Statistics of the Quantization Noise in 1-Bit Dithered Single-Quantizer Digital Delta–Sigma Modulators

Sudhakar Pamarti, *Member, IEEE*, Jared Welz, *Member, IEEE*, and Ian Galton, *Member, IEEE*

***Abstract*—An analysis of the quantization noise introduced by a widely-used class of single-quantizer digital delta–sigma ($\Delta\Sigma$) modulators with low-level, 1-bit dither is presented. Necessary and sufficient conditions are derived that ensure, in an asymptotic sense, various ensemble statistical properties of the quantization noise such as uniformity and independence from the input and delayed versions of itself. The conditions are also shown to be sufficient for a single realization of the quantization noise sequence to possess these properties in a time-averaged sense. Several of the most commonly-used digital $\Delta\Sigma$ modulators are shown to satisfy the conditions.**

***Index Terms*—Delta–sigma ($\Delta\Sigma$) modulation, dither techniques, quantization.**

I. Introduction

HIGH-PERFORMANCE analog-to-digital converters (ADCs), digital-to-analog converters (DACs), and fractional-N phase-locked loops (PLLs) based on delta–sigma ($\Delta\Sigma$) modulation—collectively referred to as $\Delta\Sigma$ data converters—are enabling components in consumer communications and entertainment products including cellular telephones, wireless LANs, modems, and MP3 players. The basic concept underlying $\Delta\Sigma$ data converters is that of performing coarse quantization within a feedback loop such that the power of the resulting quantization noise is suppressed within some frequency band of interest. This technique is known generally as *quantization noise shaping*, but in the context of data conversion is usually referred to as $\Delta\Sigma$ *modulation* for historical reasons [1]. Devices that perform $\Delta\Sigma$ modulation are referred to as $\Delta\Sigma$ modulators.

The two most basic classes of $\Delta\Sigma$ modulators are analog and digital. Analog $\Delta\Sigma$ modulators quantize continuous-amplitude input signals to generate discrete-amplitude, quantized output signals. They are used in oversampling ADCs. Digital $\Delta\Sigma$ modulators requantize discrete-amplitude input signals to generate more coarsely quantized discrete-amplitude output signals. Digital $\Delta\Sigma$ modulators are used in oversampling DACs and fractional-N PLLs, examples of which are shown in Fig. 1.

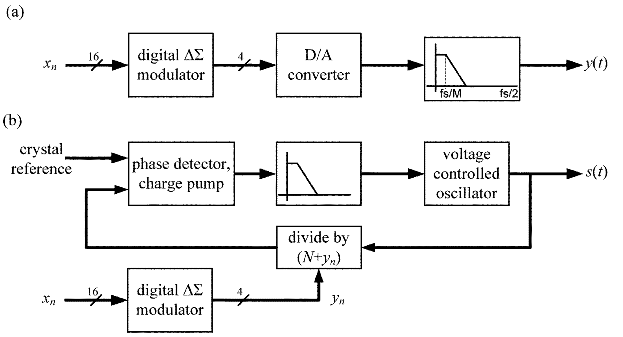

Fig. 1. Digital $\Delta\Sigma$ modulators in (a) an oversampling DAC, and (b) a fractional-N PLL.

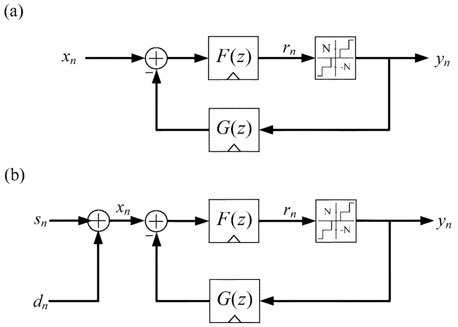

Fig. 2. (a) Generic SQDSM. (b) 1-bit dithered SQDSM.

Although the number of commercially deployed digital $\Delta\Sigma$ modulators far exceeds that of analog $\Delta\Sigma$ modulators, most of the published $\Delta\Sigma$ modulator analyses apply only to analog $\Delta\Sigma$ modulators. Interestingly, most of these analyses do not apply or even readily extend to the case of digital $\Delta\Sigma$ modulators. This paper addresses this issue by presenting a statistical analysis of the quantization noise introduced by an important family of digital $\Delta\Sigma$ modulators, i.e., single-quantizer digital $\Delta\Sigma$ modulators (SQDSMs) with low-level 1-bit input dither.

A generic SQDSM is shown in Fig. 2(a). It consists of a digital requantizer, a forward transmission filter $F(z)$, and a feedback filter $G(z)$. In most cases, it is critical in data converter applications that the error introduced by the requantizer, i.e., the *quantization noise*, be white and uncorrelated with the $\Delta\Sigma$ modulator's input sequence. Experimental evidence suggests that adding a 1-bit random sequence to the least significant bit (LSB) of the input—henceforth referred to as a *1-bit* or *LSB dithering*—imparts theses properties to the quantization noise in many digital $\Delta\Sigma$ modulators [2]–[7]. The LSB dithering technique has been very popular because it barely increases the noise floor of the digital $\Delta\Sigma$ modulator's output. In contrast, the alternative technique of adding step-size dither to the input of the quantizer dither increases the output noise floor by 3 dB.

This paper presents conditions applicable to digital $\Delta\Sigma$ modulators of the form shown in Fig. 2(a) with *LSB dithering* that

Manuscript received April 11, 2006. This paper was recommended by Associate Editor T. B. Tarim.

S. Pamarti is with the Electrical Engineering Department, University of California at Los Angeles, Los Angeles, CA 90095 USA (e-mail: spamarti@ee.ucla.edu).

J. Welz is with Broadcom Corporation, Irvine, CA 92618 USA.

I. Galton is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92037 USA (e-mail: galton@ee.ucsd.edu).

Digital Object Identifier 10.1109/TCSI.2006.887616

1549-8328/$25.00 © 2007 IEEE

are sufficient and necessary to ensure that the quantization noise is asymptotically uniform and independent of delayed versions of itself and the input sequence. The conditions are shown to be also sufficient to ensure that the quantization noise is uniform, white, and uncorrelated with the input and the dither sequences in a time-averaged sense.

The results presented in this paper are extensions of those presented in [8]–[10]. The results presented in [8] and [9] are similar to those presented in this paper, except that they are restricted to analog $\Delta\Sigma$ modulators. In [8], it is shown that no-overload dithering does not make the quantization noise in the standard first-order analog $\Delta\Sigma$ modulator white but it does so in standard multi-stage higher order analog $\Delta\Sigma$ modulators with at least two stages. In [9], conditions were derived that were sufficient for input dithering to render the quantization noise in a class of single-stage and multi-stage analog $\Delta\Sigma$ modulators asymptotically white and uncorrelated with the input. Both [8] and [9] assume that the probability distribution of the dither has an absolutely continuous component. The similarities between analog and digital $\Delta\Sigma$ modulators notwithstanding, the analog case results do not readily extend to the digital case because discrete-amplitude dither does not possess the aforementioned property of analog dither. The results presented in [10] are applicable to digital $\Delta\Sigma$ modulators, but are significantly more limited than those presented in this paper. Specifically, [10] presented conditions that were sufficient for LSB dithering to render the quantization noise in single-stage and multi-stage digital $\Delta\Sigma$ modulators asymptotically uniform and independent of the input and delayed versions of itself. This paper elaborates on the results presented in [10] and derives a more relaxed set of sufficient conditions that are also necessary.

Other published research results that characterize quantization error in $\Delta\Sigma$ modulators are restricted to analog $\Delta\Sigma$ modulators with constant or sinusoidal inputs [11]–[15] or irrational initial conditions [16]. However, these restrictions are limiting in practical applications wherein input signals deviate from ideal constant values or sinusoids, and irrational initial conditions are not feasible in digital $\Delta\Sigma$ modulators with finite bit-width data paths.

The paper consists of four main sections. Section II presents a brief overview of digital $\Delta\Sigma$ modulators and 1-bit dithering. Section III presents the aforementioned sufficient conditions. Section IV presents theoretical proof of the success of 1-bit dither in popular low-pass and bandpass digital $\Delta\Sigma$ modulators. Section V presents corroborative simulation results.

II. Digital $\Delta\Sigma$ Modulators and 1-bit Dithering

Fig. 2(b) shows a 1-bit dithered generic SQDSM. The input x_n is the sum of a desired signal sequence s_n and a *dither sequence* d_n. The dither sequence possesses the following properties:

$$\begin{aligned}&P(d_n=0)=P(d_n=1)=0.5 \qquad \forall n\in\mathbb{Z}\\ &\quad d_n \text{ is independent of } s_l, d_m \qquad \forall l,m\in\mathbb{Z}, m\neq n. \quad (1)\end{aligned}$$

Without loss of generality, all signals are assumed to be integer-valued. The integer representation follows from the digital nature of the $\Delta\Sigma$ modulator—the smallest LSB of all the signals in the $\Delta\Sigma$ modulator is *defined* to have a value of unity which implies that all signals are restricted to integer values. With this definition, it follows that the impulse responses of $F(z)\ G(z)$, and $F(z)\ G(z)$, i.e., $f_n, g\ _n$, and $(f*g)\ _n$, respectively, are integer valued.

