2}}) \overline{B(e^{-i\frac{\omega}{2}})} e^{-i(k-m)\omega} d\omega \\
& = \frac{1}{2\pi} \sum_{\ell=-\infty}^{\infty} \int_{2\pi\ell}^{2\pi(\ell+1)} \widehat{\phi}\left(\frac{\omega}{2}\right) \overline{\widehat{\phi}\left(\frac{\omega}{2}\right)} P(e^{-i\frac{\omega}{2}}) \overline{B(e^{-i\frac{\omega}{2}})} e^{-i(k-m)\omega} d\omega \\
& = \frac{1}{2\pi} \int_0^{2\pi} \left( \sum_{\ell=-\infty}^{\infty} \widehat{\phi}\left(\frac{\omega}{2} + \pi\ell\right) \overline{\widehat{\phi}\left(\frac{\omega}{2} + \pi\ell\right)} \right) \times \\
& \quad \times P((-1)^\ell z) \overline{B((-1)^\ell z)} e^{-i(k-m)\omega} d\omega \\
& = \frac{1}{2\pi} \int_0^{2\pi} \left( \sum_{n=-\infty}^{\infty} \widehat{\phi}\left(\frac{\omega}{2} + 2\pi n\right) \overline{\widehat{\phi}\left(\frac{\omega}{2} + 2\pi n\right)} \right) P(z) \overline{B(z)} e^{-i(k-m)\omega} d\omega + \\
& \quad + \frac{1}{2\pi} \int_0^{2\pi} \left( \sum_{n=-\infty}^{\infty} \widehat{\phi}\left(\frac{\omega+2\pi}{2} + 2\pi n\right) \overline{\widehat{\phi}\left(\frac{\omega+2\pi}{2} + 2\pi n\right)} \right) \times \\
& \quad \times P(-z) \overline{B(-z)} e^{-i(k-m)\omega} d\omega \\
& = \frac{1}{2\pi} \int_0^{2\pi} \left( P(z) \overline{B(z)} + P(-z) \overline{B(-z)} \right) e^{-i(k-m)\omega} d\omega = 0,
\end{aligned} \tag{6.5.21}$$

where  $z = e^{-i\omega/2}$  and the identities (6.5.18) and  $P(z)\overline{B(z)} + P(-z)\overline{B(-z)} = 0$  for  $|z| = 1$  in (6.5.3) are used. Hence, we have, for  $q = 1$  in (6.5.19),

$$\begin{aligned}
& \langle \Phi(2^j x - k, 2^j x - \ell), \widetilde{\Psi}^1(2^j x - m, 2^j y - n) \rangle = \\
& = \left( \int_{-\infty}^{\infty} \phi(2^j x - k) \overline{\widehat{\phi}(2^j x - m)} dx \right) \times \left( \int_{-\infty}^{\infty} \phi(2^j y - \ell) \overline{\widehat{\psi}(2^j y - n)} dy \right) = 0,
\end{aligned}$$

by applying (6.5.21) to the second multiplicative term. Of course, the proof of (6.5.19) for  $q = 2, 3$  is identically the same.

To prove (6.5.20), we first observe that the following  $L_2(\mathbb{R})$  inner products can be derived by applying (6.5.18) and the identities

$$P(z) \overline{B(z)} + P(-z) \overline{B(-z)} = 0,$$

$$Q(z) \overline{A(z)} + Q(-z) \overline{A(-z)} = 0,$$

and

$$Q(z) \overline{B(z)} + Q(-z) \overline{B(-z)} = 1$$

for all  $|z| = 1$  in (6.5.3), as in the derivation of (6.5.21):

$$\begin{aligned} \int_{-\infty}^{\infty} \phi(2^j x - k) \overline{\widetilde{\psi}(2^j x - n)} dx &= 0; \\ \int_{-\infty}^{\infty} \psi(2^j x - \ell) \overline{\widetilde{\phi}(2^j x - m)} dx &= 0; \\ \int_{-\infty}^{\infty} \psi(2^j x - k) \overline{\widetilde{\psi}(2^j x - n)} dx &= 2^{-j} \delta_{\ell-n}, \end{aligned} \tag{6.5.22}$$

for all  $k, \ell, m, n \in \mathbb{Z}$ .

Hence, for  $p = q = 1$  in (6.5.20), we have

$$\begin{aligned} &\langle \Phi^1(2^j x - k, 2^j y - \ell), \widetilde{\Psi}^1(2^j x - m, 2^j y - n) \rangle \\ &= \left( \int_{-\infty}^{\infty} \phi(2^j x - k) \overline{\widetilde{\phi}(2^j x - m)} dx \right) \times \left( \int_{-\infty}^{\infty} \psi(2^j y - \ell) \overline{\widetilde{\psi}(2^j y - n)} dy \right) \\ &= \left( 2^{-j} \delta_{k-m} \right) \left( 2^{-j} \delta_{\ell-n} \right) = 2^{-2j} \delta_{k-m} \delta_{\ell-n}, \end{aligned}$$

by applying the duality property of  $(\phi, \widetilde{\phi})$  and the third property in (6.5.22). The same derivation also yields (6.5.20) for  $p = q = 2, 3$ . By applying the first properties in (6.5.22), the property (6.5.20) for  $p \neq q$  also follows.

