Proceedings of the 25th Chinese Control Conference
7–10 August, 2006, Harbin, Heilongjiang

Efficient Data Representations for Signal Processing and Control: "Making Most of a Little"

Graham C. Goodwin
with
Milan S. Derpich and Daniel E. Quevedo
School of Electrical Engineering and Computer Science, The University of Newcastle, Australia
E-mail: graham.goodwin@newcastle.edu.au

Abstract: In order that signals can be stored, transmitted or processed it is necessary that they first be converted into digital form. This, in turn, raises the problem of how to digitize data so as to achieve the best trade-off between data load and performance, i.e., "how to make the most out of a little". Two issues are involved in this problem, namely temporal quantization (i.e., sampling) and spatial quantization. These two problems have traditionally been addressed separately. Indeed, there exists substantial literature dealing with the temporal quantization problem, covering both band-limited and non-band-limited signals. The usual underlying paradigm is that of an analysis filter, followed by a sampler, followed by a reconstruction filter. Various parts of this architecture can be optimized once other parts have been specified. On the other hand, spatial quantization has been studied extensively for a given sampling strategy, particularly in the framework of sigma delta conversion. Finally, it is also possible to formulate the joint design problem for sampling and spatial quantization. This typically leads to enhanced performance compared to that achievable by considering the two aspects separately.
This paper will survey the general area of sampling and quantization and analyze methods for achieving efficient data representations for signal processing and control applications. We will show how, on the one hand, contemporary control theory can contribute to the design of sampling and quantization systems and, on the other hand, how these systems impact on the performance of modern feedback control systems.

Key Words: Sampling, quantization, frames, model predictive control, constrained control, networked control systems

1 INTRODUCTION

We live in a data rich world. Most technological systems operate by first converting continuous time, continuous amplitude signals from the analog world into digital representations. This is a necessary precursor to allow signals to be stored, transmitted and processed without degradation other than that introduced by the analog-to-digital conversion itself.
The above was indeed the motivation that led Alec Reeves to invent *pulse-code modulation* (PCM) seven decades ago [1]. In his 1938 patent [2], Reeves highlighted the main benefits of PCM, namely:

1. Quality depends only on conversion steps.
2. Quality is independent of transmission media.
3. Low cost.
4. Compatibility with different media and traffic.
5. New features can easily be embedded.

These are remarkable statements for the time they were formulated. Indeed, most of these benefits have only become reality in recent times. Furthermore, the validity of the first two claims began to be formally determined years after they were formulated, and is still subject of ongoing research. In the pursuit of better quality at lower bit-rates (and lower costs), increasingly parsimonious methods are continually developed so as to acquire, process and represent signals digitally.

This topic has also motivated important theoretical results, from areas such as information theory, functional analysis, optimization, communication theory, frames, wavelet theory, etc.. As we will discuss in this paper, also control theory has much to contribute to this circle of ideas. Conversely, much of the theory and techniques from digital signal processing are highly relevant to several aspects of control, e.g., networked control, where parsimonious signal representation is a key element, see, e.g., [3][4][5].

In the present work we present some of the main strategies of sampling, quantization and reconstruction of analog, continuous-time signals. We will describe reconstruction quality and relate it to design constraints such as filter complexity, data-rate and sampling frequency. We also present some ideas concerning the joint problem of sampling-quantization, on one side, and reconstruction on the other. We limit our analysis to uniform sampling of scalar signals, sampling and reconstruction by single filters

(as opposed to filter-banks), quantizers with scalar output and we will not discuss any issues related to further symbol encoding.

The layout of the remainder of the paper is as follows: Section 2 presents the basics of PCM quantization and discusses some of the shortcomings that justify the introduction of a more general model for a sampling-quantization-reconstruction system. Section 3 poses the sampling and reconstruction processes in a frame theoretic perspective. Section 4 is a review of some recent generalized results on the sampling and reconstruction problem. In Section 5 we present some basic aspects of scalar memory-less quantization and oversampling. Section 6 describes feedback quantizers. In particular, some of the basic principles of predictive and noise shaping ($\Sigma\Delta$) analog-to-digital converters are presented. In Section 7 we present noise shaping quantizers that generalize $\Sigma\Delta$ converters based on model predictive control. Section 8 gives elements to analyze the joint problem of the quantization and sampling-reconstruction design, including some recent results and insights. In Section 9 we show how concepts related to sampling and quantization can be utilized in control problems. Section 10 draws conclusions. Finally, an Appendix is included with some of the basic concepts of frame theory necessary to understand several of the results presented in the main body of the paper.

2 AD – CONVERSION FUNDAMENTALS

In this section we will first describe PCM as a basic architecture used in AD–conversion applications. Various shortcomings of PCM will then motivate us to introduce later a more general framework.

2.1 Basic PCM Scheme

We consider the (simple) and idealized PCM system represented by the block diagram in Fig. 1.

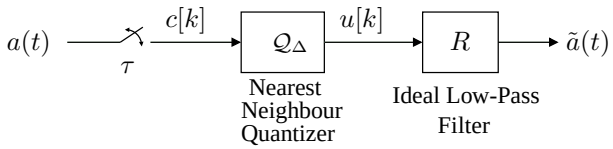

Figure 1: PCM system with ideal low-pass reconstruction filter.

The usual paradigm associated with this setup is that the input signal $a(t)$, $t \in \mathbb{R}$, is taken to be band-limited to some frequency, say, f_{max} [Hz]. Then, in accordance with the Shannon-Whittaker sampling theorem [6], the sampling step is chosen as $\tau = 1/(2f_{max})$. Since the input signal is directly sampled, we have $c[k] = a(\tau k)$, $\forall k \in \mathbb{Z}$.

The nearest neighbour scalar quantizer in Fig. 1 corresponds to the non-linear transfer function $\mathcal{Q}_\Delta(\cdot)$, defined by[1]

$$\mathcal{Q}_\Delta(\alpha) \triangleq \lceil \alpha/\Delta \rceil - \tfrac{\Delta}{2}, \quad \forall \alpha \in \mathbb{R} \tag{1}$$

where $\Delta > 0$ is the *quantization step* (see Fig. 2) and $\lceil \alpha/\Delta \rceil$ denotes rounding to the closest integer value greater than α/Δ.

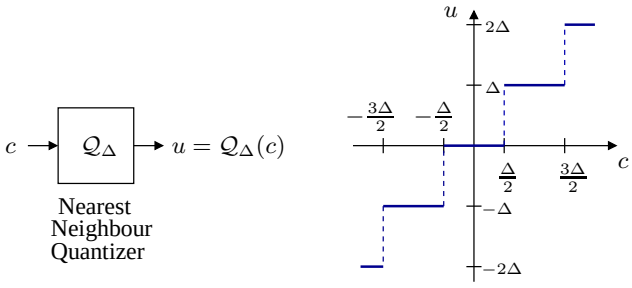

Figure 2: Nearest Neighbour Scalar Quantizer.

Thus, the output of $\mathcal{Q}_\Delta$ in Fig. 2 is the sequence of quantized values $\{u[k]\}_{k\in\mathbb{Z}}$, where

$$u[k] = \mathcal{Q}_\Delta(c[k]), \quad \forall k \in \mathbb{Z}. \tag{2}$$

The synthesis filter R in Fig. 1 is, in the simplest case, an ideal continuous time low-pass filter with cut-off frequency $f_{max} = 1/(2\tau)$ [Hz] and impulse response $\text{sinc}(2f_{max}t)$, $t \in \mathbb{R}$, where $\text{sinc}(x) \triangleq \sin(\pi x)/(\pi x)$. The output of R is the analog, continuous time signal $\tilde{a}$, given by the *mixed convolution*[2]

$$\tilde{a}(t) = \sum_{k\in\mathbb{Z}} u[k]\text{sinc}(2f_{max}t - k), \quad \forall t \in \mathbb{R} \tag{3}$$

If there were no quantization (i.e., if $\Delta = 0$), then $u[k]$ would equal $c[k]$ for all k. In this situation, $\tilde{a}(t)$ in (3) would equal *exactly* the input $a(t)$ for all $t \in \mathbb{R}$, since, by virtue of the Shannon-Whittaker sampling theorem [6], if $a(t)$ is band-limited to f_{max}, it can be reconstructed from samples by the interpolation formula

$$a(t) = \sum_{k\in\mathbb{Z}} a(k\tau)\text{sinc}(2f_{max}t - k), \quad \forall t \in \mathbb{R}. \tag{4}$$

In the presence of quantization, it is generally no longer true that $\tilde{a} = a$. Nevertheless, it is reasonable to expect that, if the quantization step is small, then the quantized samples $\{u[k]\}_{k\in\mathbb{Z}}$ will be close to the analog samples $\{a(k\tau)\}_{k\in\mathbb{Z}}$ for all k, and the output of the simple PCM system of Fig. 1 will be close (in some sense) to the analog input a. Unfortunately, this and other assumptions in the above model are often far from realistic, as discussed next.

2.2 Practical Aspects of PCM

Whilst the PCM method described above is certainly attractive, it suffers from several shortcomings that hinder its usefulness in many practical situations. In what follows, we will describe some of the main deficiencies of this architecture.

[1] In practice, all quantizers are subject to overload, i.e., there exists a saturation limit $M > 0$, such that $|\mathcal{Q}_\Delta(\alpha)| = M$, $\forall \alpha > M - \frac{\Delta}{2}$.

[2] The reconstruction formula is often written as a continuous-time convolution with input the sequence of impulses $\{u[k]\delta(t - k\tau)\}_{k\in\mathbb{Z}}$, yielding the expression in (3).

Synthesis Filter The ideal low-pass filter used in Fig. 1 for synthesis cannot be implemented in practice. Firstly, it is non-causal. A very close approximation of the ideal low-pass filter would still be non-causal, which rules it out from any delay sensitive application.
Secondly, an ideal low-pass filter has an infinite impulse response length. For practical low pass filters, the closer they mimic the ideal filter, the longer the impulse response will be. One problem with a long, slow decaying impulse response is that it affects the *stability of the reconstruction*, in the sense that bounded errors in the samples are able to produce unbounded point-wise error in the reconstructed output. As an example, consider the ideal low-pass reconstruction in (4). It is easy to show that any bounded periodical error in the samples $a(k\tau)$ of the form $\left\{\rho(-1)^k\right\}_{k\in\mathbb{Z}}$, with $|\rho| > 0$, will yield an unbounded reconstruction error in the L^∞ norm. The second difficulty with a synthesis filter with long (but finite) impulse response is cost and complexity: In applications where synthesis is accomplished via discrete-time FIR filters, longer impulse responses require higher computational complexity.
Another problem with the ideal-low pass synthesis filter model is that, in many practical applications, the synthesis filter is not a design choice, but is prescribed by other considerations. In such cases, the synthesis filter can have almost any frequency response. An important example of this situation is that of sampled-data control systems, where the plant itself can be thought of as comprising part of the synthesis filter R in Fig. 1. We will return to this situation later in Section 9.

Not Necessarily Band-Limited Input Signals The assumption of band-limitedness of the input signal a is also very restrictive. Most real applications have to deal with signals over a finite time interval (strictly speaking, any non-zero finite duration signal is not band-limited [7]). Even when processing a virtually infinite duration, perfectly band-limited signal, only a finite number of samples can be used for the reconstruction. This introduces truncation errors [8], i.e., part of the inter-sample behaviour of the input signal is not captured by the samples. On the other hand, it is often the case that the sampling rate cannot be made high enough to completely avoid aliasing. Whilst this is commonly dealt with by using a low-pass anti-aliasing filter before sampling, this paradigm may have significant shortcomings whenever the signal carries relevant information in the high frequency part of its spectrum, or when the reconstruction filter is not perfectly band-limiting (see, e.g., [9, 10]). In this case other types of analysis filters should be considered.

Availability of the Input Signal Before being able to sample the value of any physical variable, it is necessary to convert it to an electrical signal by means of a transducer, which in itself is a dynamical system. It is often the case that sampling is performed in the transducer itself. In this case, one does not have access to the underlying continuous time signal, but only to the samples taken. Depending on the situation, this can deprive further stages of knowledge of important inter-sample behaviour of the physical variable. It is then necessary to make a wise design of the synthesis stage, so that the input signal can be well approximated at the output (see, e.g., [11, 12, 13]).

Quantization, Sampling Frequency and Data-Rate In the simple PCM system of Fig. 1, quantization is done element-wise by a nearest neighbour quantizer, see (2). Thus, if one wishes to obtain a small reconstruction error, one would naturally aim at reducing the quantization step. In practice, however, the reduction of Δ is limited by cost and structural constraints. Alternatively, if the statistics of the input signal are known, then the mean square reconstruction error can often be reduced by using a quantizer in which the quantization step is not uniform along its dynamic range.
Moreover, even though the Shannon-Whittaker sampling theorem shows that when the samples are un-quantized an increase of the sampling frequency cannot improve reconstruction (since it is already perfect), the situation with quantized coefficients is different. More precisely, when quantization is introduced, sampling above the Nyquist rate (oversampling) can be utilized to reduce quantization error (see Sec. 5.2) . Thus, one often has the chance to compensate the effects of coarse magnitude quantization by means of a finer time quantization, i.e., faster sampling rate. (The reader may be well aware of this in 1 bit DAC's used in some CD players.)
In practice, the product of the sampling rate and the number of quantization levels is often constrained by data-rate limitations. This is so because, although not explicitly shown in Fig. 1, the sequence of quantized values $\{u[k]\}_{k\in\mathbb{Z}}$, in binary form, has to be stored or transmitted before reconstruction takes place at another location in time and space. This means that the total number of bits, or similarly, the data-rate, is limited. In principle, if the quantizer has $n_{\mathbb{U}} \in \mathbb{N}$ levels, then the data-rate will be approximately given by

$$\text{Bit Rate} \triangleq \frac{\log_2 n_{\mathbb{U}}}{\tau} \quad \text{[bits/s]}. \tag{5}$$

It is possible, however, to reduce the data-rate by an efficient encoding of the sequence of quantized values (compression). When such encoding is applied, the data-rate limitation translates into an information-rate limitation, precisely given by the entropy of the sequence of symbols at the output of the quantizer [14]. Systems with entropy coding are also called *variable-rate* encoders. In this paper, however, we will not consider such coding methods. Thus, we will only consider *fixed-rate* encoding, and the data-rate will be given by (5).

2.3 A More General Model for AD–Conversion

In view of the limitations of PCM conversion discussed above, a more general model for the analysis of sampling, quantization and reconstruction systems is presented in Fig. 3.

