# Optimal Noise-Shaping DPCM

Milan S. Derpich, Eduardo I. Silva, Daniel E. Quevedo and Graham C. Goodwin
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 the optimal design of Noise-Shaping Differential Pulse-Coded Modulation coders. The main contribution resides in the derivation of explicit analytic formulas for the optimal filters and the minimum achievable frequency weighted reconstruction error. A novel aspect in the analysis is the fact that we account for fed-back quantization noise and that we make no restrictions on the order of the filters deployed.**

## I. Introduction

Analog-to-Digital converters which utilize a scalar quantizer and linear, time invariant filters in a feedback loop have been extensively employed as a source coding method since the concept was first introduced in the 1960's. The generalized form of this architecture, which we denote *Noise Shaping Differential Pulse Code Modulation*[1] (NS-DPCM), can be represented as in Fig. 1. The filters in a NS-DPCM system allow one to account for the correlation between consecutive input samples, and to spectrally shape the quantization noise in the output, so as to minimize the *frequency weighted mean square reconstruction error* (FWMSE). Special cases of the NS-DPCM architecture include $\Delta$-Modulators, DPCM converters [4], and noise-shaping converters, such as one and multi-bit *Sigma-Delta* modulators [5]. NS-DPCM converters are extensively used in the context of audio compression [6], digital image half-toning [7] and oversampled A/D conversion [8].

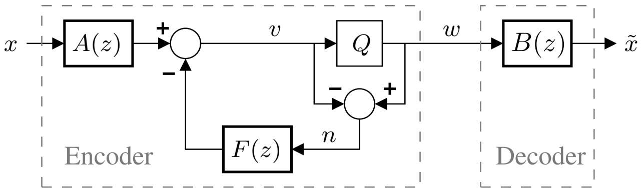

Fig. 1: Noise Shaping-DPCM Encoder and Decoder

Provided that the input power spectral density (PSD), frequency weighting error criterion, and scalar quantizer characteristics are known, the design of an NS-DPCM converter that achieves minimum FWMSE amounts to finding the corresponding optimal filters. This has been an intense area of research for at least 40 years. However, available to date results on optimal filter design for NS-DPCM encoders have been obtained assuming either fixed, finite order filters [1], [2], [8]–[10], negligible fed back quantization noise [3], [11], or have relied upon heuristic design methods [2], [9]. Since optimal performance can, in general, only be attained by arbitrary order filters designed accounting for fed back quantization noise, an exact characterization of the optimal performance (and filters) for NS-DPCM converters has remained an open problem.

In this paper we derive an explicit analytic expression for the optimal performance (and filter frequency responses) for NS-DPCM converters. We characterize the scalar quantizer via its signal-to-noise ratio, and adopt a white quantization noise model [12]. The performance bound obtained corresponds to the minimum FWMSE that can be achieved by an NS-DPCM encoder-decoder with any linear, time-invariant filters. A key departure from [3] (which, to the best of our knowledge, gives the only currently available explicit analytic solutions to the problem), is that we account for fed back quantization noise. This allows us to derive exact expressions.

Our results show that an optimal NS-DPCM converter exhibits several interesting properties. These include a spectrally flat frequency weighted error spectrum, and a white signal at the input of the scalar quantizer. We also show that, for AR Gaussian sources, the rate-distortion efficiency with the optimal filters depends only on how efficient the embedded scalar quantizer is at quantizing nearly Gaussian samples.

### *Notation and Preliminaries*

We use standard vector space notation for signals. For example, $x$ is used to denote $\{x(k)\}_{k\in\mathbb{Z}}$. We also use $z$ as the argument of the z-transform. Given two square integrable complex valued functions $f(\omega)$ and $g(\omega)$ defined over $[-\pi,\pi]$, we adopt the inner product $\langle f,g\rangle \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega)^* g(\omega)d\omega$, where $()^*$ denotes complex conjugation. We denote the usual 2-norm as $\|f\| \triangleq \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}|f(\omega)|^2 d\omega}$. If $F(z)$ is a transfer function, then we use the short hand notation $F$ to refer to the associated frequency response $F(\mathrm{e}^{j\omega})$, $\omega\in[-\pi,\pi]$. If $I$ is a set, then we will write "a.e. on $I$" (almost everywhere on $I$) as a short hand notation for "everywhere on $I$ except at most on a zero Lebesgue measure set of points".

