Article

A Concise Tutorial on Functional Analysis for Applications to Signal Processing

Najah F. Ghalyan [1,2,*], Asok Ray [1,3,*] and William Kenneth Jenkins [4,*]

1 Department of Mechanical Engineering, Pennsylvania State University, University Park, PA 16802, USA
2 Department of Mechanical Engineering, University of Kerbala, Kerbala 56001, Iraq
3 Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA
4 Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802, USA
* Correspondence: nfjasim76@gmail.com (N.F.G.); axr2@psu.edu (A.R.); jenkins@engr.psu.edu (W.K.J.)

Abstract: Functional analysis is a well-developed field in the discipline of Mathematics, which provides unifying frameworks for solving many problems in applied sciences and engineering. In particular, several important topics (e.g., spectrum estimation, linear prediction, and wavelet analysis) in signal processing had been initiated and developed through collaborative efforts of engineers and mathematicians who used results from Hilbert spaces, Hardy spaces, weak topology, and other topics of functional analysis to establish essential analytical structures for many subfields in signal processing. This paper presents a concise tutorial for understanding the theoretical concepts of the essential elements in functional analysis, which form a mathematical framework and backbone for central topics in signal processing, specifically statistical and adaptive signal processing. The applications of these concepts for formulating and analyzing signal processing problems may often be difficult for researchers in applied sciences and engineering, who are not adequately familiar with the terminology and concepts of functional analysis. Moreover, these concepts are not often explained in sufficient details in the signal processing literature; on the other hand, they are well-studied in textbooks on functional analysis, yet without emphasizing the perspectives of signal processing applications. Therefore, the process of assimilating the ensemble of pertinent information on functional analysis and explaining their relevance to signal processing applications should have significant importance and utility to the professional communities of applied sciences and engineering. The information, presented in this paper, is intended to provide an adequate mathematical background with a unifying concept for apparently diverse topics in signal processing. The main objectives of this paper from the above perspectives are summarized below: (1) Assimilation of the essential information from different sources of functional analysis literature, which are relevant to developing the theory and applications of signal processing. (2) Description of the underlying concepts in a way that is accessible to non-specialists in functional analysis (e.g., those with bachelor-level or first-year graduate-level training in signal processing and mathematics). (3) Signal-processing-based interpretation of functional-analytic concepts and their concise presentation in a tutorial format.

Keywords: statistical signal processing; adaptive signal processing; functional analysis

1. Introduction

The concept of functional analysis is built upon normed vector spaces and particularly inner product spaces, which are merged with diverse notions of topology and geometry, linear algebra, probability theory, and real and complex analysis (see, e.g., [1–5]). Topics in functional analysis include various concepts such as Banach spaces and Hilbert spaces, and linear operators and their spectral theory, as well as group and semigroup theory. Knowledge of these mathematical structures is often essential for understanding and solving a variety of analytical problems in signal processing and related fields, as well as in mathematics itself [5]. For example, in functional analysis, objects like functions are

considered as elements or points in a space of functions [6], and hence the name *functional analysis*.

Results, generated from functional analysis, form key concepts in the frameworks of advanced scientific and engineering disciplines that include the fields of statistical signal processing and adaptive signal processing. Although adaptive signal processing can be viewed as a branch of statistical signal processing [7], the special properties of this field and their roles in engineering applications have led many specialists to consider them as two separate fields. Therefore, in many universities and research institutions around the world, statistical signal processing and adaptive signal processing are taught as independent graduate courses in engineering and applied sciences, and many textbooks have been devoted to study these important fields individually (e.g., see [8,9] and references therein). Nevertheless both statistical signal processing and adaptive signal processing form the backbone of the so-called *modern signal processing*, in which signals are generally considered as random processes. Modern signal processing covers many topics of current interest, such as signal modeling and estimation, signal prediction, signal compression, adaptive lattice filtering, adaptive joint process estimation, recursive least squares lattice filtering, and spectrum estimation. The issues, related to processing of both deterministic and random signals, are further discussed below.

While an estimation error may typically converge to zero for deterministic signals, this is generally not the case for random signals [8]. Therefore, in statistical and adaptive signal processing, it is a common practice for random signals to make them unbiased (i.e., expectation of the estimation error converging to zero). As explained later, this type of convergence is of a special kind, which is known in functional analysis as *weak convergence* (see, for example, [10,11]). Therefore, many important results in functional analysis are obtained in terms of *weak convergence* and *weak topology*, which potentially have significant implications to the subfields of statistical signal processing and adaptive signal processing. Moreover, it is usually desirable in estimation theory to identify *optimal* filters, which bridges the discipline of signal processing to that of optimization theory. To this end, researchers in modern signal processing often deal with random processes for which optimization problems become more challenging, and the usage of advanced mathematical tools is justified.

From a historical perspective, the names of some of the spaces used in functional analysis are those of early-time mathematicians who had originally developed the theories of these spaces. Indeed much of the theoretical work has been associated with the names of eminent mathematicians (e.g., Gauss, Lagrange, Euler, and Kolmogorov). In fact, the Hilbert space, which is a central topic in functional analysis, is one of the most commonly used mathematical frameworks of signal processing and the associated optimization [12]. The unique features of Hilbert spaces are explained in the paper from these perspectives. However, the names of other well-known spaces (e.g., metric spaces and normed spaces) were given based on the technical properties of these spaces; many of the spaces, frequently used in functional analysis, have been named based on quite different historical backgrounds.

We have presented a concise and focused review of key concepts of functional analysis in this paper, which have strong relevance to modern signal processing. The most important spaces from the perspectives of functional analysis, considered in this paper, are metric/topological spaces, Banach spaces, and Hilbert spaces. The relations among these and other spaces are illustrated in Figure 1. Other relevant vector spaces like summable (ℓ^p), Lebesgue-integrable (L^p), and Hardy ($\mathcal{H}^p$) spaces are also introduced in the paper.
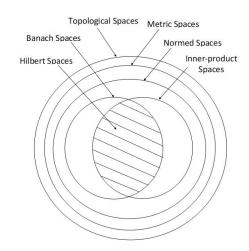
Figure 1. Relationship among different spaces in functional analysis.

The paper is organized in four sections, including the current section, and an Appendix A. Section 2 introduces Banach spaces and their relevant theorems, where special emphases are laid on the ℓ^p / L^p spaces, $\mathcal{H}^p$ spaces, spectral factorization, and weak topology in the setting of Banach spaces. Section 3 presents Hilbert spaces and their relevant features (e.g., Fourier series expansion and the orthogonality principle) along with some applications to signal processing and detection theory, such as wavelets, Karhunen-Loéve (KL) expansion, and reproducing-kernel Hilbert spaces (RKHS). Section 4 summarizes and concludes the paper. The Appendix A in this paper introduces elementary concepts and definitions in real analysis, probability theory, and topological spaces, which should be helpful for understanding the fundamental principles of functional analysis as applied to various concepts of signal processing; however the readers, who are familiar with these concepts, may only selectively refer to the Appendix A.

2. Banach Spaces for Signal Analysis

This section deals with Banach spaces for general applications to signal processing; it also introduces the concepts of Hardy spaces especially for digital signal processing. Further details on Banach spaces are provided in standard books on functional analysis such as Bachman and Narici [1] and Naylor and Sell [2].

2.1. Introduction to Banach Spaces

We start this subsection with the definition of a Banach space which is a complete normed space as defined below.

Definition 1 (Banach Spaces). *Let a vector space X be defined over a field* $\mathbb{K}$, *where examples of* $\mathbb{K}$ *are the field of real numbers* $(\mathbb{R}, +, \cdot)$ *and the field of complex numbers* $(\mathbb{C}, +, \cdot)$. *Let a function* $\| \bullet \| : X \to \mathbb{R}$, *called norm and denoted as* $x \mapsto \|x\|$, *have the following properties:*

- *(positivity)* $\forall x \in X$, $\|x\| \geq 0$, $\forall x \in X$, *and* $\|x\| = 0$ *if and only if* $x = 0$.
- *(homogeneity)* $\forall x \in X$ *and* $\forall c \in \mathbb{K}$, $\|cx\| = |c|\|x\|$.
- *(triangular inequality)* $\forall x, y \in X$, $\|x + y\| \leq \|x\| + \|y\|$.

Then, $(X, \| \bullet \|)$ *is called a normed vector space, where the norm* $\| \bullet \|$ *serves as a metric. A real (resp. complex) normed linear space that is complete (i.e., where every Cauchy sequence converges in the space) is called a real (resp. complex) Banach space.*

Example 1. *The spaces of ℓ^p sequences, $1 \leq p \leq \infty$, form an important class of Banach spaces, which are extensively used in digital signal processing. These are linear vector spaces of all real (resp. complex) sequences $x \triangleq \{x_n\}$ such that $\sum_{n=-\infty}^{\infty} |x_n|^p < \infty$, where the ℓ^p-norm is defined as:*

$$\|x\|_{\ell^p} \triangleq \left(\sum_{n=-\infty}^{\infty} |x_n|^p \right)^{1/p} \text{ if } 1 \leq p < \infty$$

$$\|x\|_{\ell^\infty} \triangleq \sup_{n \in \mathbb{Z}} |x_n| \text{ if } p = \infty$$

Some of the theorems on ℓ^p spaces [2], which are extensively used in the analyses of discrete-time signals, are presented below.

Theorem 1 (Hölder Inequality [2]). *Let $1 < p < \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$. If $f \in \ell^p$ and $g \in \ell^q$, then $f \cdot g \in \ell^1$ and $\|f \cdot g\|_{\ell^1} \leq \|f\|_{\ell^p} \|g\|_{l^q}$.*

Proof. See pp. 550–551 in Naylor and Sell [2]. □

It is noted that Hölder Inequality also holds for $p = 1$ and $p = \infty$.

Theorem 2 (Minkowski inequality [2]). *If $1 \leq p \leq \infty$ and $f, g \in \ell^p$, then $\|f + g\|_{\ell^p} \leq \|f\|_{\ell^p} + \|g\|_{\ell^p}$.*

Proof. See pp. 550–551 in Naylor and Sell [2]. □

It is noted that Lebesgue-integrable versions of ℓ^p spaces, for applications to continuous signal processing, are called L^p spaces [4]. In L^p spaces, Hölder inequality and Minkowski inequality are similar to their respective ℓ^p-versions in Theorems 1 and 2.

Next we focus on a few systems-theoretic applications of Banach spaces, which would require the operation of convolution.

Theorem 3 (Convolution inequality [13]). *For the sequences $\mathbf{u} \in \ell^p$ for $p \in [1, \infty]$ and $\mathbf{h} \in \ell^1$, the convolution product $\mathbf{h} * \mathbf{u} \in \ell^p$ and $\|\mathbf{h} * \mathbf{u}\|_{\ell^p} \leq \|\mathbf{h}\|_{\ell^1} \|\mathbf{u}\|_{\ell^p}$.*

Proof. See p. 241 in Desoer and Vidyasagar [13]. □

Lemma 1 (Barbalat Lemma). *If $\{x_n\} \in l^p$ for some $p \in [1, \infty)$, then $\lim_{n \to \infty} |x_n| = 0$.*

Proof. Let us assume that $\lim_{n \to \infty} |x_n| \neq 0$ is true. Then, there exists a subsequence $\{|x_{n_j}|\}$ bounded below by a real number $\varepsilon > 0$, which implies that $\{|x_{n_j}|^p\}$ is bounded below by ε^p so that $\sum_n |x_n|^p \to \infty$ as $n \to \infty$. This contradicts the assertion $\{x_n\} \in \ell^p$. □

Let a linear discrete-time dynamical system with an impulse response $\mathbf{h}[n, k]$ be excited by an input signal $\mathbf{u}$ to yield an output signal $\mathbf{y}$.

Definition 2 (BIBO-stability). *A system is said to be bounded-input-bounded-output (BIBO)-stable if every $\mathbf{u} \in \ell^\infty \Rightarrow \mathbf{y} \in \ell^\infty$. More generally, the system is called ℓ^p-stable if $\mathbf{u} \in \ell^p \Rightarrow \mathbf{y} \in \ell^p$, where $p \in [1, \infty]$.*

For a linear shift-invariant (LSI) system, the impulse response $\mathbf{h}[n, k]$ takes the form $\mathbf{h}[n - k]$, where the output is given by the convolution $\mathbf{y} = \mathbf{h} \star \mathbf{u}$ as [14]:

$$\mathbf{y}[n] = \sum_{k=-\infty}^{\infty} \mathbf{h}[n - k]\mathbf{u}[k] = \sum_{k=-\infty}^{\infty} \mathbf{u}[n - k]\mathbf{h}[k] \tag{1}$$

Using Theorem 3, if $\mathbf{h} \in \ell^1$ and $\mathbf{u} \in \ell^p$ for some $p \in [1, \infty]$, then it follows that [13]:

$$\|\mathbf{y}\|_{\ell^p} \leq \|\mathbf{h}\|_{\ell^1} \|\mathbf{u}\|_{\ell^p}$$

It is noted that $\mathbf{h} \in \ell^1$ is a sufficient condition for the system to be ℓ^p-stable. Furthermore, using Lemma 1, it follows that if $\mathbf{y} \in \ell^p$ for some $p \in [1, \infty)$, then $y[n] \to 0$ as $n \to \infty$. This information is useful, for example, in the design of a linear shift-invariant estimation system, where the output signal represents the estimation error. If the system impulse response is $\mathbf{h} \in \ell^1$, then the estimation error is bounded and converges asymptotically to zero if the input signal $\mathbf{u} \in \ell^p$ for some $p \in [1, \infty)$.