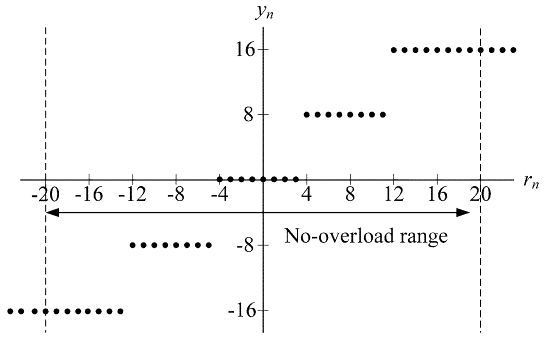

Fig. 3. Example uniform midtread requantizer.

The operation of the uniform midtread requantizer is defined as

$$y_n=\begin{cases}R_{\text{lo}}+N/2, & r_n<R_{\text{lo}}\\ N\lfloor r_n/N+1/2\rfloor, & R_{\text{lo}}\leqslant r_n<R_{\text{hi}}\\ R_{\text{hi}}-N/2, & R_{\text{hi}}\leqslant r_n\end{cases} \qquad (2)$$

where $\lfloor x\rfloor$ is the largest integer less than or equal to x, N is a positive, even integer referred to as the *step size*, and $R_{\text{lo}}<R_{\text{hi}}$ are arbitrary odd integer multiples of $N/2$.[1] Fig. 3 illustrates the operation of the uniform midtread requantizer using an example with $N=8, R_{\text{lo}}=-20$ and $R_{\text{hi}}=20$. With the quantization noise defined as

$$e_n \triangleq y_n - r_n \qquad (3)$$

it can be shown [7] that

$$y_n=\text{stf}_n * x_n + \text{ntf}_n * e_n \qquad (4)$$

where "*" is the convolution operator, and stf_n and ntf_n are, respectively, the impulse responses of the signal and noise transfer functions

$$\text{STF}(z)=\frac{F(z)}{1+F(z)G(z)};\quad \text{NTF}(z)=\frac{1}{1+F(z)G(z)}. \qquad (5)$$

Without d_n, the quantization noise, e_n, can have periodicities in its autocorrelation or can be correlated with s_n; either situation can result in undesirable artifacts such as discrete spikes in the power-spectral density (PSD) of y_n. Experimental evidence suggests that d_n removes these spikes for some special cases of the generic SQDSM. It is suspected that in such cases e_n becomes an uncorrelated sequence whose samples are uniformly distributed and uncorrelated with x_n in a time-averaged sense

$$\lim_{L\to\infty} S_L(e_n)=M_e=1/2, \quad \text{where } S_L(a_n)\triangleq\frac{1}{L}\sum_{n=0}^{L-1}a_n \qquad (6)$$

[1] The results presented in the paper apply, with minor modifications, to digital $\Delta\Sigma$ modulators with midrise requantizers.

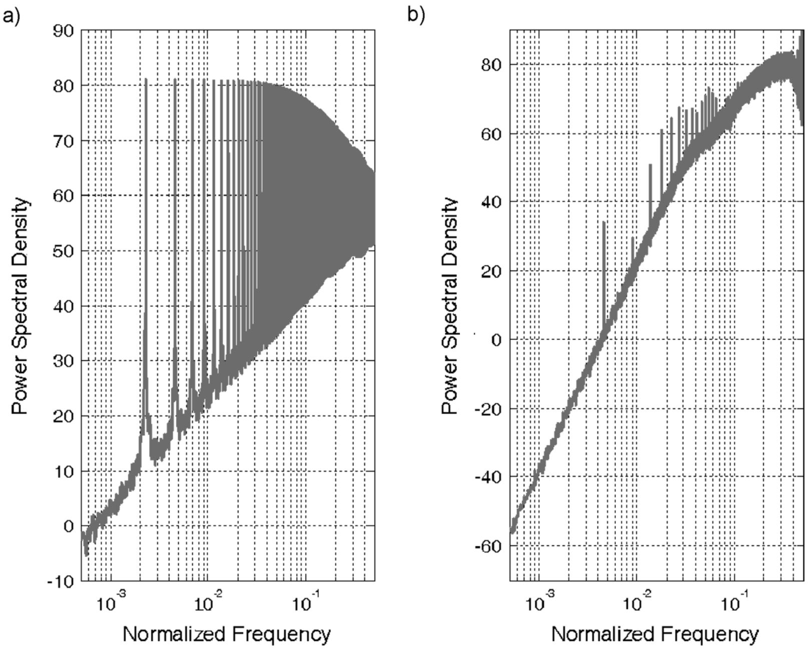

Fig. 4. Failure of 1-bit dithering: PSD of total quantization noise of a dithered SQDSM when (a) $F(z) = z^{-1}(1 - z^{-1})^{-1}, G(z) = 1$, and (b) $F(z) = z^{-3}(1 - z^{-1})^{-3}, G(z) = 3z^2 - 3z + 1$, and $R_{\text{lo}} = R_{\text{hi}} = N/2$.

$$\lim_{L\to\infty} S_L((e_n - M_e)x_{n-p}) = 0 \qquad \forall p \in \mathbb{Z} \tag{7}$$

$$\lim_{L\to\infty} S_L((e_n - M_e)(e_{n-p} - M_e)) = \sigma_{\text{ee}}^2 \delta[p] \qquad \forall p \in \mathbb{Z}$$

$$\text{where} \quad \sigma_{\text{ee}}^2 = \frac{N^2 - 1}{12}. \tag{8}$$

If (6)–(8) were true in some sense (e.g., in probability, in L^1, in L^2, or simply in mean), they could explain the disappearance of spikes in the PSD of y_n. Furthermore, they would enable accurate quantitative predictions (such as of the PSD of y_n) crucial to the use of the SQDSM. For instance, if x_n were to have a PSD, $S_{xx}(e^{j\omega})$, it can be shown[2] that the PSD of y_n is

$$S_{yy}(e^{j\omega}) = |\text{STF}(e^{j\omega})|^2 S_{xx}(e^{j\omega}) + |\text{NTF}(e^{j\omega})|^2 \sigma_{ee}^2. \tag{9}$$

Simulations suggest that (6)–(8) are not true in general as illustrated in Fig. 4. Fig. 4(a) depicts the simulated PSD of the total quantization noise [the second term in the RHS of (9)] for a popular SQDSM: $F(z) = z^{-1}(1 - z^{-1})$, and $G(z) = 1$. Fig. 4(b) depicts the same for another popular SQDSM $F(z) = z^{-3}(1-z^{-1})^{-3}, G(z) = 3z^2 - 3z + 1$, and $R_{\text{lo}} = R_{\text{hi}} = N/2$. In both cases, the requantizer step-size was chosen to be $N = 16384$ and s_n was set to be a constant. The discrete spikes in the PSD suggest that at least (8) is not true.

Conditions for which (6)–(8) are true in probability are presented in Section III. Specifically, a theorem is presented which specifies conditions on $F\,(z)$ that are necessary and sufficient for the quantization noise, e_n, to possess the following properties.

- The probability mass function (pmf) of e_n converges to that of a uniform random variable as $n \to \infty$;
- The pmf of e_n given x_{n-p} converges to that of a uniform random variable as $n \to \infty$ for every finite p;
- The joint pmf (jpmf) of e_n and e_{n-p} converges to that of a pair of independent random variables as $n \to \infty$ for every finite $p \neq 0$.

A corollary to the theorem is then presented which stipulates that the same conditions on $F\,(z)$ are sufficient for (6)–(8) to be true in probability.

The results are applicable to *non-overloading* SQDSMs, i.e., SQDSMs in which the $R_{\text{lo}} \leqslant r_n < R_{\text{hi}}$ for all n. In practice, the non-overloading condition can be guaranteed for a bounded-input, bounded output (BIBO) stable $\text{STF}(z)$ and a bounded input signal, x_n, simply by ensuring that the requantizer has sufficient number of output levels. For example, consider the digital $\Delta\Sigma$ modulator used to generate the PSD in Fig. 4(b) namely, $F(z) = z^{-3}(1 - z^{-1})^{-3}$ and $G(z) = 3z^2 - 3z + 1$. It can be shown [7] that the midtread requantizer is non-overloading provided

$$R_{\text{lo}} \leqslant \min\{s_n\} - (7N/2 - 3)$$
$$R_{\text{hi}} > \max\{s_n\} + (7N/2 - 3).$$

III. Theory of LSB Dithering

It follows from (2) that for a non-overloading SQDSM the quantization noise is given by

$$e_n = N/2 - \langle r_n + N/2 \rangle_N \tag{10}$$

where $\langle x \rangle_N = x(\text{mod} N)$ and r_n is the requantizer input given by

$$r_n = x_n * f_n - y_n * (f * g)_n. \tag{11}$$

[2] The derivation of (9) based on (3)–(8), and Fig. 2(b) is well known in the literature and is omitted here.

It further follows from (2) that y_n takes on only integer multiples of N.[3] Since $(f * g)_n$, i.e., the impulse response of $F(z)G(z)$, is integer valued, so is $y_n * (f * g)_n$, and hence, the second term in the RHS of (11) has no effect on the fractional operator. Consequently, the statistical properties of e_n do not depend on g_n.