We use $\sigma_x^2$ to denote the variance of a given wide sense stationary (w.s.s.) random process $x$, having PSD $S_x(\mathrm{e}^{j\omega})$. Note that $\sigma_x^2 \triangleq \mathcal{E}\left\{x(k)^2\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_x(\mathrm{e}^{j\omega})d\omega = \|\Omega_x\|^2$, where $\Omega_x$ is a frequency response satisfying $|\Omega_x| \triangleq \sqrt{S_x}$, $\forall\omega\in[-\pi,\pi]$. For a given function $f:[-\pi,\pi]\to\mathbb{C}$, we define $\eta_f \triangleq \exp\left(\frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|dx\right)$ (provided this integral converges). This allows one to describe the *Kolmogorov's minimal prediction error variance* [13] of a w.s.s. process $x$ via $\eta_x^2 \triangleq \eta_{S_x} = \eta_{\Omega_x}^2$. The *spectral flatness measure* of a w.s.s. process $x$ is denoted

[1] The same configuration can be found under different names in the literature, e.g.: *error feedback systems* [1], *direct feedback coders* [2] and *DPCM with noise feedback* [3].

by $\zeta_x \triangleq \frac{\eta_x^2}{\sigma_x^2}$. It is easy to show that $0 \leq \zeta_x \leq 1$, and that $\zeta_x = 1$ if and only if $S_x(\mathrm{e}^{j\omega})$ is constant a.e. on $[-\pi, \pi]$.

## II. NS-DPCM Model

As foreshadowed in the introduction, we consider the general form of an NS-DPCM architecture shown in Figure 1. In our model, the input sequence $x$ is assumed to be a zero mean, w.s.s. random process, with known PSD $S_x = |\Omega_x|^2$ satisfying $S_x(\mathrm{e}^{j\omega}) > 0,$ a.e. on $[-\pi, \pi]$. The element denoted by $\mathcal{Q}$ describes a scalar quantizer, with given and known characteristics[2]. For each input $v(k)$, $k \in \mathbb{Z}$, it outputs $w(k)$ and generates the *quantization error* $n(k) \triangleq w(k) - v(k)$. The three discrete-time filters $A(z)$, $B(z)$ and $F(z)$ in Fig. 1 are design choices.

To asses performance, we introduce the delay-compensated frequency weighted error

$$\epsilon \triangleq P(z)(\tilde{x} - z^{-\tau}x), \tag{1}$$

where $\tau \geq 0$. The *error weighting filter* $P(z)$ models the impact that reconstruction errors have on each frequency. Thus, it is application dependent.

In this paper, we restrict attention to the cases in which $\left|P(\mathrm{e}^{j\omega})\right| > 0$, $\forall \omega \in [-\pi, \pi]$, i.e., $P(z)$ has no zeros on the unit circle. Additionally, we require:

***Constraint** 1:* $A(z), B(z)$, $F(z)$ and $P(z)$ are stable. In addition, $F(z)$ is strictly causal (i.e., $\lim_{z\to\infty} F(z) = 0$). △

The first part in the above constraint is required in order to avoid unbounded signals in the NS-DPCM converter. The additional requirement on $F(z)$ is needed for the feedback loop in Fig. 1 to be well defined (see, e.g., [5, Chap. 4]).

Since the NS-DPCM architecture embeds a nonlinear element (a scalar quantizer) within a feedback loop, exact analysis of quantization errors is, in general, a formidable task [15]. This has motivated the widespread use of an additive noise model for quantization errors [1]–[3], [8]–[12]. This model allows one to study the converter via linear analysis tools. It is usually formulated as follows:

***Assumption** 1: The quantization errors are i.i.d. random variables, uncorrelated with the input signal.* △

In order not to limit our subsequent analysis to a specific type of scalar quantizer, the following is also assumed:

***Assumption** 2: The probability density function (PDF) of $v$ is not affected by the filters in the converter other than via its second moment*[3]. △

Under Assumption 2, any given type of scalar quantizer with a fixed number of quantization levels leads to quantization errors whose variance is proportional to the variance of its input. This can be stated as

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

where $\gamma$ is the *signal-to-noise ratio* of the scalar quantizer (not to be confused with that of the NS-DPCM encoder-decoder system). $\gamma$ depends on the number of quantization levels, the PDF of the signal being quantized and the companding characteristics of the scalar quantizer itself[4].