We are now ready to prove that the coefficients  $d_{1;\ell,m}^j, d_{2;\ell,m}^j, d_{3;\ell,m}^j$  in (6.5.8) for the decomposition (6.5.9), namely:

$$f_{j+1}(x, y) = f_j(x, y) + \sum_{q=1}^3 \left( \sum_k \sum_n d_{q;k,n}^j \Psi^q(2^j x - k, 2^j y - n) \right),$$

reveal the wavelet details, as described by the wavelet transforms in (6.5.17). To derive (6.5.17), we apply (6.5.19)–(6.5.20) to obtain

$$\begin{aligned} \langle f_{j+1}(x, y), \widetilde{\Psi}^p(2^j x - \ell, 2^j y - m) \rangle &= 0 + \sum_{q=1}^3 \left( \sum_k \sum_n d_{q;k,n}^j \right) 2^{-2j} \times \\ &\quad \times \delta_{k-\ell} \delta_{n-m} \delta_{p-q} \\ &= 2^{-2j} d_{p;m,n}^j. \end{aligned}$$

Hence, for  $p = 1$ , we have, from the first formula in (6.5.16)

$$\begin{aligned} d_{1;\ell,m}^j &= 2^j \int_{-\infty}^{\infty} \left( 2^j \int_{-\infty}^{\infty} f_{j+1}(x, y) \overline{\phi(2^j x - \ell)} dx \right) \overline{\psi(2^j y - m)} dy \\ &= \left( W_{\tilde{\Psi}^1} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right). \end{aligned}$$

Similarly, for  $p = 2$ , we have, from the second formula in (6.5.16),

$$\begin{aligned} d_{2;\ell,m}^j &= 2^j \int_{-\infty}^{\infty} \left( 2^j \int_{-\infty}^{\infty} f_{j+1}(x, y) \overline{\phi(2^j y - m)} dy \right) \overline{\psi(2^j x - \ell)} dx \\ &= \left( W_{\tilde{\Psi}^2} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right). \end{aligned}$$

Finally, for  $p = 3$ , we have, from the third formula in (6.5.16),

$$\begin{aligned} d_{3;\ell,m}^j &= 2^j \int_{-\infty}^{\infty} \left( 2^j \int_{-\infty}^{\infty} f_{j+1}(x, y) \overline{\psi(2^j x - \ell)} dx \right) \overline{\psi(2^j y - m)} dy \\ &= \left( W_{\tilde{\Psi}^3} f \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right). \end{aligned}$$

This completes the derivation of (6.5.17) and the proof of Theorem 6.5.1. ■

### 6.5.2 Progressive image transmission

As an application of Theorem 6.5.1, we observe that if  $f(x, y) =: f_{J+1}(x, y) \in \mathbf{V}_{J+1}$  is considered as an image defined on  $[0, b] \times [0, c]$ , then the decomposition

$$f_{J+1}(x, y) = f_{J+1-L}(x, y) + \sum_{p=1}^3 \sum_{j=J+1-L}^J \sum_{\ell,m} d_{p;\ell,m}^j \Psi^p(2^j x - \ell, 2^j y - m), \quad (6.5.23)$$

for any desired number of levels  $L \geq 1$ , provides both the image thumb-nail

$$f_{J+1-L}(x, y) = \sum_{\ell,m} c_{\ell,m}^{J+1-L} \Phi(2^{J+1-L} x - \ell, 2^{J+1-L} y - m), \quad (6.5.24)$$

with image size  $= 4^{-L}$  of the size of the given image, as well as the wavelet image details

$$\begin{aligned} d_{1;\ell,m}^j &= \left( W_{\tilde{\Psi}^1} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right) \\ d_{2;\ell,m}^j &= \left( W_{\tilde{\Psi}^2} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right) \\ d_{3;\ell,m}^j &= \left( W_{\tilde{\Psi}^3} f_{j+1} \right) \left( \frac{\ell}{2^j}, \frac{m}{2^j}, \frac{1}{2^j} \right) \end{aligned}$$

for  $j = J + 1 - L, \dots, J$ . Here, the wavelet transform  $W_{\tilde{\Psi}_1} f_{j+1}$  is applied only to the  $y$ -variable, while the smoothing operation is applied to the  $x$ -variable. Hence,  $\{d_{1;\ell,m}^j\}$  is called the  $LH$  band of  $f_{j+1}(x, y)$ . Similarly, since the wavelet transform  $W_{\tilde{\Psi}_2} f_{j+1}$  is applied only to the  $x$ -variable,  $\{d_{2;\ell,m}^j\}$  is called the  $HL$  band of  $f_{j+1}(x, y)$ . On the other hand, since the wavelet transform  $W_{\tilde{\Psi}_3} f_{j+1}$  is applied to both  $x$  and  $y$  variables of  $f_{j+1}(x, y)$  to yield  $d_{3;\ell,m}^j$ , the sequence  $\{d_{3;\ell,m}^j\}$  is called the  $HH$  band of  $f_{j+1}(x, y)$ .