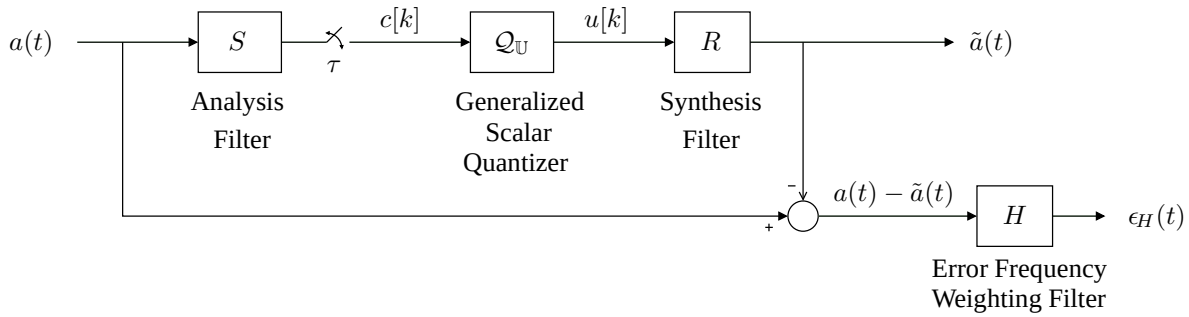

Figure 3: A more general sampling, quantization and reconstruction system.

For the remainder of this work, we will restrict our analysis to input signals a which are modeled as finite energy scalar functions of a single parameter t (i.e., $a \in L^2(\mathbb{R})$). For example, we could think of t as denoting time, for the case of time varying scalar signals. Thus, the analysis filter S in Fig. 3 accounts for all the continuous-time linear processing of the input that occurs before the sampling takes place. The sampling process is assumed uniform (i.e., regular sampling), with fixed sampling interval τ.

The synthesis filter R in Fig. 3 represents the linear processing in continuous-time (possibly with some discrete-time pre-filtering) applied to the quantized samples $\{u[k]\}_{k\in\mathbb{Z}}$. The output of R is denoted as $\tilde{a}$. It approximates a in some well defined sense.

The quantizer $\mathcal{Q}_{\mathbb{U}}$ in Fig. 3 is labeled *generalized* because it is allowed to have access to previous and future input samples during operation, and *scalar*, because it generates a sequence of scalars, one at a time. We will only discuss quantizers of this type in the remainder of this paper, which justifies a more precise definition of the class of *generalized scalar quantizers*:

Definition 1 (Generalized Scalar Quantizers)**.** *Any quantization strategy that can be devised within the following conditions*

- *The quantizer has no access to the continuous time signal a, but only to the samples $\{c[k]\}$.*
- *The quantizer outputs a sequence of scalars $\{u[k]\}$ at a constant rate, one element every τ units of time. The total elements in the output sequence equals the number of input analog samples.*
- *Each of the elements in the output sequence of the quantizer can take values only from a finite, given and fixed set of scalars $\mathbb{U}$, i.e, the output of the quantizer satisfies*
$$u[k] \in \mathbb{U}, \quad \forall k \in \mathbb{Z}. \tag{6}$$
- *The quantizer has access to all past and future analog samples.*

is said to belong to the class of Generalized Scalar Quantizers.

Note that this definition allows for the uniform, nearest neighbour scalar quantizer in (2) as a special case. The last condition in Definition 1 means that the generalized scalar quantizer in Fig. 3 is allowed, in principle, to determine the output $u[\ell]$, for any $\ell \in \mathbb{Z}$, based upon knowledge of the *entire* sequence $\{c[k]\}_{k\in\mathbb{Z}}$, i.e., it is a dynamic system. Therefore scalar quantizers with memory (such as the predictive and noise shaping quantizers to be discussed in Section 6) are special realizations of the generalized scalar quantizer [3].

In Fig. 3 an *error frequency weighting* filter H has been added. Inclusion of this filter reflects the fact that, depending on the application, the practical impact (or *cost*) of the reconstruction error is frequency dependent. Accordingly, H filters the instantaneous error $a(t) - \tilde{a}(t)$ to produce a *frequency weighted* error signal $\epsilon_H(t)$.

Based on the general setup illustrated in Fig. 3, throughout the remainder of this work the performance of the system will be assessed in terms of the squared L^2 norm of the generated signal ϵ_H:

$$\|\epsilon_H\|_{L^2}^2 \triangleq \int_{-\infty}^{\infty} (\epsilon_H(t))^2 \, dt. \tag{7}$$

3 SAMPLING AND RECONSTRUCTION FROM A FRAME THEORETIC PERSPECTIVE

As mentioned above, a paradigm which underlies many signal processing schemes consists of a *pre-filtering* (or *analysis*) stage, a sampling stage, a digital, discrete-time processing stage and a *post-filtering* (also referred to as *synthesis* or *reconstruction*) stage. It has been shown that these processes are equivalent to a sequence of mappings between Hilbert spaces (see, for example, [18, 19, 20] and [9]). This viewpoint allows one to use the powerful tools of

[3] The possibility of quantization based on *all* the future samples also admits a restricted class of vector quantizers. In this class, the reproduction codebook is restricted to be the set $\mathbb{U}^{|\mathbb{Z}|} \subsetneq \mathbb{R}^{|\mathbb{Z}|}$, where $|\mathbb{Z}|$ denotes the cardinality of the integers $\mathbb{Z}$. The asymptotic performance of infinite length vector quantizers has been the subject of intensive research, although traditionally with a different choice of reconstruction codebook, see, e.g., [15, 16, 17]. However, vector quantization becomes impractical for long vectors and large reproduction codebooks [17]. Quantizers suitable for on-line applications, based on a finite number of future samples, will be discussed later, in Section 7.

Hilbert spaces, frames and algebra of operators to study and design sampling and reconstruction systems. It allows for elegant solutions to otherwise complex design optimization problems, by using inner products and projection operators.

3.1 Historical notes

To the best of our knowledge, the first author to apply Hilbert spaces theory to the sampling problem was F. Beutler in 1961. In [21] he derived sampling theorems for random stationary processes using complex exponential Fourier expansions. Further insight and results for band-limited signals were provided by K. Yao in 1967 for other expansions, see [22]. Several publications with the Hilbert space approach to the sampling problem followed in subsequent years. Among others, a 1986 paper by Hidemitsu Ogawa [23] presented a unified approach to generalized sampling theorems. It introduced the idea of regarding the approximation of signals in a more general, finite dimensional reconstruction space, instead of restricting to perfect reconstruction by Fourier expansions. Interestingly, in [23], a finite number of samples of a filtered signal was used, as opposed to an infinite number of "raw" samples. By the early nineties, the recently arrived wavelet theory [24, 25] began to stimulate a strong revival of sampling theory (see, for example, [26, 27, 28]), by using the mathematics of basis and frames in Hilbert spaces. This framework allowed for the re-formulation of the sampling and reconstruction problem in more general and practical situations, including, inter alia, sampling and reconstruction from finite samples [23, 29], study of arbitrary input and reconstruction spaces [11, 30, 31], sampling of non-band-limited signals [27, 10], oversampling [32, 33], non-uniform sampling [34, 18], filter-banks [35, 36], and splines and interpolation [37, 38].

In the remainder of this section we will derive a representation of the sampling and reconstruction processes in a Hilbert space frame theoretic context[4]. For a more complete formal analysis, see, for example, [18, 9, 34, 28, 20, 39].

3.2 Sampling and Reconstruction as Frame Operators

It will be shown next that the analysis and sampling stages, which map continuous time signals into discrete time sequences, can be seen as the *analysis operator* of the sampling frame. This frame is made of translates of the time reversed impulse response of the analysis filter S. Similarly, the reconstruction process, which maps discrete time sequences into continuous time signals, can be seen as the *synthesis operator* of a reconstruction frame. It is made of translates of the impulse response of the reconstruction filter R.

Filtering and Sampling Consider the input signal a in the block diagram of Fig. 3. Assume that a is known to belong to some space of signals, say $\mathcal{A} \subseteq L^2$. Let y be the output of the analysis filter S, which has impulse response $\varphi(t) \in L^2$. Then, $y(t)$ is given by the convolution:

$$y(t) \triangleq (a * \varphi)(t) = \int_{-\infty}^{\infty} a(z)\varphi(t-z)dz, \quad \forall t \in \mathbb{R}.$$

If one now creates a sequence $c[k] \in \ell^2$ by taking the values of $y(t)$ at time instants $t = k\tau$, $k \in \mathbb{Z}$ (sampling process in fig. 3), one obtains

$$\begin{aligned} c[k] &= y(k\tau) = (a * \varphi)(k\tau) \\ &= \int_{-\infty}^{\infty} a(z)\varphi(k\tau - z)dz = \int_{-\infty}^{\infty} a(z)\phi(z - k\tau)dz \end{aligned} \tag{8}$$

where $\phi(t) \triangleq \varphi(-t)$, $\forall t \in \mathbb{R}$. One can see that the last integral in (8) corresponds to the conventional inner product in L^2, defined in (54), between $a(t)$ and $\phi(t - k\tau)$. If we now define the *shift operator* $T_{k\tau}$ by

$$T_{k\tau}\phi(t) \triangleq \phi(t - k\tau)\,,\ t \in \mathbb{R}, k \in \mathbb{Z}, \tag{9}$$

then it is possible to write (8) as

$$c[k] = (a * \varphi)(k\tau) = \langle a, T_{k\tau}\phi \rangle_{L^2}, \quad \forall k \in \mathbb{Z}. \tag{10}$$

Therefore, the sampled filtered input signal can be seen as the result of a sequence of inner products. From (10) and Definition 6 (see Appendix), this is indeed the process described by the analysis operator Φ^* associated to the frame $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$. As a consequence:

$$\Phi^* : L^2 \mapsto \ell^2\,, \quad \Phi^* a = \{c[k]\}_{k\in\mathbb{Z}}$$

Notice that, since $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$ is a frame for the Hilbert space

$$\mathcal{S} \triangleq \overline{\text{span}}\,\{T_{k\tau}\phi\}_{k\in\mathbb{Z}} \subset L^2,$$

it follows that $c \in \ell^2$ for all $a \in L^2$, as required[5].

Synthesis (or Reconstruction) Consider now the conversion from the discrete-time sequence $\{u[k]\}_{k\in\mathbb{Z}}$ to the continuous-time signal $\tilde{a}(t)$, see Fig. 3. If we denote the impulse response of R as $\psi(t)$, then the band-limited Shannon-Whittaker reconstruction scheme in (3) can be generalized to:

$$\tilde{a}(t) = \sum_{k\in\mathbb{Z}} u[k]\psi(t - k\tau) = \sum_{k\in\mathbb{Z}} u[k]T_{k\tau}\psi,\ \forall t \in \mathbb{R} \tag{11}$$

It is clear from Definition 5 (Appendix) that the reconstruction process (11) can be represented by the synthesis operator Ψ associated with the frame $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$:

$$\Psi : \ell^2 \mapsto \mathcal{W}\,, \quad \Psi u = \tilde{a}$$

[4] For completeness, we have included an introduction to bases and frames in Hilbert spaces in an appendix.

[5] More precisely, if B is the upper frame bound for $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$, then $\|c\|^2 \leq B\,\|a\|^2$, see (56).

In this new setting, $\psi(t)$ becomes the generating function for the principal shift invariant reconstruction space $\mathcal{W} \triangleq \overline{\text{span}}\,\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$, which is, in general, different from the space of band-limited signals[6].
The sum in (11) can be seen as a *mixed convolution* [9], i.e.,

$$\tilde{a}(t) = (u * \psi)(t), \quad \forall t \in \mathbb{R}$$

which it takes a discrete time sequence u and a continuous time function ψ, yielding a continuous time function $\tilde{a}$.
If the impulse response $\psi(t)$ is chosen such that $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ is a Bessel sequence (and therefore a frame for $\overline{\text{span}}\,\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$, see (56)), then Ψ is a bounded operator, and the output $\tilde{a}(t) = \Psi u \in L^2$ for all sequences $\{u[k]\}_{k\in\mathbb{Z}} \in \ell^2$.

The Combined Sampling and Reconstruction Process
It follows from the above that the sampling (analysis) and reconstruction (synthesis) process can be stated as a sequence of operators between Hilbert spaces:

- Analysis:

$$\Phi^* : L^2 \mapsto \ell^2, \quad c = \Phi^* a.$$

  In particular, $c[k] = \langle a, T_{k\tau}\phi\rangle$, $\forall k \in \mathbb{Z}$.

- Reconstruction:

$$\Psi : \ell^2 \mapsto \mathcal{W}, \quad \tilde{a} = \Psi u$$

Therefore, in the absence of quantization (i.e., if $u[k] = c[k]$, $\forall k \in \mathbb{Z}$), the complete process can be expressed as

$$\Psi\Phi^* : L^2 \mapsto \mathcal{W}, \quad \tilde{a} = \Psi\Phi^* a \tag{12}$$

If the sequence $\{u[k]\}_{k\in\mathbb{Z}}$ is obtained by quantization of $\{c[k]\}_{k\in\mathbb{Z}}$, then (12) becomes

$$\Psi\mathcal{Q}_{\mathbb{U}}\Phi^* : L^2 \mapsto \mathcal{W}, \quad \tilde{a} = \Psi\mathcal{Q}_{\mathbb{U}}(\Phi^* a)$$

It is interesting to note that the above results allow one to determine the ultimate limitations and capabilities of a sampling and reconstruction system in terms of the Hilbert spaces related to sampling rate and filters. More precisely, the analysis and synthesis filters alone determine, respectively, the largest class of signals that can be sensed (i.e., the sampling space) and the largest class of signals that can be generated (i.e., the reconstruction space). A rather remarkable implication is that in the intermediate (discrete-time) stages one can only design the mapping between these spaces, *but not expand the sampling and reconstruction spaces themselves*.
As a consequence, the design of an AD conversion scheme can be thought of as involving two aspects, namely:

1. Choice of the sampling and reconstruction filters (i.e., choice of spaces).
2. Design of the mapping between signals in the sampling space and signals in the reconstruction space (i.e., design of discrete-time processing, including quantization).

In what follows, we will describe aspects of the separate design of the sampling/reconstruction strategy and of the quantization method. Some aspects of the joint design problem will be discussed later in Section 8.

4 SAMPLING AND RECONSTRUCTION WITHOUT QUANTIZATION

In this section we discuss the effect that analysis and synthesis filters have on the reconstruction quality. We will assume that the input and output spaces are given and will neglect quantization effects. The implicit trade-off here is between the quality of the reconstruction and the computational complexity (and delay) incurred in the sampling and reconstruction processes.

4.1 Types of Reconstruction

As concluded in Section 3, the ultimate sampling and reconstruction capabilities of a system are limited by the sampling and reconstruction spaces. These, in turn, are entirely determined by the choice of analog filters S and R, as well as the sampling interval τ. This suggests that, whenever possible, the design of S and R should focus mostly on the sampling and reconstruction spaces that one wishes to obtain. Further refinement can be achieved by careful design of discrete-time filters which can be located right after the analysis filter S and before the synthesis filter R, see Fig. 3. Interestingly enough, it has been shown that, in general, the optimal mapping is obtained by making the sampling and reconstruction frames duals of one another [9, 40]. To achieve this for a given analysis frame , one can insert a discrete-time correction filter before the synthesis filter to make the synthesis frame the dual of the analysis frame. Although, in general, the dual frame of some given frame is not unique, there exists only one shift-invariant dual frame (i.e., a unique correction filter) for each given shift-invariant frame [40]. In what follows, we will consider the following situation:

- $\mathcal{H}$ is a non-separable Hilbert space (e.g., $L^2(\mathbb{R})$).
- $\mathcal{A} \subseteq \mathcal{H}$ is the space that contains all possible input signals.
- $\mathcal{S} = \overline{\text{span}}\,\{T_{k\tau}\phi\}_{k\in\mathbb{Z}} \subset \mathcal{H}$ is the sampling space[7].
- $\mathcal{W} = \overline{\text{span}}\,\{T_{k\tau}\psi\}_{k\in\mathbb{Z}} \subset \mathcal{H}$ is the reconstruction space.

