On Optimal Perfect Reconstruction Feedback Quantizers

Milan S. Derpich, Eduardo I. Silva, Daniel E. Quevedo, *Member, IEEE*, and Graham C. Goodwin, *Fellow, IEEE*
School of Electrical Engineering and Computer Science,
The University of Newcastle, NSW 2308, Australia
{milan.derpich, eduardo.silva}@studentmail.newcastle.edu.au, dquevedo@ieee.org,
graham.goodwin@newcastle.edu.au.

***Abstract*— This paper presents novel results on Perfect Reconstruction Feedback Quantizers (PRFQs), i.e., noise-shaping, predictive and sigma-delta A/D converters whose signal transfer function is unity. Our analysis of this class of converters is based upon an additive white noise model of quantization errors. Our key result is a formula that relates the minimum achievable MSE of such converters to the signal-to-noise ratio (SNR) of the scalar quantizer embedded in the feedback loop. This result allows us to obtain analytical expressions that characterize the corresponding optimal filters. We also show that, for a fixed SNR of the scalar quantizer, the end-to-end MSE of an optimal PRFQ which uses the optimal filters (which for this case turn out to be IIR) decreases exponentially with increasing oversampling ratio. Key departures from earlier work include the fact that fed back quantization noise is explicitly taken into account and that the order of the converter filters is not a-priori restricted.**

***Index Terms*— Differential pulse code modulation, optimization, quantization, sigma-delta modulation, source coding.**

I. Introduction

The term *Feedback Quantizer* (FQ) refers to a class of Analog-to-Digital converter architectures wherein a scalar quantizer is placed within a linear feedback loop. Well known examples of FQs include Δ-Modulators, DPCM converters [1] and Sigma-Delta modulators [2]. The latter schemes have been very successfully applied in a number of areas, including audio compression [1], [3], oversampled A/D conversion [2], [4], sub-band coding [5], digital image half-toning [6], power conversion [7], and control over networks [8].

Fig. 1 depicts a general FQ configuration. In this scheme, $\mathcal{Q}$ may take the form of a non-uniform or a uniform quantizer [9], the latter being either dithered or undithered[1] [10].

The filters $A(z)$ and $B(z)$ in an FQ system allow one to exploit the predictability of the input signal so as to reduce the variance of $\{v(k)\}_{k\in\mathbb{Z}}$. When compared with simple PCM conversion, this flexibility allows one to use a scalar quantizer with a smaller quantization step. The error-feedback filter $F(z)$ opens the possibility of spectrally shaping the effect of quantization errors on the output. In this way, one can allocate more of the quantization noise in the frequency bands where it is less harmful from a user's point of view. Accordingly, it is convenient to use a frequency weighted error criterion, via an *error frequency weighting filter* $P(z)$, and to focus on the *frequency weighted MSE* (FWMSE) (see discussion in [3], [11]).

For the sake of generality, we consider the possible use of a clipper before $\mathcal{Q}$. This device limits the value of the quantizer input signal v' so that $v' = v$ if $|v| \leq s$, and $v' = \frac{v}{|v|}s$ if $|v| > s$, where $s > 0$ is the *saturation threshold* of the clipper. This clipping technique can be used to keep $\mathcal{Q}$ from overloading, which is helpful in reducing limit-cycle oscillations (idle tones) in an FQ with high order filters, as proposed in [4]. On the other hand, if we chose s to be sufficiently large, then $v' = v$, and the clipper has no effect on the system.

If the characteristics of $\mathcal{Q}$ and the spectral properties of the input signal x are known, then the design of an FQ converter that minimizes the variance of ϵ amounts to choosing the filters $A(z)$, $B(z)$ and $F(z)$.

It is often desirable that a converter is transparent to the system in which it is inserted. This corresponds to the widespread paradigm in which the coding scheme adapts to the application that employs it, without need to modify the latter. A transparent converter is one whose signal transfer function (i.e., the transfer function from input x to output $\tilde{x}$) is unity at the frequencies of interest. The design of such *Perfect Reconstruction Feedback Quantizers* (PRFQs) constitutes the main topic of the present work. PRFQs are characterized by the property that, in the absence of quantization effects, there is no frequency weighted reconstruction error, i.e., $P(z)\tilde{x} = P(z)x$. If we denote the *power spectral density* (PSD) of x by $S_x(\mathrm{e}^{j\omega})$, then it can be seen from Fig. 1 that the latter holds if and only if

$$A(\mathrm{e}^{j\omega})B(\mathrm{e}^{j\omega}) = 1, \ \ \forall\omega \text{ such that } S_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega}) \neq 0, \quad (1)$$

Thus, in the design of an optimal PRFQ converter, only two degrees of freedom are available: the filters $F(z)$ and $A(z)$ (or, alternatively, $F(z)$ and $B(z)$).[2]

To the best of our knowledge, existing results on optimal filter design for PRFQ converters either consider finite order filters [2], [12], [13], assume (or require) that the variance of the signal $y \triangleq F(z)n$ is much smaller than that of v [1],

[1] In this case, the block $\mathcal{Q}$ in Fig. 1 represents the scalar quantizer including the dither signals.

[2] We note that this reduction in the number of degrees of freedom (in comparison with an FQ with no perfect reconstruction constraint) by no means makes the design optimization problem easier to solve. Moreover, perfect reconstruction constitutes an additional constraint that can not be added "a posteriori", i.e., after the optimization is completed.

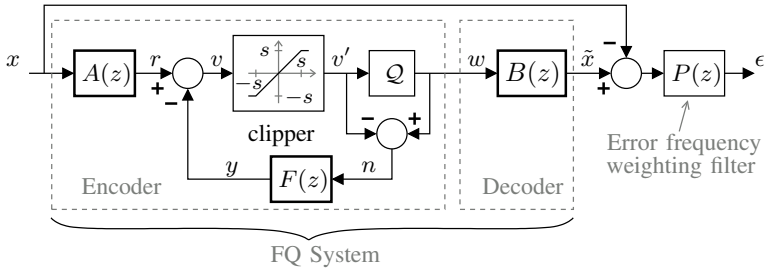


Fig. 1: Feedback Quantization system and frequency weighting filter.

[4], [14], [15], or have a heuristic component in the optimization [2], [3], [13], [16]–[19]. The only explicit analytical expressions currently available for the optimal performance (and corresponding filter frequency responses) of a PRFQ converter are those given in [14]. However, the assumption of negligible fed back quantization errors used in [14] makes these filters sub-optimal. Indeed, as we will show in the sequel, there exist situations where the filters proposed in [14] yield large fed back quantization error, *even when a fine step scalar quantizer is used*. In these situations, not only is the main assumption in [14] violated, but also an FWMSE much larger than predicted can result due to excessive quantizer overload (see, e.g., [2], [13]).

In the present work, we will show how to design optimal PRFQ converters. For this purpose, as in [12], [14], [16]–[18], we model the scalar quantizer as a linear device that introduces additive white noise whose variance is proportional to that of the signal being quantized. A key departure from [14], however, is that we explicitly take into account fed back quantization noise in the feedback loop. Our main contributions are: i) We derive one-parameter equations that relate the minimum achievable frequency weighted MSE to the *signal-to-noise ratio* (SNR) of $\mathcal{Q}$; ii) We show, within our model, that the frequency weighted MSE in an optimal PRFQ where the SNR of $\mathcal{Q}$ is fixed decreases exponentially with oversampling ratio; and iii) We derive equations that characterize the optimal filters for a PRFQ. Our results can be applied to any given number of quantization levels, and to almost arbitrary input spectra and frequency weighting criteria.

The remainder of this paper is organized as follows: In Section II, we present our analysis model for PRFQ converters. In Section III, we formulate the associated optimization problem. Section IV presents a one-parameter characterization of the solution. In Section V we discuss the main properties of an optimized PRFQ. The case of oversampled FQ is analyzed in Section VI. Section VII discusses the relationship to previous results and highlights the importance of taking account of fed back quantization noise. Section VIII presents simulation results. Section IX draws conclusions. (For ease of exposition, all proofs of our results are included in the Appendix.)

Preliminaries and Notation

We write "iff" as a short hand expression for "if and only if". The sets of all complex-valued square integrable and absolutely integrable functions on $[-\pi,\pi]$ are denoted by L^2 and L^1, respectively. Given $f(\omega)$, $g(\omega) \in L^2$, we adopt the standard inner product $\langle f,g\rangle \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)^* g(\omega) d\omega$, where $(\cdot)^*$ denotes complex conjugation. We denote the corresponding 2-norm as $\|f\| \triangleq \sqrt{\langle f,f\rangle}$. We use z as the argument of the z-transform. If $F(z)$ is a transfer function, then we use the short hand notation F to refer to the associated frequency response $F(\mathrm{e}^{j\omega})$. If I is a set, then we write "a.e. on I" (almost everywhere on I) for "everywhere on I, except on a zero Lebesgue measure subset of I". We use σ_x^2 to denote the variance of a given wide sense stationary (w.s.s.) random process $\{x(k)\}_{k\in\mathbb{Z}}$, having PSD $S_x(\mathrm{e}^{j\omega})$. We recall that if x has zero mean, then $\sigma_x^2 \triangleq \mathrm{E}\big[x(k)^2\big] = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\mathrm{e}^{j\omega})d\omega = \|\Omega_x\|^2$, where Ω_x is a frequency response satisfying $\big|\Omega_x(\mathrm{e}^{j\omega})\big| = \sqrt{S_x(\mathrm{e}^{j\omega})}$, $\forall\omega\in[-\pi,\pi]$. For any functions $f(\omega)$ or $F(\mathrm{e}^{j\omega})$ we write $\mathcal{N}_f$ and $\mathcal{N}_F$ to denote the sets $\{\omega\in[-\pi,\pi] : f(\omega) = 0\}$ and $\{\omega\in[-\pi,\pi] : F(\mathrm{e}^{j\omega}) = 0\}$, respectively.

To simplify notation, we introduce the operator $(\cdot)^{\sim 1}$, defined as follows:

$$F(\mathrm{e}^{j\omega})^{\sim 1} = \begin{cases} F(\mathrm{e}^{j\omega})^{-1} & , \quad \forall\omega \notin \mathcal{N}_F \\ \natural & , \quad \forall\omega \in \mathcal{N}_F, \end{cases} \tag{2}$$

where $F : \mathbb{C} \to \mathbb{R}$ is any given function and $\natural$ denotes any arbitrary and positive bounded value. For later use, we also recall the following definition:

***Definition** 1 (Almost Constant Function): A function $f : [-\pi,\pi] \to \mathbb{R}$ is said to be almost constant iff*

$$\int_{-\pi}^{\pi} \left| f(x) - \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)d\omega \right| dx = 0. \tag{3}$$

▲

II. PRFQ Converter Model

In this section we discuss some of the main aspects of feedback quantization. We also describe the analysis model and the constraints to be considered later in the search for the optimal filters.

A. Feedback Quantizer Equations

We begin by presenting the equations that describe the behaviour of the PRFQ shown in Fig. 1.

1) Quantization and Clipping Errors: From Fig.1, the quantization error n is given by

$$n(k) \triangleq w(k) - v'(k). \tag{4}$$

Every practical scalar quantizer has an associated constant $V > 0$ such that, if $|v'| > V$, then $\mathcal{Q}$ is said to be

overloaded. When the quantizer is not overloaded, then $n(k)$ is only *granular* quantization error, namely $\varrho(k)$, which can be bounded as $|\varrho(k)| \leq \varrho_{max}$, $\forall v'(k) \in \mathbb{R}$, for some $0 < \varrho_{max} < 2V$ (see, e.g., [9]). For example, if $\mathcal{Q}$ is a symmetric, uniform, non-dithered quantizer with N levels and quantization interval Δ, then one needs $V \leq N\Delta/2$ in order to obtain $\varrho_{max} = \frac{\Delta}{2}$.

In general, we can write

$$n(k) = \varrho(k) + \tau(k), \tag{5}$$

where

$$\tau(k) \triangleq v(k)' - \frac{v(k)'}{|v(k)'|} \min\{V, |v'(k)|\}$$

is the *overload* error. Clearly overload errors are bounded as $|\tau(k)| < |v'(k)| \leq |v(k)|$, but they cannot be bounded by a constant unless v' is bounded.

As outlined in the introduction, the clipper in Fig. 1 can be used to keep $\mathcal{Q}$ from overloading. For simplicity, we will only consider here two possibilities, namely, that $s = V$, or else $s = \infty$. The former choice guarantees that $\mathcal{Q}$ does not overload, since *clipping error*, defined as

$$\vartheta(k) \triangleq v'(k) - v(k), \quad \forall k \in \mathbb{Z}, \tag{6}$$

takes place instead. More precisely, if $s = \infty$ we have that $\vartheta(k) = 0$, and $\tau(k) = v(k) - \frac{v(k)}{|v(k)|} \min\{V, |v(k)|\}$. If, instead, $s = V$, then the latter revert to $\vartheta(k) = v(k) - \frac{v(k)}{|v(k)|} \min\{V, |v(k)|\}$ and $\tau(k) = 0$. A key point in using clipping is that, unlike overload errors, clipping errors are not fed back into $\mathcal{Q}$ through $F(z)$. This helps to avoid large limit-cycle oscillations arising from the overload of $\mathcal{Q}$, see [4]. Since such oscillations are not part of the analysis model we will use, their occurrence could increase the FWMSE significantly above the value predicted by the model.

Using the above definitions, and from Fig. 1, we can write

$$w(k) = v(k) + n(k) + \vartheta(k), \tag{7}$$

which reveals that w differs from v by the sum of the quantization and clipping errors.

2) Transfer Functions: From Fig. 1 and (7) we have that

$$v = A(z)x - F(z)n, \tag{8a}$$
$$\tilde{x} = B(z)A(z)x + B(z)[1 - F(z)]n + B(z)\vartheta, \tag{8b}$$
$$\epsilon = P(z)B(z)[1 - F(z)]n + P(z)B(z)\vartheta. \tag{8c}$$

Notice that these equations are exact and require no assumptions on the signals involved. From (8b) one can see that $A(z)B(z)$ corresponds to the *signal transfer function* (STF), from x to $\tilde{x}$, of the converter. Similarly, the product $B(z)[1-F(z)]$ is the transfer function for quantization errors, usually referred to as the *noise transfer function* (NTF) of the converter[3]. The term $[1 - F(z)]$ will play a crucial role in the derivation of the optimal filters in Section IV.

[3]In noise-shaping and $\Sigma\Delta$ literature, where $B(z)$ is typically a unit gain, the term NTF is normally used for $1 - F(z)$.

3) Stability: We say that a PRFQ is *Bounded-Input-Bounded Output* (BIBO) stable iff for any input sequence x satisfying $\|x\|_\infty \leq x_{max} < \infty$ all the signals in the converter are bounded.

If $s = V$ or if $\mathcal{Q}$ has infinitely many quantization levels, then $|n| \leq \varrho_{max}$, $\forall k \in \mathbb{Z}$, and thus all the other signals in the converter are bounded. On the other hand, if $s = \infty$, then v can be written as

$$v = \frac{A(z)}{1 - F(z)}x - \frac{F(z)}{1 - F(z)}w. \tag{9}$$

If the quantizer has a finite number of quantization levels, then w is bounded. If $F(z)$ is stable and $1 - F(z)$ is minimum-phase, then it follows from (9) that v is bounded. This in turn guarantees that n and all the other signals in the converter are bounded (see (4) and (8)). Summarizing, if all the filters in Fig. 1 are stable, and if $1 - F(z)$ has no zeros on or outside the unit circle, then the resulting PRFQ is BIBO stable.

In addition, if $A(z)$ and $F(z)$ are stable, then the ℓ_∞ norm of their impulse responses, namely A_∞ and F_∞, are bounded. Thus, if there exists a bounded $x_{max} > 0$ such that $|x(k)| \leq x_{max} < \infty$, $\forall k \in \mathbb{Z}$, then a sufficient condition to ensure $\tau(k) = \vartheta(k) = 0$, $\forall k \in \mathbb{Z}$, is that $V \geq V_{min} < \infty$, where

$$V_{min} \triangleq A_\infty x_{max} + F_\infty \varrho_{max}. \tag{10}$$

Thus, for a uniform quantizer with quantization interval Δ, it suffices to have V_{min}/Δ or more quantization levels in order to avoid clipping or overload errors.

B. Assumptions

The assumptions associated with our PRFQ model are described next.

1) Input Spectrum and Frequency Weighting: The error weighting filter $P(z)$ in Fig. 1 models the impact that reconstruction errors have at each frequency. This "performance assessment" filter is application dependent, and is assumed to be stable and given. The input signal $\{x(k)\}_{k\in\mathbb{Z}}$ is a zero-mean w.s.s. stochastic process[4] with known PSD $S_x(\omega) = \left|\Omega_x(\mathrm{e}^{j\omega})\right|^2$ and finite power, i.e., $\|\Omega_x\|^2 < \infty$. In order to simplify our subsequent analysis, we shall further restrict Ω_x and $P(z)$ to satisfy the following:

***Assumption** 1: The product $|\Omega_x P|$ is a piece-wise differentiable function having at most a finite number of discontinuities and satisfying $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right| < \infty$, $\forall\omega \in [-\pi, \pi]$. In addition, $|\Omega_x P|$ is such that one[5] of the following conditions holds:*

i) *There exists a constant $g_{min} > 0$ such that $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right| > g_{min}$, for all $\omega \in [-\pi, \pi]$, or*
ii) *$\exists\omega \in [-\pi, \pi]$ such that $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right| = 0$. Furthermore, if $\{\Gamma_i\}$ denotes the set of non-contiguous and non-overlapping intervals in $[-\pi, \pi]$ such that $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right| = 0 \Leftrightarrow \omega \in \bigcup_i \Gamma_i$, then, for every i, $\exists\zeta_i \in \Gamma_i$ such that $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right|$ is $\mathcal{O}(\omega - \zeta_i)$ as $\omega \to \zeta_i$.* ▲

[4]This excludes, for example, sinusoids or constant inputs from the analysis.
[5]Notice that conditions i) and ii) can not be met simultaneously.