## III. Formulation of the Optimization Problem

Our ultimate goal is to find the frequency responses of the filters $A$, $B$, and $F$ that minimize the variance of $\epsilon$ under Assumptions 1 and 2, and for given and known $\Omega_x$, $P$ and $\gamma$. The quantity $\sigma_\epsilon^2$ so obtained will constitute the (achievable) lower bound on the FWMSE for the NS-DPCM converter.

Towards the above goal, we first derive an expression that relates the decision variables to the error measure that we wish to minimize. From Fig. 1, equation (1), Assumption 1, and recalling that $|\Omega_x| = \sqrt{S_x}$, $\forall \omega \in [-\pi, \pi]$, we have

$$\sigma_\epsilon^2 = \sigma_n^2 \|(1 - F)BP\|^2 + \|\,(W-1)\,\Omega_x P\|^2, \tag{3}$$

where $\sigma_n^2$ is the variance of the quantization error, and

$$W(\mathrm{e}^{j\omega}) \triangleq \mathrm{e}^{j\omega\tau} A(\mathrm{e}^{j\omega}) B(\mathrm{e}^{j\omega}), \quad \forall \omega \in \mathbb{R}, \tag{4}$$

is a delay compensated version of $AB$, the frequency response from $x$ to $\tilde{x}$. The first term on the right hand side of (3) corresponds to the variance of the frequency weighted quantization error in $\epsilon$. The second term in (3) accounts for the frequency weighted linear distortion introduced by the filters in the encoder-decoder pair[5].

The variance $\sigma_n^2$ is related to $\sigma_v^2$ via (2). From Assumption 1, the latter is given by $\sigma_v^2 = \|A\Omega_x\|^2 + \sigma_n^2\|F\|^2$. Combining this result with (2) gives

$$\sigma_n^2 = \frac{\|A\Omega_x\|^2}{\gamma - \|F\|^2}. \tag{5}$$

When substituted into (3), this yields

$$\sigma_\epsilon^2 = \frac{\|A\Omega_x\|^2 \|(1-F)BP\|^2}{\gamma - \|F\|^2} + \|\,(W-1)\,\Omega_x P\|^2. \tag{6}$$

The above expression relates the filters $A(z), B(z), F(z)$, and the quantizer signal-to-noise ratio $\gamma$, to the FWMSE. Minimization of this cost functional will yield expressions for the optimal filters and performance.

For comparison, we note that the cost functional (6), together with Assumptions 1 and 2, is also part of the analysis in [1], [3] and [10], wherein equivalent optimization problems are addressed [6].

Finally, we note that, since $\sigma_\epsilon^2$ must be positive, (6) implies

***Constraint** 2:* $\|F\|^2 < \gamma$. △

[2] This may include, for example, any of the scalar quantizers described in [14].

[3] This can be expected to be a realistic approximation especially when $x$ is a first-order Gaussian AR source. Indeed, it has been shown in [16] that the prediction error in a DPCM converter with a first-order Gaussian AR input is close to Gaussian, even for as few as two quantization levels.

[4] The actual bit-rate associated with $\gamma$ will also depend on whether or not entropy coding is utilized to encode $w$, as discussed in Section V.

[5] Note that perfect reconstruction (in the absence of quantization errors) is achieved if and only if there is no linear distortion, i.e., when $W = 1$.

[6] We note that quantization noise is not assumed white in [1], and that [10] only considers $P = 1$, restricts to first order AR Gaussian inputs, and minimizes $\gamma$ for a given $\sigma_\epsilon^2$.

## IV. Optimal NS-DPCM

In this section we derive explicit analytic expressions for the optimal filters and the associated optimal performance for the NS-DPCM scheme, subject to a mild restriction. The analysis is based on a set of equations that the optimal filters must necessarily satisfy. To facilitate the flow of ideas, all proofs are given in the Appendix.