It is assumed without loss of generality throughout the remainder of the paper that all signals associated with the SQDSM are zero for $n < 0$. It follows that the quantization noise can be written as

$$e_n = N/2 - \langle z_n + N/2 \rangle_N \tag{12}$$

where

$$z_n = a_n + \sum_{m=0}^{n} d_m f_{n-m} \quad \text{and} \quad a_n = \sum_{m=0}^{n} s_m f_{n-m}. \tag{13}$$

The statistics of e_n as given by (13) have been studied extensively when z_n takes on values from a continuous range and z_n is independent of z_m for $n \neq m$ (e.g., see [17]–[19]). In contrast, the following theorem concerns the ensemble statistics of e_n when z_n is not an independent sequence but is instead $(s_n + d_n)$ convolved with f_n. The corollary to the theorem concerns the distribution of the values taken by a single instance of the quantization noise vector $\{e_0, e_1, \ldots, e_n, \ldots\}$ an aspect that is of particular interest because of the time evolution of z_n.

Definition 1: The sequence e_n is said to be *asymptotically identically distributed and independent* of the sequence x_n if for every finite integer p,

$$P_{e_n|x_{n-p}}(a, b) \overset{n\to\infty}{\longrightarrow} P_U(a) \tag{14}$$

where $P_{A|B}(a, b)$ is the probability of A given B, U is a discrete random variable, and integers a and b take on values from the ranges of A and B, respectively.

Theorem 1: Suppose the input to a non-overloading SQDSM is $x_n = s_n + d_n$, where $s_n \in [-Q/2 + 1, Q/2]$ is an arbitrary bounded integer sequence and d_n is a *1-bit dither sequence*. Let U and V be independent integer random variables that are uniformly distributed over $[-N/2 + 1, N/2]$.

i) e_n is *asymptotically identically distributed and independent* of x_n if and only if at least one of the following conditions is true for each integer $p > 0$, and each integer $k, 0 < k < N$.
   1) The sequence $\langle k f_r \rangle_N$ does not converge to zero as $r \to \infty$.
   2) A non-negative integer $s = s(p) \neq p$ exists such that $\langle k f_s \rangle_N = N/2$.

ii) (e_n, e_{n-p}) converges in distribution to (V, U) if and only if at least one of the following conditions is true for every integer $p > 0$, and each integer pair $(k_1, k_2) \neq (0, 0)$ and $0 \leqslant k_1, k_2 < N$.
   1) The sequence $\langle k_1 f_r + k_2 f_{r+p} \rangle_N$ does not converge to zero as $r \to \infty$.
   2) A non-negative integer $s = s(p) \neq p$ exists such that $\langle k_1 f_s + k_2 f_{s+p} \rangle_N = N/2$.
   3) A non-negative integer $t = t(p) < p$ exists such that $\langle k_2 f_t \rangle_N = N/2$.

Remark: The conditions of part (ii) imply those of part (i): conditions 1 and 2 of part (i) are respectively special cases of conditions 1 and 2 of part (ii) for $(k_1, k_2) = (k, 0) \neq (0, 0)$; since $(k, 0)$ does not satisfy the condition 3 of part (ii), the conditions of part (ii) imply those of part (i).

Proof of Theorem 1: The details of the proof of part (ii) are presented below. The proof of part (i) is similar so, only the differences are pointed out.

Part (ii): Each of e_{n-p}, e_n, V, and U are bounded discrete random variables, so the convergence of (e_{n-p}, e_n) in distribution to (V, U) as $n \to \infty$ is equivalent to

$$P_{e_{n-p},e_n}(a, b) \overset{n\to\infty}{\longrightarrow} P_{V,U}(a, b) \quad \forall a, b \in \mathbb{Z} \cap [-N/2 + 1, N/2]. \tag{15}$$

The N^2 values of the jpmf of (e_{n-p}, e_n) in the LHS of (15) form a unique two-dimensional discrete Fourier transform pair with the N^2 samples of the joint characteristic function (jcf) of (e_{n-p}, e_n)

$$\Phi_{e_{n-p},e_n}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right) = \sum_{a=-N/2+1}^{N/2} \sum_{b=-N/2+1}^{N/2} e^{j\left(\frac{2\pi k_1 a}{N} + \frac{2\pi k_2 b}{N}\right)} \times P_{e_{n-p},e_n}(a, b) \qquad \forall 0 \leqslant k_1, k_2 < N. \tag{16}$$

Similarly, the N^2 samples of the jcf of (V, U) form a unique two-dimensional discrete Fourier transform pair with the N^2 samples of the jpmf of (V, U). Since the characteristic functions are bounded, it follows that (15) is equivalent to the following:

$$\Phi_{e_{n-p},e_n}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right) \overset{n\to\infty}{\longrightarrow} \Phi_{V,U}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right) = \delta[k_1]\delta[k_2] \qquad \forall 0 \leqslant k_1, k_2 < N \tag{17}$$

where the product $\delta[k_1]\delta[k_2]$ represents the samples of the jcf of two independent uniform random variables, V and U. Note that the equivalence of (15) and (17) is just a special case (for discrete, finite range) of the continuity theorem [20, Th.26.3, p. 359]. Note also that the use of jcfs to prove results about the jpds is well known [8]–[12], [17]–[19]. Since both the LHS and the RHS of (17) are equal to unity for $k_1 = k_2 = 0$, (17) (and hence, (15)) are equivalent to

$$\Phi_{e_{n-p},e_n}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right) \overset{n\to\infty}{\longrightarrow} 0 \qquad \forall 0 \leqslant k_1, k_2 < N; \quad (k_1, k_2) \neq (0, 0). \tag{18}$$

It is next proved that (18) is true if an only if the conditions of part (ii) are satisfied. First, the jcf of (e_{n-p}, e_n) is expressed in terms of the jcf of (a_{n-p}, a_n) and the common characteristic function of the independent, identically distributed random variables d_n represented $\Phi_d(u)$.

The jpmf of (e_{n-p}, e_n) is related to the jpd of (z_{n-p}, z_n) as follows:

$$P_{e_{n-p},e_n}(a, b) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} P_{z_{n-p},z_n}(lN - a, mN - b). \tag{19}$$

[3] Consider the example where $r_n = 25$ is truncated by a midtread quantizer of step size $N = 8$. In some published work, the output of the quantizer is implied to be the value $y_n = 3$ instead of $y_n = 24 = 3 * 8$. For purposes of simplicity, this paper uses the latter convention.

It follows that the jcf of (e_{n-p}, e_n) is related to the jcf of (z_{n-p}, z_n) as follows:

$$\Phi_{e_{n-p},e_n}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right) = \Phi_{z_{n-p},z_n}\left(\frac{-2\pi k_1}{N}, \frac{-2\pi k_2}{N}\right) \quad \forall 0 \leqslant k_1, k_2 < N. \tag{20}$$

The details of the derivation of (20) from (19) are not presented here for the sake of brevity since they are well known in the literature (for e.g., see (17) in [21] or Lemma A1 in [9]). So, it follows from (13) that the jcf of (z_{n-p}, z_n) is

$$\begin{aligned}\Phi_{z_{n-p},z_n}(u,v) &= E\left(e^{j(uz_{n-p}+vz_n)}\right)\\ &= E\left(e^{j(ua_{n-p}+va_n)}\prod_{l=0}^{n-p} E\left(e^{j(uf_{n-p-l}+vf_{n-l})}\right)\right.\\ &\quad\times\left.\prod_{m=n-p+1}^{n-p} E\left(e^{j(vf_{n-m})}\right)\right).\end{aligned}$$

By assumption, the samples of d_n are independent, identically distributed random variables also independent of s_n. Therefore, it follows from the properties of d_n that

$$\begin{aligned}&\Phi_{z_{n-p},z_n}(u,v)\\ &\quad= \Phi_{a_{n-p},a_n}(u,v)\cdot\prod_{l=0}^{n-p}\Phi_d(uf_{n-p-l}+vf_{n-l})\\ &\quad\quad\cdot\prod_{m=n-p+1}^{n}\Phi_d(vf_{n-m})\end{aligned} \tag{21}$$

where $\Phi_{a_{n-p},a_n}(u,v)$ is the jcf of (a_{n-p}, a_n). Substituting $r = (n-p-l)$ in the first product in the RHS of (21), $r = n - m$ in the second product, and $u = -2\pi k_1/N, v = -2\pi k_2/N$, and substituting the result in (20) gives

$$\begin{aligned}&\Phi_{e_{n-p},e_n}\left(\frac{2\pi k_1}{N}, \frac{2\pi k_2}{N}\right)\\ &\quad= \Phi_{a_{n-p},a_n}\left(\frac{-2\pi k_1}{N}, \frac{-2\pi k_2}{N}\right)\\ &\quad\quad\cdot\prod_{r=0}^{n-p}\Phi_d\left(\frac{-2\pi}{N}\{k_1f_r + k_2f_{r+p}\}\right)\\ &\quad\quad\cdot\prod_{r=0}^{p-1}\Phi_d\left(\frac{-2\pi k_2 f_r}{N}\right).\end{aligned} \tag{22}$$