We note that the above is a rather weak constraint, since conditions i) and ii) include almost any product $|\Omega_x P|$ of practical or theoretical interest. In particular, condition i) covers all the cases where the product $\Omega_x(z)P(z)$ has no zeros on the unit circle. In turn, condition ii) is satisfied if $P\Omega_x$ is zero over any interval on $[-\pi, \pi]$ having non-zero measure, or if $\Omega_x(z)P(z)$ is rational and has zeros on the unit circle.

2) The Quantizer: We shall focus our analysis on the effect that granular quantization errors have on the FWMSE. For this effect to closely represent the actual FWMSE, we need to assume the following:

***Assumption** 2: The variances of overload and clipping errors are negligible, i.e.,*

$$\sigma_\tau^2 \ll \sigma_n^2, \quad \textit{if } s = \infty, \textit{ or} \tag{11a}$$

$$\sigma_\vartheta^2 \ll \sigma_n^2, \quad \textit{if } s = V. \tag{11b}$$

▲

In addition, and as stated in the introduction, we will adopt an additive white noise model for n. This model is widely used for the analysis and design of data converters (see, e.g., [1]–[5], [12]–[14], [16]–[18], [20]–[22]), being usually described as follows:

***Assumption** 3: The sequence of quantization noise $\{n(k)\}_{k\in\mathbb{Z}}$ is a zero-mean w.s.s. random process, uncorrelated with the input of the PRFQ, and having constant PSD*

$$S_n(\omega) = \sigma_n^2, \quad \forall \omega \in [-\pi, \pi],$$

where σ_n^2 is the variance of $\{n(k)\}_{k\in\mathbb{Z}}$. ▲

The above additive white noise model, although not exact, is, in general a good approximation when a signal with a smooth *probability density function* (PDF) is quantized with many levels and negligible overload (in the sense of Assumption 2), see, e.g., [2]. The model can be made exact, *even for few quantization levels*, by utilizing a uniform scalar quantizer with either subtractive or non-subtractive dither[6], provided quantizer overload does not occur, see [10]. As discussed before, one way to achieve this is to use a quantizer with a sufficiently large number of quantization levels, so as to satisfy (10). In this case, if the quantization interval is Δ and the dither sequence ν whitens n, makes n uncorrelated to x when $\mathcal{Q}$ is not overloaded and is bounded as $|\nu(k)| \leq \nu_{max}$, then any number of levels greater than or equal to $(V_{min} + 2\nu_{max})/\Delta$ will make Assumption 3 hold exactly. If a smaller number of quantization levels are employed so that $V < V_{min}$, then the use of dither with the same characteristics as before, together with clipping (i.e., setting $s = V$), will also make n satisfy Assumption 3 exactly.

Assumption 3 allows one to write the variance of $\{v(k)\}_{k\in\mathbb{Z}}$ as

$$\sigma_v^2 = \|A\Omega_x\|^2 + \sigma_n^2 \|F\|^2, \tag{12}$$

see Fig. 1. This equation describes the effect of σ_n^2 on σ_v^2 through the feedback path. However, if the scalar quantizer has a finite and fixed number of quantization levels, then another link between these two variances needs to be considered. In order to model this relationship, we will use the fixed signal-to-noise ratio model employed in, e.g., [12], [14], [16], [17], [21]:

***Assumption** 4: For a fixed number of quantization levels, the variance of quantization errors is proportional to the variance of the signal being quantized, i.e., there exists $\gamma > 0$ such that*

$$\gamma \triangleq \frac{\sigma_v^2}{\sigma_n^2}. \tag{13}$$

▲

If no clipping is used (i.e., if $s = \infty$), then γ corresponds exactly to the SNR of $\mathcal{Q}$. If $s = V$, then γ is a good a approximation of the SNR of $\mathcal{Q}$ when (11b) in Assumption 2 holds.

In our model, γ is assumed fixed and given. Strictly speaking, γ depends on the PDF of $\{v(k)\}_{k\in\mathbb{Z}}$, on the number of quantization levels of $\mathcal{Q}$, and on how quantization thresholds and levels are distributed along the dynamic range of $\mathcal{Q}$. In practice, for a given number of quantization levels, γ should be chosen such that the dynamic range of $\mathcal{Q}$ is used efficiently, whilst ensuring a low probability of quantizer overload or clipping. For example, for the often cited uniform quantizer with N levels and loading factor[7] equal to 4 we obtain $\gamma = \frac{3}{16}N^2$ (assuming that $\{n(k)\}_{k\in\mathbb{Z}}$ has a uniform PDF and neglecting overload errors). We note that for large N, and provided overload errors are negligible, a quadratic relationship between N and γ holds for most types of scalar quantizers (see, e.g., [9]). This is indeed the well known rule of "6 [dB] reduction of quantization noise variance per additional bit of quantizer resolution".

In the sequel, we refer to the model of PRFQ determined by Assumptions 2, 3 and 4 as *The Linear Model*. Summarizing, the Linear Model is exact if the PRFQ uses a dithered quantizer having enough quantization levels to avoid overload. If not enough quantization levels are available and dither is used jointly with clipping, then the model is exact in predicting the effects of granular quantization errors, and is a good approximation in predicting the total FWMSE if Assumption 2 also holds. If the scalar quantizer is undithered, has a small quantization interval (relative to $\sigma_{v'}$) and enough quantization levels to avoid overload, then the Linear Model can be expected to yield a good approximation of the total FWMSE. Perhaps surprisingly, the Linear Model turns out to predict with remarkable accuracy the FWMSE of an optimal PRFQ when few quantization levels and clipping are used with a loading factor big enough to satisfy Assumption 2, even without dither, and even for a 1-bit quantizer. This can be observed from the simulation results presented in Section VIII.

C. Optimization Constraints

The filters $A(z)$, $B(z)$ and $F(z)$ in Fig. 1 are design choices. We shall restrict the search for the optimal filters to those satisfying the following constraint:

***Constraint** 1:*

1) *$A(z)$ and $B(z)$ satisfy* (1).

[6]Here and in the sequel we assume the dither is such that n is white and uncorrelated with x when $\mathcal{Q}$ is not overloaded.

[7]The loading factor corresponds to the ratio between half the dynamic range of $\mathcal{Q}$ and the standard deviation of its input.

2) $A(z)$ *and* $B(z)$ *are stable.*
3) $F(z)$ *is stable and strictly causal (i.e.,* $\lim_{z\to\infty} F(z) = 0$*).* ▲

As foreshadowed in Section I, the first constraint enforces perfect reconstruction. As discussed in Section II-A.3, the stability constraints on $A(z)$, $B(z)$ and $F(z)$ are a necessary condition for the converter to be BIBO stable. The additional requirement on $F(z)$, namely strict causality, is needed for the feedback loop in Fig. 1 to be well defined (see, e.g., [2, Chap. 4]). Notice that we will not a priori require $1 - F(z)$ to have zeros only inside the open unit disk. Instead, we will show that the latter property arises naturally from the solution of the design optimization problem.

An additional constraint on $F(z)$ arises from the value of γ, as explained next. The ratio between the variances of v and n *imposed by the feedback* can be obtained by dividing (12) by σ_n^2, yielding

$$\frac{\sigma_v^2}{\sigma_n^2} = \frac{\|A\Omega_x\|^2}{\sigma_n^2} + \|F\|^2. \tag{14}$$

One can see from the above that if $\|F\|^2 > \gamma$, then *any* pre-filter or scaling of the quantization intervals of $\mathcal{Q}$ will yield $\sigma_v^2 > \gamma\sigma_n^2$, thus making large overload (or clipping) inevitable. This would increase overall distortion, and if no clipping is used, may lead to large limit-cycle oscillations. We thus conclude that the use of feedback imposes the following constraint:

***Constraint** 2:*

$$\|F\|^2 < \gamma.$$

▲

If the above constraint is met, then σ_n^2 can be found by substituting (13) into (14). This gives

$$\sigma_n^2 = \frac{\|A\Omega_x\|^2}{\gamma - \|F\|^2}. \tag{15}$$

III. Optimal PRFQ Design

Given the model described in the previous section, we can now evaluate the quantity that we aim to minimize, namely, the *frequency weighted mean squared error* (FWMSE). From (8c), and Assumptions 2 and 3, it follows that the FWMSE is given by $\sigma_\epsilon^2 = \sigma_n^2\|(1-F)BP\|^2$. Thus, in view of (15), the minimization of the FWMSE in the Linear Model can be stated as follows:

***Optimization Problem** 1: For given γ, and for given Ω_x and P satisfying Assumption 1, find the frequency responses F, A and B satisfying Constraints 1 and 2 that minimize*

$$\sigma_\epsilon^2 = \frac{\|A\Omega_x\|^2\|(1-F)BP\|^2}{\gamma - \|F\|^2}. \tag{16}$$

▲

The following proposition allows us to further reduce the number of unknowns in (16) by characterizing the optimal $A(z)$ for a given choice of $F(z)$.

***Proposition** 1: For any $F(z)$ satisfying Constrains 1 and 2, the infimum of the achievable FWMSE is given by*

$$\sigma_{\epsilon\,inf|F}^2 = \frac{\langle |1-F|, |\Omega_x P| \rangle^2}{\gamma + 1 - \|1-F\|^2}. \tag{17}$$

The filters that achieve the infimum, namely $A_{inf}(z)$ and $B_{inf}(z)$, satisfy

$$|A_{inf}| \triangleq \kappa\sqrt{|P|\,|\Omega_x|^{\sim 1}\,|1-F|}, \tag{18a}$$

$$|B_{inf}| \triangleq \frac{1}{\kappa}\sqrt{|P|^{\sim 1}\,|\Omega_x|\,|1-F|^{\sim 1}}, \tag{18b}$$

where $\kappa > 0$ is an arbitrary real constant. If $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right|$ satisfies condition i) in Assumption 1, then $A_{inf}(z)$ and $B_{inf}(z)$ can be chosen stable; else, if $\left|\Omega_x(e^{j\omega})P(e^{j\omega})\right|$ satisfies condition ii) in 1, then one can achieve an FWMSE arbitrarily close to $\sigma_{\epsilon\,inf|F}^2$ with causal and stable filters $A(z)$, $B(z)$ such that

$$\left|A(e^{j\omega})\right| = A^{[\varepsilon]}(\omega) \triangleq \begin{cases} \varepsilon_B & ,\forall\omega \in \mathcal{I}_{\varepsilon_B} \\ 1/\varepsilon_A & ,\forall\omega \in \mathcal{I}_{\varepsilon_A} \\ \left|A_{inf}(e^{j\omega})\right| & ,\forall\omega \notin \mathcal{I}_{\varepsilon_A} \cup \mathcal{I}_{\varepsilon_B}, \end{cases} \tag{19a}$$

$$\left|B(e^{j\omega})\right| = B^{[\varepsilon]}(\omega) \triangleq \left(A^{[\varepsilon]}(\omega)\right)^{-1}, \tag{19b}$$

a.e. on $[-\pi,\pi]$, where

$$\mathcal{I}_{\varepsilon_B} \triangleq \{\omega \in [-\pi,\pi] : \left|B_{inf}(e^{j\omega})\right| > \tfrac{1}{\varepsilon_B}\} \cup \mathcal{N}_P,$$
$$\mathcal{I}_{\varepsilon_A} \triangleq \{\omega \in [-\pi,\pi] : \left|A_{inf}(e^{j\omega})\right| > \tfrac{1}{\varepsilon_A}\} \cup \mathcal{N}_{\Omega_x},$$

by making $\varepsilon_A, \varepsilon_B \to 0$. ▲

As a consequence of Proposition 1, the optimal PRFQ design problem reduces to that of finding the filter $F(z)$ which minimizes $\sigma_{\epsilon\,min|F}^2$ in (17).

It is convenient to rewrite equation (17) more compactly by introducing the following change of variables:

$$f(\omega) \triangleq \left|1 - F(\mathrm{e}^{j\omega})\right|, \qquad \forall\omega \in [-\pi,\pi], \tag{20a}$$

$$g(\omega) \triangleq \left|\Omega_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|, \qquad \forall\omega \in [-\pi,\pi]. \tag{20b}$$

Substituting (20) into (17) allows us to rewrite the infimal FWMSE for a given choice of f as

$$\sigma_{\epsilon\,min|f}^2 = D(f) \triangleq \frac{\langle f, g\rangle^2}{\gamma + 1 - \|f\|^2}. \tag{21}$$

We next translate the restrictions on $F(z)$, stated in Constraints 1 and 2, into equivalent constraints on f. For this purpose, we note that, by definition, f needs to satisfy $f(\omega) \geq 0$, $\forall\omega \in [-\pi,\pi]$, and that, since $\|F\|^2 = \|F-1\|^2 - 1$ (see the proof of Proposition 1 in the Appendix), Constraint 2 is satisfied iff $\|f\|^2 < \gamma + 1$. In addition, a stable and strictly causal $F(z)$ (i.e., one satisfying Constraint 1) always leads to a function f, see (20), which satisfies[8]

$$0 \leq \int_{-\pi}^{\pi} \ln f(\omega) d\omega < \infty. \tag{22}$$

This result follows directly from Jensen's formula [23] (see also the Bode Integral Theorem in, e.g., [24]).

On the other hand, as we shall see in Section IV, if Assumption 1 holds, then the optimal f within the set of

[8]Notice that (22) dictates a fundamental trade-off in the noise-shaping capabilities of feedback quantizers, namely, that one can remove noise from one frequency band only at the expense of increasing it on another. This is also known as the "water bed effect". We discuss further implications of (22) in Section VII.

functions described by (22) and the requirement $\|f\|^2 < \gamma+1$ turns out to be piece-wise differentiable on $[-\pi,\pi]$, has at most a finite number of discontinuity points, and satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\omega)d\omega = 0, \quad \text{and} \tag{23a}$$

$$0 < f_{min} \leq f(\omega) \leq f_{max} < \infty,\ \forall \omega \in [-\pi,\pi]. \tag{23b}$$

Under these conditions, it is always possible to find a stable and strictly causal filter $F(z)$ such that $\left|1 - F(\mathrm{e}^{j\omega})\right|$ approximates $f(\omega)$ arbitrarily well on $[-\pi,\pi]$, as stated in the following lemma:

***Lemma** 1: Suppose that f is piece-wise differentiable on $[-\pi,\pi]$, that it has at most a finite number of discontinuity points and that it satisfies* (23)*. Then, for every $\varepsilon > 0$, there exists a (finite order) rational, strictly proper and stable $F(z)$ such that $\|f - |1-F|\,\| \leq \varepsilon$.* ▲

Using the above results, Optimization Problem 1 can be restated as follows:

***Optimization Problem** 2: For given and known $\gamma > 0$ and for g satisfying Assumption 1, find*

$$f^\star \triangleq \arg\min_{f \in \mathcal{C}_1 \cap \mathcal{C}_2} D(f), \tag{24}$$

where $D(f)$ is as in (21) *and*

$$\mathcal{C}_1 \triangleq \left\{ f : \mathbb{R} \to \mathbb{R}_0^+ \,:\, \|f\|^2 < \gamma + 1 \right\},$$

$$\mathcal{C}_2 \triangleq \{ f : \mathbb{R} \to \mathbb{R}_0^+ \,:\, 0 \leq \int_{-\pi}^{\pi} \ln f(\omega) d\omega < \infty \}.$$

▲

The optimizer $f^\star$ characterizes the optimal feedback filter, say $F^\star(z)$, via (20) (see also Lemma 1). In the following section, we will show how to solve this optimization problem.

IV. Solution of the PRFQ Optimization Problem

It would be desirable to provide an explicit analytical solution to Optimization Problem 2. Unfortunately, and as will become apparent in the discussion below, developing a closed form solution, for arbitrary functions g, appears infeasible. Nevertheless, we can provide a one-parameter characterization of the optimal function $f^\star$ in (24) as follows:

***Theorem** 1: For any given $g = |\Omega_x P|$ satisfying Assumption 1, and for any $\gamma > 0$, the function $f^\star$ in* (24) *belongs to the one-parameter family of functions $\{f_\alpha\}_{\alpha > \alpha_c}$, where*

$$f_\alpha(\omega) \triangleq \frac{\theta(\alpha)}{\sqrt{g(\omega)^2 + \alpha} + g(\omega)}, \quad \forall \omega \in [-\pi,\pi]. \tag{25a}$$

and

$$\theta(\alpha) \triangleq \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\sqrt{g(\omega)^2 + \alpha} + g(\omega) \right) d\omega \right). \tag{25b}$$

Here, $\alpha_c \triangleq \max\{0, \alpha_K\}$, is the lower bound of feasible α's, and α_K, if it exists, is the unique scalar such that $\|f_{\alpha_K}\|^2 = \gamma + 1$. If such a scalar doesn't exist, then we choose $\alpha_K = 0$. ▲

Note that the above result provides an explicit analytic expression for $f^\star$, once the optimal α, defined as

$$\alpha_{opt} \triangleq \arg\min_{\alpha \in (\alpha_c, \infty)} D(f_\alpha), \tag{26}$$

has been found, i.e., $f^\star = f_{\alpha_{opt}}$. Expression (25a) also gives insight into the structure of $f^\star$.

Theorem 1 can be used to develop an efficient algorithm to solve Optimization Problem 2. The key point is that substitution of (25a) into (21) changes the search space from the infinite-dimensional set $\mathcal{C}_1 \cap \mathcal{C}_2$ to the real interval (α_c, ∞). More precisely, Optimization Problem 2 is turned into the simpler problem of finding the minimizer of the single variable non-convex scalar function

$$\Phi(\alpha) \triangleq D(f_\alpha) = \frac{\langle f_\alpha, g \rangle^2}{\gamma + 1 - \|f_\alpha\|^2}, \ \ \alpha > \alpha_c. \tag{27}$$

We will show next that the global minimizer of $\Phi(\alpha)$, i.e., α_{opt}, (and hence the solution of Optimization Problem 2) is unique. Furthermore, α_{opt} can be obtained by finding the root of a scalar, convex, and monotonically decreasing function.