Minimization of (6) is simplified by noting that, for stable and strictly causal $F(z)$, it holds that $\|F\|^2 = \|1-F\|^2 - 1$. Substitution of this equality into (6) yields

$$\sigma_\epsilon^2 = \frac{\|A\Omega_x\|^2 \|(1-F)BP\|^2}{\gamma + 1 - \|1-F\|^2} + \|\left(W-1\right)\Omega_x P\|^2. \qquad (7)$$

We then have the following result:

*Lemma 1: For given frequency responses $F$ and $W$, the optimal $A(z)$ satisfies*

$$|A| = \kappa\sqrt{|P|\,|\Omega_x|^{-1}\,|1-F|\,|W|}, \quad \textit{a.e. on } [-\pi,\pi], \qquad (8)$$

*where $\kappa > 0$ is an arbitrary constant. This choice yields*

$$\sigma_\epsilon^2 = \frac{\langle |1-F|\,, |\Omega_x P|\,|W|\rangle^2}{\gamma+1-\|1-F\|^2} + \|\left(W-1\right)\Omega_x P\|^2. \qquad (9)$$

△

Notice that the cost functional in (9) involves only two unknown functions, namely $W$ and $F$. This makes it simpler to work with than the functional in (7).

The optimization problem can be further simplified by writing the optimal $W$ in terms of $|1-F|$. Unfortunately, the relationship between $F$ and the optimal $W$, for the general case, can only be stated implicitly, as shown next.

*Lemma 2: For a given frequency response $F$, the optimal $W$ satisfies*

$$W = \max\left\{0\ ,\ 1 - \frac{\langle |1-F|\,, |\Omega_x P|\,|W|\rangle}{\gamma+1-\|1-F\|^2}\cdot\frac{|1-F|}{|\Omega_x P|}\right\} \qquad (10)$$

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

*Remark 1: Notice that, from* (10), *$W$ is a positive, symmetric and real valued function of $\omega$. It then follows from* (4) *that the product of the optimal filters $A(z)$, $B(z)$ must exhibit linearly decreasing phase.*

In general, the presence of $|W|$ in the inner product on the right hand side of (10) makes it difficult, if not impossible, to express the optimal $W$ explicitly in terms of $F$. However, under specific conditions on $\gamma$, $\Omega_x P$ and $|1-F|$, an analytical explicit solution to (10) can be obtained, as follows:

*Lemma 3: Provided*

$$\gamma + 1 > \langle |1-F|\,, \Omega_x P\rangle \frac{|1-F|}{|\Omega_x P|}, \quad \textit{a.e. on } [-\pi,\pi], \qquad (11)$$

*then, for a given frequency response $F$, the optimal $W$ satisfies*

$$W = 1 - \frac{\langle |1-F|\,, |\Omega_x P|\rangle}{\gamma+1}\frac{|1-F|}{|\Omega_x P|}, \quad \textit{a.e. on } [-\pi,\pi]. \qquad (12)$$

△

To summarize our results so far, we have shown that, provided (11) holds, $F$ determines the optimal $W$ through (12). These two, in turn, determine the optimal $A$ and $B$ via (8) and (4), respectively.

We can now state the main result of this paper:

*Theorem 1: If*

$$\gamma + 1 > \frac{\eta^2_{\Omega_x P}}{|\Omega_x P|^2} \quad \textit{a.e. on } [-\pi,\pi], \qquad (13)$$

*then the minimum achievable frequency weighted reconstruction MSE of an NS-DPCM converter is*

$$\breve{\sigma}_\epsilon^2 \triangleq \min \sigma_\epsilon^2 = \frac{\eta^2_{\Omega_x P}}{\gamma+1}. \qquad (14)$$

*This minimum is achieved when the filters $F$, $A$ and $B$ satisfy:*

$$|1-F| = \frac{\eta_{xP}}{|\Omega_x P|}, \qquad |A| = \frac{\kappa}{|\Omega_x|}\sqrt{1 - \frac{\breve{\sigma}_\epsilon^2}{|\Omega_x P|^2}}\,, \qquad (15a)$$

$$W = 1 - \frac{\breve{\sigma}_\epsilon^2}{|\Omega_x P|^2}, \quad |B| = \frac{|\Omega_x|}{\kappa}\sqrt{1 - \frac{\breve{\sigma}_\epsilon^2}{|\Omega_x P|^2}}\,, \qquad (15b)$$

*a.e. on $[-\pi,\pi]$, where $\kappa > 0$ is an arbitrary constant.* △

## V. Discussion

The results stated in Theorem 1 have very interesting consequences. Some of these consequences are discussed below.