Since (15) is equivalent to (18), it just remains to be shown that the RHS of (22) converges to zero as $n \to \infty$ if and only if the conditions of part (ii) are satisfied. It is shown below that this is indeed the case. But first, note that the characteristic function of each d_n is

$$\begin{aligned}\Phi_{d_n}(u) = \Phi_d(u) &= e^{ju*0}P_d(0) + e^{ju*1}P_d(1)\\ &= 0.5(1+e^{\mathrm{j}u}) = e^{ju/2}\cos(u/2).\end{aligned} \tag{23}$$

Sufficiency: Suppose that for every given finite $p > 0$, and every integer pair, $(k_1, k_2) \neq (0, 0)$, where $0 \leqslant k_1, k_2 < N$, conditions 1, 2, or 3 (of part (ii)) are true. If condition 3 is true, then one of the individual terms in the last product term in the RHS of (22) equals 0

$$\left|\Phi_d\left(\frac{-2\pi k_2 f_r}{N}\right)\right|_{r=t} = \left|\cos\left(\frac{-\pi}{N}\left(lN + \frac{N}{2}\right)\right)\right| = 0, \quad \text{where} \quad l = N\lfloor k_2 f_t/N\rfloor.$$

Similarly, if condition 2 is true, then one of the terms of the first product term equals 0. If condition 1 is true, then

$$\forall r > 0, \exists s > r \text{ s.t. } k_1 f_s + k_2 f_{s+p} = l_s N + m, \quad \text{where } 0 < m < N.$$

Since $|\cos(u/2)| < 1$ for all $u \neq m\pi, m \in \mathbb{Z}$, and N is finite

$$\forall r > 0, \quad \exists s > r \text{ s.t.}\quad \left|\Phi_d\left(\frac{-2\pi}{N}\{k_1 f_s + k_2 f_{s+p}\}\right)\right| = \left|\cos\left(\frac{-2\pi m}{N}\right)\right| < 1 - h$$

for some $h > 0$. The inequality implies that the first product term in the RHS of (22) converges to zero as $n \to \infty$. Since all the terms in the RHS of (22) are bounded, the entire RHS of (22) converges to zero as well. The result is proved for all finite $p < 0$ by interchanging the roles of e_n and e_{n-p} in the preceding discussion: for a given finite $p < 0$ and $q = -p$

$$\begin{aligned}&\lim_{n\to\infty} P_{e_{n-p},e_n}(a,b)\\ &\quad= \lim_{n\to\infty} P_{e_n,e_{n-p}}(b,a) = \lim_{n\to\infty} P_{e_n,e_{n+q}}(b,a)\\ &\quad= \lim_{n\to\infty} P_{e_{n-q},e_n}(b,a)\end{aligned}$$

thereby proving the result for all finite $p < 0$ as well.

Necessity: For every given finite $p > 0$, (22) converges to zero for arbitrary s_n only if the two product series in the RHS of (22) together converge to zero i.e., if the first product series converges to zero as $n \to \infty$, or if one of the terms of the first product series equals zero or if one of the terms of the second product series equals zero. The three conditions of part (ii) correspond to the above three cases respectively. The necessity is extended to the finite $p < 0$ case just as in the sufficiency proof.

Part (i): The proof is similar to that of part (ii). Proceeding as before, it can be shown that proving (14) is equivalent to proving that

$$\Phi_{e_n|x_{n-p}}\left(\frac{2\pi k}{N}\right) \overset{n\to\infty}{\longrightarrow} \Phi_U\left(\frac{2\pi k}{N}\right) = \delta[k], \quad 0 \leqslant k < N \tag{24}$$

where $\Phi_{A|B}(u) = E(e^{\mathrm{ju A}}|B)$ is the characteristic function of A given B, and $\delta[k]$ is the characteristic function of the uniform discrete random variable U. Note that the LHS of (24) is actually a function of both k and the value of $x_{n-p}(= b)$; however, the latter argument is suppressed for the sake of simplicity, and is shown explicitly only when required. Proceeding as before, it can be shown that

$$\begin{aligned}\Phi_{e_n|x_{n-p}}\left(\frac{2\pi k}{N}\right) &= \Phi_{b_n|x_{n-p}}\left(\frac{-2\pi k}{N}\right)\cdot e^{\frac{-j2\pi k f_p x_{n-p}}{N}}\\ &\quad\cdot\prod_{r=0,r\neq p}^{n-p}\Phi_d\left(\frac{-2\pi k f_r}{N}\right)\end{aligned} \tag{25}$$

where $b_n = a_n - s_{n-p}$. Since (24) is true for $k = 0$, it remains to be shown that

$$\Phi_{b_n|x_{n-p}}\left(\frac{-2\pi k}{N}\right) e^{\frac{-j2\pi k f_p x_{n-p}}{N}} \prod_{\substack{r=0,\\ r\neq p}}^{n-p} \Phi_d\left(\frac{-2\pi k f_r}{N}\right) \xrightarrow{n\to\infty} 0 \qquad \forall 0 < k < N \quad (26)$$

if and only if the conditions of part (i) are satisfied. Note that the complex exponential term in the LHS of (26) is never zero and that b_n depends on the sequence s_n. Therefore, (26) is true for arbitrary s_n if and only if the product series term in the LHS of (26) converges to zero. The rest of the proof is similar to that of part (ii). ■

Notation:

$$S_L(q_n) \triangleq \frac{1}{L}\sum_{n=0}^{L-1} q_n.$$

Corollary:

i) The conditions of part (i) of Theorem 1 are sufficient for e_n to possess the following time-averaged properties **in probability**:

$$\lim_{L\to\infty} S_L(e_n) = M_e = 1/2 \quad (27)$$

$$\lim_{L\to\infty} S_L((e_n - M_e)x_{n-p}) = 0. \quad (28)$$

ii) The conditions of part (ii) of Theorem 1 are sufficient for e_n to possess the following time-averaged auto-covariance **in probability**:

$$\lim_{L\to\infty} S_L((e_n - M_e)(e_{n-p} - M_e)) = \sigma_{ee}^2\delta[p], \qquad \text{where } \sigma_{ee}^2 = (N^2-1)/12. \quad (29)$$

■

Proof: Equation (28) is proved below; the proofs of (27) and (29) are similar and hence, not presented here. For a fixed p, define $q_n \triangleq (e_n - M_e)x_{n-p}$. The goal is to prove that S_L (q_n) converges to 0 in probability as $L \to \infty$. Set $\eta_j = d_j$ and $\mu_j = s_j, \forall 0 \leqslant j < n$. Since e_n and x_n are bounded for all $n > 0$, it follows from (13) that q_n is a bounded, measurable function of $\eta_0, \ldots, \eta_n, \mu_0, \ldots, \mu_n$. The desired result follows from Lemma A2 presented in [22] if

$$E(q_n|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n) \xrightarrow{n-j\to\infty} 0, \quad n > j \geqslant 0 \quad (30)$$

in probability. Now, we obtain the equation shown at the bottom of the page. Since x_n is bounded for all n, (30) and the desired result follows if

$$E(e_n|x_{n-p}|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n) \xrightarrow{n-j\to\infty} M_e, \qquad n > j \geqslant 0 \quad (31)$$

in probability. To prove (31) note that (12) and (13) can be rewritten as

$$e_n = N/2 - \langle z_n + N/2\rangle_N$$

where

$$z_n = b_n + \sum_{m=0}^{n-j-1} c_m f_{n-j-1-m}$$

$$b_n \triangleq \sum_{m=0}^{n} s_m f_{n-m} + \sum_{m=0}^{j-1} d_m f_{n-m}$$

$$c_m = d_{m+j+1}. \quad (32)$$

Fixing $d_0, \ldots, d_j, s_0, \ldots, s_n$, and proceeding exactly as in the proof of part (i) of Theorem 1, with c_n and b_n respectively playing the role of d_n and a_n, it can be shown that

$$P_{e_n|x_{n-p}|d_0,\ldots,d_j,s_0,\ldots,s_n}(a) \xrightarrow{n-j\to\infty} P_U(a)$$

point wise. Since e_n and x_n are bounded, discrete random variables, it follows that the random variable

$$E((e_n|x_{n-p})|d_0,\ldots,d_j,s_0,\ldots,s_n) \xrightarrow{n-j\to\infty} E(U) = \frac{1}{2} = M_e$$

point wise and hence, in probability. ■

IV. LSB Dither in Popular SQDSMs

In many popular SQDSMs $\mathrm{STF}(z)$ is just a delay and $\mathrm{NTF}(z)$ is either a high-pass or a band-reject filter depending on the application. The former are referred to as low-pass SQDSMs and are widely used in DACs and fractional-N PLLs; the latter are called bandpass SQDSMs and are sometimes used in DACs for bandpass signal generation. Low-pass SQDSMs usually have

$$F(z) = z^{-L}(1-z^{-1})^{-L}; \quad G(z) = (1-z)^L - z^L \quad (33)$$

where L is a positive integer resulting in

$$\mathrm{STF}(z) = z^{-L}; \quad \mathrm{NTF}(z) = (1-z^{-1})^L. \quad (34)$$

Example realizations for $L = 1, 2,$ and 3[4] are shown in Fig. 5 and are referred to as the first-, second-, and third-order low-pass digital $\Delta\Sigma$ modulators, respectively. In contrast, bandpass SQDSMs often use $F\,(z)$ and $G\,(z)$ derived from their low-pass counterparts, e.g., (33), by applying the transformation, $z \to$

$$\begin{aligned}&E(x_{n-p}e_n|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n)\\ &\quad = E\left(x_{n-p}E\left(e_n|x_{n-p}\right)|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n\right)\\ &\quad = E(x_{n-p}E(e_n|x_{n-p}|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n)|\eta_0,\ldots,\eta_j,\mu_0,\ldots,\mu_n)\end{aligned}$$

[4] Note that the structures in Fig. 4 can be extended to realize low-pass digital $\Delta\Sigma$ modulators of higher order, L.