Depending on the relation between the input space $\mathcal{A}$, sampling space $\mathcal{S}$ and reconstruction space $\mathcal{W}$, we will consider three types of reconstruction notions, namely: consistent, orthogonal and perfect reconstruction.

[6] Note that this space is of countable dimension.

[7] Notice that both $\mathcal{S}$ and $\mathcal{W}$, being of countable dimension, can never be equal to an infinite, non-separable space such as $\mathcal{H}$.

Consistent Reconstruction The first and most generally attainable reconstruction goal is that of *consistent reconstruction*, first introduced in 1994 by Unser and Aldroubi, see [11][8]. A signal approximation is said to be consistent if it yields the same samples (observations) as the original signal when re-injected into the system, i.e. $\tilde{a} \in \mathcal{W}$ is a consistent approximation of $a \in \mathcal{A}$ if and only if

$$\Phi^* \tilde{a} = \Phi^* a.$$

The idea of consistent reconstruction is depicted in Fig. 4.a); in this figure, $\tilde{a}$ is projected onto $\mathcal{W}$ along $\mathcal{S}^\perp$, the null space of $\mathcal{S}$, see (49).

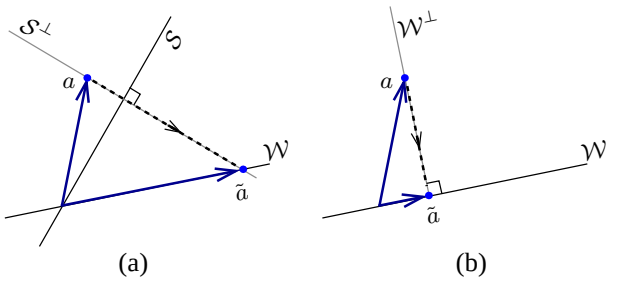

Figure 4: a) Consistent reconstruction (oblique projection); b) MSE reconstruction (orthogonal projection).

The notion of consistent reconstruction was first introduced for Riesz bases in [11], and then extended for frames in [31, 19, 13, 41, 40] and [42].

Orthogonal Reconstruction The second type of reconstruction is *orthogonal reconstruction*, also called minimum mean squared error (MMSE) reconstruction. It requires additional conditions (see next section). In this type of reconstruction, the system generates, for any input $a \in \mathcal{A}$, the output $\tilde{a} \in \mathcal{W}$ that minimizes $\|a - \tilde{a}\|_{L^2}$, i.e.:

$$\tilde{a} = \arg\min_{w \in \mathcal{W}} \|a - w\|_{L^2}.$$

It is well known that this notion is equivalent to an orthogonal projection of the signals of $\mathcal{A}$ onto the output space $\mathcal{W}$ (see Appendix A.1.1). The intuitive notion of orthogonal projection is illustrated in Fig. 4.b). Note that the $\tilde{a}$ shown in this figure is, indeed, the closest point to a in the output space $\mathcal{W}$.

Perfect Reconstruction The third, and most demanding notion is that of *perfect reconstruction*, i.e.,

$$\tilde{a} = a, \forall a \in \mathcal{A}.$$

As will be shown below, depending on the spaces $\mathcal{A}$, $\mathcal{S}$ and $\mathcal{W}$, perfect reconstruction can still be possible, even, for example, for non band-limited signals [43, 10, 27].
In the remainder of this section we will describe conditions on the sampling and reconstruction method which ensure that each of these notions can be achieved.

[8] It is worth mentioning, that Ogawa in his 1986 work [23] already referred to this concept as the *re-observation property* (of a finite sequence of samples), deriving mathematical expressions for the required synthesis method.

4.2 Conditions for Consistent, Optimal and Perfect Reconstruction

Under the assumption that the sampling and reconstruction spaces satisfy the *direct sum condition*[9]

$$\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp, \tag{13}$$

necessary and sufficient conditions have been found in order to make the sampling and reconstruction system achieve consistent reconstruction and, as particular cases, optimal and perfect reconstruction as well [19, 40, 42].
For shift invariant frames and spaces, the direct sum condition can be conveniently expressed in the frequency (Fourier) domain[10] based on the functions

$$\mathbf{A}_\psi : \mathbb{R} \mapsto \mathbb{R}, \qquad \mathbf{A}_\psi(\gamma) \triangleq \sum_{k \in \mathbb{Z}} \left| \hat{\psi}\left(\tfrac{\gamma+k}{\tau}\right) \right|^2 \tag{14}$$

$$\mathbf{A}_\phi : \mathbb{R} \mapsto \mathbb{R}, \qquad \mathbf{A}_\phi(\gamma) \triangleq \sum_{k \in \mathbb{Z}} \left| \hat{\phi}\left(\tfrac{\gamma+k}{\tau}\right) \right|^2 \tag{15}$$

and the *null sets* of $\mathbf{A}_\psi$ and $\mathbf{A}_\phi$, denoted, respectively, as $\mathcal{N}(\mathbf{A}_\psi)$ and $\mathcal{N}(\mathbf{A}_\phi)$, where

$$\mathcal{N}(f) \triangleq \{\gamma \in \mathbb{R} : f(\gamma) = 0\}, \; f : \mathbb{R} \mapsto \mathbb{R}, \tag{16}$$

by means of the following proposition:

Proposition 1 ([40, Proposition 4.8])**.** *Let $\psi, \phi \in L^2(\mathbb{R})$, and assume that $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ and $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$ are frame sequences. Then the following are equivalent:*

(i) $L^2(\mathbb{R}) = \mathcal{W} \oplus \mathcal{S}^\perp$,

(ii) $\mathcal{N}(\mathbf{A}_\psi) = \mathcal{N}(\mathbf{A}_\phi)$ *and there exists a constant* $A > 0$ *such that*

$$A \leq \left| \sum_{k \in \mathbb{Z}} \hat{\psi}(\gamma + k) \hat{\phi}^*(\gamma + k) \right|, \forall \gamma \notin \mathcal{N}(\mathbf{A}_\psi). \tag{17}$$

It is shown in [42] that, if the direct sum condition is satisfied, then $\tilde{a} \in \mathcal{W}$ is a consistent reconstruction of an input $a \in \mathcal{H}$ if and only if $\tilde{a}$ is the *oblique projection* of a onto $\mathcal{W}$ along $\mathcal{S}^\perp$, the null space of $\mathcal{S}$ (see Appendix A.1.1). Such a projector, denoted by $E_{\mathcal{W}\mathcal{S}^\perp}$, is defined as

$$E_{\mathcal{W}\mathcal{S}^\perp} : \mathcal{H} \mapsto \mathcal{W}, \qquad E_{\mathcal{W}\mathcal{S}^\perp} h = w,$$
$$\text{where } h = w + v, \text{ with } w \in \mathcal{W}, v \in \mathcal{S}^\perp$$

The following defines the concept of *oblique dual frame* and establishes its relation with the oblique projector:

Lemma 1 (from [40, Lemma 3.1])**.** *Assume that $\{f_k\}_{k\in\mathbb{Z}}$ and $\{g_k\}_{k\in\mathbb{Z}}$ are Bessel sequences in $\mathcal{H}$ and let $\mathcal{S} = \overline{\text{span}}\{g_k\}_{k\in\mathbb{Z}}$, $\mathcal{W} = \overline{\text{span}}\{f_k\}_{k\in\mathbb{Z}}$. Assume that $\mathcal{H} = \mathcal{W} \oplus \mathcal{S}^\perp$. Then the following are equivalent:*

[9] $\mathcal{S}^\perp$ is the null space of $\mathcal{S}$, see (49) in Appendix. The expression $\mathcal{H} = \mathcal{F} \oplus \mathcal{G}$ means that $\mathcal{F} \cap \mathcal{G} = \{0\}$ and that every $h \in \mathcal{H}$ can be decomposed as $f + g$, where $f \in \mathcal{F}$, $g \in \mathcal{G}$, see, e.g., [44, Def. 3.4.11, page 99].

[10] Here, and in the sequel, $\hat{\psi}$ denotes the Fourier transform of ψ defined by: $\hat{\psi}(\gamma) = \int_{-\infty}^{\infty} \psi(t) \mathrm{e}^{-2\pi i t \gamma} dt$.

a) $w = \sum_{k\in\mathbb{Z}} \langle w, g_k\rangle f_k,\ \forall w \in \mathcal{W}$.

b) $E_{\mathcal{W}\mathcal{S}^\perp} h = \sum_{k\in\mathbb{Z}} \langle h, g_k\rangle f_k,\ \forall h \in \mathcal{H}$.

c) $E_{\mathcal{S}\mathcal{W}^\perp} h = \sum_{k\in\mathbb{Z}} \langle h, f_k\rangle g_k,\ \forall h \in \mathcal{H}$.

Furthermore, if the above three equivalence conditions are satisfied, then $\{g_k\}_{k\in\mathbb{Z}}$ *is an oblique dual frame of* $\{f_k\}_{k\in\mathbb{Z}}$ *on* $\mathcal{S}$ *and* $\{f_k\}_{k\in\mathbb{Z}}$ *is an oblique dual frame of* $\{g_k\}_{k\in\mathbb{Z}}$ *on* $\mathcal{W}$.

From Lemma 1 one can see that $\Psi\Phi^*$ becomes an oblique projector if and only if $\{\phi_k\}_{k\in\mathbb{Z}}$ is an *oblique dual frame* of $\{\psi_k\}_{k\in\mathbb{Z}}$ in $\mathcal{S}$.
Although, as with the conventional case considered in Definition 7 (see Appendix A.4), the oblique dual frame within a given space is not unique, the shift-invariant oblique dual frame of a shift invariant frame *is* unique [40]. This means that, once reconstruction and sampling spaces are defined, there exists a unique analysis filter that makes the analysis frame the oblique dual of the reconstruction frame. Conversely, there exists a unique reconstruction filter that turns the reconstruction frame into the dual of the analysis frame. An expression in the Fourier domain for the oblique dual frame condition in terms of the frequency responses of the analysis and reconstruction filters is given in [40, Theorem 4.3], which, by virtue of Proposition 1, can be rewritten as follows:

Theorem 1. *Let* $\psi, \phi \in L^2(\mathbb{R})$ *and assume that* $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ *and* $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$ *are frame sequences, spanning the closed spaces* $\mathcal{W}$ *and* $\mathcal{S}$*, respectively. If* $L^2(\mathbb{R}) = \mathcal{W} \oplus \mathcal{S}^\perp$*, then the following holds:*

(i) There exists a unique function $\tilde{\psi} \in \mathcal{S}$ *such that*

$$w = \sum_{k\in\mathbb{Z}} \langle w, T_{k\tau}\tilde{\psi}\rangle T_{k\tau}\psi,\ \forall w \in \mathcal{W};$$

(ii) This unique function $\tilde{\psi} \in \mathcal{S}$ *is given in the Fourier domain by*[11]

$$\hat{\tilde{\psi}}(\gamma) = \begin{cases} \dfrac{\hat{\phi}(\gamma)}{\sum\limits_{k\in\mathbb{Z}} \hat{\psi}(\gamma+k)\hat{\phi}^*(\gamma+k)} & \text{, if } \gamma \notin \mathcal{N}(\mathbf{A}_\psi) \\ 0 & \text{, if } \gamma \in \mathcal{N}(\mathbf{A}_\psi) \end{cases} \tag{18}$$

Remark 1. *In relation to* (18)*, we note that:*

- *The function* $\hat{\tilde{\psi}}(\gamma)$ *in* (18) *is* 1*-periodic.*
- *The result in* (18) *allows one to obtain a shift invariant oblique dual frame for* $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ *on* $\mathcal{S}$ *for a given* ϕ *by inserting a continuous or discrete-time correction filter* $Q_{\phi\psi}$ *with transfer function*

$$Q_{\phi\psi}(\gamma) \triangleq \frac{\hat{\tilde{\psi}}(\gamma)}{\hat{\phi}(\gamma)}, \quad \forall \gamma \in \mathbb{R} \tag{19}$$

just before or just after the analysis filter. With such an arrangement, and provided Conditions (i) and (ii) in Theorem 1 are satisfied, the system will yield perfect reconstruction *for all inputs* $a \in \mathcal{W}$ *and* consistent reconstruction *for all inputs* $a \in L^2$*, as required.*

- *Conversely, from the reciprocity of oblique dual frames,* (18) *also allows one to obtain the oblique dual frame for* $\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$ *on* $\mathcal{W}$ *for a given* ψ*. This can be achieved by inserting a continuous (or discrete) time correction filter with transfer function* $Q_{\phi\psi}(\gamma)$ *defined in* (19) *between* S *and* R*. Notice that this correction filter does not alter the space associated to the stage in which it is inserted, i.e., if the impulse response of* $Q_{\phi\psi}$ *is* $q_{\phi\psi}(t)$*, then* $\overline{span}\{T_{k\tau}(\phi * q_{\phi\psi})\}_{k\in\mathbb{Z}} = \overline{span}\{T_{k\tau}\phi\}_{k\in\mathbb{Z}}$.

From the previous results it follows that, if the direct sum (17) and duality (18) conditions are met, then necessary and sufficient condition for each type of reconstruction can be stated as follows:

Conditions for Perfect Reconstruction Perfect reconstruction only for all inputs $a \in \mathcal{W}$ is possible, without any further requirement

Conditions for (MSE) Reconstruction (Orthogonal Projection) If, additionally, $\mathcal{S} = \mathcal{W}$, then $E_{\mathcal{W}\mathcal{S}^\perp}$ becomes an orthogonal projector onto $\mathcal{W}$, i.e., $E_{\mathcal{W}\mathcal{S}^\perp} = P_{\mathcal{W}}$, see Appendix A.1.1. This guarantees that the output signal $\tilde{a}$ will be the best approximation in $\mathcal{W}$ for the input signal $a \in \mathcal{H}$, i.e., it will minimize $\|a - \tilde{a}\|_{L^2}$.

Conditions for Consistent Reconstruction (Oblique Projection) Consistent reconstruction will be achieved for all $a \in L^2$ without further requirements.

5 QUANTIZATION

Quantization is the process of translating analog values into values which belong to a finite set. The representation of analog samples with infinite accuracy would require an infinite number of bits. Quantization allows one to achieve a controlled approximate representation of *infinite* analog values, which in turn can be represented with a finite number of bits. Hence, the main purpose of analog to digital conversion is to compress data, whilst aiming to obtain the best possible approximation of the analog signal. This is to be achieved within data-rate constraints and according to some fidelity criterion, i.e., "making most out of a little".
As already mentioned in Section 2.3, the quantizers to be discussed in this paper belong to the family of generalized scalar quantizers, see Definition 1. As such, quantizers generate an output sequence $\{u[k]\}_{k\in\mathbb{Z}}$ whose values are constrained to belong to a set of $n_{\mathbb{U}}$ elements (see (6)) , the *quantization alphabet* $\mathbb{U}$, now formally defined as:

$$\mathbb{U} \triangleq \{\mu_1, \mu_2, \ldots, \mu_{n_{\mathbb{U}}}\}, \quad \mu_i \in \mathbb{R} \tag{20}$$

[11] In (18), $\hat{\phi}^*$ denotes the complex conjugate of $\hat{\phi}$.