Example 2 (Adaptive Filtering). *In a general setting, let us consider an adaptive filtering problem in Figure 2, where a measurement vector $\mathbf{x}[n] \triangleq [x_1[n], x_2[n], \ldots, x_N[n]]^T$ is used to construct an estimate, $\hat{\mathbf{d}}[n] \triangleq (\mathbf{h} \star \mathbf{x})[n]$, of the desired signal $\mathbf{d}[n]$ by a linear shift-variant filter $\mathbf{h}[n]$ [7]. Then, the task is to synthesize an adaptive algorithm to update the filter $\mathbf{h}[n]$ such that the estimation error $\mathbf{e}[n] \triangleq (\mathbf{d}[n] - \hat{\mathbf{d}}[n]) \to 0$ as $n \to \infty$. Using Lemma 1, this could be achieved if $\mathbf{e} \in \ell^p$ for some $p \in [1, \infty)$ in the adaptive algorithm.*

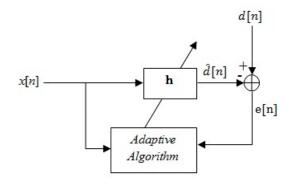**Figure 2.** An adaptive filter consisting of a shift-variant filter $\mathbf{h}$ with an adaptive algorithm for updating the filter coefficients.

If a dynamical system at any time n does not depend on the future (i.e., the system is only dependent on the past and the present) input(s), then the system is said to be *causal* [15] and the convolution in Equation (1) reduces to

$$\mathbf{y}[n] = \sum_{k=-\infty}^{n} \mathbf{h}[n-k]\mathbf{u}[k] \tag{2}$$

If, in addition, $\mathbf{u}[k] = \mathbf{0} \ \forall k < 0$, then it follows that

$$\mathbf{y}[n] = \sum_{k=0}^{n} \mathbf{h}[n-k]\mathbf{u}[k] \tag{3}$$

2.2. Hardy Spaces and Spectral Factorization for Signal Processing

This subsection introduces the concept of Hardy spaces $\mathcal{H}^p$, $1 \leq p \leq \infty$, which constitute a class of Banach spaces with a special structure; this structure is very useful for digital signal processing [3]. In particular, $\mathcal{H}^2$ and $\mathcal{H}^\infty$ spaces are of importance in robust control theory and it will be seen later in this section that the $\mathcal{H}^1$ space also plays an important role for power spectrum factorization in digital signal processing.

Recalling that, for a linear shift-invariant system with an impulse response $\mathbf{h}[n]$ and input $\mathbf{u}[n]$, the output $\mathbf{y}[n]$ is obtained by convolution [14] as: $\mathbf{y} = \mathbf{h} * \mathbf{u}$. Then, by setting $z \triangleq e^{i\omega}$ where ω is the frequency in radians, the z-transform of the impulse response $\mathbf{h}$ is defined as:

$$H(z) = \sum_{n=-\infty}^{\infty} h[n]z^{-n} \tag{4}$$

which is known as the *system transfer function* (The one-dimensional z-transform of the discrete-time impulse response $h[k]$ is the ratio of two polynomials: $H(z) \triangleq \frac{N(z)}{D(z)}$, where the degree of $N(z)$ is less than or equal to that of $D(z)$ for physically realizable systems. However, for the multi (i.e., n)-dimensional z-transform, where $n \in \mathbb{N}/\{1\}$, the resulting transfer function is given as the ratio of the numerator and denominator multinomials:

$$H(z) \triangleq \frac{N(z_1, \cdots, z_n)}{D(z_1, \cdots, z_n)}$$

The analysis of multi-dimensional *z*-transform (e.g., in signal processing of spatio-temporal processes) is significantly more complicated than that of one-dimensional z-transform, because the fundamental theorem of algebra may not be applicable to multinomials while it is always applicable to polynomials.) in the *z*-domain.

The system $H(z)$ is *stable* if the sum in Equation (4) converges, and the region of convergence (ROC) is called the *stablity region*, where all poles of $H(z)$ are located inside the unit circle with its center at zero in the complex *z*-plane. The system is said to be *minimum-phase* if all zeros of $H(z)$ are located inside the unit circle. If all zeros of $H(z)$ are located outside the unit circle, then the system is called *maximum-phase* [16], and the system is called *non-minimum-phase* if at least one zero of $H(z)$ is located outside the unit circle.

Definition 3 (Analytic Functions). *Let $D_r(z_0) \triangleq \{z \in \mathbb{C} : |z - z_0| < r\}$ be the open disc of radius $r > 0$ with center at $z_0 \in \mathbb{C}$. A complex-valued function $f(re^{i\theta})$, where $\theta \in [0, 2\pi)$, is said to be* analytic *in $D_r(z_0)$ if the derivative of $f(re^{i\theta})$ exists at each point of $D_r(z_0)$.*

Given $p \in [1, \infty]$, the Hardy space $\mathcal{H}^p$ is a set of analytic functions $f(re^{i\theta})$ with bounded $\mathcal{H}^p$-norm defined as:

$$\|f\|_{H^p} \triangleq \sup_{r \in (0,1)} \frac{1}{2\pi} \int_0^{2\pi} d\theta \, |f(re^{i\theta})|^{\frac{1}{p}} \text{ for } p \in [1, \infty)$$

$$\|f\|_{H^\infty} \triangleq \sup_{|z|<1} |f(z)|, \text{ for } p = \infty$$

The following theorem, due to Paley and Wiener [14], presents a fundamental result in the $\mathcal{H}^1$-space, which is important for spectral factorization in signal processing and for innovation representation of random processes.

Theorem 4 (Paley-Wiener). *Let $S(z)$ be a complex-valued function of the complex variable z. If $ln(S) \in H^1$, then there exists a real positive constant K_0 and a complex-valued function $H_{ca}(z)$ corresponding to a causal stable system with a causal stable inverse such that*

$$S(z) = K_0 \, H_{ca}(z) \, H_{ca}^{her}(1/\bar{z}) \tag{5}$$

where the superscript "her" indicates the Hermitian, *i.e., complex conjugate of transpose of a vector/matrix, and $\bar{z}$ is the complex conjugate of z. If, in addition, $S(z)$ is a rational polynomial, the above factors $H_{ca}(z)$ and $H_{ca}^{her}(1/\bar{z})$ are minimum-phase and maximum-phase components, respectively. This is called the* Paley-Wiener condition.

Proof. The proof of the Paley-Wiener Theorem is given in details by Therrian [14]. □

It follows from Equation (1) that, for a linear shift-invariant stable system with a deterministic LSI impulse response $\mathbf{h}[n]$ and a wide sense stationary (WSS) input signal $\mathbf{u}[n]$, the expected value of the output $\mathbf{y}[n]$ is:

$$E[\mathbf{y}[n]] = \sum_{k=-\infty}^{\infty} \mathbf{h}[n-k]E[\mathbf{u}[k]] \tag{6}$$

Since the input **u** is WSS, expected values, m_y and m_u, of the output **y** and input **u**, respectively, are related as:

$$m_y = \left(\sum_{k=-\infty}^{\infty} \mathbf{h}[k] \right) m_u \tag{7}$$

Autocorrelation of a random vector $\mathbf{x}[k]$ is denoted as $\mathbf{r}_{xx}[k] \triangleq E\left[\mathbf{x}[k]\mathbf{x}^{her}[k]\right]$, and the cross-correlation between the output **y** and the input **u** is given by

$$\mathbf{r}_{yu}[n_1, n_0] = \sum_{k=-\infty}^{\infty} \mathbf{h}[n_1 - k]\, \mathbf{r}_{uu}[k - n_0] \tag{8}$$

The above equation leads to the following important relations between correlation functions [14]:

$$r_{yu}[\ell] = \mathbf{h}[\ell] * r_{uu}[\ell] \text{ and } \mathbf{r}_{yy}[\ell] = \mathbf{h}[\ell] * \mathbf{r}_{uu}[\ell] * \mathbf{h}^{her}[-\ell] \tag{9}$$

where the superscript *her* indicates the Hermitian, i.e., the complex conjugate of transpose of a vector/matrix.

The Fourier transform of $\mathbf{r}_{xx}[k]$ for a WSS random sequence $\mathbf{x}[k]$ is called the *power spectral density function* [7], defined as:

$$S_{xx}(e^{i\omega}) \triangleq \sum_{k=-\infty}^{\infty} e^{-i\omega k} \mathbf{r}_{xx}[k] \tag{10}$$

and its inverse Fourier transform, which is equal to the autocorrelation function, is obtained as:

$$r_{xx}[k] = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega\, e^{i\omega k} S_{xx}(e^{i\omega}) \tag{11}$$

The z-transform of the autocorrelation function for a WSS random sequence $\mathbf{x}[n]$ is called the *complex spectral density function* and is defined as:

$$S_{xx}(z) \triangleq \sum_{k=-\infty}^{\infty} \mathbf{r}_{xx}[k]z^{-k} \tag{12}$$

and its inverse is given by the contour integral

$$\mathbf{r}_{xx}[k] = \frac{1}{2\pi i} \oint_C dz\, z^{k-1} S_{xx}(z) \tag{13}$$

Since the autocorrelation function of a zero-mean white noise with variance σ_w^2 is given by $\mathbf{r}_w \triangleq \sigma_w^2 \delta[k]$, the power spectral density is a constant σ_w^2 for a stationary white noise.

Using the property that the convolution in the time domain is a product in the Fourier transform domain and using Equation (9), it follows that

$$S_{yx}(e^{i\omega}) = H(e^{i\omega})S_{xx}(e^{i\omega}) \tag{14}$$

where $H(e^{i\omega})$ is the system transfer function (i.e., the Fourier transform of $\mathbf{h}[k]$). A few algebraic computations yield the following relation [14]:

$$S_{yy}(e^{i\omega}) = H(e^{i\omega})S_{xx}(e^{i\omega})H^{her}(e^{i\omega}) \tag{15}$$

In a similar manner, the following relations are obtained for the complex spectral density

$$S_{yx}(z) = H(z)S_{xx}(z)$$
$$\text{and } S_{yy}(z) = H(z)S_{xx}(z)H^{her}(1/\bar{z}) \tag{16}$$

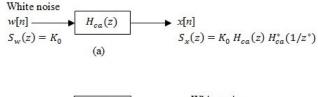Let us consider a WSS random sequence $\{x[k]\}$ whose complex spectral density satisfies the Paley-Wiener condition:

$$ln(S_{xx}) \in H^1, \text{ i.e., } \int_{-\pi}^{\pi} |\ell n\, S_{xx}(e^{i\omega})|\, d\omega < \infty \tag{17}$$

Then, by Theorem 4, there exists a real positive constant K_0 and a complex-valued transfer function $H_{ca}(z)$ of a causal stable system with a causal stable inverse such that

$$\mathbf{r}_{xx}(z) = K_0\, H_{ca}(z)\, H_{ca}^{her}(1/\bar{z}) \tag{18}$$

Remark 1. *A process, whose (complex) spectral density satisfies Equation (17), is called a regular process (see [7,14]). The spectral density factorization given by Equation (18) has important applications in signal processing. This includes what is called innovations representation of the random process [14], in view of which, any regular process can be realized as the output of a causal linear filter $H_{ca}(z)$ driven by a white noise with variance K_0 as shown in Figure 3.*

It is worth-mentioning that this type of process covers a wide range of random processes. In particular, any process whose complex spectral density is a rational function of z is a regular process.
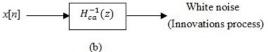

Figure 3. Innovations representation of a random process. (**a**) Signal model. (**b**) Inverse filter.

