Stochastic Analysis of the $\Sigma\Delta$ Modulator and Differential Pulse Code Modulator

Rajesh Sharma, *Member, IEEE*, James A. Bucklew, *Senior Member, IEEE*,
and William A. Sethares, *Member, IEEE*

***Abstract*— One of the most popular systems for performing high resolution analog to digital conversion is the $\Sigma\Delta$ modulator. Though common in applications, theoretical analysis of the $\Sigma\Delta$ modulator is difficult due to the presence of a discontinuous quantizer in the modulator. This paper presents asymptotic results regarding the statistical behavior of the $\Sigma\Delta$ modulator when inside the loop dithering is utilized. In some recent papers examining the stochastic behavior of the $\Sigma\Delta$, it was shown (via simulations) that the input to the quantizer can be accurately modeled as a stationary Gaussian process. Our analysis shows that both the input to the quantizer and the quantization noise are asymptotically stationary Gaussian processes, under mild assumptions on the input and the dither process. The results of this paper are derived by letting the quantizer stepsize approach zero, and the analytical approach is related to the stochastic analysis of adaptive filtering algorithms. Our analysis is valid for a large collection of stochastic input signals, including ARMA processes. Furthermore, previous stochastic analysis assumed that the quantizer never overloaded, while the present analysis does not make this assumption. It is also shown that analysis of the Differential Pulse Code Modulator is in fact analogous to the analysis of the $\Sigma\Delta$ modulator. Simulation results presented for the first-order $\Sigma\Delta$ modulator and two second-order $\Sigma\Delta$ modulators demonstrate the practicality of the analysis.**

***Index Terms*—Adaptive algorithms, analog to digital converters, differential pulse code modulators, noise shaping systems, oversampling data converters, sigma–delta ($\Sigma\Delta$) modulators.**

I. Introduction

THE PROCESS of amplitude quantization is fundamental in many modern communication systems. This is a direct consequence of the ease with which digital data can be stored and transmitted. Various systems have been proposed to quantize continuous amplitude signals into discrete amplitude sequences. Oversampling techniques for analog to digital conversion have become increasingly popular ever since the single bit delta modulation (DM) system was proposed in [14]. The popularity of these systems is directly attributable to the advances made in integrated circuit technology.

DM coders are examples of a more general class of predictive coding systems called differential pulse code modulators (DPCM), which are based on exploiting intersample dependencies in the input [10]. In DPCM systems, a prediction error (rather than the original data sequence) is quantized. Sigma–delta modulator ($\Sigma\Delta$) systems are examples of noise shaping coders [8] and have received a great deal of attention since they were first introduced more than 30 years ago [9] as a means of high resolution analog to digital conversion. $\Sigma\Delta$ modulators are well suited for VLSI implementation and are robust to circuit imperfections. As a result $\Sigma\Delta$ systems are finding widespread use in signal processing applications. recent years there has been growing interest in the use of $\Sigma\Delta$ for noise shaping in nonoversampling systems.

Exact analysis of both DPCM and $\Sigma\Delta$ is difficult for all but the simplest cases. This is primarily due to the presence of a nonlinearity (quantizer) in both systems and the fact that both systems incorporate feedback. Analyzes based on modeling the nonlinear quantizer as an uniform, signal independent, additive white noise source are carried out in numerous papers. For example, the reader is referred to [2] which contains papers covering many aspects of $\Sigma\Delta$ modulator design and analysis. Though this white noise assumption makes the analysis of these systems tractable using linear techniques, it has been shown to be lacking in simulations for lower-order systems. The assumption fails in practical systems because the quantization noise is a function of the input signal. Another approach used in the analysis of these nonlinear systems is the use describing functions. The strengths and shortcomings of this approach are detailed in [7]. In [5], analysis of the quantization noise in a first-order $\Sigma\Delta$ is carried out under the assumption that the input sequence is constant. Then, using ergodic theory, it is shown that the $\Sigma\Delta$ modulator can be modeled as a linear system in a different space. This linear model is then used to derive properties of the quantization noise. An exact analysis of the moments and spectrum of the quantization noise in the first-order $\Sigma\Delta$ with constant inputs can be found in [6] and in this paper it is shown that the quantization noise is not white. The analysis in [6] provides extensions of the results found in [1]. Analysis of higher-order $\Sigma\Delta$ is carried out in [4] under the akssumption that the input to the modulator is dithered. The results of [4] apply to constant, sinusoidal and more generally quasistationary inputs. In [3], stochastic analysis of the $\Sigma\Delta$ was carried out under the assumption that the input to the single bit quantizer, in the $\Sigma\Delta$, can be modeled as a stationary Gaussian process. Simulations were presented to demonstrate the accuracy of the Gaussian assumption.

The approach adopted in this paper is based on an asymptotic analysis of adaptive filtering algorithms [11]–[13]. sta-

Manuscript received October 26, 1995; revised Noember 28, 1996. This paper was recommended by Associate Editor K. K. Parhi.

R. Sharma is with the Advanced Information Systems Group, Envirnomental Research Institue of Michigan, Ann Arbor, MI 48113 USA.

