1
Digital Linear Systems

1.1
Introduction to Digital Phase Demodulation in Optical Metrology

In this chapter, we review the theory behind digital signals and their temporal processing using linear time-invariant (LTI) systems. The analysis of digital LTI systems is based on their impulse response $h(t)$, their Z-transform $H(z)$, their frequency transfer function (FTF) $H(\omega)$, their harmonic response, and their stability criteria. We then briefly discuss the equivalence between phase-shifting algorithms (PSAs) and quadrature linear filters tuned at the temporal phase-sampling rate of ω_0 radians per sample. Also, we analyze the aliasing phenomena produced by high-order harmonic distortion of the continuous interferogram being sampled.

In this chapter, we also discuss regularized low-pass filtering and its application to fringe-pattern denoising. Convolution spatial filters (such as the 3 × 3 averaging filter) mix up valid fringe data inside the interferogram boundaries with outside background where no fringe data is defined. This linear mixing of fringes and background distorts the modulating phase near the interferogram boundaries. In contrast, regularized linear filters optimally decouple the fringe data inside the interferogram from the outside background.

Finally, we discuss the theory behind stochastic processes to analyze the response of LTI systems to stochastic input signals $X(t)$. We define and analyze their probability density function (PDF), or $f_X(x)$, their ensemble average $E\{X\}$, and their stationary autocorrelation function $R_X(\tau)$. We then continue by defining the power spectral density (PSD) $S_X(\omega)$ for $X(t)$. This result is then used to show that the input PSD, $S_X(\omega)$ of $X(t)$, changes to $|H(\omega)|^2 S_X(\omega)$ when processed by an LTI system whose FTF is given by $H(\omega)$.

1.1.1
Fringe Pattern Demodulation as an Ill-Posed Inverse Problem

A fringe pattern is defined as a sinusoidal signal where a continuous map, analogous of the physical quantity being measured, is phase-modulated by an interferometer, Moire system, and so on. An ideal stationary fringe pattern is usually modeled by

$$I(x, y) = a(x, y) + b(x, y)\cos[\varphi(x, y)], \tag{1.1}$$

Fringe Pattern Analysis for Optical Metrology: Theory, Algorithms, and Applications, First Edition.
Manuel Servin, J. Antonio Quiroga, and J. Moisés Padilla.

where $\{x, y\} \in \mathbb{R}^2$; $a(x, y)$ and $b(x, y)$ are the background and local contrast functions, respectively; and $\varphi(x, y)$ is the searched phase function.

In physics and mathematics, an inverse problem is a general framework that is used to convert the observed measurements into information about a physical object or system under study [1]. Clearly, Eq. (1.1) represents an inverse problem, where the fringe pattern $I(x, y)$ is our measurement and the searched information is given by the phase $\varphi(x, y)$. An inverse problem is said to be well posed if the mathematical model of a given physical phenomenon fulfills the following conditions:

- A solution exists,
- The solution is unique, and
- The solution depends continuously on the data.

On analyzing Eq. (1.1), one can see that the phase function $\varphi(x, y)$ cannot be directly estimated since it is screened by two other unknown functions, namely $a(x, y)$ and $b(x, y)$. Additionally, $\varphi(x, y)$ can only be determined modulo 2π because the sinusoidal fringe pattern $I(x, y)$ depends periodically on the phase (2π phase ambiguity); and its sign cannot be extracted from a single measurement without *a priori* knowledge (sign ambiguity) because of the even character of the cosine function [$cos(\varphi) = cos(-\varphi)$]. Finally, in all practical cases, some noise $n(x, y)$ is introduced in an additive and/or multiplicative manner, and the fringe pattern may suffer from a number of distortions, degrading its quality and further screening the phase information [2, 3].

It must be noted that, even if careful experimental setups could prevent the screening of $\varphi(x, y)$ due to the unknown signals $a(x, y)$, $b(x, y)$, and $n(x, y)$, one would still have to deal with the sign ambiguity and the 2π phase ambiguity. Because of these ambiguities, the solution for this inverse problem is not unique; this is illustrated in Figure 1.1, where several phases (from an infinite number of possibilities) produce exactly the same sinusoidal signal.

In short, the phase demodulation of a fringe pattern, as the one modeled in Eq. (1.1), can be viewed as an ill-posed inverse problem where some sort of regularization process is required in order to obtain a proper phase estimation. However, despite its intrinsic difficulties, it is rather easy to visualize a possible solution for this inverse problem. First, let us rewrite Eq. (1.1) by means of the complex representation of the cosine function

$$I(x, y) = a(x, y) + \frac{1}{2} b(x, y)\{\exp[i\varphi(x, y)] + \exp[-i\varphi(x, y)]\}. \tag{1.2}$$

Now, if somehow one is able to isolate one of the analytic signals in Eq. (1.2), say, $(1/2)b(x, y)\exp[i\varphi(x, y)]$, we have

$$\tan\hat{\varphi}(x, y) = \frac{\operatorname{Im}\{(1/2)b(x, y)\exp[i\varphi(x, y)]\}}{\operatorname{Re}\{(1/2)b(x, y)\exp[i\varphi(x, y)]\}}, \tag{1.3}$$

where $b(x, y) \neq 0$. Computing the arc-tangent of the above formula, one obtains a wrapped estimation of the phase under study, that is, $\varphi(x, y) \mod 2\pi$. Thus, the

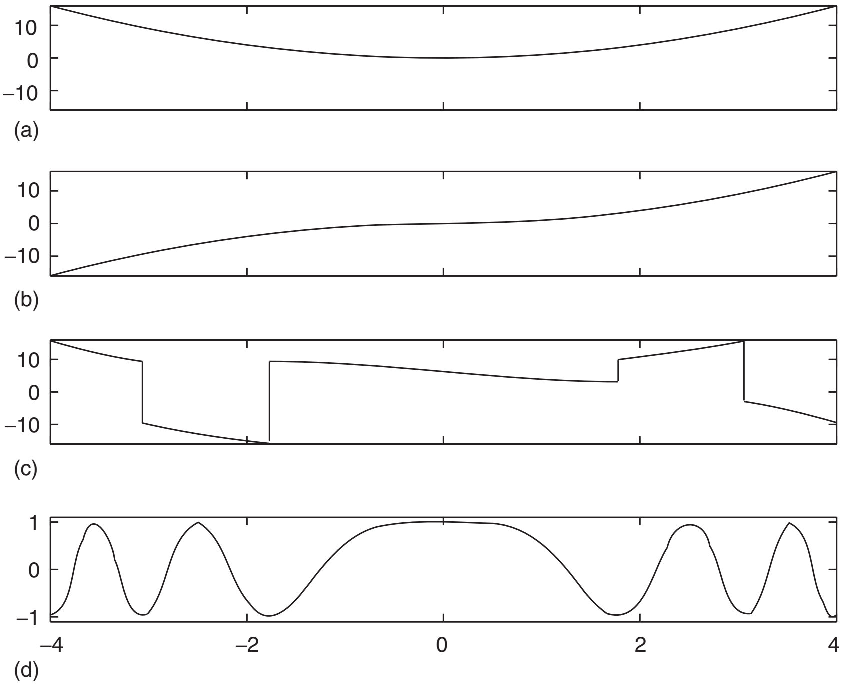

Figure 1.1 Numerical simulation of several phases (a–c) producing exactly the same sinusoidal signal (d). For ease of observation, only a horizontal slice is shown.

final step of this fringe pattern demodulation process usually involves an additional phase unwrapping process. Nevertheless, when working with good-quality data, this last step is straightforward. Next, we will illustrate the easiest way to obtain these analytic signals.

1.1.2 Adding *a priori* Information to the Fringe Pattern: Carriers

A fringe pattern obtained as the output of a measuring system may be modified by the optoelectronic/mechanical hardware (sensors and actuators) and software (virtual sensors and actuators) of the system [4]. With these modifications, one is able to introduce *known* changes in the argument of the sinusoidal signal

$$I(x,y,t) = a(x,y) + b(x,y)\cos[\varphi(x,y) + c(x,y,t)], \tag{1.4}$$

where $c(x,y,t)$ is a known function (typically a reference plane) and is called the *spatiotemporal carrier* of the interferogram. By design, a carrier must be a high-frequency signal in comparison with the searched phase $\varphi(x,y)$. That is

$$\left\|\nabla c(x,y,t)\right\| > \left\|\nabla\varphi(x,y,t)\right\|_{\max}, \tag{1.5}$$

where we define (locally) this *nabla* operator as

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial t}\right). \tag{1.6}$$

For instance, for a stationary phase (which shows no explicit time dependence) given by in $\varphi(x, y)$, and a spatial carrier $c(x, y)$, the following condition must be fulfilled:

$$\sqrt{\left(\frac{\partial c}{\partial x}\right)^2 + \left(\frac{\partial c}{\partial y}\right)^2} > \sqrt{\left(\frac{\partial \varphi}{\partial x}\right)^2 + \left(\frac{\partial \varphi}{\partial y}\right)^2}. \tag{1.7}$$

The spatial and/or temporal carriers are of extreme importance in modern interferometry: first of all, their presence allows us to solve the sign ambiguity since in general $\cos(\varphi + c) \neq \cos(-\varphi + c)$. They also allow us to isolate the analytic signal $(1/2)b(x, y)\exp[i\varphi(x, y)]$ which practically solves the phase demodulation process (the proof of this last point will be postponed until we review some basic concepts of Fourier analysis). Some typical examples of the carrier functions are as follows:

- linear temporal carrier [5, 6]

$$c_1(t) = \omega_0 t; \tag{1.8}$$

- tilted (spatial) carrier [7, 8]

$$c_2(x, y) = u_0 x + v_0 y; \tag{1.9}$$

- conic carrier [9]

$$c_3(\rho) = \omega_0 \rho; \quad \rho(x, y) = \sqrt{x^2 + y^2}; \tag{1.10}$$

- 2 × 2 pixelated carrier [10–12]

$$\exp[i\, c_4(x, y)] = \exp\left[i\omega_0 \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\right] ** \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \delta(x - 2m, y - 2n), \tag{1.11}$$

  where $\omega_0 = \pi/2$, and $**$ is the two-dimensional convolution operation;
- 3 × 3 pixelated carrier [13]

$$\exp[i\, c_5(x, y)] = \exp\left[i\omega_0 \begin{pmatrix} 1 & 2 & 3 \\ 8 & 9 & 4 \\ 7 & 6 & 5 \end{pmatrix}\right] ** \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \delta(x - 3m, y - 3n), \tag{1.12}$$

  where $\omega_0 = 2\pi/9$.

Since digital interferometry is a research area under continuous development, it is impossible to list all useful spatiotemporal carriers; again, these are just some commonly used examples. For illustrative purposes, in Figures 1.2–1.5 we show how these carriers modify the fringe pattern.

The temporal linear carrier approach (shown in Figure 1.2) allows us to demodulate closed-fringe interferograms [5, 6]. However, this method is not useful (in principle) to study fast-varying phenomena since it requires $a(x, y)$, $b(x, y)$, and $\varphi(x, y)$ to remain stationary during the phase-step acquisition.

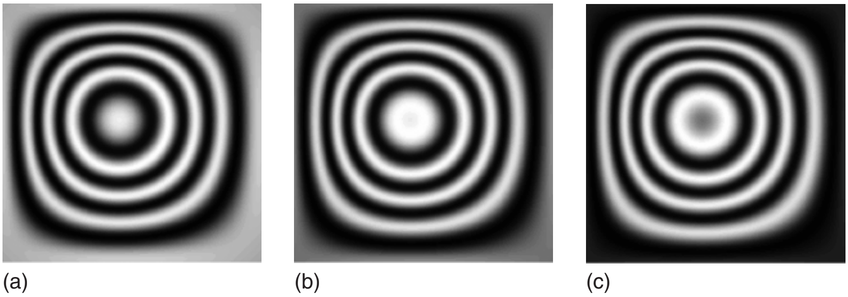

Figure 1.2 Numerical simulation of a closed-fringe interferogram, phase-modulated with a linear temporal carrier $\omega_0 t$. The piston-like phase step between successive samples is $\omega_0 = 2\pi/3$ rad.

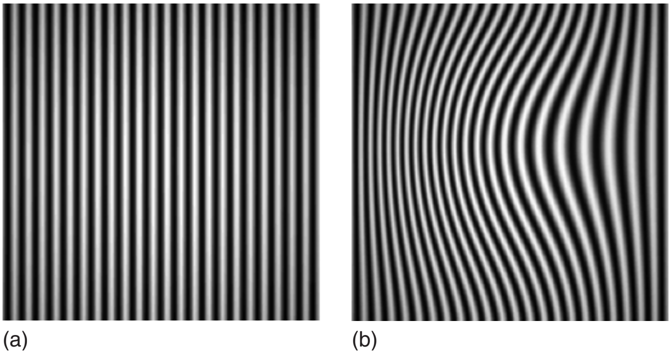

Figure 1.3 Simulation of a closed-fringe interferogram (previously shown in Figure 1.2a) phase-modulated with a linear spatial carrier (a), producing an open-fringe interferogram (b).

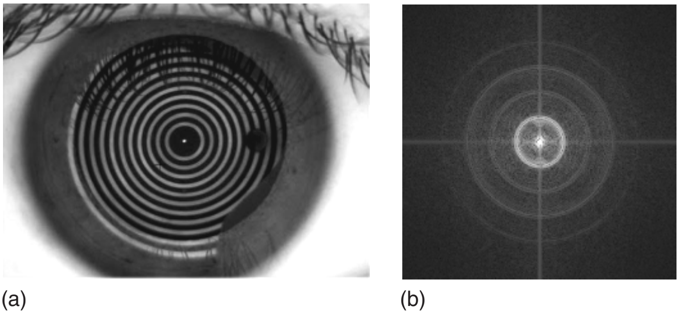

Figure 1.4 (a) Circular pattern with binary amplitude projected over an eye using a Placido mire and (b) its spectrum as obtained by the FFT2 algorithm. The larger spectral flares in the spectrum are due to the binary profile of the projected pattern, and these lead to harmonic distortion.

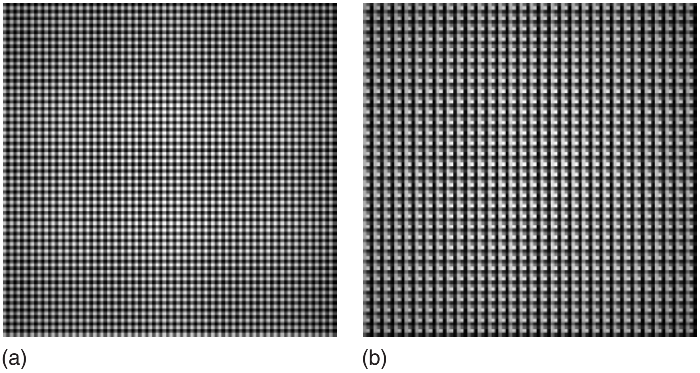

Figure 1.5 Simulation of a closed-fringe interferogram (previously shown in Figure 1.2a) phase-modulated with a four-step pixelated carrier (a), producing a 2D pixelated carrier interferogram (b).

The spatial linear carrier approach (shown in Figure 1.3) allows us to demodulate open fringe patterns from a single image, making this technique particularly useful to study fast dynamic phenomena [7, 8].

The conic carrier (shown in Figure 1.4) has been used to measure the topography irregularities of the human cornea since 1880 [14]. Traditionally, these irregularities were analyzed by means of a sparse set of estimated slope points, integrated along meridian lines to obtain the topography of the testing cornea [15]. However, recently it has been proved that these patterns of periodic concentric rings can be phase-demodulated by means of synchronous interferometric methods, providing holographic phase estimation at every point of the region under study. A detailed review of this topic is available in [9].

The 2D pixelated carrier (shown in Figure 1.5) was originally proposed as a spatial technique for the simultaneous acquisition of four phase-shifted interferograms, to be demodulated using a ''temporal'' PSA, but recently it has been shown that spatial synchronous demodulation allows higher quality measurements [10–12]. The nine-step pixelated carrier was proposed as a logical extension of this technique to allow for the analysis of nonsinusoidal signals in fast dynamic phenomena [13]. We choose to include only one illustrative example for both cases because the four-step and nine-step pixelated carrier interferograms are visually indistinguishable.

Example: Synchronous Demodulation of Open Fringes

For illustrative purposes, let us assume a vertical open-fringe interferogram phase-modulated by a linear spatial carrier in the x direction, given by

$$\begin{aligned} I(x,y) &= a(x,y) + b(x,y)\cos[\varphi(x,y) + u_0 x], \\ &= a + (b/2)\exp[i(\varphi + u_0 x)] + (b/2)\exp[-i(\varphi + u_0 x)], \end{aligned} \tag{1.13}$$

where we have omitted the spatial dependency in a, b, and φ for simplicity. Applying the spatial synchronous demodulation method, the so-called the Fourier method [7, 8], first we multiply our input signal with a complex reference signal

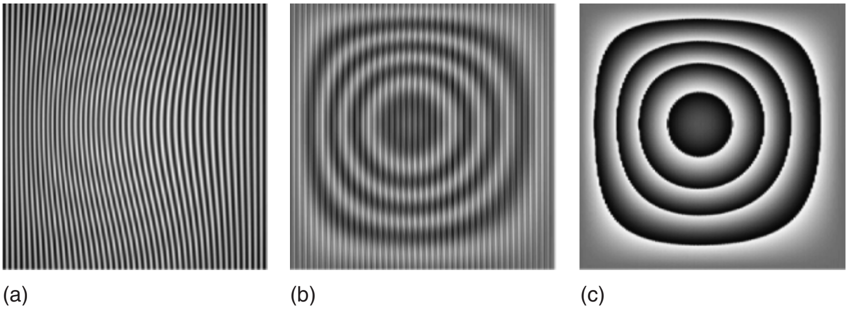

Figure 1.6 Several steps of the spatial synchronous demodulation of an open-fringe pattern interferogram. The input signal is shown in panel (a). Panel (b) shows the real part of the synchronous product $\exp(-iu_0x)I(x,y)$. The estimated phase $\hat{\varphi}(x,y)$ modulo 2π, as obtained from Eq. (1.16), is shown in panel (c).