Example 3 ([14]). *Consider a random sequence $x[n]$ with a complex spectral density function:*

$$S_{xx}(z) = \frac{-(1/a)}{z - (a + 1/a) + z^{-1}}$$

which could be re-written as:

$$S_{xx}(z) = \frac{1}{-az + (1 + a^2) - az^{-1}}$$
$$= \frac{1}{(1 - az^{-1})} \cdot \frac{1}{(1 - az)}$$

Using Paley-Wiener Theorem, $\mathbf{x}[n]$ can be realized as the output of a causally stable system, given by:

$$H_{ca}(z) = \frac{1}{1 - az^{-1}} \tag{19}$$

excited by a zero-mean white noise with unit variance $\sigma^2 = 1$. It is important to note that since $S_x(z)$ is a rational polynomial, $H_{ca}(z)$ should be minimum-phase. This is the case for the one given by Equation (19).

Since the function can be factored as:

$$S_{xx}(z) = \frac{1}{-az + (1 + a^2) - az^{-1}} = \frac{1}{(z - a)} \cdot \frac{1}{(z^{-1} - a)}$$

a possible pitfall here is to choose

$$H_{ca}(z) = \frac{1}{z-a} = \frac{z^{-1}}{1-az^{-1}} \tag{20}$$

The term in Equation (20) is not minimum-phase because it has a zero at $|z| = \infty$. Moreover, the inverse $\mathcal{H}_{ca}^{-1}(z) = z - a$ is not causal. Therefore, the spectral factorization with $H_{ca}(z)$ given by Equation (20) is not physically realizable for the given random sequence $\{x[k]\}$.

As mentioned before, any random process whose complex spectral density is a rational polynomial is a regular process, and therefore it satisfies the Paley-Wiener condition. However, this is not a necessary condition for being a regular process as seen in the following example.

Example 4 ([14]). *Let a random sequence $x[n]$ have a complex spectral density $S_x(z) = e^{z+z^{-1}}$.*
Then, the corresponding power spectral density $S_x(e^{i\omega}) = e^{2\cos\omega}$ satisfies the Paley-Wiener condition that is given as:

$$\int_{-\pi}^{\pi} d\omega \, |\ell n \, S_x(e^{i\omega})| = \int_{-\pi}^{\pi} d\omega \, |2\cos\omega| < \infty$$

Therefore, the given random sequence is regular and has an innovations representation. The spectral factorization can be done as follows:

$$S_x(z) = 1 \cdot e^{z^{-1}} \cdot e^z$$

Then, the causal factor is given by

$$H_{ca}(z) = e^{z^{-1}} \tag{21}$$

which converges everywhere except at $z = 0$. The impulse response of the filter is: $h_{ca}[n] = \frac{1}{n!} U[n]$, where $U[n]$ is the unit (discrete) step function, because

$$H_{ca}(z) = e^{z^{-1}} = \sum_{k=0}^{\infty} \frac{1}{k!} z^{-k}$$

So, the given random sequence can be realized as the output of a system, with a transfer function given by Equation (21), which is driven by a zero-mean white noise with a unit variance (i.e., $\sigma^2 = 1$).

In fact, a regular process is related to the corresponding *predictable process* that can be predicted with zero error. The relation between these two processes are given by the following fundamental theorem [7].

Theorem 5 (Wold Decomposition Theorem). *A general random sequence $\mathbf{x}[n]$ can be written as the sum of two processes as:*

$$\mathbf{x}[n] = \mathbf{x}_r[n] + \mathbf{x}_p[n] \tag{22}$$

where $\mathbf{x}_r[n]$ is a regular process and $\mathbf{x}_p[n]$ is a predictable process, with $\mathbf{x}_r[n]$ being orthogonal to $\mathbf{x}_p[n]$, i.e., $E\{\mathbf{x}_r[m]\mathbf{x}_p^{her}[n]\} = 0 \ \forall m, n$.

Proof. The proof is given in [7]. □

2.3. Weak Topology in a Banach Space

It follows from the Appendix A that an appropriate collection of open sets in a metric space defines its topology, and such a topology is called a *metric topology* or *strong topology*. In fact, a base for the strong topology on a Banach space X is the collection of all open balls, i.e., sets of the form:

$$\{f \in X : \|f - g\| < r\}, \tag{23}$$

where the center g is a vector/function in X and the radius r is a positive real number. In this topology, convergence of a sequence, $\{f_n\}$, of functions in X to a limit g in X is referred to as *strong convergence*, which implies that $\|g - f_n\| \to 0$ and is denoted by $f_n \xrightarrow{s} g$. Besides strong convergence, other notions of convergence (e.g., *weak convergence* and *uniform convergence*) have been introduced in the literature, which play significant roles in the theory of Banach algebra [1].

We now introduce the notions of weak convergence and weak topology. Given a Banach space X over a field $\mathbb{K}$, let $\mathcal{F} \triangleq \{F_1, F_2, \cdots\}$ be a set of bounded linear functionals (A functional is a mapping of a vector space X into its field $\mathbb{K}$. Then, the set of all linear bounded (equivalently, linear continuous) functionals in X is called the dual space X^*.) on X, i.e., each F_i is an element in the dual space $X^\star$ and hence $\mathcal{F} \subset X^\star$. Given an $\varepsilon > 0$ and a vector/function $f_0 \in X$, let us define the set:

$$\Omega(\mathcal{F}; f_0, \varepsilon) \triangleq \{f \in X; |F_i(f) - F_i(f_0)| < \varepsilon, \forall F_i \in \mathcal{F}\} \tag{24}$$

A class of such sets is obtained by varying $\varepsilon \to 0^+$ in Equation (24) to establish the notions of weak convergence and weak topology. Some of these convergence concepts in the space of linear bounded operators are briefly explained in the following definitions, which are introduced for different notions of convergence of sequences $\{T^k\}$ of bounded linear operators in Banach spaces.

Definition 4 (Convergence in operator norm or uniform convergence). *Let $T^k \in BL(V, V)$ be a bounded linear operator from V into V. Then, the sequence $\{T^k\}$ converges to some $T \in BL(V, V)$ in the operator norm (also called uniform convergence) if the induced norm $\|(T - T^k)\|_{ind} \triangleq \lim_{k \to \infty} \sup_{\|x\|_V = 1} \|(T - T^k)x\|_V = 0$, which is denoted as: $T^k \xrightarrow{u} T$.*

Definition 5 (Strong convergence). *Let $T^k \in BL(V, V)$ be a bounded linear operator from V into V. Then, the sequence $\{T^k\}$ converges strongly to some $T \in BL(V, V)$ if $\lim_{k \to \infty} \|(T - T^k)x\|_V = 0 \ \forall x \in V$, which is denoted as $T^k \xrightarrow{s} T$.*

Definition 6 (Weak convergence). *Let $T^k \in BL(V, V)$ be a bounded linear operator from V into V. Then, the sequence $\{T^k\}$ converges weakly to some $T \in BL(V, V)$ if*

$$\forall F \in V^\star \ \forall x \in V, \ \lim_{k \to \infty} |(F(Tx) - F(T^k x))| = 0,$$

which is denoted as $T^k \xrightarrow{w} T$.

Remark 2 (Convergence in operator norm). *$\Rightarrow$ (Strong convergence) $\Rightarrow$ (Weak Convergence). The converse is not true, in general.*

To show $(T^k \xrightarrow{u} T) \Rightarrow (T^k \xrightarrow{s} T)$, we proceed as:
$\forall x \in V, \|(T - T^k)x\|_V \leq \|(T - T^k)\|_{ind} \|x\|_V$ implies that, given $T^k \xrightarrow{u} T$, i.e., $\lim_{k \to \infty} \|(T - T^k)\|_{ind} = 0$, it follows that $\lim_{k \to \infty} \|(T - T^k)x\|_V = 0 \ \forall x \in V$, i.e., $T^k \xrightarrow{s} T$.

To show $(T^k \xrightarrow{s} T) \Rightarrow (T^k \xrightarrow{w} T)$:, we proceed as:
$T^k \xrightarrow{s} T \Rightarrow \lim_{k \to \infty} \|(T - T^k)x\|_V = 0 \ \forall x \in V$. Let $f \in V^\star$; then, it follows from linearity and boundedness of the functional f that $|f(x)| \leq \|f\|_{ind} \|x\|_V$. Therefore, $\forall x \in V \ \forall f \in V^\star$, $\lim_{k \to \infty} |(f(Tx) - f(T^k x))| = 0 \Rightarrow T^k \xrightarrow{w} T$
We demonstrate the falsity of the converse by two counterexamples, one for each case.

(Strong convergence) $\nRightarrow$ (Convergence in operator norm): Let us define $x \triangleq \{\xi_n : n \in \mathbb{N}\}$ and a sequence of bounded linear operators $T^k : \ell_2 \to \ell_2 \ \forall k \in \mathbb{N}$ as:

$$T^k x \triangleq \{\underbrace{0, 0, 0, \cdots, 0}_{first\ k\ terms}, \xi_{k+1}, \xi_{k+2}, \cdots\}$$

Therefore, T^k is a bounded linear operator, i.e., $T^k \in BL(\ell_2, \ell_2)$. Since $x \in \ell_2$, it follows that

$$\lim_{k \to \infty} \|T^k x\|_{\ell_2} = 0 \Rightarrow \lim_{k \to \infty} \|T^k\| =_s \mathbf{0}_{BL(\ell_2, \ell_2)}$$

However, the limit may not converge in the induced norm, $\lim_{k \to \infty} \sup_{\|x\|_{\ell_2} = 1} \|T^k x\|_{\ell_2} = 1$ as seen by choosing $x = \{\underbrace{0, 0, 0, \cdots, 0}_{\text{first } k \text{ terms}}, \xi_{k+1}, \xi_{k+2}, \cdots\}$ with $\|x\|_{\ell_2} = 1 \Rightarrow \lim_{k \to \infty} \|T^k\| \neq_u$ $\mathbf{0}_{BL(\ell_2, \ell_2)}$.

Therefore, (Strong convergence) $\not\Rightarrow$ (Convergence in operator norm).

(Weak convergence) $\not\Rightarrow$ (Strong convergence): Let us define a sequence of bounded linear operators $T^k : \ell_2 \to \ell_2 \ \forall k \in \mathbb{N}$ as:

$$T^k x = \{\underbrace{0, 0, 0, \cdots, 0}_{\text{first } k \text{ terms}}, \xi_1, \xi_2, \cdots\}$$

where $x \triangleq \{\xi_n : n \in \mathbb{N}\}$. It is given that $\{T^k\}$ is a sequence of bounded linear operators, i.e., each $T^k \in BL(\ell_2, \ell_2)$. Furthermore, in this Hilbert space setting, it follows from the Riesz Representation Theorem that every $f \in \ell_2^\star$ can be represented as:

$$f(x) = \langle x, y \rangle_{\ell_2} = \sum_{n=1}^{\infty} \xi_n \eta_n, \ \text{where } y = \{\eta_k : k \in \mathbb{N}\}$$

It follows by Cauchy-Schwarz inequality that, as $k \to \infty$

$$|f(T^k x)|^2 = |\langle T^k x, y \rangle|^2 \le \sum_{n=1}^{\infty} |\xi_n|^2 \sum_{m=k+1}^{\infty} |\eta_m|^2 \to 0$$

However,

$$\|T^k x\|_{\ell_2} = \|x\|_{\ell_2} \ \forall k \in \mathbb{N} \Rightarrow \exists x \neq \mathbf{0}_{\ell_2} \text{such that}$$

$$\lim_{k \to \infty} \|T^k x\|_{\ell_2} \neq 0 \ \Rightarrow \lim_{k \to \infty} T^k \neq_s \mathbf{0}_{BL(\ell_2, \ell_2)}$$

Therefore, (Weak convergence) $\not\Rightarrow$ (Strong convergence).

Remark 3. *It is noted that, for finite-dimensional vector spaces, the notions of strong convergence and weak convergence are indistinguishable. Equivalently, we make the following statement:*

In a finite-dimensional Banach space V, the weak topology generated by $V^\star$ is the same as the strong topology generated by V.

However, in the analysis of stochastic processes, we deal with infinite-dimensional spaces of signal functions, which may not have the same criteria for weak convergence and strong convergence. This is especially applicable to statistical signal processing, where the expectation of the estimation error is required to weakly converge to zero without having the strong convergence of the error signal itself to zero.

Based on the concept of weak convergence, weak topology is defined as follows:

Definition 7 (Convergence in weak topology). *Given a Banach space X, let there be a class of bounded linear functionals $\mathcal{F} \subseteq X^\star$, and let $\Im(\mathcal{F})$ be the topology in X generated by $\mathcal{F}$. Then, for a given vector/function $g \in X$, a sequence $\{f_n\} \subset X$ is said to converge to g in the weak topology $\Im(\mathcal{F})$, denoted as $f_n \xrightarrow{w} g$ in $\Im(\mathcal{F})$, provided that $F_\alpha(f_n)$ converges strongly to $F_\alpha(g)$, denoted as $F_\alpha(f_n) \xrightarrow{s} F_\alpha(g) \ \forall F_\alpha \in \mathcal{F}$.*

Weak convergence in Definition 7 is a generalization of weak convergence as introduced in the functional analysis literature, which implies that a sequence $\{f_n\} \subset X$ converges weakly to some $g \in X$ if $G(f_n) \xrightarrow{s} G(g) \ \forall G \in X^\star$ [10].

