A Digital Requantizer With Shaped Requantization Noise That Remains Well Behaved After Nonlinear Distortion

Ashok Swaminathan, *Student Member, IEEE*, Andrea Panigada, *Student Member, IEEE*, Elias Masry, *Fellow, IEEE*, and Ian Galton, *Member, IEEE*

***Abstract*—A major problem in oversampling digital-to-analog converters and fractional-N frequency synthesizers, which are ubiquitous in modern communication systems, is that the noise they introduce contains spurious tones. The spurious tones are the result of digitally generated, quantized signals passing through nonlinear analog components. This paper presents a new method of digital requantization called Successive Requantization, special cases of which avoids the spurious tone generation problem. Sufficient conditions are derived that ensure certain statistical properties of the quantization noise, including the absence of spurious tones after nonlinear distortion. A practical example is presented and shown to satisfy these conditions.**

***Index Terms*—Dither techniques, nonlinearities, quantization.**

I. Introduction

OVERSAMPLING digital-to-analog converters (DACs) and fractional-N phase-locked loops (PLLs) are each enabling components in modern communication systems [1]–[3]. In both components, a *digital delta-sigma* ($\Delta\Sigma$) *modulator*, i.e., a $\Delta\Sigma$ modulator implemented with digital logic, is used to coarsely quantize a constant or slowly varying digital sequence. The quantized sequence can be viewed as the sum of the original sequence plus spectrally shaped *quantization noise* that has most of its power outside of a given low-frequency *signal band*. Ultimately, the quantized sequence is converted to an analog signal and further processed by analog circuitry including a low-pass filter to suppress quantization noise outside of the signal band.

In most communications applications, it is critical that any spurious tones in the noise introduced by DACs and fractional-N PLLs have very low power [2], [4]. In principle, dither applied to a $\Delta\Sigma$ modulator can prevent the quantization noise from containing any spurious tones whatsoever [5], [6]. Nevertheless, in practice digital $\Delta\Sigma$ modulators are major sources of spurious tones in oversampling DACs and fractional-N PLLs [7], [8]. Regardless of how dither is applied, all $\Delta\Sigma$ modulator architectures known to the authors give rise to spurious tones when their quantization noise is subjected to nonlinear distortion. This is particularly problematic in fractional-N PLLs wherein the input to the $\Delta\Sigma$ modulator usually is a constant and the output sequence from the $\Delta\Sigma$ modulator is converted to analog form and subjected to various nonlinear operations because of nonideal circuit behavior. Heretofore, the only known solution was to make the analog circuitry very linear so that the spurious tones have sufficiently low power for the given application. Unfortunately, this limits design options and results in higher analog circuit power consumption than would be required if fewer linear analog circuits could be tolerated.

This paper presents a new type of digital quantizer, referred to as a *Successive Requantizer*, that addresses this problem. The paper presents sufficient conditions on the successive requantizer's design parameters to ensure certain statistical properties of the requantization noise and the running sum of the requantization noise. These properties include the absence of spurious tones under application of nonlinear distortion. An example is presented that satisfies the conditions and is demonstrated via computer simulation. The work borrows ideas from dc-free codes [9], [10] and dynamic element matching tree structured encoders [11], [12]. In particular, the work in [12] reflects the operation of a successive requantizer in a limited context.

The paper consists of three main sections. Section II presents the principle of successive requantization, as well as an example that illustrates the appearance of spurious tones when the quantized sequence is subjected to nonlinear distortion. Section III presents the sufficient conditions mentioned above. Section IV presents an example successive requantizer that satisfies the sufficient conditions.

II. Successive Requantization

A. Spectral Properties of Interest

As outlined above, fractional-N PLLs and delta-sigma DACs ultimately generate analog waveforms. Each such waveform contains components corresponding to digitally generated quantization noise, $s[n]$, and, in the case of fractional-N PLLs, its running sum

$$t[n] = \sum_{k=0}^{n} s[k]. \tag{1}$$

Manuscript received June 26, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. David J. Miller. This work was supported by the National Science Foundation under Award 0515286 and by the University of California Communications UC Discovery Program.

A. Swaminathan was with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA. He is now with NextWave Broadband, San Diego, CA 92130 USA (e-mail: swami@ece.ucsd.edu).

A. Panigada, E. Masry, and I. Galton are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (e-mail: panigada@ece.ucsd.edu; masry@ece.ucsd.edu; galton@ece.ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2007.899385

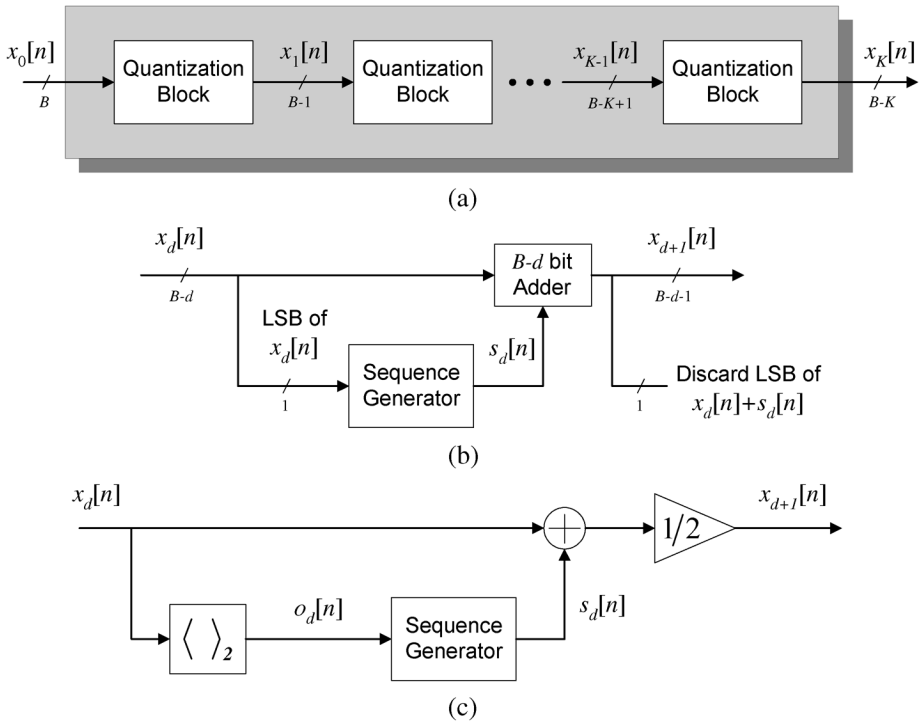


Fig. 1. (a) High-level block diagram of the successive requantizer. (b) Quantization block details. (c) Signal processing model.

Moreover, inevitable nonideal analog circuit behavior generally causes nonlinear distortion. The distortion can be any nonlinear function, but for almost all practical applications can be represented by a memoryless, truncated power series. This gives rise to components in the output waveform corresponding to $s^p[n]$ for $p = 1, 2, 3, \ldots, h_s$, and $t^p[n]$ for $p = 1, 2, 3, \ldots, h_t$, where h_s and h_t are the highest significant orders of distortion for the given application applied on $s[n]$ and $t[n]$, respectively.

Most communication system standards specify the required performance of such systems in terms of quantities that can be measured using spectrum analyzers, so the properties of the waveforms typically are quantified in the laboratory using spectrum analyzers. Although the waveforms themselves are considered to be random processes in most cases, spectrum analyzers can only average over time, not over ensemble. Therefore, in such applications the properties of the periodograms of $s^p[n]$ and $t^p[n]$ given by

$$I_{s^p,L}(\omega) = \frac{1}{L}\left|\sum_{n=0}^{L-1} s^p[n]e^{-j\omega n}\right|^2 \tag{2}$$

and

$$I_{t^p,L}(\omega) = \frac{1}{L}\left|\sum_{n=0}^{L-1} t^p[n]e^{-j\omega n}\right|^2 \tag{3}$$

are of particular interest, rather than traditional power spectral density (PSD) functions [13]. It is well known that in certain cases the expected values of the periodograms converge to the true PSD functions in the limit as $L \to \infty$, but in the applications mentioned above this is not a requirement, or even relevant to the measured performance. Hence, the results presented in this paper focus on the properties of the periodograms given by (2) and (3).

B. Signal Processing Model of the Successive Requantizer

The proposed successive requantizer architecture is shown in Fig. 1(a). Its input is a sequence of B-bit numbers, $x_0[n]$, and its output is a sequence of $B-K$-bit numbers, $x_K[n]$, where $n = 0, 1, 2, \ldots$, is the time index of the sequences. The successive requantizer consists of K *quantization blocks*, each of which quantizes its input by one bit, so the successive requantizer quantizes K bits overall.[1]

The high-level details of each quantization block are shown in Fig. 1(b) and the signal-processing model is shown in Fig. 1(c). Each quantization block generates a *quantization sequence*, $s_d[n]$, with the property that $x_d[n]+s_d[n]$ is an even number for each n, where $x_d[n]$ is the quantization block's input sequence. The quantization block adds $s_d[n]$ to $x_d[n]$ and discards the least significant bit (LSB) to implement the 1-bit quantization. Without loss of generality, numbers within the successive requantizer are taken to be integers with a two's-complement binary number representation. Since $x_d[n] + s_d[n]$ is an even number for each n, its LSB is zero, so discarding the LSB does not incur a truncation error. Hence, the quantization noise of the successive requantizer is a weighted sum of the $s_d[n]$ sequences

$$s[n] = \sum_{d=0}^{K-1} 2^d s_d[n]. \tag{4}$$

[1]Quantization blocks that quantize their input sequences by more than one bit could be used. However, it is straightforward to show that this is a trivial extension of the one-bit-per-stage case.

So far, the only restriction on the $s_d[n]$ sequences is that $x_d[n] + s_d[n]$ must be an even integer for each n and d. This leaves considerable flexibility in the design of the $s_d[n]$ sequences which is exploited in the remainder of the paper to achieve the desired quantization noise properties.