(which is a value stored in the digital computer) oscillating at the same frequency as our lineal carrier:

$$\begin{aligned} f(x,y) &= \exp(-iu_0x)I(x,y), \\ &= a\exp(-iu_0x) + (b/2)\exp(i\varphi) + (b/2)\exp[-i(\varphi + 2u_0x)]. \end{aligned} \tag{1.14}$$

In general, the spatial variations of the phase are small in comparison to the carrier (Eq. 1.5), $|\nabla\varphi|_{\max} \ll u_0$, so the only low-frequency term in the above equation is the analytic signal $(b/2)\exp(i\varphi)$. Thus, applying a low-pass filter to Eq. (1.14), we have

$$\mathrm{LP}\{f(x,y)\} = (1/2)b(x,y)\exp[i\varphi(x,y)], \tag{1.15}$$

where the low-pass filter $\mathrm{LP}\{\cdot\}$ is preferentially applied in the Fourier domain for more control in the filtering process. Taking the ratio between the imaginary and real part of this complex-valued analytic signal, we have

$$\tan\hat{\varphi}(x,y) = \frac{\mathrm{Im}\{(1/2)b(x,y)\exp[i\varphi(x,y)]\}}{\mathrm{Re}\{(1/2)b(x,y)\exp[i\varphi(x,y)]\}}, \tag{1.16}$$

where $b(x,y) \neq 0$. Computing the arc-tangent of the above equation, the estimated phase $\hat{\varphi}(x,y)$ is wrapped within the principal branch $(-\pi,\pi]$; so there is a 2π phase ambiguity as illustrated in Figure 1.6. Usually, *a priori* knowledge of the phenomenon indicates that $\hat{\varphi}(x,y)$ should be continuous so the final step in the demodulation process is to apply a regularization condition that removes this 2π ambiguity.

1.1.3 Classification of Phase Demodulation Methods in Digital Interferometry

To summarize our previous discussion, the main objective of fringe pattern analysis is to estimate a usually continuous phase map $\varphi(x,y)$ from the input intensity values $I(x,y,t)$. This means solving an ill-posed inverse problem where the signal of interest is masked by unknown functions, plus the sign ambiguity and the 2π phase ambiguity problems. The simplest way of action is to actively modify the fringe pattern in order to provide additional information, that is, introducing spatial or temporal carriers.

The inclusion of phase carriers not only solves the sign ambiguity problem but it also provides spectral isolation between the unknown signals in the interferogram (this will be discussed in detail in Chapters 2 and 4). On the other hand, the 2π phase ambiguity is intrinsic to fringe-pattern analysis, so some unwrapping method is usually required as the last step of a phase-demodulation process [16, 17]. Nevertheless, there are notable exceptions that estimate nearly directly the absolute phase without 2π phase ambiguity, such as the temporal heterodyning technique [18], as well as phase demodulation methods that directly estimate the unwrapped phase, such as the linear phase-locked loop [19], temporal phase unwrapping [20], hierarchical absolute phase measurement [21], and the regularized phase tracking [22].

According to the above, a possible classification for the phase demodulation methods in fringe pattern analysis is as follows: whether a phase carrier is required; whether this carrier is a spatial and/or temporal one; and whether the estimated phase is wrapped within a single branch (requiring an additional unwrapping processing) or without 2π ambiguity. In Figure 1.7, we present a schematic representation of this proposed classification for some commonly used

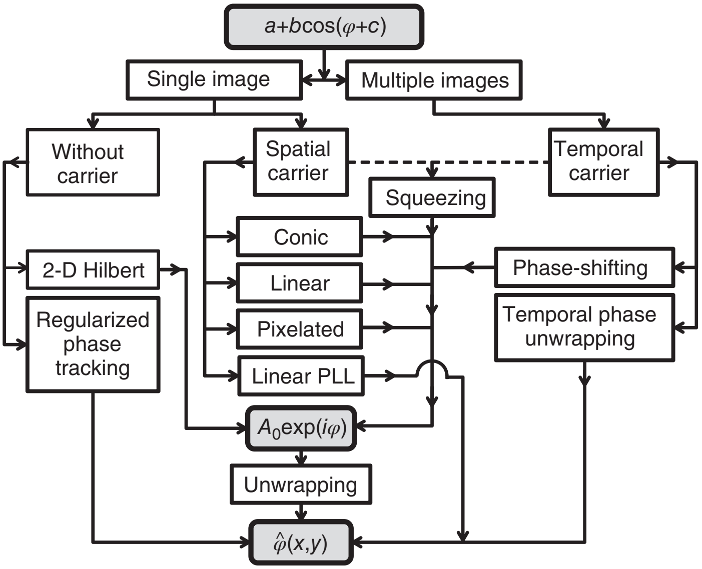

Figure 1.7 Schematic classification of some commonly used phase estimation methods in modern fringe pattern analysis. Here we try to illustrate that the intermediate target in most methods is to isolate the analytic signal $A_0 \exp[i\varphi(x, y)]$, from where one can straightforwardly compute the wrapped phase $\hat{\varphi}(x, y)$ modulo 2π. On the other hand, some methods combine both the fringe demodulation and the phase unwrapping processes, obtaining directly the estimated phase $\hat{\varphi}(x, y)$ without the 2π ambiguity.

phase estimation methods. We want to stress that this scheme is illustrative and by no means exhaustive.

In the following chapters, we will analyze several methods to estimate the analytic signal $(1/2)b(x, y)\exp[i\varphi(x, y)]$ highlighting their positive features and drawbacks. However, in order to do this, first we need to review some basic mathematical tools. For a beginner to this topic, this will serve as a quick reference guide for linear systems theory. Advanced readers can skip the rest of this chapter and return to it only in specific cases that we will refer to whenever we are unable to keep the discussion self-contained in the following chapters.

1.2 Digital Sampling

Despite the fact that (analog) macroscopic phenomena are properly modeled as continuous functions, nowadays virtually any processing required is done on digital computers. Thus, typically one of the very first steps in fringe pattern analysis is to perform some analog-to-digital (A/D) conversion, the so-called digital sampling process. In this section, we analyze some mathematical functions commonly used to model digital signals and systems. This will allow us to understand and cope with many problematic phenomena (e.g., spectral overlap with high-order distorting harmonics) that arise in fringe pattern analysis as consequence of the digital sampling process.

It is noteworthy that we will use t for the independent variable when working with unidimensional (1D) signals and systems; thus we will refer to continuous-time and discrete-time functions. Nevertheless, this is just a convention and the following theory also applies for 1D spatial processing.

1.2.1 Signal Classification

By definition, a signal is everything that contains information. Signals in engineering systems are typically classified in five different groups:

1) Continuous-time or discrete-time
2) Complex or real
3) Periodic or aperiodic
4) Energy or power
5) Deterministic or random.

Continuous-time and discrete-time signals. A signal is defined to be a continuous-time signal if the domain of the function defining the signal contains intervals of the real line $f(t)$ where $t \in \mathbb{R}$. A signal is defined to be a discrete-time signal if the domain of the signal is a countable subset of the real line $\{f(n)\}$ or $f[n]$, where $n \in \mathbb{Z}$.

In most cases, discrete signals arise from uniform sampling of continuous-time signals. However, these sampled signals can also be represented by continuous functions (as we will see in Section 1.2.3). Thus, the following definitions and conventions apply to both continuous and sampled signals:

Real and complex signals. In optics, we often work with complex (analytic) signals of real arguments. In general, a complex signal is given by

$$f(t) = \mathrm{Re}[f(t)] + i\mathrm{Im}[f(t)], \tag{1.17}$$

where $i = \sqrt{-1}$. Or, in polar form, the modulo of the signal is defined by

$$|f(t)| = \sqrt{f(t)f^*(t)} = \sqrt{\{\mathrm{Re}[f(t)]\}^2 + \{\mathrm{Im}[f(t)]\}^2}, \tag{1.18}$$

and its phase (modulo 2π) is given by

$$\text{angle } [f(t)] = \arctan \frac{\mathrm{Im}\{f(t)\}}{\mathrm{Re}\{f(t)\}}. \tag{1.19}$$

A word of caution: in modern programming languages, this operation is called atan2(·) and it uses two arguments. Unlike the single argument arc-tangent function, atan2(·) is able to retrieve the searched angle without sign ambiguity within $(0, 2\pi)$.

Periodic and aperiodic signals. A signal is said to be periodic if repeats itself in time. The function $f(t)$ represent a periodic signal when

$$f(t) = f(t + kT), \quad \forall \quad k \in \mathbb{Z} \tag{1.20}$$

and the fundamental frequency of a periodic signal is given by $1/T$.

Energy and power of signals. The energy of a signal $f(t)$ is a real and nonnegative quantity given by

$$U\{f(t)\} = \int_{-\infty}^{\infty} |f(t)|^2 \, dt. \tag{1.21}$$

If $U\{f(t)\}$ exceeds every bound, we say that $f(t)$ is a signal of infinite energy. For such cases, it is useful to calculate the power of the signal, which represents the energy per unit time. It is defined by

- aperiodic signals

$$P\{f(t)\} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |f(t)|^2 \, dt; \tag{1.22}$$

- periodic signals

$$P\{f(t)\} = \frac{1}{T} \int_{t}^{t+T} |f(\tau)|^2 \, d\tau. \tag{1.23}$$

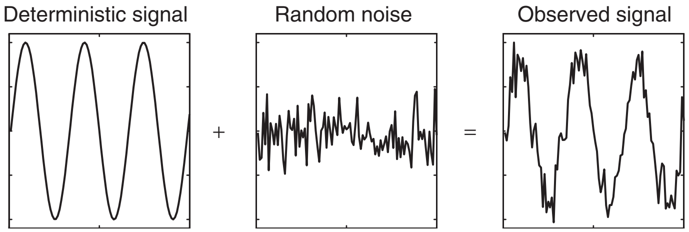

Figure 1.8 Signals observed in nature, which are usually composed by deterministic signals distorted by some degree of random noise.

Deterministic and random signals. Most of the time, we deal with deterministic signals distorted in some degree by random noise (Figure 1.8). The kind of noise typically observed in fringe pattern analysis can be modeled as a well-known stochastic process; however, the theory of stochastic processes is so vast that a detailed review is beyond the scope of this book. In Section 1.9, we will briefly review some basic aspects of this theory, but for now we will assume that we are dealing with purely deterministic signals.

1.2.2 Commonly Used Functions

Dirac delta function. Also called the *unit-impulse function*, the Dirac delta is (informally) a generalized function on the real number line that is zero everywhere except at zero where its value tends to infinity. However, quite often is better to define the Dirac delta function by its properties as

$$\int_{-\infty}^{\infty} f(t)\delta(t-t_0)dt = f(t_0), \tag{1.24}$$

which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1. \tag{1.25}$$

For convenience, some algebraic properties of the Dirac delta function are listed in Table 1.1.

The Dirac delta function is graphically represented as a vertical line with an arrow at the top. The height of the arrow is usually used to specify the value of any multiplicative constant, which will give the area under the function; another convention is to write the area next to the arrowhead (Figure 1.9).

Unit step function. Also called *Heaviside's step function*, it may be defined by means of the Dirac delta as

$$u(t) = \int_{-\infty}^{t} \delta(\tau)\, d\tau = \begin{cases} 0 & \text{for } t < 0, \\ 1 & \text{for } t > 0. \end{cases} \tag{1.26}$$

Table 1.1 Properties of the Dirac delta function.

Properties	Observations
$\delta(t-t_0)=0$	For all $t \neq t_0$
$\delta(-t)=\delta(t)$	Dirac delta is an even function
$\delta(at)=(1/\lvert a\rvert)\delta(t)$	Scaling property
$\int_{-\infty}^{\infty} f(t)\delta(t-t_0)dt=f(t_0)$	Definition as a measure
$\delta\left(g(x)\right)=\sum_i \delta(x-x_i)/\lvert g'(x_i)\rvert$	Where x_i are the roots of $g(x)$
$f(t)\delta(t-t_0)=f(t_0)\delta(t-t_0)$	Valid under the integration symbol
$f(t)*\delta(t-t_0)=f(t-t_0)$	Shifting property
$\delta(x,y,z,\dots)=\delta(x)\delta(y)\delta(z)\dots$	n-dimension generalization

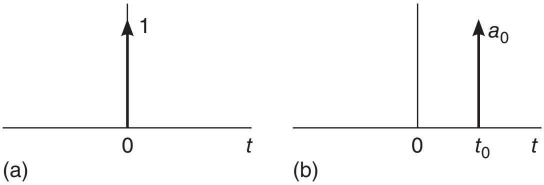

Figure 1.9 (a) Usual representations of the impulse function $\delta(t)$ and (b) the shifted (and escalated) impulse function $a_0\delta\left(t-t_0\right)$.

Rectangle function. The rectangle function of unit height and base is defined by

$$\text{II}(t)=\begin{cases}0 & \text{if } |t|>1/2,\\ 1 & \text{if } |t|<1/2.\end{cases} \tag{1.27}$$

This function can also be represented by means of the unit step function as

$$\text{II}(t)=u\left(t+\frac{1}{2}\right)-u\left(t-\frac{1}{2}\right). \tag{1.28}$$

The step function and the rectangle function are illustrated in Figure 1.10.

Dirac comb. The so-called Dirac comb is a periodic distribution of Dirac delta functions that plays an important role in the sampling process:

$$\text{III}(t)=\sum_{n=-\infty}^{\infty}\delta(t-n). \tag{1.29}$$

This generalized function is illustrated in Figure 1.11.

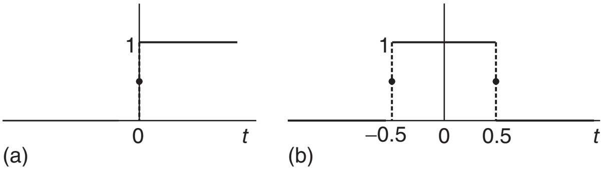

Figure 1.10 Unit step function (a) and rectangle function (b).

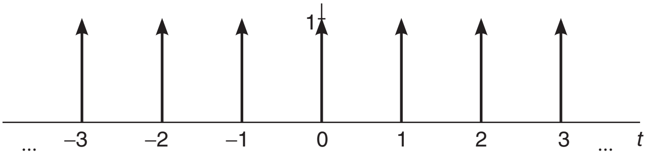

Figure 1.11 Dirac comb or sampling function, III(t).

1.2.3 Impulse Sampling

In order to process continuous-time analog signals in digital systems, an A/D conversion is required. This process maps the analog signals into a set of discrete values, both for time and space.

A uniformly sampled signal is the result of examining an analog signal at periodic intervals. In this book, we will work exclusively with unit sampling because, in general, it is unnecessary to know how much time elapses between successive samples. Assuming that the temporal width of each sample approaches zero, these samples can be represented as a sequence of impulse functions. For instance, considering the unit sampling of a continuous-time analog signal $f(t)$, we have

$$f(t)\text{III}(t) = \{f(n)\} = \sum_{n=-\infty}^{\infty} f(n)\delta(t-n) \tag{1.30}$$

where

$$f(n) = f(t)|_{t=n}, \quad n \in \mathbb{Z}. \tag{1.31}$$

As illustrated in Figure 1.12, this means that the sampled signal is composed of a series of equally spaced impulse functions, whose weights represent the values of the original signal at the sampling instants.

Because of the properties of Dirac's delta, only the information observed at the discrete times $\{t : t = n, n \in \mathbb{Z}\}$ remains after the sampling process. For this reason, the sampled signal $f(t)\text{III}(t)$ is often reduced to a discrete sequence of data

$$\{f(n)\} = \{f(0), f(1), f(2), \ldots\}. \tag{1.32}$$

In general, a sampled signal contains less information than its continuous-time counterpart unless certain conditions are fulfilled during the sampling process. These conditions are set by the Nyquist–Shannon sampling theorem discussed next.

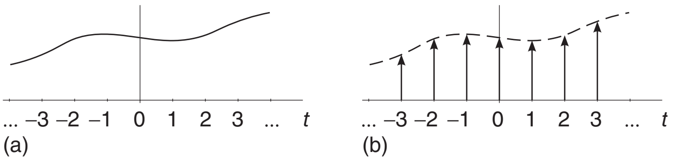

Figure 1.12 Continuous-time analog signal (a) and its unit sampled counterpart (b).

1.2.4 Nyquist–Shannon Sampling Theorem

The first part of the Nyquist–Shannon theorem states that a band-limited signal $f(t)$ that contains no frequencies higher than F_0 hertz is completely determined by its sample values if the sampling frequency $1/T_s$ is greater than twice the bandwidth of $f(t)$, that is

$$\frac{1}{T_s} > 2F_0, \tag{1.33}$$

where the sampling rate of $2F_0$ is called the *Nyquist rate*. Rewriting this condition in terms of an angular bandwidth B,we have

$$B = 2\pi F_0 < \frac{\pi}{T_s}, \tag{1.34}$$

which under unit sampling (as typically assumed in fringe pattern analysis) is reduced to

$$B < \pi. \tag{1.35}$$

The second part of the Nyquist–Shannon sampling theorem states that the band-limited continuous signal $f(t)$ can be reconstructed from its discrete samples $\{f(n)\}$ using the following interpolation formula:

$$f(t) = \sum_{n=-\infty}^{\infty} f(n)\,\mathrm{sinc}(t-n), \tag{1.36}$$

where $\mathrm{sinc}(t) = [\sin(\pi t)]/(\pi t)$. This means that under proper conditions an analog signal and its digital sampling contain the same information.