J. A. Bucklew and W. A. Sethares are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706 USA.

Publisher Item Identifier S 1057-7130(97)07646-5.

tistical properties of the quantization noise and the quantizer input in the single bit ΣΔ modulator are derived under the assumption that inside the loop dithering is utilized. In our analysis, the distribution of the dither is used to smooth the discontinuous quantization function. The results are asymptotic in the sense that they are exact only as the quantizer stepsize approaches zero. Furthermore, in the analysis, it is not assumed that the quantizer is never overloaded. Both the quantization noise process and the quantizer input process are shown to be asymptotically stationary Gaussian processes and the results are valid for a large class of stochastic input signals with arbitrary distribution (not necessarily Gaussian).

In Section II, brief introductions to ΣΔ and DPCM modulators are presented, and the evolution of both the quantization noise and the quantizer input process are described in terms of recursive equations. In Section III, the central limit result of [13] is applied to the recursive equations for the ΣΔ modulator. Application of the central limit result to the recursive equations describing DPCM systems can be carried out in a similar manner and is thus omitted. In Section IV, simulation results for first-order single bit ΣΔ modulator and two second-order single bit ΣΔ modulators are presented. Section V presents conclusions and extensions for future work. The Appendix contains technical proofs.

II. Preliminaries and Background

Pulse code modulation (PCM) is the most widely used digital coding system. PCM systems are based on instantaneous quantization of the current input sample using a memoryless quantizer. If $x_k \in \mathbb{R}$ is the input to a memoryless quantizer, then the output is $Q(x_k)$, where $Q(\cdot)$ is an increasing right continuous function which maps $\mathbb{R}$ into a finite set $\{a_1, a_2, \cdots, a_K\} \subset \mathbb{R}$. This paper assumes $Q(x) = \mu\mathrm{sgn}(x)$, where

$$\mathrm{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

and μ is some positive number, called the quantizer stepsize. Practical PCM systems are characterized by a high ratio of bits per input sample and are particularly inefficient if the continuous amplitude sequence $\{x_k\}$ contains intersample dependencies. quantization error or quantization noise q_k is defined as $q_k = \mathrm{sgn}(x_k) - x_k$. Considerable research been devoted to the calculation of the statistical properties of this discrete time process under various assumptions (see, for example, the list of references in [7]).

Differential Pulse Code Modulators

DPCM systems are designed to exploit the dependencies in the continuous amplitude sequence $\{x_k\}$ and yield quantizers of superior performance [10, ch. 6]. In DPCM systems, dependencies are removed prior to quantization and the prediction error is quantized rather than the original sequence x_k. Refer to the block diagram shown in Fig. 1. The general DPCM system estimates x_k based on past information using a FIR filter $\boldsymbol{H}$ with impulse response $\{0, h_1, \cdots, h_M\}$. Then quantizes the error $d_k = x_k - \hat{x}_k$ plus the dither D_k. Let $q_k = d_k - u_k$. This section shows that the evolution of $\{q_k\}$ and $\{d_k\}$ can be described in terms of a recursive equation of the form

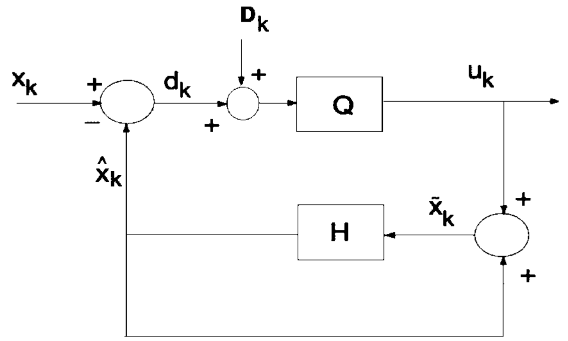

Fig. 1. DPCM coder.

$$W_{k+1} = AW_k + \mu G(W_k, Y_k, U_{k+1}, \mu). \tag{1}$$

Equation (1) is in the form needed to apply [13, Th. 2] to yield the results of Section III.

For the general DPCM system shown in Fig. 1, assume that the input x_k is premultiplied by μ, a positive number, and the quantization function is $\mu\mathrm{sgn}(x)$. Thus, both the input and the quantization function are scaled by μ. The dither D_k is assumed to be a sequence of i.i.d random variables. Observe that

$$\begin{aligned} q_{k+1} &= d_{k+1} - u_{k+1} \\ &= d_{k+1} - \mu\,\mathrm{sgn}(d_{k+1} + D_{k+1}) \\ &= \mu x_{k+1} - \hat{x}_{k+1} - \mu\,\mathrm{sgn}(\mu x_{k+1} - \hat{x}_{k+1} + D_{k+1}). \end{aligned}$$

Since

$$\begin{aligned} \tilde{x}_k &= u_k + \hat{x}_k \\ &= \mu x_k - d_k + u_k \\ &= \mu x_k - q_k \end{aligned}$$

the quantization noise $\{q_k\}$ satisfies the recursive equation

$$\begin{aligned} q_{k+1} = \sum_{j=1}^{M} h_j q_{k+1-j} - \mu \Bigg[\mathrm{sgn}\Bigg(\mu \Bigg(x_{k+1} - \sum_{j=1}^{M} h_j x_{k+1-j} \Bigg) \\ + \sum_{j=1}^{M} h_j q_{k+1-j} + D_{k+1} \Bigg) \\ - \Bigg(x_{k+1} - \sum_{j=1}^{M} h_j x_{k+1-j} \Bigg) \Bigg]. \end{aligned}$$