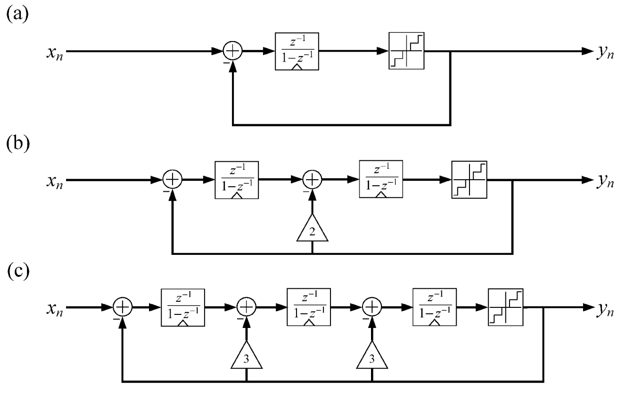
(a)
x_n
$\frac{z^{-1}}{1-z^{-1}}$
y_n
(b)
x_n
$\frac{z^{-1}}{1-z^{-1}}$
$\frac{z^{-1}}{1-z^{-1}}$
2
y_n
(c)
x_n
$\frac{z^{-1}}{1-z^{-1}}$
$\frac{z^{-1}}{1-z^{-1}}$
$\frac{z^{-1}}{1-z^{-1}}$
3
3
y_n

Fig. 5. Example of single-stage digital $\Delta\Sigma$ modulators. (a) First-order low-pass $\Delta\Sigma$ modulator. (b) Second-order low-pass $\Delta\Sigma$ modulator. (c) Third-order low-pass $\Delta\Sigma$ modulator.

$-z^2$. For example, the Lth-order bandpass SQDSM derived from the Lth-order low-pass SQDSM has

$$F(z) = (-1)^L z^{-2L}(1+z^{-2})^{-L}$$
$$G(z) = (1+z^2)^L - (-1)^L z^{2L} \quad (35)$$
$$\mathrm{STF}(z) = (-1)^L z^{-2L}; \quad \mathrm{NTF}(z) = (1+z^{-2})^L. \quad (36)$$

This section proves that the $F(z)$ in (33) and (35) satisfy the conditions of Theorem 1 provided $L > 1$ and requantizer step-sizes are positive powers of 2 ($N = 2^M, M \in \mathbb{Z}, M > 0$). It also proves that if a low-pass SQDSM satisfies the conditions of Theorem 1 then so does the bandpass SQDSM derived from it using the $z \to -z^2$ transformation. Other SQDSMs not of the type shown in (33)–(36) have also been reported. However, for the sake of brevity, the discussion is limited to the aforementioned cases.

Note that step-sizes that are positive powers of two are the most common choices owing to their ease of implementation. This is illustrated in Fig. 6 for the case of $N = 2^3$—the midtread requantization is achieved using just one 2-bit adder and some re-wiring.

A. First-Order Low-Pass SQDSM

In the first-order low-pass $\Delta\Sigma$ modulator, $F(z) = z^{-1}(1-z^{-1})^{-1}$, which does not satisfy the conditions of part (ii) of the Theorem 1 as shown below. The impulse response of $F(z) = z^{-1}(1-z^{-1})^{-1}$ is $f_r = u_{r-1}$, where u_r is the unit step function. So, for $k_1 = k \neq N/2$ and $k_2 = N-k$

$$\langle k_1 f_r + k_2 f_{r+p}\rangle_N = \langle k u_{r-1} + (N-k)u_{r+p-1}\rangle_N = 0 \quad \forall r \geqslant 1$$

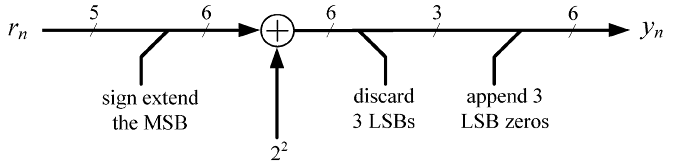

Fig. 6. Two's complement implementation of requantizer with $N = 2^M, M > 1$.

and so, neither condition 1 nor 2 is satisfied. Furthermore, $\langle k_2 f_r\rangle_N = k \neq N/2, \forall r > 0$ and so, condition 3 is also not satisfied. However, condition 1 of part (i) is satisfied since

$$\langle k f_r\rangle_N = \langle k u_{r-1}\rangle_N = k \quad \forall 0 < k < N, r \geqslant 1.$$

Consequently, e_n converges to a uniform random variable that is *asymptotically independent* of x_n; but (e_n, e_{n-p}) does not converge to (U, V) in distribution. The simulations described later in Section V and shown in Fig. 7 corroborate this claim.

B. Second–Order Low-Pass SQDSM

Theorem 2: Suppose that $F(z) = z^{-2}(1-z^{-1})^{-2}$. Then, the impulse response of $F(z)$ satisfies the conditions of part (ii) of the Theorem 1 for $N = 2^M$, where M is a positive integer.

Proof: The impulse response of $F(z)$ is $f_r = (r-1)u_{r-2}$. Condition 1 of part (ii) is satisfied for most positive values of p as shown below. The rest of the proof identifies the situations in which condition 1 is not satisfied and shows that in such situations, either condition 2 or condition 3 is satisfied.

In the following note that $0 \leqslant k_1, k_2 < N, (k_1, k_2) \neq (0,0)$, and $p > 0$. Substituting this in the expression for condition 1 results in (37), shown at the bottom of the page. So, condition 1 is not satisfied only if

$$(k_1+k_2)(r-1) + k_2 p = m_r N, \quad m_r \in \mathbb{Z}, r \geqslant r_0 \geqslant 2. \quad (38)$$

To determine the cases where condition 1 is not satisfied, suppose that (38) is true. Then, there exist two consecutive integers $r_*, r_*+1 > r_0$, which satisfy (38)

$$(k_1+k_2)(r_*-1) + k_2 p = m_a N$$
$$\text{and} \quad (k_1+k_2)(r_*) + k_2 p = m_b N, \qquad m_a, m_b \in \mathbb{Z}. \quad (39)$$

Subtracting the first of the two equations in (39) from the second and substituting the result back in either equation in (39) results in

$$k_1 + k_2 = N; \quad k_2 p = m_d N, \qquad m_d \in \mathbb{Z}; \quad m_d > 0. \quad (40)$$

So, condition 1 is true for all triplets (k_1, k_2, p) except those specified by (40). As shown below, these triplets satisfy condition 2 for $p = 2$ and condition 3 for the remaining p.

$$\langle k_1 f_r + k_2 f_{r+p}\rangle_N = \begin{cases} 0, & \forall r < 2-p \\ \langle k_2(r+p-1)\rangle_N, & \forall 2-p \leqslant r < 2 \\ \langle (k_1+k_2)(r-1) + k_2 p\rangle_N, & \forall r \geqslant 2 \end{cases} \quad (37)$$

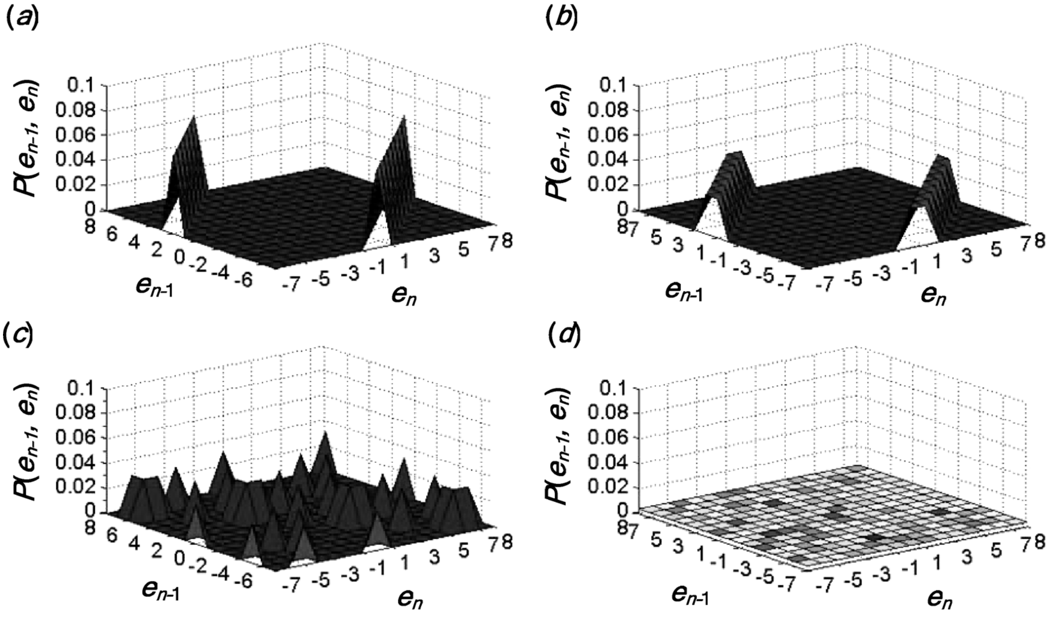

Fig. 7. Simulated jpmf of e_n and e_{n-1} in (a) first-order SQDSM without LSB dither, (b) first-order SQDSM with LSB dither, (c) second-order SQDSM without LSB dither, and (d) second-order SQDSM with LSB dither.