From now on, unless explicitly indicated otherwise, we will assume that all the discrete functions under study are sampled according to the Nyquist criterion (Eqs. 1.33–1.35). The demonstration of both parts of the Nyquist–Shannon sampling theorem is a consequence of the spectral characteristics of the discrete signals, so we will return to this topic after reviewing the Fourier transform in Section 1.5.1. With this, we end our short review about digital sampling, and now we will proceed to the digital linear systems theory.

1.3 Linear Time-Invariant (LTI) Systems

In this section, we review the basic theory of LTI systems commonly used in modern fringe pattern analysis. For a much more complete study of this topic, we recommend [23, 24].

1.3.1 Definition and Properties

A system is a mathematical model of a physical process that relates the input (or excitation) signal to the output (or response) signal.

Let $I(\cdot)$ and $f(\cdot)$ be, respectively, the input and output signals of a system. Then the system is viewed as a transformation (or mapping) of $I(\cdot)$ into $f(\cdot)$. This transformation is represented by the mathematical notation

$$\mathbf{L}\{I(\cdot)\} = f(\cdot), \tag{1.37}$$

where $\mathbf{L}\{\cdot\}$ is the operator representing some well-defined rule by which $I(\cdot)$ is transformed into $f(\cdot)$. If the input and output are continuous-time signals $I(t)$ and $f(t)$, respectively, then the system is called a *continuous-time system* (Figure 1.13a). If the input and output are discrete-time signals or sequences $\{I(n)\}$ and $\{f(n)\}$, respectively, then the system is called a *discrete-time system* (Figure 1.13b).

An operator $\mathbf{L}(\cdot)$ that satisfies the following condition is called a *linear operator*, and the system represented by a linear operator is called a *linear system*: given that $\mathbf{L}\left(I_1\right) = f_1$ and $\mathbf{L}\left(I_2\right) = f_2$, then

$$\mathbf{L}\left(\alpha_1 I_1 + \alpha_2 I_2\right) = \alpha_1 f_1 + \alpha_2 f_2, \tag{1.38}$$

where α_1 and α_2 are arbitrary scalars. Equation (1.38) is known as the *superposition property*.

A system is called *time-invariant* if a time shift in the input signal causes the same time shift in the output signal. Thus, for continuous-time signals, the system is time-invariant if

$$\mathbf{L}[I(t-t_0)] = f(t-t_0) \tag{1.39}$$

for any real value of t_0. If the system is linear and also time-invariant (Eqs. 1.38–1.39), then it is called a *linear time-invariant* (*LTI*) system.

1.3.2 Impulse Response of LTI Systems

In signal processing, the impulse response function of a dynamic system is the output obtained when the input signal is a unitary impulse function. The unitary impulse can be modeled as a Dirac delta function for continuous-time systems.

$$h(t) = \mathbf{L}\{\delta(t)\}. \tag{1.40}$$

Any LTI system is completely characterized by its impulse response; for any input function, the output function can be calculated in terms of the input and the

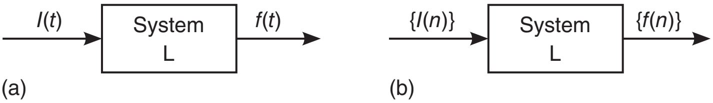

Figure 1.13 (a) Continuous-time and (b) discrete-time linear systems.

impulse response. For instance, assuming the input $I(t)$ as a sampled function

$$f(t) = \mathbf{L}\left\{ I(t) \sum_{n=-\infty}^{\infty} \delta(t-n) \right\} = \mathbf{L}\left\{ \sum_{n=-\infty}^{\infty} I(n)\delta(t-n) \right\}. \tag{1.41}$$

Since the system is linear, we have

$$f(t) = \sum_{n=-\infty}^{\infty} I(n)\mathbf{L}\left\{\delta(t-n)\right\}. \tag{1.42}$$

Applying the time-invariant condition, we have

$$h(t-n) = \mathbf{L}\{\delta(t-n)\}. \tag{1.43}$$

Finally, substituting Eq. (1.43) in Eq. (1.42), we obtain

$$f(t) = \sum_{n=-\infty}^{\infty} I(n)h(t-n) = I(t) * h(t). \tag{1.44}$$

The result stated in Eq. (1.44) is valid even with continuous (nonsampled) signals and the demonstration follows the same steps (Eqs. 1.42–1.44).

Example: Impulse Response of a Three-Step Averaging System

Consider the three-step averaging system illustrated in Figure 1.14 where the output $f(t)$ is given by the average value between the current input $I(t)$ and the two previous input values. That is

$$f(t) = (1/3)[I(t) + I(t-1) + I(t-2)]. \tag{1.45}$$

Applying the shifting property of Dirac's delta (Table 1.1), it is straightforward to see that

$$\begin{aligned} f(t) &= I(t) * (1/3)[\delta(t) + \delta(t-1) + \delta(t-2)] = I(t) * h_3(t),\\ h_3(t) &= (1/3)[\delta(t) + \delta(t-1) + \delta(t-2)]. \end{aligned} \tag{1.46}$$

This kind of averaging system is commonly used in signal processing (particularly for noise rejection) and, as we will show in Section 1.5, it corresponds to a normalized low-pass filter.

Example: Centered Three-Step Averaging System

In digital processing, it is very common to use an averaging mask for low-pass filtering. For instance, the system illustrated in Figure 1.15 corresponds to the

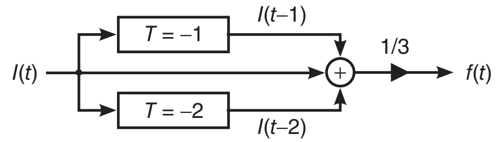

Figure 1.14 Block diagram of a three-step averaging linear filter.

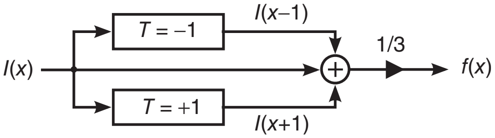

Figure 1.15 Block diagram of a centered three-step averaging linear filter.

convolution of the input signal with a unidimensional averaging mask, with an input response given by

$$h(x) = (1/3)[\delta(x-1) + \delta(x) + \delta(x+1)]. \tag{1.47}$$

Clearly, this system is almost identical to the one described in Eq. (1.46), except that this "centered" system is not causal: every pixel in the output is given by the average with its neighboring pixels (both the previous and the next). Nevertheless, this noncausality is not an issue since nowadays we can process in delayed time. Similarly, for the linear systems theory is perfectly equivalent to work with centered or noncentered impulse responses.

1.3.3 Stability Criterion: Bounded-Input Bounded-Output

The "bounded-input bounded-output" (BIBO) criterion is one of the most commonly used criteria in the study of linear systems. This criterion establishes that a system is considered stable if for any bounded input

$$|I(\cdot)| \le k_1, \tag{1.48}$$

the corresponding output is also bounded, that is

$$|f(\cdot)| \le k_2, \tag{1.49}$$

where k_1 and k_2 are finite real constants.

Consider a discrete-time invariant linear system, where the output $f(t)$ is given by the convolution between the input signal $I(t)$ and the system's impulse response $h(t)$. Assuming BIBO stability, we have

$$|f(t)| = \left|\sum_{n=-\infty}^{\infty} I(n)h(t-n)\right| \le \sum_{n=-\infty}^{\infty} |I(n)||h(n)|, \tag{1.50}$$

where we have applied the triangle inequality and the time-invariant property over the impulse response. Substituting the bounded-input condition (Eq. 1.48) results in

$$|f(t)| \le k_1 \sum_{n=-\infty}^{\infty} |h(n)|. \tag{1.51}$$

And, since we are supposing that the output is also bounded, this means that a discrete LTI system will be BIBO stable *if and only if* its impulse response is absolutely summable, that is

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty. \tag{1.52}$$

For any digital linear system, Eq. (1.52) is a necessary and sufficient condition for system stability. Similarly, it can be proved that a continuous LTI system will be BIBO stable *if and only if* its impulse response is absolute integrable [23], that is

$$\int_{-\infty}^{\infty} |h(t)|dt < \infty. \tag{1.53}$$

In practice, the analysis of linear systems almost always requires the application of integral transforms. This means that, instead of the impulse response of a linear system, one usually finds its integral transform, the so-called the transfer function. Nevertheless, the BIBO stability criterion can also be easily evaluated in such transformed space, as we will show in the following sections.

1.4 Z-Transform Analysis of Digital Linear Systems

The Z-transform is a useful tool in the analysis of discrete-time signals and systems that may be defined as the discrete-time counterpart of the Laplace transform. The Z-transform may be used to solve constant-coefficient difference equations, to evaluate the response of a LTI system to a given input, and to design linear filters.

In fringe pattern analysis, the pioneering works in the application of Z-transform for the analysis of PSAs are due to Surrel [25–27] (although the connection between both formalisms was not explicit in his publications).

1.4.1 Definition and Properties

The bilateral Laplace transform is defined by

$$\mathcal{L}\{f(t)\} = \int_{-\infty}^{\infty} f(t)\exp(-st)dt, \tag{1.54}$$

where $s \in \mathbb{C}$, so $\exp(s) = \exp(\alpha + i\omega) = r\exp(i\omega)$. Considering a sampled function (as in Eq. 1.30) and taking its Laplace transform, we have

$$\begin{aligned}\mathcal{L}\{f(t)\text{III}(t)\} &= \int_{-\infty}^{\infty} f(t) \sum_{n=-\infty}^{\infty} \delta(t-n)\exp(-st)dt,\\ &= \sum_{n=-\infty}^{\infty} \int_{-\infty}^{\infty} f(t)\delta(t-n)\exp(-st)dt,\\ &= \sum_{n=-\infty}^{\infty} f(n)\exp(-sn).\end{aligned} \tag{1.55}$$

Making the change of variable $z = \exp(s)$, one finds the most commonly used expression for the bilateral Z-transform of a discrete signal

$$\begin{aligned}\mathcal{Z}\{f(t)\} &= \mathcal{L}\{f(t)\mathrm{III}(t)\},\\ F(z) &= \sum_{n=-\infty}^{\infty} f(n)z^{-n},\end{aligned} \tag{1.56}$$

where, once again, $f(n) = f(t)|_{t=n}$ and $z = r\exp(i\omega)$. For cases where the function $f(t)$ is defined only for $t \geq 0$ (for instance, in a causal system), the single-sided or unilateral Z-transform is defined as

$$\mathcal{Z}\{f(t)\} = F(z) = \sum_{n=0}^{\infty} f(n)z^{-n}. \tag{1.57}$$

Should be noted that the Z-transform is usually defined only for discrete sequences $\mathcal{Z}\{f[n]\}$. However, we are explicitly extending its definition to describe the Laplace transform of sampled functions.

1.4.2 Region of Convergence (ROC)

In general, the Z-transforms are infinite series, so convergence is a very important aspect to take into consideration. The region of convergence (ROC) is defined as the set of points in the complex plane for which a Z-transform summation converges. That is

$$\mathrm{ROC} = \left\{ z \; : \; \left| \sum_{n=-\infty}^{\infty} f(n)z^{-n} \right| < \infty \right\}. \tag{1.58}$$

Using the triangle inequality, we can rewrite the above summation condition as

$$\left| \sum_{n=-\infty}^{\infty} f(n)z^{-n} \right| \leq \sum_{n=-\infty}^{\infty} \left| f(n)z^{-n} \right|. \tag{1.59}$$

Furthermore, using the polar form in the right-hand side of the inequality [24], we have

$$\begin{aligned}\sum_{n=-\infty}^{\infty} \left| f(n)z^{-n} \right| &= \sum_{n=-\infty}^{-1} \left| f(n)r^{-n} \right| + \sum_{n=0}^{\infty} \left| f(n)r^{-n} \right|,\\ &= \sum_{n=1}^{\infty} \left| f(-n)r^{n} \right| + \sum_{n=0}^{\infty} \left| f(n)r^{-n} \right|.\end{aligned} \tag{1.60}$$

If the first sum in Eq. (1.60) converges, there must exist some region where the sequence $\{f(-n)r^n\}$ is absolutely summable; this region will be given by the points in the complex plane inside a circle of some radius r_1. On the other hand, if the second sum converges, there must exist some region where the sequence $\{f(n)r^{-n}\}$ is absolutely summable; this region will be given by the points in the complex plane outside a circle of some radius r_2. Therefore, the ROC for both summations will be given by some annular region defined by $r_2 < r < r_1$. Following our chain of inequalities (Eqs. 1.58–1.60), we know that such annular region also guarantees the convergence of $|F(z)|$. Thus, in general, the ROC for a given Z-transform can be described in the form

$$\text{ROC} = \left\{z \ : \ |F(z)| < \infty\right\} = \left\{z \ : \ r_2 < |z| < r_1\right\}. \tag{1.61}$$

Example: Z-Transform of a Finite-Duration Sequence

Consider a sequence of data $\{f(n)\}$ where $f(n) \neq 0$ only for a finite number of values $n_1 < n < n_2$. So

$$F(z) = \sum_{n=n_1}^{n_2} f(n)z^{-n}. \tag{1.62}$$

Convergence of this expression simply requires that $|f(n)| < \infty$ for $n_1 < n < n_2$. Then z may take on all values except $z = \infty$ if $n_1 < 0$ and $z = 0$ if $n_2 > 0$. Thus, we conclude that the ROC of finite-duration sequence is at least $0 < |z| < \infty$, and it may include either $z = 0$ or $z = \infty$.

1.4.3 Poles and Zeros of a Z-Transform

Many signals or systems of interest have Z-transforms that are rational functions of z. That is

$$F(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{q} b_k z^{-k}}{\sum_{k=0}^{p} a_k z^{-k}}. \tag{1.63}$$

Factoring the numerator and denominator polynomials, that is, $A(z)$ and $B(z)$, a rational Z-transform may be expressed as follows:

$$F(z) = c_0 \frac{\prod_{k=1}^{q}(1 - \beta_k z^{-1})}{\prod_{k=1}^{p}(1 - \alpha_k z^{-1})}. \tag{1.64}$$

The roots of the numerator polynomial β_k are referred to as the *zeros* of $F(z)$, and the roots of the denominator polynomial α_k are referred to as the *poles* of $F(z)$. The poles and zeros uniquely define the functional form of a rational Z-transform to within a constant c_0. Therefore, they provide a concise representation for $F(z)$ which is often represented pictorially in terms of a pole-zero plot in the Z-plane. In a pole-zero plot, the location of the poles is indicated by crosses (×) and the location the zeros by circles (∘), with the ROC indicated by shading the appropriate region of the z-plane. The multiplicity of (m-order) poles or zeros is usually indicated by a number close to the corresponding cross or circle (for $m > 1$).

Example: Z-Transform of an Exponential Function

Consider a discrete-time exponential function defined for $t > 0$ as $f(t)\text{III}(t) = \sum_{n=0}^{\infty} a^n \delta(t - n)$. Then, following our definition (Eq. 1.56), its Z-transform is

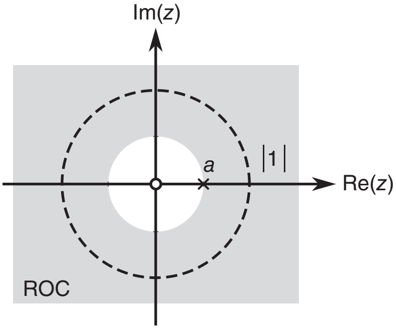

Figure 1.16 Pole-zeros plot for Z-transform in Eq. (1.65) for $a < 1$.

given by

$$\begin{aligned} F(z) &= \sum_{n=-\infty}^{\infty} f(n)z^{-n} = \sum_{n=0}^{\infty}(a/z)^n, \\ &= \frac{1}{1-(a/z)} = \frac{z}{z-a}, \end{aligned} \tag{1.65}$$

where the summation converges for $|(a/z)| < 1$. Therefore, the ROC is exterior to the circle defined by the points in the complex plane given by $|z| = |a|$ as illustrated in Figure 1.16. Note that, if $|a| < 1$, the unit circle is inside the ROC.

1.4.4 Inverse Z-Transform

The inverse Z-transform is formally defined by

$$\{f(n)\} = \mathcal{Z}^{-1}\{F(z)\} = \frac{1}{i2\pi}\oint_C F(z)z^{n-1}dz, \tag{1.66}$$

where C is a counterclockwise closed path encircling the origin and all of the poles of $F(z)$, and entirely within the ROC . However, in practice we rarely rely in the contour integration method (Eq. 1.66) to find an inverse Z-transform. Instead, we usually perform some algebraic manipulations (e.g., partial-fraction expansion) to solve in function of well-known Z-transform pairs. In Tables 1.2 and 1.3, we summarize some useful properties of the most commonly used Z-transform pairs [28].

For illustrative purposes, next we demonstrate one of the properties presented in Table 1.2.

Example: Z-Transform of the Time-Shifting Operator

Consider a discrete sequence given by the sampling of some analog Signal, $\{f(n)\} = f(t)\text{III}(t)$ with $n = \{0, 1, 2, \ldots\}$. Then, the sequence $\{f(n-k)\}$ models the sampled data with a temporal shifting of k samples. Taking its Z-transform, we

Table 1.2 Some properties of the Z-transform.