Traditionally, quantization has been analyzed only in terms of discrete-time performance, usually looking at the MSE between input samples and quantized samples. Denoting the input and output sequences of the quantizer as $\{c[k]\}_{k\in\mathbb{Z}}$ and $\{u[k]\}_{k\in\mathbb{Z}}$, the MSE is given by $\|c-u\|^2_{\ell^2}$:

$$\|c-u\|^2_{\ell^2} \triangleq \sum_{k\in\mathbb{Z}} \left(c[k]-u[k] \right)^2 \tag{21}$$

We will next briefly discuss the simplest realization of the generalized scalar quantizer in Fig. 3: the zero-memory scalar quantizer. Its performance will be analyzed in terms of the MSE as defined in (21). Other realizations of the generalized scalar quantizer, such as quantization with memory (by means of feedback) and quantization with memory and "preview", will be analyzed in Sections 6 and 7, respectively. For a more comprehensive analysis of quantization see, e.g., [45, 16, 17].

5.1 Scalar Quantization

Scalar quantization is also referred to as *zero-memory* quantization, since each analog sample is quantized ignoring previous or future samples. Scalar quantizers partition the real line into a set of $n_{\mathbb{U}}$ disjoint and consecutive intervals $\mathbb{I} = \{\mathrm{I}_1, \ldots, \mathrm{I}_{n_{\mathbb{U}}}\}$, $\mathrm{I}_i \subset \mathbb{R}$. A unique scalar from $\mathbb{U}$ is associated to each interval in $\mathbb{I}$, usually satisfying $\mu_i \in \mathrm{I}_i, i = 1, \ldots, n_{\mathbb{U}}$. Depending on the choice of the partition intervals, either a uniform or a non-uniform scalar quantizer is obtained.

Uniform Quantizer The simplest scalar quantizer is the nearest neighbour uniform quantizer introduced in Section 2.1, where the partition of the input space (the real line) is given by (1) and the elements of $\mathbb{U}$ satisfy

$$\mu_{i+1} - \mu_i = \Delta, \quad i = 1, \ldots, n_{\mathbb{U}} - 1$$

Defining the positive constants *extreme output value* M and *extreme input value* C as

$$M \triangleq -\mu_1 = \mu_{n_{\mathbb{U}}} \tag{22}$$
$$C \triangleq M + \Delta/2, \tag{23}$$

the quantizer is said to be *overloaded* if the input $|x| > C$. If the probability density function of the analog samples is smooth and the quantization step is small enough, then the quantization error can be approximately modeled as a random variable with uniform distribution over $[-\Delta/2, \Delta/2]$ (see [46] for precise conditions), and the mean squared error between the input x and the output $u = \mathcal{Q}_{\mathbb{U}}(x)$ of the quantizer is given by the distortion measure:

$$D \triangleq \mathrm{E}\left[\left(x - \mathcal{Q}_{\mathbb{U}}(x) \right)^2 \right] = \Delta^2/12,$$

where $\mathrm{E}[X]$ denotes the expected value of the random variable X.
In terms of the number of bits utilized to represent each sample, we first note that

$$\Delta = 2C\, 2^{-B}$$

Thus, the distortion depends on the *number of bits per sample* B as

$$D = \frac{C^2}{3} 2^{-2B} \simeq \frac{M^2}{3} 2^{-2B}, \text{ for large } B\,. \tag{24}$$

Non-Uniform Quantizer For a given number of bits per sample, the distortion D can be further reduced if the probability density function (PDF) of the analog samples is known. This can be achieved by utilizing a non uniform quantization step. Any form of non-uniform quantization can be accomplished by placing complementary non linear elements before and after a nearest neighbour quantizer. The first block is a *compressor*, and its transfer function $\mathscr{C}(x)$ is a monotonically increasing function satisfying

$$\mathscr{C}(-C) = -C, \quad \mathscr{C}(C) = C, \quad \mathscr{C}(0) = 0$$

The complementary block placed after the quantizer is called *expander*, and has a transfer function $\mathscr{C}^{-1}$.
Adapting an expression first derived in [47], one has that, for a non uniform quantizer with a large number of quantization levels, compressor characteristic $\mathscr{C}(x)$ and without overload, the MSE due to quantization is given by

$$D_{\mathscr{C}} = \frac{M^2}{3} 2^{-2B} \int_{X_{min}}^{X_{max}} \frac{f_{\mathbf{x}}(x)}{\left[\mathscr{C}'(x)\right]^2} dx \tag{25}$$

where $f_{\mathbf{x}}(x)$ is the PDF of the analog samples and $\mathscr{C}'(x) \triangleq d\mathscr{C}/dx$. The no overload assumption implies $-C \le X_{min}$ and $X_{max} \le C$, and that $f_{\mathbf{x}}(x) = 0$, $\forall x \notin [X_{min}, X_{max}]$. Notice that for $\mathscr{C}'(x) = 1$ (i.e., with a uniform quantization step), (25) becomes (24).
Clearly, minimization of $D_{\mathscr{C}}$ in (25) requires a compressor curve $\mathscr{C}$ matched to the PDF of the input signal. The optimal compressor characteristic $\mathscr{C}^*$ is given by the solution to

$$\frac{d\mathscr{C}^*(x)}{dx} = \alpha \left[f_{\mathbf{x}}(x) \right]^{1/3} \tag{26}$$

where α is a constant such that $\mathscr{C}(C) = C$. When the solution of (26) is inserted into (25), the MSE without overload and for large B is found to be

$$D_{\mathscr{C}^*} = \frac{\sigma^2}{12} \left\{ \int_{X_{min}/\sigma}^{X_{max}/\sigma} \left[f_{\mathbf{x_N}}(x) \right]^{1/3} dx \right\}^3 \cdot 2^{-2B} \tag{27}$$

where σ^2 and $f_{\mathbf{x_N}}(x)$ are the variance and the normalized PDF of an individual input analog sample, respectively. In relation to (27), it must also be pointed out that C (see (23)) must be made several times larger than σ for the no-overload assumption and (27) to be valid. For more details about the derivation and applications of this and other results related to scalar quantization, see, e.g., [48] and the references therein.

5.2 Oversampling

It is possible to further reduce the reconstruction MSE, while keeping the quantization step constant, by increasing the sampling frequency above the Nyquist frequency f_N. This technique is called *oversampling*. For oversampling ratio $r \triangleq 1/(\tau f_N)$ not too large, the mean square error is reduced as r^{-1}, i.e.,

$$D_r = D_1 r^{-1} \tag{28}$$

where D_1 is the MSE when $r = 1$ [47]. Notice that this can also be seen as a particular case of the resilience properties of redundant frame expansions discussed in Appendix A.5 (see also, e.g., [32, 49]). However, as the sampling frequency is increased, quantization noise becomes more and more correlated and the decrease rate of D_r diminishes. Furthermore, D_r asymptotically approaches a lower, strictly positive limit. The bigger Δ is, the higher this limit becomes. A larger quantization step also causes the decrease rate of D_r to depart from (28) "sooner" as r is increased [47].

The reconstruction error can be further reduced, for a given oversampling ratio, by the use of feedback[12]. Furthermore, feedback A/D converters yield a MSE that decreases steadily as r is increased. Thus, one can obtain an arbitrarily low MSE, for a given Δ, by sampling fast enough. These converters are briefly described in the next section.

6 AD CONVERTERS WITH FEEDBACK

Quantization schemes that use feedback can be grouped into two main families: *predictive quantizers* and *noise shaping quantizers*. Examples of the first type are the *delta modulator* and *differential pulse code modulator* (DPCM) (see, e.g., [50]). The popular $\Sigma\Delta$ (*sigma-delta*) converter, see, e.g., [51], belongs to the latter type.

The following is a basic description of the main characteristics of both converter families, based mainly on the approach proposed in [52]. In the sequel, the quantization process is modeled as additive noise, corresponding to the quantization error of a scalar quantizer.

6.1 Predictive Quantizers

The general form of a predictive quantizer is shown in Fig. 5.

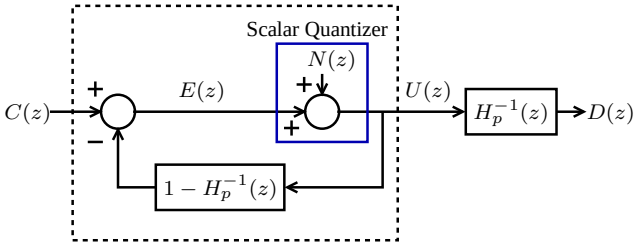

Figure 5: A predictive quantizer

[12] Here we begin to see control theory impacting signal processing.

In this diagram, $U(Z)$ and $C(z)$ correspond, respectively, to the Z-transforms of the analog samples sequence $\{c[k]\}$ and the quantized output sequence $\{u[k]\}$ depicted in Fig. 3. Thus, the quantizer contained in the dashed line rectangle in Fig. 5 is a particular realization of the generalized scalar quantizer in Fig. 3. The filter $H_p^{-1}(z)$ included at the end of the chain in Fig. 5 can be considered as part of the reconstruction stage in Fig. 3. The terms $E(Z)$ and $D(Z)$ in Fig. 5 correspond to the Z-transforms of the discrete-time signals in each of the respective nodes. $N(z)$ is the Z-transform of the error introduced by the scalar quantizer, i.e., $N(z) = U(z) - E(z)$. From Fig. 5, the expression for the output $U(z)$ is found to be

$$U(z) = H_p(z)\left[C(z) + N(z)\right]. \tag{29}$$

Thus, the filtered output $D(z)$ satisfies

$$D(z) = C(z) + N(z) \tag{30}$$

The key to the noise reducing capabilities of the predictive quantizer rests on the prediction filter $H_p(z)$. This filter is designed to minimize the variance of the prediction error

$$E(z) = H_p(z)C(z) + [1 - H_p(z)]N(z), \tag{31}$$

see Fig. 5. It is common to assume that the quantization noise is uncorrelated to any of the signals in the loop [51][13]. Thus, $H_p(z)$ is chosen so as to reduce the contribution of $C(z)$ to $E(z)$ in (31). By doing so, the variance (energy per sample) of the analog sequence that enters the quantizer is reduced. This in turn allows one to reduce the quantization step Δ in the embedded scalar quantizer, without increasing the number of quantization levels needed to avoid overload. Thus, by reducing a measure of the term $H_p(z)C(z)$ in (31), one is also reducing the quantization noise contribution, and the MSE is reduced accordingly.

Of course, how much distortion reduction is achieved will ultimately depend on how predictable the sequence $\{c[k]\}$ is, i.e., on the autocorrelation of $\{c[k]\}$. It will also depend on how well the prediction filter $H_p(z)$ is able to capture this predictability

It has been shown [52] that the MSE of the scheme in Fig. 5 decreases with the oversampling ratio not "faster" than $r^{-(2n_p)}$, where n_p is the order of the filter $H_p(z)$. If an additional ideal low pass filter with cut-off frequency $f_N/2$ is placed after $H_p^{-1}(z)$ (see Fig. 5), then the MSE is reduced at most as $r^{-(2n_p+1)}$. A common choice of $H_p(z)$ is of the form $(1 - z^{-1})^{n_p}$.

Note that the predictive quantizer in Fig. 5 can reduce distortion even if signals are sampled at Nyquist frequency, as long as the input analog samples are correlated. If the input samples are uncorrelated (white noise), then the predictive quantizer is unable to yield any MSE reduction at all. It is the increase in the autocorrelation of the input samples produced by oversampling which allows for the r^{-2n_p} behaviour in the MSE reduction rate.

[13] Other analysis methods of quantization noise consider more sophisticated spectral and probabilistic models (see, e.g., [53, 54]), as well as non-linear deterministic models (see, e.g., [55, 56, 57, 58]).

6.2 Noise- Shaping ($\Sigma\Delta$ Quantizers)

The second main category of feedback quantizers corresponds to the noise-shaping quantizers such as $\Sigma\Delta$ A/D converters, first proposed by Inose and Yasuda in [59]. One possible form to represent a noise shaping quantizer is depicted in Fig. 6. Again, $C(z)$ and $U(z)$ correspond, respectively, to the Z-transforms of $\{c[k]\}$ and $\{u[k]\}$ in Fig. 3. The noise shaping quantizer within the dashed line rectangle in Fig. 6 is a particular realization of the generalized scalar quantizer in Fig. 3.

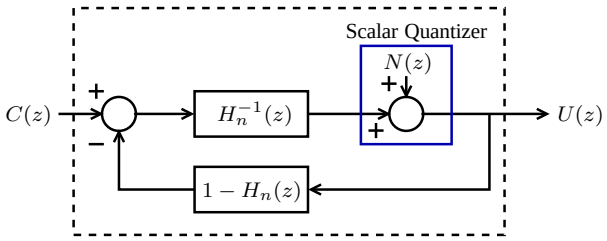

Figure 6: A noise-shaping quantizer.

From this figure, it is easy to see that the output $U(z)$ is given by

$$U(z) = C(z) + H_n(z)N(z) \tag{32}$$

where the noise shaping filter $H_n(z)$ constitutes a degree of freedom in the design process. Since $C(z)$ is band-limited, and because of oversampling, it is generally convenient to choose $H_n(z)$ to be a high-pass filter, see, e.g., [51]. With this choice, the quantization noise is attenuated within the signal band whilst increased outside of it (see Fig. 7). This compensatory increase in the off-band quantization noise is unavoidable, as determined by the Bode integral theorem [60] [14]. Because of the frequency shaping of the quantization noise, most of its energy can be suppressed by low pass filtering $U(z)$, leaving only the in-band portion of the quantization noise. By doing so, it is verified in [52] that the MSE decays by increasing oversampling ratio at most as $r^{-(2n_n+1)}$, where n_n is the order of the noise shaping filter $H_n(z)$. Most common choices for $H_n(z)$ have the form $(1-z^{-1})^{n_n}/P(z)$, where $P(z)$ is an FIR filter.

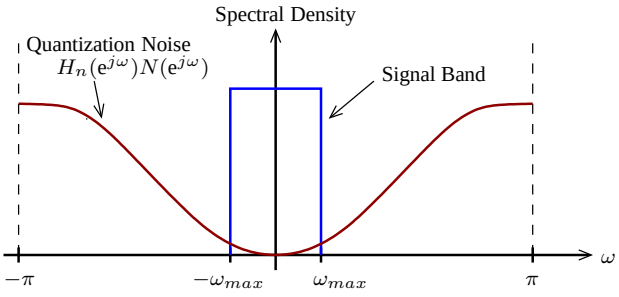

Figure 7: Quantization Noise Shaping.

As in control systems, one of the beneficial aspects of using feedback in analog-to-digital converters is the increased robustness of the resultant system. Indeed, if properly designed, feedback converters allow one to achieve high accuracy quantization despite the use of inaccurate building blocks (such as the scalar quantizer itself, which can be allowed to have a very coarse and uncertain quantization step). This makes feedback quantizers the preferred choice for many practical applications.