***Theorem** 2: Let $g = |\Omega_x P|$ satisfy Assumption 1, and suppose that g is not almost constant, see* (3)*. Then, for any $\gamma > 0$, the parameter α_{opt} defined in* (26) *satisfies*

$$\gamma + 1 = e^{\frac{1}{\pi}\int_{-\pi}^{\pi} \ln\left[\sqrt{g(\omega)^2 + \alpha_{opt}} + g(\omega)\right] d\omega} \Big/ \alpha_{opt}. \tag{28}$$

On the other hand, if g is almost constant, then any $\alpha \in (\alpha_c, \infty)$ is optimal. ▲

***Theorem** 3: The right hand side of* (28) *is a convex and strictly decreasing function of α_{opt}. Furthermore, the following holds*

$$\lim_{\alpha \to 0} e^{\frac{1}{\pi}\int_{-\pi}^{\pi} \ln\left[\sqrt{g(\omega)^2 + \alpha} + g(\omega)\right] d\omega} \Big/ \alpha = \infty, \tag{29}$$

$$\lim_{\alpha \to \infty} e^{\frac{1}{\pi}\int_{-\pi}^{\pi} \ln\left[\sqrt{g(\omega)^2 + \alpha} + g(\omega)\right] d\omega} \Big/ \alpha = 1\,. \tag{30}$$

▲

Thus, γ and α_{opt} are related through a bijective function. Moreover, it follows from Theorems 2 and 3 that, for any g satisfying Assumption 1, and for any $\gamma > 0$, the global minimizer of (27) exists and is unique[9]. In addition, these results guarantee that α_{opt} can be easily found by solving (28), via, for example, the bisection algorithm [25], or any other convex optimization method [26].

We can now express $f^\star$ and the minimum achievable FWMSE, namely $D^\star$, in terms of g, γ and α_{opt}. Indeed, combining (28) and (25a) with (21) yields (after some algebraic simplification), that

$$\begin{aligned} D^\star &\triangleq \min_{f \in \mathcal{C}_2 \cap \mathcal{C}_1} D(f) \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi} \left(\sqrt{g(\omega)^2 + \alpha_{opt}} - g(\omega) \right) g(\omega) d\omega, \end{aligned} \tag{31a}$$

whilst the associated optimal feedback filter is characterized via:

$$f^\star(\omega) = \sqrt{\frac{\gamma+1}{\alpha_{opt}}} \left(\sqrt{g(\omega)^2 + \alpha_{opt}} - g(\omega) \right), \tag{31b}$$

$\forall \omega \in [-\pi,\pi]$, see (20). Note that applying (49b) (see the Appendix) to the above it follows that $f^\star(\omega) < \sqrt{\gamma+1}$, $\forall \omega \in$

[9]If g is almost constant, then α_{opt} is not unique. Nevertheless, in this case, the minimizer of $D(f)$ is unique. (It is $f(\omega) \equiv 1$, see the paragraph immediately after (78) in the proof of Theorem 1 in the Appendix.)

$[-\pi, \pi]$. Thus, as expected, Constraint 2 is satisfied. Notice also from (31b) that if $|\Omega_x P|$ satisfies Assumption 1, then $f^\star$ satisfies the conditions of Lemma 1.

It can be seen from (31a) that $D^\star$ is a monotonically increasing function of α_{opt}. In view of Theorem 3, this implies that, as expected, $D^\star$ is monotonically decreasing with increasing γ. As a consequence, the converse of Optimization Problem 1, namely, finding the optimal filters and minimum required SNR of $\mathcal{Q}$ for a given target distortion, can be solved by using (28) and (31). Moreover, since the right hand side of (31a) is a concave, monotonically increasing function of α_{opt}, this parameter can be easily found by using standard iterative algorithms, as in the original optimization problem.

It is also interesting to note that (28) and (31a), which relate γ and $D^\star$ via the parameter α_{opt}, have a structure akin to the well known *reverse water-filling* equations (see, e.g., [27, pp. 108-123], and [28]). The latter characterize the rate-distortion function for Gaussian sources.

To summarize, we have given an explicit analytic expression for the optimal $|1 - F|$ and D, once α_{opt} has been determined. Furthermore, we have shown that the parameter α_{opt} always exists, is unique, and can be easily found using simple numerical methods.

In the following sections, we will provide additional insight into the consequences of these results, as well as into some properties of optimal PRFQs,

V. Properties of Optimal PRFQ

In the sequel, we say that a PRFQ is *optimal* or *optimized* if its filters $A(z)$, $B(z)$ satisfy (19) for negligibly small values of ε_A and ε_B, and $F(z)$ is such that $\left|1 - F(\mathrm{e}^{j\omega})\right| = f^\star(\omega)$, a.e. on $[-\pi, \pi]$, with $f^\star$ as defined by (24).

A. The Effect of the SNR of $\mathcal{Q}$

It follows from Theorems 2 and 3 that, for any given $\Omega_x P$ satisfying Assumption 1, f_α in (25a) describes the family of all noise shaping characteristics that are optimal for *some* $\gamma > 0$.

As we will show, adjusting α from 0 to ∞ (equivalently, γ from ∞ to 0) allows one to undergo a smooth progression from "full" noise-shaping to no noise-shaping, in an optimal manner. An example of this progression is shown in Fig. 2. Note in this figure how $|1 - F|$ (solid lines) approaches a unit transfer function as γ (the quantizer SNR for which $\alpha = \alpha_{opt}$ in the figure), becomes smaller (and α_{opt} gets larger). It can also be observed that $|1 - F|$ approaches the inverse of $|\Omega_x P|$ as γ is increased. Such asymptotic convergence does indeed take place in general, as the following theorem shows:

***Theorem** 4: For any $g = |\Omega_x P|$ satisfying Assumption 1, the functions $f_\alpha(\omega)$ defined in* (25a) *converge uniformly to*

$$f_\infty(\omega) \triangleq 1 \tag{32}$$

as $\gamma \to 0$. Similarly, for any function g satisfying condition i) in Assumption 1, the functions $f_\alpha(\omega)$ defined in (25a) *converge uniformly to*

$$f_0(\omega) \triangleq \left[e^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln g(\omega) d\omega}\right] [g(\omega)]^{-1} \tag{33}$$

as $\gamma \to \infty$. ▲

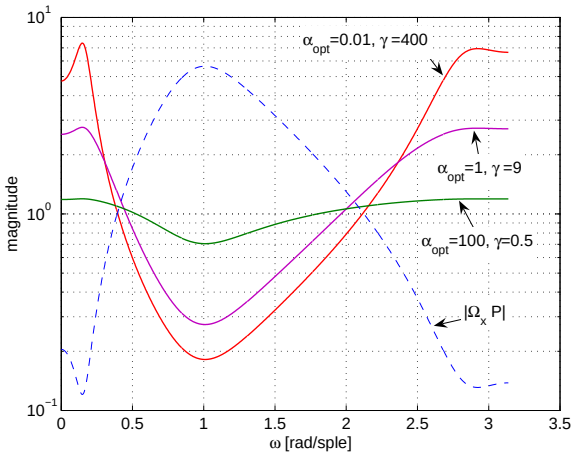


Fig. 2: Progression of $f_\alpha(\omega)$ (solid lines) for $\alpha \in \{0.01, 1, 100\}$. In this example, $\Omega_x(z)P(z) = \frac{z^4 - 0.3549z^3 - 1.313z^2 + 0.1723z + 0.5776}{z^4 - 1.223z^3 + 0.8192z^2 - 0.196z}$

Note that f_∞ in (32) corresponds to the choice of *no feedback* ($F(z) = 0$), which reduces the PRFQ to a PCM converter. In view of (30), this no-noise shaping scenario is asymptotically optimal as $\gamma \to 0$. In turn, f_0 defined in (33) corresponds to the full whitening feedback filters proposed in [1], [14], [15]. From (29) and (33), f_0 is optimal iff $\gamma \to \infty$. See also the discussion in Section VII.

B. Signal Spectra

1) The Output of the Quantizer: By looking at Fig. 1 and using Assumption 3, we find that the PSD of $\{w(k)\}_{k\in\mathbb{Z}}$ in an optimized PRFQ is given by $S_w(\omega) = |\Omega_x(\omega)A(\omega)|^2 + \sigma_n^2 f^\star(\omega)^2$, $\forall \omega \in [-\pi, \pi]$. Applying (18) to the latter result yields

$$S_w(\omega) = f^\star(\omega)\left[\kappa^2 g(\omega) + \sigma_n^2 f^\star(\omega)\right], \quad \forall \omega \in [-\pi, \pi]. \tag{34}$$

Comparing (15) and (16), it is easy to see that $\sigma_n^2 = D(f)/\|fBP\|^2$. If $B(z)$ satisfies (18), then we have $\sigma_n^2 = \kappa^2 D(f)/\langle f, g\rangle$. With the choice $f = f^\star$, and using (31) and (28), we conclude that the variance of the quantization noise in an optimized PRFQ is given by[10]

$$\sigma_n^2 = \frac{\kappa^2}{2}\sqrt{\frac{\alpha_{opt}}{\gamma + 1}} = \frac{\kappa^2}{2}\mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left[\sqrt{g(\omega)^2 + \alpha_{opt}} - g(\omega)\right] d\omega}. \tag{35}$$

Substitution of this expression into (34) yields $S_w(\omega) = \frac{\kappa^2 f^\star(\omega)}{2}\left[\sqrt{g(\omega)^2 + \alpha_{opt}} + g(\omega)\right]$, $\forall \omega \in [-\pi, \pi]$, where (25a) has been used. Substitution of (49a) and (28) into this expression leads to

$$S_w(\omega) = \frac{\kappa^2 f^\star(\omega)\alpha_{opt}}{2[\sqrt{g(\omega)^2 + \alpha_{opt}} - g(\omega)]} = \frac{\kappa^2\sqrt{(\gamma + 1)\alpha_{opt}}}{2}, \tag{36}$$

[10]Note from (35) that if σ_n^2 is fixed, then the value of κ is no longer arbitrary. This ensures that (13) is satisfied.

which is independent of ω. Therefore, under Assumptions 3 and 4, the output of the quantizer in an optimized PRFQ *is white*. This suggests that near optimal coding of the quantizer output can be achieved with a memory-less entropy coder.

2) The Frequency Weighted Reconstruction Error: The PSD of the frequency weighted reconstruction error is given by $S_\epsilon(w) = \sigma_n^2 \left|f(\omega)B(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|^2$, $\forall \omega \in [-\pi, \pi]$. Substitution of (18) into the above yields $S_\epsilon(\omega) = \frac{\sigma_n^2}{\kappa^2} g(\omega) f(\omega)$. Applying (35) to the latter we obtain

$$S_\epsilon(\omega) = \tfrac{1}{2}\left(\sqrt{g(\omega)^2 + \alpha_{opt}} - g(\omega)\right) g(\omega), \quad \forall \omega \in [-\pi, \pi]. \tag{37}$$

Thus, we conclude that the frequency weighted quantization error in an optimized PRFQ is *not white*. This fact stands in stark contrast to the conclusions reached when the FQ filters are optimized *without the perfect reconstruction constraint* (1), see, e.g., [22]. It also differs from the result obtained when the feedback filter is optimized *ignoring fed back quantization error*, as in [14] and [15]. Note that, as γ is made larger, ϵ not only becomes smaller, but its PSD asymptotically approaches[11] a constant function over the frequencies $\{\omega : \left|\Omega_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right| > 0\}$.

VI. Oversampled Feedback Quantization

It is well known that oversampling (i.e., sampling a band-limited continuous-time signal at a frequency above its Nyquist rate) allows one to achieve a smaller MSE error for a given, fixed number of quantization levels. For instance, the MSE of simple scalar quantization (without feedback) is known to decrease as λ^{-1}, see [29], where λ is the *oversampling ratio*, given by

$$\lambda \triangleq \frac{\text{Sampling Frequency}}{\text{Nyquist Frequency}}.$$

In turn, it has been shown in [4] that feedback quantizers can attain an MSE that is $\mathcal{O}(\lambda^{-2(m+1)})$ as $\lambda \to \infty$, where m is the order of the feedback filter (see also recent work in [20]). From a rate-distortion viewpoint, the inversely polynomial error decay of this error estimate is "too slow" to compensate for the increase in the overall bit-rate due to oversampling (which is proportional to λ). To be more precise, let us consider a scalar quantizer with $N = 2^b$ quantization levels, where b denotes the quantization resolution in bits per sample. If the additional bit-rate caused by oversampling was utilized instead to increase N, then the MSE would decay as $\mathcal{O}(2^{-2b\lambda})$, i.e., exponentially[12].

A faster decay of the MSE of oversampled FQ with λ can be achieved by selecting a different feedback filter (with possibly different order) for each oversampling ratio. An example of such a family (of 1-bit $\Sigma\Delta$ converters) was given in [31]. Here, the continuous-time reconstruction error can be uniformly bounded by $\lambda^{-\rho \log \lambda}$, where $\rho > 0$ is independent of λ. This bound guarantees an MSE that decays with λ as $\mathcal{O}(\lambda^{-2\rho \log \lambda})$, which is faster than any inverse polynomial, but still far from exponential. Based on this result, the family of 1-bit $\Sigma\Delta$ converters reported in [32] achieve an MSE that is $\mathcal{O}(2^{-0.14\lambda})$, i.e., exponentially decaying with increasing λ. Notably, the results in [31] and [32] were obtained using an exact, deterministic model of quantization.

We will next show that, within the Linear Model, if the optimal infinite order filters characterized in Section IV are used for each value of λ, then one can achieve an exponential decay of $D^\star$ with the oversampling ratio, provided γ is kept constant.

If the input sequence $\{x(k)\}_{k\in\mathbb{Z}}$ is obtained from sampling a band-limited analog signal, oversampling would cause g (defined in (20)) to vary with λ. To capture this effect, we replace g by the family of functions g_λ, defined as

$$g_\lambda(\omega) \triangleq \begin{cases} \sqrt{\lambda}\, g_1(\lambda\omega) & , \text{ if } \quad |\omega| < \omega_c, \\ 0 & , \text{ if } \omega_c \le |\omega| \le \pi. \end{cases} \tag{38}$$

In (38), g_1 denotes the square root of the PSD of the frequency weighted input without oversampling, and $\omega_c \triangleq \frac{\pi}{\lambda}$. Notice that $\|g_\lambda\|^2$, that is, the total power of g_λ (in units of variance per sample), remains constant for all $\lambda \ge 1$. This ensures a uniform comparison basis for the distortion figures.

We can now make explicit the dependence of $D^\star$ on γ and λ by writing

$$D^\star(K, \lambda) \triangleq \min_{\substack{f\in\mathcal{C}_2\cap\mathcal{C}_1 \\ g=g_\lambda}} D(f) = \min_{f\in\mathcal{C}_2\cap\mathcal{C}_1} \frac{\langle f, g_\lambda\rangle^2}{K - \|f\|^2}, \tag{39}$$

see (21), where

$$K \triangleq \gamma + 1 = \frac{\sigma_w^2}{\sigma_n^2} \tag{40}$$

corresponds to the *output-SNR of* $\mathcal{Q}$. Interestingly, it is possible to establish a precise "exchange" formula for K and λ. Indeed, in terms of minimal achievable distortion, the effect of increasing oversampling is equivalent to an exponential increase in the output-SNR of $\mathcal{Q}$. This is shown in the next theorem:

***Theorem** 5: Under the Linear Model described in Section II-B, for any function $g_1(\omega)$, and for any $K > 1$, $\lambda \ge 1$, the minimum achievable FWMSE satisfies:*

$$D^\star(K, \lambda) = D^\star(K^\lambda, 1). \tag{41}$$

▲

If we assume that γ depends exponentially on the number of bits per sample, then Theorem 5 suggests an FWMSE that decays exponentially with λ, provided the Linear Model holds and that optimal filters $A(z)$, $B(z)$ and $F(z)$ (characterized by (18), (25a) and (28)) are employed for each λ. The following simple example illustrates this idea:

***Example** (Flat Weighted Input Spectrum) Consider an input signal $\{x(k)\}_{k\in\mathbb{Z}}$ and a weighting filter $P(z)$ such that $|\Omega_x P|$ is constant $\forall \omega \in [-\pi, \pi]$, without oversampling. For this setup, the optimal $F(z)$ for our model of PRFQ is $F(z) \equiv 0$*

[11] Substitution of (49a) into (37) yields $S_\epsilon(\omega) = \frac{\alpha_{opt}}{4} \frac{2g(\omega)}{\sqrt{g(\omega)^2+\alpha_{opt}}+g(\omega)}$. Thus, $S_\epsilon(\omega) < \alpha_{opt}/4$ for all $\omega \in [-\pi, \pi]$, and $S_\epsilon(\omega) \to \alpha_{opt}/4$ as $\alpha_{opt} \to 0^+$, $\forall \omega$ such that $g(\omega) > 0$.

[12] Strictly speaking, this only holds for signals whose PDFs have finite support. Indeed, it has been shown that for several infinite support PDFs, the MSE of uniform quantization decreases asymptotically with b not faster than $(\ln 2)^{2/a}\, b^{\frac{2}{a}} 2^{-2b}$, where $a > 0$ is a constant independent of b, see [30].