Case $p = 2$*:* The only triplet that satisfies (40) for $p = 2$ is $(k_1 = N/2, k_2 = N/2, p = 2)$. Choosing $s = 0 \neq p$ in (37) implies that the condition 2 is satisfied.

Case $p \neq 2$*:* Note that the second equality in (40) is satisfied only if $p = q * 2^{M-b}$ and $k_2 = 2^b\,(2l+1)$ where q is a positive integer, and b and l are non-negative integers such that $b < M$. The integer $t = 1 + N/2^{b+1}$ satisfies the condition 3 if $t < p$

$$\langle k_2 f_t \rangle_N = \langle k_2(t-1) \rangle_N = (2l+1)\frac{N}{2} \bmod N = \frac{N}{2}.$$

Since

$$1 + \frac{N}{2^{b+1}} = 1 + \frac{2^{M-b}}{2} < q \cdot 2^{M-b} \qquad \forall q > 1$$

$t < p$ except when $q = 1$ and $b = M - 1$. Note that $q = 1$ and $b = M - 1$ corresponds to the triplet $(k_1 = N/2, k_2 = N/2, p = 2)$ which has shown to satisfy condition 2. Therefore, the triplets of (40) satisfy condition 3 for $p \neq 2$. ■

Remark: Note that the above proof shows that the impulse response, f_n, of $F(z) = z^{-2}(1 - z^{-1})^{-2}$ does **not** satisfy condition 1 of part (ii) for all $p > 0$, i.e., it does not satisfy the sufficient conditions presented in [10]. The additional conditions 2 and 3 are required to prove that e_n of the second-order low-pass SQDSM possesses the desired time-averaged statistics in probability. It is interesting to note that f_n satisfies the equivalent of condition 1 for all $p > 0$ for the counterpart analog $\Delta\Sigma$ modulators presented in [9].

The proven result extends to SQDSMs employing any other $G(z)$ and even with a slightly different $F(z)$ namely, $F(z) = z^{-R}(1 - z^{-1})^{-2}$, where R is an integer. The impulse response of $F(z)\,G(z)$ has to be an integer sequence and the constraints of no-overload and $N = 2^M$ need to be satisfied. The extension to other $G(z)$ follows from the discussion in Section III [specifically, (12) and (13)]. That the result is true for $R \neq 2$ can be seen by redrawing the SQDSM with $F(z) = z^{-2}(1 - z^{-1})^{-2}$ and two additional delay elements namely, $z^{-(R-2)}$ each: one outside the modulator's feedback loop and the other merged with $G(z)$.

C. Lth-Order Low-Pass SQDSM

It is shown below that $F_L\,(z)$ satisfies the *impulse response conditions* for all integers $L > 2$.

Theorem 3: The impulse response of $F(z) = z^{-L}(1 - z^{-1})^{-L}$ satisfies the condition 1 of part (ii) of Theorem 1 for all integers $p > 0, L > 2$, and $N = 2^M$, where M is a positive integer.

Proof: Define $F_S(z) \triangleq z^{-S}(1 - z^{-1})^{-S}$ for positive integers S and suppose $f_{S,r}$ is its impulse response. Then, $f_{S,r}$ and f_{S-1}, r are related by the difference equation

$$f_{S,r} - f_{S,r-1} = f_{S-1,r}. \tag{41}$$

The result will be proved by the mathematical induction on the index S using (41).

$\boldsymbol{S = 3}$*:* Suppose that condition 1 is not satisfied for some $p > 0$. Since the impulse response of $F_3\,(z)$ is

$$f_{3,r} = \frac{1}{2}(r-2)(r-1)u_{r-3}$$

for that particular p and some integer r_0,

$$k_1\frac{(r-2)(r-1)}{2} + k_2\frac{(r+p-2)(r+p-1)}{2} = m_r N.$$
$$m_r \in \mathbb{Z}; \quad \forall r \geqslant r_0 \geqslant 3. \tag{42}$$

The set of equations in (42), in the three unknowns k_1, k_2, and p, can be reduced by considering these equations for any three consecutive values of $r \geqslant r_0$, resulting in the following:

$$k_1 + k_2 = m_a N; \quad k_2 p = m_b N, \quad \text{and} \quad p = 1 + 2m_c,$$
$$\text{where } m_a, m_b, m_c \in \mathbb{Z}. \tag{43}$$

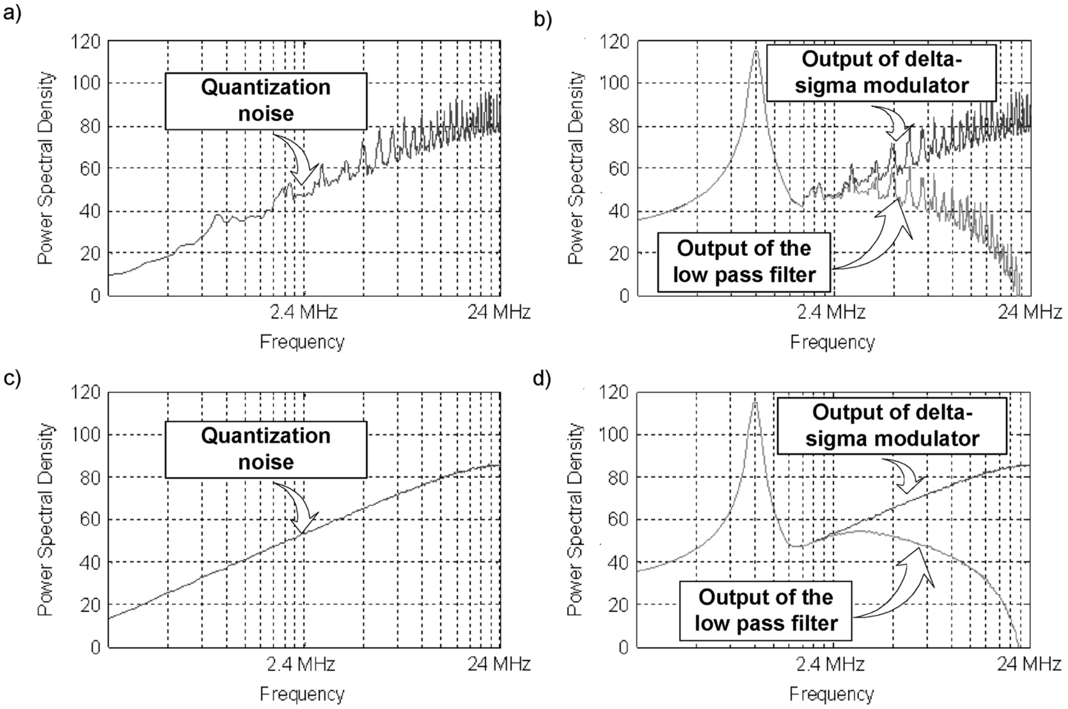

Fig. 8. Illustration of the effect and utility of LSB dither in an oversampled DAC.

The equations in (43) can not all be simultaneously true: since p is odd, and N is greater than one and a power of 2, k_2 has to be a multiple of N to satisfy the second equation; but this contradicts the first equation in (43). So it is proved by contradiction that $f_{3,n}$ satisfies condition 1.

***S* + 1 Given *S*:** It is proved below by contradiction that if $f_{S,r}$ satisfies condition 1 for a given $p > 0$, so does $f_{S+1,r}$.