The versions of the successive requantizer considered in this paper partially exploit this flexibility to have *first-order high-pass shaped quantization noise*, i.e., they are designed such that the running sum of each $s_d[n]$ sequence

$$t_d[n] = \sum_{k=0}^{n} s_d[k] \tag{5}$$

is bounded over all n and that the estimated power spectrum of $s_d[n]$ has a high-pass spectral shape. It follows from (4) that the overall quantization noise, $s[n]$, inherits the spectral shape of the $s_d[n]$ sequences, and similarly that the running sum of the quantization noise

$$t[n] = \sum_{d=0}^{K-1} 2^d t_d[n] \tag{6}$$

is bounded.

The restriction to first-order high-pass shaped quantization noise still leaves flexibility in the design of the $s_d[n]$ sequences. This flexibility is exploited in the remainder of the paper to ensure that $s^p[n]$ for $p = 1, 2, \ldots, h_s$, and $t^p[n]$ for $p = 1, 2, \ldots, h_t$ are free of spurious tones, where h_s and h_t are positive integers. By definition, if $s^p[n]$ and $t^p[n]$ contain spurious tones at a frequency ω_n, then (2) and (3), respectively, are expected to be unbounded in probability at $\omega = \omega_n$ as $L \to \infty$. Therefore, to establish that there are no spurious tones in either $s^p[n]$ or $t^p[n]$, it is sufficient to show that (2) and (3) are bounded in probability for all $|\omega| \leq \pi$ as $L \to \infty$. A spurious tone at $\omega = 0$ is just a constant offset. Many practical systems are able to tolerate, or compensate for this offset so this case is excluded from consideration. Theorems 1 and 2 in the next section present sufficient conditions on the $s_d[n]$ sequences for (2) and (3) to be bounded in probability for every $L \geq 1$ and $0 < |\omega| \leq \pi$, thereby ensuring the absence of spurious tones in $s^p[n]$ and $t^p[n]$.

C. Example Successive Requantizer, Appearance of Spurious Tones, and Comparison to Prior Art

As shown in [14], first-order high-pass quantization noise is achieved with quantization blocks that implement

$$s_d[n] = \begin{cases} 0, & x_d[n] = \text{even} \\ r_d[n], & x_d[n] = \text{odd}, \quad t_d[n-1] = 0 \\ 1, & x_d[n] = \text{odd}, \quad t_d[n-1] = -1 \\ -1, & x_d[n] = \text{odd}, \quad t_d[n-1] = 1 \end{cases} \tag{7}$$

where $r_d[n]$ is an independent random sequence that takes on the values 1 and -1 with equal probability. The results presented in [15] imply that neither $s_d[n]$ nor $t_d[n]$ contain spurious tones. Therefore, $s[n]$ and $t[n]$ inherit these properties provided the $r_d[n]$ sequences for $d = 0, \ldots, K-1$ are independent. This is demonstrated by the estimated power spectra shown in Fig. 2

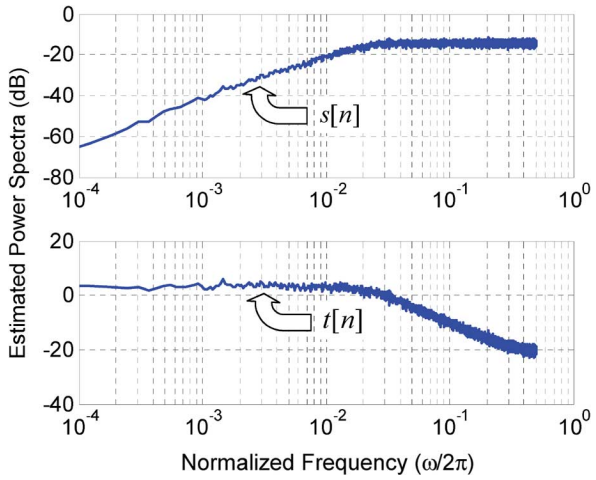


Fig. 2. Estimated power spectra of the quantization noise and its running sum for the successive requantizer presented in Section II.

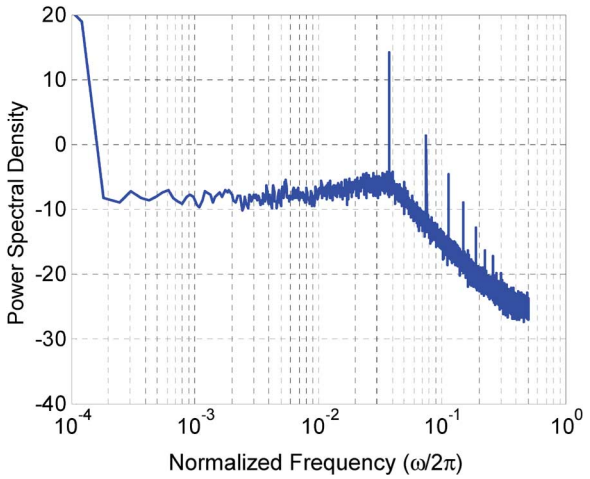


Fig. 3. Estimated power spectra of the square of the running sum of the quantization noise for the successive requantizer presented in Section II.

which correspond to a simulated successive requantizer with $K = 16$, $x_0[n] = 2457$, and quantization blocks that implement (7).

However, if the quantization noise or its running sum is subjected to nonlinear distortion, spurious tones can be induced. For instance, Fig. 3 shows the estimated power spectrum of $t^2[n]$ for the simulation example described above. Discrete spikes are evident in the plot, and it can be shown that the spikes grow without bound in proportion to the periodogram length. Therefore, the spikes represent spurious tones. The presence of spurious tones implies that subjecting $t[n]$ to second-order distortion is sufficient to induce spurious tones even though $t[n]$ is known to be free of spurious tones.

The spur generation mechanism can be understood by considering the first quantization block. Suppose the input to the successive requantizer is an odd-valued constant and $t_0[n-1] = 0$ for some value of n. Then (7) implies that $(s_0[n], s_0[n+1])$ is either $(-1, 1)$ or $(1, -1)$ depending on the polarity of $r_0[n]$. It follows from (5) that $(t_0[n], t_0[n+1])$ is either $(-1, 0)$ or $(1, 0)$, and, by induction, $t_0[n]$ has the

form $\{\ldots, 0, \pm 1, 0, \pm 1, 0, \pm 1, 0, \ldots\}$. Therefore, $t_0^2[n]$ has the form $\{\ldots, 0, 1, 0, 1, 0, 1, 0, \ldots\}$ which is periodic. A similar, but more involved analysis can be used to show that the $t_d^2[n]$ sequences for $d > 0$ also contain periodic components. These periodic components cause the spurious tones visible in Fig. 3.

Other methods of implementing noise-shaped quantizers are presented in [16]–[19]. However, these methods specifically focus on stabilizing noise-shaped coders and do not address the effect of nonlinearities on the quantization error. Successive requantization distinguishes itself from these methods in that it eliminates spurious tones that arise after subjecting the quantization error to nonlinear distortion.

III. Theory for Tone-Free Quantization Sequences

It is assumed throughout the remainder of the paper that the input to the quantizer, $x_0[n]$, is an integer-valued and deterministic sequence for $n = 0, 1, \ldots$, and that the successive requantizer is designed such that the following properties are satisfied:

Property 1: $x_{d+1}[n] = (s_d[n] + x_d[n])/2$ is integer-valued for $n = 0, 1, \ldots$, and $d = 0, 1, \ldots, K-1$.

Property 2: there exists a positive constant B such that $|t_d[n]| < B$, for $n = 0, 1, 2, \ldots$.

Property 3: $t_d[0] = 0$, and

$$t_d[n] = f\left(t_d[n-1], r_d[n], o_d[n]\right) \tag{8}$$

where $\{r_d[n],\ d = 0, 1, \ldots, K-1,\ n = 1, 2, \ldots\}$ is a set of independent identically distributed (iid) random variables, and

$$o_d[n] = x_d[n] \bmod 2 = \begin{cases} 1, & \text{if } x_d[n] \text{ is odd} \\ 0, & \text{if } x_d[n] \text{ is even} \end{cases} \tag{9}$$

is called the *parity sequence* of the dth quantization block. Furthermore, $f(x, y, z)$ is a deterministic function which does not depend on n.

Property 2 guarantees first-order spectral shaping of the quantization error by ensuring that $t_d[n]$ takes on a finite number of values for all n. However, it need not be an optimal bound on the quantization error of the successive requantizer. Much of the literature concerning noise-shaped coders is focused on minimizing some error function, which typically results in a minimization of the quantization error [17]. This paper posits that by relaxing this bound, and hence incurring more quantization error power, useful properties can be obtained such as the removal of spurious tones under nonlinearities.

Property 1 and the assumption that $x_0[n]$ is integer-valued imply that $s_d[n]$ is an even integer when $x_d[n]$ is even, and an odd integer otherwise. Therefore, (5) implies that $t_d[n]$ is integer-valued, and Property 2 further implies that it is restricted to a finite set of values. Let $T_1, T_2, \ldots, T_N$ denote these values. Therefore, the function f in Property 3 takes on values restricted to the set $\{T_1, T_2, \ldots, T_N\}$.

It follows from Properties 1, 2, and 3 that $x_{d+1}[n]$, $s_d[n]$, and $t_d[n]$, for $d = 0, 1, \ldots, K-1$, and $n = 1, 2, \ldots$, depend only on the set of iid random variables $\{r_d[n],\ d = 0, 1, \ldots, K-1,\ n = 0, 1, 2, \ldots\}$ and the deterministic successive requantizer input sequence, $\{x_0[n],\ n = 1, 2, \ldots,\}$. Therefore, the sample description space of the underlying probability space is the set of all possible values of the random variables $\{r_d[n],\ d = 0, 1, \ldots, K-1,$ and $n = 0, 1, 2, \ldots\}$.