Time domain	Z-domain	ROC
$f(t)$	$F(z)$	R
$f_1(t)$	$F_1(z)$	R_1
$f_2(t)$	$F_2(z)$	R_2
$a_1 f_1(t) + a_2 f_2(t)$	$a_1 F_1(z) + a_2 F_2(z)$	$R_1 \cap R_2$
$f(t-k)$	$z^{-k} F(z)$	$R \cap \{0 < \lvert z\rvert < \infty\}$
$z_0^t f(t)$	$F(z/z_0)$	$\lvert z_0\rvert R$
$e^{i\omega_0 t} f(t)$	$F(e^{i\omega_0} z)$	R
$f(-t)$	$F(1/z)$	$1/R$
$t f(t)$	$-zF'(z)$	R
$f_1(t) * f_2(t)$	$F_1(z)\, F_2(z)$	$R_1 \cap R_2$

Table 1.3 Commonly used Z-transform pairs.

Time domain	Z-domain	ROC
$\delta(t)$	1	All z
$a^n u(t)\mathrm{III}(t)$	$\frac{z}{a-z}$	$\lvert z\rvert > \lvert a\rvert$
$t a^t u(t)\mathrm{III}(t)$	$\frac{az^{-1}}{(1-az^{-1})^2}$	$\lvert z\rvert > \lvert a\rvert$
$\cos(\omega_0 t)u(t)\mathrm{III}(t)$	$\frac{1-\cos(\omega_0)z^{-1}}{1-2\cos(\omega_0)z^{-1}+z^{-2}}$	$\lvert z\rvert > \lvert 1\rvert$
$\sin(\omega_0 t)u(t)\mathrm{III}(t)$	$\frac{\sin(\omega_0)z^{-1}}{1-2\cos(\omega_0)z^{-1}+z^{-2}}$	$\lvert z\rvert > \lvert 1\rvert$

have

$$\mathcal{Z}\left\{f(t-k)\right\} = \sum_{n=-\infty}^{\infty} f(n-k)z^{-n}. \tag{1.67}$$

Introducing a shifted change of the variable, that is, $j = n - k$, results in

$$\mathcal{Z}\left\{f(t-k)\right\} = z^{-k} \sum_{j=-\infty}^{\infty} f(j)z^{-j} = z^{-k}F(z), \tag{1.68}$$

and the ROC of $z^{-k}F(z)$ is the same as that of $F(z)$ except for $z = 0$ if $k > 0$, or $z = \infty$ if $k < 0$.

1.4.5 Transfer Function of an LTI System in the Z-Domain

As stated in Section 1.3, any LTI system can be fully described in the temporal domain as (Eq. 1.44)

$$f(t) = I(t) * h(t), \tag{1.69}$$

where $I(t)$, $f(t)$, and $h(t)$ represent the input, output, and the impulse response of the system, respectively. Taking the Z-transform of the above equation and applying the convolution property (Table 1.2), we have

$$F(z) = I(z)H(z). \tag{1.70}$$

Now, since Eq. (1.70) is an algebraic one, it is possible to solve for the ratio $F(z)/I(z)$ to find the so-called transfer function of the LTI, $H(z)$:

$$\frac{F(z)}{I(z)} = H(z). \tag{1.71}$$

And, since, by definition, the transfer function $H(z)$ is the Z-transform of the impulse response function of the system $h(t)$, we have

$$H(z) = \frac{F(z)}{I(z)} = \sum_{n=-\infty}^{\infty} h(n)z^{-n}. \tag{1.72}$$

A we see from Eqs. (1.71) and (1.72), the transfer function $H(z)$ will be in general a rational Z-transform. Thus, the correspondent ROC will be defined by the location of its poles (Eqs. 1.63 and 1.64). To further illustration of this point, we present some illustrative examples at the end of this section, together with the multitude of linear filters that we will analyze in the rest of the book.

1.4.6 Stability Evaluation by Means of the Z-Transform

As discussed in Section 1.3.3, a discrete-time linear system is said to be BIBO stable if its impulse response function $h(t)$ is absolutely summable. That is (Eq. 1.52)

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty. \tag{1.73}$$

From Eqs. (1.72) and (1.73), it is straightforward to see that the BIBO stability criterion is equivalent to requiring the transfer function $F(z)$ to be absolutely summable in the unit circle of the Z-domain:

$$\left\{ \sum_{n=-\infty}^{\infty} |h(n)| = \sum_{n=-\infty}^{\infty} |h(n)z^{-n}|_{z\in U} \right\} < \infty, \tag{1.74}$$

where the unit circle is defined by

$$U(z) = \{z \ : \ |z| = 1\}. \tag{1.75}$$

In other words, a linear system will be BIBO stable *if and only if* the ROC of its transfer function include the unit circle. Furthermore, following the ROC definition (Eq. 1.58) one may prove that a causal system will be BIBO stable *if and only if* all the poles of its transfer function are located *inside* the unit disc defined by

$$\overline{U}(z) = \{z \ : \ |z| < 1\}. \tag{1.76}$$

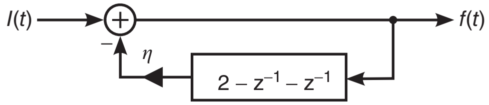

Figure 1.17 Block diagram of a second-order recursive filter.

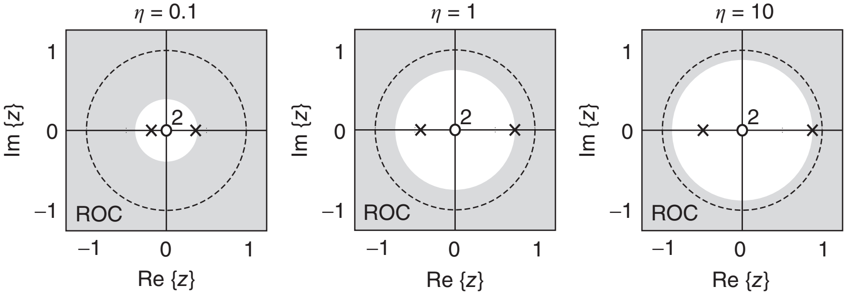

Figure 1.18 Pole-zeros plot for the second-order recursive filter discussed in Eq. (1.77) with $\eta = 0.1, 1, 10$. Note that, for $\eta \gg 1$, each pole approaches asymptotically $-1/2$ and 1, respectively.

Example: Stability Evaluation of a Recursive Digital Filter

Consider the following difference equation that describes the second-order recursive filter illustrated in Figure 1.17:

$$f(t) = I(t) - \eta[2f(t) - f(t-1) - f(t-2)]. \tag{1.77}$$

Taking the Z-transform of the above equation and applying the time-shifting property from Table 1.2, it is straightforward to find that

$$H(z) = \frac{F(z)}{I(z)} = \frac{1}{1+\eta(2-z^{-1}-z^{-2})} = \frac{z^2}{(1+2\eta)z^2 - \eta z - \eta}. \tag{1.78}$$

As illustrated in Figure 1.18, applying the quadratic formula in the right-side denominator, we find that both poles of $H(z)$ are located inside the unit disc $\overline{U}(z)$ for $0 < \eta < \infty$. Thus, since the unit circle is part of the ROC of $H(z)$, this filter is said to be *BIBO stable*.

1.5 Fourier Analysis of Digital LTI Systems

In this section, we apply Fourier transformation to analyze signals and systems. The Fourier transform allows us convert the mathematical representation of a signal in time into a that of the signal in frequency, known as its *frequency spectrum*. As in most textbooks, the term *Fourier transform* will refer to both the transform operation and to the complex-valued function it produces [29].

Table 1.4 Fourier transforms for some common mathematical operations.

Operation	Time function $f(t)$	Transform $F(\omega)$
Linearity	$af_1(t) + bf_2(t)$	$aF_1(\omega) + bF_2(\omega)$
Reversal	$f(-t)$	$F(-\omega)$
Symmetry	$F(t)$	$f(-\omega)$
Scaling	$f(at)$	$(1/\|a\|)\, F(\omega/a)$
Time delay	$f(t{-}t_0)$	$F(\omega)\exp(-i\omega t_0)$
Time differentiation	$f^{(n)}(t)$	$(i\omega)^n F(\omega)$
Frequency translation	$f(t)\exp[i\omega_0 t]$	$F(\omega - \omega_0)$
Convolution	$f(t) * h(t)$	$F(\omega)H(\omega)$
Multiplication	$f(t)h(t)$	$F(\omega) * H(\omega)$
Energy	$\int_{-\infty}^{\infty} \|f(t)\|^2 dt$	$\int_{-\infty}^{\infty} \|F(\omega)\|^2 d\omega$

1.5.1 Definition and Properties of the Fourier Transform

There are several conventions for defining the Fourier transform of an integrable function. For a continuous-time signal, we will adopt the following convention to define the direct Fourier transform and its inverse:

$$\mathcal{F}\{f(t)\} = F(\omega) = \int_{-\infty}^{\infty} f(t)\exp(-i\omega t)dt, \tag{1.79}$$

$$\mathcal{F}^{-1}\{F(\omega)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp(i\omega t)d\omega. \tag{1.80}$$

The conditions for the existence of the Fourier transform (Eq. 1.79) and its inverse (Eq. 1.80) represent a very broad topic, so we will not discuss them here. Nevertheless, it is important to highlight that, if $f(t)$ is absolutely integrable, that is

$$\int_{-\infty}^{\infty} |f(t)|\, dt < \infty, \tag{1.81}$$

this is a sufficient condition for the existence of $\mathcal{F}\{f(t)\}$. For convenience, in Table 1.4 we summarize other useful properties of the Fourier transform, and in Table 1.5 some commonly used Fourier transform pairs.

1.5.2 Discrete-Time Fourier Transform (DTFT)

When working with sampled signals, Eq. (1.79) may be reduced to the so-called discrete-time Fourier transform (DTFT), that is

Table 1.5 Commonly used Fourier transform pairs.

Time function $f(t)$	Transform $F(\omega)$
1	$\delta(\omega)$
$\exp(i\omega_0 t)$	$\delta(\omega - \omega_0)$
$\cos(\omega_0 t)$	$(1/2)[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]$
$\sin(\omega_0 t)$	$(1/2i)[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)]$
$\exp(-\pi t^2)$	$\exp(-\pi\omega^2)$
$\text{II}(t) = \text{rect}(t)$	$\text{sinc}(\omega) = \sin(\pi\omega)/\pi\omega$
$\delta(t - t_0)$	$\exp(-i\omega t_0)$
$\sum_{n=-\infty}^{\infty} \delta(t - n)$	$\sum_{n=-\infty}^{\infty} \delta\,(\omega - 2\pi n)$

$$\begin{aligned}\mathcal{F}\{f(t)\text{III}(t)\} &= \mathcal{F}\left\{\sum_{n=-\infty}^{\infty} f(n)\delta(t-n)\right\},\\ &= \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} f(n)\delta(t-n)\exp(-i\omega t)dt,\\ F(\omega) &= \sum_{n=-\infty}^{\infty} f(n)\exp(-i\omega n). \qquad (1.82)\end{aligned}$$

Note that, according to the above equation, the Fourier spectrum of a sampled signal (or a discrete sequence) is periodic and continuous in the frequency domain. Also note that the analysis equation (Eq. 1.82) converges if the discrete sequence $\{f(n)\}$ is absolutely summable. That is

$$\sum_{n=-\infty}^{\infty} |f(n)| < \infty, \qquad (1.83)$$

is a sufficient condition for the existence of the DTFT.

1.5.3
Relation Between the DTFT and the Z-Transform

As hinted before, the DTFT can be considered a particular case of the Z-transform. To illustrate this, let us reintroduce the change of variable $z = r\exp(i\omega)$ in the Z-transform formula, as

$$\begin{aligned}F(z) &= \sum_{n=-\infty}^{\infty} f(n)[r\exp(i\omega)]^{-n} \qquad (1.84)\\ &= \sum_{n=-\infty}^{\infty} [f(n)r^{-n}]\exp(-i\omega n) = \mathcal{F}\{f(n)r^{-n}\}.\end{aligned}$$

This means that the Z-transform can be seen as the DTFT of an exponentially weighted sequence. Likewise, the DTFT of can be seen as the Z-transform evaluated

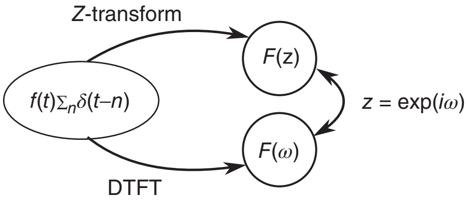

Figure 1.19 Relation between the DTFT and the Z-transform.

in the unit circle $U(z) = \{z \ : \ |z| = 1\}$. That is

$$F(z)|_{z \in U} = \sum_{n=-\infty}^{\infty} f(n) \exp(i\omega) = F(\omega), \tag{1.85}$$

assuming, of course, that $U(z)$ is part of the ROC of $F(z)$. These relations are schematically illustrated in Figure 1.19.

1.5.4
Spectral Interpretation of the Sampling Theorem

As discussed in Section 1.2.4, the Nyquist–Shannon sampling theorem is composed of two parts. According to the first part, a band-limited signal $f(t)$ is completely determined by its sampled values if the sampling frequency $(1/T_s)$ is greater than twice the bandwidth of $f(t)$ (Eq. 1.33):

$$\frac{1}{T_s} > 2F_0. \tag{1.86}$$

Or, in terms of the angular bandwidth under unit sampling (Eq. 1.35):

$$B < \pi. \tag{1.87}$$

This is the so-called Nyquist criterion. The validity of the first part this theorem can conceptually tested as follows: consider a band-limited analog signal $f(t)$ such that

$$\mathcal{F}\{f(t)\} = F(\omega) = 0 \quad \text{for} \quad |\omega| > B. \tag{1.88}$$

Assuming unit sampling is obtained, the discrete sequence of data $\{f(n)\}$ with a DTFT given by

$$\mathcal{F}\{f(t)\mathrm{III}(t)\} = F(\omega) * \sum_{n=-\infty}^{\infty} \delta(\omega - 2\pi n) = \sum_{n=-\infty}^{\infty} F(\omega - 2\pi n). \tag{1.89}$$

That is, the spectrum of a sampled function is given by copies of the spectrum of the continuous function, shifted by multiples of 2π and combined by addition. Thus, since such spectral copies will not overlap if Eqs. (1.86) and (1.87) are fulfilled, one can reconstruct $F(\omega)$ from the DTFT by applying a brick-wall low-pass filter

$$\mathrm{II}\,(\omega/2\pi)\,\mathcal{F}\{f(t)\mathrm{III}(t)\} = \mathrm{II}\,(\omega/2\pi) \sum_{n=-\infty}^{\infty} F(\omega - 2\pi n) = F(\omega). \tag{1.90}$$

This ideal filtering is translated to the temporal domain as

$$f(t) = \frac{\sin(\pi t)}{\pi t} * \sum_{n=-\infty}^{\infty} f(n)\delta(t-n) = \sum_{n=-\infty}^{\infty} f(n)\,\mathrm{sinc}(t-n), \tag{1.91}$$

which is precisely the interpolation formula presented in Eq. (1.36); this is the second part of the Nyquist–Shannon sampling theorem.

On the other hand, if the Nyquist criterion is not fulfilled, which is called *sub-Nyquist sampling*, the spectral replicas overlap, producing spectral distortion and making the isolation of the spectrum of the analog signal impossible. This is illustrated in Figure 1.20.

To summarize, the spectrum of a sampled signal comprises multiple copies of the spectrum of the analog signal, shifted by multiples of 2π and combined by addition. Assuming the sampling process fulfilled the Nyquist criterion, this enables us to fully recover the analog signal from its sampled counterpart by applying a brick-wall low-pass filter in the Fourier domain (or, alternatively, an extrapolation in the temporal domain). However, if the Nyquist criterion is not fulfilled (sub-Nyquist sampling), the spectral replicas overlap with each other producing distortion.

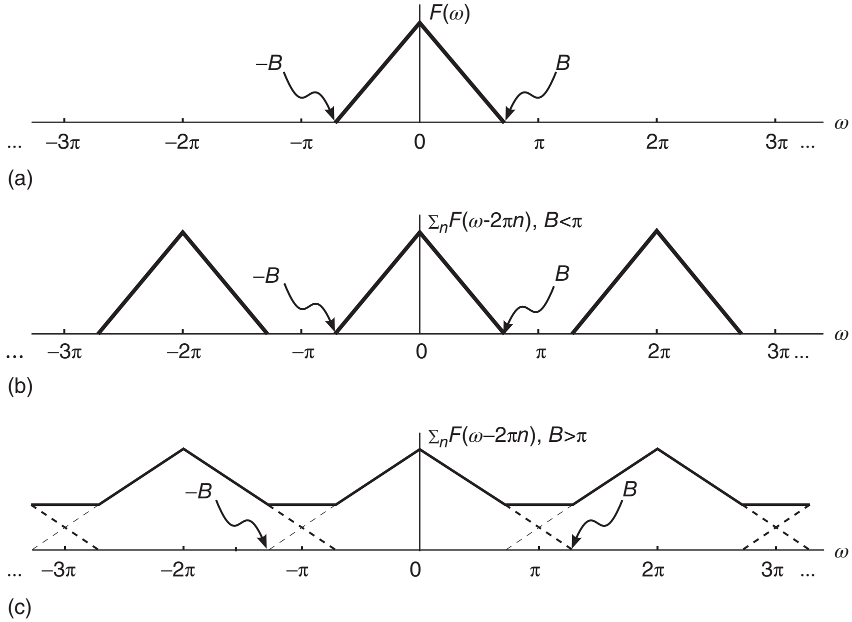

Figure 1.20 Hypothetical spectra of a band-limited analog signal (a) and its discrete counterpart fulfilling the Nyquist criterion (b) and under sub-Nyquist sampling (c).