Note that the above recursion can be put in the form of (1) as follows. Let $W_k = [q_k, q_{k-1}, \cdots, \cdots, q_{k-M+1}]^T, Y_{k-1} = x_{k+1} - \Sigma_{j=1}^{M} x_{k+1-j}$, $U_k = D_k$ and

$$A = \begin{pmatrix} h_1 & h_2 & h_3 & \cdots & h_M \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{2}$$

Then

$$W_{k+1} = AW_k + \mu G(W_k, Y_k, U_{k+1}, \mu) \tag{3}$$

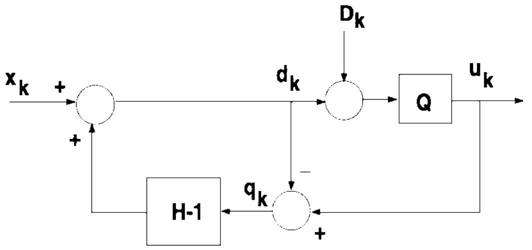

Fig. 2. Block diagram of the sigma–delta modulator.

where

$$G(w,y,U,\mu) = -[\text{sgn}(U+\mu y+h^T w)-, 0, \cdots, 0]^T$$

with $w = (w_1, \cdots, w_M)^T, y \in \mathbb{R}, U \in \mathbb{R}$ and $h = (h_1, \cdots, h_M)^T$. Similarly, it can be shown that the process $\{d_k\}$ satisfies the recursive equation

$$\begin{aligned} d_{k+1} = &\sum_{j=1}^{M} h_j d_{k+1-j} - \sum_{j=1}^{M} h_j u_{k+1-j} \\ &+ \mu\left(x_{k+1} - \sum_{j=1}^{M} h_j x_{k+1-j}\right). \end{aligned}$$

With $W_k = [d_k, d_{k-1}, \cdots, d_{k-M+1}]^T$,

$$Y_k = [x_{k+1}, x_k, \cdots, x_{k-M+1}, D_{k-1}, D_{k-2}, \cdots, D_{k-M+1}]^T$$

and $U_{k+1} = D_k$ we get

$$W_{k+1} = AW_k + \mu G(W_k, Y_k, U_{k+1}) \tag{4}$$

with G defined in a similar manner as before.

$\Sigma\Delta$ Modulator

Fig. 2 shows the error feedback topology of a $\Sigma\Delta$ modulator. Referring to Fig. 2, $\boldsymbol{H}$ has impulse response $\{1, h_1, \cdots, h_M\}$ and is called the noise transfer function, since it filters the quantization noise. $\boldsymbol{H} - \mathbf{1}$ denotes a FIR filter with impulse response $\{0, h_1, h_2, \cdots, h_M\}$ and M is called the order of the $\Sigma\Delta$. A typical example of a low pass $\boldsymbol{H}$ has transfer function $H(z) = (1 - z^{-1})^M$. $\Sigma\Delta$ modulators belong to the family of noise shaping coders. This is apparent by noting that the output u_k can be expressed as

$$u_k = x_k + \boldsymbol{H}[\{e_k\}].$$

Thus, the output is simply the sum of the input x_k and the quantization noise filtered by $\boldsymbol{H}$.

Next, the quantization noise process in the $\Sigma\Delta$ and the quantizer input process $d_k = x_k + f_k$ are expressed in terms of a recursive equation where f_k is the output of the FIR filter. Suppose that the input x_k in Fig. 2 is premultiplied by μ and a single bit quantizer $\mu\text{sgn}(\cdot)$ is used. Then

$$\begin{aligned} u_k &= \mu\,\text{sgn}(d_k + D_k) \\ q_k &= u_k - d_k \\ d_k &= \mu x_k + h_1 q_{k-1} + h_2 q_{k-2} + \cdots + h_M q_{k-M}. \end{aligned}$$

So

$$\begin{aligned} q_{k+1} = &\,\mu\,\text{sgn}(\mu x_{k+1} + D_{k+1} + h_1 q_k + h_2 q_{k-1} \\ &+ \cdots + h_M q_{k+1-M}) \\ &- \mu x_{k+1} - h_1 q_k - h_2 q_{k-1} - \cdots - h_M q_{k+1-M}. \end{aligned}$$

Define $W_k = [q_k, q_{k-1}, \cdots, q_{k-M+1}]^T$,

$$A = \begin{pmatrix} -h_1 & -h_2 & -h_3 & \cdots & -h_M \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \tag{5}$$

$Y_k = x_{k+1}, U_k = D_k$ and

$$\begin{aligned} G(w,y,U,\mu) = &[\text{sgn}\,(\mu y + U + h_1 w_1 + h_2 w_2 \\ &+ \cdots + h_M w_M) - 0, \cdots, 0]^T. \end{aligned}$$

Then

$$W_{k+1} = AW_k + \mu G(W_k, Y_k, U_{k+1}, \mu). \tag{6}$$

Similarly

$$\begin{aligned} d_{k+1} = &-\sum_{j=1}^{M} h_j d_{k+1-j} \\ &+ \mu\left(\sum_{j=1}^{M} h_j \text{sgn}\,(d_{k+1-j} + D_{k+1-j}) + x_{k+1}\right) \end{aligned}$$

which can be put in a form of (4) with A as in (5), $W_k = [d_k, d_{k-1}, \cdots, d_{k-M+1}]^T, Y_k = [x_{k+1}, D_{k-1}, D_{k-2}, \cdots, D_{k-M+1}]^T, U_{k+1} = D_k$ and G defined in an obvious way.