Equation (5) implies that

$$s_d[n] = t_d[n] - t_d[n-1]. \tag{10}$$

Therefore, it follows from Property 1 that

$$x_d[n] = \left(t_{d-1}[n] - t_{d-1}[n-1] + x_{d-1}[n]\right)/2 \tag{11}$$

for $1 \le d < K$. Recursively substituting (11) into itself and applying (9) yields

$$o_d[n] = \frac{1}{2}\left[x_0[n] + \sum_{k=0}^{d-1} 2^{-k}\left(t_k[n] - t_k[n-1]\right)\right] \bmod 2. \tag{12}$$

Recursively substituting (8) into itself implies that for any integer $n > 0$,

$$t_d[n] = g_n(r_d[n], r_d[n-1], \ldots, r_d[1], o_d[n], o_d[n-1], \ldots, o_d[1]) \tag{13}$$

where g_n is a deterministic, memoryless function. Similarly, for any pair of integers $n_2 > n_1 > 0$, recursively substituting (8) into itself $m = n_2 - n_1 - 1$ times implies that

$$t_d[n_2] = h_m\left(t_d[n_1], r_d[n_1+1], r_d[n_1+2], \ldots, r_d[n_2], o_d[n_1+1], o_d[n_1+2], \ldots, o_d[n_2]\right) \tag{14}$$

where h_m is a deterministic, memoryless function.

Repeatedly substituting (12) into (13) to eliminate the variables $\{o_d[n], \ldots, o_d[1]\}$ and then recursively substituting the result into itself to eliminate the variables $\{t_k[m],\ k = 0, \ldots, d-1,\ m = 1, \ldots, n\}$ shows that $t_d[n]$ is a random variable that depends only on $x_0[n]$ (which is deterministic), and the random variables $\{r_k[m],\ k = 0, 1, \ldots, d,\ m = 1, 2, \ldots, n\}$. This in conjunction with (12) implies that $o_d[n]$ is a random variable that depends only on $x_0[n]$, and the random variables $\{r_k[m],\ k = 0, 1, \ldots, d-1,\ m = 1, 2, \ldots, n\}$. In particular since the random sequence $\{o_d[n],\ n = 0, 1, 2, \ldots\}$ does not depend on the random sequence $\{r_d[n],\ n = 0, 1, 2, \ldots\}$ and since all the random variables $\{r_k[m],\ k = 0, 1, \ldots, d-1,\ m = 0, 1, 2, \ldots, n\}$ are statistically independent by Property 3, it follows that $\{o_d[n],\ n = 0, 1, 2, \ldots\}$ and $\{r_d[n],\ n = 0, 1, 2, \ldots\}$ are statistically independent random sequences. By similar reasoning, the random variable $t_d[n]$ is statistically independent of the random variables $\{r_d[m],\ m = n+1, n+2, \ldots\}$.

Hence, (14) implies that $t_d[n_2]$ conditioned on the random variables $t_d[n_1], o_d[n_1+1], o_d[n_1+2], \ldots, o_d[n_2]$ is a function only of the statistically independent random variables $r_d[n_1], r_d[n_1+1], \ldots, r_d[n_2]$. By definition, for $i \ne j$ the random variables $\{r_i[n_1], r_i[n_1+1], \ldots, r_i[n_2]\}$ are statistically independent of the random variables $\{r_j[n_1], r_j[n_1+1], \ldots, r_j[n_2]\}$. Therefore, for $i \ne j$ the random variables $t_i[n_2]$ and $t_j[n_2]$ conditioned on $t_i[n_1], t_j[n_1], o_i[n_1+1], o_i[n_1+2], \ldots, o_i[n_2],\ o_j[n_1+$

$1], o_j[n_1+2], \ldots, o_j[n_2]$ are statistically independent. Consequently, for any positive real numbers $p_0, \ldots, p_{K-1}$,

$$E\left[\prod_{j=0}^{K-1} t_j^{p_j}[n_2] | t_d[n_1], o_d[n]; \; d=0,\ldots,K-1,\; n=n_1+1,\ldots,n_2\right]$$
$$= \prod_{j=0}^{K-1} E\left[t_j^{p_j}[n_2] | t_d[n_1], o_d[n]; \; d=0,\ldots,K-1,\; n=n_1+1,\ldots,n_2\right]$$
$$= \prod_{j=0}^{K-1} E\left[t_j^{p_j}[n_2] | t_j[n_1], o_j[n]; n=n_1+1,\ldots,n_2\right] \quad (15)$$

where the second equality follows from (8) and the independence of the $\{r_d[n],\ n = 1,2,\ldots,\}$ sequences for $d = 0,\ldots,K-1$. This implies that the probability mass function (pmf) of the random variable $t_i[n_2]$ conditioned on $t_i[n_1], o_i[n_1+1], o_i[n_1+2], \ldots, o_i[n_2]$ is independent of any additional conditioning by $t_j[n_1], o_j[n_1+1], o_j[n_1+2], \ldots, o_j[n_2]$ for $i \neq j$.

The statistical independence of $o_d[n]$ and $r_d[n]$ together with (8) imply that $\{t_d[n],\ n = 0,1,\ldots\}$ is a discrete-valued Markov random sequence conditioned on the sequence $\{o_d[n],\ n = 0,1,\ldots\}$. Whenever $x_d[n]$ is odd, the one-step state transition matrix for $t_d[n]$ is given by

$$\mathbf{A_o} = \left[P\left\{t_d[n] = T_j | t_d[n-1] = T_i, o_d[n] = 1\right\}\right]_{N\times N}. \quad (16)$$

Similarly, whenever $x_d[n]$ is even the one-step state transition matrix for $t_d[n]$ is given by

$$\mathbf{A_e} = \left[P\left\{t_d[n] = T_j | t_d[n-1] = T_i, o_d[n] = 0\right\}\right]_{N\times N}. \quad (17)$$

The function f in Property 3 is independent of n and d, so neither matrix is a function of n and d.

Equation (10) implies that each possible value of $s_d[n]$ is given by $T_j - T_i$ for some pair of integers i and j, $1 \leq i, j \leq N$, so

$$P\left\{s_d[n] = T_j - T_i | t_d[n-1] = T_i, o_d[n] = 1\right\}$$
$$= P\left\{t_d[n] = T_j | t_d[n-1] = T_i, o_d[n] = 1\right\}. \quad (18)$$

Given that $t_d[n]$ is restricted to N possible values, $s_d[n]$ is restricted to N' possible values where $N' \leq N^2$. With identical reasoning to that used to proceed from (11)–(15), it follows that

$$E\left[\prod_{j=0}^{K-1} s_j^{p_j}[n_2] | t_0[n_1], \ldots, t_{K-1}[n_1], o_d[n]; \; d=0,\ldots,K-1, n=n_1+1,\ldots,n_2\right]$$
$$= \prod_{j=0}^{K-1} E\left[s_j^{p_j}[n_2] | t_j[n_1], o_j[n]; n=n_1+1,\ldots,n_2\right]. \quad (19)$$

Given that $\{t_d[n],\ n = 0,1,\ldots\}$ is a discrete-valued Markov random sequence conditioned on the sequence $\{o_d[n],\ n = 0,1,\ldots\}$, the conditional pmf of $t_d[n_2]$ given $t_d[n_1]$ and $o_d[n]$ is equal to the conditional pmf of $t_d[n_2]$ given $t_d[n_1], t_d[n_1-1]$ and $o_d[n]$. Therefore, (10) implies that (19) is equivalent to

$$E\left[\prod_{j=0}^{K-1} s_j^{p_j}[n_2] | s_0[n_1], \ldots, s_{K-1}[n_1], t_0[n_1], \ldots, t_{K-1}[n_1], o_d[n]; d=0,\ldots,K-1, n=n_1+1,\ldots,n_2\right]$$
$$= \prod_{j=0}^{K-1} E\left[s_j^{p_j}[n_2] | t_j[n_1], o_j[n]; n=n_1+1,\ldots,n_2\right]. \quad (20)$$

The following definitions are used by the theorems presented below. In analogy to the matrices $\mathbf{A_o}$ and $\mathbf{A_e}$, let

$$\mathbf{S_o} = \left[P\left\{s_d[n] = S_j | t_d[n-1] = T_i, o_d[n] = 1\right\}\right]_{N\times N'} \quad (21)$$

and

$$\mathbf{S_e} = \left[P\left\{s_d[n] = S_j | t_d[n-1] = T_i, o_d[n] = 0\right\}\right]_{N\times N'} \quad (22)$$

where $\{S_i, 1 \leq i \leq N'\}$ is the set of all possible values of $s_d[n]$. Property 3 ensures that neither matrix is a function of n and d. It follows from (18) that each nonzero element of $\mathbf{S_o}$ or $\mathbf{S_e}$ is equal to an element in $\mathbf{A_o}$ or $\mathbf{A_e}$, respectively. For example, if $S_k = T_j - T_i$, then the element in the ith row and kth column of $\mathbf{S_o}$ is equal to the element in the ith row and jth column of $\mathbf{A_o}$. In this fashion, once $\mathbf{A_o}$ and $\mathbf{A_e}$ are known, $\mathbf{S_o}$ and $\mathbf{S_e}$ can be deduced.

Let

$$\mathbf{1} \triangleq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{t}^{(p)} \triangleq \begin{bmatrix} (T_1)^p \\ \vdots \\ (T_N)^p \end{bmatrix}, \quad \text{and} \quad \mathbf{s}^{(p)} \triangleq \begin{bmatrix} (S_1)^p \\ \vdots \\ (S_{N'})^p \end{bmatrix}. \quad (23)$$

Suppose a sequence of vectors, $\mathbf{b}[n] = [b_1[n], \ldots, b_N[n]]^{\mathrm{T}}$ converges to a constant vector, $b\mathbf{1}$, as $n \to \infty$. Then the convergence is said to be *exponential* if there exist constants $C \geq 0$ and $0 \leq \alpha < 1$ such that

$$|b_i[n] - b| \leq C\alpha^n \quad (24)$$

for all $1 \leq i \leq N$ and $n \geq 0$.