($f(\omega) \equiv 1$), i.e., a PCM converter. From (21)*, the minimum FWMSE without oversampling (i.e., with $\lambda = 1$) becomes*

$$D^\star(K,1) = \frac{\sigma_{xP}^2}{\gamma} = \frac{\sigma_{xP}^2}{K-1},$$

where $\sigma_{xp}^2 \triangleq \|\Omega_x P\|^2$. To analyze oversampling behaviour of $D^\star$ in this case, we apply Theorem 5 to the above expression. This gives that $D^\star(K,\lambda) = \frac{\sigma_{xP}^2}{K^\lambda - 1}$, and, thus,

$$\sigma_{xP}^2 K^{-\lambda} \leq D^\star(K,\lambda) \leq \left(\frac{\sigma_{xP}^2}{1-K^{-1}}\right) K^{-\lambda} \qquad (42)$$

for all $\lambda \geq 1$. Note that, to achieve (42)*, $F(z)$ needs to be synthesized according to* (31b) *and* (20)*. Therefore, for this example, the MSE of an optimized PRFQ with fixed γ exhibits an exponential decay with the oversampling ratio (since, by definition, $K > 1$).*

If we further assume K to depend on the number of bits per sample b as $K = \frac{3}{16}2^{2b} + 1$ (which would correspond to $\mathcal{Q}$ being a uniform quantizer with many levels and operating with a loading factor of 4*), then* (42) *becomes*

$$\begin{aligned} &\sigma_{xP}^2\, 2^{-\left[\log_2(\frac{3}{16}+2^{-2b})+2b\right]\lambda} \leq \\ &\qquad D^\star(K,\lambda) < \left(\tfrac{\sigma_{xP}^2}{1-K^{-1}}\right) 2^{-\left[\log_2(\frac{3}{16}+2^{-2b})+2b\right]\lambda}. \end{aligned} \qquad (43)$$

The term $\log_2(\frac{3}{16}+2^{-2b})$ in (43) *is negative for all $b \geq 1$. This implies that the decrease of $D^\star$ with λ, although exponential, is slower than $2^{-2b\lambda}$. Thus, the use of oversampling in this case is rate-distortion inefficient. In particular, taking $b = 1$, and supposing that Assumptions 3 and 4 hold, we obtain from* (43) *that $D^\star(K,\lambda)$ is lower and upper bounded by terms proportional to $2^{-0.807\lambda}$. For loading factor values of* 6*,* 10 *and* 20*, the exponent in the latter expression changes to -0.41λ, -0.1635λ and -0.0426λ, respectively.* ▲

The next theorem shows that the exponential decay of the FWMSE obtained in the example above can be extended to arbitrary (band-limited) input signals and frequency weighting criteria.

***Theorem** 6: For any $K > 1$ and function $g_1(\omega)$ satisfying Assumption 1, the following holds:*

$$D^\star(K,\lambda) \leq \frac{K^2 \alpha_{opt}(K,1)}{4(K-1)} K^{-\lambda}, \quad \forall K > 1, \forall \lambda \geq 1, \quad (44)$$

where $\alpha_{opt}(K,1)$ denotes the optimal α for $\lambda = 1$. ▲

Thus, under the Linear Model, we have that the FWMSE of an optimized PRFQ decays exponentially with λ.

***Remark** 1:* We recall that Theorem 6 is exact within the Linear Model described in Section II-B. Here it is convenient to present some further observations regarding the validity of that model when the oversampling ratio tends to infinity, for different implementations of a PRFQ.

1) As already mentioned in Section II-B, if x is bounded and a sufficiently large number of quantization levels to avoid overload is used together with dither, then the Linear Model is exact. Nevertheless, there is no guarantee that the number of necessary quantization levels to avoid overload remains constant as λ increases. If such number increases with λ, then γ can only be kept constant by increasing the number of quantization levels in the quantizer.
2) If the number of quantization levels is insufficient to avoid clipping/overload errors, and if dither and clipping are used with a fixed loading factor, then there exists a certain finite value of λ beyond which Assumption 2 is violated. This arises from the fact that, for any fixed loading factor, the effect of clipping errors in the output does not decay with λ, thus becoming the dominant component in the FWMSE for sufficiently high oversampling ratios. Further reduction of the FWMSE would then require one to balance clipping and granular quantization errors by increasing the loading factor. If the number of quantization levels is fixed, this would necessarily reduce the value of γ, clearly increasing the component of the FWMSE due to granular quantization errors[13]. Nevertheless, if clipping and dither are used (with $s = V$), then the Linear Model and Theorem 6 is exact in describing the FWMSE due to *granular* quantization errors. ▲

VII. The Importance of Taking Account of Fed Back Quantization Noise

If one tried to optimize the filters of a PRFQ neglecting fed back quantization noise, i.e., by trying to minimize $\frac{\|A\Omega_x\|^2\|(1-F)BP\|^2}{\gamma}$ (compare to (16)), then one would obtain a (sub optimal) feedback filter, namely $F_0(z)$, which satisfies

$$|1 - F_0| = \eta_{xP} \left|\Omega_x P\right|^{-1}, \qquad (45a)$$

where

$$\eta_{xP} \triangleq \mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|\Omega_x(\mathrm{e}^{j\omega})\right| P(\mathrm{e}^{j\omega}) d\omega}, \qquad (45b)$$

provided $|\Omega_x P| > 0$, $\forall \omega \in [-\pi, \pi]$ (see (87) in the proof of Theorem 1). This corresponds to the result obtained in [14], which was restricted to the cases where $\gamma \gg \|F_0\|^2$. For the case $\Omega_x(\mathrm{e}^{j\omega}) \equiv 1$, the noise transfer function magnitude $|1 - F_0(z)|$ is also equivalent to that derived in [15]. The latter is optimal in the sense of minimizing the ratio $\sigma_\epsilon^2/\sigma_n^2$, but not in the sense of minimizing σ_ϵ^2 for a fixed quantizer SNR γ.

As shown in Theorem 4, $f^\star$, in general, does approach $f_0 = |1 - F_0|$ as $\gamma \to \infty$. One can then expect F_0 to be *near* optimal in situations where $\gamma \gg \|F_0\|^2$, see (16). The latter is often satisfied at high bit-rates (i.e., when many quantization levels are available). However, for any given number of quantization levels, it is easy to find practical situations where $\Omega_x P$ is such that $\|F_0\|^2$ is comparable to (or greater than) γ. More precisely, from (22), and recalling that $\|F - 1\|^2 = \|F\|^2 + 1$ (see Appendix B), one can show that, if $|\Omega_x P| < m$ over a set of frequencies in $[-\pi, \pi]$ with measure Γ, where m is some positive scalar, then

$$\|F_0\|^2 \geq \left(\frac{\eta_{xP}}{m}\right)^{\Gamma/\pi} - 1. \qquad (46)$$

[13] As an extension of the results presented in this section, the authors have recently derived an asymptotic decay rate of the FWMSE with λ that includes the effect of clipping errors. For Gaussian inputs, this asymptotic decay rate is faster than any inverse polynomial. These results are beyond the scope of the current paper.

This means that *a large* $\|F_0\|^2$ *is obtained for any product* $\Omega_x P$ *whose magnitude becomes significantly small (in relative terms) over certain frequency bands*. (An example is included in Section VIII below.) A direct consequence is that, for these cases, and in view of (16), trying to match $|1-F|$ to $\eta_{xP}\,|\Omega_x P|^{-1}$ will yield a performance far from optimal, also increasing the risk of incurring large limit-cycle oscillations if no clipping is employed (see, e.g., [2], [13]).

The (possibly unbounded) increase of $\|F\|^2$ as $|1-F|$ approaches $\eta_{xP}\,|\Omega_x P|^{-1}$ was already observed in [12]. Several heuristic solutions have been proposed since then (see, e.g., [2], [3], [13], [15], [17], [18]). In contrast to these approaches, the method derived in the present work allows one to characterize the true optimal filters, by explicitly taking into account $\|F\|^2$ in the cost functional to be minimized (see (16)). Our method not only guarantees that $\|F\|^2 < \gamma$, but also yields the actual optimal filters. Our proposal also has the advantage of being applicable to arbitrary input spectra and frequency weighting functions, regardless of how small the quantizer SNR γ may be, within the scope of validity of the Linear Model.

VIII. Simulation Study

To illustrate our results, we have designed the filters of a PRFQ aimed at digitally encoding audio signals in a psycho-acoustically optimal manner. The details of the simulation model, as well as the results of both the simulations and the numerical optimizations are given below.

A. Simulation Setup

The PSD of audio signals was modeled as unit-variance zero mean white Gaussian noise filtered through $\Omega_x(z) = 0.09315\left(\frac{z+0.6773}{z-0.8588}\right)$. The magnitude of the frequency response of $\Omega_x(z)$ is depicted in Fig. 3 (solid line). The frequency weighting filter $P(z)$ considered had a frequency response magnitude which approximated the psycho-acoustic curve derived in [3, Table 1], thus modeling the sensitivity of human hearing to noise[14]. The corresponding frequency response is plotted with dotted line in Fig. 3 (the sampling frequency is 44.1 [kHz]). The resulting $g = |\Omega_x P|$ for these Ω_x and $P(z)$ is also shown in the same figure (dashed line). For this choice of g, and in view of (46), one could expect the norm of a full whitening feedback filter to be very large. This is indeed the case: $\|F_0\|^2 = 2.2\times10^{11}$. Thus, the sub-optimal feedback filter characterized by (45) requires the use of a scalar quantizer with at least 18 bits in order to become feasible (see Constraint 2).

In the simulations, $\mathcal{Q}$ was chosen to be a uniform mid-rise quantizer with quantization interval $\Delta = 1$. Several values of γ were considered for the simulations, calculated as $\gamma = \frac{3}{(O.F.)^2}2^{2b}$, where $b \in \{1, 2, \ldots, 16\}$ and where $O.F. \triangleq \frac{N\Delta}{2\sigma_v}$ denotes the *loading factor*. Two different loading factors were considered: 4 and 6. The latter choice yields a slightly lower γ than the usual loading factor of 4. However, this regime has the benefit of making overload errors smaller and more infrequent.

[14]The coefficients of $P(z)$ can be found at http://msderpich.no-ip.org/research

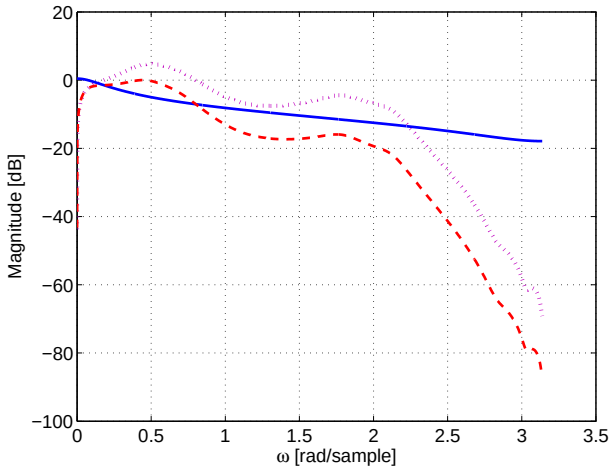


Fig. 3: Frequency response magnitudes for $\Omega_x(z)$ (solid line), $P(z)$ (dotted line) and $g(\omega) = \left|\Omega_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})\right|$ (dashed line).

As the simulation results will show, for our choices of Ω_x and P, this more conservative loading factor yields lower overall distortion when b takes values above 6 bits per sample.

For each b (and corresponding two values for γ, one for each loading factor), the filters of the converter were designed according to the following:

1) The parameter α_{opt} was calculated by numerically solving (28).
2) The optimal $|1-F|$, $|A|$ and $|B|$ were obtained via (31b) and (18).
3) These functions were then approximated[15] with rational IIR transfer functions $A(z)$, $B(z)$ (of order 7) and $F(z)$ (of order 15).
4) An appropriate value for the parameter κ in (18) was chosen via $\kappa^2 = 2\sigma_n^2\sqrt{\frac{K}{\alpha_{opt}}}$, see (35), assuming $\sigma_n^2 = 1/12$ (recall that $\Delta = 1$ for all the simulations). This ensures that $\sigma_v^2 = \gamma\sigma_n^2$.

For each combination of b and $O.F.$, the resulting PRFQ converter was simulated utilizing two different architectures.

1) *Non Overloading* $\mathcal{Q}$*:* This scheme is as depicted in Fig. 1, with $\mathcal{Q}$ having (virtually) infinitely many levels. Thus $|n(k)| \leq \frac{\Delta}{2}$ for all k (neither clipping nor overload errors occur).
2) *Overloading* $\mathcal{Q}$ *and Clipped* n*:* Here, $\mathcal{Q}$ has $N = 2^b$ levels, which yields a scalar quantizer with a finite input dynamic range $[-N\frac{\Delta}{2}, N\frac{\Delta}{2}]$. As a consequence, any value $|v(k)| > N\frac{\Delta}{2}$ would overload $\mathcal{Q}$ (if $s = \infty$) or produce clipping error (if $s = V$). To avoid large limit-cycle oscillations, this variant was simulated using clipping (i.e., $s = V$).

Each simulation with the non-overloading PRFQ comprised $100,000$ samples. For the overloading converter, five $100,000$ samples simulations were performed for each combination of $O.F.$ and b.

[15]The optimization routines utilized are based upon the Matlab optimization toolbox and can be found at http://msderpich.no-ip.org/research.

B. Results

The results of the numerical optimizations and the simulations are discussed next.

1) Comparison between $D^\star$ and the Rate-Distortion Function: The information theoretic lower bound (see [28]) for the FWMSE associated with the given source $\{x(k)\}_{k\in\mathbb{Z}}$ and filter $P(z)$ is plotted in Fig. 4 (solid line). This corresponds to Shannon's quadratic Distortion-Rate function $D(R)$ when $R = b$. As the bit-rate is increased, the gap between $D^\star$ and this absolute lower bound decreases to approx 7.5 [dB] for $O.F. = 4$ and 11 [dB] for $O.F. = 6$, at $b = 16$. This difference can be attributed to the rate-distortion inefficiency of the uniform scalar quantizer[16]. On the other hand, the larger performance gap observed at lower bit-rates can be attributed to the perfect reconstruction constraint.[17] Recall that, at low bit rates, the achievement of Shannon's rate-distortion function demands the suppression of relatively less significant bands of the PSD of the input signal (see, e.g., [27], and [28]). This linear distortion, which a PRFQ cannot achieve, is more severe at lower bit-rates. Thus, the performance gap increases as b is reduced.

2) Non Overloading $\mathcal{Q}$: The FWMSE of this converter variant is presented in four of the plots in Fig. 4, with labels beginning with "σ_ϵ^2 opt. PRFQ, Non Overloading". These differ in the loading factor, and in the meaning of b in each case. For the plots whose labels do not have the ending "E.C." (entropy coding), b is simply the number utilized to generate the value $\gamma = \frac{3}{(O.F.)^2}2^{2b}$ for which the filters were optimized. The plots whose labels end in "E.C." correspond to the same simulations, but for each point the value of b is the *scalar* entropy of the quantized output of the converter. It can be seen in Fig. 4 that the FWMSE obtained for the non overloading $\mathcal{Q}$ without entropy coding is remarkably close to the theoretical value $D^\star$ predicted by (31a). More importantly, even for bit-rates as small as $b = 2$, each observed ratio σ_v^2/σ_n^2 deviates from its nominal value of γ by less than 2%. (For the extreme situation $b = 1$, the observed σ_v^2 was slightly lower than predicted, while σ_n^2 was 55% higher than $1/12$ due to the highly non-uniform PDF of the resulting sequence $\{n(k)\}_{k\in\mathbb{Z}}$.) It can also be seen that the scalar entropy of the quantized output of the PRFQ in these cases is very close to Shannon's $R(D)$ function for a given distortion. This agrees with the observation that the output of $\mathcal{Q}$ in an optimized PRFQ is white, see the comment at the end of Section V-B.1. The difference between these quantities is bigger for lower values of b, for the same reason discussed in Section VIII-B.1 above.

3) Overloading $\mathcal{Q}$: For the overloading PRFQ using an $O.F.$ of 4, the FWMSE diminished along with the corresponding $D^\star$ for $b \in \{1,\ldots,6\}$. However, the measured FWMSE varied very little for $b \geq 7$, staying several dB higher than $D^\star$ over that range of bit-rates. This performance degradation can be attributed to clipping errors. The fact that overload errors become noticeable only for high bit rates (many quantization levels) might seem, at first, surprising. However, this phenomenon can be easily explained by noting that the size of the tails of the PDF of $\{v(k)\}_{k\in\mathbb{Z}}$ that fall outside the dynamic range of $\mathcal{Q}$ remains approximately constant in relation to $N\Delta = 2^b\Delta$ for all b. (This is a direct consequence of the loading factor rule.) In contrast, granular (non-overloading) quantization error is proportional to Δ^2 (which is held constant in the simulations). Therefore, the ratio between clipping and granular quantization errors grows approximately as 2^b and clipping errors become dominant for sufficiently high bit-rates.

Because of the reduced occurrence (and magnitude) of clipping errors, the optimized PRFQ with overloading $\mathcal{Q}$ and $O.F. = 6$ exhibits an FWMSE smaller than that of its counterpart with $O.F. = 4$ for $b \geq 7$. Furthermore, this more conservative loading factor allows the converter to perform almost exactly as predicted by our analytical expression for $D^\star$.[18]

4) Comparison with PCM: The theoretical FWMSE of a PCM A/D converter, denoted by D_{PCM}, can be found from (16) by making $A(z) \equiv B(z) \equiv 1$ and $F \equiv 0$, which gives $D_{PCM} = \|\Omega_x\|^2\|P\|^2/\gamma$. For the chosen input PSD and frequency weighting filter, and calculating γ as $\frac{3}{16}2^{2b}$, the value of D_{PCM} varies with b as shown in Fig. 4 (dotted line). As seen in this figure, the gap between $D^\star$ and D_{PCM}, for each value of $O.F.$, gets smaller as the bit-rate decreases. This agrees with the fact that the optimal PRFQ approaches a PCM converter as $\gamma \to 0$, see Section V-A. It can also be seen in Fig 4 that the optimized PRFQ with overloading and $O.F. = 6$ exhibits an improvement of 32 [dB] over PCM at $b = 16$. Equivalently, in order to obtain the same FWMSE as that of PCM at 16 bits, the PRFQ converter with $O.F. = 6$ requires less than 12 bits. At lower bit-rates, the improvement of the optimal PRFQ over PCM is also significant. For example, the overloading PRFQ with $O.F. = 4$ and $b = 2$ has a lower FWMSE than the PCM converter with $b = 4$, thus achieving a data rate compression of 50% (see Fig. 4).