In the next section the asymptotic behavior of the quantization noise process $\{q_k\}$ and the quantizer input process $\{d_k\}$ in the $\Sigma\Delta$ is analyzed using the weak convergence results of [13]. Results for the DPCM systems are similar and thus omitted.

III. Analysis

In [13], asymptotic results for (4) were presented. These results can be extended to recursions of the type (1) (see the Appendix). Application of these results to the recursive equations for the quantization noise process $\{q_k\}$ and the quantization input process $\{d_k\}$ of the $\Sigma\Delta$ modulator show that both processes are asymptotically stationary Gaussian random processes. Asymptotically, both q_k and d_k are zero mean Gaussian random variables with mean zero and variance α^2. The expression for α^2 depends on the filter coefficients $\{h_j\}_{j=1}^{M}$, the autocorrelation of the input process $\{x_k\}$, the dither process $\{D_k\}$ and the quantizer stepsize μ. In [3], it was verified via simulations that $\{d_k + D_k\}$ can be modeled as a stationary Gaussian process. Our theoretical analysis yields that this approximation is certainly valid asymptotically under mild condition on the input $\{x_k\}$ and the dither process $\{D_k\}$.

We next list the technical conditions under which our results hold.

A1) $\{x_k\}$ is a zero mean, ergodic sequence of real valued random variables, and $\{D_k\}$ is a sequence of symmetric, zero mean i.i.d random variables independent of $\{x_k\}$. Furthermore, the density of D_k is bounded and continuous (or continuous everywhere except at a finite number of points not including the origin) and both $E[x_k^2]$ and $E[D_k^2]$ are finite.

A2) Let $\Delta(\lambda) = \lambda^M + h_1\lambda^{M-1} + \cdots + h_M$. Assume that the roots of Δ lie inside the unit circle in the complex plane with a single root at $\lambda = 1$.

A3) Assume

$$\sigma^2 = E[x_0 x_0] + 2\sum_{k=1}^{\infty} E[x_0 x_k] < \infty$$

and $\sqrt{\mu}\Sigma_{k=0}^{[t/\mu]-1} x_k \Rightarrow B$ as $\mu \to 0$, where B is standard Brownian motion with variance σ^2 (see [16] for the definition of $\Rightarrow$ and Brownian motion).

A4) Assume $q_0 = q_{-1} = \cdots = q_{-(M-1)} = k_1$ and $z_0 = z_{-1} = \cdots = z_{-(M-1)} = k_2$ for some constants k_1 and k_2.

The ergodic assumption in A1 is general enough to include a large collection of stochastic signals. For example, suppose $\{v_k\}$ is a sequence of zero mean square integrable i.i.d random variables independent of $\{D_k\}$ and suppose $\{c_k\}$ is a real valued square summable impulse response of a LTI system. Let $x_k = \Sigma_{l=-\infty}^{\infty} c_l v_{k-l}$. Then $\{x_k\}$ is an ergodic sequence and is independent of $\{D_k\}$. Assumption A2 is not satisfied by higher-order modulators for which $\Delta(\lambda)$ has multiple zeros at $\lambda = 1$. For example, the second-order lowpass modulator with $h_1 = -2$ and $h_2 = 1$ gives $\Delta(\lambda) = (\lambda - 1)^2$, which does not satisfy A2. Nevertheless, A2 holds for many modulators including the very popular first-order modulator, for which $\Delta(\lambda) = \lambda - 1$. Assumption A3 is rather technical but holds under a wide variety of conditions on the input process $\{x_k\}$. For example, in [16, ch. 4] it is shown that A3 will hold for various ARMA $\{x_k\}$ processes. For examples see Section IV.

Let $f(x) = E[\mathrm{sgn}(x + D_1)]$. Thus

$$f(x) = 1 - 2F_D(-x),$$

where $F_D(x)$ is the cumulative distribution function of D_k. Let $\alpha^2 = \mu\beta(1 + \sigma^2)/2f'(0)$, where

$$\beta = \frac{1}{M + \sum_{j=1}^{M-1}(M - j)h_j}.$$

Theorem 1: Assume A1–A4 hold and both $f'(0)$ and β are strictly positive. For small μ, q_k and d_k are asymptotically (k large) zero mean Gaussian random variables with variance α^2. Furthermore, for small μ, both $\{q_k\}$ and $\{d_k\}$ are asymptotically (k large) stationary Gaussian processes with autocorrelation function

$$R(k) = \alpha^2 e^{-\beta f'(0)k\mu}.$$

Proof: See the Appendix. □

IV. Applications

In this section, simulation results for first and second-order ΣΔ modulators are presented. The simulation results obtained are compared with the theoretical results of the previous section.