It should also be noted that the above mentioned decay rate of the MSE with increasing oversampling ratio is not fast enough to be rate-distortion efficient. Indeed, oversampling AD converters require, in general, a higher data-rate than a system with finer quantization and no oversampling to achieve the same distortion. This can be seen by noting that, for feedback converters, the MSE decays only polynomially with increasing the oversampling ratio, as[15] $O(r^{-(2n+1)})$, while the MSE decreases with increasing the bits per sample (i.e., reducing Δ) as $O(2^{-2B})$, i.e., exponentially. Nevertheless, recent results show that the L^∞ norm of the reconstruction error in $\Sigma\Delta$ converters can be reduced as $O(\kappa^{-r})$, $\kappa > 0$, by selecting for each oversampling ratio an appropriate noise shaping filter from an infinite set of filters [64]. Following a different approach, quantization schemes based on threshold crossings exhibit a reconstruction MSE that decays exponentially with increasing oversampling ratio [65, 66], and are thus rate-distortion efficient.

7 MOVING HORIZON QUANTIZATION

Interestingly, control theory can be used to design the generalized scalar quantizer in Fig. 3. More precisely, since the output of the quantizer is constrained to belong to a finite alphabet of values, the situation can be regarded as a control problem with input constraints. This point of view motivated us to apply *Moving Horizon Optimization* (MHO) tools to achieve a more effective noise shaping quantizer. This paradigm uses *Model Predictive Control*, which has proved to be a powerful tool for dealing with constrained systems [67, 61, 68, 69, 70, 63]. The quantization scheme so obtained, named *Multi Step Optimal Converter* (MSOC) [71], typically outperforms $\Sigma\Delta$ quantizers, while embedding the latter as a particular case. We will present next some of the fundamental principles behind the MSOC. The remainder of this section has been basically adapted from [71].

7.1 Noise Shaping Quantization as an Optimization Problem

A more general formulation to analyze the discrete-time performance of noise shaping quantization can be derived from the block diagram depicted in Fig. 8.

In Fig. 8, $\{c[k]\}$ and $\{u[k]\}$ represent, respectively, the input analog samples and the quantized output sequence. The motivation for quantization noise-shaping has been incor-

[14] Note that $H_n(z)$ in (32) corresponds to the closed loop sensitivity of the system in Fig. 6 (see, e.g., [61, 62, 63]).

[15] If x is a variable that tends to some limit and $g(x)$ is a positive function, the expression $f(x) = O(g(x))$ means that there exists a finite constant Λ such that $|f(x)| < \Lambda g(x)$ for all values of x.

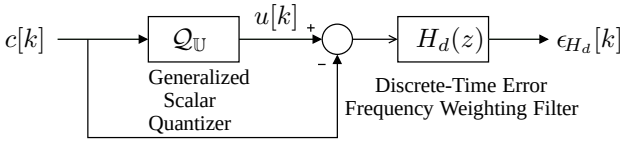

Figure 8: Scheme to generate the frequency weighted quantization error sequence $\epsilon_{H_d}[\cdot]$.

porated by introducing a frequency weighted reconstruction error sequence, denoted by

$$\epsilon_{H_d}[k] \triangleq H_d(z)\,(c[k] - u[k])\,, \quad k \in \mathbb{Z}, \tag{33}$$

compare to (7).
In (33), H_d is a stable, causal, linear, time-invariant filter, which can be characterized via[16]:

$$H_d(z) \triangleq 1 + C(zI - A)^{-1}B, \tag{34}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$, $C \in \mathbb{R}^{1\times n}$ and $n \in \mathbb{N}$ is the state dimension, i.e. the order of the filter H_d. This filter can, e.g., represent the typical low-pass filter utilized in oversampled conversion, see e.g. [72], in order to decimate the converter output. In audio applications it makes sense to choose H_d as a psycho-acoustic model of the human hearing, compare also with work in [73, 74].
The performance of the quantization process in Fig. 8 will be evaluated by the measure

$$V \triangleq \sum_{k\in\mathbb{Z}} [\epsilon_{H_d}[k]]^2. \tag{35}$$

The cost V penalizes the distortion introduced in the conversion process in a frequency-selective manner.
If the generalized scalar quantizer in Fig. 8 is designed to minimize the performance measure V, then its quantized output u will approximate the input c, while the un-filtered quantization error, $a - u$, will tend to have a spectrum similar to that of the inverse of the filter H_d. Thus, the method will shape the quantization noise spectrum, just as the $\Sigma\Delta$ converter discussed in Section 6 does.
Unfortunately, minimization of V by using expression (35) is not possible in practical applications, due to the complexity of solving the associated combinatorial optimization problem. Furthermore, in the general case, an optimal quantizer would need to *pre-view* the entire signal c. This is clearly unsuitable for on-line applications.

7.2 Multi Step Optimal Converter

In order to obtain a more practical method to minimize the cost in (35), it is convenient develop a recursive conversion method, which can be implemented on-line. For that purpose, we will first introduce a cost measure over a finite horizon, to deploy later the concept of *moving horizon approximation*, see [63].

[16] Here, and in the remainder of this paper, z denotes the forward shift operator, $zv[k] = v[k+1]$.

Finite Horizon Formulation A practical conversion scheme, suitable for online applications, must operate sequentially, evaluating a restricted number of decision variables and considering a moderate number of future values of c. For this purpose, it is convenient to characterize ϵ_{H_d} as the output in a state space representation of H_d

$$\begin{aligned} x[k+1] &= Ax[k] + B(c[k] - u[k]) \\ \epsilon_{H_d}[k] &= Cx[k] + \quad c[k] - u[k]. \end{aligned} \tag{36}$$

This relation follows directly from (34). In (36), $x \in \mathbb{R}^n$ is the state vector. Note that, due to the Markovian structure of (36), at time $k = \ell$ the impact of the past trajectories of c and u on future values of ϵ_{H_d} is exactly summarized by means of the *present* state, $x[k]$.
Given the above, we next replace the infinite horizon cost function (35) by the *finite* horizon cost:

$$V_N(\ell) \triangleq x^T P x + \sum_{k=\ell}^{\ell+N-1} \left(\epsilon_{H_d}[k]\right)^2. \tag{37}$$

In (37), $N \in \mathbb{N}$ determines the prediction horizon and P is a given positive semidefinite matrix.
With a given and known current state value $x[\ell]$ (see (36)), V_N is a measure of the filtered distortion ϵ_{H_d} over the prediction horizon plus a measure of the *final* state, $x[\ell + N]$. These predicted quantities are formed based upon the model (36).
The finite horizon cost $V_N(\ell)$ proposed in (37) takes into account only a finite number N of constrained values. The value of N determines the computational complexity required for the minimization of $V_N(\ell)$. This should be compared with the infinite number of decision variables in the original cost V. Using a finite horizon N also reduces the required pre-viewing of c to $N - 1$ samples. Since N is a design parameter, it can be chosen so that the minimization can be carried out on-line.

Moving Horizon Approach As noted above, the optimizer to $V_N(\ell)$, say $\vec{u}_\ell^\star$, contains a feasible output sequence for time instants $\ell \le k \le \ell + N - 1$. Thus, in principle, one could think of an implementation *in blocks*, where the minimization is carried out every N sampling instants. Unfortunately, the last few elements of $\vec{u}_\ell^\star$ depend only on a small window of the filtered distortion, ϵ_{H_d}. To improve performance, the multi-step optimal converter utilizes only the *first* element of $\vec{u}_\ell^\star$, say $u^\star[\ell] \in \mathbb{U}$. It becomes the ℓ-th element of the converter output sequence, by setting:

$$u[\ell] \longleftarrow u^\star[\ell] \tag{38}$$

It is also utilized to update the state according to (36), i.e.:

$$x[\ell + 1] = Ax[\ell] + B\,(c[\ell] - u^\star[\ell])\,. \tag{39}$$

At the next sampling instant, this new state value is used to minimize the cost $V_N(\ell + 1)$, yielding $u[\ell + 1]$. This procedure is repeated *ad-infinitum*. As illustrated in Fig. 9

for the case $N = 3$, the prediction horizon of the criterion $V_N(k)$ moves (slides) forward as k increases. The past is propagated forward in time via the state sequence x, thus, yielding a recursive scheme.
The resultant architecture defines the MSOC. It constitutes an analog-to-digital converter architecture which optimizes the frequency weighted conversion distortion, based upon Model Predictive Control principles.

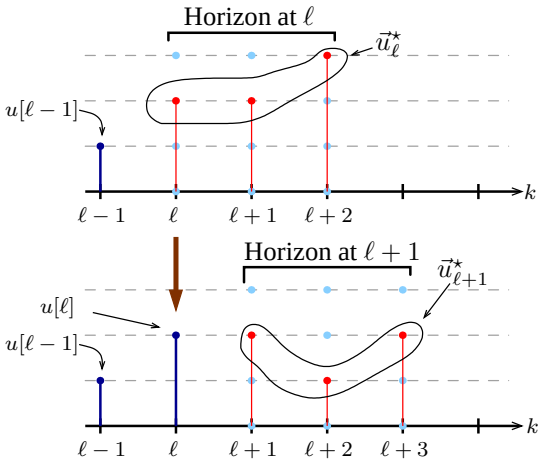

Figure 9: Moving horizon principle, $N = 3$.

Interestingly, it has been shown that the MSOC with $N = 1$ and $P = 0$ reduces to the $\Sigma\Delta$ converter, see [71]. However, it is easy to see that, in general, larger values for N provide better performance, since more data is taken into account in the decision process of allocating scalars from $\mathbb{U}$ to the elements in the sequence u. In fact, one can expect that, if N is chosen large enough relative to the time scale of H_d, then the effect of $u[\ell]$ on $\epsilon_{H_d}[j]$ for $j \geq \ell + N$ will be negligible and the performance of the MSOC will approach that obtained if the infinite horizon measure of (35) were to be minimized directly (which, for the reasons explained above, is impractical). This asymptotic behaviour has been experimentally confirmed, see [71].
In summary, the prediction horizon N allows the designer to trade-off performance versus on-line computational effort. Interestingly enough, excellent performance can often be achieved with relatively small horizons (see, e.g., [71]), thus rendering the scheme quite easy to implement in practical cases.
Another advantage of the MSOC when compared to the $\Sigma\Delta$ converter resides in that the matrix P in (37) can be designed to ensure *stability like* properties of the MSOC, see [71].

8 SAMPLED-DATA QUANTIZATION

Given that digital signal processing systems have to interact with the real, physical world, the design of a quantization scheme should take into account the sampling (continuous to discrete-time) and reconstruction (discrete to continuous-time) stages between which it is to be inserted. Unfortunaly, there exists only partial understanding of how sampling-reconstruction strategies interact with a given quantization method in terms of the resulting, overall reconstruction error. Furthermore, most literature analyzes the performance of quantizers in terms of how close the input analog *samples* are approximated by the output, quantized *samples*, and not by comparing the analog, continuous-time underlying signal entering the system against the analog, continuous-time reconstructed signal that comes out of the reconstruction stage of the system. Accordingly, performance is most often measured by the ℓ^2 norm of the sample approximation error, see (21). Similarly, traditional works on sampling and reconstruction theory build their analysis based first upon ideal, non-quantized samples, incorporating later the effect of quantization viewed as the corruption of ideal samples by white additive noise. Although it has been shown that this white-noise model of quantization is indeed accurate for small quantization steps and input samples whose PDF satisfies certain rather weak requirements, it is certainly not accurate, for example, when quantization steps are large, or when feedback structures are deployed. As presented in Sections 6 and 7, it is often the case that quantization noise is deliberately made non-white by the quantizer so as to minimize a frequency weighted measure of the reconstruction error.
Within the setup depicted in Fig. 3, we aim to present in this section some results and additional insight related to the joint problem of designing systems that make use of the pre-filtering, sampling, quantization and reconstruction paradigm[17].

8.1 Decomposition of the Reconstruction Error

The Hilbert spaces model of the sampling and reconstruction process described in Section 3 leads to a somewhat trivial but nevertheless important result: it allows for a decomposition of the final reconstruction MSE between the analog input a and the analog output $\tilde{a}$ (see Fig. 3) of a sampling-quantization-reconstruction system into two terms. The first term corresponds to the error due to the "*spaces mismatch*", i.e., the non coincidence of input and output signal spaces. The second error term comes from the deviation of the discrete-time processing (both linear and non-linear) from the optimal mapping between input and output vectors in the sampling and reconstruction spaces, respectively. The following proposition formalizes this idea:

Proposition 2. *Let $\psi(\cdot) \in L^2$ be the impulse response of the reconstruction filter R, such that $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ is frame for $\mathcal{W} \triangleq \overline{span}\,\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$, and let τ be the sampling interval. Then, the mean square reconstruction error between any input signal $a \in L^2$ and an approximation $\tilde{a} \in \mathcal{W}$ generated by the reconstruction stage can always be de-*

[17] Other approaches to the joint problem which fall outside this framework, such as sparse representations [75, 76], non-linear reconstruction [57, 32], sub-band coding [77, 20, 78, 36] and threshold crossing quantization [65, 66], are not discussed here.

composed as follows

$$\|a-\tilde{a}\|_{L^2}^2 = \|a-P_{\mathcal{W}}\,a\|_{L^2}^2 + \|P_{\mathcal{W}}\,a-\tilde{a}\|_{L^2}^2\,, \quad (40)$$

where $P_{\mathcal{W}}\,a$ is the orthogonal projection of a onto $\mathcal{W}$.

Proof. Define $w \triangleq \tilde{a} - P_{\mathcal{W}}\,a$. Then we can write

$$\begin{aligned}\|a-\tilde{a}\|_{L^2}^2 &= \langle a-w-P_{\mathcal{W}}\,a\,,\, a-w-P_{\mathcal{W}}\,a\rangle \\ &= \|a-P_{\mathcal{W}}\,a\|_{L^2}^2 - 2\langle a-P_{\mathcal{W}}\,a, w\rangle + \|w\|_{L^2}^2\end{aligned}$$

Since $(a-P_{\mathcal{W}}\,a) \in \mathcal{W}^{\perp}$, and because $w \in \mathcal{W}$, we have that $\langle a-P_{\mathcal{W}}\,a, w\rangle = 0$ (see (50)), and (40) follows. □

Corollary 1. *From Proposition 2, it follows that for any $a \in L^2$, choice of quantization scheme and/or discrete time processing, the continuous time reconstruction error is lower bounded by*

$$\|a-\tilde{a}\|_{L^2}^2 \geq \|a-P_{\mathcal{W}}\,a\|_{L^2}^2 \quad (41)$$

We emphasize that the lower bound in (41) corresponds to the minimum continuous-time error attainable by *any* discrete-time scheme, once the output space is given, even if and no quantization is applied to the samples.

