A Bound on the MSE of Oversampled Dithered Quantization With Feedback

Milan S. Derpich

***Abstract*—We analyze the behavior of the mean squared error (MSE) achievable by oversampled, uniform scalar quantization using feedback, pre- and post-filters of unrestricted order, when encoding wide-sense stationary discrete-time random sources having (possibly) unbounded support. Our results are based upon the use of subtractively dithered uniform scalar quantizers. We consider the number of quantization levels, N, to be given and fixed, which lends itself to fixed-rate encoding, and focus on the cases in which N is insufficient to avoid overload. In order to guarantee the stability of the closed-loop, we consider the use of a clipper before the scalar quantizer. Our results are valid for zero-mean sources having independent innovations whose moments satisfy some mild requirements, which are met by infinite-support distributions such as Gaussian and Laplacian. We show that, for fixed N, the MSE can be made to decay with the oversampling ratio λ as $\mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}})$ when λ tends to infinity, where $c_0 \triangleq [0.5(N-1)]^{2/3}$. We note that the latter bound is asymptotic in λ but not in N, and that it includes clipping errors.**

***Index Terms*—Oversampling, quantization, $\Sigma\Delta$ converters.**

I. Introduction

It is well know that oversampling can reduce the magnitude of the reconstruction error that originates from quantizing the samples of an analog source, see., e.g., [1]–[3]. This reduction is exploited by *analog-to-digital converters* (ADCs) such as *sigma-delta* ($\Sigma\Delta$) modulators, which have been successfully utilized in audio and image quantization [1].

It was shown in [2] that the MSE of $\Sigma\Delta$ modulation is $\mathcal{O}(\lambda^{-2[M+1]})$, as $\lambda \to \infty$, where λ is the oversampling ratio and M denotes the order of the feedback filter (assumed fixed for all values of λ). In their analysis, the authors of [2] utilized an *additive noise model* (ANM) [4], in which quantization errors are assumed to form a *wide sense stationary* (w.s.s.) random process, white and uncorrelated with the input samples. Also using the ANM, it was recently shown in [5] that by using different filters (of unrestricted order) for each value of λ, the MSE can be made to decay as $\mathcal{O}([\gamma+1]^{-\lambda})$, where γ denotes the signal-to-noise ratio of the scalar quantizer. The analysis in [2] and [5] restrict to the cases where the effect of quantizer overload errors is negligible, which cannot be guaranteed if $\lambda \to \infty$ when the source has unbounded support unless infinitely many quantization levels are available. Indeed, an important body of literature related to oversampled quantization avoids overload errors either by careful design of the converters or by simply assuming there exist enough quantization levels to avoid overload, see e.g., [3], [6] and the references therein.

Manuscript received December 03, 2008; revised February 23, 2009. Current version published April 29, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yuriy V. Zakharov.

The author is with the Department of Electronic Engineering, Universidad Técnica Federico Santa María, Casilla 110-V, Valparaíso, Chile (e-mail: milan.derpich@usm.cl).

Digital Object Identifier 10.1109/LSP.2009.2017475

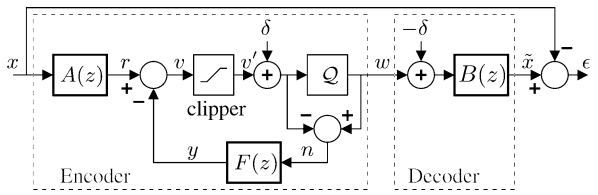

Fig. 1. Scalar feedback quantization scheme with subtractive dither.

Families of 1-bit (two-level) $\Sigma\Delta$ converters in which the quantizer is guaranteed to never overload have been found in [7], [8], by following a deterministic approach. The converters in [7] yield a continuous-time reconstruction error that can be uniformly bounded by a term proportional to $\lambda^{-c\log\lambda}$, where $c>0$ is independent of λ. In turn, the continuous-time reconstruction error with the converters constructed in [8] can be uniformly bounded as $\mathcal{O}(2^{-0.07\lambda})$ when $\lambda \to \infty$. This leads immediately to an MSE that behaves as $\mathcal{O}(2^{-0.14\lambda})$, when $\lambda \to \infty$. To the best of the author's knowledge, the latter is the fastest decay rate of the reconstruction error with λ available in the literature. [1] However, the results in [7] and [8] have not been extended to $\Sigma\Delta$ modulators with more than two quantization levels, and rely upon the input samples being uniformly bounded. On the other hand, available results on the quantization of unbounded sources including the effects of overload errors do not consider oversampling, see, e.g., [9] and the references therein.