First-Order ΣΔ Modulator

From (6) it follows that the recursion for the quantization noise in the first-order ΣΔ with a single bit quantizer satisfies

$$q_{k+1} = q_k + \mu[\mathrm{sgn}(D_{k+1} + \mu x_{k+1} - q_k) - x_{k+1}]$$

and the recursion for the quantizer input process $\{d_k\}$ satisfies

$$d_{k+1} = d_k - \mu[\mathrm{sgn}(d_k + D_k) - x_{k+1}].$$

Assume that:

E1) all initial conditions of the modulator are set to zero;

E2) the dither process $\{D_k\}$ is a sequence of i.i.d random variables distributed uniformly on [−0.5, 0.5];

E3) $\{x_k\}$ is a first-order AR process given by

$$x_{k+1} = 0.8x_k + v_k$$

where $\{v_k\}$ is a sequence of random variables distributed uniformly on [−0.2, 0.2] and independent of $\{D_k\}$.

Next, it is shown assumptions A1–A4 of Theorem 1 hold. It follows from [15, ch. 6] that $\{x_k, D_k\}$ is a sequence of ergodic $\mathbb{R}^2$-valued random variables. Note that the density of D_k is discontinuous at only two points, −0.5 and 0.5. Hence, A1 holds. Since $\Delta(\lambda) = \lambda - 1$, A2 holds. Note

$$\sigma^2 = E[x_0 x_0] + 2\sum_{k=1}^{\infty} E[x_0 x_k] = \tfrac{1}{3}$$

and as shown in [16, ch. 4], A3 holds. Since all the initial conditions are set to zero, A4 holds. It follows that $f'(0) = 2, \beta = 1$ and $\alpha^2 = \mu(1 + 1/3)/4$.

In the simulations, 100 000 samples of $\{q_k\}$ and $\{d_k\}$ were computed. Histograms of bin width

$$\frac{\max\{q_k\}_{k=0}^{100000} - \min\{q_k\}_{k=0}^{100000}}{100}$$

and

$$\frac{\max\{d_k\}_{k=0}^{100000} - \min\{d_k\}_{k=0}^{100000}}{100}$$

were used to compute the densities of the quantization noise and the quantizer input process, respectively. These simulated densities where then plotted along with the theoretical densities derived in Section III. Fig. 3 shows simulated and theoretical densities for d_k for quantizer stepsizes $\mu = 1$ and $\mu = 0.5$. Fig. 4 shows simulated and theoretical densities for q_k for quantizer stepsizes $\mu = 1$ and $\mu = 0.5$. In both Fig. 3 and Fig. 4, the agreement between theoretical and experimental results is better for $\mu = 0.5$. This agreement is further improved as the quantizer stepsize is made even smaller. Note that in the case of $\mu = 0.5$ the dither D_k spans the entire interval $[-\mu, \mu]$.

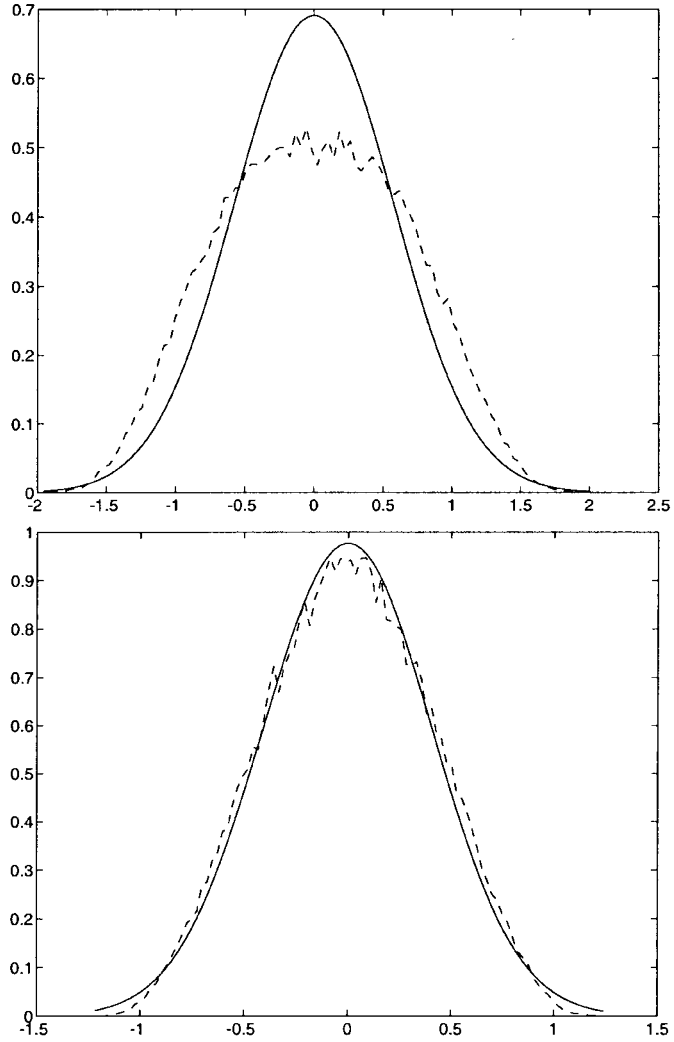

Fig. 3. Theoretical (solid line) and simulated (dashed line) densities of d_k in the first-order $\Sigma\Delta$ modulator with $\mu = 1.0$ and $\mu = 0.5$.