To transmit or acquire an image  $f(x, y) = f_{J+1}(x, y)$ , the thumb-nail  $f_{J+1-L}(x, y)$  is most essential, followed by the  $LH$  and  $HL$  bands,  $g_{1,J+1-L}(x, y)$  and  $g_{2,J+1-L}(x, y)$ , and then the  $HH$  band  $g_{3,J+1-L}(x, y)$ . These four sub-images constitute a larger image thumb-nail  $f_{J+2-L}(x, y)$  of size  $= 4^{-L+1}$  of that of the given image. To acquire a higher-resolution image, the  $LH$  and  $HL$  bands, followed by the  $HH$  band,  $g_{1,J+2-L}(x, y)$ ,  $g_{2,J+2-L}(x, y)$  and  $g_{3,J+2-L}(x, y)$ , in this order, may be transmitted and combined with the previous thumb-nail  $f_{J+2-L}(x, y)$  to yield yet a larger thumb-nail  $f_{J+3-L}(x, y)$ , and so forth.

For image compression (before transmission or storage), the wavelet image details  $d_{p;\ell,m}^j$  for  $p = 1, 2, 3$  and all  $j, m$  can be quantized, and even thresholded (i.e. replaced by zero) for small values. This is analogous to DCT quantization as studied in Subunit 2.5, though different quantization tables should be created. As to the thumb-nail image  $f_{J+1-L}(x, y)$  on the  $L^{\text{th}}$  level, since it is an image, it can be compressed by any other schemes, such as DCT or  $8 \times 8$  tiled DCT as studied in Subunit 2.5, by treating the coefficients in  $\{c_{\ell,m}^{J+1-L}\}$  of (6.5.24) as a digital image.

Entropy coding, studied in Subunits 2.3.2–2.3.4, such as the Huffman encoding scheme, discussed in Subunit 2.5, can be applied to the quantized values of the wavelet image details.

In the following, we extend the wavelet decomposition algorithm in (6.3.47) and wavelet reconstruction algorithm in (6.3.49) to two-dimensions, for efficient computation of the wavelet image details  $\{d_{p;\ell,m}^j\}, p = 1, 2, 3; j = J + 1 - L, \dots, J; \ell, m \in \mathbb{Z}$ , as well as the thumb-nail  $\{c_{\ell,m}^{J+1-L}\}$ ; and efficient reconstruction from the de-coded quantized values of  $\{d_{p;\ell,m}^j\}$  and de-compressed digital thumb-nail image  $\{c_{\ell,m}^{J+1-L}\}$ .

## Two-dimensional wavelet decomposition

For  $j = J, J-1, \dots, J+1-L$ , compute

$$c_{\ell,m}^j = \frac{1}{4} \sum_{k_2} \bar{a}_{k_2-2m} \left( \sum_{k_1} \bar{a}_{k_1-2\ell} c_{k_1,k_2}^{j+1} \right);$$

$$d_{1;\ell,m}^j = \frac{1}{4} \sum_{k_2} \bar{b}_{k_2-2m} \left( \sum_{k_1} \bar{a}_{k_1-2\ell} c_{k_1,k_2}^{j+1} \right);$$

$$d_{2;\ell,m}^j = \frac{1}{4} \sum_{k_2} \bar{a}_{k_2-2m} \left( \sum_{k_1} \bar{b}_{k_1-2\ell} c_{k_1,k_2}^{j+1} \right);$$

$$d_{3;\ell,m}^j = \frac{1}{4} \sum_{k_2} \bar{b}_{k_2-2m} \left( \sum_{k_1} \bar{b}_{k_1-2\ell} c_{k_1,k_2}^{j+1} \right),$$

as follows:

$$\begin{array}{ccccccc} c_{\ell,m}^{J+1} & \longrightarrow & c_{\ell,m}^J & \longrightarrow & \cdots & \longrightarrow & c_{\ell,m}^{j+2-L} \longrightarrow c_{\ell,m}^{J+1-L} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ d_{n;\ell,m}^J & & d_{n;\ell,m}^{J-1} & & & & d_{n;\ell,m}^{J+1-L} \\ (n=1,2,3) & & (n=1,2,3) & & & & (n=1,2,3) \end{array}$$