In this letter, we study the behaviour of the MSE with increasing oversampling ratio when the source is a (possibly unbounded) *wide sense stationary* (w.s.s.) band-limited process. Our analysis is based upon the use of a *subtractively dithered uniform scalar quantizer* (SDUSQ) [10], preceeded by a clipper, together with feedback, pre- and post-filters of unrestricted order (see Fig. 1). We focus on the cases in which the number of quantization levels, N, is insufficient to avoid quantizer overload. We show that, for this architecture, the MSE can be made to decay with λ as $\mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}})$, where $c_0 \triangleq [0.5(N-1)]^{2/3}$, provided the following holds.

Assumption 1: The source process has independent innovations $\xi(k)$, with zero mean and symmetric *probability density function* (PDF). Moreover, there exists a constant $H<\infty$ such that the n-th moments of each $\xi(k)$ satisfy

$$\left|\mu_n^{(k)}\right| \le 0.5(n!)H^{n-1}\left|\mu_2^{(k)}\right|, \quad \text{for all } n>1, k\in\mathbb{Z}. \tag{1}$$

▲

This letter extends the work in [5] by taking account of clipping errors in the analysis.

[1] Krahmer, Güntürk and Deift have recently obtained a faster decay rate, but this result does not seem to be available at the time of this publication.

II. Preliminaries and Problem Statement

Our results are related with the feedback quantization architecture shown in Fig. 1. In this scheme, the samples $\{x(k)\}(k \in \mathbb{Z})$ form a zero-mean w.s.s. process, obtained from sampling a w.s.s. band-limited analog signal. For each oversampling ratio $\lambda \geq 1$, the *power spectral density* (PSD) of $\{\mathrm{x}(k)\}$ can be written as $S_{\mathrm{x}}(\mathrm{e}^{jw}) = g_\lambda(\omega)^2$, where

$$g_\lambda(\omega) \triangleq \begin{cases} \sqrt{\lambda} g_1(\lambda\omega), & \text{if } |\omega| < \omega_c, \\ 0, & \text{if } \omega_c \leq |\omega| \leq \pi. \end{cases} \quad (2)$$

In (2), $\omega_c \triangleq (\pi)/(\lambda)$, and g_1 is the square root of the PSD of the input process when $\lambda = 1$. It is assumed that the input process has finite power, i.e., that $(1)/(2\pi)\int_{-\pi}^{\pi} g_1(\omega)^2 dw < \infty$. For simplicity, we shall further restrict the analysis to the cases in which $g_1(\omega) > 0, \forall \omega \in [-\pi, \pi]$. Notice also from (2) that the total power of g_λ (in units of variance per sample), remains constant for all $\lambda \geq 1$.

In Fig. 1, $\mathcal{Q}$ represents a uniform scalar quantizer, with quantization interval Δ and N reconstruction levels. The dither $\{\delta(k)\}$ is a random process with i.i.d. samples independent of $\{\mathrm{x}(k)\}$ and uniformly distributed over the interval $[-\Delta/2, \Delta/2]$. Adding dither to the input of the quantizer reduces the range for the input signal over which quantizer overload cannot occur. Since the dither is distributed over $[-\Delta/2, \Delta/2]$, this range is $(N-1)\Delta$. It is well known that such a dither signal yields a quantization error process $\{\mathrm{n}(k)\}$ with i.i.d. samples which are also independent of the source [10], [11], provided $|v'(k)| < (N-1)\Delta/2, \forall k \in \mathbb{Z}$, i.e., as long as $\mathcal{Q}$ does not overload. Quantization error samples appear in the output as the stationary process $\epsilon_n(k) \triangleq [1 - F(z)]B(z)n(k)$. In order to keep $\mathcal{Q}$ from overloading, we consider the use of a clipper before $\mathcal{Q}$, as shown in Fig. 1. The clipper limits the value of the input signal v' so that $v'(k) = v(k)$, if $|v(k)| \leq (N-1)\Delta/2$, or $v'(k) = (v(k)/|v(k)|)(N-1)\Delta/2$, if $|v(k)| > (N-1)\Delta/2$, thus ensuring stability, see [5]. The key point here is that, unlike overload errors, clipping errors, given by $\vartheta(k) \triangleq v'(k) - v(k)$, are not injected into the feedback loop. Instead, clipping errors appear in the output after being filtered by $B(z)$, to yield the process $\epsilon_\vartheta(k) \triangleq B(z)\vartheta(k), \forall k \in \mathbb{Z}$. Unless the source $\{\mathrm{x}(k)\}$ is a stationary process, one cannot guarantee that the samples of the clipping error will form a stationary, or even a w.s.s., random process. In order to quantify the contribution of clipping errors to the MSE for not-necessarily stationary sources, we define the *average power of clipping errors in the output* as