From Proposition 2, it is clear that the performance of discrete-time processing (e.g., discrete-time filtering and quantization) should be evaluated in terms of the second term of the right hand side of (40), that is, the L^2 norm of $P_{\mathcal{W}}\,a-\tilde{a}$. In relation to the design of quantizers, this gives rise to the question of what information is needed by a generalized scalar quantizer to minimize $\|P_{\mathcal{W}}\,a-\tilde{a}\|_{L^2}^2$. We have addressed this question in [79]. A summary of the analysis and results therein is presented below.

8.2 Optimality

As noted above, the reduction of the continuous time MSE by discrete-time processing takes place by minimizing the second term on the right hand side of (40). For the general system under study (see Fig. 3), the signal to be approximated is actually a convolved with $h \in L^2$, the impulse response of H:

$$\alpha(t) \triangleq (a*h)(t), \quad \forall t \in \mathbb{R}, \quad (42)$$

as shown in Fig. 10.

Defining λ as the impulse response of R, the approximation of α generated by the system becomes

$$\tilde{\alpha}(t) \triangleq (u*\psi)(t), \quad \forall t \in \mathbb{R}$$

where ψ is now redefined as the impulse response of the filter $W \triangleq HR$, i.e.:

$$\psi(t) \triangleq (\lambda*h)(t), \quad \forall t \in \mathbb{R},$$

see, Fig. 10. The impulse response of W determines the reconstruction frame $\{\psi_k\}_{k\in\mathbb{Z}}$, which spans the reconstruction Hilbert space

$$\mathcal{W} \triangleq \overline{\text{span}}\,\{\psi_k\}_{k\in\mathbb{Z}}$$

As described in Section 4.2, the generation of the optimal output $P_{\mathcal{W}}\,\alpha$ can be accomplished by applying the pre-frame operator Ψ associated with $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ to the sequence of scalars $\left\{\langle T_{k\tau}\mathring{\psi}, a\rangle\right\}_{k\in\mathbb{Z}}$, i.e.

$$P_{\mathcal{W}}\,\alpha = \Psi\mathring{\Psi}^*\alpha, \quad \forall \alpha \in L^2,$$

where $\mathring{\Psi}^*$ is the analysis operator associated to $\left\{T_{k\tau}\mathring{\psi}\right\}_{k\in\mathbb{Z}}$, the canonical dual frame of $\{T_{k\tau}\psi\}_{k\in\mathbb{Z}}$ (see Definition 7 in the Appendix). We will denote this optimal, un-quantized sequence of samples by

$$u^{\circ} = \{u^{\circ}[k]\}_{k\in\mathbb{Z}} \triangleq \left\{\langle T_{k\tau}\mathring{\psi}, \alpha\rangle\right\}_{k\in\mathbb{Z}}. \quad (43)$$

It is clear from the above that any quantization algorithm that attempts to minimize the continuous time error $\|P_{\mathcal{W}}\,\alpha-\tilde{\alpha}\|_{L^2}^2$ needs to be able, in the first place, to obtain the *target* sequence u° in (43). From the results presented in Section 4.2, this implies that the first necessary condition for the feasibility of optimal quantization is that sampling and reconstruction stages be matched for orthogonal (MSE) reconstruction.

If we now suppose that the quantizer has access to u°, then the problem of optimal quantization is that of choosing the *optimal quantized sequence* $u^{\star}$, defined as

$$u^{\star} = \arg\min_{u[k]\in\mathbb{U},\,\forall k\in\mathbb{Z}} \|P_{\mathcal{W}}\,\alpha-\Psi u\|_{L^2}^2 \quad (44)$$

The solution to (44) requires one to solve a continuous-time optimization problem with discrete-time, quantized decision variables. It is shown in [79] that this can be converted into an equivalent discrete time optimization problem. More precisely,

$$\begin{aligned}\|P_{\mathcal{W}}\,\alpha-\Psi u\|_{L^2}^2 &= \|\Psi(u^{\circ}-u)\|_{L^2}^2 \\ &= \langle \Psi(u-u^{\circ})\,,\,\Psi(u-u^{\circ})\rangle_{L^2} \\ &= \langle u-u^{\circ}\,,\,\Psi^*\Psi(u-u^{\circ})\rangle_{\ell^2}\end{aligned} \quad (45)$$

The operator $\Psi^*\Psi : \ell^2 \mapsto \ell^2$ is characterized by the *Gram* matrix (see [80, sec. 3.5]) of the reconstruction frame, which is defined element-wise as

$$\mathbf{G}_{\psi\,j,k} = \langle T_{j\tau}\psi, T_{k\tau}\psi\rangle_{L^2}, \quad j,k \in \mathbb{Z}$$

This matrix allows one to re write (45) in matrix notation as

$$\|P_{\mathcal{W}}\,\alpha-\Psi u\|_{L^2}^2 = (\vec{u}-\vec{u}^{\circ})^T\mathbf{G}_{\psi}(\vec{u}-\vec{u}^{\circ}) \quad (46)$$

where $\vec{u}$ and $\vec{u}^{\circ}$ are the vector representations of the sequences u and u°, respectively.

The direct consequence of (46) is that a quantizer can determine the optimal output sequence without full knowledge of the inter-sample behaviour of the impulse responses of the reconstruction filter. Indeed, quantization performance can be measured by the *weighted* ℓ^2 norm implicitly defined in (46). Note that the design of an optimal quantizer is not possible without knowledge of the matrix $\mathbf{G}_{\psi}$.

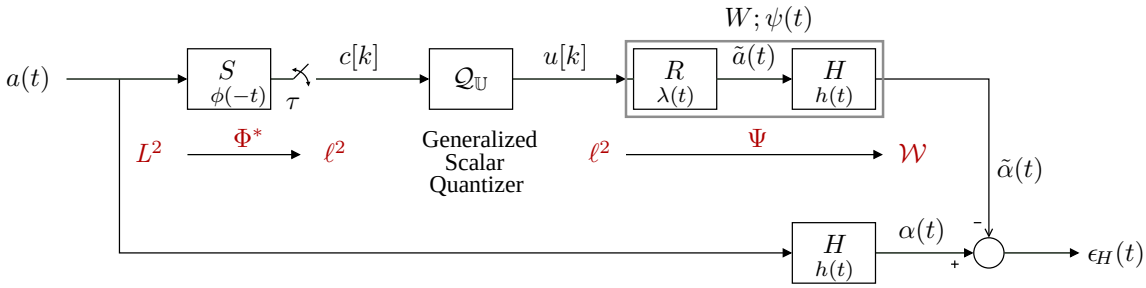

Figure 10: The sampling, quantization and reconstruction system from Fig. 3 revisited. Impulse responses and frame operators are shown for each filter.

8.3 Moving Horizon Conversion

In general, minimization of (46) would require one to evaluate it for every sequence $\{u[k]\}_{k\in\mathbb{Z}}$, $u[k] \in \mathbb{U}\,\forall k \in \mathbb{Z}$, that can be generated by the quantizer. This optimization programme, however, becomes intractable for sufficiently long sequences. Given the similarity of (46) and (35), one can use the ideas introduced in Section 7 and optimize over a short moving horizon of samples. Details can be found in [81, 79], where a sampled-data multi step optimal converter is proposed. Preliminary results show that, interestingly, significant distortion reduction is obtained even when converting non band-limited signals. Indeed, since the focus is on the reduction of the total continuous-time reconstruction error, if the sampling rate is lower than the Nyquist rate, the resultant converter will attempt to reduce not only quantization noise, but also aliasing noise. Furthermore, as the horizon is made larger, the output of the converter approaches the optimal feasible output sequence, defined in (44).

9 APPLICATIONS TO CONTROL

In previous sections of this work we have illustrated that the power of feedback can be used in the design of AD-conversion schemes. In particular, we have shown in Sections 7 and 8 that careful deployment of elements of Model Predictive Control may lead to high-performance conversion techniques. The purpose of the present section is to highlight the role played by sampling and especially quantization in feedback control applications.

Efficiency in data representations plays a central role in any control system where parsimony aspects need to be taken into account. Thus, quantization and sampling are worth investigating, for example, in the following situations:

- when signals need to be transmitted over a digital network, i.e., in Networked Control Systems (NCS's) [3, 82, 4];
- when plant inputs need to be quantized (e.g., relay feedback, on-off control, digital control, or also due to the presence of a human operator)[70];
- in large scale systems, such as those related to mining operations and supply chain management.

In the following, we will briefly describe how concepts surrounding sampling and quantization translate into the design of these types of control systems.

Sampling and Reconstruction In the design of a sampling/reconstruction scheme for a control system, traditional reconstruction criteria should be complemented with more appropriate performance notions. Indeed, reconstruction quality is only of secondary importance. The main objective is measured at the plant output. In particular, as shown in [83, 5] for NCS's, open loop performance measures should be replaced by closed loop ones. This can be achieved through consideration of frequency weighted measures such as (7).

Quantization Interestingly, the noise shaping ideas described in Sections 6-8 can also be applied to control systems where signals are quantized; see, e.g., [70]. For example, when focusing on the design of controllers for plants with quantized inputs, a key point resides in realizing that the AD-conversion scheme of Fig. 3 is related to a quantized control system with plant H: The plant input $u[k]$ is to be chosen such that the plant output $H\tilde{a}(t)$ tracks the reference signal $Ha(t)$. Thus, performance can be measured via the frequency weighted error signal $\epsilon_H(t)$, see (7) and also (35).

Details on how to apply principles of Moving Horizon to NCS's can be found, for example, in [84]. It is interesting to note that the framework can also be enriched to incorporate dynamic scheduling into NCS's. The resultant methodology can be regarded as incorporating sampling and quantization *on demand* and is, thus, highly efficient from a data representation perspective, see [84].

10 CONCLUSIONS

This paper has reviewed basic results and methods related to the process of sampling, quantization and reconstruction of scalar signals. With the introduction of a frame theoretic viewpoint, three notions of sampling and reconstruction have been discussed. We have described several generalized scalar quantization schemes, and have showed how control theory has contributed to signal processing theory. Furthermore, we have given insights into the joint

problem of sampling, quantization and reconstruction, and have outlined how these stages interact. Finally, we have examined the role played by sampling and quantization in control systems.

A APPENDIX

A.1 Background on Hilbert Spaces, Riesz Bases and Isomorphisms

Definition 2 (Hilbert Space)**.** Let $\mathcal{W}$ be a vector space with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ and the induced norm $\|\cdot\|_{\mathcal{W}} \triangleq \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{W}}}$. If such a space is complete under its norm then it is a *Hilbert Space*.

Definition 3 (Riesz Basis)**.** A sequence of vectors (functions) $\{\psi_k\}_{k\in\mathbb{K}}$, $\mathbb{K}\subseteq\mathbb{Z}$, in a Hilbert space $\mathcal{W}$ is a a *Riesz basis* for $\mathcal{W}$ if and only if $\mathcal{W} = \overline{\text{span}}\,\{\psi_k\}_{k\in\mathbb{K}}$ [18] and there exist two constants $0 < m \leq M < \infty$ such that[19]

$$\forall\, c\in\ell^2,\ m\,\|c\|_{\ell^2}^2 \leq \left\|\sum_{k\in\mathbb{K}} c[k]\psi_k\right\|_{\mathcal{W}}^2 \leq M\,\|c\|_{\ell^2}^2 \quad (47)$$

This and other equivalent definitions can be found in [80, Theorem 3.6.6] .

Remark 2. *From Definition 3 one can observe that:*

- *The elements ψ_k in* (47) *are* orthogonal *if and only if $m = M$ and* orthonormal *if and only if $m = M = 1$.*
- *The lower bound in* (47) *is equivalent to saying that $\{\psi_k\}_{k\in\mathbb{Z}}$ is a set of linearly independent vectors.*
- *The higher bound in* (47) *guarantees that $\sum_{k\in\mathbb{K}} c[k]\psi_k$ will be bounded for any choice of $c\in\ell^2$.*

A.1.1 Orthogonal Projection

If $\mathcal{W}\subseteq\mathcal{H}$ is a Hilbert space, then the *best approximation* in $\mathcal{W}$ (in the sense of the norm $\|\cdot\|_{\mathcal{H}} = \sqrt{\langle\cdot,\cdot\rangle_{\mathcal{H}}}$) of any $h\in\mathcal{H}$ is given by the *orthogonal projection* of h onto $\mathcal{W}$, denoted by $P_{\mathcal{W}}\,h$, and defined as the operator

$$P_{\mathcal{W}} : \mathcal{H}\mapsto\mathcal{W};\quad P_{\mathcal{W}}\,h \triangleq \arg\min_{w\in\mathcal{W}} \|h-w\|_{\mathcal{H}} \quad (48)$$

The orthogonal projection from a Hilbert space $\mathcal{H}$ onto $\mathcal{W}\subseteq\mathcal{H}$ implicitly defines the *null space* of $\mathcal{W}$:

$$\mathcal{W}^{\perp} \triangleq \{h\in\mathcal{H} : P_{\mathcal{W}}\,h = 0\}. \quad (49)$$

[18] The span of a set of vectors is the vector space consisting of all possible linear combinations of the set. The closed span, written as $\overline{\text{span}}$, of a set of vectors, is the *closure* of the span of these vectors. Open and closed spans of a finite set of vectors are equal. However, the open and closed spans of an infinite set of vectors are in general different. The closure of the span becomes mandatory in such cases, since Hilbert spaces are closed spaces.

[19] Throughout this section, $\mathbb{K}\subseteq\mathbb{Z}$.

It is easy to verify that

$$\langle\beta, w\rangle_{\mathcal{H}} = 0,\ \forall\beta\in\mathcal{W}^{\perp},\ \forall w\in\mathcal{W}. \quad (50)$$

If $\{\upsilon_k\}_{k\in\mathbb{K}}$, is an orthonormal basis of $\mathcal{W}$, then the orthogonal projection operator can be explicitly written as

$$P_{\mathcal{W}}\,h = \sum_{k\in\mathbb{K}} \langle h,\upsilon_k\rangle_{\mathcal{H}}\upsilon_k, \quad \forall h\in\mathcal{H} \quad (51)$$

Orthogonal projection permits elegant solutions to some otherwise complex optimization problems in functional analysis. This makes Hilbert spaces and operators a natural framework for studying the problem of efficient sampling and quantization.

A.1.2 Isomorphism

A fundamental property of Hilbert spaces and operators is that they are able to define a precise form of equivalence between two different Hilbert spaces. It is called *isomorphism*: two different Hilbert spaces are *isomorphic* if they have the same dimension[20]. An isomorphism is indeed any linear invertible[21] operator from one space onto the other. Of particular interest for our analysis are the isomorphisms between any separable Hilbert space $\mathcal{W}\subset L^2$ (function space) of dimension $|\mathbb{K}|$, where $\mathbb{K}\subseteq\mathbb{Z}$, and $\mathbb{R}^{|\mathbb{K}|}$ (Euclidian space). Such an isomorphism can be stated by considering any orthonormal basis of $\mathcal{W}$, namely $\{\upsilon_k\}_{k\in\mathbb{K}}$, and constructing the associated analysis operator

$$\Upsilon^* : \mathcal{W}\mapsto\mathbb{R}^{|\mathbb{K}|};\quad \Upsilon^* w \triangleq \{\langle w,\upsilon_k\rangle_{\mathcal{W}}\}_{k\in\mathbb{K}} \quad (52)$$

The analysis operator Υ^* defined in (52) is an *unitary isomorphism*. This means that *the respective images in $\mathbb{R}^{|\mathbb{K}|}$ through Υ^* of any group of vectors in $\mathcal{W}$ preserve their respective norms and relative orientations*, i.e.

$$\langle\Upsilon^* w_1, \Upsilon^* w_2\rangle_{\ell^2} = \langle w_1, w_2\rangle_{\mathcal{W}} \quad (53)$$

This remarkable property of isomorphic spaces allows one to study the relation between elements of a Hilbert space by looking at their images through Υ^* in another, more convenient Hilbert space. Actually, one can argue that all digital signal processing (including digital control) is made possible because of the existence of isomorphism between signal spaces and subspaces of ℓ^2.