### Two-dimensional wavelet reconstruction

For  $j = J+1-L, \dots, J$ , compute

$$\begin{aligned} c_{\ell,m}^{j+1} &= \sum_{k_2} p_{m-2k_2} \left( \sum_{k_1} p_{\ell-2k_1} c_{k_1,k_2}^j \right) + \\ &\quad + \sum_{k_2} q_{m-2k_2} \left( \sum_{k_1} p_{\ell-2k_1} d_{1;k_1,k_2}^j \right) + \\ &\quad + \sum_{k_2} p_{m-2k_2} \left( \sum_{k_1} q_{\ell-2k_1} d_{2;k_1,k_2}^j \right) + \\ &\quad + \sum_{k_2} q_{m-2k_2} \left( \sum_{k_1} q_{\ell-2k_1} d_{3;k_1,k_2}^j \right) \end{aligned}$$

as follows:

$$\begin{array}{ccccccc} c_{\ell,m}^{J+1-L} & \longrightarrow & c_{\ell,m}^{J+2-L} & \longrightarrow & \cdots & \longrightarrow & c_{\ell,m}^{J+1} \\ & & \uparrow & & \uparrow & & \uparrow \\ & & d_{n;\ell,m}^{J+1-L} & & & & d_{n;\ell,m}^J \\ & & (n=1,2,3) & & & & (n=1,2,3) \end{array}$$

### 6.5.3 Lossless JPEG-2000 compression

As already studied in Subunits 2.3–2.5, there are two types of image compression schemes, namely: (1) lossless (or reversible) compression, and (2) lossy (or non-reversible) compression. For lossy compression, the DCT transform



is the most popular data transformation, since it facilitates isolating high-frequency content for effective application of the quantization scheme. On the other hand, DCT does not help in lossless compression at all, since multiplying by DCT coefficients which are not dyadic (i.e. not of the form  $k/2^n$  for some integers  $k$  and  $n > 0$ ) increase bit-depths, often significantly. For example, a multiple of  $\frac{1}{3}$  changes an 8-bit pixel value, which is not divisible by 3, to infinite bit-depth (without the benefit of round-off truncation). On the other hand, a multiple of  $\frac{1}{2^n}$ , for any positive integer  $n$ , only requires shifting the bits by  $n$  places for the binary representation. Therefore, for JPEG image compression studied in Subunit 2.5, while  $8 \times 8$  DCT is used to transform  $8 \times 8$  image blocks to the frequency domain (to facilitate effective quantization) for lossy compression, the transform in the lossless image compression mode of the JPEG standard is only DPCM, by coding differences of pixel values.

The birth of wavelets for image compression some two decades ago gave us hope to unify lossy and lossless compression by using the same transform, namely convolution with the same wavelet filters, with dyadic filter taps, followed by downsampling. In Subunit 6.3.3 we gave such an example in Example 6.3.7, where

$$P(z) = \left(\frac{1+z}{2}\right)^2, A(z) = \left(\frac{1+z}{2}\right)^2 \left(\frac{-z+4-z^{-1}}{2}\right),$$

$$B(z) = -z \left(\frac{1-z^{-1}}{2}\right)^2, Q(z) = -z \left(\frac{1-z^{-1}}{2}\right)^2 \left(\frac{z+4+z^{-1}}{2}\right).$$

These are Laurent polynomials with dyadic coefficients and satisfy the matrix extension property:

$$\begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix} \begin{bmatrix} A(z^{-1}) & B(z^{-1}) \\ A(-z^{-1}) & B(-z^{-1}) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (6.5.25)$$

for  $|z| = 1$ .

Observe that in the above example, while  $(P(z), A(z))$  are 2-dual, in that

$$P(z)A(z^{-1}) + P(-z)A(-z^{-1}) = 1, \quad (6.5.26)$$

the other Laurent polynomial symbols satisfy

$$\begin{aligned} B(z) &= -zP(-z^{-1}); \\ Q(z) &= -zA(-z^{-1}). \end{aligned} \quad (6.5.27)$$

Indeed, that  $B(z)$  and  $Q(z)$  can be constructed from the 2-dual pair  $(P(z), A(z))$  by applying (6.5.27) is always valid, as follows.

**Theorem 6.5.2** *Let  $(P(z), A(z))$  be a 2-dual Laurent polynomial pair with real coefficients as defined by (6.5.26). Then the polynomials  $B(z)$  and  $Q(z)$  as in (6.5.27) satisfy the matrix identity (6.5.25).*