1.5.5 Aliasing: Sub-Nyquist Sampling

Aliasing refers to an effect that causes different signals to become indistinguishable under sub-Nyquist sampling. That is, if two continuous signals produce the same set of data when sampled (at least one of them without fulfilling the Nyquist criterion), we say that such signals are aliases of each other.

Actual signals have finite duration and their frequency content, as defined by the Fourier transform, has no upper bound [24]. Thus, the Nyquist criterion cannot be strictly fulfilled in real-life applications and some negligible amount of aliasing is always to be expected.

Aliasing in sinusoidal signals. To illustrate the aliasing effect, consider the following analog sinusoidal Signals:

$$\begin{aligned} f_1(t) &= \cos(\omega_1 t), \\ f_2(t) &= \cos(\omega_2 t), \end{aligned} \tag{1.92}$$

given in the Fourier domain as

$$\begin{aligned} F_1(\omega) &= (1/2)[\delta(\omega+\omega_1)+\delta(\omega-\omega_1)], \\ F_2(\omega) &= (1/2)[\delta(\omega+\omega_2)+\delta(\omega-\omega_2)]. \end{aligned} \tag{1.93}$$

If the following equation holds for their angular frequency, that is

$$\omega_2 = \omega_1 + 2\pi, \tag{1.94}$$

the unit sampling of both continuous signals will produce the same set of data. This is illustrated in Figure 1.21 (with $\omega_1 = 2\pi/3$ purely for illustrative purposes).

Furthermore, calculating the DTFT for $f_2(t)$ we have

$$\begin{aligned} \mathcal{F}\{f_2(t)\text{III}(t)\} &= \sum_{n=-\infty}^{\infty} \cos[(\omega_1 + 2\pi)n]\exp(-i\omega n) \\ &= \sum_{n=-\infty}^{\infty} \cos(\omega_1 n)\exp(-i\omega n) = \mathcal{F}\{f_1(t)\text{III}(t)\}. \end{aligned} \tag{1.95}$$

Thus, even when both sinusoidal signals are completely different in the continuous domain, their sampling produces exactly the same set of data and therefore the same DTFT. This is illustrated in Figure 1.22.

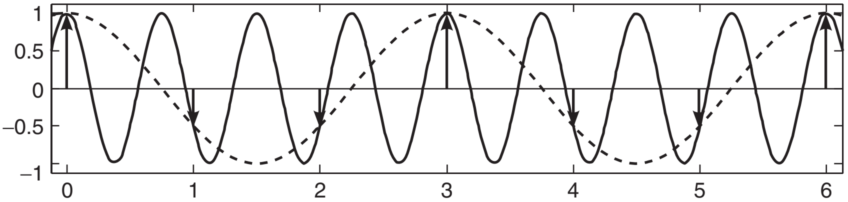

Figure 1.21 Two different continuous sinusoidal signals that fit the same set of samples as an illustration of the aliasing effect.

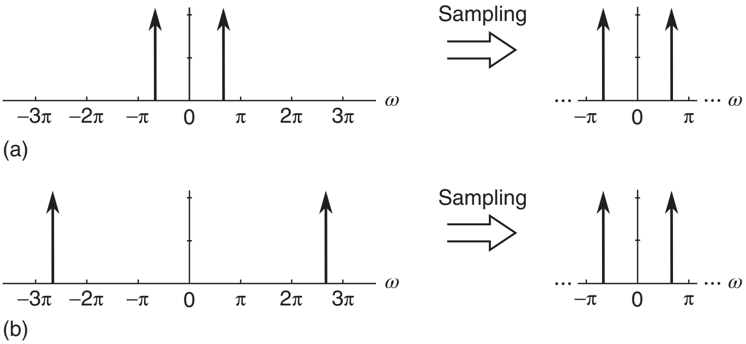

Figure 1.22 Continuous sinusoidal signals presented in Fig. 1.21 having different spectra but, since the Nyquist criterion is not fulfilled in (b), their DTFT is exactly the same. Recall that the DTFT has a 2π periodicity, so in the right-side plot we have shown only the principal branch.

From Fourier analysis we know that the set of sinusoidal functions form an orthonormal base for square-integrable functions in the domain $(-\pi, \pi]$. Thus, because of the 2π periodicity of the DTFT, we can think both sinusoidal functions of our example as discrete components resulting from the sampling of some general continuous signal. Considering the above, we can extend this result to a more general interpretation of the aliasing effect which should help us to understand many problematic phenomena in fringe pattern analysis (such as high-order harmonic distortion, wrapped phase inconsistencies, etc.).

During the sampling process of any continuous signal, the energy of those spectral components with angular frequency $\{\omega : |\omega| > \pi\}$ will be distributed to its alias on the principal branch $(-\pi, \pi]$.

Anti-alias filtering. In many areas of signal processing, it is common to apply an continuous low-pass filter prior to the sampling process, restricting the bandwidth of the continuous signal under study to more or less satisfy the Nyquist criterion (as illustrated in Figure 1.23). These are called *anti-alias filters*.

Aliasing can be either spatial or temporal. However, nowadays one is able to acquire a two-dimensional set of data with millions of samples from a single frame, which usually allow us to neglect the influence of spatial aliasing and to apply some anti-aliasing filtering whenever required. On the other hand, for many techniques of

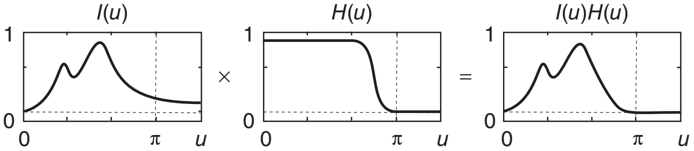

Figure 1.23 Fourier domain representation of anti-alias filtering to remove those components that do not fulfill the Nyquist criterion (as defined in Eq. 1.87).

fringe patter analysis, particularly in phase-shifting interferometry, we are usually restricted to work with a few temporal samples so temporal anti-aliasing filtering may not be feasible.

1.5.6 Frequency Transfer Function (FTF) of an LTI System

Now, let us apply the mathematical concepts briefly reviewed earlier to the analysis of linear systems. As discussed in Section 1.3, a time-invariant linear system is completely characterized by its impulse response function $h(t)$. That is, for every input $I(t)$, the corresponding output $f(t)$ is given by

$$f(t) = I(t) * h(t). \tag{1.96}$$

Taking the Fourier transform of Eq. (1.96), we have

$$F(\omega) = I(\omega)H(\omega). \tag{1.97}$$

The Fourier transform of the impulse response function $h(t)$, that is, the spectrum $H(\omega)$, is called the *FTF* (frequency transfer function) and in general it can be evaluated as the ratio

$$H(\omega) = \frac{F(\omega)}{I(\omega)}, \tag{1.98}$$

where $I(\omega) \neq 0$. As with any Fourier transform, the FTF is in general a complex-valued function

$$H(\omega) = Hr(\omega) + i\,Hi(\omega), \tag{1.99}$$

where, by definition

$$Hi(\omega) = \mathrm{Im}\{H(\omega)\}, \tag{1.100}$$

$$Hr(\omega) = \mathrm{Re}\{H(\omega)\}. \tag{1.101}$$

Another representation for any complex-valued FTF can be made in terms of its amplitude and phase, as

$$H(\omega) = |H(\omega)| \exp\{i\,\mathrm{angle}[H(\omega)]\}, \tag{1.102}$$

where

$$|H(\omega)| = \sqrt{[Hi(\omega)]^2 + [Hr(\omega)]^2} \tag{1.103}$$

and the phase(mod2π) is given by

$$\mathrm{angle}[H(\omega)] = \tan^{-1}\left[\frac{Hi(\omega)}{Hr(\omega)}\right]. \tag{1.104}$$

In some instances – particularly when plotting – one may prefer to work only with real functions. Of course, Eqs. (1.100) and (1.101) and Eqs. (1.103) and (1.104) represent real functions but in general it is more useful to describe an FTF by means of its amplitude and phase:

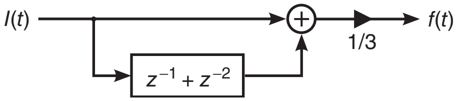

Figure 1.24 Block diagram of a three-step averaging linear filter.

- The amplitude of the FTF is time-invariant:

$$\begin{aligned}\mathcal{F}\left\{h(t+t_0)\right\} &= H(\omega)\exp(-i\omega t_0),\\ \left|H(\omega)\exp(-i\omega t_0)\right| &= |H(\omega)|\left|\exp(-i\omega t_0)\right| = |H(\omega)|\,.\end{aligned} \tag{1.105}$$

- By plotting $|H(\omega)|$, one can easily find the real zeros of $H(\omega)$, that is, those frequencies at which the system present null response:

$$\left|H(\omega_0)\right| = 0 \quad \Leftrightarrow \quad \mathrm{Re}\{H(\omega_0)\} = \mathrm{Im}\{H(\omega_0)\} = 0. \tag{1.106}$$

Example: FTF for a Three-Step Averaging System

Consider the three-step averaging system illustrated in Figure 1.24 (previously discussed in Eq. (1.46) and replicated here for convenience) where the output $f(t)$ is given by the average value between the current input $I(t)$ and the two previous input values. The impulse response of this filter is given by

$$h(t) = (1/3)[\delta(t) + \delta(t-1) + \delta(t-2)]. \tag{1.107}$$

By definition, the FTF of this filter is given by the Fourier transform of its impulse response. That is

$$\begin{aligned}\mathcal{F}\left\{h(t)\right\} &= (1/3)\mathcal{F}\left\{[\delta(t) + \delta(t-1) + \delta(t-2)]\right\},\\ H(\omega) &= (1/3)\left[1 + \exp(-i\omega) + \exp(-i2\omega)\right].\end{aligned} \tag{1.108}$$

From Eq. (1.108), calculating the amplitude and phase of the FTF results in

$$\begin{aligned}|H(\omega)| &= (1/3)\left[3 + 4\cos(\omega) + 2\cos(2\omega)\right]^{1/2},\\ \mathrm{angle}[H(\omega)] &= \tan^{-1}\left[\frac{\sin(\omega)+\sin(2\omega)}{1+\cos(\omega)+\cos(2\omega)}\right]\end{aligned} \tag{1.109}$$

which are graphically presented in Figure 1.25. Note that, as predicted in Eq. (1.82), the FTF of this system has a 2π periodicity, so we only need to plot the principal branch $(-\pi, \pi)$.

As can be seen from Figure 1.25, this three-step averaging system represents a symmetrical low-pass filter with its maximum (within the principal branch) at $\omega = 0$ and null frequency response at $\omega = \pm(2\pi/3)$.

A note about linear and semilog plots for the FTF: In some areas of signal processing, it is common to present the spectral plots using a logarithmic scale in the vertical axis and linear scale in the horizontal one; this is called a *semilog plot*. However, in fringe pattern analysis we are more interested in the study of the stop-band region than the pass-band region, and the stop-band region behavior goes out of range

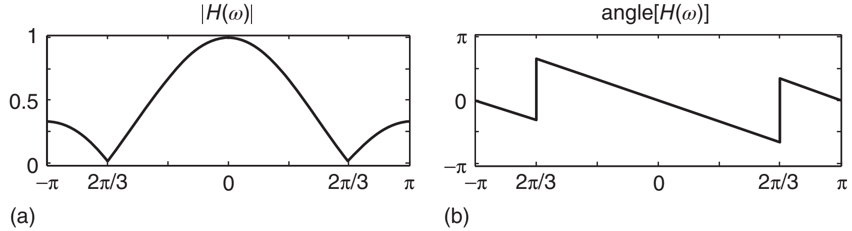

Figure 1.25 (a) Absolute value and (b) phase of the frequency transfer function of the three-step averaging system discussed in Eq. (1.108).

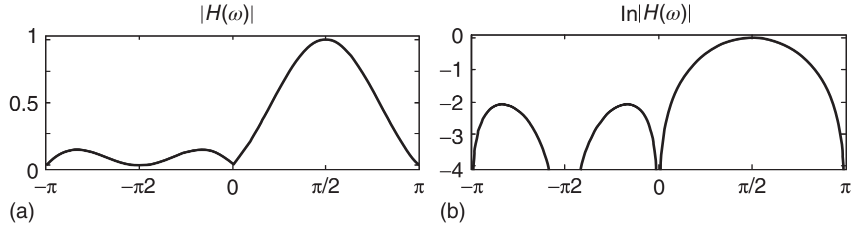

Figure 1.26 (a) Linear plot and (b) semilog plot for the FTF of the five-step band-pass filtering system discussed in Eq. (1.110).

in a semilog plot since $\log(x)$ diverges for $x = 0$. To illustrate this, consider the following five-step band-pass filter with an impulse response function given by

$$h(t) = (1/8)[\delta(t) - 2i\delta(t-1) - 2\delta(t-2) + 2i\delta(t-3) + \delta(t-4)]. \tag{1.110}$$

Taking its DTFT and factorizing the result, we find the following FTF:

$$H(\omega) = (1/8)\left[1 - e^{i\omega}\right]\left[1 - e^{-i(\omega-\pi/2)}\right]^2\left[1 - e^{i(\omega-\pi)}\right]. \tag{1.111}$$

The linear and semilog plots for $|H(\omega)|$ are presented in Figure 1.26.

As can be seen from Figure 1.26, both the pass-band and stop-band regions are clearly represented in the linear plot of $|H(\omega)|$, while the stop-band region is out of the chart in the semilog plot. For this reason, we prefer to work exclusively with linear plots of $|H(\omega)|$ for the filters that we study.

1.5.7 Stability Evaluation in the Fourier Domain

In Section 1.3.3, we showed that an LTI discrete system is said to be BIBO stable *if* its impulse response is absolutely summable (Eq. 1.52, replicated here for convenience):

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty. \tag{1.112}$$

As shown in Section 1.4.6, the above equation is fulfilled *if and only if* the unit circle of the Z-domain is part of the ROC of the transfer function. Furthermore (according to Eq. 1.85), the transfer function an LTI system evaluated in the unit circle $U(z)$ is equal to its FTF. That is

$$H(z)|_{z\in U} = \sum_{n=-\infty}^{\infty} h(n)\exp(i\omega) = H(\omega). \tag{1.113}$$

Thus, since $|\exp(i\omega)| = 1$, this means that an LTI system will be BIBO stable *if and only if* its FTF $H(\omega)$ is absolutely summable (finite) within the interval $\omega \in (-\pi, \pi)$, that is

$$|H(\omega)| \leq \left\{ \sum_{n=-\infty}^{\infty} \left|h(n)\exp(i\omega)\right| = \sum_{n=-\infty}^{\infty} |h(n)| \right\} < \infty. \tag{1.114}$$

To summarize, an LTI system is said to be BIBO stable *if any* of the following (equivalent) conditions is fulfilled:

- Its impulse response $\{h(n)\}$ is absolutely summable;
- Its FTF $H(\omega) = \mathcal{F}\{h(n)\}$ is absolutely summable;
- The unit circle of the Z-domain $U(z)$ is part of the ROC of its transfer function $H(z) = \mathcal{Z}\{h(n)\}$.

1.6 Convolution-Based One-Dimensional (1D) Linear Filters

In signal processing, a finite impulse response (FIR) filter is a linear system whose impulse response is of finite duration, because it settles to zero in finite time. In contrast, infinite impulse response (IIR) filters may continue to respond indefinitely because of some internal feedback. In this section we present the Z-transform and FTF analysis of one-dimensional (1D) FIR and IIR filters, which allows us to analyze the filter's stability along with its spectral frequency behavior.

1.6.1 One-Dimensional Finite Impulse Response (FIR) Filters

For an FIR filter, the output is a weighted sum of the current value and a finite number of previous values of the input (Figure 1.27). This operation is described by the following equation:

$$f(t) = b_0 I(t) + b_1 I(t-1) + \cdots + b_N I(t-N) \tag{1.115}$$

$$= \sum_{n=0}^{N} b_n I(t-n) = I(t)\sum_{n=0}^{N} b_n \delta(t-n), \tag{1.116}$$

where N is the filter order, $I(t)$ is the input signal, $f(t)$ is the output signal, and b_n are the filter coefficients that make up the impulse response. Then, the impulse response for an FIR filter is given by

$$h(t) = \sum_{k=0}^{N} b_n \delta(t-n) \tag{1.117}$$

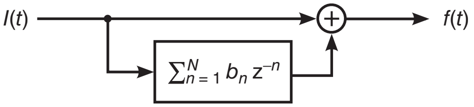

Figure 1.27 General diagram for a finite impulse response (FIR) filter.

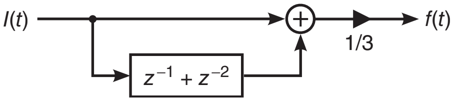

Figure 1.28 Diagram of a three-step averaging system.

and its Z-transform yields the transfer function of the FIR filter:

$$H(z) = Z\{h(t)\} = \sum_{n=-\infty}^{\infty} h(n)z^{-n} = \sum_{n=0}^{N} b_n z^{-n}. \tag{1.118}$$

The impulse response of an Nth-order discrete-time FIR filter lasts for $N+1$ samples (where $b_n \neq 0$), and then settles to zero. This means that *all FIR filters are BIBO stable* since their ROCs include at least $\{z\ :\ 0 < |z| < \infty\}$.