Theorem 1: Suppose that the state transition matrices $\mathbf{A_e}$ and $\mathbf{A_o}$ satisfy

$$\mathbf{A_e A_o} = \mathbf{A_o A_e} \quad (25)$$

and there exists an integer $h_t \geq 1$ such that for each positive integer $p \leq h_t$

$$\lim_{n\to\infty} \mathbf{A_e}^n \mathbf{t}^{(p)} = b_p \mathbf{1} \quad \text{and} \quad \lim_{n\to\infty} \mathbf{A_o}^n \mathbf{t}^{(p)} = b_p \mathbf{1} \quad (26)$$

where b_p is a constant and the convergence of both vectors is exponential. Then for every $L \geq 1$

$$E\left[I_{t^p,L}(\omega)\right] \leq C(\omega) < \infty \tag{27}$$

for each $0 < |\omega| \leq \pi$. Moreover, the bound $C(\omega)$, which is independent of L, is uniform in ω for all $0 < \varepsilon < |\omega| \leq \pi$.

By Markov's Inequality [20], this immediately leads to:

Corollary 1: Under the assumptions of Theorem 1, $I_{t^p,L}(\omega)$ is bounded in probability for all $L \geq 1$ and for each ω satisfying $0 < |\omega| \leq \pi$.

Proof of Theorem 1: The expectation of $I_{t^p,L}(\omega)$ can be expressed as

$$\begin{aligned} E\left[I_{t^p,L}(\omega)\right] &= \frac{1}{L}\sum_{n_1=0}^{L-1}\sum_{n_2=0}^{L-1} E\left[t^p[n_1]t^p[n_2]\right] e^{-j\omega(n_1-n_2)} \\ &= \frac{1}{L}\sum_{n_1=0}^{L-1} E\left[t^{2p}[n]\right] \\ &\quad + \frac{1}{L}\sum_{\substack{n_1=0\\ n_1\neq n_2}}^{L-1}\sum_{n_2=0}^{L-1} E\left[t^p[n_1]t^p[n_2]\right] e^{-j\omega(n_1-n_2)} \\ &\triangleq J_1 + J_2. \end{aligned} \tag{28}$$

The notation above means that J_1 and J_2 are defined as the first and second terms, respectively, to the left of the $\triangleq$ symbol. Property 2 states that $|t_d[n]| \leq B$, so it follows from (6) that $t[n] \leq B_1$ for some finite constant B_1. Therefore, $J_1 \leq B_1^{2p}$. The crux of the proof is showing that there exists a constant C_{t^p}, positive constants D_1, D_2, and a constant $0 < \alpha < 1$ such that for $n_1 \neq n_2$

$$\left|E\left[t^p[n_1]t^p[n_2] - C_{t^p}\right]\right| \leq D_1\alpha^{|n_2-n_1|} + D_2\alpha^{n_1}. \tag{29}$$

The proof of (29), is outlined in Lemma 1 in the Appendix. Here (29) is used to complete the proof of the theorem. From (28), J_2 can be expressed as

$$\begin{aligned} J_2 &= \frac{1}{L}\sum_{\substack{n_1=0\\ n_1\neq n_2}}^{L-1}\sum_{n_2=0}^{L-1} \left(E\left[t^p[n_1]t^p[n_2] - C_{t^p}\right]\right) e^{-j\omega(n_1-n_2)} \\ &\quad + \frac{1}{L}\sum_{\substack{n_1=0\\ n_1\neq n_2}}^{L-1}\sum_{n_2=0}^{L-1} C_{t^p} e^{-j\omega(n_1-n_2)} \\ &\triangleq J_{2,1} + J_{2,2}. \end{aligned} \tag{30}$$

From (29) it is seen that

$$\begin{aligned} |J_{2,1}| &\leq \frac{1}{L}\sum_{\substack{n_1=0\\ n_1\neq n_2}}^{L-1}\sum_{n_2=0}^{L-1} \left(D_1\alpha^{|n_1-n_2|} + D_2\alpha^{n_1}\right) \\ &\leq \frac{D_1}{L}\sum_{n_1=0}^{L-1}\sum_{n_2=0}^{L-1} \alpha^{|n_1-n_2|} + D_2\sum_{n_1=0}^{L-1}\alpha^{n_1} \\ &\leq 2(D_1+D_2)\frac{1-\alpha^L}{1-\alpha} \leq 2(D_1+D_2)\frac{1}{1-\alpha} \end{aligned} \tag{31}$$

and the bound is independent of L. Similarly, $J_{2,2}$ can be bounded by

$$\begin{aligned} |J_{2,2}| &\leq \frac{|C_{t^p}|}{L}\left|\left|\sum_{n=0}^{L-1} e^{-j\omega n}\right|^2 - L\right| \\ &\leq \frac{|C_{t^p}|}{L}\left|\left|\frac{1-e^{-j\omega L}}{1-e^{-j\omega}}\right|^2 - L\right| \\ &= \frac{|C_{t^p}|}{L}\left|\left|\frac{\sin(\omega L/2)}{\sin(\omega/2)}\right|^2 - L\right| \\ &\leq |C_{t^p}|\left(1 + \frac{1}{\sin^2(\omega/2)}\right) \end{aligned} \tag{32}$$

which is finite, independent of L, for each ω satisfying $0 < |\omega| \leq \pi$; the bound is uniform for all ω satisfying $0 < \varepsilon < |\omega| \leq \pi$ since $\sin(\omega/2) > \sin(\varepsilon/2)$. The result of the theorem then follows from (28)–(32). ■

Theorem 2: Suppose that the state transition matrices $\mathbf{A_e}$ and $\mathbf{A_o}$ satisfy

$$\mathbf{A_e}\mathbf{A_o} = \mathbf{A_o}\mathbf{A_e} \tag{33}$$

and there exists an integer $h_s \geq 1$ such that for each positive integer $p \leq h_s$, the sequence transition matrices $\mathbf{S_e}$ and $\mathbf{S_o}$ satisfy

$$\begin{aligned} \lim_{n\to\infty} \mathbf{A_e}^n\mathbf{S_e}\mathbf{s}^{(p)} &= \lim_{n\to\infty} \mathbf{A_e}^n\mathbf{S_o}\mathbf{s}^{(p)} = \lim_{n\to\infty} \mathbf{A_o}^n\mathbf{S_e}\mathbf{s}^{(p)} \\ &= \lim_{n\to\infty} \mathbf{A_o}^n\mathbf{S_o}\mathbf{s}^{(p)} = c_p\mathbf{1} \end{aligned} \tag{34}$$

where c_p is a constant and the convergence of all vectors are exponential. Then for every $L \geq 1$

$$E\left[I_{s^p,L}(\omega)\right] \leq D(\omega) < \infty \tag{35}$$

for each $0 < |\omega| \leq \pi$. Moreover, the bound $D(\omega)$, which is independent of L, is uniform in ω for all $0 < \varepsilon < |\omega| \leq \pi$.

By Markov's Inequality, this immediately leads to:

Corollary 2: Under the assumptions of Theorem 2, $I_{s^p,L}(\omega)$ is bounded in probability for all $L \geq 1$ and for each ω satisfying $0 < |\omega| \leq \pi$.

Proof of Theorem 2: The proof is identical to that of Theorem 1. Replacing $t^p[n_1]$ and $t^p[n_2]$ with $s^p[n_1]$ and $s^p[n_2]$ respectively, the crux of the proof is showing that there exists a constant C_{s^p}, positive constants E_1, E_2, and a constant $0 < \beta < 1$ such that for $n_1 \neq n_2$

$$\left|E\left[s^p[n_1]s^p[n_2] - C_{s^p}\right]\right| \leq E_1\beta^{|n_2-n_1|} + E_2\beta^{n_1}. \tag{36}$$

With (36) proven in Lemma 2 in the Appendix, the remainder of the proof follows directly from Theorem 1. ■

IV. A Successive Requantizer That Satisfies Theorems 1 and 2

A. Verification of Example Matrices

Matrices $\mathbf{A_e}$, $\mathbf{A_o}$, $\mathbf{S_e}$, and $\mathbf{S_o}$ which can be used with the successive requantizer to generate quantized sequences and satisfy

the conditions of Theorems 1 and 2 for $h_t = 3$ and $h_s = 5$ are presented in this section.

For a state $t_d[n]$ whose possible values are $\{-2,-1,0,1,2\}$, define

$$\mathbf{t}^{(p)} = \begin{bmatrix}(-2)^p & (-1)^p & 0 & 1^p & 2^p\end{bmatrix}^T \tag{37}$$

and the proposed state transition matrices as

$$\mathbf{A_o} = \begin{bmatrix} 0 & 3/4 & 0 & 1/4 & 0 \\ 3/16 & 0 & 3/4 & 0 & 1/16 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 1/16 & 0 & 3/4 & 0 & 3/16 \\ 0 & 1/4 & 0 & 3/4 & 0 \end{bmatrix} \quad \text{and}$$

$$\mathbf{A_e} = \begin{bmatrix} 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 5/8 & 0 & 3/8 & 0 \\ 1/8 & 0 & 3/4 & 0 & 1/8 \\ 0 & 3/8 & 0 & 5/8 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \end{bmatrix}. \tag{38}$$

From (10) all possible $s_d[n]$ values are $\{-4,-3,-2,-1,0,1,2,3,4\}$, and further define

$$\mathbf{s}^{(p)} = \begin{bmatrix}(-4)^p & (-3)^p & (-2)^p & (-1)^p & 0 & 1^p & 2^p & 3^p & 4^p\end{bmatrix}^T. \tag{39}$$

Applying (18) yields

$$\mathbf{S_o} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 3/4 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 3/16 & 0 & 3/4 & 0 & 1/16 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 1/16 & 0 & 3/4 & 0 & 3/16 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 3/4 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and}$$

$$\mathbf{S_e} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1/4 & 0 & 3/4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5/8 & 0 & 3/8 & 0 & 0 \\ 0 & 0 & 1/8 & 0 & 3/4 & 0 & 1/8 & 0 & 0 \\ 0 & 0 & 3/8 & 0 & 5/8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{40}$$