Remark 4. *The concept of topological spaces and weak topology are important for learning using statistical invariants (LUSI). In a machine learning paradigm, learning machines often compute statistical invariants for specific problems with the objective of reducing the expected values of errors in a such way that preserves these invariants. In contrast to classical machine learning that employs the mechanism of strong convergence for approximations to the desired function, LUSI can significantly increase the rate of convergence by combining the mechanisms of strong convergence and weak convergence [17]. Furthermore, the notion of weak topology is also important when dealing with shift spaces for signal analysis that uses symbolic dynamics, as explained in [18,19].*

3. Hilbert Spaces for Signal Processing

This section introduces the concept of Hilbert spaces, which forms the backbone in the disciplines of signal processing and other fields of engineering. Details are provided in many textbooks such as Naylor and Sell [2].

Definition 8 (Hilbert Spaces). *Let a vector space X be defined over a field $\mathbb{K}$, which is either $\mathbb{R}$ or $\mathbb{C}$. A function $\langle \bullet, \bullet \rangle : X \times X \to \mathbb{K}$ is called an inner product if, for $\forall x, y, z \in X$ and $\forall \alpha \in \mathbb{K}$, the following conditions hold:*

1. *(positive definiteness) $\langle x, x \rangle > 0$ when $x \neq \mathbf{0}$;*
2. *(additivity) $\langle (x + y), z \rangle = \langle x, z \rangle + \langle y, z \rangle$;*
3. *(homogeneity) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$;*
4. *(symmetry) $\langle x, y \rangle = \overline{\langle y, x \rangle}$*

Then, $\left(X, \langle \bullet, \bullet \rangle \right)$ is called a inner product space or a pre-Hilbert space, and a complete inner product space (i.e., where every Cauchy sequence converges in the space) is called a real (resp. complex) Hilbert space, depending on whether the vector space is defined over $\mathbb{R}$ (resp. $\mathbb{C}$).

The following two properties are immediate consequences of the four properties in Definition 8:

- $\langle x, (y + z) \rangle = \langle x, y \rangle + \langle x, z \rangle$;
- $\langle x, \alpha y \rangle = \overline{\alpha} \langle x, y \rangle$;

It is also noted that every inner product space is a normed space with the norm $\|x\| \triangleq \sqrt{\langle x, x \rangle} \ \forall x \in X$ [2].

Example 5. *An example of a Hilbert space is the ℓ^2 space of square-summable sequences. Given two sequences $x = \{x_n\}$ and $y = \{y_n\}$ in ℓ^2, the inner product is given by $\langle x, y \rangle \triangleq \sum_{n=-\infty}^{\infty} \overline{x}_n y_n$. Two vectors x and y in a Hilbert space H are said to be orthogonal if $\langle x, y \rangle = 0$. Given a subspace $V \subset H$, its orthogonal complement is denoted as: $V^{\perp} \triangleq \{ u \in H : \langle u, v \rangle = 0 \ \forall v \in V \}$; consequently, $V \oplus V^{\perp} = H$*

Hilbert spaces have many common interesting properties that make them to be important in optimization theory [12]. As we will see in the sequel, these properties form the core of many fundamental results in adaptive and statistical signal processing, and they are established through the following theorem.

Theorem 6 (Riesz Representation Theorem [2,5]). *Let H be a a Hilbert space. Then, for every bounded linear functional $f : H \to \mathbb{C}$, there exists a unique $y \in H$ such that $f(x) = \langle y, x \rangle_H$ $\forall x \in H$.*

Proof. The proof is given in [2] in pp. 345–346. □

Remark 5. *For the Hilbert spaces ℓ^2 (resp. L^2), this result can be obtained by using a theorem [5], which states that, given $p \in [1, \infty)$, ℓ^q (resp. L^q) is isometrically isomorphic to the dual space of ℓ^p (resp. L^p) provided that $\dfrac{1}{p} + \dfrac{1}{q} = 1$, where q is called the conjugate of p. Since the conjugate of $p = 2$ is $q = 2$, it follows that ℓ^2 is isometrically isomorphic to $\left(\ell^2\right)^*$, and similar relations hold for L^2 and $(L^2)^*$ (for example, see [3]); hence, ℓ^2 and L^2 are reflexive. Generalization of this fact is stated as the following theorem.*

Theorem 7. *Every Hilbert space is reflexive, i.e., H is isometrically isomorphic to its dual space H^*.*

The proof of this theorem is given in many textbooks on functional analysis (e.g., [5,10,20]).

Another important property of Hilbert spaces, which is widely used in signal processing in combination with the previous two properties (see Theorens 6 and 7), is given as the following theorem [5]:

Theorem 8 (Orthogonal Projections). *Let H be a Hilbert space, and let $V \subset H$ be a closed subspace of H, implying that V is also a Hilbert space. Then, it follows that*
(i) $H = V \oplus V^\perp$. That is, given $x \in H$, there exists a unique pair $v \in V$ and $u \in V^\perp$ such that $x = u + v$.
(ii) $v_V(x)$ is the unique vector in V having minimal distance from a vector $x \in H$, while $u_{V^\perp}(x)$ is the unique vector in $V^\perp$ having minimal distance from x.
(iii) The orthogonal projections $x \mapsto v_V(x)$ and $x \mapsto u_{V^\perp}$ are linear continuous operators, with norm ≤ 1.

Proof. The proof is given in [2] in pp. 300–305.　□

Remark 6. *It follows from Theorem 8 that the decomposition in Equation (22) in Section 2 is indeed unique. Based on this fact, any random process generally consists of two unique orthogonal components; a predictable component and an unpredictable component. That is, if one wants to predict $x[n]$ by using N past observations $\{x_{n-N}, x_{n-N+1}, \ldots, x_{n-1}\}$, then let $\hat{x}[n] = \sum_{k=1}^{N} a_k x[n-k]$ denote an optimal linear prediction of $x[n]$. Such a prediction can be obtained by applying the orthogonality principle, where the prediction error is given by $e[n] \triangleq x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{N} a_k x[n-k]$, and the process $x[n]$ can be expressed as:*

$$x[n] = \hat{x}[n] + e[n] = \hat{x}[n] + \left(x[n] - \sum_{k=1}^{N} a_k x[n-k] \right)$$

Hence, the part $\hat{x}[n]$ represents the predictable part of $x[n]$, which corresponds to $x_p[n]$ in Wold decomposition Theorem in Equation (22), while the error $e[n]$ represents the unpredictable part of $x[n]$, which corresponds to $x_r[n]$ in Wold decomposition Theorem. That is, the regular process represents the difference *between the random process and its optimal prediction. Therefore, the output of $H_{ca}^{-1}(z)$ represents only the* new *part of information, brought by $x[n]$, which cannot be extracted from the past observations. Therefore, the output of $H_{ca}^{-1}(z)$ is called* innovations process *as depicted in Figure 3b.*

Another interesting result on Hilbert spaces is stated in the following theorem [2,10].

Theorem 9 (Bessel Inequality). *Let H be a Hilbert space and let $V = span\{e_1, e_2, \ldots\}$ be a subspace of H. If $P_V : H \rightarrow V$ denotes the orthogonal projection of elements in H into V, then, the Bessel inequality*

$$\sum_{k \geq 1} |(x, e_k)|^2 = \|P_V x\|^2 \leq \|x\|^2 \tag{25}$$

holds for every $x \in H$. Moreover,

$$\sum_{k \geq 1} (x, e_k) e_k = P_V x \qquad (26)$$

This theorem has an important implication to signal processing as explained in the following subsection.

3.1. Fourier Series Expansion in a Hilbert Space

In the Hilbert space $L^2([t_0, t_0 + T])$, which is the space of all square-integrable periodic functions $f : [t_0, t_0 + T] \rightarrow \mathbb{C}$ with $t_0 \in \mathbb{R}$ with a period $T \in (0, \infty)$, an inner product is defined as:

$$\langle f, g \rangle \triangleq \int_{t_0}^{t_0 + T} dt \, \overline{f}(t) g(t) \quad \forall f, g \in L^2([t_0, t_0 + T]) \qquad (27)$$

Let the set of functions $S \triangleq \{\varphi_n(t), n \in \mathbb{Z}\}$, where

$$\varphi_n(t) \triangleq \frac{e^{i\omega_n t}}{\sqrt{T}}; \; \omega_n \triangleq 2\pi n / T \; \forall n \in \mathbb{Z} \qquad (28)$$

Then, it follows by setting $T = 2\pi$ that

$$\langle \varphi_m, \varphi_n \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} dt \, \overline{e^{imt}} \, e^{int} = \begin{cases} 1 & \text{if } m = n \\ 0, & \text{if } m \neq n \end{cases}$$

and thus $\langle \varphi_m, \varphi_n \rangle = \delta_{mn} \; \forall m, n \in \mathbb{Z} \qquad (29)$

where δ_{mn} is called the Kronecker delta. Moreover, it turns out that $span(S)$ is dense in $L^2([t_0, t_0 + T])$, i.e., the completion $\overline{span(S)} = L^2([t_0, t_0 + T])$ (see [10]). Therefore, given any $f \in L^2([t_0, t_0 + T])$, there exists a sequence of scalars $\{c_k\} \subset \mathbb{C}$ such that

$$\lim_{N \to \infty} \|f - \sum_{k=-N}^{N} c_k \, \varphi_k \|_{L^2} = 0$$

That is,

$$f(t) \overset{ms}{=} \sum_{k=-\infty}^{\infty} c_k \varphi_k(t) \qquad (30)$$

Therefore, $\{\varphi_n(t); n \in \mathbb{Z}\}$ is in fact an orthonormal basis of $L^2([t_0, t_0 + T])$. The infinite sum $\sum_{k=-\infty}^{\infty} c_k \varphi_k$ is called the *Fourier series expansion* of f, where $\{c_k\}$ are the *Fourier coefficients*. Using Equation (29) and taking the inner product of both sides of Equation (30) by φ_n, for any $n \in \mathbb{Z}$, yields

$$c_n \overset{ms}{=} \langle \varphi_n, f(t) \rangle \qquad (31)$$

Hence, Equation (30) can be rewritten in the following form:

$$f(t) \overset{ms}{=} \sum_{n=-\infty}^{\infty} \langle \varphi_n, f(t) \rangle \, \varphi_n(t) \qquad (32)$$

Moreover, it follows from Equation (26) in Theorem 9 that

$$\|f\|_{L_2}^2 = \sum_{k=-\infty}^{\infty} |c_k|^2 \qquad (33)$$

In view of Equation (32), Fourier expansion of any square integrable signal $f \in L^2([t_0, t_0 + T])$ can be decomposed as a linear combination of *harmonic modes* φ_k with frequencies ω_k [21], where each Fourier coefficient c_k represents the signal's component associated with each mode φ_k. Furthermore, Equation (33) reveals how signal's energy $\|f\|_{L_2}^2$ is distributed over the signal's components $\{c_k\}$ and demonstrates an important

fact that, for each component c_k, the value $|c_k|^2$ represents a part of the signal's energy contributed by the component c_k. This fact plays a central role in *signal compression*, where a signal $f \in L^2([t_0, t_0 + T])$ is approximated by using as few Fourier coefficients as possible; this is accomplished with a minimum approximation error by considering those values of $\{c_k\}$ with large magnitudes and by discarding those coefficients with small magnitudes.

Now we summarize the main results of Fourier series expansion of periodic functions as a theorem.