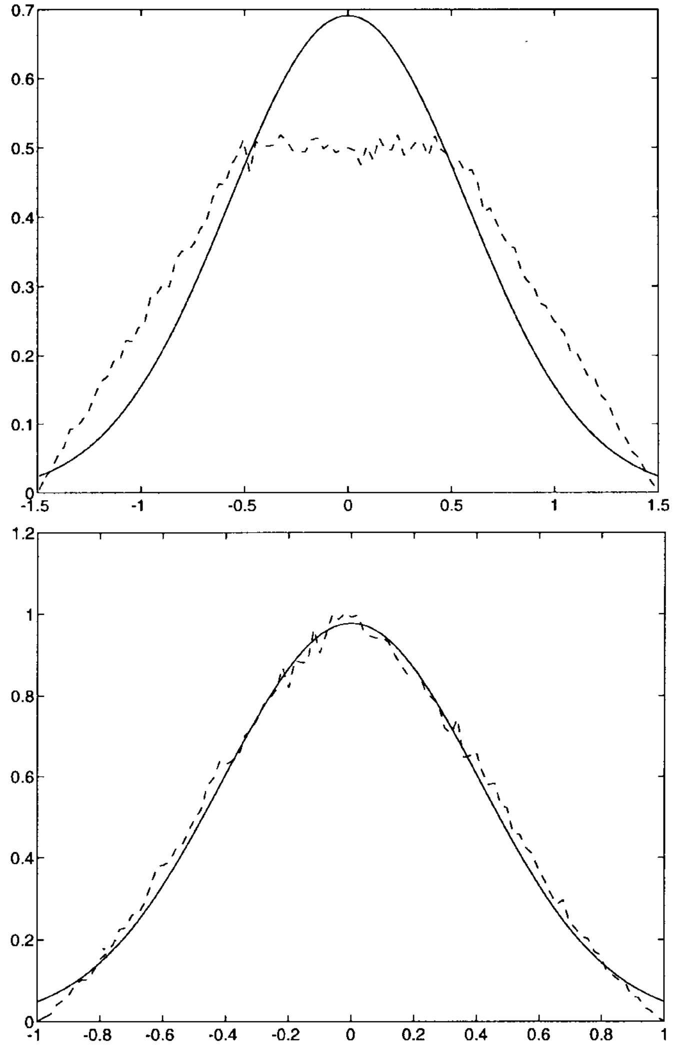

Fig. 4. Theoretical (solid line) and simulated (dashed line) densities of q_k in the first-order $\Sigma\Delta$ modulator for $\mu = 1.0$ and $\mu = 0.5$.

Second-Order $\Sigma\Delta$ Modulators

From (6) it follows that the recursion for the quantization noise $\{q_k\}$ in the second-order $\Sigma\Delta$ satisfies

$$\begin{aligned} q_{k+1} = &-h_1 q_k - h_2 q_{k-1} + \mu[\operatorname{sgn}(D_{k+1} + \mu x_{k+1} \\ &+ h_1 q_k + h_2 q_{k-1}) - x_{k+1}] \end{aligned}$$

and the recursion for the quantizer input process $\{d_k\}$ satisfies

$$\begin{aligned} z_{k+1} = &-h_1 d_k - h_2 z_{k-1} + \mu[h_1 \operatorname{sgn}(d_k + D_k) \\ &+ h_2 \operatorname{sgn}(z_{k-1} + D_{k-1}) + x_{k+1}]. \end{aligned}$$

Assume E1–E3 hold. First consider the modulator for which $h_1 = -1.5$ and $h_2 = 0.5$. In which case $\Delta(\lambda) = \lambda^2 - 1.5\lambda + 0.5 = (\lambda - 1)(\lambda - 0.5)$. It can be verified that A1–A4 continue to hold and thus Theorem 1 applies. Fig. 5 shows simulated and theoretical densities for q_k for stepsizes $\mu = 1$ and $\mu = 0.5$. Fig. 6 shows simulated and theoretical densities of d_k for quantizer stepsizes $\mu = 1$ and $\mu = 0.5$. Again note in Figs. 5 and 6 the agreement between the theoretical and simulated densities is better for $\mu = 0.5$.

Next, consider the second-order modulator with $h_1 = -0.5$ and $h_2 = -0.5$. Fig. 7 shows simulated and theoretical densities of q_k for quantizer stepsizes $\mu = 1$ and $\mu = 0.5$. Fig. 8 shows simulated and theoretical densities of d_k for quantizer stepsizes $\mu = 1$ and $\mu = 0.5$.