IX. Conclusions

This paper has studied perfect reconstruction feedback quantizers based on an additive white noise model for quantization errors. We have derived results that relate the minimum frequency weighted MSE and the signal-to-noise ratio of the scalar quantizer embedded in the converter. We have also provided closed form expressions for the optimal frequency responses of the filters in the converter and have derived several properties of optimal PRFQs. In particular, we have shown that the optimal frequency response magnitudes of the filters are unique, that the frequency weighted errors of an optimal PRFQ are non white, and that consecutive samples of

[16]From Shannon's Rate-Distortion function for memoryless Gaussian sources, the maximum SNR for a bit-rate b is 2^{2b}. The SNR (neglecting overload errors) for a uniform scalar quantizer with loading factor $O.F.$ is given by $\frac{3}{(O.F.)^2}2^{2b}$. Thus, the theoretical performance gaps for $O.F. = 4$ and 6 are $10\log_{10}(3/16) = 7.3$ [dB] and $10\log_{10}(3/36) = 10.8$ [dB], respectively.

[17]The quadratic Gaussian rate-distortion function with the constraint that the end-to-end distortion is uncorrelated to the source has recently been characterized in [33].

[18]There exist several results on the optimal balance between overload and granular error variances for stand-alone scalar quantizers (see, e.g., [30] and the references therein). However, for feedback quantizers the question seems to be open.

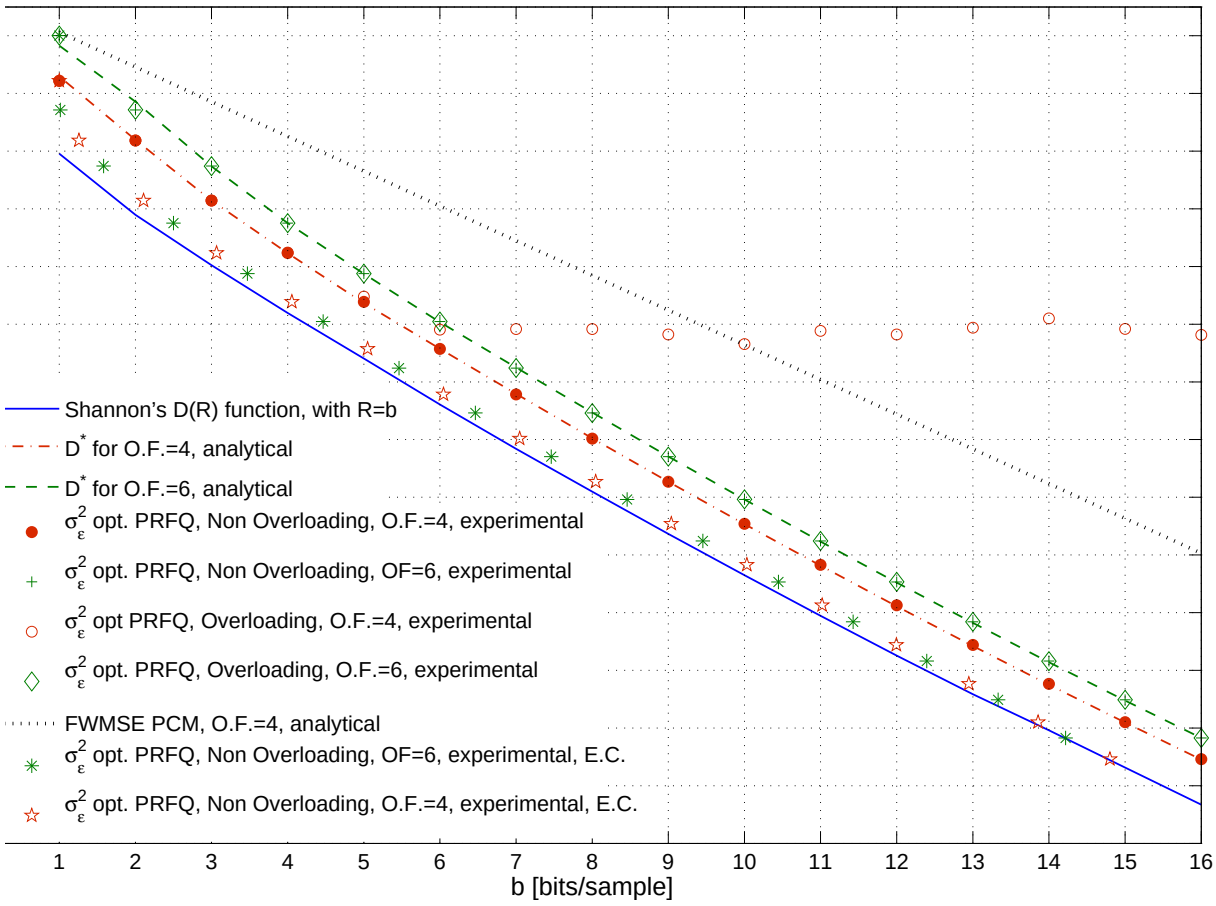


Fig. 4: Frequency weighted MSE for $b \in \{1, \ldots, 16\}$.

the output sequence of the scalar quantizer are uncorrelated. We have also shown that, within our model, the frequency weighted MSE of an optimal, oversampled PRFQ, decreases exponentially with oversampling ratio.

APPENDIX

A. Preliminary Results

The following preliminary results are necessary to prove the theorems stated in the previous sections. We begin by introducing the following definition:

***Definition** 2 (Similarly/Oppositely Functionally Related): We say that two functions $\phi, \psi : [a,b] \to \mathbb{R}$ are* similarly functionally related *iff there exists a monotonically increasing function $G(\cdot)$ such that $\phi(x) = G(\psi(x))$, for all $x \in [a,b]$, and write $\phi \uparrow\uparrow \psi$. Similarly, if there exists a monotonically decreasing function $G(\cdot)$ such that $\phi(x) = G(\psi(x))$, for all $x \in [a,b]$, we say that ϕ and ψ are* oppositely functionally related*, and write $\phi \uparrow\downarrow \psi$.* ▲

***Theorem** 7:* [19] *If $\phi, \psi : [a,b] \to \mathbb{R}$ are similarly functionally related, then*

$$[b-a]\int_a^b \phi(x)\psi(x)dx \geq \int_a^b \phi(x)dx \int_a^b \psi(x)dx. \qquad (47)$$

[19]This theorem is related to the variant of Tchebyshev's Integral Inequality given in [34, Theorem 236]. It departs from the latter in that the integrands must be functionally dependent, which allows us to state necessary and sufficient conditions for equality.

If ϕ and ψ are oppositely functionally related, then the inequality in (47) *is reversed. In either case, equality is achieved iff ψ (and therefore ϕ) is almost constant.* ▲

Proof: We will examine the difference between the right and left hand side in (47). We obtain

$$\int_a^b \phi(x)\psi(x)dx - \overline{\psi}\int_a^b \phi(x)dx = \int_a^b \phi(x)\left[\psi(x) - \overline{\psi}\right]dx,$$

where $\overline{\psi} \triangleq \frac{1}{b-a}\int_a^b \psi(x)dx$. Note that we have divided both sides by $b-a$. Suppose $\phi \uparrow\uparrow \psi$. (The proof for $\phi \uparrow\downarrow \psi$ proceeds in a similar way.) Then there exists a monotonically increasing function $G(\cdot)$ such that $\phi = G(\psi)$, and a value ϕ_0 such that $\phi(x) > \phi_0 \iff \psi(x) > \overline{\psi}$ and $\phi(x) < \phi_0 \iff \psi(x) < \overline{\psi}$. It then follows that

$$\int_a^b \phi(x)\left[\psi(x) - \overline{\psi}\right]dx \geq \int_a^b \phi_0\left[\psi(x) - \overline{\psi}\right]dx = 0,$$

with equality iff

$$\int_{\psi > \overline{\psi}} [\psi(x) - \overline{\psi}]dx = 0 = \int_{\psi < \overline{\psi}} [\psi(x) - \overline{\psi}]dx,$$

i.e., iff ψ (and therefore ϕ as well) is almost constant. ■

***Proposition** 2: Define*

$$p(\omega) \triangleq r(\omega) - g(\omega) \qquad (48a)$$

$$q(\omega) \triangleq r(\omega) + g(\omega) \qquad (48b)$$

$$r(\omega) \triangleq \sqrt{g(\omega)^2 + \alpha}, \qquad (48c)$$

with $\alpha > 0$ and $g \in L^1$, $g : [-\pi, \pi] \to \mathbb{R}_0^+$. Then, the following results hold:

$$p(\omega) = \frac{\alpha}{q(\omega)}, \tag{49a}$$

$$p(\omega)^2 \leq \alpha, \ \forall \omega \in [-\pi, \pi], \tag{49b}$$

$$p(\omega) \leq \frac{\alpha}{2g(\omega)}, \ \forall \omega \in [-\pi, \pi]. \tag{49c}$$

$$\tfrac{1}{2\pi} \int_{-\pi}^{\pi} \ln(q(\omega)) d\omega \leq \ln\left(\sqrt{\overline{g}^2 + \alpha} + \overline{g}\right). \tag{49d}$$

where $\overline{g} \triangleq \frac{1}{2\pi} \int_{-\pi}^{\pi} g(\omega) d\omega$. Equality in (49b) is obtained iff ω is such that $g(\omega) = 0$.

Proof: (49a), (49b) and (49c) follow directly by algebraic manipulation. In order to show (49d), we define the functions

$$Q(x) \triangleq \sqrt{x^2 + \alpha} + x; \qquad W(x) \triangleq \ln(Q(x)), \tag{50}$$

where $\alpha > 0$ and $x \geq 0$. We have that $\frac{d^2 W(x)}{dx^2} = -x(x^2 + \alpha)^{-3/2} \leq 0$, and thus $W(x)$ is a concave function. Then, applying Jensen's inequality, we obtain $\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln(q(\omega)) d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(g(\omega)) d\omega \leq W\Big(\frac{1}{2\pi}\int_{-\pi}^{\pi} g(\omega) d\omega\Big)$, which leads directly to (49d). ∎

***Proposition** 3: Define $f_\alpha(\omega)$ as in* (25a)*, with g satisfying Assumption 1. Then, if $\exists \zeta \in \mathcal{N}_g$ such that $g(\omega)$ is $\mathcal{O}(\omega - \zeta)$ as $\omega \to \zeta$, the following holds:*

$$\lim_{\alpha \to 0^+} \|f_\alpha\|^2 = \infty.$$

Proof: The interval $[-\pi, \pi]$ can be partitioned into two disjoint sets $\mathcal{H} \triangleq \{\omega : g(\omega) > \delta\}$ and $\mathcal{I} \triangleq \{\omega : g(\omega) \leq \delta\}$ by utilizing an arbitrary "threshold" $\delta > 0$. Then, substituting (48) into (25a), we obtain

$$\|f_\alpha\|^2 = \frac{\frac{1}{2\pi}\int_{\mathcal{I}} p(\omega)^2 d\omega + \frac{1}{2\pi}\int_{\mathcal{H}} p(\omega)^2 d\omega}{\mathrm{e}^{\frac{1}{2\pi}\int_{\mathcal{I}} \ln p(\omega)^2 d\omega} \ \mathrm{e}^{\frac{1}{2\pi}\int_{\mathcal{H}} \ln p(\omega)^2 d\omega}}. \tag{51}$$

Using (49b) and (49c), we have that

$$\sqrt{\alpha} \geq p(\omega), \ \forall w \in \mathcal{I}; \qquad \frac{\alpha}{2\delta} \geq p(\omega), \ \forall w \in \mathcal{H}. \tag{52}$$

Substituting of (52) into (51) we obtain

$$\|f_\alpha\|^2 \geq \frac{\frac{1}{2\pi}\int_{\mathcal{I}} p(\omega)^2 d\omega}{\alpha^{\frac{|\mathcal{I}|}{2\pi}} \left(\frac{\alpha}{2\delta}\right)^{\frac{|\mathcal{H}|}{\pi}}} = \left[\frac{(2\delta)^{\frac{|\mathcal{H}|}{\pi}}}{2\pi}\right] \left[\frac{\int_{\mathcal{I}} p(\omega)^2 d\omega}{\alpha^{\frac{|\mathcal{H}|}{2\pi}+1}}\right], \tag{53}$$

where $|\mathcal{H}|$ and $|\mathcal{I}|$ denote the Lebesgue measures of $\mathcal{H}$ and $\mathcal{I}$, respectively.

We will next show the divergence of the last expression on the right hand side of (53) as $\alpha \to 0^+$. For this purpose, we consider two scenarios, characterized by $|\mathcal{N}_g|$, the Lebesgue measure of $\mathcal{N}_g$.

- Case i): $|\mathcal{N}_g| > 0$. Since $p(\omega)^2 = \alpha$, $\forall \omega \in \mathcal{N}_g$, and $\mathcal{N}_g \subseteq \mathcal{I}$ for any $\delta > 0$, we can obtain from (53) that

$$\|f_\alpha\|^2 \geq \left[\frac{(2\delta)^{\frac{|\mathcal{H}|}{\pi}}}{2\pi}\right] \left[\frac{\int_{\mathcal{N}_g} p(\omega)^2 d\omega}{\alpha^{\frac{|\mathcal{H}|}{2\pi}+1}}\right] = \left[\frac{(2\delta)^{\frac{|\mathcal{H}|}{\pi}}}{2\pi}\right] \left[\frac{|\mathcal{N}_g|\,\alpha}{\alpha^{\frac{|\mathcal{H}|}{2\pi}+1}}\right] = \left[\frac{(2\delta)^{\frac{|\mathcal{H}|}{\pi}}}{2\pi}\right] \frac{|\mathcal{N}_g|}{\alpha^{\frac{|\mathcal{H}|}{2\pi}}},$$

  which clearly tends to ∞ as $\alpha \to 0^+$.

- Case ii): $|\mathcal{N}_g| = 0$. The conditions of the proposition ensure the existence of scalars $\varepsilon > 0$, $L < \infty$ such that $g(\omega) \leq L\,|w - \zeta|$, if $|\omega - \zeta| < \varepsilon$. This implies that for any $\delta > 0$ there exists $\mu \in (0, \varepsilon)$ such that $[\zeta, \zeta + \mu] \subset \mathcal{I}$ and $g(\omega) \leq L\,|w - \zeta|\,, \forall \omega \in [\zeta, \zeta + \mu]$. Applying this result, and noting that $p \uparrow\downarrow g$, we have

$$\begin{aligned}\int_{\mathcal{I}} p(\omega)^2 d\omega &\geq \int_{\zeta}^{\zeta+\mu} p(\omega)^2 d\omega \\ &\geq I(\alpha) \triangleq \int_0^{\mu} (\sqrt{L^2x^2 + \alpha} - Lx)^2 dx \\ &= \mu\alpha + \tfrac{2}{3L}\left[L^3\mu^3 + \alpha^{3/2}\right] - \tfrac{2}{3L}\left[L^2\mu^2 + \alpha\right]^{3/2}.\end{aligned}$$

  After substituting the above inequality into the right hand side of (53), choosing $\delta > 0$ small enough so as to ensure[20] $|\mathcal{H}| > \pi$, it is easy to verify that

$$\lim_{\alpha \to 0^+} \|f_\alpha\|^2 \geq \lim_{\alpha \to 0^+} \left[\frac{(2\delta)^{\frac{|\mathcal{H}|}{\pi}}}{2\pi}\right] \frac{I(\alpha)}{\alpha^{\frac{|\mathcal{H}|}{2\pi}+1}} = \infty.$$

This completes the proof. ∎

B. Proof of Proposition 1

From the fact that $F(z)$ is stable and strictly causal, we have that $\|F\|^2 = \|1 - F\|^2 - 1$. Therefore, the denominators of the right hand side terms of (16) and (17) are equal. Denote the numerator of the right side term of (16) as $N \triangleq \|\Omega_x A\|^2 \|(1 - F)PB\|^2$. Applying Cauchy-Schwartz inequality we get

$$\begin{aligned}N &\geq \langle |\Omega_x A|\,, |(1-F)PB| \rangle^2 \\ &= \langle |\Omega_x P|\,|AB|\,, |1-F| \rangle^2 = \langle |\Omega_x P|\,, |1-F| \rangle^2,\end{aligned} \tag{54}$$

where the last equality in (54) follows from (1). Substituting the last term on the right hand side of (54) into (16) yields (17), which is obtained iff equality holds in (54). In turn, equality in (54) is achieved iff $|\Omega_x A| = \kappa^2\,|(1 - F)PB|\,,$ a.e. on $[-\pi, \pi]$, for arbitrary $\kappa^2 \in \mathbb{R}^+$. This equation, when combined with (1) and (2), leads directly to (18).