Theorem 10 (Fourier Series Theorem)**.** *Let $\{\varphi_n\}$ be an orthonormal set in a Hilbert space H. Then, the following statements are equivalent:*

1. *$\{\varphi_n\}$ is an orthonormal basis of H, i.e., $\{\varphi_n\}$ is a complete orthonormal set in H.*
2. *(Fourier series expansion) Any vector $x \in H$ can be expanded as: $x \stackrel{ms}{=} \sum_{n \in \mathbb{N}} \langle x, \varphi_n \rangle \varphi_n$. Note: The inner products $\langle x, \varphi_n \rangle$ are called Fourier coefficients of the vector x.*
3. *(Parseval Equality) For any two vectors $x, y \in H$, the inner product: $\langle x, y \rangle = \sum_{n \in \mathbb{N}} \langle x, \varphi_n \rangle \overline{\langle y, \varphi_n \rangle}$*
4. *The norm: $\|x\|^2 = \sum_{n \in \mathbb{N}} |\langle x, \varphi_n \rangle|^2 \ \forall x \in H$*
5. *Let U be a subspace of H such that U contains the sequence $\{\varphi_n\}$. Then, U is dense in H, i.e., $\overline{U} = H$.*

Proof. The proof is given in [2] in pp. 307–312. $\square$

3.2. Fourier Transform and Inverse Fourier Transform

For decomposition by Fourier series expansion, a function needs to be periodic as seen in Section 3.1. To extend this analysis to non-periodic functions, we first consider square-integrable periodic functions $f : [-T/2, T/2] \to \mathbb{C}$ and let $T \to \infty$ so that the restriction of periodic functions can be removed. Then, a combination of Equations (28) and (31) yields:

$$c_n \stackrel{ms}{=} \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} dt \, e^{-i\omega_n t} f(t) \tag{34}$$

and we define:

$$\hat{f}_T(\omega_n) \triangleq \sqrt{T} c_n \tag{35}$$

Having $n \to \infty$ and $\omega_n \to \omega$ as $T \to \infty$, it follows that

$$\hat{f}(\omega) \triangleq \lim_{T, n \to \infty} \hat{f}_T(\omega_n) = \int_{-\infty}^{\infty} dt \, e^{-i\omega t} f(t) \tag{36}$$

Now, we have *Fourier transform* of a signal $f \in L_2(\mathbb{R})$. Since $L_2(\mathbb{R})$ is the completion $\overline{L_1(\mathbb{R}) \cap L_2(\mathbb{R})}$, we impose a mild restriction: $f(t)$ to be both absolute-value integrable and square-integrable. Nevertheless, this restriction is satisfied if f is an analytic function [21].

To obtain the inverse Fourier transform, we substitute Equations (28) and (35) into Equation (30), which yields

$$f(t) \stackrel{ms}{=} \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{i\omega_n t} \hat{f}_T(\omega_n) \tag{37}$$

By defining $\omega_n \triangleq \frac{n}{T}$, we have $\Delta\omega_n \triangleq \omega_{n+1} - \omega_n = \frac{1}{T}$. Then, substitution of $\frac{1}{T}$ into Equation (37) for $\Delta\omega_n$ yields

$$f(t) \stackrel{ms}{=} \sum_{n=-\infty}^{\infty} \Delta\omega_n \, \hat{f}_T(\omega_n) \, e^{i\omega_n t} \tag{38}$$

In the limits $T \to \infty$ and $n \to \infty$, Equation (38) becomes the *inverse Fourier transform* by using the Riemann sum [4]:

$$f(t) \overset{ms}{=} \int_{-\infty}^{\infty} d\omega\, e^{i\omega t} \hat{f}(\omega) \tag{39}$$

Equation (33) can also be rewritten as:

$$\|f\|_T^2 = \sum_{n=-\infty}^{\infty} \Delta\omega_n\, |\hat{f}_T(\omega_n)|^2 \tag{40}$$

This formula shows that a signal $f(t) \in L^2(\mathbb{R})$ has, at any given time t, (possibly) uncountably many harmonic components distributed over the frequency range $-\infty < \omega < \infty$, and the magnitude of the harmonic component at a frequency ω is given by the signal's Fourier transform $\hat{f}(\omega)$. By taking the limits $T \to \infty$ and $n \to \infty$, it follows from Equations (33) and (34) that

$$\|f\|_{L^2}^2 = \int_{-\infty}^{\infty} dt\, |f(t)|^2 = \int_{-\infty}^{\infty} d\omega\, |\hat{f}(\omega)|^2 = \|\hat{f}\|_{L^2}^2 \tag{41}$$

The above relation is known as *Plancherel's theorem* [3], which implies that the total energy of the signal, obtained in the time domain $t \in \mathbb{R}$ is re-distributed over the frequency domain $\omega \in \mathbb{R}$ such that the energy density at each frequency ω is $|\hat{f}(\omega)|^2$. It is worthmentioning that the inner products of two functions f and g in the time domain and the frequency domain is related by:

$$\langle f, g \rangle_{L^2} = \langle \hat{f}, \hat{g} \rangle_{L^2} \tag{42}$$

which is known as *Parseval's identity* [2].

In many signal processing applications, the signal f is complex-valued with a discrete domain, i.e., $f : \mathbb{Z} \to \mathbb{C}$. Then the *discrete-time Fourier transform* (DTFT) is given by:

$$\hat{f}(\omega) = \sum_{n=-\infty}^{\infty} f(n) e^{-in\omega} \tag{43}$$

From Equation (43), it follows that

$$\begin{aligned} |\hat{f}(\omega)| &= \left| \sum_{n=-\infty}^{\infty} f(n) e^{-in\omega} \right| \\ &\leq \sum_{n=-\infty}^{\infty} |f(n) e^{-in\omega}| = \sum_{n=-\infty}^{\infty} |f(n)| \end{aligned} \tag{44}$$

Therefore, a sufficient condition that guarantees the DTFT to be well-defined is that $f \in \ell^1$. The original sequence can be recovered from its DTFT by the *inverse discrete-time Fourier transform* (IDTFT)

$$f(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega\, \hat{f}(\omega) e^{in\omega} \tag{45}$$

Expressing the frequency from radians/sec to Hertz, i.e., by setting $\omega = 2\pi\xi$, it follows that

$$f(n) = \int_{-1}^{1} d\xi\, \hat{f}(\xi) e^{i2\pi n\xi} \tag{46}$$

Although the Fourier transform plays a central role in signal analysis, it considers the time-averaged frequency behavior of the signal by integrating over the entire time domain $-\infty < t < \infty$. This property reduces the capability of capturing abrupt (i.e., rapid) changes which may occur in the signal; capturing such rapid changes in the signal

is crucial in many applications, such as detection of faults and anomalies. In order to remedy this shortcoming of Fourier transform, the signal is integrated over a time window, instead of integrating over the entire time domain. This gives rise to the so-called *windowed Fourier transform* (WFT) which augments Fourier transform with a *time-localization* property that would provide information about the signal simultaneously in time and frequency domains [21,22]; a quantum-mechanics-based explanation of time-frequency localization is briefly explained in [23].

3.3. Windowed Fourier Transform in a Hilbert Space

A function $g : \mathbb{R} \to \mathbb{C}$ is said to have a *compact support* $B \subset \mathbb{R}$, and we say $supp(g) = B$ if g vanishes outside its compact domain B, i.e., $g(t) = 0 \ \forall t \notin B$. Given a function $f : \mathbb{R} \to \mathbb{C}$ and a $t \in \mathbb{R}$, let us define

$$f_t(u) \triangleq \overline{g}(u - t)f(u) \tag{47}$$

where $\overline{g}(\cdot)$ is the complex conjugate of $g(\cdot)$, and $supp(g) \subset [-T, 0]$ for a positive real number T. Hence f_t is a *localized version* of f and $supp(f_t) \subseteq [t - T, t]$ [21]. Then, the *windowed Fourier transform* (WFT) of f is the Fourier transform of f_t, which is given as:

$$\tilde{f}(\omega, t) \triangleq \hat{f}_t(\omega) = \int_{-\infty}^{\infty} du \ e^{-i\omega u} f_t(u) \tag{48}$$

and the inverse WFT is obtained as:

$$f(t) = \frac{1}{\|g\|^2} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} dt \ g_{\omega,t}(u)\tilde{f}(\omega, t) \tag{49}$$

where $g_{\omega,t}(u) \triangleq e^{i\omega u} g(u - t)$.

By following Equation (48) and Equation (49), an inner product is defined as:

$$\langle g_{\omega,t}, f \rangle \triangleq \tilde{f}(\omega, t) \tag{50}$$

Using Parseval's identity in Equation (42), we have

$$\langle g_{\omega,t}, f \rangle = \langle \hat{g}_{\omega,t}, \hat{f} \rangle = \tilde{f}(\omega, t) \tag{51}$$

Example 6. *An example of the window function g is:*

$$g(u) = \begin{cases} 1 + cos(\pi u), & -1 \leq u \leq 1 \\ 0, & otherwise \end{cases}$$

As mentioned in a previous subsection, if the signal is discrete, its DTFT is used to provide a frequency representation of the signal. Although the signal $f(n)$ in this case is discrete in the time domain, its DTFT is a continuous function of the frequency ω. However, most of the devices used in signal processing are digital, and therefore it is more convenient to deal with a discrete frequency representation of the signal. Moreover, the discrete signal $f(n)$ in many cases represent a measurement data provided by some sensors, and such signals $f(n)$ are usually of finite length. The *discrete Fourier transform* (DFT) is a useful tool in signal processing that accommodates for these two issues as explained below.

Given a finite-length discrete signal $f : \{0, N - 1\} \to \mathbb{C}$, the DFT is given as:

$$\tilde{f}[k] = \sum_{n=0}^{N-1} f[n]e^{-i2\pi kn/N} \tag{52}$$

The *inverse discrete Fourier transform* IDFT is:

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} \tilde{f}[k] e^{i2\pi kn/N} \tag{53}$$

Returning to the continuous WFT, if a function f is windowed over a time interval, the resulting WFT would have a time-localization property as seen earlier in the end of Section 3.2. Moreover, Equation (51)) shows that WFT of f localizes $\hat{f}(\omega)$ to a neighborhhod of ω. Therefore, WFT has both time and frequency localizations. However, due to the *uncertainty principle*, these two kinds of localization have different physical interpretations and are mutually exclusive in the sense that making a WFT $\tilde{f}(\omega, t)$ sharper in time makes it *more* flat in frequency and vice versa [21,23]. Moreover, WFT is not efficient in scanning signals that involve time intervals much shorter or much longer than the window length T [21]. To address these issues, the notion of *wavelet transform* has been introduced in Section 3.4, which produce an efficient analysis tool to capture signal features occurring over short and long intervals.

3.4. Wavelet Transform in a Hilbert Space

The continuous wavelet transform (CWT) of a function $f : \mathbb{R} \to \mathbb{C}$ is defined as:

$$\tilde{f}(s, t) \triangleq \int_{-\infty}^{\infty} du \, \overline{\psi}_{s,t}(u) f(u) = \langle \psi_{s,t}, f \rangle \tag{54}$$

where $\psi_{s,t}$, called the *wavelet*, is a scaled and translated version

$$\psi_{s,t}(u) = |s|^{-p} \psi\left(\frac{u-t}{s}\right) \text{ where } s \neq 0 \tag{55}$$

of what is called a *mother* (or *basic*) *wavelet* $\psi \triangleq \psi_{1,0}$; and $\overline{\psi}_{s,t}$ is the complex conjugate of $\psi_{s,t}$.

It is noted from Equation (55) that when $|s| > 1$, $\psi_{s,t}$ is a *stretched* version of ψ, and when $|s| < 1$, $\psi_{s,t}$ is a *compressed* version of ψ. Moreover, if $s < 0$ then $\psi_{s,t}$ is a *reflected* version of ψ. For example, these stretching, compression, and reflection processes can be conveniently done on the time axis. The exponent term p in Equation (55) is a real number that *stretches* or *compresses* ψ along the vertical axis. The idea of using p in Equation (55) is to keep a desired norm unchanged when scaling the wavelet $\psi_{s,t}$. For example, if $p = 1$, then both ψ and $\psi_{s,t}$ have the same L^1 norm; and if $p = 1/2$, then ψ and $\psi_{s,t}$ have the same L^2 norm [21].

Using Parseval's identity, Equation (54) can be written as

$$\tilde{f}(s, t) = \langle \psi_{s,t}, f \rangle = \langle \hat{\psi}_{s,t}, \hat{f} \rangle \tag{56}$$

where $\hat{\psi}_{s,t}$ is the Fourier transform of $\psi_{s,t}$. This equality shows that wavelets transform localizes signals in both time and frequency domains, where sharpness of these localizations is controlled by the scaling factor s and the choice of the mother wavelet ψ.

Example 7. *Morlet wavelet is a (frequency-modulated) mother wavelet which is given in the time domain as:*

$$\psi(u) = e^{-i2\pi\xi_0 u} \, e^{-u^2/2} \tag{57}$$

whose Fourier transform is

$$\hat{\psi}(\xi) = e^{-(\xi-\xi_0)^2/2} \tag{58}$$

where ξ_0 is the center frequency around which the signal is localized in the frequency domain.

Various forms of the mother wavelet ψ have been reported in the wavelet literature [21,22]. All of these wavelet forms should satisfy the *admissibility condition*:

$$C_\psi \triangleq \int_{-\infty}^{\infty} d\xi \, \frac{|\hat{\psi}(\xi)|^2}{|\xi|} < \infty \tag{59}$$

where $\hat{\psi}(\xi)$ is the Fourier transform of $\psi(t)$.