Multiplying the matrices in either order yields

$$\mathbf{A_e A_o} = \mathbf{A_o A_e} = \begin{bmatrix} 0 & 9/16 & 0 & 7/16 & 0 \\ 9/64 & 0 & 3/4 & 0 & 7/64 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 7/64 & 0 & 3/4 & 0 & 9/64 \\ 0 & 7/16 & 0 & 9/16 & 0 \end{bmatrix} \tag{41}$$

so the matrices commute. Direct computation reveals that the eigenvectors of both $\mathbf{A_e}$ and $\mathbf{A_o}$ are linearly independent, and therefore $\mathbf{A_e}$ and $\mathbf{A_o}$ are diagonalizable [21]. Specifically, $\mathbf{A_e^n} = \mathbf{V_e \Lambda_e^n V_e^{-1}}$, where

$$\mathbf{\Lambda_e^n} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/4^n & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1/4^n \end{bmatrix},$$

$$\mathbf{V_e} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & -1/3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & -1 & 1 & 0 & 0 \end{bmatrix},$$

$$\text{and } \mathbf{V_e^{-1}} = \begin{bmatrix} 1/8 & 0 & 3/4 & 0 & 1/8 \\ 1/2 & 0 & 0 & 0 & -1/2 \\ 3/8 & 0 & -3/4 & 0 & 3/8 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & -1/2 & 0 & 1/2 & 0 \end{bmatrix} \tag{42}$$

and $\mathbf{A_o^n} = \mathbf{V_o \Lambda_o^n V_o^{-1}}$, where

$$\mathbf{\Lambda_o^n} = \begin{bmatrix} (-1)^n & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & (-1/4)^n & 0 & 0 \\ 0 & 0 & 0 & 1/4^n & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{V_o} = \begin{bmatrix} 1 & 1 & 1 & -1 & 1 \\ -1 & 1 & -1/2 & -1/2 & 0 \\ 1 & 1 & 0 & 0 & -1/3 \\ -1 & 1 & 1/2 & 1/2 & 0 \\ 1 & 1 & -1 & 1 & 1 \end{bmatrix},$$

$$\text{and } \mathbf{V_o^{-1}} = \begin{bmatrix} 1/16 & -1/4 & 3/8 & -1/4 & 1/16 \\ 1/16 & 1/4 & 3/8 & 1/4 & 1/16 \\ 1/4 & -1/2 & 0 & 1/2 & -1/4 \\ -1/4 & -1/2 & 0 & 1/2 & 1/4 \\ 3/8 & 0 & -3/4 & 0 & 3/8 \end{bmatrix}. \tag{43}$$

By inspection of (42), $\mathbf{\Lambda_e^n}$ converges to

$$\mathbf{\Lambda_e}, 1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{44}$$

The vector given by $\mathbf{V_e \Lambda_{e,1} V_e^{-1} t}^{(p)}$ is equal to $b_p \mathbf{1}$, where $b_p = 0$, 1 and 0 for $p = 1$, 2, and 3, respectively, which is of the form required by Theorem 1. To show exponential convergence, consider

$$\begin{aligned} \left\| \mathbf{A_e^n t}^{(p)} - b_p \mathbf{1} \right\| &= \left\| \left(\mathbf{A_e^n} - \mathbf{V_e \Lambda_{e,1} V_e^{-1}} \right) \mathbf{t}^{(p)} \right\| \\ &\le \left\| \mathbf{A_e^n} - \mathbf{V_e \Lambda_{e,1} V_e^{-1}} \right\| \left\| \mathbf{t}^{(p)} \right\| \end{aligned} \tag{45}$$

where $\|\ \|$ is the l_2 norm, and $p = 1, 2$ or 3. Evaluating $\|\mathbf{t}^{(p)}\|$ for $p = 3$, and $\|\mathbf{A_e^n} - \mathbf{V_e \Lambda_{e,1} V_e^{-1}}\|$ yields $\sqrt{130}$ and $\sqrt{2}(1/4)^n$, respectively, therefore the right side of (45) is equal to

$$\sqrt{260}(1/4)^n \tag{46}$$

and therefore each element of the vector given by $\mathbf{A_e^n t}^{(p)} - b_p \mathbf{1}$ converges exponentially to zero.

By inspection of (43), $\mathbf{\Lambda_o^n}$ does not converge, however, it is sufficient to show that the vector $\mathbf{V_o \Lambda_o^n V_o^{-1} t}^{(p)}$ converges. Consider $\mathbf{A_o^n} = \mathbf{V_o \Lambda_{o,1}^n V_o^{-1}} + \mathbf{V_o \Lambda_{o,2}^n V_o^{-1}}$ where

$$\mathbf{\Lambda_{o,1}^n} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & (-1/4)^n & 0 & 0 \\ 0 & 0 & 0 & 1/4^n & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and}$$

$$\mathbf{\Lambda_{o,2}^n} = \begin{bmatrix} (-1)^n & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{47}$$

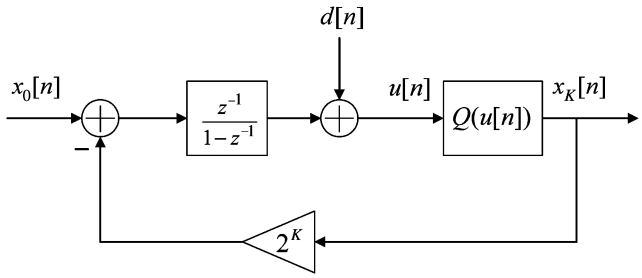


Fig. 4. First-order $\Delta\Sigma$ modulator.

Multiplying $\mathbf{V_o}\mathbf{\Lambda}_{\mathbf{o},2}^n\mathbf{V_o}^{-1}$ by $\mathbf{t}^{(p)}$ for $p = 1$, 2 or 3 results in a vector with all zero elements for all $n \geq 1$. Therefore, for all $n \geq 1$ and $p = 1$, 2, or 3, $\mathbf{A_o}^n\mathbf{t}^{(p)} = \mathbf{V_o}\mathbf{\Lambda}_{\mathbf{o},1}^n\mathbf{V_o}^{-1}\mathbf{t}^{(p)}$. By inspection, $\mathbf{\Lambda}_{\mathbf{o},1}^n$ converges to

$$\mathbf{\Lambda}_{\mathbf{o},3} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{48}$$

The vector given by $\mathbf{V_o}\mathbf{\Lambda}_{\mathbf{o},3}\mathbf{V_o}^{-1}\mathbf{t}^{(p)}$ is equal to $b_p\mathbf{1}$, where $b_p = 0$, 1 and 0 for $p = 1$, 2 and 3 respectively. Replacing $\mathbf{A_e}^n$, $\mathbf{V_e}$, $\mathbf{\Lambda}_{\mathbf{e},1}$, and $\mathbf{V_e}^{-1}$ in (45) with $\mathbf{A_o}^n$, $\mathbf{V_o}$, $\mathbf{\Lambda}_{\mathbf{o},3}$, and $\mathbf{V_o}^{-1}$ respectively shows that $\|\mathbf{A_o}^n\mathbf{t}^{(p)} - b_p\mathbf{1}\|$ converges exponentially to $b_p\mathbf{1}$. Therefore the state transition matrices given by (38) satisfy the conditions of Theorem 1 for $h_t = 3$.

Using the decomposition in (42) and (43) and the sequence transition matrices given by (40), it can be shown by direct computation that $\mathbf{A_o}^n\mathbf{S_e}\mathbf{s}^{(p)}$, $\mathbf{A_o}^n\mathbf{S_o}\mathbf{s}^{(p)}$, $\mathbf{A_e}^n\mathbf{S_e}\mathbf{s}^{(p)}$ and $\mathbf{A_e}^n\mathbf{S_o}\mathbf{s}^{(p)}$ converges to $c_p\mathbf{1}$, where $c_p = 0$, 1.5, 0, 6, and 0 for $p = 1$, 2, 3, 4 and 5, respectively. Furthermore, the convergence of each vector at index n can be bounded using (45), replacing $\|\mathbf{t}^{(p)}\|$ alternately with $\|\mathbf{S_e}\mathbf{s}^{(p)}\|$ and $\|\mathbf{S_o}\mathbf{s}^{(p)}\|$, which implies that the convergence of $\mathbf{A_o}^n\mathbf{S_e}\mathbf{s}^{(p)}$, $\mathbf{A_o}^n\mathbf{S_o}\mathbf{s}^{(p)}$, $\mathbf{A_e}^n\mathbf{S_e}\mathbf{s}^{(p)}$ and $\mathbf{A_e}^n\mathbf{S_o}\mathbf{s}^{(p)}$ are exponential. Therefore, the matrices $\mathbf{A_e}$, $\mathbf{A_o}$, $\mathbf{S_e}$, and $\mathbf{S_o}$ given in (38) and (40) also satisfy the conditions of Theorem 2 for $h_s = 5$.

B. Simulation Results and Comparisons

The successive requantizer presented above performs first-order quantization noise shaping. Therefore, it is reasonable to compare its quantization noise characteristics to those of a first-order $\Delta\Sigma$ modulator of the type shown in Fig. 4. The $\Delta\Sigma$ modulator consists of a discrete-time integrator and a mid-tread quantizer enclosed in a negative feedback loop. A random iid dither sequence, $d[n]$, is added to the output of the discrete-time integrator prior to the quantizer to ensure that the quantization noise sequence introduced by the quantizer is white (and therefore free of spurious tones) [22]. The quantizer implements

$$x_k[n] = \left\lfloor \frac{u[n]}{2^K} + \frac{1}{2} \right\rfloor$$

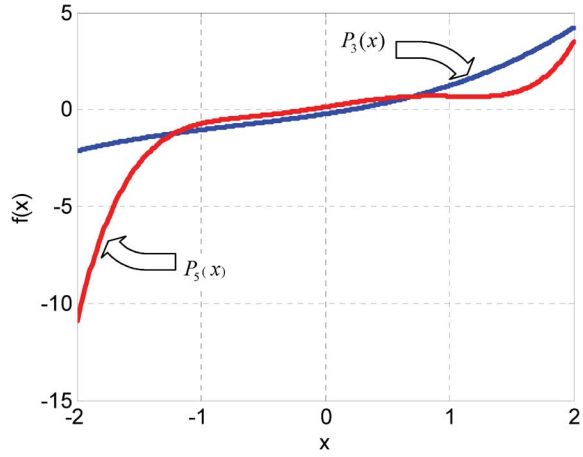


Fig. 5. Distortion polynominals applied on $t[n]$ and $s[n]$.

where $\lfloor\ \rfloor$ is the floor function, and the dither sequence has a triangular pmf with support on $\{0, 2^{K+1} - 2\}$.