*a) Optimality of Scalar Quantization Without Feedback:* It is easy to verify from the results in Theorem 1 that scalar quantization without feedback is optimal if and only if $|\Omega_x P|$ is constant. In particular, it follows from (15) that if $|\Omega_x P| = 1$, a.e. on $[-\pi,\pi]$, then the optimal NS-DPCM converter reduces to a PCM converter with a fully whitening pre-filter and a post-filter satisfying $|A| = \kappa\,|\Omega_x|^{-1}$ and $|B| = |A|^{-1}\,\gamma/(\gamma+1)$.

*b) Comparison with [3]:* The minimum FWMSE for an NS-DPCM system derived by Noll in [3], neglecting fed back quantization noise, is $\frac{\eta^2_{\Omega_x P}}{\gamma}$. Perhaps surprisingly, Theorem 1 shows that the optimal performance is slightly better (compare with (14)). Moreover, the corresponding optimal filters $A_N$, $B_N$ and $F_N$ derived in [3] satisfy $|A_N| \triangleq \kappa\,|\Omega_x|^{-1}$, $|1-F_N| \triangleq \eta_{\Omega_x P}\,|\Omega_x P|^{-1}$ and $B_N = A_N^{-1}$, respectively. Substituting these expressions into (9) actually yields an FWMSE $\sigma^2_{\epsilon N} \triangleq \breve{\sigma}^2_\epsilon \cdot \frac{\zeta_{(xP)^{-1}}}{\zeta_{(xP)^{-1}} - \frac{1}{\gamma+1}}$, where $\zeta_{(xP)^{-1}}$ is the spectral flatness measure of $(\Omega_x P)^{-1}$. It then follows that $\sigma^2_{\epsilon N} > \breve{\sigma}^2_\epsilon$ for *any* finite $\gamma$, and that $\sigma^2_{\epsilon N} \to \breve{\sigma}^2_\epsilon$ as $\gamma \to \infty$.

*c) Total Frequency Weighted Distortion is White:* It follows from (14) and (15) that, in an optimized NS-DPCM system, the PSDs of frequency weighted quantization noise and linear distortion are, respectively

$$\begin{aligned} S_{n'} &\triangleq \sigma_n^2\,|1-F|^2\,|B|^2\,P^2 = \breve{\sigma}_\epsilon^2\left[1 - \breve{\sigma}_\epsilon^2/\,|\Omega_x P|^2\right] \\ S_L &\triangleq (W-1)^2\,|\Omega_x P|^2 = (\breve{\sigma}_\epsilon^2)^2/\,|\Omega_x P|^2\,. \end{aligned}$$

From the above, one can see that when the condition of Theorem 1 holds, the noise shaping effected by an optimal NS-DPCM system is not "complete", i.e., frequency weighted quantization noise is not white. However, the PSD of *the total frequency weighted error is white*, since $S_\epsilon = S_{n'} + S_L = \breve{\sigma}_\epsilon^2$.

*d) Relation with the Reverse Water-Filling Paradigm:* The parametric Rate-Distortion formula for a Gaussian w.s.s. process and FWMSE as the distortion measure is given by the well known reverse water-filling paradigm (see, e.g., [17]). For $\sigma_\epsilon^2 \leq \min_\omega (S_x \left|P\right|^2)$, it predicts total frequency weighted distortion to be equally distributed over frequency. It also predicts the input signal to appear at the output with PSD $S_x \left|P\right|^2 - \sigma_\epsilon^2$, i.e., less significant spectral components of $x$ suffer higher attenuation. Interestingly, (13) is equivalent to $S_x \left|P\right|^2 \geq \breve{\sigma}_\epsilon^2$, a.e. on $[-\pi, \pi]$. Furthermore, $S_\epsilon$ is flat, as discussed in c) above, and (15) yields $S_x \left|W\right|^2 \left|P\right|^2 = S_x \left|P\right|^2 - \breve{\sigma}_\epsilon^2$, in full agreement with the above prediction.