At a fixed scale s, the CWT of a signal $f(u)$ yields information relevant to the feature contained in the signal at the scale s, and the behavior of this feature over time is captured by translating $\psi_{s,t}$ over t. Then this process is repeated for different scales by changing s to capture other signals' features that are relevant to different scales.

Given a CWT of a signal $f \in L^2$, the original signal f can be reconstructed by

$$f(u) \overset{ms}{=} \frac{1}{C_\psi} \int_{-\infty}^{\infty} dt \int_{-\infty}^{\infty} ds \, \frac{\tilde{f}(s,t)}{s^2} \psi_{s,t}(u) \tag{60}$$

where C_ψ is a constant depending on the wavelet ψ, and Equation (60) shows that any signal $f \in L^2$ can be represented as a superposition of shifted and dilated wavelets [24].

For a discrete signal $f : \mathbb{Z} \to \mathbb{C}$, the *discrete wavelet transform* (DWT) is used with a discrete wavelet as:

$$\psi_{s,t}[u] = |s|^{-p} \, \psi\left[\frac{u-t}{s}\right] \tag{61}$$

where s is the scaling parameter and t is the shifting parameter. The most commonly used discrete wavelets have the following values of the parameters:

$$s = 2^j, t = k2^j, \text{ and } p = 1/2$$

where j is an integer that controls the scaling parameter and specifies the level of wavelet decomposition of the signal, and k is another integer which controls the shifting parameter. Substitution of these values into Equation (61) yields the most common form of the discrete wavelet

$$\psi_{j,k}[n] = \frac{1}{\sqrt{2^j}} \psi\left[\frac{n-k2^j}{2^j}\right] \tag{62}$$

Notice that large values of j result in large scaling parameters which *stretch* the wavelet function and let the DWT capture low-frequency features in the signal. On the other hand, small values of j would make the DWT more capable of capturing high-frequency features by decreasing the scaling parameter [21,22].

Given a wavelet level j, the DWT of a sequence $\{f[n]\}$ consists of the following two parts:

The *average coefficients* $\{A_j[k2^j]\}$ are given by:

$$\begin{aligned} A_j[k2^j] &\triangleq \sum_{n=-\infty}^{\infty} f[n]\phi_{j,k}[n] \\ &= \sum_{n=-\infty}^{\infty} f[n] \frac{1}{\sqrt{2^j}} \phi\left[\frac{n-k2^j}{2^j}\right] \end{aligned} \tag{63}$$

and the *detail coefficients* $\{D_j[k2^j]\}$ are described by:

$$\begin{aligned} D_j[k2^j] &\triangleq \sum_{n=-\infty}^{\infty} f[n]\psi_{j,k}[n] \\ &= \sum_{n=-\infty}^{\infty} f[n] \frac{1}{\sqrt{2^j}} \psi\left[\frac{n-k2^j}{2^j}\right] \end{aligned} \tag{64}$$

where the scaling function $\phi_{j,k}[n]$ is associated with the wavelet function $\psi_{j,k}[n]$; full details are given in [21].

Let us now consider a special case of DWT, where the analyses (i.e., computation of $\tilde{f}(\omega, t)$ (see Equation (50)) or $\tilde{f}(s, t)$ (see Equation (54)) or their discrete samples) are made directly from relevant integration with necessary values of time-frequency or time-scale parameters, Around 1980, a new method for performing DWT was created, which is known as *Multiresolution Analysis* (MRA). This method is completely recursive and is therefore ideal for computation, as succinctly described below.

In MRA, we may think of level-1 DWT of $f[n]$ as the output of two filters connected in parallel, consisting of a low-pass filter with the impulse response g and a high-pass filter with the impulse response h, as seen in Figure 4. This is known as the *filter bank* implementation of DWT, consisting of different levels j. The cutoff frequency of each filter in the filter bank equals to a half of the bandwidth of the respective input signal. Hence, the output of each filter has a half of the bandwidth of the original sequence $f[n]$ so that it is subsampled by 2. That is,

$$\text{The average } A_1[n] \triangleq \sum_{k=-\infty}^{\infty} f[k]\, g[2n - k] \tag{65}$$

$$\text{The detail } D_1[n] \triangleq \sum_{k=-\infty}^{\infty} f[k]\, h[2n - k] \tag{66}$$

Therefore, given a level-j DWT of a discrete-time signal $f[n]$, if $A_j[k2^j]$ in the sequence of average coefficients is passed through a parallel combination of identically structured filters g and h, then the output is a sequence of level-$(j+1)$ DWT of $f[n]$ as seen in Figure 4. The features associated with different frequency components of the signal $f[n]$ can be captured by using a multilevel wavelet decomposition of $f[n]$ via iterative implementation of filter banks in the setting of time and frequency localization (see, for example, [21,23,24]).
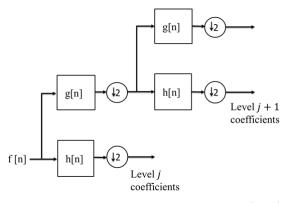

Figure 4. Implementation of level-j and level-$(j+1)$ MRA filter banks

Example 8. *Let us consider a function $f(t)$ having a wavelet transform $\tilde{f}(s, t)$, which can be interpreted as the "details" contained at fixed scales $s \neq 0$. This interpretation is especially useful in the discrete case for understanding the principles of MRA as seen below.*

Let $\phi(u)$ be a zero-mean unit-variance probability density function, which has the following properties:

- $\phi(u) \geq 0\ \forall u \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} du\, \phi(u) = 1$;
- $\int_{-\infty}^{\infty} du\, \phi(u)\, u = 0$;
- $\int_{-\infty}^{\infty} du\, \phi(u)\, u^2 = 1$;

Assuming that $\phi \in C^n$, i.e., ϕ is at least n times differentiable, where $n \in \mathbb{N}$., it follows that $\lim_{n \to \pm\infty} \phi^{(n-1)}(u) = 0$. Now letting $\psi^n(u) \triangleq (-1)^n \phi^{(n)}(u)$, we have

$$\int_{-\infty}^{\infty} du\, \psi^n(u) = (-1)^n \left[\phi^{(n-1)}(\infty) - \phi^{(n-1)}(-\infty) \right] = 0$$

Thus, ψ^n satisfies the admissibility condition in Equation (59) and hence can be used to define a CWT.

For $s \neq 0$ and $t \in \mathbb{R}$, let $\phi_{s,t}(u) = |s|^{-1}\phi\left(\frac{u-t}{s}\right)$ and $\psi_{s,t}^n(u) = |s|^{-1}\psi^n\left(\frac{u-t}{s}\right)$. Then, $\phi_{s,t}$ is a probability density with mean t and standard deviation $|s|$; and $\psi_{s,t}^n$ is qualified to be a wavelet family $\{\psi^n\}$ by setting $p = 1$ in Equation (61).

As a numerically explicit example, let ϕ represent the zero-mean unit-variance Gaussian density, i.e., $\phi(u) = \frac{\exp(-u^2/2)}{\sqrt{2\pi}}$. Since $\phi \in C^\infty$, n can be taken to be any positive integer. For instance, $\psi^1(u) = -\phi^{(1)}(u) = \frac{u\,\exp(-u^2/2)}{\sqrt{2\pi}}$ and $\psi^2(u) = \phi^{(2)}(u) = \frac{(u^2-1)\,\exp(-u^2/2)}{\sqrt{2\pi}}$, and so on. Because of the shape of the graph, $-\psi^2$ is popularly known as the Mexican hat mother wavelet, which is often used in engineering applications.

3.5. Karhunen-Loéve Expansion of Random Signals

Karhunen-Loéve (K-L) expansion is a powerful tool that generalizes Fourier series expansion for analysis of random time-dependent signals. The K-L expansion is frequently used in statistical signal processing and detection theory by using deterministic time-dependent orthonormal functions and random-variable coefficients.

Theorem 11 (Karhunen-Loéve Expansion). *Let $X(t)$ be a zero-mean, second-order random process, defined over $[-T/2, T/2]$ where $T \in (0, \infty)$, with a continuous covariance function $K_{XX}(t, \tau)$. Then, it follows that*

$$X(t) \stackrel{ms}{=} \sum_{n=1}^{\infty} X_n \phi_n(t) \ \forall t \in [-T/2, T/2] \tag{67}$$

where the (countable) sequence of (deterministic) functions $\{\phi_n(t)\}$ is a complete orthonormal set of solutions to the following integral equation:

$$\int_{-T/2}^{T/2} d\tau\, K_{XX}(t, \tau)\phi_n(\tau) = \lambda_n \phi_n(t)\ \forall t \in [-T/2, T/2] \tag{68}$$

and the random coefficients $X_n \triangleq \int_{-T/2}^{T/2} dt\, X(t)\phi_n^{her}(t)$ are mutually statistically orthogonal, i.e.,

$$E[X_n X_m^{her}] = \lambda_n \delta_{mn}\ \text{ with the Kronecker delta } \delta_{mn}$$

Proof. The proof is given in [25,26]. □

Remark 7. *The deterministic functions $\phi_n(t)$ are orthonormal in the following sense:*

$$\int_{-T/2}^{T/2} dt\, \phi_n(t)\phi_m^{her}(t) = \lambda_n \delta_{mn}$$

Example 9 (K-L expansion of white noise). *Let the covariance function of zero-mean stationary white noise $w(t)$ be $K_{ww}(t, \tau) = \sigma^2\delta(t - \tau)$. Then, the orthonormal functions $\phi_n(t)$ satisfy the K-L integral equation, for all $n \in \mathbb{N}$, as:*

$$\int_{-T/2}^{T/2} d\tau\, K_{ww}(t, \tau)\phi_n(\tau) = \sigma^2 \int_{-T/2}^{T/2} d\tau\, \delta(t - \tau)\phi_n(\tau)$$

It is also true that $\int_{-T/2}^{T/2} d\tau\, K_{ww}(t, \tau)\phi_n(\tau) = \lambda_n\phi_n(t)$, which implies that $\lambda_n\phi_n(t) = \sigma^2\phi_n(t)$ $\forall n \in \mathbb{N}$. Thus, the choice of these orthonormal functions is arbitrary and all λ_n's are identically equal to σ^2. It is concluded that, for any zero-mean white noise, the K-L expansion functions $\{\phi_n(t)\}$ can be any set of orthonormal functions with all eigenvalues $\lambda_n = \sigma^2$.

Example 10 (K-L expansion as an application to detection theory [25]). *Let us assume that a waveform $X(t)$ is observed over a finite time interval $[-T/2, T/2]$ to decide whether it contains a recoverable signal buried in noise, or the signal is completely noise-corrupted (i.e., the signal cannot be recovered). In this regard, we formulate a binary hypothesis testing problem with the hypothesis H_1 of having a recoverable signal and the hypothesis H_0 of complete noise capture, i.e.,*

$$X(t) = \begin{cases} s(t) + w(t): & \text{if } H_1 \text{ is true} \\ w(t): & \text{if } H_0 \text{ is true} \end{cases}$$

where the signal $s(t)$ is a deterministic function of time, and the noise $w(t)$ is modeled as zero-mean, unit-variance, white Gaussian. Using the K-L expansion, we simplify the above decision problem by replacing the waveform $X(t)$ with a sequence $\{X_n\}$, which reduces to a sequence of simpler problems as:

$$X_n = \begin{cases} s_n + \omega_n: & \text{if } H_1 \text{ is true} \\ \omega_n: & \text{if } H_0 \text{ is true} \end{cases}$$

where s_n and ω_n are the respective (at most countably many) K-L coefficients of the signal $s(t)$ and noise $w(t)$.

Now we take the K-L transform (instead of Fourier transform) of the received signal $X(t)$, where the transform space is the space of sequences of K-L coefficients that are mutually statistically orthogonal random variables. By taking advantage of the facts that the noise is zero-mean Gaussian and that the K-L coefficients are mutually statistically orthogonal, the random variables ω_n become jointly independent., i.e., $\{\omega_n\}$ is a sequence of independent and identically distributed (iid) random variables. By selecting the first orthonormal function as:

$$\phi_1(t) = \frac{s(t)}{\sqrt{\int_{-T/2}^{T/2} d\theta \, s^2(\theta)}}$$

we can complete the rest of the orthonormal set $\{\phi_n(t)\}$ in a valid way. We also notice that all of the random coefficients s_k, with the exception of s_1, will be zeros, i.e., X_t is affected by the presence or absence of the recoverable signal. Thus, the distributed detection problem is reduced to the following scalar detection problem:

$$X_1 = \begin{cases} \sqrt{\int_{-T/2}^{T/2} d\theta \, s^2(\theta)} + \omega_1: & \text{if } H_1 \text{ is true} \\ \omega_1: & \text{if } H_0 \text{ is true} \end{cases}$$

We note that the scalar X_1 can be computed as:

$$X_1 = \frac{\int_{-T/2}^{T/2} d\theta \, X(\theta)s(\theta)}{\sqrt{\int_{-T/2}^{T/2} d\theta \, s^2(\theta)}}$$

which is commonly referred to as a matching operation. In fact, this operation can be performed by sampling the output of a filter whose impulse response is:

$$h(t) = \frac{s(T-t)}{\sqrt{\int_{-T/2}^{T/2} d\theta \, s^2(\theta)}}$$

where the parameter T should be chosen sufficiently large to make the impulse response causal. The output of the physically realizable filter at time T is then X_t. This filter is called a matched filter and is widely used in the disciplines of communications and pattern recognition.