Example: A Three-Step Averaging Filter

Consider the three-step averaging system presented in Figure 1.28, where the output signal $f(t)$ is given by

$$f(t) = \frac{1}{3}[I(t) + I(t-1) + I(t-2)], \tag{1.119}$$

or in terms of its (finite) impulse response

$$\begin{aligned} f(t) &= I(t) * h_3(t), \\ h_3(t) &= \sum_{n=0}^{2}(1/3)\delta(t-n). \end{aligned} \tag{1.120}$$

Thus the Z-transfer function for this FIR filter, $H(z) = F(z)/I(z)$, is

$$H(z) = \mathcal{F}\{h_3(t)\} = \frac{1}{3}(1 + z^{-1} + z^{-2}) = \frac{1}{3z^2}(1 + z + z^2). \tag{1.121}$$

As illustrated in Figure 1.29a, this ROC (given by $0 < |z| \leq \infty$) includes the unit circle so this system is said to be BIBO stable (as expected). Furthermore, its FTF, $H(z = e^{i\omega})$, exists and it is given by

$$H(\omega) = (1/3)[1 + \exp(-i\omega) + \exp(-i2\omega), \tag{1.122}$$

which corresponds to a low-pass filter as illustrated in Figure 1.29b.

Example: A Three-Step Band-Pass (Quadrature) Filter

Consider the three-step averaging system described in Eq. (1.120). According to the so-called frequency translation property of the Fourier transform (Table 1.4), one can “displace” the spectral response of this linear filter by multiplying its

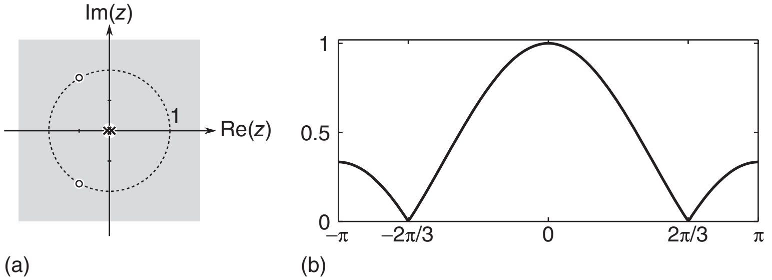

Figure 1.29 (a) Pole-zero plot for the three-step averaging filter discussed in this example and (b) the absolute value of its FTF.

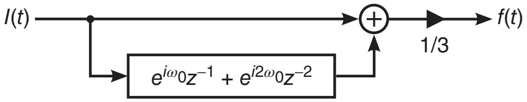

Figure 1.30 Diagram of a three-step band-pass (quadrature) system.

impulse response by a complex sinusoidal signal, thus making it a band-pass filter (Figure 1.30):

$$h(t) = \exp(i\omega_0 t) \sum_{n=0}^{2} (1/3)\delta(t-n),$$
$$H(\omega) = (1/3)\left\{1 + \exp[-i(\omega-\omega_0)] + \exp[-i2(\omega-\omega_0)]\right\}. \tag{1.123}$$

The corresponding pole-zero plot and the FTF plot are presented in Figure 1.31, with $\omega_0 = 2\pi/3$ for illustrative purposes.

From the panel (b), it should be noted that with the proper selection of ω_0 this FTF fulfills the so-called quadrature conditions [30], given by

$$H(\omega_0) \neq 0, \quad H(0) = H(-\omega_0) = 0, \tag{1.124}$$

so we can also say that this is a quadrature filter. This kind of filters play an extremely important role in the analysis of fringe patterns (as we will show in

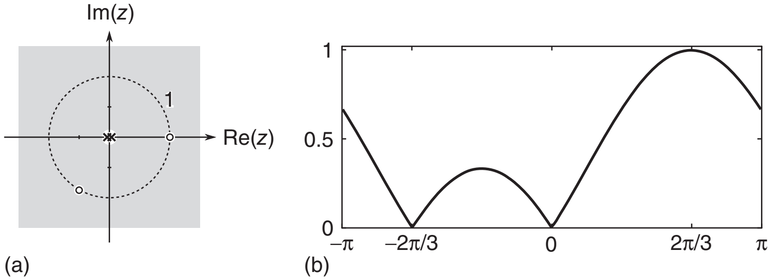

Figure 1.31 (a) Pole-zero plot for the frequency-displaced three-step averaging filter discussed in this example and (b) the absolute value of its FTF.

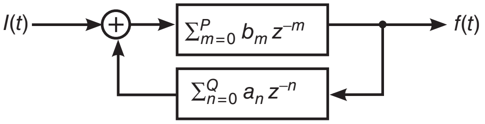

Figure 1.32 General diagram for an infinite impulse response (IIR) filter.

Chapter 2). For now, it is enough to note that for some cases they may be obtained by multiplying the impulse response of a low-pass filter with a complex sinusoidal function.

1.6.2 One-Dimensional Infinite Impulse Response (IIR) Filters

For IIR, filters the output at a given time is given by a weighted sum of the current and a finite number of previous values of both the input and the output, as illustrated in Figure 1.32. IIR filters are often described and implemented in terms of a difference equation

$$\begin{aligned} f(t) = \frac{1}{a_0} \{ & b_0 I(t) + b_1 I(t-1) + \cdots + b_P I(t-P) \\ & -a_1 f(t-1) - a_2 f(t-2) - \cdots - a_Q f(t-Q) \} \end{aligned} \tag{1.125}$$

with P and Q being the feed-forward and the feedback filter order, respectively. A condensed form of this difference equation is given by

$$\sum_{n=0}^{Q} a_n f(t-n) = \sum_{m=0}^{P} b_m I(t-m), \tag{1.126}$$

where a_n and b_m are, respectively, the feedback and feed-forward filter coefficients. To find the transfer function of this filter, first we take the Z-transform of each side of the above equation, where we use the time-shift property to obtain

$$\sum_{n=0}^{Q} a_n z^{-n} F(z) = \sum_{m=0}^{P} b_m z^{-m} I(z). \tag{1.127}$$

Solving for the transfer function results in

$$H(z) = \frac{F(z)}{I(z)} = \frac{\sum_{m=0}^{P} b_m z^{-m}}{\sum_{n=0}^{Q} a_n z^{-n}}. \tag{1.128}$$

In most IIR filter designs, $a_0 = 1$ so the transfer function is often expressed as

$$H(z) = \frac{\sum_{m=0}^{P} b_m z^{-m}}{1 + \sum_{n=1}^{Q} a_n z^{-n}}. \tag{1.129}$$

Clearly, the transfer function of an IIR filter has poles. So, in order to determine whether an IIR filter is BIBO stable, we have to locate these poles and find whether the unit circle $U(z)$ is part of its ROC.

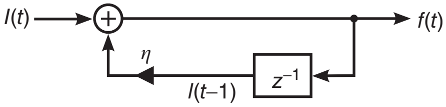

Figure 1.33 Diagram of a first-order recursive linear filter.

Example: A First-Order Recursive Low-Pass Filter

The first-order recursive linear filter illustrated in Figure 1.33 can be described by the following recursive equation:

$$f(t) = \eta f(t-1) + I(t). \tag{1.130}$$

Taking the Z-transform of the above equation, we have

$$F(z) = \eta z^{-1} F(z) + I(z). \tag{1.131}$$

And solving for the transfer function, $H(z) = F(z)/I(z)$, results in

$$H(z) = \frac{1}{1 - \eta z^{-1}} = \frac{z}{z - \eta}. \tag{1.132}$$

As illustrated in Figure 1.34a, the transfer function $H(z)$ contains a simple zero at $z = 0$ and a simple pole at $z = \eta$. So the ROC will contain the unit circle (producing a stable system) *if and only if* $\eta < 1$. Furthermore, its FTF, $H(z = e^{i\omega})$, exists and it is given by

$$H(\omega) = \frac{1}{1 - \eta \exp(-i\omega)}, \tag{1.133}$$

which corresponds to a low-pass filter as illustrated in Figure 1.34b.

Example: A First-Order Recursive Band-Pass Filter

By changing $\eta \Rightarrow \eta \exp(i\omega_0)$ in our previous example (Figure 1.35), it is straightforward to see that the transfer function is now given by

$$H(z) = \frac{z}{z - \eta \exp(i\omega_0)}, \tag{1.134}$$

while the corresponding FTF is given by

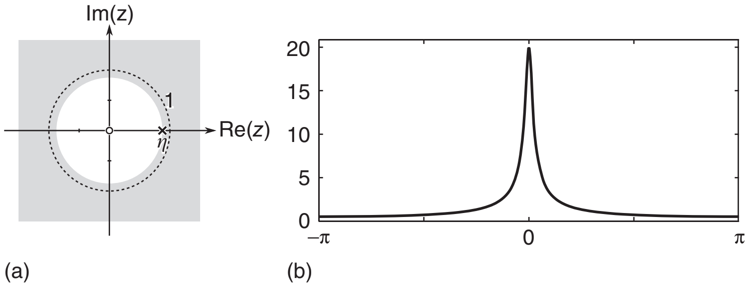

Figure 1.34 (a) ROC for the first-order recursive filter discussed in this example and (b) its FTF. Here $\eta = 0.95$ for illustrative purposes.

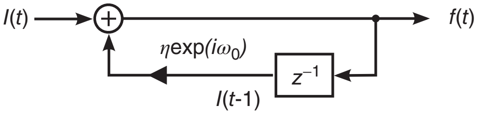

Figure 1.35 Diagram of a quadrature first-order recursive band-pass filter.

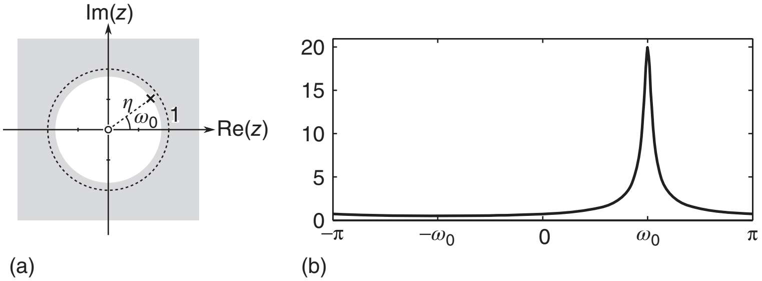

Figure 1.36 (a) ROC for the first-order recursive filter discussed in this example and (b) its FTF.

$$H(\omega) = \frac{1}{1 - \eta \exp[-i(\omega - \omega_0)]}. \tag{1.135}$$

As illustrated in Figure 1.36, this recursive system remains stable for $\eta < 1$ and represents a narrow band-pass filter centered around ω_0.

1.7 Convolution-Based two-dimensional (2D) Linear Filters

The input data in fringe pattern analysis are typically described at every given time by discrete arrays that depend on two independent variables (x and y). Generally speaking, fringe patterns mostly contain a low-frequency signal along with a high-frequency degrading noise (multiplicative or additive). Therefore, low-pass filtering of a fringe pattern may remove a substantial amount of noise, making the demodulation process more reliable.

In this section, we discuss the generalization for two-dimensional (2D) FIR and IIR filters. This analysis should allow us to understand higher dimensional generalizations, especially when considering linearly independent variables (e.g., spatiotemporal digital filters). Luckily, most properties previously discussed in our review of the 1D linear filter theory can be directly generalized to 2D linear filters.

1.7.1 Two-Dimensional (2D) Fourier and Z-Transforms

As in the 1D case, the analysis of 2D linear filters is usually carried out entirely in the frequency domain. Thus, we need to define at least the “direct” formula for the 2D Fourier transform and 2D Z-transform, that is

$$\mathcal{F}\{f(x,y)\} = F(u,v) = \iint_{\mathbb{R}^2} f(x,y)\exp\left[-i(ux+vy)\right] dxdy, \tag{1.136}$$

$$\mathcal{Z}\{f(x,y)\} = F(z_x,z_y) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} f(n,m) z_x^{-n} z_y^{-m}. \tag{1.137}$$

Similarly, the 2D DTFT can be found by evaluating the 2D Z-transform in $z_x = \exp(iu)$ and $z_y = \exp(iv)$. That is,

$$\mathcal{F}\{f(n,m)\} = F(u,v) = \sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty} f(n,m)\exp[-i(nu+mv)], \tag{1.138}$$

where, as before, $f(n,m) = f(x,y)\big|_{(x,\,y)=(n,\,m)}$. Once again, nowadays virtually any processing required is done on digital systems, so in the practice the formulas in Eqs. (1.137–1.138) are the ones we actually implement.

1.7.2
Stability Analysis of 2D Linear Filters

The general form for a 2D digital linear filter of input $I(x,y)$ and output $f(x,y)$ is given by

$$\sum_{n=-N/2}^{N/2}\sum_{m=-M/2}^{M/2} a_{n,m} f(x-n,y-m) = \sum_{n=-N/2}^{N/2}\sum_{m=-M/2}^{M/2} b_{n,m} I(x-n,y-m). \tag{1.139}$$

Taking its Z-transform and solving for the transfer function results in

$$H(z_x,z_y) = \frac{F(z_x,z_y)}{I(z_x,z_y)} = \frac{\sum_{n=-N/2}^{N/2}\sum_{m=-M/2}^{M/2} b_{n,m} z_x^{-n} z_y^{-m}}{\sum_{n=-N/2}^{N/2}\sum_{m=-M/2}^{M/2} a_{n,m} z_x^{-n} z_y^{-m}}. \tag{1.140}$$

The ROC consists of the 2D set of points (z_x, z_y) for which $H(z_x, z_y)$ is absolutely summable, which in turn translates into finding the location of its poles and zeros. A two-dimensional linear system is BIBO stable if its transfer function $H(z_x, z_y)$ has no singularities within the unit bidisc, defined by the set

$$\overline{U}^2 = \{(z_x^{-1}, z_y^{-1}) \;:\; \left|z_x^{-1}\right| \le 1, \left|z_y^{-1}\right| \le 1\}. \tag{1.141}$$

According to Shank's theorem [31], by expressing the transfer function as the (causal) rational function

$$H(z_x,z_y) = \frac{\overline{N}(z_x^{-1},z_y^{-1})}{\overline{D}(z_x^{-1},z_y^{-1})} = \frac{\sum_{j=0}^{N}\sum_{k=0}^{M} b_{jk} z_x^{-j} z_y^{-k}}{\sum_{j=0}^{M}\sum_{k=0}^{M} a_{jk} z_x^{-j} z_y^{-k}}; \quad a_{00} = 1 \tag{1.142}$$

the corresponding 2D linear system will be BIBO stable if $\overline{N}(z_x^{-1}, z_y^{-1})$ and $\overline{D}(z_x^{-1}, z_y^{-1})$ have no common factor, and if

$$\overline{D}(z_x^{-1},z_y^{-1}) \ne 0, \quad \text{for} \quad (z_x^{-1},z_y^{-1}) \in \overline{U}^2. \tag{1.143}$$

However, since the zeros of polynomials of two complex variables are not isolated points, in general there will be an infinite number of singularities and verifying the previous condition may be rather difficult and cumbersome. A more convenient approach for our purposes can be stated as follows (Strintzis' theorem [31]): a 2D digital filter is BIBO stable *if and only if* the following conditions are fulfilled:

- $\overline{D}(1, z_y^{-1}) \neq 0, \quad \text{for } |z_y^{-1}| \leq 1,$
- $\overline{D}(z_x^{-1}, 1) \neq 0, \quad \text{for } |z_x^{-1}| \leq 1,$
- $\overline{D}(z_x^{-1}, z_y^{-1}) \neq 0, \quad \text{for } (z_x^{-1}, z_y^{-1}) \in U^2.$

Here, the unit bicircle U^2 is given by

$$U^2 = \left\{ \left(z_x^{-1},\, z_y^{-1} \right) \,:\, |z_x^{-1}| = 1,\, |z_x^{-1}| = 1 \right\}. \tag{1.144}$$

The first and second conditions translate as locating the poles of 1D digital filters, whereas the third condition means that the FTF must remain bounded: $|H(u, v)| < \infty$. Thus, the stability of 2D filters can be assessed by means of DTFT plots and the 1D filters' theory previously analyzed.

Example: A 3 × 3 Averaging Convolution Filter

The convolution averaging window is by far the most used low-pass filter in fringe analysis. Convolution with an averaging window represents a 2D FIR filter, so we know from our previous analysis that it is always BIBO stable. The discrete impulse response of the convolution averaging window is typically represented by a matrix. For instance, consider the 3 × 3 averaging filter, given by

$$h(x, y) = \frac{1}{9} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}. \tag{1.145}$$

The application of this particular filter can be expressed in functional form as

$$\begin{aligned} f(x, y) = I(x, y) * h(x, y) = I(x, y) * \sum_{m=-1}^{1} \sum_{n=-1}^{1} \frac{1}{9} \delta(x - n, y - m), \\ = \frac{1}{9} \sum_{m=-1}^{1} \sum_{n=-1}^{1} I(x - m, y - n). \end{aligned} \tag{1.146}$$

The 2D frequency response of this 3 × 3 convolution matrix is given by

$$\begin{aligned} H(u, v) = (1/9) \Big[1 + 2\cos u + 2\cos v + 2\cos \sqrt{2}(u + v) \\ +2\cos \sqrt{2}(u - v) \Big]. \end{aligned} \tag{1.147}$$

Clearly, this 2D FTF is bounded for all $(u, v) \in \mathbb{R}^2$, as expected.

Small-size convolution filters may be used several times to decrease the band-pass frequency; this also changes the spectral shape of the filter. In general, the FTF of a sequence of identical low-pass filters will approach a Gaussian-shaped response, as can be seen in Figure 1.37.

To summarize, all convolution-based 2D linear filters have an FIR. For examples of 2D IIR filters, see Section 1.8.2.