$$\sigma^2_{\epsilon_\vartheta} \triangleq \lim_{\ell\to\infty} \frac{1}{2\ell+1} \sum_{k=-\ell}^{\ell} \mathrm{E}[\epsilon_\vartheta(k)^2] \quad (3)$$

where $\mathrm{E}[\cdot]$ denotes expectation. Similarly, we define the *average power of the reconstruction error* as

$$\sigma^2_\epsilon \triangleq \lim_{\ell\to\infty} \frac{1}{2\ell+1} \sum_{k=-\ell}^{\ell} \mathrm{E}[(\epsilon_n(k) + \epsilon_\vartheta(k))^2]. \quad (4)$$

Two important parameters characterizing the conditions under which the combination of clipper and quantizer operate are the *signal-to-noise ratio* (SNR)

$$\gamma \triangleq \sigma^2_{\mathrm{v}}/\sigma^2_{\mathrm{n}} \quad (5)$$

and the *loading factor*

$$\rho \triangleq [N-1]\Delta/(2\sigma_{\mathrm{v}}). \quad (6)$$

It follows from (5) and (6) that, if N and Δ are kept fixed, then ρ can only be increased at the expense of reducing the SNR at which the clipper and $\mathcal{Q}$ operate.

For the scheme of Fig. 1, it was shown in [5] that the reconstruction MSE *due to granular quantization errors only*, which here corresponds to $\sigma^2_{\epsilon_n}$, can be upper bounded as

$$\sigma^2_{\epsilon_n} \leq \frac{\alpha_o(K,1)}{4\gamma} K^{2-\lambda}, \quad \forall \gamma > 0, \forall \lambda \geq 1 \quad (7)$$

where $K \triangleq \gamma + 1$ and the scalar function $\alpha_o(K,1) > 0$ yields the unique value of α that satisfies

$$\frac{1}{2}\ln(\gamma+1) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left(\sqrt{\frac{g_\lambda(\omega)^2}{\alpha}+1} + \frac{g_\lambda(\omega)}{\sqrt{\alpha}}\right) d\omega \quad (8)$$

when $\lambda = 1$. Upon defining $f(\omega) \triangleq |1 - F(\mathrm{e}^{jw})|, \forall \omega \in [-\pi, \pi]$, it is also shown in [5] that, for (7) to hold, the frequency responses of $A(z)$, $B(z)$ and $F(z)$ must satisfy

$$f(\omega) = \sqrt{K\alpha}/[\sqrt{g_\lambda(\omega)^2 + \alpha} + g_\lambda(\omega)], \quad (9)$$

$$|B(\mathrm{e}^{jw})|^2 = g(\omega)[\sqrt{g_\lambda(\omega)^2+\alpha} + g_\lambda(\omega)]/(\kappa^2\sqrt{K\alpha}) \quad (10)$$

$$|A(\mathrm{e}^{jw})|^2 = \kappa^2\sqrt{K\alpha}/[g_\lambda(\omega)(\sqrt{g_\lambda(\omega)^2+\alpha} + g_\lambda(\omega))] \quad (11)$$

i.e., on $[-\pi, \pi]$, for every $\lambda \geq 1$, where α satisfies (8). In (10) and (11), κ can be any bounded, nonzero gain. With the optimal filters in (10)–(11), α relates to γ, κ and the variance σ^2_{n} via [5]

$$\sigma^2_{\mathrm{n}} = (\kappa^2/2)\sqrt{\alpha/K}. \quad (12)$$

Notice the upper bound on the MSE due to granular quantization errors in (7) decays exponentially with λ. However, the behaviour of the average power of clipping errors with increasing λ is unknown. Therefore, in view of (4), the exponential decay of $\sigma^2_{\epsilon_n}$ given by (7) does not necessarily hold for σ^2_ϵ, for λ sufficiently large. In the next section we find an upper bound to the total average power of the reconstruction error σ^2_ϵ, including clipping errors.

III. Main Result

We start with the following technical lemma:

Lemma 1: Let $\varsigma_1, \varsigma_2, \ldots$ be independent random variables with moments $\mu^i_n \triangleq E[\varsigma_i^n]$, and let $\sigma^2 \triangleq \sum_i \mu^i_2 < \infty$. If there exists a constant H such that (1) is satisfied, then

$$\Pr\Big\{\sum_i \varsigma_i > u\sigma\Big\} \leq \mathrm{e}^{-(\sigma)/(2H)u}, \quad \forall u \geq \sigma/(2H) \quad (13)$$

where $\Pr\{\cdot\}$ denotes probability.