A.2 Illustrative Example

Some of the basic concepts of Hilbert spaces of signals and bases presented so far will be illustrated by the following simple example.

Let $\mathcal{W}$ be the space of all real valued functions $w(t)$ satisfying the following conditions:

- $w(t)$ is continuous.
- $w(t) = 0,\ \forall t\notin\mathbf{I}\triangleq[0,3\tau]$.

[20] For the case of infinite dimensional spaces, all separable spaces (i.e., spaces with infinite but countable dimension) are isomorphic.

[21] Hence the need for both spaces to have equal dimension.

- $\int_0^{3\tau} w^2(t)dt < \infty,\ \forall s \in \mathcal{S}$ (i.e., $w(\cdot)$ is square integrable over $\mathbf{I}$).
- The derivatives of $w(t)$ are constant over any of the open intervals $i_k = (k\tau, k\tau + \tau), k = 0, 1, 2$.

Fig. 11 a) shows three functions, $w_1(t), w_2(t), w_3(t)$ that belong to this space.

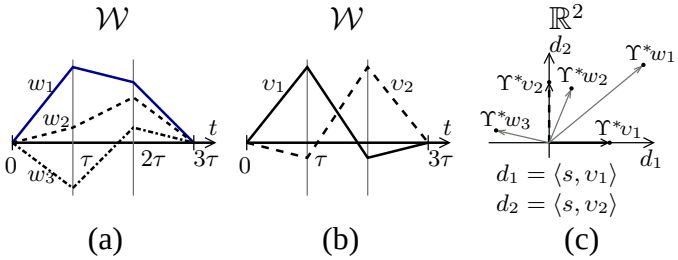

Figure 11: Example of a functional space, an orthonormal base and a unitary analysis operator. a) Functions w_1, w_2 and w_3 belong to the Hilbert space $\mathcal{W}$; b) The functions $v_1, v_2 \in \mathcal{S}$ constitute an orthonormal basis for $\mathcal{W}$; c) Image of the functions w_1, w_2 and w_3 in $\mathbb{R}^2$ through the analysis operator Υ^*.

With the addition of the standard L^2 inner product, defined as[22]

$$\langle w_1, w_2 \rangle_{L^2} \triangleq \int_{-\infty}^{\infty} w_1(t) w_2(t) dt, \quad \forall w \in \mathcal{W} \qquad (54)$$

$\mathcal{W}$ becomes a Hilbert space. The inner product (54) also defines a norm in $\mathcal{W}$, given by

$$\|w\|_{L^2} \triangleq \sqrt{\langle w, w \rangle_{L^2}}$$

It is easy to show that $\mathcal{W}$ is a two-dimensional space. This can be intuitively verified by noting that any function $w \in \mathcal{W}$ is completely determined by exactly two parameters, such as, for example, the values of the functions evaluated at τ and 2τ. A basis for a Hilbert space of dimension two contains two elements. Figure 11.b) shows a pair of orthonormal functions v_1, v_2 in $\mathcal{W}$ which form an orthonormal basis for $\mathcal{W}$.
Figure 11.c) shows the images of w_1, w_2, w_3, v_1 and v_2 through the analysis operator Υ^* (see (52)) in $\mathbb{R}^2$. As expected, the images of the orthonormal functions v_1 and v_2 are orthonormal vectors in $\mathbb{R}^2$. How "close" is w_1 to w_2 in their space's norm?. Since the analysis operator Υ^* is a unitary isomorphism between $\mathcal{W}$ and $\mathbb{R}^2$, we have, from (53)

$$\|w_1 - w_2\|_{L^2}^2 = \|\Upsilon^* w_1 - \Upsilon^* w_2\|_{\ell^2}^2 = \left(\langle w_1, v_1 \rangle - \langle w_2, v_1 \rangle\right)^2 + \left(\langle w_1, v_2 \rangle - \langle w_2, v_2 \rangle\right)^2$$

i.e., $\|w_1 - w_2\|_{L^2}$ is given by the Euclidian distance between $\Upsilon^* w_1$ and $\Upsilon^* w_2$.

[22] Since all the signals considered here are real, and for ease of notation, we will write the inner products in L^2 and in ℓ^2 without complex conjugation of one of the arguments.

Consider now the case of a function $h(t)$, $t \in \mathbb{R}$, that belongs to a space $\mathcal{H} \subset L^2$, such that $\mathcal{W} \subsetneq \mathcal{H}$ and $h \notin \mathcal{W}$. An example of such a function is shown in Fig. 12.a).

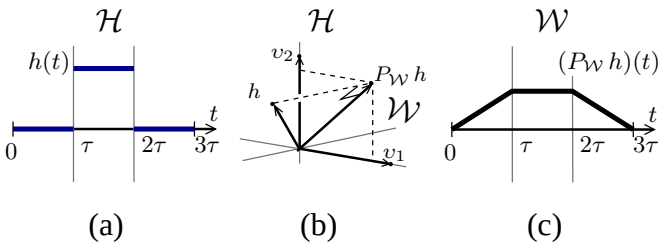

Figure 12: a) Orthogonal projection onto $\mathcal{W} \subset \mathcal{H}$. a) Function $h(t)$ belongs to $\mathcal{H}$. b) Relative positions between h, v_1, v_2 and $P_{\mathcal{W}} h$ represented in an isomorphic Euclidian space. c) Orthogonal projection of $h(t)$ onto $\mathcal{W}$ in function representation.

The magnitudes and relative directions of h with respect to an orthogonal basis for $\mathcal{W}$ such as $\{v_k\}_{k=1}^2$ are shown in the 3 dimensional representation of Fig. 12.b). Here it can be seen that h is outside $\mathcal{W}$ but has a non zero orthogonal projection onto $\mathcal{W}$. This orthogonal projection is the closest vector to h in $\mathcal{W}$, in accordance with (48), and is given by (51). Consequently, the best approximation (in an L^2 sense) of h in $\mathcal{W}$ is, expressed as a function of time

$$(P_{\mathcal{W}} h)(t) = \langle h, v_1 \rangle_{L^2} v_1(t) + \langle h, v_2 \rangle_{L^2} v_2(t)$$

Figure 12.c) shows a plot of $(P_{\mathcal{W}} h)(t)$.

A.3 Frames

Despite the computational convenience of bases, one often needs to study spaces generated by a set of linearly *dependent* vectors (over-complete basis). The concept of frames, introduced by Duffin and Schaffer [85], allows one to analyze such cases. Situations with over-complete bases arise in practice not only by chance. It has been shown that the redundancy of frames is beneficial, for it can reduce the effect of errors in the expansion coefficients, see [39] and Appendix A.5. The definition and some properties of frames are given next.

Definition 4 (Frame)**.** A sequence $\{\psi_k\}_{k \in \mathbb{K}}$ of elements in a Hilbert space $\mathcal{W}$ is a *frame* for $\mathcal{W}$ if there exist constants $A, B > 0$ such that

$$A \|w\|^2 \leq \sum_{k \in \mathbb{K}} |\langle w, \psi_k \rangle|^2 \leq B \|w\|^2, \quad \forall w \in \mathcal{W} \qquad (55)$$

The largest number A and smallest number B that satisfy (55) are called *frame bounds*. Some important remarks about frames are:

- If $\{\psi_k\}_{k \in \mathbb{K}}$ is a frame for a Hilbert space $\mathcal{W}$, then $\overline{\text{span}}\, \{\psi_k\}_{k \in \mathbb{K}} = \mathcal{W}$.

- A frame is said to be *tight* if one can choose $A = B$ as frame bounds. If A=B=1, it is called a *Parseval frame*.
- If a frame ceases to be a frame when an arbitrary element is removed, it is called an *exact frame*. An exact frame is equivalent to a Riesz basis.
- A frame $\{\psi_k\}_{k\in\mathbb{K}}$ in which $\|\psi_k\| = 1$ for all $k \in \mathbb{K}$ is called a *normalized frame*.
- If the elements of a normalized frame are linearly independent then $A \leq 1 \leq B$ (see [39]).
- A frame with linearly dependent elements is said to be *redundant*.
- The upper frame bound B of a frame $\{\psi_k\}_{k\in\mathbb{K}}$ is greater than $\max_{k\in\mathbb{K}} \|\psi_k\|^2$.

The *redundancy* of a frame with $|\mathbb{K}|$ vectors for a space $\mathcal{W}$ is defined as the ratio

$$r \triangleq \frac{|\mathbb{K}|}{\dim \mathcal{W}}$$

It is easy to show that, for a normalized tight frame, $r = A$, where A is the lower frame bound in (55).
Another important property of the elements of a frame $\{\psi_k\}_{k\in\mathbb{K}}$ is that they are also a *Bessel sequence*, i.e., they satisfy

$$\left\| \sum_{k\in\mathbb{K}} c[k]\psi_k \right\|_{\mathcal{W}}^2 < B \|c\|_{\ell^2}^2 , \quad \forall c \in \ell^2 \tag{56}$$

where B is the upper frame bound in (55).
From remark 2 and the above properties, orthogonal bases are a special type of Riesz basis, whilst Riesz bases are exact frames. Thus, by basing our analysis on frames, one is also including orthogonal and Riesz bases as special cases.

A.4 Frames and their Operators

Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{W} = \overline{\text{span}}\{\psi_k\}_{k\in\mathbb{K}} \subseteq \mathcal{H}$.

Definition 5 (Synthesis Operator)**.** *The* synthesis *(or* preframe*) operator for a frame* $\{\psi_k\}_{k\in\mathbb{K}}$ *is defined as*

$$\Psi : \ell^2 \mapsto \mathcal{H} , \quad \Psi \{c[k]\}_{k\in\mathbb{Z}} = \sum_{k\in\mathbb{K}} c[k]\, \psi_k.$$

Since every frame sequence is a Bessel sequence (see (56)), the synthesis operator for a frame with frame bounds A, B is bounded, with operator norm $\|\Psi\| = B$, i.e., B is the minimum constant such that $\|\Psi c\|_{\mathcal{W}}^2 \leq B \|c\|_{\ell^2}^2$, $\forall c \in \ell^2$.

Definition 6 (Analysis Operator)**.** *The* analysis *operator for a frame* $\{\psi_k\}_{k\in\mathbb{K}}$ *is defined as*

$$\Psi^* : \mathcal{H} \mapsto \ell^2 , \quad \Psi^* h = \{\langle h, \psi_k\rangle\}_{k\in\mathbb{K}}$$

Remark 3. *The analysis operator* Ψ^* *is the* adjoint *of* Ψ*, i.e., it satisfies* $\langle w, \Psi c\rangle = \langle \Psi^* w, c\rangle$, $\forall c \in \mathcal{R}(\Psi^*)$, $\forall w \in \mathcal{W}$.

Definition 7 (Dual Frame)**.** *Let* $\{\psi_k\}_{k\in\mathbb{K}}$ *be a frame for a Hilbert space* $\mathcal{W}$*. Another frame for* $\mathcal{W}$*, namely,* $\{g_k\}_{k\in\mathbb{K}}$ *that satisfies*

$$w = \sum_{k\in\mathbb{K}} \langle w, g_k\rangle \psi_k, \quad \forall w \in \mathcal{W} \tag{57}$$

is said to be a dual frame *of* $\{\psi_k\}_{k\in\mathbb{K}}$ *in* $\mathcal{W}$*.*

As can be seen in (57), a dual frame provides an explicit method for representing any signal $w \in \mathcal{W}$ in terms of coefficients (samples), from which w can be exactly recovered through the synthesis frame $\{\psi_k\}_{k\in\mathbb{K}}$.

Definition 8 (Frame Operator)**.** *The* frame operator *of a frame* $\{\psi_k\}_{k\in\mathbb{K}}$ *is defined as*

$$S : \mathcal{H} \mapsto \mathcal{H} , \quad Sh = \Psi\Psi^* h = \sum_{k\in\mathbb{K}} \langle h, \psi_k\rangle \psi_k \tag{58}$$

Lemma 2 (from [80, Lemma 5.1.5])**.** *Let* $\{\psi_k\}_{k\in\mathbb{K}}$ *be a frame with frame operator* S *and frame bounds* A, B*. Then the following holds:*

(i) S is bounded, invertible, self-adjoint, and positive.

(ii) $\left\{S^{-1}\psi_k\right\}_{k\in\mathbb{Z}}$ *is a frame with bounds* B^{-1}*,* A^{-1}*. The frame operator for* $\left\{S^{-1}\psi_k\right\}_{k\in\mathbb{Z}}$ *is* S^{-1}

Since $\|\Psi w\|^2 = \langle Sw, w\rangle$, one can derive from Lemma 2 , (55) and (56) that:

$$A\|w\| \leq \|Sw\| \leq B\|w\| \tag{59}$$
$$B^{-1}\|w\| \leq \left\|S^{-1}w\right\| \leq A^{-1}\|w\| \tag{60}$$

The frame operator defined in (58) is of particular importance for the problem of sampling and reconstruction, since it provides an explicit way to obtain a dual frame (see (57)). More precisely, with S as defined in (58), if $\{\psi_k\}_{k\in\mathbb{K}}$ is a frame for $\mathcal{W}$, then the frame $\left\{S^{-1}\psi_k\right\}_{k\in\mathbb{Z}}$ is a dual frame for $\{\psi_k\}_{k\in\mathbb{K}}$ in $\mathcal{W}$, i.e.

$$w = \sum_{k\in\mathbb{K}} \langle w, S^{-1}\psi_k\rangle \psi_k, \quad \forall w \in \mathcal{W} \tag{61}$$

and

$$w = \sum_{k\in\mathbb{K}} \langle w, \psi_k\rangle S^{-1}\psi_k, \quad \forall w \in \mathcal{W} \tag{62}$$

The frame $\left\{S^{-1}\psi_k\right\}_{k\in\mathbb{Z}}$ is called the *canonical dual frame* of $\{\psi_k\}_{k\in\mathbb{K}}$ in $\mathcal{W}$. This is a reciprocal relation, i.e., $\{\psi_k\}_{k\in\mathbb{K}}$ is the canonical dual of $\left\{S^{-1}\psi_k\right\}_{k\in\mathbb{Z}}$ in $\mathcal{W}$ as well.