**Proof** Multiplication of the first row of  $M_{P,Q}(z) = \begin{bmatrix} P(z) & P(-z) \\ Q(z) & Q(-z) \end{bmatrix}$  to the first column of  $M_{A,B}(z^{-1}) = \begin{bmatrix} A(z^{-1}) & B(z^{-1}) \\ A(-z^{-1}) & B(-z^{-1}) \end{bmatrix}$  yields the left-hand side of (6.5.26), which is equal to 1 for  $|z| = 1$  by the 2-duality assumption. Hence, by (6.5.27), the other row-column multiplications of  $M_{P,Q}(z)$  to  $M_{A,B}(z^{-1})$  are:

$$\begin{aligned} P(z)B(z^{-1}) + P(-z)B(-z^{-1}) \\ = P(z)(-z^{-1}P(-z)) + P(-z)(z^{-1}P(z)) = 0; \end{aligned}$$

$$\begin{aligned} Q(z)A(z^{-1}) + Q(-z)A(-z^{-1}) \\ = (-zA(-z^{-1}))A(-z^{-1}) + (zA(z^{-1}))A(-z^{-1}) = 0; \end{aligned}$$

$$\begin{aligned} Q(z)B(z^{-1}) + Q(-z)B(-z^{-1}) \\ = (-zA(-z^{-1}))(-z^{-1}P(-z)) + (zA(z^{-1}))z^{-1}P(z^{-1}) \\ = A(-z^{-1})P(-z) + A(z^{-1})P(z) = 1, \end{aligned}$$

where the last equality is obtained by applying (6.5.26). ■

**Remark 6.5.2** From the first formula in (6.5.27), we may write

$$P(z) = zB(-z^{-1}), \quad |z| = 1. \quad (6.5.28)$$

Hence, if  $A(z)$  and  $B(z)$  are given, so that the pair  $(P(z), A(z))$ , with  $P(z)$  defined by (6.5.28), satisfies the 2-duality condition (6.5.26), then by defining  $Q(z)$  as in (6.5.27), we have the wavelet decomposition sequence pair

$$\left( \left\{ \frac{1}{2}a_k \right\}, \left\{ \frac{1}{2}b_k \right\} \right) \quad (6.5.29)$$

and wavelet reconstruction sequence pair

$$(\{p_k\}, \{q_k\}) \quad (6.5.30)$$

for image decomposition and reconstruction respectively, by applying the “Two-dimensional wavelet decomposition” scheme, and “Two-dimensional wavelet reconstruction” scheme, respectively, discussed in Subunit 6.5.2. The formulation of (6.5.29)–(6.5.30) is a result of

$$\begin{aligned} A(z) &= \frac{1}{2} \sum_k a_k z^k, \quad B(z) = \frac{1}{2} \sum_k b_k z^k; \\ P(z) &= \frac{1}{2} \sum_k p_k z^k, \quad Q(z) = \frac{1}{2} \sum_k q_k z^k. \end{aligned} \quad (6.5.31)$$

In the following example, we list the sequence pairs  $(\{a_k\}, \{b_k\})$ , and call them (Low, High) pairs. Then by applying (6.5.28)–(6.5.31), the reconstruction sequence pair  $(\{p_k\}, \{q_k\})$  can be obtained easily.

**Example 6.5.1** (Low, High) pairs  $(\{a_k\}, \{b_k\})$  for unified lossless and lossy image compression:

- (1) 2/2 decomposition filters (Haar)

$$\{a_k\} = \{1, 1\}$$

$$\{b_k\} = \{-1, 1\}$$

- (2) 2/10 decomposition filters

$$\{a_k\} = \{1, 1\}$$

$$\{b_k\} = \frac{1}{2^7} \{-3, -3, 22, 22, -128, 128, -22, -22, 3, 3\}$$

- (3) 2/6 decomposition filters

$$\{a_k\} = \{1, 1\}$$

$$\{b_k\} = \frac{1}{2^3} \{1, 1, -8, 8, -1, -1\}$$

- (4) 5/11 decomposition filters

$$\{a_k\} = \frac{1}{4} \{-1, 2, 6, 2, -1\}$$

$$\{b_k\} = \frac{1}{2^7} \{-1, 2, 7, 0, -70, 124, -70, 0, 7, 2, -1\}$$

- (5) 5/3 decomposition filters

$$\{a_k\} = \frac{1}{4} \{-1, 2, 6, 2, -1\}$$

$$\{b_k\} = \frac{1}{2} \{-1, 2, -1\}$$

- (6) 9/3 decomposition filters

$$\{a_k\} = \frac{1}{2^6} \{3, -6, -16, 38, 90, 38, -16, -63\}$$

$$\{b_k\} = \frac{1}{2} \{-1, 2, -1\}$$

- (7) 9/7 decomposition filters

$$\{a_k\} = \frac{1}{2^5} \{1, 0, -8, 16, 46, 16, -8, 0, 1\}$$

$$\{b_k\} = \frac{1}{2^4} \{1, 0, -9, 16, -9, 0, 1\}$$

**Remark 6.5.3** The 5/3 decomposition filter pair in (5), or equivalently the symbols  $A(z), B(z)$  in Example 6.3.3 of Subunit 6.3.2, was selected by the JPEG-2000 image compression standard for lossless image compression.