Proof: From one of Bernstein's inequalities, given in [12, Sec. 5.5], we have that

$$\Pr\Big\{\sum_i \varsigma_i > u\,\sigma\Big\} \leq \mathrm{e}^{-2(1-c)u^2} \quad (14)$$

$\forall u > 0$ and $\forall c \in (0,1)$ such that $u \leq (c\sigma)/(2(1-c)H)$. For every $u > 0$, the tightest bound for the first inequality in (14) is obtained with $c = u/(\sigma/(2H) + u)$. Substituting this into (14) yields $\Pr\{\sum_i \varsigma_i > u\sigma\} \leq \mathrm{e}^{-2u^2/(1+2H/(\sigma u))}$. The latter, together with the fact that $2u/(1+[2H/\sigma]u) \geq \sigma/(2H)$, $\forall u \geq \sigma/(2H)$ leads directly to (13), completing the proof. ■

The following theorem provides an upper bound for $\sigma^2_{\varepsilon_\vartheta}$ applicable (but not restricted) to situations in which the source has unbounded support.

Theorem 1: Suppose there exists a scalar $\hat{g} < \infty$ such that $g_1(\omega) \leq \hat{g}, \forall \omega \in [-\pi, \pi]$ [see (2)]. Suppose that Assumption

1 holds, and that the innovations of $\{\mathrm{x}(k)\}$ satisfy (1) with $H = H_\xi$, for some constant H_ξ. Then

$$\sigma^2_{\varepsilon_\vartheta} \leq 16\frac{\hat{g}^2}{\nu^2}\lambda \mathrm{e}^{-\nu\rho}, \quad \forall \lambda \geq 1 \tag{15}$$

where ρ is the loading factor defined in (6), and where

$$\nu \triangleq \frac{1}{2}\min\left\{\left(\frac{\gamma\lambda}{K}\right)^{1/2}\frac{\sigma_\xi}{H_\xi}, \frac{\sigma_\mathrm{n}}{H_n}\right\}. \tag{16}$$

Proof: We have from (3) that

$$\sigma^2_{\epsilon_\vartheta} \leq B^2_{\max}\sigma^2_\vartheta \tag{17}$$

where $B^2_{\max} \triangleq \max_{\omega\in[-\pi,\pi]}|B(\mathrm{e}^{jw})|^2$, and where $\sigma^2_\vartheta \triangleq \lim_{\ell\to\infty} 1/(2\ell+1)\sum_{k=-\ell}^{\ell}\mathrm{E}[\vartheta(k)^2]$. We will first upper bound $B_{\max}$ and then σ^2_ϑ.

a) Bounding $B^2_{\max}$: From (9) and (10), we have

$$|B(\mathrm{e}^{jw})|^2 = g_\lambda(\omega)/[\kappa^2 f(\omega)], \quad \forall\omega \in [-\pi,\pi]. \tag{18}$$

From (12), $\kappa^2 = 2\sigma^2_\mathrm{n}[K/\alpha_o(K,\lambda)]^{1/2}$. Substitution of this into (9) yields

$$\begin{aligned}\kappa^2 f(\omega) &= 2\sigma^2_\mathrm{n}\frac{K}{\alpha_o(K,\lambda)}[\sqrt{g_\lambda(\omega)^2+\alpha_o(K,\lambda)} - g_\lambda(\omega)]\\ &= 2\sigma^2_\mathrm{n}K/[\sqrt{g_\lambda(\omega)^2+\alpha_o(K,\lambda)} + g_\lambda(\omega)].\end{aligned}$$

Noting that $\alpha_o(K,\lambda) = \lambda\alpha_o(K^\lambda,1)$ (see the Proof of Theorem 5 in [5]), we obtain

$$\kappa^2 f(\omega) = \frac{2\sigma^2_\mathrm{n}K/\sqrt{\lambda}}{\sqrt{g_1(\lambda\omega)^2+\alpha_o(K^\lambda,1)}+g_1(\lambda\omega)}.$$

Substitution of this last equation and (2) into (18) yields

$$|B(\mathrm{e}^{jw})|^2 = \frac{\lambda[\sqrt{g_1(\lambda\omega)^2+\alpha_o(K^\lambda,1)}+g_1(\lambda\omega)]g_1(\lambda\omega)}{2\sigma^2_\mathrm{n}K} \tag{19}$$