To verify that the results do indeed improve as μ decreases to zero, the second-order modulator with $\mu = 0.25$ and $h_1 = -0.5$ and $h_2 = -0.5$ was simulated. Fig. 9 shows the simulated and theoretical densities for q_k and d_k. Note in this case the dither spans the interval $[-2\mu, 2\mu]$.

V. Summary and Conclusion

In this paper asymptotic results were presented regarding the statistical properties of the quantization noise process $\{q_k\}$ and quantizer input process $\{d_k\}$ in $\Sigma\Delta$ modulators. The analysis technique is related to the asymptotic analysis of adaptive filtering algorithms. Simulation results were presented for first and second-order $\Sigma\Delta$ modulators. The results for the first-order and second-order $\Sigma\Delta$ modulators were fairly accurate even for quantizer stepsizes μ as large as one. The simulations verified the conclusion that both the quantization noise and quantizer input are asymptotically Gaussian. The analysis and simulations were carried out under the assumption that in loop

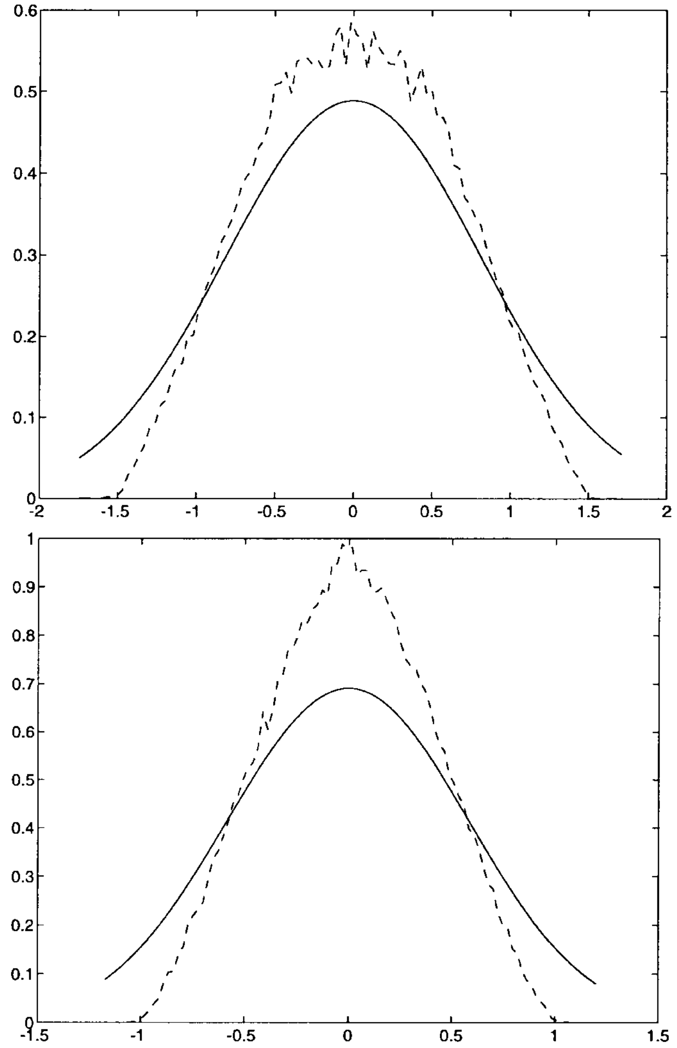

Fig. 5. Theoretical (solid line) and simulated (dashed line) densities of q_k in the second-order ΣΔ modulator for $\mu = 1$ and $\mu = 0.5$.

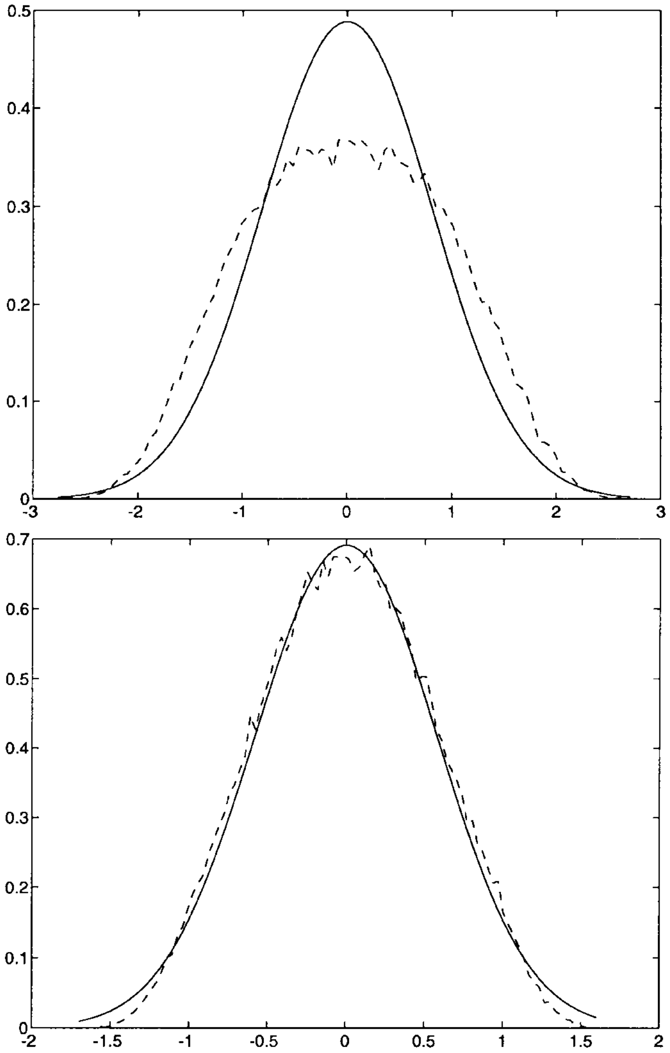

Fig. 6. Theoretical (solid line) and simulated (dashed line) densities of d_k in the second-order ΣΔ modulator with $\mu = 1$ and $\mu = 0.5$.