# Index

- 2-duality" condition, 243
- $B$ -frames, 94
- $B$ -pictures, 94
- $I$ -frames, 94
- $I$ -slices, 94
- $L_2$  inner-product, 5
- $L_2$ -closure, 246
- $P$ -frames, 94
- $P$ -pictures, 94
- $RGB$ , 92
- $YC_bC_r$ , 92
- $YIQ$ , 92, 93
- $YUV$ , 93
- $\mathbb{F}_u$ , 151
- $\ell_2$  inner-product, 4
- $\ell_2$ -approximation, 51
- $\ell_2$ -norm, 25
- $s$ -dimensional convolution, 179
- $s$ -dimensional heat equation, 180
- 2-Dimensional DCT, 64
- 2-duality, 244
- 24-bit color, 72
- 8-Point DCT, 63
- identify precious metals, 43
- Leonhard Euler, 186
- Neumann diffusion PDE, 193
- A/D (analog-to-digital) converter, 206
- ac (alternate current), 89
- AC (alternating current), 217
- adjoints of bounded linear operators, 10
- admissible wavelet, 219
- admissible window function, 167
- agriculture, 48
- Airborne Visible/Infrared Imaging Spectrometer, 48
- analog signal, 142
- analysis wavelet, 243, 253
- anisotropic diffusion, 198
- anisotropic diffusion PDE, 195
- anisotropic PDE, 196
- Approximation by lower-rank matrices, 26
- Archaeology research, 48
- Archimedes Palimpsest, 47
- arithmetic coding, 83
- authentication, 43
- AVC (for Advanced Video Coding), 94
- AVC-Intra video codec, 94
- average code-word length, 81
- backward diffusion, 209
- Balian-Low restriction, 160
- band-limited functions, 223
- bandwidth, 223
- Basel Problem, 132
- Basel problem, 129, 135
- Bernoulli brothers, 132
- Bernoulli numbers, 129, 135
- Bernoulli polynomials, 129, 136
- Bernoulli probability distribution, 69
- Bessel's inequality, 119, 129
- best approximation, 39
- bi-directional frame prediction, 94
- bi-infinite sequences, 4
- bi-orthogonal, 242
- bi-orthogonal wavelets, 245
- binary codes, 75, 78
- binary coding, 71
- binary digits, 71

- binary representation, 75
- Binomial probability distribution, 69
- bit, 71
- bit-stream, 71, 78
- boundary conditions, 184
- boundary value problems, 194
- bounded linear functionals, 8
- bounded linear operators, 8
- bounded linear transformation, 8
- broadcasting, 71
- Césaro means, 114, 116
- cancer drug research, 43
- Cardinal  $B$ -spline, 238
- Cardinal  $B$ -spline, 226
- Cauchy principal value, 214
- Cauchy-Schwarz Inequality, 3
- Cauchy-Schwarz inequality, 219
- CCITT, 93
- centered data-matrix, 27
- channel coding, 76
- chroma-subsampling, 92
- Claude Shannon, 71, 83
- code alphabet, 77, 78
- code-table, 71, 78
- code-word, 80
- color calibration, 37
- complete inner-product space, 9
- completeness, 118
- computed tomography, 43
- conductivity function, 195
- confocal microscopy, 43
- continuous turning tangent, 195
- continuous wavelet transform, 252
- converting, 75
- convex combination, 204
- convolution filter, 143
- convolution operation, 113
- cosine and sine series, 102
- covariance matrix, 27
- criminal identification, 43
- crystallography, 43
- CWT, 214, 252
- data matrix, 51
- data-dependent basis, 26
- Daubechies orthonormal scaling functions, 257
- Daubechies wavelets, 240
- DC (direct current), 217
- dc (direct current), 89
- DCT matrix, 207
- DCT-II, 87, 89
- de-Noising, 205
- De-quantization, 87
- decoding, 71
- delta frames, 86
- delta function, 173
- detection of counterfeit, 43
- DFT, 58
- diffusion partial differential equation, 172
- diffusion process, 172
- diffusion system, 208
- digital image, 76
- digital video, 76
- dimension-reduced data, 51
- dimensionality reduction, 37
- Dirac delta distribution, 173
- direct sum, 247
- Direct TV, 94
- Dirichlet kernel, 113
- Dirichlet's kernel, 114
- Dirichlet's kernels, 99
- Discrete Cosine Transform (DCT), 63
- discrete Fourier transform (DFT), 58
- discrete probability distribution, 68
- discrete wavelet transform, 253
- divergence operator, 194
- Divergence Theorem, 197
- DNA matching, 43
- downsampling, 248
- DPCM, 89
- DPCM (differential pulse code modulation), 70, 72, 74
- dual pair, 9
- dual refinable function, 243
- DWT, 253
- DWT (discrete wavelet transform), 87