$\forall\omega \in [-\pi,\pi]$. The fact that $\alpha_o(K^\lambda,1)$ decreases monotonically with increasing λ, together with (19), leads to

$$|B(\mathrm{e}^{jw})|^2 \leq \frac{\lambda}{2\sigma^2_\mathrm{n}K}(\sqrt{\hat{g}^2+\alpha_o(K,1)}+\hat{g})\hat{g} \tag{20}$$

$\forall\omega \in [-\pi,\pi]$. In order to get rid of $\alpha(K,1)$ in the right hand side of (20), we will use an upper bound for $\alpha_o(K,1)$ instead of the latter. More precisely, since $K = \gamma+1$, it follows directly from (8) that

$$\alpha_o(K,1) \leq 4\hat{g}^2(\gamma+1)/\gamma^2 \tag{21}$$

and thus $(\sqrt{\hat{g}^2+\alpha(K,1)}+\hat{g})\hat{g} \leq (\sqrt{1+4(\gamma+1)/\gamma^2}+1)\hat{g}^2 = 2\hat{g}^2(\gamma+1)/\gamma$. Substitution of the latter into (20) yields

$$|B(\mathrm{e}^{jw})|^2 \leq B^2_{\max} \leq \hat{g}^2\lambda/\left(\sigma^2_\mathrm{n}\gamma\right), \quad \forall\omega \in [-\pi,\pi]. \tag{22}$$

b) Bounding σ^2_ϑ: For every $k\in\mathbb{Z}$, $v(k)$ is a linear combination of the i.i.d. random variables $\{\mathrm{n}(i)\}_{i=-\infty}^{k-1}$, and the independent random variables $\{\xi(i)\}_{i=-\infty}^{k}$. Notice also that, due to the use of subtractive dither, the random variables $\mathrm{n}(i)$ and $\xi(k)$ are independent for all $i,k\in\mathbb{Z}$. More explicitly, at any instant k, we can write

$$\mathrm{r}(k) = \sum_{i=0}^{\infty} c_\xi(i)\xi(k-i) \tag{23a}$$

$$\mathrm{y}(k) = \sum_{i=1}^{\infty} c_n(i)\mathrm{n}(k-i) \tag{23b}$$

$$\mathrm{v}(k) = \sum_{i=0}^{\infty} \varsigma(i) \tag{23c}$$

where the sequence

$$\varsigma(i) \triangleq \begin{cases} c_\xi\left(\frac{i}{2}\right)\xi\left(k-\frac{i}{2}\right), & i \text{ even}\\ c_n\left(\frac{i+1}{2}\right)\mathrm{n}\left(k-\frac{i+1}{2}\right), & i \text{ odd}\end{cases} \tag{24}$$

is made of independent random variables and where $\{c_\xi(i)\}_{i=0}^{\infty}$ and $\{c_n(i)\}_{i=1}^{\infty}$ are constants. We will upper bound σ^2_ϑ by applying Lemma 1 to $\sum_{i=0}^{\infty}\varsigma(i)$. To do so, we need to find a value for H, say H_ς, for which the random variables $\{\varsigma_i\}_{i=0}^{\infty}$ satisfy (1). This can be done by upper bounding the coefficients c_ξ and c_n. From (23) and Fig. 1, we have that

$$\sigma^2_\xi\sum_{i=0}^{\infty}c_\xi(i)^2 = \sigma^2_\mathrm{r} = \frac{1}{2\pi}\int_{-\pi}^{\pi}|A(\mathrm{e}^{jw})|^2 g_\lambda(\omega)^2. \tag{25}$$

From (19), and since $A(\mathrm{e}^{jw}) = B(\mathrm{e}^{jw})^{\sim 1}$, the frequency response magnitude of the pre-filter $A(z)$ can be upper bounded as $|A(\mathrm{e}^{jw})|^2 \leq \sigma^2_\mathrm{n}(\gamma+1)/(g_\lambda(\omega)^2\lambda)$, which, when substituted into the right hand side of (25), yields $\sum_{i=-\infty}^{k}c_\xi(i)^2\sigma^2_\xi \leq [\sigma^2_\mathrm{n}(\gamma+1)]/\lambda = \sigma^2_\mathrm{v}(\gamma+1)/(\gamma\lambda)$. The latter yields

$$c_\xi(i)^2 \leq (\gamma+1)\sigma^2_\mathrm{v}/\left(\sigma^2_\xi\gamma\lambda\right), \quad \forall i\in\mathbb{Z}^+_0. \tag{26}$$