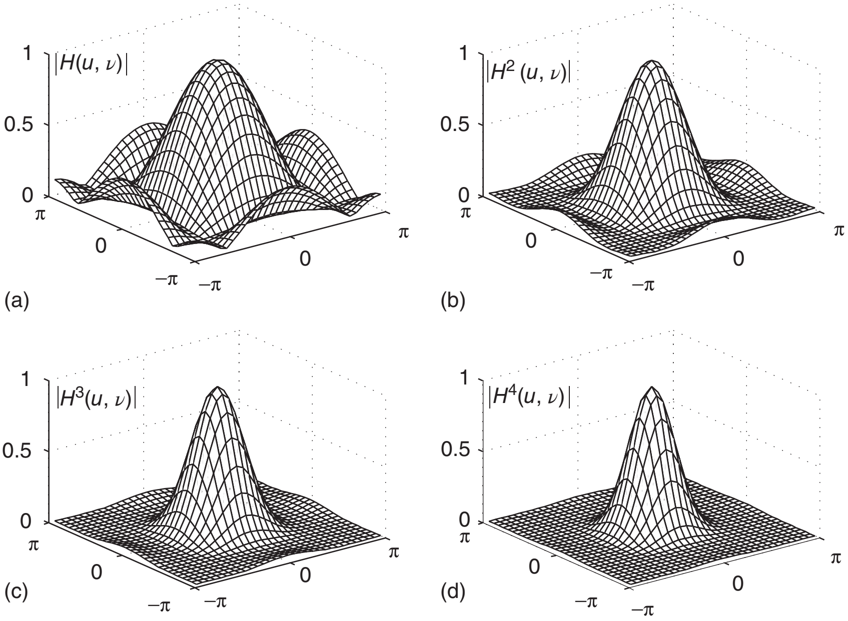

Figure 1.37 (a–d) Frequency transfer function of a 3 × 3 averaging window convoluted with itself n times. Note how the frequency response tends to a Gaussian as the number of convolutions increase.

1.8 Regularized Spatial Linear Filtering Techniques

Convolution-based linear filtering is the most basic operation in digital signal processing but is not always the best option in fringe pattern analysis: interferometric signals are bounded by spatial, temporal, or spatiotemporal pupils (diaphragms, finite sequences of sampled data, etc.), and right at the edge of these bounding pupils, convolution filters mix up well-defined interferometric data with invalid background outside the pupils where no fringe data is available or defined.

The edge distortion may be so significant that some people shrink the interferogram's area during spatial filtering to avoid those unreliable pixels near the edge. While this approach may be practical and easy to implement, it implies discarding valid data. In this section, we will show that classical regularization techniques are viable alternatives for convolution linear filtering, which allow us to cope with bounding pupils reducing the phase distortion near the edges.

1.8.1 Classical Regularization for Low-Pass Filtering

The prototypical example for classical regularization is the low-pass filtering process of a noisy signal. According to Marroquin *et al.* [32], it may be stated as follows: given the observations $I(\cdot)$, find a smooth function $f(\cdot)$ defined on a two-dimensional field L, which may be modeled by

$$I(x, y) = f(x, y) + n(x, y), \quad \forall(x, y) \in S \tag{1.148}$$

where $n(\cdot)$ is a high-frequency noise field (e.g., a white Gaussian noise), and S is the subset of L where observations have a good signal-to-noise ratio.

The low-pass filtering process may be seen as an optimizing inverse problem, in which one strikes a compromise between obtaining a smooth filtered field $f(x, y)$ and keeping a good fidelity to the observed data $I(x, y)$. In the continuous domain, it can be stated as the minimization of the following energy functional:

$$\begin{aligned} U\left[f(x, y)\right] = \iint_{(x,y)\in S} \Bigg\{ & \left[f(x, y) - I(x, y)\right]^2 \\ & + \eta \left[\frac{\partial f(x, y)}{\partial x}\right]^2 + \eta \left[\frac{\partial f(x, y)}{\partial y}\right]^2 \Bigg\} dxdy. \end{aligned} \tag{1.149}$$

On the right-hand side of the above equation, the first term measures the fidelity between the smoothed field $f(x, y)$ and the observed data $I(x, y)$ in a least-squares sense. The second term (the regularizer) penalizes the departure from smoothness of the filtered field $f(x, y)$ by restricting the solution within the space of continuous functions up to the first derivative (the C^1 functional space); this is known as a *first-order membrane regularizer* because it corresponds to the mechanical energy of a 2D membrane, $f(x, y)$, attached by linear springs to the observations $I(x, y)$. The parameter η measures the stiffness of the membrane model; a high stiffness value will lead to a smoother filtered field (this will be demonstrated in the following subsection).

Another widely used energy functional is constructed using a second-order or metallic thin-plate regularizer, which restricts the filtered field $f(x, y)$ within the C^2 functional space (i.e., the space of continuous functions up to the second derivative). In the continuous domain, this energy functional may be stated as

$$\begin{aligned} U\left[f(x, y)\right] = \iint_{(x,y)\in S} \Bigg\{ & \left[f(x, y) - I(x, y)\right]^2 + \eta \left[\frac{\partial^2 f(x, y)}{\partial x^2}\right]^2 \\ & + \eta \left[\frac{\partial^2 f(x, y)}{\partial y^2}\right]^2 + \eta \left[\frac{\partial^2 f(x, y)}{\partial x \partial y}\right]^2 \Bigg\} dxdy. \end{aligned} \tag{1.150}$$

Similar to the first-order regularizer, this energy functional corresponds to a metallic thin plate $f(x, y)$ attached by linear springs to the observations $I(x, y)$, where the parameter η indicates the stiffness of these linear springs. The difference between both optimizing systems is schematically illustrated in Figure 1.38 (showing just a horizontal slice for ease of observation).

In the discrete version of the energy functionals shown before (the ones actually used on a digital computer), the functions $f(x, y)$ and $I(x, y)$ are now defined on the nodes of a regular lattice L and the integrals become sums over the domain of

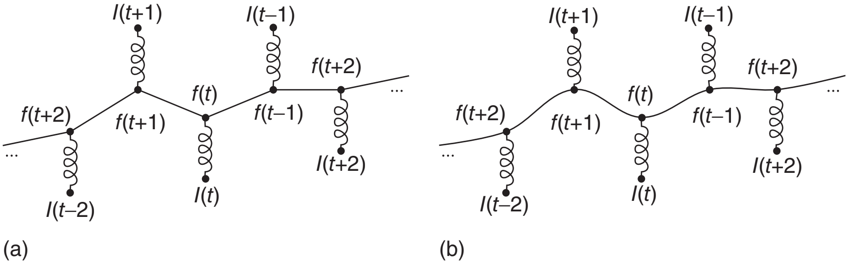

Figure 1.38 Diagram of the estimated fields obtained with (a) first-order membrane and (b) second-order metallic thin-plate regularizers.

interest, that is,

$$U\left[f(x,y)\right] = \sum_{(x,y)\in S} \left\{ \left[f(x,y) - I(x,y)\right]^2 + \eta R\left[f(x,y)\right] \right\} \tag{1.151}$$

where S is the subset of L where observations are available. The discrete version of the first-order regularizer $R_1\left[f(x,y)\right]$ may be approximated by

$$R_1\left[f(x,y)\right] = \left[f(x,y) - f(x-1,y)\right]^2 + \left[f(x,y) - f(x,y-1)\right]^2 \tag{1.152}$$

and the second-order regularizer $R_2\left[f(x,y)\right]$ may be approximated by

$$\begin{aligned} R_2\left[f(x,y)\right] = & \left[f(x+1,y) - 2f(x,y) - f(x-1,y)\right]^2 \\ & + \left[f(x,y+1) - 2f(x,y) - f(x,y-1)\right]^2 \\ & + \Big[[f(x+1,y+1) - f(x-1,y-1) \\ & + f(x-1,y+1) - f(x+1,y-1)\Big]^2. \end{aligned} \tag{1.153}$$

A simple way to optimize the discrete energy functionals stated in this section is by gradient descent

$$\begin{aligned} f^0(x,y) &= I(x,y), \\ f^{k+1}(x,y) &= f^k(x,y) - \mu \frac{\partial U\left[f(x,y)\right]}{\partial f(x,y)}, \end{aligned} \tag{1.154}$$

where k is the iteration number and $\mu \ll 1$ is the step size of the gradient search. However, gradient descent is a slow procedure, especially for high-order regularizers, so one may need to implement more complex but faster approaches (conjugate gradient method, Newton's method, etc.).

Up to this point, we have only established the groundwork of regularized low-pass filtering. Now we will see a practical way of implementing it in a digital computer using an irregularly shaped domain S. Let us define an indicator function $m(x,y)$ in the lattice L having $N \times M$ nodes:

$$m(x,y) = \begin{cases} 1 & \forall (x,y) \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{1.155}$$

Using this indicator field, the filtering problem with a first-order regularizer may be rewritten as

$$U\left[f(x,y)\right] = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\left\{\left[f(x,y) - I(x,y)\right]^2 m(x,y) + \eta R\left[f(x,y)\right]\right\}, \qquad (1.156)$$

where

$$\begin{aligned} R\left[f(x,y)\right] = {} & \left[f(x,y) - f(x-1,y)\right]^2 m(x,y)m(x-1,y) \\ & + \left[f(x,y) - f(x,y-1)\right]^2 m(x,y)m(x,y-1). \end{aligned} \qquad (1.157)$$

Then the derivative may be found as

$$\begin{aligned} \frac{\partial U\left[f(x,y)\right]}{\partial f(x,y)} = {} & \left[f(x,y) - I(x,y)\right]^2 m(x,y) + \eta\left[f(x,y) - f(x-1,y)\right] m(x,y)m(x-1,y) \\ & + \eta\left[f(x+1,y) - f(x,y)\right] m(x,y)m(x+1,y) \\ & + \eta\left[f(x,y) - f(x,y-1)\right] m(x,y)m(x,y-1) \\ & + \eta\left[f(x,y) - f(x,y+1)\right] m(x,y)m(x,y+1). \end{aligned} \qquad (1.158)$$

Note that only the difference terms lying completely within the region of valid fringe data marked by $m(x,y)$ survive. In other words, the indicator field $m(x,y)$ is the function that actually decouples valid fringe data from its surrounding background. A numerical comparison of this regularizing low-pass filtering approach versus traditional convolution-based low-pass filtering is shown in Figure 1.39.

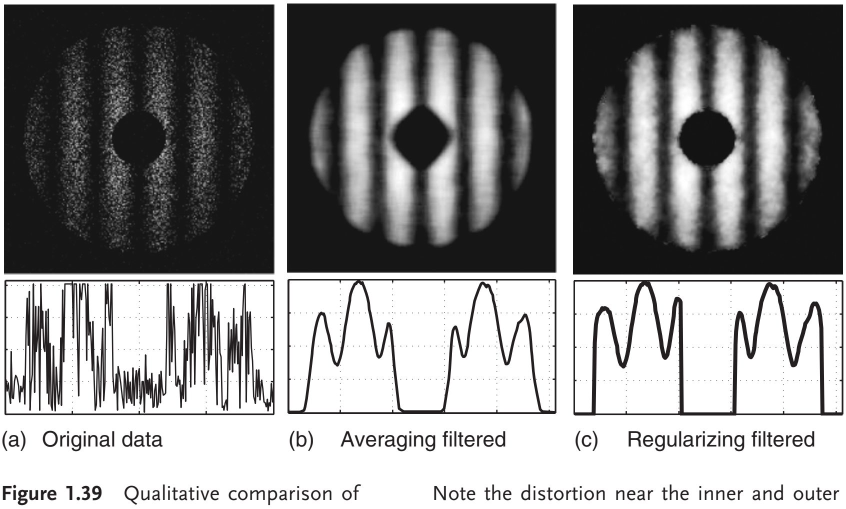
(a) Original data (b) Averaging filtered (c) Regularizing filtered

Figure 1.39 Qualitative comparison of convolution-based low-pass filtering versus the proposed regularizing low-pass filtering. In panel (a), we have a noisy fringe pattern bounded by two circular pupils. In panel (b), we have the smoothed field as obtained with convolution based low-pass filtering. Note the distortion near the inner and outer boundaries due to the mixing with the surrounding background. In panel (c), we have the estimated field obtained with first-order regularizing low-pass filtering. Here, the fringe data was properly decoupled from the background.

Extrapolation and/or interpolation. Classical regularization techniques also allow us to extrapolate and/or interpolate data in a well-defined way, simply by defining two different indicator functions: one for the region with valid data, $m_1(x, y)$; and other for the region where the estimated field $f(x, y)$ will be extrapolated and/or interpolated, $m_2(x, y)$. For instance, using the first-order regularizer, we have

$$\begin{aligned} U\left[f(x,y)\right] = \sum_{x=0}^{N-1}\sum_{y=0}^{M-1}\Big\{ & \left[f(x,y) - I(x,y)\right]^2 m_1(x,y) \\ & + \eta\left[f(x,y) - f(x-1,y)\right]^2 m_2(x,y)m_2(x-1,y) \\ & + \eta\left[f(x,y) - f(x,y-1)\right]^2 m_2(x,y)m_2(x,y-1)\Big\}. \end{aligned} \tag{1.159}$$

It is important to remark that $m_2(x, y)$ must be a superset of $m_1(x, y)$; in other words, $m_2(x, y) = 1$ for *at least* all the regions where $m_1(x, y) = 1$.

As before, the first term in the above equation measures the fidelity between the input data $I(x, y)$ and the estimated field $f(x, y)$ in the least-squares sense; this is done only in the region with valid data, where $m_1(x, y) = 1$. The remaining terms restrict the estimated field $f(x, y)$ within the C^1 functional space for all the points where $m_2(x, y) = 1$; this includes certain regions where the input data is undefined. The extrapolation and/or interpolation takes place because of the regularizing restrictions: for this first-order (rubber membrane) regularizer, $f(x, y)$ for $\{(x, y) : m_1(x, y) = 0, m_2(x, y) = 1\}$ is estimated enforcing the continuity of this 2D field with $f(x \pm 1, y \pm 1)$ for $\{(x, y) : m_1(x, y) = 1, m_2(x, y) = 1\}$. Similarly, if we apply a second-order (metallic thin-plate) regularizer, $f(x, y)$ would be estimated by preserving the curvature of the 2D field. Finally, if we set a very low value for the stiffness parameter $\eta \ll 1$, we may extrapolate and/or interpolate the input data without any noticeable low-pass filtering.

1.8.2 Spectral Response of 2D Regularized Low-Pass Filters

From the above discussion, we know that the 2D field $f(x, y)$ that minimizes the energy functionals seen in the previous section smooths out the input data $I(x, y)$. But in order to have a quantitative idea of the amount of smoothing, we need to find the frequency response of these regularizing low-pass filters [32–34]. For the first-order regularizer (from Eqs. 1.151 and 1.152), considering an infinite 2D lattice and setting the gradient to zero, we have

$$\begin{aligned} & f(x,y) - I(x,y) + \eta\left[-f(x-1,y) + 2f(x,y) - f(x-1,y)\right] \\ & \quad + \eta\left[-f(x,y-1) + 2f(x,y) - f(x,y-1)\right] = 0. \end{aligned} \tag{1.160}$$

Taking its Z-transform and solving for the transfer function results in

$$H_1(z_x, z_y) = \frac{F(z_x, z_y)}{I(z_x, z_y)} = \frac{1}{1 + \eta(4 - z_x^{-1} - z_x - z_y^{-1} - z_y)}, \tag{1.161}$$

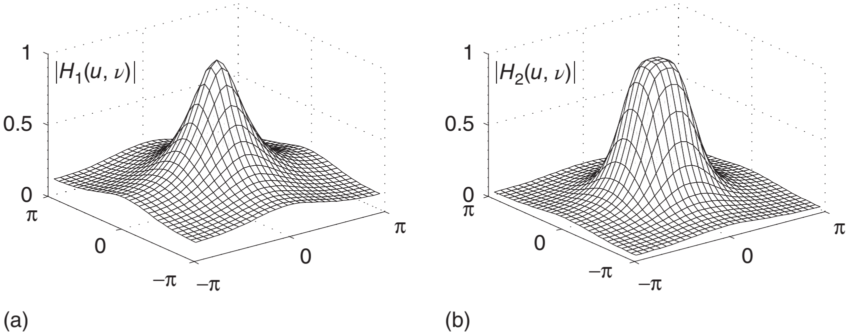

Figure 1.40 Frequency transfer function $|H(u, \nu)|$ for (a) the first-order regularizer and (b) the second-order regularizer. In both cases, $\eta = 5$.

with an ROC given by

$$\text{ROC} = \{(z_x, z_y) \ : \ |z_x| < \infty, |z_y| < \infty\}, \quad \text{for} \quad \eta > 0. \tag{1.162}$$

Since the ROC clearly includes the biunit circle (Eq. 1.144), this 2D IIR filter is found to be BIBO stable and its FTF is obtained by substituting $z_x = \exp(iu)$, and $z_y = \exp(iv)$:

$$H_1(u, v) = \frac{F(u, v)}{I(u, v)} = \frac{1}{1 + 2\eta(2 - \cos u - \cos v)}. \tag{1.163}$$

Similarly, for the second-order regularizer (from Eqs. 1.151 and 1.153), we obtain the following FTF:

$$H_2(u, v) = \frac{1}{1 + 2\eta[8 - 6(\cos u + \cos v) + \cos 2u + \cos 2v + 2\cos u \cos v]}. \tag{1.164}$$

As illustrated in Figures 1.40 and 1.41, these FTFs behave somewhat like 2D Lorentzian functions, where the bandwidth of these low-pass filters is controlled by the parameter η.