- eigen-space, 204
- eigenspace, 198
- Eigenvalue Problems, 6
- eigenvalue-eigenvector pair, 14
- electromagnetic radiation, 41
- electromagnetic radiation (EMR), 37
- Encoder - Decoder (Codec), 90
- encoding, 71
- encryption, 76
- enhanced JPEG, 211
- entropy, 73, 74, 82
- entropy coding, 86, 87, 269
- EOB, end-of-block, 92
- Euclidean space, 2
- Euler's formula, 98, 129, 135, 139
- expected value, 70
- fabrication, 44
- fast Fourier transform (FFT), 58
- Fejér kernel, 115
- Fejér's kernels, 99, 114
- FFT, 57
- finger-print image, 43
- fluoroscopy, 43
- FM radio signal, 41
- Fourier series, 98
- Fourier series expansion, 100
- Fourier series representations, 99
- Fourier transform, 142
- Fourier-coefficients, 98
- frame, 163
  - tight, 163
- frame bounds, 163
- Frobenius norm, 23
- Frobenius norm of a matrix, 23
- Fubini's theorem, 222
- full singular value decomposition, 20
- full SVD, 23
- Gabor transform, 150, 152
- Gamma rays, 42
- Gaussian function, 144
- Gaussian Kernel, 143
- Gaussian model, 209
- general binomial probability distribution, 69
- generalized inverse, 30
- generating function, 136
- GIF image, 87
- global warming, 178
- gradient operator, 191, 194
- Gram matrix, 14, 27
- Gram-Schmidt orthonormalization procedure, 202
- Gram-Schmidt process, 21
- H.264, 94
- Haar wavelet, 237
- Hardamard transform, 87
- hat function, 226
- Hermitian, 12
- Hertz, 41
- hierarchy of sub-images, 262
- histogram, 70, 73
- homeland security, 49
- Huffman coding, 80
- Huffman table, 92
- Hyperion sensor, 47
- Hyperion system, 47
- hyperspectral imaging, 45
- I, P, and B video frames, 93
- IDCT, 90
- ideal bandpass filters, 223
- IEC, 93
- IFT, 146, 150
- image coding, 259
- image enhancement, 209
- image features, 43
- image thumb-nail, 268
- infinite sequences, 4
- information, 71
- information coding, 66
- information source, 71, 73
- infrared, 42
- initial values, 186
- initial-valued Neumann PDE, 191
- inner product, 2
- inner-product space, 2
- instantaneous, 78
- instantaneous code-table, 78

- instantaneous code-tables, 83
- inter-macroblocks, 94
- intra-macroblocks, 94
- inverse DCT, 90
- inverse Fourier transform, 146, 150
- inverse transformation, 87
- inverse wavelet transform, 221
- ISO, 93
- isotropic heat diffusion, 206
- iTune stores, 94
  
- Jacob Bernoulli, 135
- John Tukey, 71
- Joint Photographic Experts Group, 88
- Joseph Fourier, 187
- JPEG, 88, 90, 93
- JPEG compressed image, 72
- JPEG compression, 212
- JPEG quantization tables, 212
- JPEG standard, 92
- JPEG-2000, 258
- JPEG-2000 image compression standard, 273
  
- Kraft inequality, 80
- Kraft's inequality, 80
- Kraft-McMillan's inequality, 80
- Kronecker symbol, 244
- Ky Fan norm, 25
  
- lagged anisotropic diffusion, 195
- lagged anisotropic transform, 202
- Lagrange multipliers, 82
- Lanczos matrix factorization, 58
- Landsat satellites, 46
- Laplace operator, 180, 181, 200
- Laplacian operator, 191
- Laurent polynomial, 240, 244, 245
- Laurent series, 230
- least-squares estimation, 35
- Lebesgue's dominated convergence theorem, 177
- LFT, 151
- LIFT, 151, 152
- linear transformation, 7
  
- Linear Transformations, 6
- linearly independent integer shifts, 246
- local basis functions, 167
- local coordinates, 201
- localized Fourier transform, 151
- localized inverse Fourier transform, 151, 152
- lossless and lossy compression, 79
- lossless compression, 86
- lossless JPEG-2000 compression, 270
- lowpass filters, 216
- luminance - chrominance formats, 93
- LZW compression, 86
  
- Malvar wavelets, 169
- mammography, 43
- matrix
  - covariance, 27
- matrix adjoint, 10, 12
- matrix extension, 238, 240, 245, 254
- mean-square approximation, 110
- measurement of best approximation, 39
- medical applications, 43
- medical imaging, 43
- method of separation of variables, 187
- microwave, 42
- mineralogy, 49
- minimum-norm least-squares estimation, 30
- motion search and compensation, 93
- motion vector, 94
- Moving Pictures Expert Group, 93
- MPEG, 93
- MPEG-1, 93
- MPEG-2, 94
- MPEG-4 Part 4, 94
- MRA, 222
- MRA wavelet, 247
- multiresolution analysis, 222
- multispectral image, 44
  