Similarly, from (23), and since $\sigma^2_\mathrm{n} \leq \sigma^2_\mathrm{v}$, we have that $\sum_{i=-\infty}^{k-1}c_n(i)^2\sigma^2_\mathrm{n} \leq \sigma^2_\mathrm{v}$, which leads directly to

$$c_n(i)^2 \leq \sigma^2_\mathrm{v}/\sigma^2_\mathrm{n}, \quad \forall i\in\mathbb{Z}^+. \tag{27}$$

Since, for any random variable x and scalar c, $H_{c\mathrm{x}} = cH_\mathrm{x}$, it follows from (24), (26), and (27) that

$$H_\varsigma \leq \hat{H}_\varsigma \triangleq \max\left\{\max_i\{c_\xi(i)\}H_\xi, \max_i\{c_n(i)\}H_n\right\} \tag{28}$$

$$\leq \max\left\{\left(\frac{\gamma+1}{\gamma\lambda}\right)^{1/2}\frac{H_\xi}{\sigma_\xi}, \frac{H_n}{\sigma_\mathrm{n}}\right\}\sigma_\mathrm{v}. \tag{29}$$

Substituting H by H_ς into (13), we obtain

$$\begin{aligned}\Pr\{\mathrm{v} > u\sigma_\mathrm{v}\} &\leq \mathrm{e}^{-\frac{\sigma_\mathrm{v}}{2\hat{H}_\varsigma}u}\\ &\leq \mathrm{e}^{-\frac{\sigma_\mathrm{v}}{2\hat{H}_\varsigma}u}, \quad \forall u \geq \sigma_\mathrm{v}/(2H_\varsigma).\end{aligned} \tag{30}$$

From (30), we have that the variance of ϑ cannot be larger than that obtained if v were a random variable with cumulative PDF given by

$$F_{\mathrm{v}_{\max}}(x) \triangleq \begin{cases}\mathrm{e}^{\frac{1}{2\hat{H}_\varsigma}x}, & \text{if } x < -2\hat{H}_\varsigma\ln(2)\\ 1/2, & \text{if } |x| \leq 2\hat{H}_\varsigma\ln(2)\\ 1-\mathrm{e}^{-\frac{1}{2\hat{H}_\varsigma}x}, & \text{if } x > 2\hat{H}_\varsigma\ln(2).\end{cases} \tag{31}$$

Hence, for $\rho > (2\hat{H}_\varsigma/\sigma_\mathrm{v})\ln(2)$, the variance of overload errors can be upper bounded as

$$\sigma^2_\vartheta \leq \int_{\rho\sigma_\mathrm{v}}^{\infty}(t-\rho\sigma_\mathrm{v})^2\left[\frac{d}{dt}F_{v_{\max}}(t)\right]dt \tag{32}$$

$$= \frac{2}{2\hat{H}_\varsigma}\int_{\rho\sigma_\mathrm{v}}^{\infty}\left(t^2-2\rho\sigma_\mathrm{v}t+\rho^2\sigma^2_\mathrm{v}\right)\mathrm{e}^{-\frac{1}{2\hat{H}_\varsigma}t}dt \tag{33}$$

$$= 16\hat{H}^2_\varsigma\mathrm{e}^{-\frac{\sigma_\mathrm{v}}{2\hat{H}_\varsigma}\rho} = 16\hat{H}^2\mathrm{e}^{-\nu\rho} \tag{34}$$

since $\sigma_\mathrm{v}/(2\hat{H}_\varsigma) = (1/2)\min\{(\gamma\lambda/(\gamma+1))^{1/2}\sigma_\xi/H_\xi, \sigma_\mathrm{n}/H_n\}$
$= \nu$, see (16). Substituting (34) and (22) into (17), we obtain

$$\begin{aligned}\sigma^2_{\epsilon_\vartheta} &\leq 16\frac{\hat{g}^2}{\sigma^2_\mathrm{n}\gamma}\lambda\hat{H}^2_\varsigma\mathrm{e}^{-\nu\rho}\\ &= 16\frac{\hat{g}^2}{\nu^2}\lambda\mathrm{e}^{-\nu\rho}, \quad \forall\lambda\geq 1\end{aligned} \tag{35}$$

where (5) and (16) were used. This completes the proof. ■

Thus, we have obtained an upper bound on the MSE due to clipping errors that grows linearly with λ and decays exponentially with ρ (provided the product $\gamma\lambda$ does not tend to zero as $\lambda \to \infty$, see (16)).