In order to prove the second part of the proposition, we note that for any $\varepsilon_A, \varepsilon_B > 0$, the functions $A^{[\boldsymbol{\varepsilon}]}, B^{[\boldsymbol{\varepsilon}]} \in L^2$, and $A^{[\boldsymbol{\varepsilon}]}(\omega), B^{[\boldsymbol{\varepsilon}]}(\omega) > 0$, $\forall \omega \in [-\pi, \pi]$. As a consequence, one can always find causal, rational and stable filters $A(z)$ and $B(z)$ satisfying (19). Secondly, the difference between $\sigma^2_{\epsilon\, inf|F}$ and σ^2_ϵ when $\left|A(\mathrm{e}^{j\omega})\right|$ and $\left|B(\mathrm{e}^{j\omega})\right|$ satisfy (19) is given by

$$\sigma^2_\epsilon - \sigma^2_{\epsilon\, inf|F} = \frac{N^{[\boldsymbol{\varepsilon}]} - N_{inf}}{\gamma - \|F\|^2}, \tag{55}$$

where $N^{[\boldsymbol{\varepsilon}]} \triangleq \|\Omega_x A^{[\boldsymbol{\varepsilon}]}\|^2 \|P(1 - F)B^{[\boldsymbol{\varepsilon}]}\|^2$ and $N_{inf} \triangleq \|\Omega_x A_{inf}\|^2 \|P(1 - F)B_{inf}\|^2$. Defining

$$\begin{aligned}e_A(\mathrm{e}^{j\omega}) &\triangleq A^{[\boldsymbol{\varepsilon}]}(\omega) - \left|A_{inf}(\mathrm{e}^{j\omega})\right|, \\ e_B(\mathrm{e}^{j\omega}) &\triangleq B^{[\boldsymbol{\varepsilon}]}(\omega) - \left|B_{inf}(\mathrm{e}^{j\omega})\right|,\end{aligned}$$

[20] This is always possible since $|\mathcal{N}_g| = 0$.

and $f(\omega)$ as in (20), we can write

$$\begin{aligned}
N^{[\varepsilon]} - N_{inf} &= \|\Omega_x(|A_{inf}| + e_A)\|^2 \|fP(|B_{inf}| + e_B)\|^2 \\
&\qquad - \|\Omega_x A_{inf}\|^2 \|fPB_{inf}\|^2 \\
&= \|\Omega_x A_{inf}\|^2 \left(\|fPe_B\|^2 + 2\langle |P|^2 f^2 |B_{inf}|, e_B\rangle\right) \\
&\quad + \|fPB_{inf}\|^2 \left(\|\Omega_x e_A\|^2 + 2\langle |\Omega_x|^2 |A_{inf}|, e_A\rangle\right) \\
&= N_{inf}^{\frac{1}{2}}\Big[\|fP\, e_B\|^2 + \|\Omega_x e_A\|^2 + 2\langle |P|^2 f^2 |B_{inf}|, e_B\rangle \\
&\qquad + 2\langle |\Omega_x|^2 |A_{inf}|, e_A\rangle\Big].
\end{aligned}$$

Each of the terms above can be upper bounded as follows

$$\begin{aligned}
\|fP\, e_B\|^2 &\overset{(a)}{\leq} \int_{\mathcal{I}_{\varepsilon_A}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\omega)^2 \varepsilon_A^2 d\omega \\
&\quad + \int_{\mathcal{I}_{\varepsilon_B}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\omega)^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right|^2 d\omega \\
&\overset{(b)}{\leq} \varepsilon_A^2 \|fP\|^2 + \int_{\mathcal{I}_{\varepsilon_B}} \left|P(\mathrm{e}^{j\omega})\right| \left|\Omega_x(\mathrm{e}^{j\omega})\right| f(\omega) d\omega \\
&\overset{(c)}{\leq} \varepsilon_A^2 \|fP\|^2 + \varepsilon_B^2 \|\Omega_x\|^2/\kappa^2. \\
\|\Omega_x e_A\|^2 &\overset{(d)}{\leq} \int_{\mathcal{I}_{\varepsilon_B}} \left|\Omega_x(\mathrm{e}^{j\omega})\right|^2 \varepsilon_B^2 d\omega \\
&\quad + \int_{\mathcal{I}_{\varepsilon_A}} \left|\Omega_x(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right|^2 d\omega \\
&\overset{(e)}{\leq} \varepsilon_B^2 \|\Omega_x\|^2 + \int_{\mathcal{I}_{\varepsilon_A}} \left|\Omega_x(\mathrm{e}^{j\omega})\right| \left|P(\mathrm{e}^{j\omega})\right| f(\omega) d\omega \\
&\overset{(f)}{\leq} \varepsilon_B^2 \|\Omega_x\|^2 + \varepsilon_A^2 \|fP\|^2 \kappa^2.
\end{aligned}$$

$$\begin{aligned}
\langle |P|^2 f(\omega)^2 |B_{inf}|, e_B\rangle &\overset{(g)}{\leq} \int_{\mathcal{I}_{\varepsilon_A}} \left|P(\mathrm{e}^{j\omega})\right|^2 f(\omega)^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right| \varepsilon_A d\omega \\
&\overset{(h)}{\leq} \varepsilon_A^2 \|fP\|^2. \\
\langle |\Omega_x|^2 |A_{inf}|, e_A\rangle &\overset{(i)}{\leq} \int_{\mathcal{I}_{\varepsilon_B}} \left|\Omega_x(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right| \varepsilon_B d\omega \\
&\overset{(j)}{\leq} \varepsilon_B^2 \|\Omega_x\|^2
\end{aligned}$$

In the above, (a) follows from the fact that

$$\left|e_B(\mathrm{e}^{j\omega})\right| \leq \varepsilon_A, \forall \omega \in \mathcal{I}_{\varepsilon_A}, \quad \text{and} \tag{56a}$$

$$-\left|B_{inf}(\mathrm{e}^{j\omega})\right| < e_B(\mathrm{e}^{j\omega}) < 0, \forall \omega \in \mathcal{I}_{\varepsilon_B}. \tag{56b}$$

(b) follows from the fact that

$$\begin{aligned}
&\left|P(\mathrm{e}^{j\omega})\right|^2 \left|1 - F(\mathrm{e}^{j\omega})\right|^2 \left|B_{inf}(\mathrm{e}^{j\omega})\right|^2 \\
&\qquad = \left|\Omega_x(\mathrm{e}^{j\omega})\right|^2 \left|A_{inf}(\mathrm{e}^{j\omega})\right|^2 \\
&\qquad = \left|P(\mathrm{e}^{j\omega})\right| \left|\Omega_x(\mathrm{e}^{j\omega})\right| \left|1 - F(\mathrm{e}^{j\omega})\right|,
\end{aligned} \tag{57}$$

$\forall \omega \in [-\pi, \pi]$, see (18), and from $\mathcal{I}_{\varepsilon_A} \subset [-\pi, \pi]$. Inequality (c) follows from the fact that

$$\left|\Omega_x(\mathrm{e}^{j\omega})\right| < \varepsilon_A^2 \kappa^2 \left|P(\mathrm{e}^{j\omega})\right| f(\omega), \quad \forall \omega \in \mathcal{I}_{\varepsilon_A}; \tag{58a}$$

$$\left|P(\mathrm{e}^{j\omega})\right| < \varepsilon_B^2 \kappa^{-2} \left|\Omega_x(\mathrm{e}^{j\omega})\right| f(\omega)^{\sim 1}, \quad \forall \omega \in \mathcal{I}_{\varepsilon_B}, \tag{58b}$$

which is readily obtained from (18) and (19). Inequality (d) follows from

$$\left|e_A(\mathrm{e}^{j\omega})\right| \leq \varepsilon_B, \forall \omega \in \mathcal{I}_{\varepsilon_B}, \quad \text{and} \tag{59a}$$

$$-\left|A_{inf}(\mathrm{e}^{j\omega})\right| < e_A(\mathrm{e}^{j\omega}) < 0, \forall \omega \in \mathcal{I}_{\varepsilon_A}. \tag{59b}$$

Inequality (e) is due to (57) and to the fact that $\mathcal{I}_{\varepsilon_B} \subset [-\pi, \pi]$. Inequality (f) stems from (58). Inequality (g) follows from (56), while (h) follows from the fact that $\left|B_{inf}(\mathrm{e}^{j\omega})\right| \leq \varepsilon_A$, $\forall \omega \in \mathcal{I}_{\varepsilon_A}$. Inequality (i) stems from (59), while (j) follows from the fact that $\left|A_{inf}(\mathrm{e}^{j\omega})\right| \leq \varepsilon_B$, $\forall \omega \in \mathcal{I}_{\varepsilon_B}$. Therefore,

$$\begin{aligned}
&N^{[\varepsilon]} - N_{inf} \\
&\qquad \leq N_{inf}^{1/2} \left[(3 + \kappa^2)\|fP\|^2 \varepsilon_A^2 + (3 + \kappa^{-2})\|\Omega_x\|^2 \varepsilon_B^2\right],
\end{aligned}$$

which completes the proof. ∎

C. Proof of Lemma 1

Define the partition $-\pi = \omega_0 < \omega_1 < \cdots < \omega_p = \pi$, where $\{\omega_i\}_{i=1}^{p-1}$ correspond to the discontinuity points (if any) of f. Since f is piece-wise differentiable, its first derivative over all open intervals (ω_i, ω_{i+1}), $i \in \{0, \ldots, p-1\}$ is bounded by a constant $0 \leq S < \infty$. For each $m > S$, we define the set $\mathcal{R}_m$, consisting of all *continuous* functions $h : [-\pi, \pi] \to \mathbb{R}^+$ satisfying

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \log h(\omega) d\omega = 0, \tag{60a}$$

$$f_{min} \leq h(\omega) \leq f_{max}, \quad \forall \omega \in [-\pi, \pi], \text{ and} \tag{60b}$$

$$\left|\frac{d}{d\omega} h(\omega)\right| \leq m, \ \forall \omega \in [-\pi, \pi]. \tag{60c}$$

For each m, the function

$$h_m \triangleq \arg\min_{h \in \mathcal{R}_m} \|f - h\|. \tag{61}$$

is the element in $\mathcal{R}_m$ "closest" to f. From (23b), and from the fact that f is piece-wise differentiable, it follows that for every $\varepsilon_0 > 0$, there exists a bounded $T \geq S$ such that

$$\|f - h_m\| \leq \varepsilon_0, \quad \forall m > T. \tag{62}$$

(Indeed, it is easy to obtain the bound $\|f(\omega) - h_m(\omega)\| \leq (f_{max} - f_{min})^2 p/m$). Notice that if f had no discontinuity points and if $m \geq S$, then $h_m \equiv f$ (see (60c)), yielding $\|f - h_m\| = 0$.

Since $h_m(\omega)$ is continuous and piece-wise differentiable, its Fourier series converges uniformly over $[-\pi, \pi]$. Thus, for every $\varepsilon_1 > 0$, there exists an N-th order (where $N < \infty$ is odd and depends on ε_1) rational transfer function $H_N(z)$ (the Z-transform of the coefficients of the $\frac{N-1}{2}$-th partial sum of the Fourier series of f) such that

$$\left|h_m(\omega) - H_N(\mathrm{e}^{j\omega})\right| < \varepsilon_1, \ \forall \omega \in [-\pi, \pi]. \tag{63}$$

$H_N(z)$ can be written as $H_N(z) = G_1 z^{-\frac{N+1}{2}} \prod_{i=1}^{N}(z - c_i)$, where $G_1 \in \mathbb{R}$. Thus, the transfer function

$$H'_N(z) \triangleq H_N(z) \frac{G_1}{|G_1|} z^{-\frac{N-1}{2}} \prod_{\substack{i=1, \\ |c_i|>1}}^{N} \frac{c_i}{|c_i|} \left(\frac{c_i^* z - 1}{z - c_i}\right)$$

is clearly biproper, stable, minimum-phase and such that $\left|H'_N(\mathrm{e}^{j\omega})\right| = \left|H_N(\mathrm{e}^{j\omega})\right|$, $\forall \omega \in [-\pi, \pi]$, with the first value of its impulse response being

$$\chi \triangleq \lim_{z\to\infty} H'_N(z) > 0.$$

Define $\tilde{H}_N(z) \triangleq \frac{1}{\chi} H'_N(z)$, so that $\lim_{z\to\infty} \tilde{H}_N(z) = 1$ and

$$\left|\tilde{H}_N(\mathrm{e}^{j\omega})\right| = \frac{1}{\chi}\left|H_N(\mathrm{e}^{j\omega})\right|, \quad \forall \omega \in [-\pi, \pi]. \tag{64}$$

With the choice $F(z) = 1 - \tilde{H}_N(z)$, we have

$$\begin{aligned}\|f - |1-F|\| = \left\|f - |\tilde{H}_N|\right\| &\le \|f - h_m\| + \left\|h_m - |\tilde{H}_N|\right\| \\ &\le \varepsilon_0 + \varepsilon_1 + \left\|h_m - |\tilde{H}_N|\right\|. \end{aligned} \tag{65}$$

We now proceed to upper bound the last term in the above inequality. From (63) and (64), we have that

$$\begin{aligned}\left\|h_m - |\tilde{H}_N|\right\| &\le \|h_m - |H_N|\| + \left\||H_N| - |\tilde{H}_N|\right\| \\ &\le \varepsilon_1 + \left|1 - \frac{1}{\chi}\right| \|H_N\| = \varepsilon_1 + \frac{|\chi - 1|}{\chi}\|H_N\|. \end{aligned} \tag{66}$$

From Jensen's formula (see, e.g., [23]), and since $H'_N(z)$ is stable and minimum phase, we obtain

$$\log \chi = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|H'_N(\mathrm{e}^{j\omega})\right| d\omega. \tag{67}$$

Recalling from (60a) and (61) that $\frac{1}{2\pi}\int_{-\pi}^{\pi} \log h_m(\omega) d\omega = 0$, we can write (67) as

$$\begin{aligned}\log \chi &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\frac{\left|H_N(\mathrm{e}^{j\omega})\right|}{h_m(\omega)}\right) d\omega \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\frac{h_m(\omega) + e(\omega)}{h_m(\omega)}\right) d\omega, \end{aligned} \tag{68}$$

where $e(\omega) \triangleq \left|H_N(\mathrm{e}^{j\omega})\right| - h_m(\omega)$. From (63), we have that

$$|e(\omega)| = \left|h_m(\omega) - \left|H_N(\mathrm{e}^{j\omega})\right|\right| \le \left|h_m(\omega) - H_N(\mathrm{e}^{j\omega})\right| \le \varepsilon_1.$$

Thus, choosing $\varepsilon_1 < f_{min}$, the last integral in (68) can be upper and lower bounded as

$$\begin{aligned}\log\left(\frac{f_{min} - \varepsilon_1}{f_{min}}\right) &\le \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\frac{h_m(\omega) + e(\omega)}{h_m(\omega)}\right) d\omega \\ &\le \log\left(\frac{f_{min} + \varepsilon_1}{f_{min}}\right). \end{aligned}$$

It then follows from (68) that

$$1 - \frac{\varepsilon_1}{f_{min}} \le \chi \le 1 + \frac{\varepsilon_1}{f_{min}} \iff |\chi - 1| \le \frac{\varepsilon_1}{f_{min}}$$

Substituting the latter into (66), we obtain

$$\begin{aligned}\left\|h_m - |\tilde{H}_N|\right\| &\le \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1}\|H_N\| \\ &\le \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1}(\|f\| + \varepsilon_0 + \varepsilon_1), \end{aligned} \tag{69}$$

where the last inequality stems from (62) and (63). Substitution of (69) into (65) yields

$$\|f - |1-F|\| \le \varepsilon_0 + \varepsilon_1 + \frac{\varepsilon_1}{f_{min} - \varepsilon_1}(\|f\| + \varepsilon_0 + \varepsilon_1). \tag{70}$$

Since $\|f\|$ is bounded, and from (23b), it follows from (70) that for any $\varepsilon > 0$, one can always choose sufficiently large (bounded) values for T (see (62)) and N (see (63)) so that ε_0 and ε_1 are small enough to yield $\|f - |1-F|\| < \varepsilon$. This completes the proof. ∎

D. Proof of Theorem 1

Denote the squared norm of $f^\star$ (see (24)) via $c_{opt} \triangleq \|f^\star\|^2$, and define the set of all the $f \in \mathcal{C}_2$ having the same norm as $f^\star$ by $\mathcal{M}_{c_{opt}} \triangleq \{f \in \mathcal{C}_2 : \|f\|^2 = c_{opt}\}$. Define

$$\mathcal{B}_0 \triangleq \left\{ f \in \mathcal{C}_2 : \int_{-\pi}^{\pi} \ln\left(f(\omega)\right) d\omega = 0 \right\} \subset \mathcal{C}_2. \tag{71}$$

It is easy to show[21] that $f^\star$ must belong to $\mathcal{B}_0$. From this, and since $\{\mathcal{B}_0 \cap \mathcal{M}_{c_{opt}}\} \subset \{\mathcal{C}_2 \cap \mathcal{C}_1\}$, it follows that $f^\star = \arg\min_{f\in\mathcal{B}_0\cap\mathcal{M}_{c_{opt}}} D(f)$. Minimization of $D(f)$ subject to $f \in \mathcal{B}_0 \cap \mathcal{M}_{c_{opt}}$ can be stated as the following problem[22]:

$$\text{minimize :} \qquad J(f) \triangleq \int_{-\pi}^{\pi} f(\omega)g(\omega)d\omega, \tag{72}$$

$$\text{subject to :} \qquad i)\ M(f) \triangleq \int_{-\pi}^{\pi} f(\omega)^2 d\omega = c_{opt}, \tag{73}$$

$$ii)\ H(f) \triangleq \int_{-\pi}^{\pi} \ln(f(\omega)) d\omega = 0. \tag{74}$$

The problem described by (72)-(74) falls within the category of *isoperimetrical problems*, well known in variational calculus (see, e.g., [35] and [36]). The standard solution of these problems is based upon the fact that any f that extremizes J (see (72)) *needs* to satisfy

$$\frac{\partial}{\partial f} L(f(\omega)) = 0, \quad \forall \omega \in [-\pi, \pi], \tag{75}$$

where the Lagrangian $L(f(\omega))$, in our case, is given by

$$L(f(\omega)) \triangleq f(\omega)g(\omega) + \lambda_1 f(\omega)^2 + \lambda_2 \ln(f(\omega)) \tag{76}$$

and λ_1 and λ_2 are the Lagrange multipliers, to be found by enforcing (75) and the constraints (74). Substitution of (76) into (75) yields $g(\omega) + 2\lambda_1 f(\omega) + \lambda_2 f(\omega)^{-1} = 0$, a.e. on $[-\pi, \pi]$, or, equivalently,

$$f(\omega) = \begin{cases} \beta\left(\pm\sqrt{g(\omega)^2 + \alpha} - g(\omega)\right) & , \text{if } \lambda_1 \neq 0, \\ \frac{-\lambda_2}{g(\omega)} & , \text{if } \lambda_1 = 0, \end{cases} \tag{77}$$

a.e. on $[-\pi, \pi]$, where the scalars

$$\alpha \triangleq -8\lambda_1\lambda_2, \quad \beta \triangleq \frac{1}{4\lambda_1} \tag{78}$$

[21] If $f^\star$ was such that $\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\omega) d\omega = \Psi > 0$, then $f' = f\mathrm{e}^{-\Psi} \in \mathcal{B}_0$ would clearly yield a smaller D in (21), thus contradicting the optimality of f.