3.6. Reproducing Kernel Hilbert Spaces

This subsection develops the concept of reproducing kernel Hilbert spaces (RKHS) [14], in which each point in the space is a linear bounded (equivalently, a linear continuous)

functional. The continuity (or boundedness) of linear functions implies that if two linear bounded functions f and g are close to each other (i.e., $\|f - g\|$ is small in the function space), then f and g are also close to each other pointwise, i.e., $|f(t) - g(t)|$ is also small for all t.

The RKHS has many engineering and scientific applications, including those in harmonic analysis, wavelet analysis, and quantum mechanics. In particular, functions from RKHS have special properties that make them useful for function estimation problems in high-dimensional spaces, which is critically important in the fields of statistical learning theory and machine learning [17]. In fact, every function in RKHS that minimizes an empirical risk functional can be expressed as a linear combination of the kernel functions evaluated at the training points. This procedure potentially simplifies the handling of the problem from infinite-dimensional to finite-dimensional.

We now present a formal definition of reproducing kernel Hilbert spaces (RKHS). The presented theory is often applied to real-valued Hilbert spaces and can be extended to complex-valued Hilbert spaces; examples of complex-valued RKHS are spaces of analytic functions.

Definition 9 (Reproducing Kernel Hilbert Spaces). *Let T be an arbitrary non-empty set (e.g., the time domain or the spatial domain of a function) and let H be a Hilbert space of real-valued (resp. complex-valued) functions on T, equipped with pointwise vector addition and pointwise scalar multiplication, and the continuous functions in H are evaluated at each point $t \in T$. Then, H is defined to be a reproducing kernel Hilbert space (RKHS) if there exist a positive real M_t and a continuous linear functional $\mathcal{L}_t$ on H such that $|\mathcal{L}_t(f)| = |f(t)| \leq M_t \|f\|_H \quad \forall t \in T \ \forall f \in H$. [Note: Although M_t is constrained to be a positive real, it is possible that $\sup_{t \in T} M_t = \infty$.]*

Remark 8. *Definition 9 is rather a weak condition to ensure the existence of an inner product and the evaluation of every functional on H at every point in the domain T. From the application perspectives, a more useful definition would be to construct an inner product of a given function $f \in H$ with another function $K_t \in H$, which is the so-called* reproducing kernel function *for the Hilbert space H; the RKHS has taken its name from here.*

To make Definition 9 more useful for many applications, we make use of Reisz representation theorem (Theorem 6) which states that there exists a unique $K_t \in H$ with the following reproducing property for each $f \in H$, which takes values at any given $t \in T$ as:

$$f(t) = \mathcal{L}_t(f) = \langle K_t, f \rangle_H$$

Since, for a given $t \in T$, the function $K_t \in H$ takes values in $\mathbb{R}$ (resp. $\mathbb{C}$) and having another $K_\tau \in H$ associated with the parameter $\tau \in T$ and a corresponding functional L_τ on H, it follows that

$$K_t(\tau) = \mathcal{L}_\tau(K_t) = \langle K_\tau, K_t \rangle_H$$

The above situation can be interpreted as follows: K_τ is a time translation of K_t from t to τ if the set T is the time domain of the functions in the Hilbert space. This allows us to redefine the reproducing kernel of the Hilbert space H as a function $K : T \times T \rightarrow \mathbb{R}$ (resp. $\mathbb{C}$) as: $K(t, \tau) \triangleq \langle K_\tau, K_t \rangle_H$.

Example 11 (Bandlimited approximation of Dirac delta function in the RKHS setting). *Let us consider the space of continuous signals that are also band-limited with frequencies under the compact support, i.e., in the range of $[-2\pi\Omega, 2\pi\Omega]$, where the cutoff frequency $\Omega \in (0, \infty)$. It is noted that $K_t(\bullet)$ is a bandlimited version of the Dirac delta function, because $K_t(\tau)$ converges to the delta distribution, expressed as $\delta(\tau - t)$ in the weak sense, as the cutoff frequency Ω tends to infinity.*

Let us define $T = \mathbb{R}$ and $H = \left\{ f \in C^0(T) : supp(\hat{f}) \subseteq [-\Omega, \Omega] \right\}$, where $C^0(T)$ is the space of continuous functions whose domain is T, and the Fourier transform of f is: $\hat{f}(\xi) \triangleq$

$\int_{\mathbb{R}} dt \, \exp(-i2\pi\xi t) \, f(t)$ *and the inverse Fourier transform of* $\hat{f}(\xi)$ *is:* $f(t) \triangleq \int_{\mathbb{R}} d\xi \, \exp(i2\pi\xi t) \, \hat{f}(\xi)$. *Then, it follows by Cauchy-Schwarz inequality and Plancherel theorem that:*

$$|f(t)|^2 \leq \int_{-\Omega}^{\Omega} d\xi \, |e^{i2\pi\xi}|^2 \int_{-\Omega}^{\Omega} d\xi \, |\hat{f}(\xi)|^2 = 2\Omega \|f\|_H^2$$

i.e., $|f(t)| \leq \sqrt{2\Omega} \, \|f\|_H$.

It follows from the relation: $f(t) = \mathcal{L}_t(f) = \langle K_t, f \rangle_H$, *established earlier, that the functional* $\mathcal{L}_t$ *and the RKHS kernel function* K_t *are bounded. Therefore, H is indeed an RKHS.*

By choosing the kernel function in this case as: $K_t(\tau) = sinc \left(2\pi\Omega \left(\tau - t \right) \right) \triangleq \frac{\sin \left(2\pi\Omega(\tau-t) \right)}{2\pi\Omega(\tau-t)}$, *and by taking* $\lim_{\Omega\to\infty} K_t(\tau) = \delta(\tau - t)$, *it follows that as* $\Omega \to \infty$, *the Fourier transform of the kernel* $K_t(\tau)$ *becomes*

$$\hat{K}_t(\xi)) = \int_{-\infty}^{\infty} d\tau \, \exp(-i2\pi\xi\tau) \, K_t(\tau)$$

$$= \int_{-\infty}^{\infty} d\tau \, \exp(-i2\pi\xi\tau) \, \delta(\tau - t)) = \exp(-i2\pi\xi t)$$

This is a consequence of frequency modulation due to the time-shifting property of Fourier transform. Then, it follows by using Plancherel theorem that:

$$\langle K_t, f \rangle_H = \int_{-\infty}^{\infty} d\tau \, \overline{K}_t(\tau) f(\tau) = \int_{-\infty}^{\infty} d\xi \, \overline{\hat{K}}_t(\xi) \hat{f}(\xi)$$

$$= \int_{-\infty}^{\infty} d\xi \, \hat{f}(\xi) \, \exp(i2\pi\xi t) = f(t)$$

Thus, the reproducing property of the kernel is established as the cutoff frequency $\Omega \to \infty$.

4. Summary and Conclusions

This paper sheds light on some of the key concepts from functional analysis, which provide a unified mathematical framework for solving problems in engineering and applied sciences, and especially in modern signal processing. Additionally, the simple (and yet elegant) way, by which this framework facilitates formulation of different topics in signal processing along with several relevant examples, enables solving many problems in science and engineering by utilizing the concepts from the discipline of functional analysis. Some of the important results from functional analysis can find their ways to contribute for further advances in statistical signal processing and adaptive signal processing. Nevertheless one of the main difficulties in doing so is the existing gap in the terminologies and technological languages used in these two (apparently different) fields; and this paper attempts to (at least partially) bridge this gap.

Author Contributions: Conceptualization, N.F.G., A.R. and W.K.J.; methodology, N.F.G., A.R. and W.K.J.; software, N.F.G. and A.R.; formal analysis, N.F.G. and A.R.; model preparation and validation, N.F.G. and A.R.; data curation, N.F.G. and A.R.; writing—original draft preparation, N.F.G., A.R. and W.K.J.; writing—review and editing, N.F.G., A.R. and W.K.J.; funding acquisition, A.R. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Appendix A. Preliminary Concepts

This appendix introduces several preliminary (but critical) concepts from real analysis, probability theory, and topology, mainly taken from Naylor and Sell [2], Royden [4], Stark and Woods [25], Ash [27], and Munkres [28]. These references are expected to be helpful, along with their cited materials therein, for appropriate understanding of key concepts that are frequently encountered in the discipline of modern signal processing.

Appendix A.1. Metric Spaces and Topological Spaces

This subsection introduces rudimentary concepts of metric and topological spaces. Details are available in the afore-mentioned standard textbooks.

Definition A1 (Metric Spaces). *Let X be a non-empty set. A function $\rho : X \times X \to \mathbb{R}$, where $\mathbb{R}$ is the space of real numbers, is called a metric (or a distance function) on X if the following conditions hold for all $x, y, z \in X$:*

(i)   *Positivity: $\rho(x, y) \geq 0$, and $\rho(x, y) = 0$ iff $x = y$;*
(ii)   *Symmetry: $\rho(x, y) = \rho(y, x)$;*
(iii) *Triangular Inequality: $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.*

The pair (X, ρ) is called a metric space; if there is no ambiguity on ρ, the metric space is denoted only by X.

Example A1. *A well-known example of a metric space is the n-dimensional Euclidean space $(\mathbb{R}^n, \rho)$, where n is a positive integer, $\rho(x, y) \triangleq \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$ for every vector $x = [x_1, \ldots, x_n]^T$ and $y = [y_1, \ldots, y_n]^T$ in $\mathbb{R}^n$.*

Remark A1. *The set X, upon which the metric ρ can operate, is an arbitrary nonempty set. The conditions (i)–(iii) in Definition A1 are obvious if ρ operates on $\mathbb{R}^n$. However, in general, there can be other types of metric operators and X may not be $\mathbb{R}^n$; an example is the Hamming distance defined on sets of symbol sequences, which is widely used in error correction theory to measure the distance between two code words.*

Definition A2 (Open and Closed Sets). *A set $E \subseteq X$ in a metric space (X, ρ) is called open if, for all $y \in E$, there exists an open ball $B_\varepsilon(y) \triangleq \{x \in E : \rho(x, y) < \varepsilon\}$, which is of radius $\varepsilon > 0$ with center at y. A set $F \subseteq X$ is called closed if the complement $X \setminus F$ is open in (X, ρ).*

Definition A3 (Cauchy Sequence). *A sequence $\{x_n\}$ in a metric space (X, ρ) is called a Cauchy sequence if $\forall \varepsilon > 0 \, \exists n(\varepsilon) \in \mathbb{N}$ such that $\rho(x_k, x_\ell) < \varepsilon \; \forall k, \ell > n$. In other words, $\rho(x_k, x_\ell) \to 0$ as $k, \ell \to \infty$.*

Definition A4 (Completeness of a metric space). *A metric space is called complete if every Cauchy sequence converges in the metric space,*

Definition A5 (Sequential Compactness). *A metric space (X, ρ) is said to be sequentially compact if every sequence of points $\{x_1, x_2, \ldots\}$ in (X, ρ) contains a convergent subsequence $\{x_{n1}, x_{n2}, \ldots\}$, and it is called sequentially precompact if every sequence of points $\{x_1, x_2, \ldots\}$ in (X, ρ) contains a Cauchy subsequence $\{x_{n1}, x_{n2}, \ldots\}$ (see Definition A3 and also see [29]).*

In contrast to metric spaces, where a distance function ρ on a set X is used to introduce key concepts (e.g., neighborhood, open set, closed set, and convergence), a more powerful approach is to specify a system in terms of open sets to introduce such properties; this leads to the notion of a topological space (e.g., [28]).