Now we can upper bound the total MSE:

Theorem 2: Suppose the conditions of Theorem 1 hold. If the loading factor ρ varies with the oversampling ratio λ as

$$\rho = 4^{-1/3}\sqrt{3}(N-1)^{2/3}\lambda^{1/3} \tag{36}$$

then σ_ϵ^2, the MSE including overload errors, satisfies

$$\sigma_\epsilon^2 = \mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}}), \quad \text{as } \lambda \to \infty \tag{37}$$

where the constant $c_0 \triangleq [0.5(N-1)]^{2/3}$. ▲

Proof: We have that

$$\begin{aligned}\mathrm{E}[(\epsilon_n(k)+\epsilon_\vartheta(k))^2] &\le \mathrm{E}[2(\epsilon_n(k)^2+\epsilon_\vartheta(k)^2)] \\ &= 2(\sigma_{\epsilon_n}^2 + \mathrm{E}[\epsilon_\vartheta(k)^2]).\end{aligned} \tag{38}$$

Substitution of (3) and (38) into (4) yields

$$\sigma_\epsilon^2 \le 2\left(\sigma_{\epsilon_n}^2 + \sigma_{\epsilon_\vartheta}^2\right). \tag{39}$$

By substituting (21) into (7), the upper bound to the MSE due to granular quantization errors in (7) becomes

$$\sigma_{\epsilon_n}^2 \le \hat{g}^2([\gamma+1]/\gamma)^3\mathrm{e}^{-\ln(\gamma+1)\lambda}. \tag{40}$$

Upon substituting (35) and (15) in (39), we obtain the following upper bound:

$$\sigma_\epsilon^2 \le 2\hat{g}^2([\gamma+1]/\gamma)^3\mathrm{e}^{-\ln(\gamma+1)\lambda} + 32\frac{\hat{g}^2}{\nu^2}\lambda\mathrm{e}^{-\nu\rho}. \tag{41}$$

The above upper bound for σ_ϵ does not tend to zero with increasing λ unless one makes the loading factor ρ grow with λ fast enough. Substituting $\sigma_{\mathrm{n}}^2 = \Delta^2/12$ and (6) into (5) we obtain $\gamma = (3/\rho^2)(N-1)^2$. From the latter, we have that $\gamma = \eta/\rho^2$, where $\eta \triangleq 3(N-1)^2$. Thus, the term due to clipping errors in (41) can be reduced only at the expense of having $\mathcal{Q}$ operate at a lower SNR. This, in turn, makes the term due to granular errors decay more slowly with increasing λ.

For example, if one makes the loading factor ρ grow with λ as $\rho = \varpi\lambda^p$, where $p > 0$ and $\varpi > 0$ are constants to be chosen, then the RHS of (41) becomes

$$2\hat{g}^2\left[1+\frac{\varpi^2\lambda^{2p}}{\eta}\right]^3\mathrm{e}^{-\ln(\frac{\eta}{\varpi^2}\lambda^{-2p}+1)\lambda} + 32\frac{\hat{g}^2}{\nu^2}\lambda\mathrm{e}^{-\nu\varpi\lambda^p}. \tag{42}$$

The optimal decay rate when $\lambda \to \infty$ is achieved by choosing p and ϖ so as to make granular and clipping error terms decay at the same asymptotic rate. This is achieved if and only if p and ϖ are chosen so that

$$c \triangleq \lim_{\lambda\to\infty}\frac{\lambda\ln\left(\frac{\eta}{\varpi^2\lambda^{2p}}+1\right) - 3\ln\left(1+\frac{\varpi^2\lambda^{2p}}{\eta}\right) - \ln 2}{\nu\varpi\lambda^p - \ln(\lambda) - 2\ln(4\hat{g}/\nu) - \ln 2} \tag{43}$$

equals 1. Before evaluating the above limit, note that from (16) we obtain $\nu = \breve{\nu} \triangleq 4/\sqrt{3}$, $\forall\lambda \ge 2\sqrt{2}((\gamma+1)/\gamma)H_\xi^2/\sigma_\xi^2$, since n, being a random variable uniformly distributed over $[-\Delta/2, \Delta/2]$, has standard deviation $\sigma_{\mathrm{n}} = \Delta/(2\sqrt{3})$ and satisfies (1) with $H_n = \Delta/8$. Applying l'Hôpital's rule to (43) twice and substituting ν by $\breve{\nu}$, we obtain that $c = \lim_{\lambda\to\infty}\mathcal{N}(\lambda)/\mathcal{D}(\lambda)$, where

$$\begin{aligned}\mathcal{N}(\lambda) &\triangleq -2\eta p\lambda(\eta+\varpi^2\lambda^{2p}) - 6\eta p(\eta+\varpi^2\lambda^{2p}) \\ &\quad + 4\varpi^2\eta p^2[1-3\lambda^{-1}]\lambda^{2p+1} + 6p(\eta+\varpi^2\lambda^{2p})^2 \\ \mathcal{D}(\lambda) &\triangleq (\eta^2+\varpi^4\lambda^{4p}+2\varpi^2\eta\lambda^{2p})(\breve{\nu}\varpi p(p-1)\lambda^p+1).\end{aligned}$$