Simulation results for the successive requantizer presented above and the $\Delta\Sigma$ modulator with $K = 16$ and a constant input of $x_0[n] = 2048$ are shown in Fig. 6. The quantization noise, as well as its running sum for both the successive requantizer and the $\Delta\Sigma$ modulator are subjected to the following distortion polynomials graphically shown in Fig. 5

$$\begin{aligned} P_3(t[n]) &= 0.15\,(t[n])^3 + 0.32\,(t[n])^2 + 0.99\,(t[n]) - 0.23 \\ P_5(s[n]) &= 0.32\,(s[n])^5 - 0.27\,(s[n])^4 - 0.64\,(s[n])^3 \\ &\quad + 0.12\,(s[n])^2 + 1.03\,(s[n]) + 0.13 \end{aligned} \tag{49}$$

which represent levels of distortion typically found in fractional-N PLLs [23]. Fig. 6 shows the estimated power spectra of the quantization noise and integrated quantization noise before and after application of the distortion polynomials. The estimated power spectra of the sequences, $t^p[n]$ or $s^p[n]$, are taken to be the average of the periodograms of the M *windowed* sequences, $t^p[n-kL]w[n-kL]$ and $s^p[n-kL]w[n-kL]$, for $k = 1, 2, \ldots, M$, where $w[n]$ is a Hanning window of length L. As expected from the theoretical results presented above, no spurious tones are apparent in the figures for the successive requantizer before or after application of the distortion polynomials. In contrast, spikes, which imply the presence of spurious tones, are evident in the estimated power spectra of the quantization noise from the $\Delta\Sigma$ modulator after application of the distortion polynomials.

The distortion polynomials in (49) are applied to the quantization noise and integrated quantization noise of the noise-shaped coding schemes in [16]–[19], with the result shown in Fig. 7. The example noise-shaped coders used to generate the data in Fig. 7 are a third-order single-bit delta-sigma modulator with $p = 0.99$ [16], a fifth-order coder with a quantizer step size, of 1/8 [17], a first-order single loop delta-sigma modulator with a time horizon of two samples [18], and a second-order coder

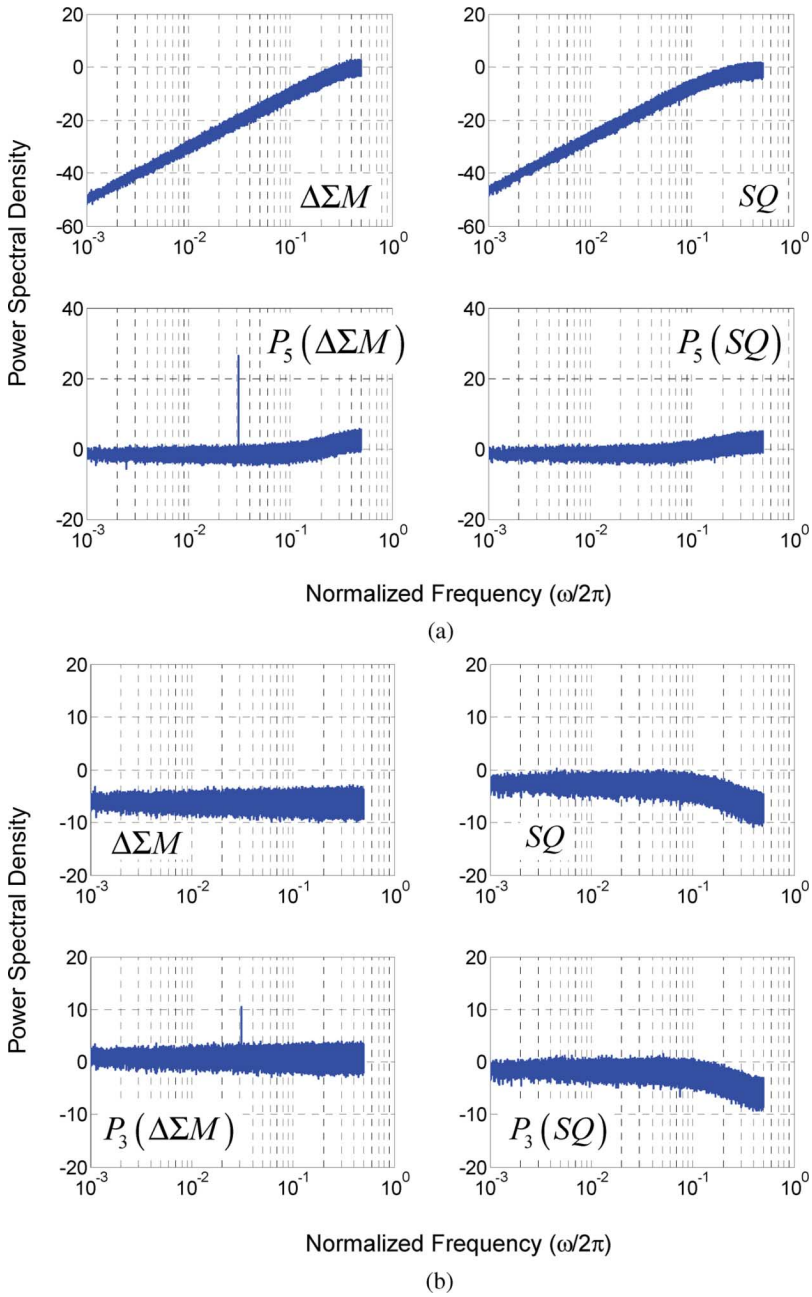


Fig. 6. Estimated power spectra of (a) the quantization noise sequences, and (b) the running sums of the quantization noise sequences of the first-order $\Delta\Sigma$ modulator and the successive requantizer presented in Section IV before and after application of nonlinear distortion.

with $M = 7/3$ [19]. Spurious tones are evident in both the quantization noise, and integrated quantization noise. Furthermore, the same distortion polynominals are applied to the quantization noise and integrated quantization noise of the successive requantizer presented above with a full-scale sinusoidal input at a frequency of 1/64 the sample rate. As expected from the theoretical results, a periodic input also does not result in spurious tones in the quantization noise before or after application of the distortion polynominals.

APPENDIX

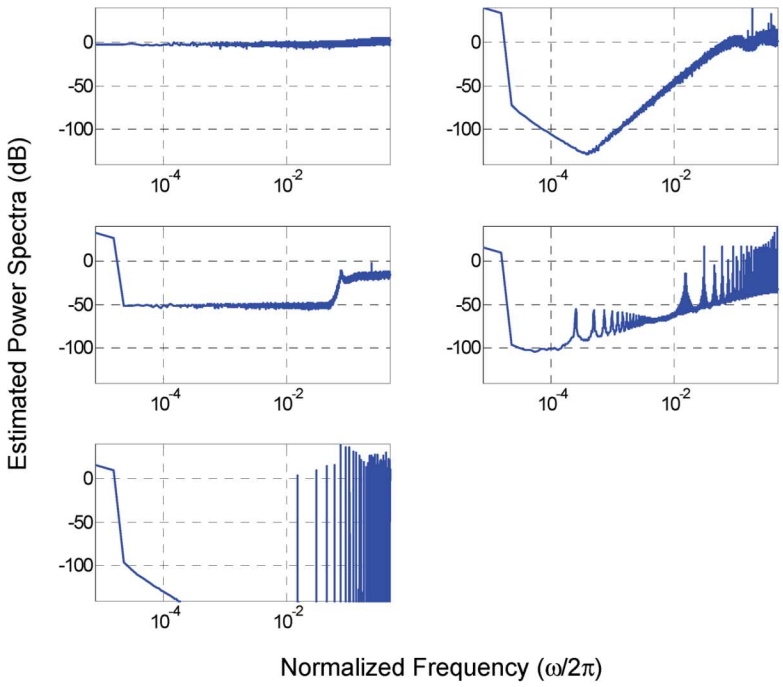
This Appendix contains the proof for Lemmas 1 and 2, used in Theorems 1 and 2, respectively.

Lemma 1: Suppose the conditions of Theorem 1 are satisfied. Then there exists a constant C_{t^p}, positive constants D_1, D_2, and a constant $0 < \alpha < 1$ such that for $n_1 \neq n_2$

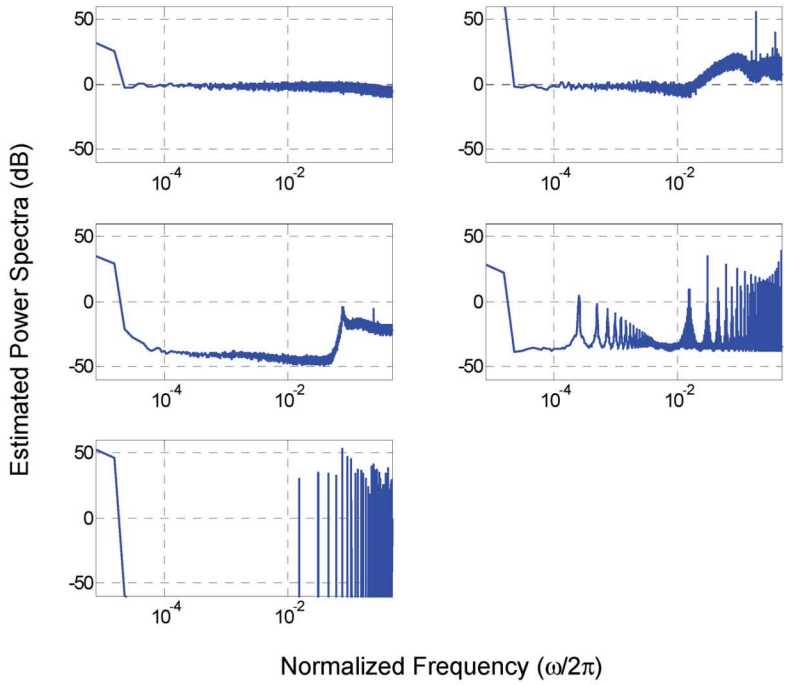
$$|E\left[t^p[n_1]t^p[n_2] - C_{t^p}\right]| \leq D_1\alpha^{|n_2-n_1|} + D_2\alpha^{n_1}. \quad (50)$$