*e) Output of the Scalar Quantizer is White:* It can be seen from (15a) that, unless $|\Omega_x P|$ is constant, the optimal $A$ is not a full whitening filter for $\Omega_x$. Interestingly, however, it is straightforward to verify that the optimal filters in (15) render a sequence $w$ (see Fig. 1) with flat PSD. More precisely, $S_w \triangleq \left|\Omega_x\right|^2 \left|A\right|^2 + \sigma_n^2 \left|1-F\right|^2 = \kappa^2$, where $\kappa$ is the same arbitrary constant that appears in (15a). A remarkable implication is that the quantized output of the optimized NS-DPCM converter can be efficiently translated into bits by means of a first-order entropy coder.

*f) Rate-Distortion Analysis:* The rate-distortion efficiency of any source encoding scheme (with quadratic error as distortion measure) can be established by comparing its

$$SNR \triangleq \frac{\sigma_x^2}{\sigma_\epsilon^2}$$

against the upper bound derived by O'Neal in [18]. For the case $\sigma_\epsilon^2 \leq \min_\omega S_x(\mathrm{e}^{j\omega})P(\mathrm{e}^{j\omega})$, and restricting to Gaussian inputs, this bound [18, eq. (6)] can be written as[7]

$$SNR_{maxdB} \triangleq 6R - 10\log\zeta_x^2 - 10\log\eta_P^2, \qquad (16)$$

where $R$ denotes the bit-rate (in bits per sample), $\zeta_x^2$ is the *spectral flatness measure* of $x$ and $\eta_P^2$ is the minimum variance associated with $P$ (see Section I). Notice that $6R$ is Shannon's upper bound [19] for the SNR (in decibels) of encoding a Gaussian memoryless source.

On the other hand, under the conditions of Theorem 1 and using (14), the best achievable $SNR$ of an NS-DPCM system is given by

$$SNR = \frac{\sigma_x^2(\gamma+1)}{\eta_{\Omega_x P}^2} = \frac{\gamma+1}{\zeta_x^2\eta_P^2}.$$

In decibels, this ratio is

$$SNR_{dB} = 10\log(\gamma+1) - 10\log\zeta_x^2 - 10\log\eta_P^2. \qquad (17)$$

By comparing[8] (17) and (16), we see that the SNR of the NS-DPCM converter optimized via Theorem 1 departs from the information-theoretic upper bound (16) as follows:

$$SNR_{dB} - SNR_{maxdB} = 10\log(\gamma+1) - 6R \approx 10\log\gamma - 6R.$$

The difference $\Delta_{SNR} \triangleq 10\log\gamma - 6R$ for of Gaussian sources has long been known for a variety of scalar (memoryless) quantizer types (see, e.g. [14] and the references therein). Assuming $v$ to be Gaussian in the optimized NS-DPCM converter, $\Delta_{SNR}$ can be approximated by $-1.5$, $-2.45$, $-4.35$ and $-7.3$ for a uniform quantizer with entropy coding (E.C.), non-uniform quantizer with E.C., non-uniform quantizer optimized for MSE without E.C., and uniform quantization without E.C. and a loading factor of 4, respectively[9] (see [14]).

## VI. Conclusions

This paper has derived explicit analytic expressions for the best achievable performance (and optimal filters) for noise-shaping DPCM encoders. These expressions, which we believe to be novel, were found by accounting for fed back quantization noise in the optimization. The results presented in this paper simplify the analysis and design of NS-DPCM converters, and provide valuable insight into the trade-offs inherent in linear feedback quantizers.

## VII. Appendix

### A. Preliminary Result

*Lemma 4: Let $g \in L^2$ be a given function such that $g(\omega) > 0$, $\forall \omega \in [-\pi, \pi]$ and $e^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(g(\omega))d\omega}$ is finite. Then*

$$\arg\min_{f \in \mathcal{B}_+} \langle\, f, g \,\rangle = e^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(g(\omega))d\omega} g^{-1},$$

*where*

$$\mathcal{B}_+ \triangleq \left\{ f : \mathbb{R} \to \mathbb{R}^+ : 0 \leq \int_{-\pi}^{\pi} \ln\left(f(\omega)\right) d\omega < \infty \right\} \qquad (18)$$