For a given $p > 0$, and some integers k_1, k_2 such that $0 \leqslant k_1, k_2 < N$ and $k_1 + k_2 \neq 0$, suppose that $f_{S,r}$ satisfies condition 1, but $f_{S+1,r}$ does not. Then $\langle k_1 f_{S,r} + k_2 f_{S,r+p} \rangle_N$ does not converge to zero as $r \to \infty$ but for some integer r_0,

$$k_1 f_{S+1,r} + k_2 f_{S+1,r+p} = mN, \qquad m \in \mathbb{Z}; \qquad \forall r \geqslant r_0. \tag{44}$$

For all pairs of consecutive integers $r_* + 1$ and $r_* + 2$ where $r_* \geqslant r_0$, (44) implies the following:

$$\begin{aligned} k_1 f_{S+1,r_*+1} + k_2 f_{S+1,r_*+1+p} &= m_1 N \\ k_1 f_{S+1,r_*+2} + k_2 f_{S+1,r_*+2+p} &= m_2 N. \end{aligned} \tag{45}$$

Subtracting the first of the two equations in (45) from the seconds results in

$$\begin{aligned} k_1(f_{S+1,r_*+1} - f_{S+1,r_*}) + k_2(f_{S+1,r_*+p+1} - f_{S+1,r_*+p}) \\ = (m_2 - m_1)N. \end{aligned} \tag{46}$$

Substituting (41) in (46) results in

$$k_1 f_{S,r_*} + k_2 f_{S,r_*+p} = m_3 N, \quad m_3 \in \mathbb{Z}, r_* \geqslant r_0. \tag{47}$$

So, it is proved by contradiction that $f_{S+1,r}$ satisfies condition 1 if $f_{S,r}$ does. ■

D. Bandpass SQDSMs

To the best of the authors' knowledge, there has been no published work analyzing the effects of dither in bandpass $\Delta\Sigma$ modulators, digital or analog except for [9]. As mentioned before, bandpass $\Delta\Sigma$ modulators are typically derived from low-pass $\Delta\Sigma$ modulators by applying the transformation, $z \to -z^2$, on all the concerned filters. Theorem 4 below shows that the $F(z)$ of such bandpass SQDSMs satisfy the conditions of part (ii) of Theorem 1 for $N = 2^M, M \in \mathbb{Z}, M > 0$.

Theorem 4: Suppose the impulse response of $F(z)$ satisfies the conditions of part (ii) of Theorem 1. Then, so does the impulse response of

$$H(z) = F(z)|_{z=-z^2}.$$

Proof: Suppose f_r were the impulse response of $F\ (z)$. Then, the impulse response of $H\ (z)$ is

$$h_r = \begin{cases} (-1)^{r/2} f_{r/2}, & r \text{ even} \\ 0, & r \text{ odd.} \end{cases}$$

Case p Odd: Suppose that $p = 2q + 1$ where q is a nonnegative integer. Then,

$$\begin{aligned} &k_1 h_r + k_2 h_{r+p} \\ &\quad = \begin{cases} (-1)^l k_1 f_l, & r = 2l \\ (-1)^{q+l+1} k_2 f_{q+l+1}, & r = 2l+1 \end{cases} \forall l \geqslant 0 \end{aligned} \tag{48}$$

where l is an integer. By assumption, f_l satisfies the conditions of part (ii) and so, for the pair $(0, k_1)$, where $k_1 \neq 0$, either $\langle k_1 f_l \rangle_N$ does not converge to 0 as $l \to \infty$ or equals $N/2$ for some $l = s\ (p) \neq p$. While the former implies that the even terms of the LHS of (48) (modulo- N) do not converge, the latter implies that one of the even terms of the LHS of equals $N/2$ (modulo-N) for some $r = 2s \neq p$. Consequently, h_r satisfies conditions of part (ii) for odd p.

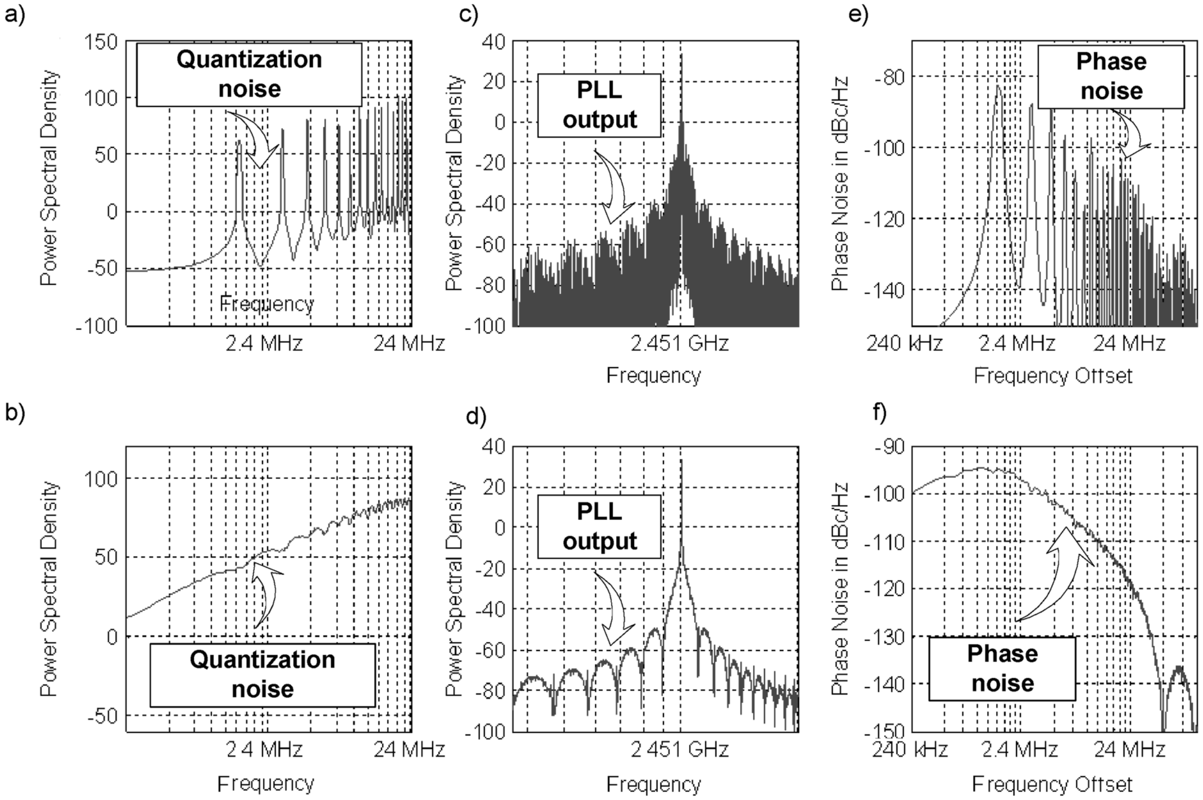

Fig. 9. Illustration of the effect and utility of LSB dither in a fractional-N PLL.

Case $p = 2q, q$ *Even,* $q \geqslant 0$*:* In this case

$$
\begin{aligned}
&k_2 h_{r+p} \\
&\quad = \begin{cases} (-1)^l (-1)^q k_2 f_{l+q}, & r = 2l, \\ 0, & r = 2l+1 \end{cases} \\
&k_1 h_r + k_2 h_{r+p} \\
&\quad = \begin{cases} (-1)^l (k_1 f_l + (-1)^q k_2 f_{l+q}), & r = 2l \\ 0, & r = 2l+1 \end{cases}
\end{aligned}
\tag{49}
$$

where l is a non-negative integer. Fix non-negative integer pair $(k_1, k_2) \neq (0,0)$ such that $0 \leqslant k_1, k_2 < N$. By assumption, f_l satisfies at least one of the three conditions of part (ii) for the given q and (k_1, k_2). If f_l satisfies the condition 1 then

$$\forall l > 0, \qquad \exists s > l \ \text{ s.t. } \langle k_1 f_s + k_2 f_{s+q} \rangle_N \neq 0.$$

Since $\langle -x \rangle_N \neq 0$ if $\langle x \rangle_N \neq 0$, it follows that

$$\forall l > 0, \qquad \exists s > l \ \text{ s.t. } \langle (-1)^s (k_1 f_s + k_2 f_{s+q}) \rangle_N \neq 0.$$

Substituting in (49) implies that the even terms of $\langle k_1 h_r + k_1 h_{r+p} \rangle_N$ do not converge to 0 as $r \to \infty$ for the given (k_1, k_2). So, h_r satisfies the condition 1. Instead, if f_l satisfies the conditions 2 or 3 then

$$
\begin{aligned}
&\exists s \neq q \ \text{ s.t. } \langle k_1 f_s + k_2 f_{s+q} \rangle_N = N/2 \\
\text{or } &\exists t < q \ \text{ s.t. } \langle k_2 f_s \rangle_N = N/2.
\end{aligned}
$$

Note that $\langle -x \rangle_N = N/2$ implies $\langle x \rangle_N = N/2$. Furthermore, since r = 2 s and $p = 2q$, the existence of $s \neq q$ implies the existence of $r = 2s \neq p$ and the existence of $t < q$ implies the existence of $r = 2t < p$. So, it follows that h_r satisfies the conditions 2 or 3 for the particular (k_1, k_2) pair. Therefore, for $p = 2q, q$ even, h_r satisfies at least one of the three conditions of part (ii) of Theorem 1.

Case $p = 2q, q$ *Odd,* $q > 0$*:* Equation (49) is applicable in this case too. Fix (k_1, k_2) as before. Note that

$$\langle k_1 f_l + k_2 (-1)^q f_{l+q} \rangle_N = \langle k_1 f_l + (N - k_2) f_{l+q} \rangle_N.$$

Note that f_l satisfies at least one of the conditions of part (ii) of Theorem 1 for the pair $(k_1, N - k_2)$. Applying the same arguments as in the case of even q for $(k_1, N - k_2)$ instead of (k_1, k_2) proves that h_r satisfies at least one of the conditions of part (ii) of Theorem 1. ■

V. Simulations

Computer simulation results of the first- and second-order low-pass SQDSMs in stand-alone configuration and in DAC and fractional-N PLL applications are presented in this section to demonstrate the results presented in Section IV and the utility of LSB dither in these applications.