Proof of Lemma 1: To establish (50), it suffices to assume that $n_2 > n_1$. Using (6), $E[t^p[n_2]t^p[n_1]]$ can be expressed as

$$E\left[t^p[n_2]t^p[n_1]\right] = \sum_{c_1=0}^{K-1}\cdots\sum_{c_p=0}^{K-1}\sum_{d_1=0}^{K-1}\cdots\sum_{d_p=0}^{K-1} 2^{c_1+\cdots+c_p+d_1+\cdots+d_p} \times E\left[\prod_{i=1}^{p} t_{c_i}[n_2]\prod_{j=1}^{p} t_{d_j}[n_1]\right]. \quad (51)$$



(a)



(b)

Fig. 7. Estimated power spectra of (a) the quantization noise sequences, and (b) the running sums of the quantization noise sequences of (clockwise from upper left) the successive requantizer presented in Section IV with a periodic full-scale sinusoidal input, the noise shaped coders from [16]–[19].

It is seen that the above expression is a finite sum of terms of the form

$$Q(n_1, n_2) = E\left[\prod_{j=0}^{K-1} \left(t_j^{p_j}[n_2] t_j^{q_j}[n_1]\right)\right] \tag{52}$$

where p_j and q_j are positive integers less than or equal to p. It thus suffices to establish a bound for $Q(n_1, n_2)$ of the form

$$|Q(n_1, n_2) - C_3| \leq C_1 \alpha^{n_2 - n_1} + C_2 \alpha^{n_1}. \tag{53}$$

The right side of (52) is computed by conditional expectation as follows:

$$Q(n_1, n_2) = E\left\{\prod_{i=0}^{K-1} t_i^{p_i}[n_1] E\left(\prod_{j=0}^{K-1} t_j^{q_j}[n_2] | t_d[n_1], o_d[n], \right.\right.$$
$$\left.\left. d = 0, 1, \ldots, K-1,\ n = n_1 + 1, \ldots, n_2\right)\right\}. \tag{54}$$

Substituting (15) into the inner conditional expectation of (54) yields

$$Q(n_1,n_2)=E\Bigg\{\prod_{j=0}^{K-1}\left(t_j^{p_j}[n_1]E\left[t_j^{q_j}[n_2]|t_j[n_1],o_j[n],\right.\right.\\ \left.\left.n=n_1+1,\ldots,n_2\right]\right)\Bigg\}. \quad (55)$$

Since $\{t_d[n],\ n=0,1,\ldots\}$ is a Markov process for any given parity sequence, $\{o_d[n]=o_{d,n},\ n=0,1,\ldots\}$ where $o_{d,n}\in\{0,1\}$, it follows from (16) and (17) that the m-step state transition matrix corresponding to $t_d[n]$ from time n to time $n+m$ can be written as

$$\mathbf{A}_d[n,m]=\prod_{k=n+1}^{n+m}\left[\mathbf{A_o}o_{d,k}+\mathbf{A_e}(1-o_{d,k})\right] \quad (56)$$

where $\mathbf{A}_d[n,m]$ is an $N\times N$ matrix with elements of the form

$$P\left\{t_d[n+m]=T_j|t_d[n]=T_i,o_d[n+1]=o_{d,n+1},\right.\\ \left.o_d[n+2]=o_{d,n+2},\ldots,o_d[n+m]=o_{d,n+m}\right\}. \quad (57)$$

Since $o_{d,n}$ is either 1 or 0 for each n, (25) can be used to write (56) as

$$\mathbf{A}_d[n,m]=\mathbf{A}_\mathbf{e}^{y_m}\mathbf{A}_\mathbf{o}^{m-y_m}\\ =\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{A}_\mathbf{e}^{y_m},\ \text{where } y_m=\sum_{k=n+1}^{n+m}o_{d,k}. \quad (58)$$

By definition, $y_m\geq m/2$ or $m-y_m\geq m/2$ depending on the given parity sequence. It follows from the exponential convergence of (26) that there exists positive numbers $C_{p,e}$ and $C_{p,o}$ and positive numbers $\alpha_{p,e}$ and $\alpha_{p,o}$ less than unity such that each element of

$$\mathbf{A}_\mathbf{e}^{y_m}\mathbf{t}^{(p)}-b_p\mathbf{1} \quad (59)$$

is less than $C_{p,e}\alpha_{p,e}^{m/2}$ for $y_m\geq m/2$, and each element

$$\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{t}^{(p)}-b_p\mathbf{1} \quad (60)$$

is less than $C_{p,o}\alpha_{p,o}^{m/2}$ for $m-y_m\geq m/2$.

The matrices $\mathbf{A}_\mathbf{o}^{m-y_m}$ and $\mathbf{A}_\mathbf{e}^{y_m}$ are stochastic matrices, so $\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{1}=\mathbf{1}$, $\mathbf{A}_\mathbf{e}^{y_m}\mathbf{1}=\mathbf{1}$ and

$$\mathbf{A}_\mathbf{o}^{m-y_m}\left(\mathbf{A}_\mathbf{e}^{y_m}\mathbf{t}^{(p)}-b_p\mathbf{1}\right)=\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{A}_\mathbf{e}^{y_m}\mathbf{t}^{(p)}-b_p\mathbf{1} \quad (61)$$

$$\mathbf{A}_\mathbf{e}^{y_m}\left(\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{t}^{(p)}-b_p\mathbf{1}\right)=\mathbf{A}_\mathbf{e}^{y_m}\mathbf{A}_\mathbf{o}^{m-y_m}\mathbf{t}^{(p)}-b_p\mathbf{1}. \quad (62)$$

Since the elements of the vectors in (59) and (60) are exponentially bounded, the same must be true for the vectors in (61) and (62). From (58) it follows that the right side of either (61) or (62) is equal to

$$\mathbf{A}_d[n,m]\mathbf{t}^{(p)}-b_p\mathbf{1}. \quad (63)$$

Therefore, in general each element of (63) has a magnitude less than $C\alpha^{m/2}$ where $C=\max\{C_{p,e},C_{p,o}\}$ and $\alpha=\max\{\alpha_{p,e},\alpha_{p,o}\}$, which implies that

$$E\left[t_d^p[n+m]|t_d[n],o_d[n+j]=o_{d,n+j},\ j=1,\ldots,m\right]\to b_p \quad (64)$$

as $m\to\infty$ uniformly in n where the convergence is also exponential. This result is independent of the given deterministic sequence $\{o_{d,n},\ n=0,1,\ldots\}$, so it implies that

$$E\left[t_d^p[n+m]|t_d[n],o_d[n+j],\ j=1,\ldots,m\right]\to b_p \quad (65)$$

almost surely as $m\to\infty$ uniformly in n where the convergence is also exponential.

Thus, the inner conditional expectation in (55) converges exponentially to b_{r_j} as $n_2-n_1\to\infty$ with probability one so that

$$Q(n_1,n_2)\to\prod_{j=0}^{K-1}b_{q_j}E\left\{\prod_{i=0}^{K-1}t_i^{p_i}[n_1]\right\}. \quad (66)$$

More precisely, the exponential convergence of (66) implies that for every $n_2>n_1$

$$\left|E\left[t_j^{q_j}[n_2]|t_j[n_1],o_j[n],\ n=n_1+1,\ldots,n_2\right]-b_{q_j}\right|\\ \leq C(q_j)\alpha^{n_2-n_1} \quad (67)$$

with probability one where $C(q_j)$ is a constant that depends on q_j. For every $n_2>n_1$

$$\left|Q(n_1,n_2)-\prod_{j=0}^{K-1}b_{q_j}E\left\{\prod_{i=0}^{K-1}t_i^{p_i}[n_1]\right\}\right|\\ \leq E\Bigg\{\prod_{j=0}^{K-1}|t_j[n_1]|^{p_j}\left|E\left[t_j^{q_j}[n_2]|t_j[n_1],o_j[n],\right.\right.\\ \left.\left.n=n_1+1,\ldots,n_2\right]-b_{q_j}\right|\Bigg\}\\ \leq\prod_{j=0}^{K-1}C(q_j)B^{p_j}\alpha^{n_2-n_1}\triangleq C_1\alpha^{n_2-n_1} \quad (68)$$

where B is given from Property 2. By similar reasoning, it can be established that

$$\left|E\left\{\prod_{j=0}^{K-1}t_j^{q_j}[n_1]\right\}-\prod_{j=0}^{K-1}b_{q_j}\right|\leq C_2\alpha^{n_1}. \quad (69)$$

Hence, the above two bounds imply there exist positive constants C_1 and C_2 such that for all $n_2>n_1$

$$\left|Q(n_1,n_2)-\prod_{i=0}^{K-1}b_{p_i}\prod_{j=0}^{K-1}b_{q_j}\right|\\ \leq\left|Q(n_1,n_2)-E\left\{\prod_{i=0}^{K-1}t_i^{p_i}[n_1]\right\}\prod_{j=0}^{K-1}b_{q_j}\right|\\ +\left|E\left\{\prod_{i=0}^{K-1}t_i^{p_i}[n_1]\right\}\prod_{j=0}^{K-1}b_{q_j}-\prod_{i=0}^{K-1}b_{p_i}\prod_{j=0}^{K-1}b_{q_j}\right|\\ \leq C_1\alpha^{n_2-n_1}+C_2\alpha^{n_1}. \quad (70)$$

Consequently, there exists a constant C_3 such that

$$|Q(n_1, n_2) - C_3| \leq C_1 \alpha^{n_2 - n_1} + C_2 \alpha^{n_1} \tag{71}$$

which is of the required form. ∎

Lemma 2: Suppose the conditions of Theorem 2 are satisfied. Then there exists a constant C_{s^p}, positive constants E_1, E_2, and a constant $0 < \beta < 1$ such that for $n_1 \neq n_2$

$$|E\left[s^p[n_1]s^p[n_2] - C_{s^p}\right]| \leq E_1 \beta^{|n_2 - n_1|} + E_2 \beta^{n_1}. \tag{72}$$

Proof of Lemma 2: The proof is similar to that of Lemma 1, so only the nontrivial differences with respect to the proof of Lemma 1 are presented.