[22] As will become evident in the derivation, the additional constraint $f(\omega) \ge 0$, $\forall \omega \in [-\pi, \pi]$, imposed by the definition of f (see (20)) turns out to be non-binding.

are such that the constraints in (74) are met.

We note that for the trivial case in which g is almost constant (see Definition 1), $f^\star$ is also almost constant. Applying this to constraint i) in (74) yields that, for this case, $f^\star$ is such that $f(\omega) \equiv 1$. Thus, the remainder of the proof addresses only the cases in which g is not almost constant.

In order to find $f^\star$, we will next discard the possible solutions of (77) which do not correspond to global minimizers of $D(f)$ in $\mathcal{C}_2 \cap \mathcal{C}_1$. The unique remaining function, which is obtained with $\alpha > 0$ and $\beta > 0$ in (77), will characterize the solution of Optimization Problem 2.

The Case $\lambda_1 \neq 0$: Fore this case, substitution of (77) into (74) yields that β needs to satisfy

$$|\beta| = \exp\left(-\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|\pm\sqrt{g(\omega)^2+\alpha} - g(\omega)\right| d\omega\right), \quad (79)$$

so that β can be obtained explicitly from α. Note that α can not be zero in the above expression, otherwise β would be undefined. From this, the feasible[23] sign combinations for α, the $\pm$ sign before the square root, and β in (77) are:

a) $\beta < 0$, $-\sqrt{}$, $\alpha \neq 0$;
b) $\beta < 0$, $+\sqrt{}$, $\alpha < 0$;
c) $\beta > 0$, $+\sqrt{}$, $\alpha > 0$.

We will next show that only option c) characterizes the optimum.

Discarding Option a): We show next that any solution obtained by applying option a) in (77), say f_a, yields a greater FWMSE than the choice $f(\omega) \equiv 1$. In relation to the numerator on the right hand side of (21), we have:

$$\begin{aligned}
\langle f_a, g\rangle &= \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sqrt{g(\omega)^2+\alpha} + g(\omega)\right)g(\omega)d\omega}{\mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left(\sqrt{g(\omega)^2+\alpha}+g(\omega)\right)d\omega}} \\
&\overset{(a)}{>} \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sqrt{g(\omega)^2+\alpha} + g(\omega)\right)g(\omega)d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi}\sqrt{g(\omega)^2+\alpha} + g(\omega)d\omega} \\
&\overset{(b)}{>} \frac{\frac{1}{2\pi}\int_{-\pi}^{\pi}\sqrt{g(\omega)^2+\alpha} + g(\omega)d\omega \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\omega)d\omega}{\frac{1}{2\pi}\int_{-\pi}^{\pi}\sqrt{g(\omega)^2+\alpha} + g(\omega)d\omega} \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\omega)d\omega = \langle 1, g\rangle.
\end{aligned} \quad (80)$$

Inequality (a) above stems from Jensen's inequality. Inequality (b) follows by applying Theorem 7 to the numerator of (80), together with (48) and the fact that $\sqrt{g(\omega)^2+\alpha} \uparrow\uparrow g(\omega)$. Both inequalities are strict since g is not almost constant (see Theorem 7 and Definition 1).

On the other hand $\|f_a\|^2 \geq \|1\|^2 = 1$. From the above, it follows that $V(f_a) > V(1)$ (see (21)), discarding, for all non A.E. flat g, the global optimality of the solutions associated to Option a).

Discarding Option b): The candidate solutions are now characterized by options b) and c) only. Applying (49a) to (77) and (79), these solutions take the form

$$f_\alpha(\omega) \triangleq \frac{\theta(\alpha)}{\sqrt{g(\omega)^2+\alpha} + g(\omega)}, \quad (81)$$

[23]There exist other four sign combinations, which yield $f(\omega) < 0$, $\forall \omega \in [-\pi, \pi]$, i.e., inadmissible solutions.

where $\theta(\alpha)$ is as defined in (25b), with $\alpha \in [\alpha_{min}, 0) \cup (0, \infty)$ and $\alpha \geq \alpha_{min}$, where

$$\alpha_{min} \triangleq -\min_{\omega\in[-\pi,\pi]} g(\omega)^2.$$

We will discard the optimality of option b) by showing that, if $\alpha \in [\alpha_{min}, 0)$, then $D(f_\alpha) > D(f_0)$, if $\|f_0\|^2 < K$, or else $\|f_\alpha\|^2 > K$, where K and f_0 are as defined in (40) and (87), respectively. For this purpose, define the function

$$C(\alpha) \triangleq M(f_\alpha) = \|f_\alpha\|^2 = \frac{\theta(\alpha)^2}{2\pi}\int_{-\pi}^{\pi} q(\omega)^{-2}d\omega, \quad (82)$$

with $q(\omega)$ as defined in (48). Differentiation of $C(\alpha)$ yields

$$\begin{aligned}
&\frac{dC}{d\alpha} \\
&= \frac{\theta(\alpha)^2}{2\pi}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{d\omega}{q(\omega)r(\omega)}\int_{-\pi}^{\pi}\frac{d\omega}{q(\omega)^2} - \int_{-\pi}^{\pi}\frac{d\omega}{q(\omega)^3 r(\omega)}\right],
\end{aligned} \quad (83)$$

where $r(\omega) = \sqrt{g(\omega)^2+\alpha}$ (see (48)). Application of Theorem 7 to (83) yields

$$\frac{dC}{d\alpha} \leq 0, \ \forall \alpha \in [\alpha_{min}, \infty), \quad (84)$$

with equality iff g is almost constant. In turn, option b) is feasible iff $\alpha_{min} < 0$, i.e., only if $\min_{\omega\in[-\pi,\pi]} g(\omega)^2 > 0$. We then have from (81) that $f_\alpha \to f_0$ uniformly as $\alpha \to 0$. Thus

$$\lim_{\alpha\to 0} C(\alpha) = C_0 \triangleq \|f_0\|^2. \quad (85)$$

Combining this result with (84), and considering g to be non almost constant, we obtain

$$C(\alpha) > C_0, \quad \forall \alpha \in (\alpha_{min}, 0). \quad (86)$$

If $\|f_0\|^2 < K$, then (85) and (86), combined with the fact that $\langle f_0, g\rangle = \min_{f\in\mathcal{B}_0\cap\mathcal{C}_1}\langle f, g\rangle$, yield $V(f_\alpha) = \frac{\langle f,g\rangle^2}{K-C(\alpha)} > \frac{\langle f_0,g\rangle^2}{K-C_0} = V(f_0)$, for all $\alpha \in (\alpha_{min}, 0)$. Thus, if $\|f_0\|^2 < K$, then option b) is not globally optimal. If, on the other hand, $\|f_0\|^2 \geq K$, then (86) implies that $\|f_\alpha\|^2 > K$ for all $\alpha \in (\alpha_{min}, 0)$. In this case, sign combination b) would be infeasible. It thus follows that sign combination b) cannot be globally optimal.

The Case $\lambda_1 = 0$: For the case $\lambda_1 = 0$, substitution of (77) into constraint ii) in (74) yields

$$f(\omega) = f_0(\omega) \triangleq \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln g(x)dx\right)}{g(\omega)}, \ \text{ a.e. on } [-\pi, \pi]. \quad (87)$$

Notice that $f_0 = |1 - F_0|$, i.e., the optimal noise shaping frequency response magnitude in the absence of fed back quantization noise (recall (45)). This is not surprising, since taking $\lambda_1 = 0$ amounts to removing constraint i) (which restricts the power gain of fed back quantization noise, see Fig. 1).

We will discard this option and its associated solution f_0 by showing that f_0 is either infeasible or that there exists $\alpha > 0$ such that $D(f_\alpha) < D(f_0)$.

If g did satisfy the second condition of Assumption 1, then it is easy to show that f_0 and $D(f_0)$ would not be well defined[24]. Else, if g satisfies the first condition of Assumption 1, we have $g(\omega) > 0$ for all $\omega \in [-\pi, \pi]$. This implies that f_α converges uniformly to f_0 as $\alpha \to 0^+$, and thus

$$\lim_{\alpha \to 0^+} D(f_\alpha) = D(f_0). \tag{88}$$

Recall that $D(f_\alpha) = \Phi(\alpha)$, see (27), and write

$$\Phi(\alpha) = \frac{N(\alpha)^2}{K - C(\alpha)}, \quad \alpha \in (\alpha_c, \infty), \text{ where} \tag{89}$$

$$N(\alpha) \triangleq \langle f_\alpha, g \rangle = \frac{\theta(\alpha)}{2\pi} \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)} d\omega \tag{90}$$

and $C(\alpha)$ is as defined in (82). The continuity of $\Phi(\alpha)$ stated in (88) implies that if $\frac{d\Phi}{d\alpha} < 0$ at $\alpha = 0$, then f_0 can not be the minimizer of $D(f)$. We next show that this is indeed the case. Differentiation of (89) with respect to α gives

$$\frac{d\Phi}{d\alpha} = \frac{2N(\alpha)\left[K - C(\alpha)\right] \frac{dN}{d\alpha} + N(\alpha)^2 \frac{dC}{d\alpha}}{\left[K - C(\alpha)\right]^2}. \tag{91}$$

Differentiating (90) we get

$$\frac{dN}{d\alpha} = \frac{\theta(\alpha)}{4\pi} \times \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega) r(\omega)} \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)} d\omega - \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)^2 r(\omega)} d\omega \right] \tag{92}$$

and, therefore, $\frac{dN}{d\alpha}\big|_{\alpha=0} = 0$. Substitution of this into (91), together with the fact that $0 < N(0) < \infty$ and that $\frac{dC}{d\alpha}|_{\alpha=0} < 0$ for a non almost constant g, yields that

$$\frac{d\Phi}{d\alpha}\Big|_{\alpha=0} < 0, \tag{93}$$

thus discarding the optimality of f_0.

As a result, the global optimum is characterized sign combination c), i.e, by (81) with $\alpha > 0$. Finally, the uniqueness of α_K follows directly from (84). This completes the proof. ■

Proof of Theorem 2

Since the functions $N(\alpha)$, $C(\alpha)$ and are continuously differentiable $\forall \alpha \in (\alpha_c, \infty)$, so is $\Phi(\alpha)$. We therefore have that if

$$\lim_{\alpha \to \alpha_c^+} \frac{d\Phi}{d\alpha} \leq 0 \text{ and } \lim_{\alpha \to \infty} \frac{d\Phi}{d\alpha} \geq 0, \tag{94}$$

then α_{opt}, the minimizer of $\Phi(\alpha)$, needs to satisfy

$$\frac{d\Phi}{d\alpha}\Big|_{\alpha=\alpha_{opt}} = 0. \tag{95}$$

We will first elaborate upon (95) to derive (28). Then we will prove that (94) holds.

From (91), one can see that $\frac{d\Phi}{d\alpha} = 0$ iff $K = C(\alpha) - \frac{N(\alpha)}{2\frac{dN}{d\alpha}} \frac{dC}{d\alpha}$, provided $\frac{dN}{d\alpha} \neq 0$ and $\frac{dC}{d\alpha} \neq 0$. If g is not almost constant, then this is guaranteed (apply Theorem 7 to (92) and (83)). Else, if g is A.E. flat, then $f_\alpha(\omega) = 1 \ \forall \omega \in [-\pi, \pi]$, $\forall \alpha \in [\alpha_{min}, \infty)$, and all $\alpha \in (\alpha_c, \infty)$ are optimal. Thus, for a non almost constant g, we have

$$K = \theta(\alpha)^2 \times \left[\frac{\int_{-\pi}^{\pi} \frac{d\omega}{q(\omega)^3 r(\omega)} \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)} d\omega - \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega)^2} \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)^2 r(\omega)} d\omega}{\int_{-\pi}^{\pi} \frac{d\omega}{q(\omega) r(\omega)} \int_{-\pi}^{\pi} \frac{g(\omega}{q(\omega)} d\omega - 2\pi \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)^2 r(\omega)} d\omega} \right], \tag{96}$$

where $q(\omega)$ and $r(\omega)$ are as defined in (48). Application of the identity $\frac{1}{q(\omega)^2} = \frac{1}{\alpha}\left(1 - 2\frac{g(\omega)}{q(\omega)}\right)$, (which follows from (48) and (49a)) to the numerator on the right hand side of (96) yields (28).

Now we will prove (94).

The Sign of $\lim_{\alpha \to \alpha_c^+} \frac{d\Phi}{d\alpha}$*:* Since $\alpha_c = \min\{0, \alpha_K\}$, this limit needs to be analyzed for two possible scenarios, depending on whether or not α_K is positive.

- *The Case* $\alpha_K \leq 0$ For this case, $\alpha_c = 0$, so we need to prove that $\lim_{\alpha \to 0^+} d\Phi/d\alpha < 0$. From Proposition 3 it follows that the first condition in Assumption 1 ($g(\omega) > 0 \ \forall \omega \in [-\pi, \pi]$) must necessarily hold in order to obtain$\alpha_K \leq 0$. Thus, $\Phi(\alpha)$ and its first derivatives are continuous. Therefore, in view of (93), we get $\lim_{\alpha \to 0^+} \frac{d\Phi}{d\alpha} = \frac{d\Phi}{d\alpha}\big|_{\alpha=0} < 0$.
- *The Case* $\alpha_K > 0$. For this case, we need to prove that $\lim_{\alpha \to \alpha_K^+} d\Phi/d\alpha < 0$. Rewrite (91) as

$$\frac{d\Phi}{d\alpha} = \frac{N(\alpha)}{K - C(\alpha)} \left[2\frac{dN}{d\alpha} + \frac{N(\alpha)\frac{dC}{d\alpha}}{K - C(\alpha)} \right]. \tag{97}$$

From (92), it easy to see that $\frac{dN}{d\alpha} \leq \frac{\theta(\alpha)}{2\alpha}, \quad \forall \alpha \geq 0$. On the other hand, from (49d), and given that $g \in L^1$ and $\alpha > 0$, we conclude that $\theta(\alpha)$ is bounded. Thus, $\frac{dN}{d\alpha}$ is bounded. From this, and recalling that $\frac{dC}{d\alpha} \leq 0$, $\forall \alpha > 0$, it is clear from (97) that there a value for α greater than α_K under which $K - C(\alpha)$ is small enough to render $\frac{d\Phi}{d\alpha}$ negative. Therefore, $\lim_{\alpha \to \alpha_K} d\Phi/d\alpha < 0$.

The Sign of $\lim_{\alpha \to \infty} \frac{d\Phi}{d\alpha}$*:* Substitution of (92) and (83) into (91) yields

$$\begin{aligned} \frac{d\Phi}{d\alpha} &= \frac{N(\alpha)}{[K - C(\alpha)]^2} \left[2S(\alpha) \frac{dN}{d\alpha} + N(\alpha) \frac{dC}{d\alpha} \right] \\ &= U(\alpha) \times \\ &\left[S(\alpha) \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega) r(\omega)} \int_{-\pi}^{\pi} \frac{g(\omega)}{q(\omega)} d\omega - \int_{-\pi}^{\pi} \frac{g(\omega) d\omega}{q(\omega)^2 r(\omega)} \right) \right. \\ &\left. + T(\alpha) \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega) r(\omega)} \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega)^2} - \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega)^3 r(\omega)} \right) \right] \\ &= U(\alpha) \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{d\omega}{q(\omega) r(\omega)} \int_{-\pi}^{\pi} \frac{h(\omega)}{q(\omega)} d\omega - \int_{-\pi}^{\pi} \frac{h(\omega)}{q(\omega)^3 r(\omega)} d\omega \right], \end{aligned} \tag{98}$$

[24]Or else, if we extend the support of the function $\ln(\cdot)$ by defining $\ln(0) = -\infty$, then we obtain $\eta_{xP} = 0$. This would imply $f_0(\omega) = 0$ for all ω such that $g(\omega) > 0$. Thus, since f_0 must belong to $\mathcal{C}_2$, the integral of $\ln f_0(\omega)$ over the remaining frequencies needs to be infinite. Since $\ln(x) < x, \forall x \in \mathbb{R}^+$, this implies that $\|f_0\|^2 = \infty$ (infeasible) and $D(f_0) = \infty$.

with

$$h(\omega) \triangleq S(\alpha)g(\omega)+\tfrac{T(\alpha)}{q(\omega)}; \tag{99a}$$
$$U(\alpha) \triangleq \tfrac{N(\alpha)\theta(\alpha)}{2\pi[K-C(\alpha)]^2}; \tag{99b}$$
$$S(\alpha) \triangleq K-C(\alpha); \tag{99c}$$
$$T(\alpha) \triangleq N(\alpha)\theta(\alpha). \tag{99d}$$

Direct application of Theorem 7 to (98) allows one to conclude that

$$\frac{h(\omega)}{q(\omega)} \uparrow\downarrow \frac{1}{q(\omega)r(\omega)} \implies \frac{d\Phi}{d\alpha} \geq 0. \tag{100}$$

Since $q(\omega) = Q(g(\omega))$, $h(\omega) = H(g(\omega))$ and $r(\omega) = R(g(\omega))$, where

$$\begin{aligned} Q(x) &\triangleq R(x) + x; \\ R(x) &\triangleq \sqrt{x^2+\alpha}; \\ H(x) &\triangleq S(\alpha)x + T(\alpha)/Q(x) \end{aligned}$$

are continuous functions, we have from (100) that

$$\frac{d}{dg}\tfrac{H(g)}{Q(g)} \times \frac{d}{dg}\tfrac{1}{Q(g)R(g)} < 0 \Longrightarrow \frac{d\Phi}{d\alpha} \geq 0. \tag{101}$$

Since clearly $\frac{d}{dg}\frac{1}{Q(g)R(g)} < 0$, it is only left to determine the sign of

$$\begin{aligned} \frac{d}{dg}\left(\tfrac{H(g)}{Q(g)}\right) &= \frac{d}{dg}\left([Sg+TQ^{-1}]Q^{-1}\right) \\ &= SQ^{-1} + [2TQ^{-1}+Sg]\tfrac{d}{dg}(Q^{-1}) \\ &= SQ^{-1} - [2TQ^{-1}+Sg]\left(\frac{g}{R}+1\right)Q^{-2} \\ &= \left(SQ[R-g]-2T\right)/(Q^2R) \\ &= \left(S-2T/\alpha\right)\alpha/(Q^2R). \end{aligned}$$

It follows directly from the last equation that the sign of $\frac{d}{dg}(H/Q)$ corresponds to the sign of $S(\alpha) - 2T(\alpha)/\alpha$. Thus, (101) translates into

$$\begin{aligned} & S(\alpha) - 2T(\alpha)/\alpha > 0 \\ \Longleftrightarrow & K - C(\alpha) - 2N(\alpha)\theta(\alpha)/\alpha > 0 \\ \Longrightarrow & \frac{d\Phi}{d\alpha} > 0, \end{aligned} \tag{102}$$

see (99). From inequality (49d), the left hand side of (102) is lower bounded as

$$\begin{aligned} K - \left(\sqrt{\overline{g}^{\,2}/\alpha+1}+\overline{g}/\sqrt{\alpha}\right)^2 & \left[1+2\,\overline{g}\,\sqrt{\alpha}\right] \\ & \leq K - C(\alpha) - 2N(\alpha)\theta(\alpha)/\alpha, \end{aligned}$$

where $\overline{g} = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(\omega)d\omega$. From this inequality, and since $K > 1$ and $\overline{g} < \infty$, it is clear that $\lim_{\alpha\to\infty}(S(\alpha)-2T(\alpha)) > 0$. It then follows from (102) that $\lim_{\alpha\to\infty}\frac{d\Phi}{d\alpha} \geq 0$.