To summarize, regularization filters may be considered to be a more robust approach than convolution filters in the following sense:

- They prevent the mixing of valid fringe data with the background signal (with this, the distorting effect at the boundary is minimized). This is especially important when dealing with irregular-shaped regions and finite samples sequences;
- They tolerate missing observations because of the capacity of these filters to extrapolate and/or interpolate over regions of missing data with a well-defined behavior. This behavior is controlled by the order of the regularization term.

Furthermore, one may obtain many different types of filters by modifying the potentials in the cost function (as will be shown in Chapter 4). For instance, if $I(x, y)$ represents an interferogram phase-modulated with a generalized carrier $c(x, y)$, one may low-pass filter the synchronous product $I(x, y) \exp[ic(x, y)]$ following the classical regularized approach discussed in this section to produce a regularized quadrature band-pass filter [32–35].

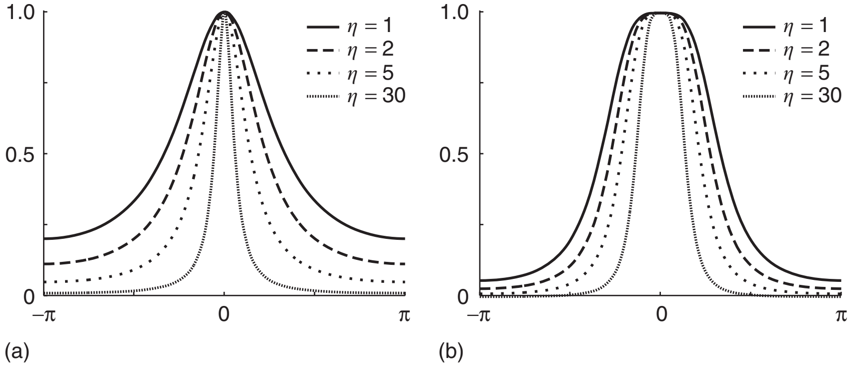

Figure 1.41 Horizontal slice (along the axis $\nu = 0$) of the frequency transfer function $|H(u, \nu)|$ for the first-order regularizer (a) and the second-order regularizer (b), using several values for the parameter η.

1.9 Stochastic Processes

In this section we present a brief review of the theory of stochastic processes. This will allow us to consider more realistic models of our signals under study in the following chapters; it will also establish the basis for a better assessment of many algorithms commonly used in fringe pattern analysis. For a thorough review of this topic, we recommend the books by Artés-Rodríguez *et al.* [23], B. P. Lathi [36], and Papoulis and Unnikrishna Pillai [37].

1.9.1 Definitions and Basic Concepts

A stochastic process is an indexed collection of random variables where the index is conventionally associated with the time [36]; basically, it is a process in which the outcome at any given time instance is given by a random variable.

A continuous random variable X can be characterized by its PDF (probablity density function), given by a nonnegative function f_X that describes the relative likelihood for X to take on a given value, and its statistical averages, given by

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x) f_X(x) dx. \tag{1.165}$$

Note that in this section x represents all possible values that can take the random variable X; it does not mean spatial dependency. Some statistical averages of special interest are the mean (or expected value) μ_X, and the variance σ^2, given by

$$\mu_X = E\{X\}, \tag{1.166}$$

$$\sigma^2 = E\{X^2\} - E^2\{X\}. \tag{1.167}$$

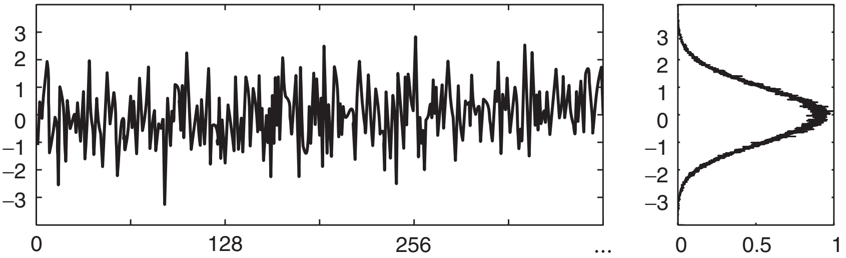

Figure 1.42 Computer-simulated realization of a white stochastic process with normal distribution.

The most commonly observed PDF is the Gaussian one, given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_X)^2}{2\sigma^2}\right], \tag{1.168}$$

where the mean matches with the parameter μ_X and its variance is given by σ^2. The Gaussian distribution is a valid model for many random processes observed in nature; particularly, it is a good model for the electronic noise [23]. In Figure 1.42, we present a numerical simulation of a very large sequence of random values and their (normalized) frequency distribution. As can be seen, this sequence of random values clearly follows a Gaussian distribution.

The abundance of the Gaussian distribution in nature may be explained as a consequence of the central limit theorem [36]. In its simple form, this theorem states that, given a set of independent random variables $\{X_1, X_2, \ldots, X_N\}$ with mean μ and variance $\sigma > 0$, then the sequence of random variables

$$Y_n = \frac{\Sigma_{k=1}^{n}(X_k - \mu)}{\sigma\sqrt{n}} \tag{1.169}$$

converges in distribution to a normal random variable (i.e., Gaussian with $\mu_Y = 0$, $\sigma_Y^2 = 1$).

Sum of two random processes. The PDF of the sum of two independent random variables X and Y is obtained as the convolution of their PDFs [37], that is

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x) f_Y(x-y)dy = f_X(x) * f_Y(x). \tag{1.170}$$

Ensemble averages for a stochastic process. For a continuous-time process, we say that the variable $X(t)$ describes a stochastic process if $X(t)$ is a random variable for all $t \in \mathbb{R}$ [23]. It should be noted that, for a time-dependent process (where $X(t)$ can take continuous or discrete values), its ensemble averages are also time dependent, that is

$$E\{g[X(t)]\} = \int_{-\infty}^{\infty} g(x) f_X(x,t)dx. \tag{1.171}$$

To characterize a stochastic process, it is insufficient to characterize each of its random variables; one also needs to characterize their statistical dependence, that is, their ensemble averages. In digital signal processing, we are interested in those process that can be characterized by their mean and autocorrelation functions.

The mean of a stochastic process $X(t)$ is defined by

$$\mu_X(t) = E\{X(t)\} = \int_{-\infty}^{\infty} x f_X(x,t)dx. \tag{1.172}$$

And the autocorrelation function, which gives a description of how rapidly the random process $X(t)$ is changing with time, is defined by

$$R_X(t_1,t_2) = E\{X(t_1)X^*(t_2)\}. \tag{1.173}$$

Stationary random processes. A stochastic process is said to be stationary if its statistical properties do not change with time. However, all processes are nonstationary since they must begin at some finite time and must terminate at some finite time (just like a purely sinusoidal signal does not exist in real life). A stochastic process is said to be wide-sense (or weakly) stationary if its mean value and autocorrelation functions are independent of a finite shift in the time origin [36], that is, if

$$E\{X(t)\} = E\{X(t+\tau)\}, \tag{1.174}$$

$$R_X(t_1,t_2) = R_X(t_1+\tau, t_2+\tau). \tag{1.175}$$

For such processes, where the mean is constant and the autocorrelation depends only in the time difference, the autocorrelation is simply denoted as

$$R_X(\tau) = R_X(t+\tau, t). \tag{1.176}$$

White process. A particular case of stationary stochastic processes is that in which the autocorrelation of the samples at two different instants is zero. This way, a stochastic stationary process $X(t)$ is said to be white if its autocorrelation function takes the form

$$R_X(\tau) = c_0\delta(\tau) \tag{1.177}$$

with c_0 being a constant. The most common type of noise in in digital communications is the thermal noise [23] with power density $c_0 = \eta/2$, which can be characterized as a white stochastic process with normal distribution [37].

Deterministic process. A deterministic signal may be considered a degenerated stochastic process in which its realizations always take the same values. For instance, consider a process $X(t)$ that takes with unitary probability the value

$$X(t) = g(t), \forall\, t \in \mathbb{R}. \tag{1.178}$$

This process has a PDF given by

$$f_X(x,t) = \delta\left[x - g(t)\right], \tag{1.179}$$

from where it is straightforward to calculate its ensemble mean and autocorrelation function as [37]

$$\begin{aligned}\mu(t) &= g(t),\\ R_X(t_1, t_2) &= g(t_1)g(t_2).\end{aligned} \tag{1.180}$$

1.9.2 Ergodic Stochastic Processes

A stochastic process is said to be ergodic if its ensemble averages are equal to the (long-enough) temporal averages of any sample function [37], that is, if

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} g[X(t)]dt = \int_{-\infty}^{\infty} g(x)f_X(x,t)dx = E\left\{g[X(t)]\right\}. \tag{1.181}$$

Particularly, a stochastic process $X(t)$ is said to be ergodic in its mean if

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t)dt = \mu_X, \tag{1.182}$$

and is said to be ergodic in its autocorrelation function if

$$\lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} X(t+\tau)X^*(t)dt = R_X(\tau). \tag{1.183}$$

The ergodicity notion is extremely important since in practice we do not have infinitely many sample functions available to compute ensemble averages. But if the process is known to be ergodic, then we only need long-enough realizations. As illustrated in Figure 1.43, ergodicity of a stochastic process is an even more restrictive property than the stationary property (which, as we mentioned before, is already difficult to prove analytically). Luckily for us, the stochastic processes observed in fringe pattern analysis (usually additive distorting noise) are found to be stationary and ergodic.

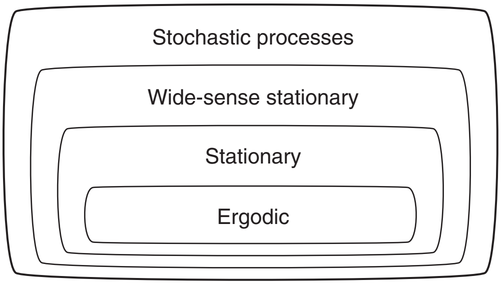

Figure 1.43 Venn diagram of the stochastic processes classification.

1.9.3 LTI System Response to Stochastic Signals

For an LTI system with impulse response function given by $h(t)$, the output $Y(t)$ for a random input signal $X(t)$ is given by Artés-Rodríguez *et al.* [23]

$$Y(t) = h(t) * X(t). \tag{1.184}$$

The linearity property of the convolution operation allows us to easily calculate some ensemble averages of the output signal $Y(t)$. For instance, the expectation value $\mu_Y(t) = E\{Y(t)\}$ is given by

$$\begin{aligned} E\{Y(t)\} &= E\{h(t) * X(t)\} \\ &= h(t) * E\{X(t)\} = h(t) * \mu_X(t), \end{aligned} \tag{1.185}$$

where the property $E\{h(t)\} = h(t)$ was applied since $h(t)$ represents a deterministic signal. The autocorrelation of the output signal $Y(t)$ can also be determined from the autocorrelation function of the input $X(t)$ and the system input response $h(t)$. In general, $R_Y(t_1, t_2)$ is given by Artés-Rodríguez *et al.* [23]

$$R_Y(t_1, t_2) = [h(t_1) * R_X(t_1, t_2)] * h^*(t_2), \tag{1.186}$$

where $h^*(t)$ stands for the complex conjugate of $h(t)$. Assuming that the input signal $X(t)$ represents a stochastic process, Eq. (1.186) is reduced to [23]

$$R_Y(\tau) = R_X(\tau) * [h(\tau) * h^*(-\tau)]. \tag{1.187}$$

These equations show that the ensemble averages of the output depend exclusively on the input response function of the linear system and the ensemble averages of the input.

1.9.4 Power Spectral Density (PSD) of a Stochastic Signal

To translate stochastic processes to the Fourier domain, there are at least two major difficulties: a (stationary) stochastic process is not absolutely integrable, so strictly speaking its Fourier transform does not exist; and, although the spectral representation of a truncated realization does exist, in general it varies between successive samples [36].

When working with (stationary) stochastic processes, we actually deal with gated realizations since it is impossible to observe any processes for infinite periods of time. Considering a stochastic process $X(t)$, we can define its gated realization by

$$X_T(t) = X(t)\text{II}(t/T) = \begin{cases} X(t) & \text{for } |t| \le T/2, \\ 0 & \text{for } |t| > T/2 \end{cases} \tag{1.188}$$

where T is the observation period. Now, since $X_T(t)$ is absolutely integrable, we can calculate its Fourier transform as

$$X_T(\omega) = \int_{-T/2}^{T/2} x(t)\exp(-i\omega t)dt, \tag{1.189}$$

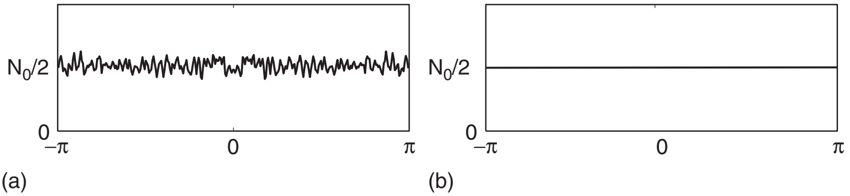

Figure 1.44 Power spectral density of a computer-simulated 1024-sample realization of a white stochastic process (a) and the ideally expected result (b) for reference.

which it is also a stochastic process. Assuming this process to be ergodic, we can compute the ensemble average of the PSDs of all the sample functions to obtain its PSD, given by

$$S_X(\omega) = \lim_{T\to\infty} E\left\{\frac{1}{T}|X_T(\omega)|^2\right\} = \int_{-\infty}^{\infty} R_X(\tau)\exp(-i\omega\tau)d\tau. \tag{1.190}$$

PSD of white noise. Applying the relations stated in Eq. (1.190) for a white noise stochastic process $N(t)$ with autocorrelation function given by $R_N(t) = (N_0/2)\delta(t)$, its PSD is straightforward to find:

$$S_N(\omega) = \mathcal{F}\left\{\frac{N_0}{2}\delta(t)\right\} = \frac{N_0}{2}. \tag{1.191}$$

That is, the white noise has a uniform PSD as illustrated in Figure 1.44. This result is very important for our purposes because the most commonly observed corrupting noise in fringe pattern analysis is modeled by a white stochastic process, namely the additive white Gaussian noise (AWGN).

PSD of a linear system output. The other case of major interest in fringe pattern analysis (and digital signal processing in general) is to find the PSD of a stochastic process at the output of a linear filter [37].

From Eq. (1.187), by applying the convolution property, we have

$$\mathcal{F}\{R_Y(\tau)\} = \mathcal{F}\{R_X(\tau)\}[H(\omega)H^*(\omega)], \tag{1.192}$$

and by applying Eq. (1.190) results in

$$S_Y(\omega) = S_X(\omega)\,|H(\omega)|^2\,. \tag{1.193}$$

That is, the PSD of the output is given by the PSD of the input times the square of the filter's FTF. For instance, in Figure 1.45 we show the PSD for the white noise (previously shown in Figure 1.44) after being filtered by the well-known three-step averaging filter. With this ends our brief review on the theory of stochastic processes.

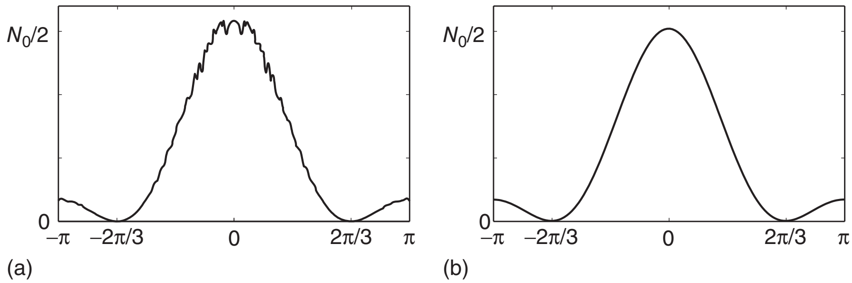

Figure 1.45 Power spectral density of a computer-simulated 1024-sample realization for a white stochastic process after being filtered with a three-step averaging system (a) and the ideally expected result (b) for reference.

1.10 Summary and Conclusions

In this chapter, we reviewed the main theoretical results in digital linear systems theory that are used in the rest of the book. In particular we discussed the following:

- We showed that the general problem of phase estimation from a single interferogram image is, in general, an ill-posed inverse problem. That is because infinitely many estimated phases may be compatible with the observed interferogram data (Figure 1.1).
- We introduced the field of digital phase demodulation process through some motivational examples (Figures 1.2–1.6).
- We gave a general schematic classification of the various strategies that one may follow to actively modify the interferogram by introducing high frequency spatial and/or temporal carriers, as also an overview of the main techniques used to demodulate the measured phase in optical metrology (Figure 1.7).
- We then introduced the main signal classification schemes used in this book, including continuous and discrete, complex and real, and deterministic and random signals, among others. Then we introduced the main space sets of functions used in the mathematical theory of digital signal processing, such as the Dirac delta function and its wide use in digital signal processing. We continued by introducing the concepts of the spectra and other characteristics and limitations of the sampling process of a continuous signal.
- We then proceeded to study digital LTI systems along with their impulse response functions. Afterward, some standard stability criteria applied to LTI systems were discussed, such as the ROC and BIBO criteria.
- We discussed the DTFT and the Z-transform of sampled temporal signals and LTI systems, highlighting their intrinsic relationship ($z = \exp(i\omega)$). These results were generalized to two dimensions (2D) signals and LTI systems.
- The regularizing linear filtering paradigm was then introduced along with their spectral response. Two standard linear regularizers were introduced: the membrane and the thin-plate ones. Also we discussed how these regularized

filters decouple in an optimal way the interferometric data just inside the interferogram fringes from the undefined data outside it.

- Finally, in Section 1.9, the basic theory of stochastic process was reviewed and applied to the analysis of noise in LTI systems. In this section, we also discussed the concepts of stationarity and ergodicity of stochastic processes. We also introduced the autocorrelation of a stochastic process and its Fourier transform, which is the PSD of the stochastic process.