By comparing the powers of λ in $\mathcal{N}(\lambda)$ and $\mathcal{D}(\lambda)$, it is clear that c is either 0 or ∞ unless $p = 1/3$. With this choice, we get $c = \eta/(\breve{\nu}\varpi^3)$, and thus $c = 1 \iff \varpi = (\eta/\breve{\nu})^{1/3}$. Therefore, the right-hand side of (43) equals 1 iff $p = 1/3$ and $\varpi = (\eta/\breve{\nu})^{1/3}$, which yields (36). Substituting these values into (42) and (41) we obtain $\sigma_\epsilon^2 \le \mathrm{e}^{-h_1(\lambda)} + \mathrm{e}^{-h_2(\lambda)}$, where

$$\begin{aligned}h_1(\lambda) &\triangleq \ln[c_1\lambda^{-2/3}+1]\lambda \\ &\quad - 3\ln[1+c_1^{-1}\lambda^{2/3}] - \ln(2\hat{g}^2), \\ h_2(\lambda) &\triangleq c_1\lambda^{1/3} - \ln(\lambda) - \ln(32\hat{g}^2\nu^{-2})\end{aligned} \tag{44}$$

$\forall\lambda > 1$, and where $c_1 \triangleq \eta^{1/3}\breve{\nu}^{2/3} = (3/16)^{1/3}\eta^{1/3} = (3/16)^{1/3}[3(N-1)^2]^{1/3} = [(3/4)(N-1)]^{2/3}$. From (44), it is straightforward to show that

$$\lim_{\lambda\to\infty}\frac{h_1(\lambda)}{\lambda^{1/3}} = \lim_{\lambda\to\infty}\frac{h_2(\lambda)}{\lambda^{1/3}} = c_1. \tag{45}$$

From (45) we have that, for any constant $c < c_1$, the following holds:

$$\lim_{\lambda\to\infty}\left(\frac{\mathrm{e}^{-h_1(\lambda)}}{\mathrm{e}^{-c\lambda^{1/3}}}\right)^{\frac{1}{c\lambda^{1/3}}} = \lim_{\lambda\to\infty}\mathrm{e}^{1-\frac{h_1(\lambda)}{c\lambda^{1/3}}} = \mathrm{e}^{1-\frac{c_1}{c}} < 1.$$

From the definition of limit, this implies that $\lim_{\lambda\to\infty}(\mathrm{e}^{-h_1(\lambda)})/(\mathrm{e}^{-c\lambda^{1/3}}) < 1$. A similar limit is obtained for $h_2(\lambda)$. The result then follows by noting that $c_0 < c_1$. ■

IV. Conclusion

We have studied the asymptotic behaviour of the reconstruction MSE of fixed rate dithered quantization with feedback as the oversampling ratio λ tends to infinity, for w.s.s. sources having possibly unbounded support. It was shown that with the proper choice of filters and loading factor for each λ, the MSE can decrease with λ at least as fast as $\mathcal{O}(\mathrm{e}^{-c_0\lambda^{1/3}})$, where c_0 does not depend on λ.

References

[1] , S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma Data Converters: Theory, Design and Simulation*. Piscataway, NJ: IEEE Press, 1997.

[2] 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, Jul. 1978.

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

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

[5] M. S. Derpich, E. I. Silva, D. E. Quevedo, and G. C. Goodwin, "On optimal perfect reconstruction feedback quantizers," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pt. 2, pp. 3871–3890, Aug. 2008.

[6] R. Zamir and M. Feder, "Rate-distortion performance in coding bandlimited sources by sampling and dithered quantization," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 141–154, Jan. 1995.

[7] 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. Math.*, vol. 158, no. 2, pp. 679–710, 2003.

[8] C. S. Güntürk, "One-bit Sigma-Delta quantization with exponential accuracy," *Commun. Pure Appl. Math.*, vol. 56, no. 11, pp. 1608–1630, 2003.

[9] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 957–977, Mar. 2001.

[10] J. Ziv, "On universal quantization," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 344–347, May 1985.

[11] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun.*, vol. 12, no. 12, pp. 162–165, Dec. 1964.

[12] H. J. Godwin, , S. M. G. Kendall and M. A. , Eds., *Inequalities on Distribution Functions*, 1st ed. London: Charles Griffin, 1964.