*is the set of non-negative log-integral functions.* △

*Proof:* Since $\ln(\cdot)$ is a monotonically increasing function, minimization of $\int fg$ is equivalent to minimizing $\ln\left(\int fg\right)$. From Jensen's inequality and the constraint $f \in \mathcal{B}_+$, we obtain

$$\ln\left(\int fg\right) \overset{(a)}{\geq} \int \ln fg = \int \ln f + \int \ln g \overset{(b)}{\geq} \int \ln g.$$

Equality is obtained in $(a)$ if and only if $f = \eta g^{-1}$, a.e. on $[-\pi, \pi]$, for some $\eta > 0$. Inequality $(b)$ becomes equality if and only if $\eta = \mathrm{e}^{\frac{1}{2\pi}\int_{-\pi}^{\pi}\ln(g(\omega))d\omega} g^{-1}$. This completes the proof. ■

### B. Proofs of Lemmas 1-3 and Theorem 1

*Proof:* [Lemma 1] The numerator of the first term on the right side of (7), denoted here by $N$, is given by

$$N \triangleq \|A\Omega_x\|^2 \|(1-F)BP\|^2. \qquad (19)$$

We can use the Cauchy-Schwartz inequality to obtain

$$N \geq \langle\, |A\Omega_x|\,, |(1-F)BP|\,\rangle^2 = \langle\, |\Omega_x P|\,, |1-F|\,|W|\,\rangle^2. \quad (20)$$

Substituting (20) into (7) yields (9), which is obtained with equality in (20). The latter is achieved if and only if $|A\Omega_x| = \kappa^2\left|(1-F)BP\right|$, a.e. on $[-\pi, \pi]$, for arbitrary $\kappa^2 \in \mathbb{R}^+$, or, equivalently, $\left|A\right|^2\left|\Omega_x\right| = \kappa^2\left|(1-F)P\right|\left|W\right|$, a.e. on $[-\pi, \pi]$.

[7] Hereafter, log denotes the base 10 logarithm.

[8] Notice that the observations made in d) above validate this comparison.

[9] These figures are good approximations only for many quantization levels. Better estimates for $\Delta_{SNR}$ with few quantization levels can obtained from [20].

This last equation leads directly to (8), completing the proof. ■

*Proof:* [Lemma 2] Expanding the squared norm of the last term in (3) we obtain

$$\sigma_\epsilon^2 = \sigma_n^2 \|(1-F)PB\|^2 + \|\Omega_x PAB\|^2 - 2Re\left\{\langle \Omega_x^2 P^2, W\rangle\right\} + \|\Omega_x P\|^2. \quad (21)$$

Substituting (4) into the above equation yields

$$\sigma_\epsilon^2 = \|TPB\|^2 - 2Re\left\{\langle \Omega_x^2 P^2 \mathrm{e}^{-j\omega\tau} A^*, B\rangle\right\} + \|\Omega_x P\|^2, \quad (22)$$

where $T \triangleq (\sigma_n^2 \left|1-F\right|^2 + \Omega_x^2 \left|A\right|^2)^{1/2}$. Rearranging terms,

$$\sigma_\epsilon^2 = \left\| \left(T^2 B - \Omega_x^2\, \mathrm{e}^{-j\omega\tau} A^*\right) \frac{P}{T} \right\|^2 + \left\| \frac{\sigma_n^2 \Omega_x P \left|1-F\right|}{T} \right\|^2,$$

which is clearly minimized if and only if $B = \Omega_x^2\, \mathrm{e}^{-j\omega\tau} \frac{A^*}{T^2}$ a.e. on $[-\pi,\pi]$. Multiplying the latter by $A$ yields[10]

$$W = \frac{\left|\Omega_x\right|^2 \left|A\right|^2}{\sigma_n^2 \left|1-F\right|^2 + \left|\Omega_x\right|^2 \left|A\right|^2}, \quad \text{a.e. on } [-\pi,\pi], \quad (23)$$

where $\sigma_n^2$ is given by (5). By direct substitution of (8) into (23), and after some algebra, the optimal $W$ is found to satisfy (10). This completes the proof. ■

*Proof:* [Lemma 3] Define

$$\alpha \triangleq \frac{\langle \left|1-F\right|, \left|\Omega_x P\right| W\rangle}{\gamma + 1 - \|1-F\|^2}, \quad (24)$$

and suppose that

$$\alpha \left|1-F\right| \left|\Omega_x P\right|^{-1} \leq 1, \quad \text{a.e. on } [-\pi,\pi]. \quad (25)$$