In one set of simulations, first-order and second-order low-pass SQDSMs [see Fig. 4(a) and (b)] with $N = 16$ and a constant input were simulated without and with LSB dithering. The jpmf of e_n and e_{n-1}, i.e., $\mathrm{P}(e_n, e_{n-1})$, in the first-order SQDSM without and with LSB dithering are plotted in Fig. 7(a) and (b) respectively; whereas, $\mathrm{P}(e_n, e_{n-1})$ in the second-order SQDSM without and with LSB dithering are plotted in Fig. 7(c) and (d). Note that $\mathrm{P}(e_n, e_{n-1})$ is expected to be a constant if e_n and e_{n-1} are asymptotically independent and uniform. As is evident from the figures, it is the case only in the second-order

SQDSM with LSB dithering; LSB dithering is not successful in making $\mathrm{P}(e_n, e_{n-1})$ a constant in the first-order SQDSM. These simulations confirm the results from Section IV-A and B.

In the next set of simulations, the second-order low-pass SQDSM [see Fig. 4(b)] is simulated in the context of a DAC and a fractional-N PLL. In both applications, the SQDSM employs a 16-bit x_n, a requantizer with $N = 16384$, and 5 output levels. The output of the SQDSM for a sinusoidal s_n was fed to a 4-bit DAC[5] and the result was low-pass filtered by a third-order Butterworth filter with a -3 dB bandwidth of 6 MHz as shown in Fig. 1(a). The SQDSM and the DAC were clocked at 48 MHz. Fig. 8(a) and (c) depicts the simulated PSD of the SQDSMs total quantization noise for a 2 MHz s_n without and with *LSB dithering* respectively. Fig. 8(b) and (d) depicts the PSDs of the outputs of the SQDSM and the low-pass filter without and with *LSB dithering* respectively. The absence of the spikes and the 20 dB/decade high-pass shape of the quantization noise PSD in Fig. 8(b) confirm the claims of Theorem 2. The elimination of tones is critical to the DAC's use in applications such as high-fidelity audio or video. Without *LSB dithering* the spikes are perceptible in such applications even after the filtering provided by $\mathrm{NTF}(z)$ and the low-pass filter as shown in Fig. 8(b).

The output of the SQDSM modulator for a constant s_n is fed to a frequency divider of a fractional-N PLL operating on a 48 MHz reference as shown in Fig. 1(b). Successive frequency division ratios are chosen according to the 4-bit output sequence, y_n. The instantaneous frequency of the PLL's output tracks the changes in the $\Delta\Sigma$ modulator's output thereby allowing frequency synthesis of arbitrary precision at the expense of conversion of the SQDSM's total quantization noise into PLL phase noise. The 16-bit s_n was chosen to be a constant such that the PLL's output has a fixed frequency of 2.451 GHz. Fig. 9(a) and (b) depicts the PSD of the SQDSM's total quantization noise without and with *LSB dithering* respectively. Fig. 9(c) and (d) depicts the PSD of the PLL's output without and with *LSB dithering* respectively; Fig. 9(e), and (f) depicts the PSD of the output's phase noise without and with *LSB dithering* respectively. The elimination of spikes in the PSDs enables the use of this fractional-N PLL for frequency synthesis in many high-performance applications such as wireless transceivers.

VI. Conclusion

Necessary and sufficient conditions for the quantization noise in *1-bit dithered* non-overloading SQDSMs to be asymptotically uniformly distributed and independent of delayed versions of itself and the input sequence have been presented. They are also sufficient to ensure that the quantization noise is uniform, white, and uncorrelated with the input in time-averaged sense. Many popular non-overloading low-pass and bandpass digital $\Delta\Sigma$ modulators have been shown to satisfy the presented conditions.

References

[1] R. S. Gabor and C. Temes, *Understanding Delta–Sigma Data Converters*. New York: Wiley, 2005.

[2] J. C. Candy and G. C. Temes, "Oversampling methods for A/D and D/A conversion," in *Oversampling Delta–Sigma Data Converters Theory, Design and Simulation*. New York: IEEE Press, 1992, pp. 1–25.

[3] M. Annovazzi, "A low-power 98-dB multibit audio DAC in a standard 3.3–V 0.35-mm CMOS technology," *IEEE J. Solid-State Circuits*, vol. 37, no. 7, pp. 825–834, Jul. 2002.

[4] S. Willingham, "An integrated 2.5-GHz $\Delta\Sigma$ frequency synthesizer with 5-μs settling and 2 Mb/s closed loop modulation," in *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.* , Feb. 2000, vol. 43, pp. 200–201.

[5] B. De Muer and M. S. J. Steyaert, "A CMOS monolithic $\Sigma\Delta$—controlled fractional-N frequency synthesizer for DCS-1800," *IEEE J. Solid-State Circuits*, vol. 37, no. 7, pp. 835–844, Jul. 2002.

[6] S. Pamarti, L. Jansson, and I. Galton, "A wide-band 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.

[7] S. R. Norsworthy, R. Schreier, and G. C. Temes, *Delta–Sigma Data Converters Theory, Design, and Simulation*. New York: IEEE Press, 1996, pp. 75–121.

[8] 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. 500–514, May 1991.

[9] I. Galton, "Granular quantization noise in a class of delta–sigma modulators," *IEEE Trans. Inf. Theory*, vol. 40, no. 3, pp. 848–859, May 1994.

[10] I. Galton, "One-bit dithering in delta–sigma modulator-based D/A conversion," in *Proc. IEEE Int. Symp. on Circuits Syst.*, 1993, pp. 1310–13.

[11] R. M. Gray, W. Chou, and P. W. Wong, "Quantization noise in a single-loop sigma-delta modulation with sinusoidal inputs," *IEEE Trans. Commun.*, vol. , no. 9, pp. 956–968, Sep. 1989.

[12] P. W. Wong and R. M. Gray, "Sigma-delta modulation with i.i.d. Gaussian inputs," *IEEE Trans. Inf. Theory*, vol. 36, no. 4, pp. 784–788, Jul. 1990.

[13] N. He, F. Kuhlmann, and A. Buzo, "Multi-loop sigma-delta modulation," *IEEE Trans. Inf. Theory*, vol. 38, no. 3, pp. 1015–1028, May 1992.

[14] C. S. Gunturk and N. T. Thao, "Refined error analysis in second-order sigma-delta modulation with constant inputs," *IEEE Trans. Inf. Theory*, vol. 50, no. 5, pp. 839–860, May 2004.

[15] R. M. Gray, "Spectral analysis of quantization noise in single-loop sigma-delta modulation with dc inputs," *IEEE Trans. Commun.*, pp. 588–599, Jun. 1989.

[16] M. Kozak and I. Kale, "Rigorous analysis of delta–sigma modulators for fractional-N PLL frequency synthesis," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 6, pp. 1148–1162, Jun. 2004.

[17] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 10, pp. 442–448, Oct. 1997.

[18] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 805–811, May 1993.

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

[20] P. Billingsley, *Probability and Measure*. New York: Wiley, 1986.

[21] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.

[22] I. Galton, "Granular quantization noise in the first-order delta–sigma modulator," *IEEE Trans. Inf. Theory*, vol. 39, no. 6, pp. 1944–1956, Nov. 1993.

Sudhakar Pamarti (S'98–M'03) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, San Diego, in 1999 and 2003, respectively.

Since 2005, he has been an Assistant Professor of Electrical Engineering at the University of California at Los Angeles (UCLA), where he teaches and conducts research in the design of mixed-signal circuits for wireless and wire-line communication systems. Prior to joining UCLA, he worked with Rambus Inc. (2003–2005) designing high-speed chip-to-chip electrical interfaces and with Hughes Software Systems (1995–1997) developing real-time embedded software for a wireless communication system. His current research interests include mixed-signal circuits, signal processing, and digital communications.

[5] Note that only five of the 4-bit DAC's input levels are used.

Jared Welz (S'98–M'02) received a B.S.E.E. degree from University of California at Irvine, the M.S.E.E. degree from University of California at Los Angeles (UCLA) with an emphasis in communication theory, and the Ph.D. degree in electrical engineering from University of California at San Diego (UCSD) in 1993, 1994, and 2002, respectively.

He worked as a Cell Site Engineer for AirTouch International (now Vodafone AirTouch), Pacific Bell Wireless (now Cingular Wireless), and L.A. Cellular (now Cingular Wireless) between 1994 and 1997. From 1997 to 2002, he was a Graduate Student Researcher at UCSD. After graduating, he designed mixed-signal circuits at Northrop Grumman Space Technology (formerly TRW) from 2002 to 2005, and has since been employed as a Wireless Systems Engineer at Broadcom Corporation, Irvine, CA. His interests include mixed-signal circuits, signal processing, communication systems, and probability theory.

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, 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 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 the 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, and as a member of the IEEE International Solid-State Circuits Conference Technical Program Committee.