Definition A6 (Topological Spaces). *A topology on a nonempty set X is a collection $\Im$ of subsets of X, which has the following properties:*

1. *$\Im$ must contain the empty set $\varnothing$ and the set X.*
2. *Any union of the members in $\Im$, belonging to an arbitrary (i.e., finite, countable, or uncountable) (A set is defined to be countable if it is bijective to the set, $\mathbb{N}$, of positive integers; and a finite or a countably infinite set is often called at most countable. An infinite set, which is not countable, is called uncountable. For example, the set of integers, $\mathbb{Z} \subsetneq \mathbb{R}$, is countable, while an interval $(a, b) \triangleq \{x \in \mathbb{R} : a < x < b\}$ is uncountable. These concepts lead to the fundamental difference between "continuous-time (CT) analog" and "discrete-time (DT) digital" signal processing.) subcollection of sets must be contained in $\Im$.*
3. *The intersection of the members of any finite subcollection of $\Im$ must be contained in $\Im$.*

Then, the pair $(X, \Im)$ is called a *topological space*, and the members of $\Im$ are called *open sets* of $(X, \Im)$; if B is an open set in $(X, \Im)$, the complement of B (i.e., $X \setminus B$) is called a closed set in $(X, \Im)$. If there is no confusion regarding $\Im$, then $\Im$ is often omitted from $(X, \Im)$ and only X is referred to as a topological space.

Definition A7 (Topological Basis). *A basis $\mathcal{B}$ for a topology $(X, \Im)$ is a collection of open sets in $\Im$, called basis elements, if the following two conditions hold:*

1. *For each $x \in X$, there exists at least one basis element $B \in \mathcal{B}$ such that $x \in B$.*
2. *If $x \in B_1 \cap B_2$, where $B_1, B_2 \in \mathcal{B}$, then there exists a basis element $B_3 \in \mathcal{B}$ such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$.*

Appendix A.2. Random Variables and Stochastic Processes

This subsection introduces rudimentary concepts of random variables and stochastic processes. Details are available in textbooks such as [25,27,30].

Definition A8 (algebra and σ-algebra). *Let $\mathcal{F}$ be a (non-empty) collection of subsets of a (non-empty) set Ω having some or all of the following properties.*

(a) *$\Omega \in \mathcal{F}$.*
(b) *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, where $A^c \triangleq \Omega \setminus A$.*
(c) *If $A_1, A_2, \ldots, A_n \in \mathcal{F}$ then $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$.*
(d) *If $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$,*

Then, $\mathcal{F}$ is called an algebra *(or a* field*) if the properties (a), (b) and (c) are true. If, in addition, the property (d) is true, then $\mathcal{F}$ is called a σ-algebra (or a σ-field).*

Remark A2. *The largest σ-algebra of a nonempty set Ω is the collection of all subsets of Ω, which is the* power set *2^{Ω}. On the other hand, the smallest σ-algebra consists of two sets ϕ and Ω, i.e., the indiscrete σ-algebra $\{\phi, \Omega\}$.*

Definition A9 (Borel Sets). *Given a non-empty collection $\mathcal{D}$ of subsets of Ω, the smallest σ-algebra containing $\mathcal{D}$ is called the σ-algebra* generated by $\mathcal{D}$. The Borel *σ-algebra $\mathcal{B}(\mathbb{R})$ is the σ-algebra generated by the collection of all open intervals $\{(a, b) : a, b \in \mathbb{R}\}$ in the usual topology of $\mathbb{R}$. Members of $\mathcal{B}(\mathbb{R})$ are called* Borel sets.

Definition A10 (Measure). *A countably additive measure μ on a σ-algebra $\mathcal{F}$ is a non-negative, extended real-valued function on $\mathcal{F}$ such that if $\{A_1, A_2, \ldots\}$ forms an at most countable (i.e., finite or countably infinite) collection of disjoint sets in $\mathcal{F}$, then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$. A*

measurable space *is a pair* $(\Omega, \mathcal{F})$, *and a* measure space *is a triple* $(\Omega, \mathcal{F}, \mu)$, *where* Ω *is a non-empty set,* $\mathcal{F}$ *is a* σ-*algebra of subsets of* Ω, *and* μ *is a measure on* $\mathcal{F}$. *The sets in* $\mathcal{F}$ *are called* measurable sets.

Example A2. *Let* $\Omega = \mathbb{R}^n$, *where* $n \in \mathbb{N}$ *and the Borel set* $\mathcal{B}(\mathbb{R}^n)$ *is the associated* σ-*algebra. Then,* $\mu : \mathcal{B}(\mathbb{R}^n) \rightarrow [0.\infty]$ *is called the n-dimensional* Lebesgue measure, *and* $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mu)$ *is called the n-dimensional* Lebesgue measure space. *For* $n = 1$, *i.e., in the 1-dimensional real space* $\mathbb{R}$, *given an interval* $S \in \mathcal{B}(\mathbb{R})$, *the measure* $\mu(S)$ *is the length of the interval S. Similarly, for two-dimensional (i.e.,* $n = 2$) *and three-dimensional (i.e.,* $n = 3$) *Lebesgue measures,* $\mu(S)$ *denotes the area and volume measures, respectively.*

Definition A11 (Probability Spaces). *If* $\mu(\Omega) = 1$, *then* μ *is called a* probability measure, *usually denoted by P, and the triplet* $(\Omega, \mathcal{F}, P)$ *is called a* probability space.

Definition A12 (Measurable Functions). *Let* $(\Omega_1, \mathcal{F}_1)$ *and* $(\Omega_2, \mathcal{F}_2)$ *be two measurable spaces. A function* $f : (\Omega_1, \mathcal{F}_1) \rightarrow (\Omega_2, \mathcal{F}_2)$ *is called* $(\mathcal{F}_1 - \mathcal{F}_2)$ *measurable if the inverse image* $f^{-1}(A) \in \mathcal{F}_1 \, \forall A \in \mathcal{F}_2$. *If* $\Omega_2 = \mathbb{R}$ *and* $\mathcal{F}_2 = \mathcal{B}(\mathbb{R})$ *then f is said to be* Borel measurable.

Definition A13 (Random Variables and Random Sequences). *A* random variable *X on a probability space* $(\Omega, \mathcal{F}, P)$ *is a Borel measurable function from* Ω *to* $\mathbb{R}$. *Similarly, a sequence of random variables* $\{X_1, X_2, \ldots\}$ *is called a discrete random process.*

Remark A3. *A function* $f : (\mathbb{R}, \mathcal{U}) \rightarrow (\mathbb{R}, \mathcal{U})$ *is continuous in the usual topology if* $f^{-1}(A)$ *is open for every open set* $A \in \mathcal{U}$. *Therefore, any continuous function is Borel measurable. Furthermore, a function* $f : \mathbb{R} \rightarrow \mathbb{R}$ *which is continuous almost everywhere on* $\mathbb{R}$ *(i.e., except on a set of measure zero) is also Borel measurable. As another example, the unit step function on a compact set* $S \subsetneq \mathbb{R}$, *which is discontinuous in* $\mathbb{R}$ *in the usual topology, is also a Borel-measurable function.*

Now, we introduce the concept of the expected value of a random variable in a probability measure space $(\Omega, \mathcal{F}, P)$. Given a random variable X, the expected value of X is denoted as $E[X]$ (see [25] or [30]). Along this line, two random variables X and Y on $(\Omega, \mathcal{F}, P)$ are said to be equal in the mean square (*ms*) sense, denoted as: $X \overset{ms}{=} Y$ if $E[|X - Y|^2] = 0$ (see [25]). Similarly, two random variables X and Y on are said to be equal in the almost sure (*as*) sense, denoted as: $X \overset{as}{=} Y$, if $X(\zeta) \neq Y(\zeta)$ is allowed $\forall \zeta \in S \subseteq \Omega$ such that $P[S] = 0$.

Given a random process $x(t)$, the *autocorrelation* is defined as:

$$r_x(t, \tau) \triangleq E[x(t) x^{her}(\tau)] \tag{A1}$$

and the *autocovariance* is defined as

$$c_x(t, \tau) \triangleq E[(x(t) - E[x(t)]) \, (x(\tau) - E[x(\tau)])^{her}] \tag{A2}$$

where the superscript *her*, called *Hermitian*, indicates the complex conjugation of a complex variable, or the complex conjugation of transpose of a complex vector/matrix.

A random process $x(t)$ is called stationary (in the strict sense) if its statistics are not affected by a time translation [25], i.e., $x(t)$ and $x(t + \varepsilon)$ have the same statistics for any real number ε. A random process $x(t)$ is said to be *wide-sense stationary* [7,25] if

1. The expected value $E[x(t)]$ is a constant for all t;
2. The autocorrelation $r_x(t, \tau)$ depends only on the difference $(t - \tau)$, not explicitly on both t and τ.

References

1. Bachman, G.; Narici, L. *Functional Analysis*; Academic Press: New York, NY, USA, 1966.
2. Naylor, A.; Sell, G. *Linear Operator Theory in Engineering and Science*, 2nd ed.; Springer-Verlag: New York, NY, USA, 1982.

3. Rudin, W. *Real and Complex Analysis*; McGraw-Hill: Boston, MA, USA, 1987.

4. Royden, H. *Real Analysis*, 3rd ed.; Macmillan: New York, NY, USA, 1989.

5. Kreyszig, E. *Introductory Functional Analysis with Applications*; John Wiley & Sons: Hoboken, NJ, USA, 1978.

6. Bobrowski, A. *Functional Analysis for Probability and Stochastic Processes*; Cambridge University Press: Cambridge, UK, 2005.

7. Hayes, M. *Statistical Digital Signal Processing and Modeling*, 1st ed.; Wiley: Hoboken, NJ, USA, 1996.

8. Haykin, S. *Adaptive Filter Theory*, 4th ed.; Prentice Hall: Englewood Cliffs, NJ, USA, 2002.

9. Farhang-Boroujeny, B. *Adaptive Filters Theory and Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2003.

10. Bressan, A. *Lecture Notes on Functional Analysis with Applications to Linear Partial Differential Equations*; American Mathematical Society: Providence, RI, USA, 2013.

11. Reed, M.; Simon, B. *Methods of Modern Mathematical Physics Part 1: Functional Analysis*; Academic Press: Cambridge, MA, USA, 1980.

12. Luenberger, D. *Optimization by Vector Space Methods*; John Wiley & Sons: Hoboken, NJ, USA, 1969.

13. Desoer, C.; Vidyasagar, M. *Feedback Systems: Input-Output Properties*; Academic Press: Cambridge, MA, USA, 1975.

14. Therrien, C. *Discrete Random Signals and Statistical Signal Processing*; Prentice Hall: Englewood Cliffs, NJ, USA, 1992.

15. Proakis, J.; Manolakis, D. *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd ed.; Macmillan Publishing Company: New York, NY, USA, 1998.

16. Oppenheim, A.; Schafer, R. *Discrete-Time Signal Processing*; Prentice Hall: Englewood Cliffs, NJ, USA, 1989.

17. Vapnik, V.; Izmailov, R. Rethinking statistical learning theory: Learning using statistical invariants. *Mach. Learn.* **2019**, *108*, 381–423. [CrossRef]

18. Ghalyan, N.F.; Ray, A. Symbolic Time Series Analysis for Anomaly Detection in Measure-invariant Ergodic Systems. *J. Dyn. Syst. Meas. Control.* **2020**, *142*, 061003. [CrossRef]

19. Ghalyan, N.F.; Ray, A. Measure invariance of symbolic systems for low-delay detection of anomalous events. *Mech. Syst. Signal Process.* **2021**, *159*, 107746. [CrossRef]

20. Lorch, E. *Spectral Analysis*; Oxford University Press: New York, NY, USA, 1962.

21. Kaiser, G. *A Friendly Guide to Wavelets*; Birkhauser: Boston, MA, USA, 1994.

22. Mallat, S. *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed.; Academic Press: Amsterdam, The Netherlands, 2009.

23. Ray, A. On State-space Modeling and Signal Localization in Dynamical Systems. *ASME Lett. Dyn. Syst. Control.* **2022**, *2*, 011006. [CrossRef]

24. Vetterli, M.; Kovacevic, J. *Wavelets and Subband Coding*; Prentice-Hall, Inc.: Hoboken, NJ, USA, 1995.

25. Stark, H.; Woods, J. *Probability and Random Processes with Applications to Signal Processing*; Prentice-Hall: Upper Saddle River, NJ, USA, 2002.

26. Helstrom, C. *Elements of Signal Detection and Estimation*; Prentice Hall: Englewood Cliffs, NJ, USA, 1995.

27. Ash, R. *Real Analysis and Probability*; Academic Press: Boston, MA, USA, 1972.

28. Munkres, J. *Topology*, 2nd ed.; Prentice-Hall: Upper Saddle River, NJ, USA, 2000.

29. Shilov, G. *Elementary Real and Complex Analysis*; Dover Publication Inc.: Mineola, NY, USA, 1996.

30. Papoulis, A. *Probability, Random Variables, and Stochastic Processes*, 2nd ed.; McGraw-Hill, Inc.: Boston, MA, USA, 1984.