Then, from (10),

$$W = 1 - \alpha \left|1-F\right| \left|\Omega_x P\right|^{-1}, \quad \text{a.e. on } [-\pi,\pi]. \quad (26)$$

Substituting (26) into (24) yields $\alpha = \frac{\langle |1-F|, |\Omega_x P|\rangle - \alpha \|1-F\|^2}{\gamma+1-\|1-F\|^2}$. Thus,

$$\alpha = \frac{\langle \left|1-F\right|, \left|\Omega_x P\right|\rangle}{\gamma+1}. \quad (27)$$

Substituting this into (26) yields (12). Notice that (11) guarantees that the denominator on the right hand side of (24) is strictly positive. The proof is completed by noting that substitution of (27) into (11) gives the inequality (25), thus validating our initial supposition. ■

*Proof:* [Theorem 1] Suppose the optimal $F$ is such that (11) in Lemma 3 holds. Then, one can substitute (12) into (9) to obtain $\sigma_\epsilon^2 = \frac{\langle |1-F|\,,\, |\Omega_x P| - \frac{\langle |1-F|, |\Omega_x P|\rangle}{\gamma+1} |1-F|\rangle^2}{\gamma+1-\|1-F\|^2} + \frac{\langle |1-F|, |\Omega_x P|\rangle^2}{(\gamma+1)^2} \|1-F\|^2$. After some algebra, this becomes

$$\sigma_\epsilon^2 = \frac{\langle \left|1-F\right|, \left|\Omega_x P\right|\rangle^2}{\gamma+1}. \quad (28)$$

Requiring $F(z)$ to be stable and strictly causal (from Constraint 1) is equivalent to requiring the function $\left|1-F(\mathrm{e}^{j\omega})\right|$ to belong to the set of *non-negative log-integral* functions defined in (18), see, e.g. [21, Theorem 3.4.4] and [22]. Then, it follows from Lemma 4 that the optimal $|1-F|$ is as in (15a). Substitution of the latter into (28) yields (14). It also follows from (15a) that the inequality in Lemma 3 is equivalent to the condition required by the theorem. This validates our initial supposition. Notice also that the latter inequality also guarantees that $\|1-F\|^2 < \gamma+1$, as required by Condition 2. Finally, substituting (15a) into (12), (8) and (4) yields the remaining equalities of (15). This completes the proof. ■

[10] Notice that $W$ in (23) describes the anti-causal form of the Wiener filter for a w.s.s. signal with PSD $S_x\left|A\right|^2$ corrupted by uncorrelated additive noise with PSD $\sigma_n^2\left|1-F\right|^2$. This filter is known to minimize error variance.

## References

[1] E. Kimme and F. Kuo, "Synthesis of optimal filters for a feedback quantization system." *IEEE Trans. Circuit Theory*, vol. CT-10, pp. 405–413, September 1963.

[2] 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.

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

[4] N. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, no. 5, pp. 611–633, May 1974.

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

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

[7] 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.

[8] 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.

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

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

[11] 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.

[12] D. Marco and D. L. Neuhoff, "The validity of the additive noise model for uniform scalar quantizers," *IEEE Trans. Inf. Theory*, vol. 51, no. 5, pp. 1739–1755, May 2005.

[13] U. Grenander and G. Szegö, *Toeplitz forms and their applications.* Berkeley, Calif.: University of California Press, 1958.

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

[15] R. M. Gray, "Spectral analysis of quantization noise in a single-loop Sigma–Delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, no. 6, pp. 588–599, June 1989.

[16] D. Arnstein, "Quantization error in predictive coders," *IEEE Trans. Commun.*, vol. COM-23, no. 4, pp. 423–429, April 1975.

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

[18] J. O'Neal Jr., "Bounds on subjective performance measures for source encoding systems," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 3, pp. 224–231, May 1971.

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

[20] P. Noll and R. Zelinski, "Bounds on quantizer performance in the low bit-rate region," vol. COM-26, no. 2, pp. 300–304, February 1978.

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

[22] M. R. Aaron, R. A. McDonald, and E. Protonotarios, "Entropy power loss in linear sampled data filters," *Proc. IEEE (letters)*, vol. 55, no. 6, pp. 1093–1094, June 1967.