A.5 Noise Reduction by Redundancy of the Frame

If the frame coefficients $\{\langle w, \psi_k \rangle\}_{k\in\mathbb{K}}$ in (62) were contaminated by additive noise $e[k]$, $k \in \mathbb{K}$, then the reconstruction formula (62) would yield a reconstruction error

$$w_e \triangleq \sum_{k\in\mathbb{K}} \left(\langle w, \psi_k \rangle + e[k]\right) S^{-1}\psi_k - w = \sum_{k\in\mathbb{K}} e[k] S^{-1}\psi_k \tag{63}$$

Early references to the fact that the redundancy of the frame reduces the reconstruction error were provided in [24], whilst proofs can be found in [32] and [39]. Due to the importance of this property of redundant frames, we present next an adaptation of the result in [32], which is also illustrative of the importance of the frame bounds.

Proposition 3. *Let $\{\psi_k\}_{k\in\mathbb{K}}$ be a frame of unit-norm vectors with frame bounds $0 < A \leq B$, and let $e[k]$, $k \in \mathbb{K}$ be a sequence of independent random variables with mean zero and variance σ^2. Then the mean square value of w_e in* (63) *satisfies*

$$\frac{|\mathbb{K}|\sigma^2}{B^2} \leq E\left[\|w_e\|_{L^2}^2\right] \leq \frac{|\mathbb{K}|\sigma^2}{A^2}$$

Proof. If $e[k]$, $k \in \mathbb{K}$ is a sequence of independent random variables with zero mean and variance σ^2, we have

$$\mathrm{E}\left[\|w_e\|_{L^2}^2\right] = \mathrm{E}\left[\left\|\sum_{k\in\mathbb{K}} e[k] S^{-1}\psi_k\right\|^2\right] = \sigma^2 \sum_{k\in\mathbb{K}} \left\|S^{-1}\psi_k\right\|^2 \tag{64}$$

From (60) one can derive that

$$B^{-2}\|\psi_k\|^2 \leq \left\|S^{-1}\psi_k\right\|^2 \leq A^{-2}\|\psi_k\|^2$$

which simplifies to

$$B^{-2} \leq \left\|S^{-1}\psi_k\right\|^2 \leq A^{-2} \tag{65}$$

because $\{\psi_k\}_{k\in\mathbb{K}}$ is a normalized frame. Combining (64) with (65) gives the result. □

Corollary 2. *If the frame in Proposition 3 is also tight, then*

$$E\left[\|w_e\|_{L^2}^2\right] = \frac{(\dim \mathcal{W})\,\sigma^2}{r}$$

References

[1] D. Robertson, "Alec Reeves 1902–1971," Available from http://www.privateline.com/TelephoneHistory2/reeves.html, 2002.

[2] A. H. Reeves, French patent 852,185, 3rd October 1938, (Invention patent for PCM). Assigned to ITT.

[3] P. Antsaklis and J. Baillieul, "Guest editorial special issue on networked control systems," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1421–1423, Sept. 2004.

[4] G. C. Goodwin, D. E. Quevedo, and E. I. Silva, "An introduction to networked control systems," in *Proc. Asian Control Conference, Bali, Indonesia*, 2006.

[5] ——, "Filter banks in networked control," in *Proc. 17th Int. Symp. Mathematical Theory of Networks and Systems (MTNS2006), Kyoto, Japan*, 2006.

[6] C. E. Shannon, "Communication in presence of noise," in *Proc. IRE*, vol. 37, 1949, pp. 10–21.

[7] D. Slepian, "On bandwidth," *Proc. IEEE*, vol. 64, no. 3, pp. 292–300, March 1976.

[8] H. Helms and J. Thomas, "Truncation error of sampling-theorem expansions," *Proc. IRE.*, vol. 50, pp. 179–184, February 1962.

[9] A. Aldroubi and M. Unser, "Sampling procedures in function spaces and asymptotic equivalence with Shannon's sampling theory," *Numer. Funct. Anal. Optimizat.*, vol. 15,, no. 1-2, pp. 1–21, Feb 1994.

[10] M. Unser and J. Zerubia, "A generalized sampling theory without bandlimiting constraints," *IEEE Trans. Circuits Syst.*, vol. 45, no. 8, pp. 959–969, Aug. 1998.

[11] M. Unser and A. Aldroubi, "A general sampling theory for nonideal acquisition devices," *IEEE Trans. Signal Processing*, vol. 42, no. 11, pp. 2915–2925, Nov. 1994.

[12] Y. C. Eldar and T. Dvorkind, "A minimum squared-error framework for generalized sampling," to appear in IEEE Trans. Signal Proc.

[13] Y. C. Eldar, "Sampling without input constraints: consistent reconstruction in arbitrary spaces," in *Sampling, wavelets and tomography*, J. Benedetto and A. Zayed, Eds. Brikhäuser, 2004, pp. 33–59.

[14] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana: Univ. of Illinois Press, 1949.

[15] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, Jul. 1979.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic, 1992.

[17] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[18] J. Benedetto, "Irregular sampling and frames," in *Wavelets-a tutorial in theory and applications*, C. Chui, Ed. Boca Raton, FL.: CRC Press, 1992, ch. VII, pp. 445–507.

[19] Y. C. Eldar, "Sampling with arbitrary sampling and reconstruction spaces and oblique dual frame vectors," *J. Fourier Anal. Appl.*, vol. 9, no. 1, pp. 77–96, 2003.

[20] M. Vetterli and J. Kovačević, *Wavelets and subband coding*, A. V. Oppenheim, Ed. Prentice Hall, 1995.

[21] F. Beutler, "Sampling theorems and bases in a Hilbert space," *Information and Control*, vol. 4, pp. 97–117, 1961.

[22] K. Yao, "Applications of reproducing kernel Hilbert spaces - bandlimited signal models," *Information and Control*, vol. 11, pp. 429–444, 1967.

[23] H. Ogawa, "A unified approach to generalized sampling theorems," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 11, April 1986, pp. 1657–1660.

[24] I. Daubechies, *Ten lectures on wavelets*. Philadephia, PA: Society for industrial and applied mathematics, 1992.

[25] G. Walter, "A sampling theorem for wavelet subspaces," *IEEE Trans. Inform. Theory*, vol. 38, pp. 881–884, March 1992.

[26] M. Unser, "Sampling – 50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, April 2000.

[27] P. Vaidyanathan, "Sampling theorems for nonbandlimited signals," in *Sampling, wavelets, and tomography*, J. Benedetto and A. Zayed, Eds. Brikhäuser, 2004, pp. 115–135.

[28] A. Zayed, "A prelude to sampling, wavelets, and tomography," in *Sampling, wavelets and tomography*, J. Benedetto and A. Zayed, Eds. Brikhäuser, 2004, pp. 1–30.

[29] Y. C. Eldar, "Reconstruction from finitely many samples in the presence of noise," *Int. workshop on sampling theory and app. (SampTA 2005)*.

[30] S. Ramani, D. Van De Ville, and M. Unser, "Sampling in practice: is the best reconstruction space bandlimited?" in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, 2005, pp. 153–156.

[31] Y. C. Eldar and A. Oppenheim, "Nonredundant and redundant sampling with arbitrary sampling and reconstruction spaces," in *Proc. of SampTA*, May 2001, pp. 229–234.

[32] V. K. Goyal, M. Vetterli, and N. T. Thao, "Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms," *IEEE Trans. Inform. Theory*, vol. 44, pp. 16–31, 1998.

[33] H. Bölcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 155–172, Jan. 2001.

[34] A. Aldroubi and K. Gröochenig, "Nonuniform sampling and reconstruction in shift-invariant spaces," *SIAM Review*, vol. 43, no. 4, pp. 585–620, 2001.

[35] J. Benedetto and S. Li, "The theory of multiresolution analysis frames and applications to filter banks," *Appl. Comp. Harm. Anal.*, vol. 5, pp. 389–427, 1998.

[36] H. Bölcskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Processing*, vol. 46, no. 12, pp. 3256–3268, Dec. 1998.

[37] L. Condat, T. Blu, and M. Unser, "Beyond interpolation: Optimal reconstruction by quasi-interpolation," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2005, pp. 33–36.

[38] A. Aldroubi, M. Unser, and M. Eden, "Cardinal spline filters: Stability and convergence to the ideal sinc interpolator," *Signal Process.*, vol. 28, no. 2, pp. 127–138, 1992.

[39] S. Mallat, *A wavelet tour of signal processing*, 2nd ed. Academic Press, 1999.

[40] O. Christensen and Y. C. Eldar, "Oblique dual frames and shift-invariant spaces," *Appl. Comput. Anal.*, vol. 17, no. 1, pp. 46–48, Jul. 2004.

[41] Y. C. Eldar and T. Dvorkind, "Minmax sampling with arbitrary spaces," in *Proc. IEEE Int. Conf. Electronic Circuit Syst.*, December 2004, pp. 559–562.

[42] Y. C. Eldar and T. Werther, "General framework for consistent sampling in Hilbert spaces," *In. J. Wavelets, Multiresolution and Inf. Proc.*, vol. 3, no. 4, pp. 497–509, 2005.

[43] P. Vaidyanathan and B. Vrcelj, "On sampling theorems for non bandlimited signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 6, May 2001, pp. 3897–3900.

[44] A. L. Brown and A. Page, *Elements of Functional Analysis*, E. Davies, Ed. London: Van Nostrand Reinohold, 1970.

[45] P. Elias, "Bounds on performance of optimum quantizers," *IEEE Trans. Inform. Theory*, vol. IT-16, pp. 172–184, Mar. 1970.

[46] A. B. Sripad and D. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white." *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 5, pp. 442–448, Oct. 1977.

[47] W. R. Bennet, "Spectrum of quantized signals," *Bell Syst. Tech J.*, vol. 27, pp. 446–472, July 1948.

[48] A. Gersho, "Principles of quantization," *IEEE Trans. Circuits Syst.*, vol. 25, no. 7, pp. 427– 436, Ju. 1978.

[49] Cvetković, "Resilience properties of redundant expansions under additive noise and quantization," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 644–656, Mar. 2003.

[50] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, pp. 611–632, 1974.

[51] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta–Sigma Data Converters: Theory, Design and Simulation*. Piscataway, N.J.: IEEE Press, 1997.

[52] S. K. Tewksbury and R. W. Hallock, "Oversampled, linear predictive and noise-shaping coders of order $N > 1$," *IEEE Trans. Circuits Syst.*, vol. 25, no. 7, pp. 436–447, July 1978.

[53] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, no. 6, pp. 1220–1244, Nov. 1990.

[54] N. T. Thao and S. Güntürk, "Generalized spectral theory for $\Sigma\Delta$ quantization with constant inputs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2006.

[55] A. G. Clavier, P. F. Panter, and D. D. Grieg, "Distortion in a pulse count modulation system," *AIEE Trans.*, vol. 66, pp. 989–1005, 1947.

[56] ——, "PCM distortion analysis," *Elect. Eng.*, pp. 1110–1122, Nov. 1947.

[57] N. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Processing*, vol. 42, pp. 519–531, Mar. 1994.

[58] O. Feely, "A tutorial introduction to non-linear dynamics and chaos and their application to Sigma-Delta modulators," *Int. J. Circuit Theory Appl.*, vol. 25, pp. 347–367, 1997.

[59] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, vol. 51, pp. 1524–1535, Nov. 1963.

[60] M. M. Serón, J. H. Braslavsky, and G. C. Goodwin, *Fundamental Limitations in Filtering and Control*. Springer-Verlag, London, 1997.

[61] G. C. Goodwin, S. F. Graebe, and M. E. Salgado, *Control System Design*. Prentice-Hall, 2001.

[62] M. Gerzon and P. G. Craven, "Optimal noise shaping and dither of digital signals," in *87th Convention of the AES, New York, NY, preprint 2822*, Oct. 1989.

[63] G. C. Goodwin, M. M. Serón, and J. A. De Doná, *Constrained Control & Estimation – An Optimization Perspective*. London: Springer-Verlag, 2005.

[64] C. S. Güntürk, "One-bit sigma-delta quantization with exponential accuracy." *Communications on pure and appl. math.*, vol. LVI, pp. 1608–1630, 2003.

[65] Z. Cvetković and M. Vetterli, "Error-rate characteristics of oversampled analog-to-digital conversion," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1961–1964, 1998.

[66] Z. Cvetković and I. Daubechies, "Single-bit oversampled A/D compression with exponential accuracy in the bit-rate," in *Proc. Data Comp. Conf.*, Mar 2000.

[67] J. M. Maciejowski, *Predictive Control with Constraints*. Prentice-Hall, 2002.

[68] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Optimality and stability," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[69] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Contr. Eng. Pract.*, vol. 11,

pp. 733–764, 2003.

[70] D. E. Quevedo, G. C. Goodwin, and J. A. De Doná, "Finite constraint set receding horizon quadratic control," *Int. J. Robust Nonlin. Contr.*, vol. 14, no. 4, pp. 355–377, Mar. 2004.

[71] D. E. Quevedo and G. C. Goodwin, "Multistep optimal analog-to-digital conversion," *IEEE Trans. Circuits Syst. I*, vol. 52, Issue 3, pp. 503–515, March 2005.

[72] L. Lo Presti, "Efficient modified-sinc filters for Sigma–Delta A/D converters," *IEEE Trans. Circuits Syst. II*, vol. 47, no. 11, pp. 1204–1213, Nov. 2000.

[73] C. Dunn and M. Sandler, "Psychoacoustically optimal Sigma–Delta modulation," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 212–223, Apr. 1997.

[74] G. C. Goodwin, D. E. Quevedo, and D. McGrath, "Moving-horizon optimal quantizer for audio signals," *J. Audio Eng. Soc.*, vol. 51, no. 3, pp. 138–149, Mar. 2003.

[75] S. O. Aase, J. H. Husøy, K. Skretting, and K. Engan, "Optimized signal expansions for sparse representation," *IEEE Trans. Signal Processing*, vol. 49, no. 5, pp. 1087–1096, May 2001.

[76] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[77] A. N. Akansu and R. A. Haddad, *Multiresolution Signal Decomposition: Transforms, Subbands, and Wavelets*. San Diego, CA: Academic Press, 1992.

[78] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, New Jersey: Prentice-Hall, 1993.

[79] M. S. Derpich, D. E. Quevedo, and G. C. Goodwin, "Optimal A-D conversion via sampled-data receding horizon control theory." Chinese Control Conference 2006 (to appear).

[80] O. Christensen, *An introduction to frames and Riesz bases*. Boston, MA: Birkhäuser, 2003.

[81] M. S. Derpich, D. E. Quevedo, G. C. Goodwin, and A. Feuer, "Quantization and sampling of not necessarily band-limited signals," to appear in Proc. of the International Conf. on Audio, Speech and Signal Proc. ICASSP-2006.

[82] H. Ishii and B. A. Francis, *Limited Data Rate in Control Systems with Networks*. Springer, 2002.

[83] D. E. Quevedo, J. S. Welsh, G. C. Goodwin, and M. McLeod, "Networked PID control," in *Proc. IEEE Conf. Contr. Appl.*, 2006.

[84] G. C. Goodwin, H. Haimovich, D. E. Quevedo, and J. S. Welsh, "A moving horizon approach to networked control system design," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1427–1445, Sept. 2004.

[85] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series." *Trans. Amer. Math. Soc.*, vol. 72, pp. 341–366, 1952.