Similarly to the proof of Lemma 1, it is necessary to show that

$$E\left[s_d^p[n+m] | t_d[n], o_d[n+j],\ j = 1, \ldots, m\right] \to c_p \tag{73}$$

almost surely as $m \to \infty$ uniformly in n where the convergence is also exponential. With this result and $s_d[n]$, c_p, and (20) playing the roles of $t_d[n]$, b_p, and (15) in the proof of Lemma 1, respectively, the proof of Lemma 2 is almost identical that of Lemma 1. Therefore, it is sufficient to prove (73).

Since the random variables $t_d[n-1]$ and $o_d[n]$ are statistically independent, for any given parity sequence, $\{o_d[n] = o_{d,n},\ n = 0, 1, \ldots\}$ where $o_{d,n} \in \{0, 1\}$, it follows from (21), (22), and (57) that

$$\mathbf{S}_d[n, m+1] = \mathbf{A}_d[n, m]\left[\mathbf{S_o} o_{d,n+m+1} + \mathbf{S_e}(1 - o_{d,n+m+1})\right] \tag{74}$$

where $\mathbf{S}_d[n, m+1]$ is an $N \times N'$ matrix with elements of the form

$$P\left\{s_d[n+m+1] = S_j | t_d[n] = T_i, o_d[n+1] = o_{d,n+1}, \ldots, o_d[n+m+1] = o_{d,n+m+1}\right\} \tag{75}$$

where i is the row index and j is the column index. By similar reasoning to that used in the proof of Lemma 1, (33) and (34) together imply that there exists a positive number D and a positive number β less than unity such that each element of the vector

$$\mathbf{S}_d[n, m+1]\mathbf{s}^{(p)} - c_p \mathbf{1} \tag{76}$$

has a magnitude less than $D \cdot \beta^{m/2}$. Thus, (76) implies that

$$E\left[s_d^p[n+m] | t_d[n], o_d[n+j] = o_{d,n+j},\ j = 1, \ldots, m\right] \to c_p \tag{77}$$

as $m \to \infty$ uniformly in n where the convergence is also exponential. This result is independent of the given deterministic sequence $\{o_{d,n},\ n = 0, 1, \ldots\}$, so it implies that (73) holds almost surely as $m \to \infty$ uniformly in n where the convergence is also exponential. ∎

ACKNOWLEDGMENT

The authors would like to acknowledge S. Pamarti and J. Welz for helpful discussions relating to this work.

REFERENCES

[1] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. New York: Wiley-IEEE Press, 2004.
[2] B. Razavi, *Phase-Locking in High-Performance Systems: From Devices to Architectures*. New York: Wiley-Interscience, 2003.
[3] I. Galton, "Delta-sigma data conversion in wireless transceivers," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 1, pp. 302–315, Jan. 2002.
[4] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[5] S. Pamarti, J. Welz, and I. Galton, "Statistics of the quantization noise in one-bit dithered single-quantizer digital delta-sigma modulators," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 54, no. 3, pp. 492–503, Mar. 2007.
[6] W. Chou and R. M. Gray, " Dithering and its effects on sigma-delta and multistage sigma-delta modulation," *IEEE Trans. Inf. Theory*, vol. 37, no. 3, pp. 500–513, May 1991.
[7] S. Pamarti, L. Jansson, and I. Galton, "A wideband 2.4-GHz delta-sigma fractional-N PLL with 1-Mb/s in-loop modulation," *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 49–62, Jan. 2004.
[8] B. De Muer and M. Steyaert, "A CMOS monolithic $\Delta\Sigma$-controlled fractional-N frequency synthesizer for DCS-1800," *IEEE J. Solid-State Circuits*, vol. 37, no. 7, pp. 835–844, Jul. 2002.
[9] B. H. Marcus and P. H. Siegel, "On codes with spectral nulls at rational submultiples of the symbol frequency," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 4, pp. 557–568, Jul. 1987.
[10] G. L. Pierobon, "Codes for zero spectral density at zero frequency," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 2, pp. 435–439, Mar. 1984.
[11] I. Galton, "Spectral shaping of circuit errors in digital-to-analog converters," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 44, no. 10, pp. 808–817, Oct. 1997.
[12] E. Fogleman and I. Galton, "A digital common-mode rejection technique for differential analog-to-digital conversion," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 48, no. 3, pp. 255–271, Mar. 2001.
[13] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.
[14] J. Welz, I. Galton, and E. Fogleman, "Simplified logic for first-order and second-order mismatch-shaping digital-to-analog converters," *IEEE Trans. Circuits Syst. II: Analog Digit. Signal Process.*, vol. 48, no. 11, pp. 1014–1027, Nov. 2001.
[15] J. Welz and I. Galton, "A tight signal-band power bound on mismatch noise in a mismatch shaping digital-to-analog converter," *IEEE Trans. Inf. Theory*, vol. 50, no. 4, pp. 593–607, Apr. 2004.
[16] A. J. Magrath and M. B. Sandler, "Efficient dithering of sigma-delta modulators with adaptive bit flipping," *Electron. Lett.*, vol. 31, no. 11, pp. 846–847, May 1995.
[17] S. H. Yu, "Noise-shaping coding through bounding the frequency-weighted reconstruction error," *IEEE Trans. Circuits Syst. II: Expr. Briefs*, vol. 53, no. 1, pp. 67–71, Jan. 2006.
[18] D. E. Quevedo and G. C. Goodwin, "Multistep optimal analog-to-digital conversion," *IEEE Trans. Circuits and Systems I: Regular Papers*, vol. 52, no. 3, pp. 503–515, Mar. 2005.
[19] 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.
[20] A. Papoulis and S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes*. New York: McGraw-Hill, 2002.
[21] R. A. Horn and C. A. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
[22] R. M. Gray and T. G. Stockham, Jr., "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.
[23] A. Swaminathan, K. J. Wang, and I. Galton, "A wide-bandwidth 2.4GHz ISM-band fractional-N PLL with adaptive phase-noise cancellation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, Feb. 2007, pp. 302–303, 604.

Ashok Swaminathan (S'96) received the B.A.Sc. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, the M.Eng. degree in electronics engineering from Carleton University, Ottawa, ON, Canada, and the Ph.D. degree from the University of California at San Diego, La Jolla, in 1994, 1997, and 2006, respectively.

From 1997 to 2000, he was with Philsar Semiconductor, which was later acquired by Skyworks Solutions, developing analog and mixed-signal circuits for low-power wireless transceivers. Since 2006, he has been with NextWave Broadband, San Diego, CA, designing high-performance frequency synthesizers for WiMAX applications.

Andrea Panigada (S'05) received the Laurea degree in electrical engineering from the University of Pavia, Pavia, Italy, in 1999.

From 2000 to 2004, he worked for STMicroelectronics as a Design Engineer at the Studio di Microelettronica of Pavia, a design center founded by STMicroelectronics with the cooperation of the Electronic department of the University of Pavia. There he conducted research in algorithms for the digital calibration of analog-to-digital converters and in the design of CMOS prototypes of sigma-delta and pipelined ADCs. In 2003, he spent one year as a visiting scholar at the University of California at San Diego (UCSD), where he worked at the Integrated Signal Processing Group (ISPG). Since 2005, he has been pursuing the Ph.D. degree at UCSD, where he joined the ISPG. His research interests are in the field of mixed-signal integrated circuits and systems, including data converters.

Elias Masry (S'64–M'68–SM'83–F'86) received the B.Sc. and M.Sc. degrees from the Technion–Israel Institute of Technology, Haifa, Israel, in 1963 and 1965, respectively, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1966 and 1968, respectively, both in electrical engineering.

From 1955 to 1959, he served in the Israel Defense Forces. From 1963 to 1965, he was a Teaching Assistant at the Technion. Since 1968, he has been on the faculty of the University of California at San Diego, La Jolla, CA, and is currently a Professor of electrical engineering. His current research interests include spectral, probability density, and regression functions estimation in a time series context, function estimation from nonequally spaced data, analysis of adaptive filtering algorithms and wireless communication systems.

Dr. Masry was the Associate Editor for Stochastic Processes of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1980 to 1983.

Ian Galton (M'92) received the Sc.B. degree from Brown University, Providence, RI, in 1984, and the M.S. and Ph.D. degrees from the California Institute of Technology, Pasadena, CA, in 1989 and 1992, respectively, all in electrical engineering.

Since 1996, he has been a Professor of electrical engineering at the University of California at San Diego, where he teaches and conducts research in the field of mixed-signal integrated circuits and systems for communications. Prior to 1996, he was with the University of California at Irvine, and prior to 1989, he was with Acuson and Mead Data Central. His research involves the invention, analysis, and integrated circuit implementation of critical communication system blocks such as data converters, frequency synthesizers, and clock recovery systems. In addition to his academic research, he regularly consults at several semiconductor companies and teaches industry-oriented short courses on the design of mixed-signal integrated circuits.

Dr. Galton has served on a corporate Board of Directors, on several corporate Technical Advisory Boards, as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, as a member of the IEEE Solid-State Circuits Society Administrative Committee, as a member of the IEEE Circuits and Systems Society Board of Governors, as a member of the IEEE International Solid-State Circuits Conference Technical Program Committee, and as a member of the IEEE Solid-State Circuits Society Distinguished Lecturer Program.