- nanometers, 41
- narrow-band, 37



- NASA, 47
- Neumann condition, 188
- noiseless coding, 83
- Noiseless Coding Theorem, 84, 87
- non-deterministic function, 68
- norm, 3
- norm of a linear transformation, 7
- norm-1 matrices, 37
- normalized eigenvectors, 19
- normalized Gaussian, 157
- normalized orthogonality, 233
- NTSC standard, 92
- orthogonal projection, 107, 109
- orthogonal sum, 247
- orthogonal wavelet, 229
- orthonormal, 231
- orthonormal bases, 15
- orthonormal basis, 9, 165, 169
- PAL standard, 92
- Parallelogram Law, 109, 110
- Parseval's formula, 219, 220
- Parseval's identity, 129, 130, 164, 203
- partial sums, 98
- PCA, 51
- PCA dimensionality reduction, 51
- PDE, 172
- permutation matrix, 60
- Perona-Malik model, 210
- phase modulation, 166
- Plancherel's formula, 147, 148, 153, 158
- Plancherel's identity, 266
- Planck's constant, 41
- positive approximate identity, 99, 116
- positive semi-definite, 14
- PostScript, 87
- precise error, 39
- prefix code, 77
- prefix-code, 78
- principal component analysis (PCA), 28
- principal components, 26, 50
- Principle of superposition, 184
- probability distributions, 66
- progressive image transmission, 268
- pseudo-inverse, 30
- pseudo-inverses, 29
- Pythagorean Theorem, 110, 119
- Pythagorean theorem, 107
- QMF, 236
- QMF pair, 236
- quadrature mirror filter, 230, 236
- quantization, 87
- quantizer, 88, 205
- quantizers, 90, 91
- radio frequencies, 41
- radiography, 43
- random variable, 68, 70
- rank, 14
- rank-1 decomposition, 38
- Rayleigh quotient, 198
- reduced singular value decomposition, 17
- refinable function, 224
- refinement equation, 224
- refinement sequence, 224
- remote sensing, 46
- representer of the linear functional, 9
- reproducing kernel, 222
- Riesz Lemma, 256
- right-hand and left-hand limits, 122
- RLE (run-length encoding), 72
- rotationally invariant, 200
- round-off function, 88
- run-length encoding (RLE), 86
- Sampling Theorem, 143, 224
- scalar field, 2
- Schatten norm, 25
- SECAM, 93
- self-adjoint, 12
- self-dual, 9
- semi-definite, 14
- semiconductor device, 44
- separable function, 182
- separation of variables, 183
- Shannon wavelet, 238

- short-time Fourier transform, 151
- simple knots, 36
- simply connected domain, 195
- simultaneous time and frequency localization, 153
- singular values, 14
- singular-vector pair, 15
- smoothing transform, 262
- source alphabet, 76
- source coding, 76
- sparse, 36
- sparse matrix decomposition, 25
- Specim LWIR-C imager, 49
- spectral decomposition, 14
- spectro-colorimeters, 37
- St. Petersburg Academy of Science, 186
- STFT, 151
- sum rule, 231
- surveillance, 49
  
- Taylor representation, 243
- tensor product, 264
- terahertz frequency band, 44
- terahertz radiation, 44
- ternary codes, 78
- thermal imaging, 43
- thermographic cameras, 44
- thumb-nail, 254, 262
- TIFF, 87
- tight frame, 165
- time-domain, 142
- time-frequency localization window, 158
- time-scale analysis, 213
- trace, 24
- trace norm, 25
- trichromatic, 37
- trigonometric series, 102
- Two-dimensional wavelet decomposition, 269
- TV model, 209, 211
- two-dimensional wavelet decomposition, 272
- two-dimensional wavelet reconstruction, 270, 272
- two-scale relation, 224
- two-scale sequence, 224
- two-scale symbol, 225, 230, 238
  
- ultraspectral imaging, 45
- ultraviolet, 42
- uncertainty principle, 158
- uniform error, 117
- uniform probability distribution, 69
- unit normal vector, 191
- unit outer normal, 197
- unit tangent-normal pair, 199
- unitary matrix, 14
- upsampling, 249
- USGS, 47
- UV imaging, 43
  
- variance, 70
- variational method, 82
- vibrating string, 186
- video streaming, 71
- visible light, 37, 42
  
- wavelet decomposition, 248
- wavelet decomposition algorithm, 254
- wavelet details, 262, 264
- wavelet domain, 262
- wavelet reconstruction, 249
- wavelet reconstruction algorithm, 254
- wavelet transform, 214
- wide spectrum, 41
- Wilhelm Roentgen, 43
- window center, 155
- window width, 155
  
- X-ray imaging, 43
- X-rays, 42
  
- YouTube, 94
- Zig-zag ordering, 91
- ZRL, zero run length, 92