Thus (94) holds, and α_{opt} needs to satisfy (95) and (28). This completes the proof. ∎

E. Proof of Theorem 3

Monotonicity: Denote the right hand side of (28) as $k(\alpha) = \theta(\alpha)^2/\alpha$. Then we have

$$\begin{aligned} \frac{dk}{d\alpha} &= \frac{\theta(\alpha)^2}{\alpha^2}\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\alpha}{q(\omega)r(\omega)}d\omega - 1\right) \\ &= -k(\alpha)\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{g(\omega)}{r(\omega)}d\omega < 0, \end{aligned} \tag{103}$$

wherein (49a) has been used. This proves the second claim in Theorem 3.

Convexity: Differentiation of (103) yields $\frac{d^2k}{d\alpha^2} = k(\alpha)\left[\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{g(\omega)}{r(\omega)}d\omega\right)^2 + \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{g(\omega)}{r(\omega)^3}d\omega\right]$, which is clearly positive for all $\alpha > 0$. This shows that the right hand side of (28) is a convex function, proving the first claim of the theorem.

Limits: In order to show that the limits (29) and (30) in Theorem 3 hold, we write $k(\alpha)$ as

$$\begin{aligned} k(\alpha) &= \theta(\alpha)^2/\alpha \\ &= \exp\left(\frac{1}{\pi}\int_{-\pi}^{\pi}\ln\left[\frac{\sqrt{g(\omega)^2+\alpha}+g(\omega)}{\sqrt{\alpha}}\right]d\omega\right). \end{aligned} \tag{104}$$

We will first prove the validity of $\lim_{\alpha\to 0^+}\theta(\alpha)^2/\alpha = \infty$. Clearly, if $g(\omega) > 0$ for all $\omega \in [-\pi,\pi]$ (condition i) of Assumption 1), then the right hand of the above equation tends to ∞ as $\alpha \to 0^+$. If this wasn't the case, then the second condition of Assumption 1) must be satisfied, and therefore the conditions of Proposition 3 are met. Applying Proposition 3 and the fact that $\theta(\alpha)^2/\alpha > \|f_\alpha\|^2$ (see (81)), it follows that $k(\alpha)$ tends to ∞ as $\alpha \to 0^+$. This proves the validity of (29).

In order to show that $\lim_{\alpha\to\infty}k(\alpha) = 1$ (i.e., (30)) holds, we first note from (104) that $\theta(\alpha)^2/\alpha \geq 1$ for all $\alpha > 0$. On the other hand, applying (49d), we obtain

$$\theta(\alpha)^2/\alpha \leq \left(\sqrt{\overline{g}^{\,2}+\alpha}+\overline{g}\right)^2/\alpha, \tag{105}$$

where $\overline{g} \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi}g(\omega)d\omega$. Since $\overline{g} < \infty$ (as required by Assumption 1), the right hand side of (105) clearly tends to 1 as $\alpha \to \infty$. Therefore, $\lim_{\alpha\to\infty}k(\alpha) = 1$. ∎

F. Proof of Theorem 4

In view of Theorem 3, it suffices to proof the limits for $\alpha \to 0^+$ and $\alpha \to \infty$, respectively. The uniform convergence of f_α to f_0 as $\alpha \to 0$ if $g(\omega) > 0\ \forall\omega \in [-\pi,\pi]$ was already shown in the proof of Theorem 1. In order to show that f_α tends uniformly to f_∞ as $\alpha \to \infty$, we write

$$f_\alpha(\omega) = \frac{\exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln\left[\frac{\sqrt{g(x)^2+\alpha}+g(x)}{\sqrt{\alpha}}\right]dx\right)\sqrt{\alpha}}{\sqrt{g(\omega)^2+\alpha}+g(\omega)}. \tag{106}$$

If $g(\omega) < \infty$, $\forall\omega \in [-\pi,\pi]$, then $\frac{\sqrt{g(\omega)^2+\alpha}+g(\omega)}{\sqrt{\alpha}}$ tends uniformly to 1 as $\alpha \to \infty$. Applying this result to (106) yields that f_α tends uniformly to $1 = f_\infty$ as $\alpha \to \infty$. ∎

G. Proof of Theorem 5

From (28) and (38) we have

$$K = \exp\left(\frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} \ln\left[\frac{\sqrt{g(\omega)^2+\alpha_{opt}}+g(\omega)}{\sqrt{\alpha_{opt}}}\right]^2 d\omega\right)$$
$$= \exp\left(\frac{1}{2\pi}\int_{-\omega_c}^{\omega_c} \ln\left[\frac{\sqrt{\lambda g_1(\omega)^2+\alpha_{opt}}+\sqrt{\lambda}g_1(\omega)}{\sqrt{\alpha_{opt}}}\right]^2 d\omega\right).$$

With the change of variable $u = \lambda\omega$, this becomes $K = \exp\left(\frac{1}{2\pi\lambda}\int_{-\pi}^{\pi} \ln\left[\frac{\sqrt{g_1(u)^2+\alpha_{opt}/\lambda}+g_1(u)}{\sqrt{\alpha_{opt}/\lambda}}\right]^2 du\right)$. Thus, by writing α_{opt} as the function $\alpha_{opt}(K,\lambda)$, we conclude that $\alpha_{opt}(K,\lambda) = \lambda\alpha_{opt}(K^\lambda,1)$. Substituting the latter and (38) into (37) we obtain

$$\begin{aligned} D^\star(K,\lambda) &= \frac{1}{4\pi}\int_{-\omega_c}^{\omega_c}\left[\sqrt{g_\lambda(\omega)^2+\alpha_{opt}(K,\lambda)}-g_\lambda(\omega)\right]g_\lambda(\omega)d\omega \\ &= \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\sqrt{g_1(u)^2+\alpha_{opt}(K^\lambda,1)}-g_1(u)\right]g_1(u)du \\ &= D^\star(K^\lambda,1). \end{aligned}$$

This completes the proof. ∎

H. Proof of Theorem 6

Applying (49c) to (37) one can write

$$\begin{aligned} D^\star(K,1) &\le \frac{\alpha_{opt}(K,1)}{4} \\ &= \frac{(K-1)\alpha_{opt}(K,1)}{4} D^{\star h}(K,1) \\ &\le \frac{K\alpha_{opt}(K,1)}{4} D^{\star h}(K,1), \end{aligned}$$

where $D^{\star h}(K,1) \triangleq \frac{1}{K-1}$ is the minimum FWMSE corresponding to $g_1(\omega) \equiv 1$ when $\lambda = 1$. Substitution of (49a) into (28) yields

$$K\alpha_{opt}(K,1) = \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\sqrt{g(\omega)+\alpha_{opt}(K,1)}+g(\omega)\right)^2 d\omega\right).$$

Since $\alpha_{opt}(K,1)$ is monotonically decreasing (see Theorem 3), it follows that $K\alpha_{opt}(K,1)$ decreases with increasing K. Since $K > 1$, this leads directly to $D^\star(K^\lambda,1) \le GD^{\star h}(K^\lambda,1)$, $\forall\lambda \ge 1$, where $G \triangleq \frac{K\alpha_{opt}(K,1)}{4}$ is independent of λ. Applying Theorem 5 to both sides of the latter inequality, we obtain $D^\star(K,\lambda) = D^\star(K^\lambda,1) \le GD^{\star h}(K^\lambda,1) = GD^{\star h}(K,\lambda)$. Since $D^{\star h}(K,\lambda)$ corresponds to the minimum FWMSE for a constant g, by virtue of (42) we have that $D^{\star h}(K,\lambda) \le K^{-\lambda}/(1-K^{-1})$. Substitution of this into the last inequality yields $D^\star(K,\lambda) \le \frac{G}{1-K^{-1}}K^{-\lambda}$. This completes the proof. ∎

REFERENCES

[1] N. Jayant and P. Noll, *Digital coding of waveforms. Principles and approaches to speech and video.* Englewood Cliffs, NJ: Prentice Hall, 1984.

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

[3] R. A. Wannamaker, "Psycho-acoustically optimal noise shaping," *J. Audio Eng. Soc.*, vol. 40, no. 7/8, pp. 611–620, July/Aug. 1992.

[4] 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.

[5] 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.

[6] F. Baqai, J.-H. Lee, A. Agar, and J. Allebach, "Digital color halftoning," *Signal Processing Magazine, IEEE*, vol. 22, no. 1, pp. 87–96, Jan 2005.

[7] C. H. Bae, J. H. Ryu, and K. W. Lee, "Suppression of harmonic spikes in switching converter output using dithered Sigma–Delta modulation," *IEEE Trans. Ind. Applicat.*, vol. 38, no. 1, pp. 159–166, Jan./Feb. 2002.

[8] E. I. Silva, G. C. Goodwin, D. E. Quevedo, and M. S. Derpich, "Optimal noise shaping for networked control systems," in *Proc. Europ. Contr. Conf.*, Kos, Greece, 2007.

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

[10] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A theory of non-subtractive dither," *IEEE Trans. Signal Processing*, vol. 48, no. 2, pp. 499–516, Feb. 2000.

[11] 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.

[12] H. Spang, III and P. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Comm. Syst.*, vol. CS-10, no. 4, pp. 373–380, Dec. 1962.

[13] R. Schreier and G. Temes, *Understanding Delta-Sigma data converters*. Wiley-IEEE Press, 2004.

[14] P. Noll, "On predictive quantizing schemes," *Bell. Syst. Tech. J.*, vol. 57, no. 5, pp. 1499–1532, May-June 1978.

[15] 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.

[16] R. Brainard and J. Candy, "Direct-feedback coders: design and performance with television signals." *Proc. IEEE*, vol. 57, no. 7, pp. 776–786, July 1969.

[17] B. S. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 3, pp. 247–254, June 1979.

[18] D. Stacey, R. Frost, and G. Ware, "Error spectrum shaping quantizers with non-ideal reconstruction filters and saturating quantizers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 3, 1991, pp. 1905–1908.

[19] S. P. Lipschitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally audible noise shaping," *J. Audio Eng. Soc.*, vol. 39, no. 11, pp. 836–852, Nov. 1991.

[20] H. Bölcskei, "Noise shaping quantizers of order $L > 1$ for "general" frame expansions." 2006, presented at the Workshop on Coarsely Quantized Redundant Representations of Signals, Banff , Alberta, Canada, 2006.

[21] J. Tuqan and P. P. Vaidyanathan, "Statistically optimum pre- and postfiltering in quantization." *IEEE Trans. Circuits Syst. II*, vol. 44, no. 1, pp. 1015–1031, December 1997.

[22] O. Guleryuz and M. Orchard, "On the DPCM conpression of Gaussian autoregressive sequences," *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 945–956, March 2001.

[23] G. F. Carrier, M. Krook, and C. Pearson, *Functions of a complex variable: theory and technique*. Ithaca, N.Y.: Hod Books, 1983.

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

[25] R. L. Burden and J. D. Faires, *Numerical analysis*, ser. The Prindle, Weber & Schmidt series in mathematics. Boston: PWS-Kent Pub. Co., 1993.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Springer, 2004.

[27] T. Berger, *Rate distortion theory: a mathematical basis for data compression*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

[28] J. O'Neal Jr., "Bounds on subjective performance measures for source encoding systems," *IEEE Trans. Inform. Theory*, vol. IT-17, no. 3, pp. 224–231, May 1971.
[29] W. R. Bennet, "Spectrum of quantized signals," *Bell Syst. Tech J.*, vol. 27, pp. 446–472, July 1948.
[30] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization." *IEEE Trans. Inform. Theory*, vol. 47, no. 3, pp. 957–977, March 2001.
[31] I. Daubechies and R. DeVore, "Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order." *Ann. of Math.*, vol. 158, no. 2, pp. 679–710, 2003.
[32] C. S. Güntürk, "One-bit sigma-delta quantization with exponential accuracy," *Commun. Pure Appl. Math.*, vol. 56, no. 11, pp. 1608–1630, 2003.
[33] M. S. Derpich, J. Østergaard, and G. C. Goodwin, "The quadratic Gaussian rate-distortion function for source uncorrelated distortions," in *Proc. Data Compression Conf.*, Snowbird, UT, March 2008, pp. 73–82.
[34] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*. Cambridge Univ. Press, 1959.
[35] R. Weinstock, *Calculus Of Variations With Applications to Physics and Engineering*. New York: Dover Publications, Inc., 1974.
[36] I. M. Gelfand and S. V. Fomin, *Calculus of variations*. Englewood Cliffs, New Jersey: Prentice Hall Inc., 1963.

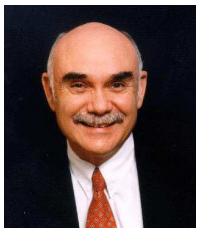
Milan S. Derpich received the Ingeniero Civil Electrónico degree from the Universidad Técnica Federico Santa María (UTFSM), Valparaíso, Chile in 1999. During his time at the university he was supported by a full scholarship from the alumni association and upon graduating received several university-wide prizes. Since 2005 he has been with the School of Electrical Engineering and Computer Science, The University of Newcastle, Australia, working toward a Ph.D. under the supervision of Professor Graham Goodwin and Dr. Daniel Quevedo. Mr. Derpich also worked by the electronic circuit design and manufacturing company Protonic Chile S.A. between 2000 and 2004. He received the Guan Zhao-Zhi Award at the Chinese Control Conference 2006. His main research interests include sampling, quantization, rate-distortion theory, communications and networked control systems.

Eduardo I. Silva was born in Valdivia, Chile, in 1979. He received the Ingeniero Civil Electrónico and Magister en Ingeniería Electrónica degrees from the Universidad Técnica Federico Santa María (UTFSM), Valparaso, Chile in 2004. He worked as a Research Assistant at the UTFSM during 2005. He is currently working toward the Ph.D at the School of Electrical Engineering and Computer Science, The University of Newcastle, Australia. His research interests include multivariate control systems, performance limitations, decentralized control, networked control and signal processing.

Daniel E. Quevedo received Ingeniero Civil Electrónico and Magister en Ingeniería Electrónica degrees from the Universidad Técnica Federico Santa María, Valparaíso, Chile in 2000. In 2005, he received the Ph.D. degree from The University of Newcastle, Australia, where he currently holds a research academic position.

He has lectured at the Universidad Técnica Federico Santa María and The University of Newcastle. He also has working experience at the VEW Energie AG, Dortmund, Germany and at the Cerro Tololo Inter-American Observatory, La Serena, Chile. His research interests cover several areas of automatic control, signal processing, and communications.

Daniel was supported by a full scholarship from the alumni association during his time at the Universidad Técnica Federico Santa María and received several university-wide prizes upon graduating. He received the IEEE Conference on Decision and Control Best Student Paper Award in 2003 and was also a finalist in 2002.



Graham C. Goodwin obtained a B.Sc (Physics), B.E (Electrical Engineering), and Ph.D from the University of New South Wales. He is currently Laureate Professor of Electrical Engineering at the University of Newcastle, Australia and is Director of an Australian Research Council Centre of Excellence for Complex Dynamic Systems and Control. He holds Honorary Doctorates from Lund Institute of Technology, Sweden and the Technion Israel. He is the co-author of eight books, four edited books, and many technical papers. Graham is the recipient of Control Systems Society 1999 Hendrik Bode Lecture Prize, a Best Paper award by IEEE Transactions on Automatic Control, a Best Paper award by Asian Journal of Control, and 2 Best Engineering Text Book awards from the International Federation of Automatic Control in 1984 and 2005. In 2008 he will receive the Quazza Medal from the International Federation of Automatic Control. He is a Fellow of IEEE; an Honorary Fellow of Institute of Engineers, Australia; a Fellow of the International Federation of Automatic Control, a Fellow of the Australian Academy of Science; a Fellow of the Australian Academy of Technology, Science and Engineering; a Member of the International Statistical Institute; a Fellow of the Royal Society, London and a Foreign Member of the Royal Swedish Academy of Sciences.