dithering is utilized, and the theoretical results hold for a large collection of stochastic input signals. Similar results can easily be obtained for DPCM systems.

One topic for future work is to generalize the theoretical results to systems in which dithering is not employed or the distribution of the dither is μ dependent. The present analysis used the fixed distribution of the dither to smooth the $\mathrm{sgn}(\cdot)$ function. Since agreement between theoretical results and experimental results improves as $\mu \to 0$, it is desirable from a practical viewpoint to have the dither amplitude approach zero as $\mu \to 0$ or not to use dithering at all.

Our results hold under the assumption that $\Delta(\lambda)$ does not have multiple zeros at $\lambda = 1$. However, this assumption is not satisfied by certain higher-order low pass noise transfer functions. For example, the second-order low pass noise transfer function $H(z) = (1 - z^{-1})^2$ yields $\Delta(\lambda) = (\lambda - 1)^2$, which violates A2. The behavior of this modulators can be approximated by a modulator which places a single zero at one and the other very close to one, say 0.9. Our theoretical results apply to this modulator since for this modulator $\Delta(\lambda) = (\lambda - 1)(\lambda - 0.9)$. However, with the dither process of Section IV, the agreement between theoretical and experimental results is for μ much smaller than 0.5. Therefore, it is of interest to generalize the present results to modulators in which dithering is not used and which have multiple zeros at one.

Appendix

Proof of Theorem 1: We carry out the details for the quantization noise process $\{q_k\}$ under the assumption that the density of U_k is bounded and continuous. The results for the quantizer input process are obtained is a similar manner. The results for discontinuous densities follow in a similar manner with more technical details. Recall, $W_k = [q_k, q_{k-1}, \cdots, q_{k-M+1}]^T$,

$$A = \begin{pmatrix} -h_1 & -h_2 & -h_3 & \cdots & -h_M \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}. \tag{7}$$

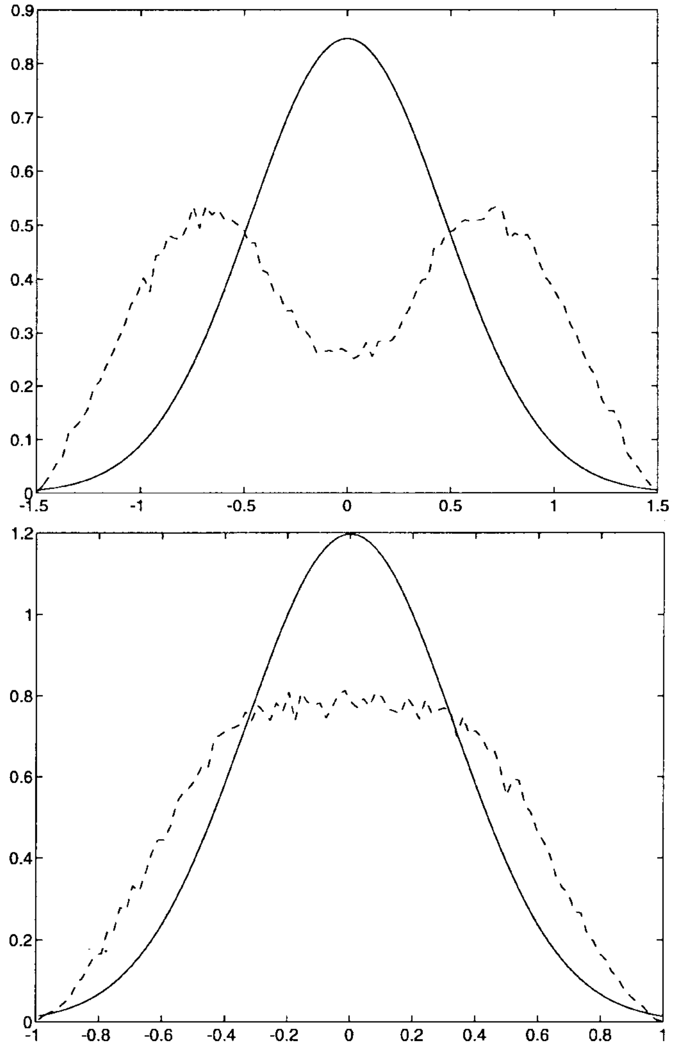

Fig. 7. Theoretical (solid line) and simulated (dashed line) densities of q_k in the second-order $\Sigma\Delta$ modulator for $\mu = 1$ and $\mu = 0.5$.

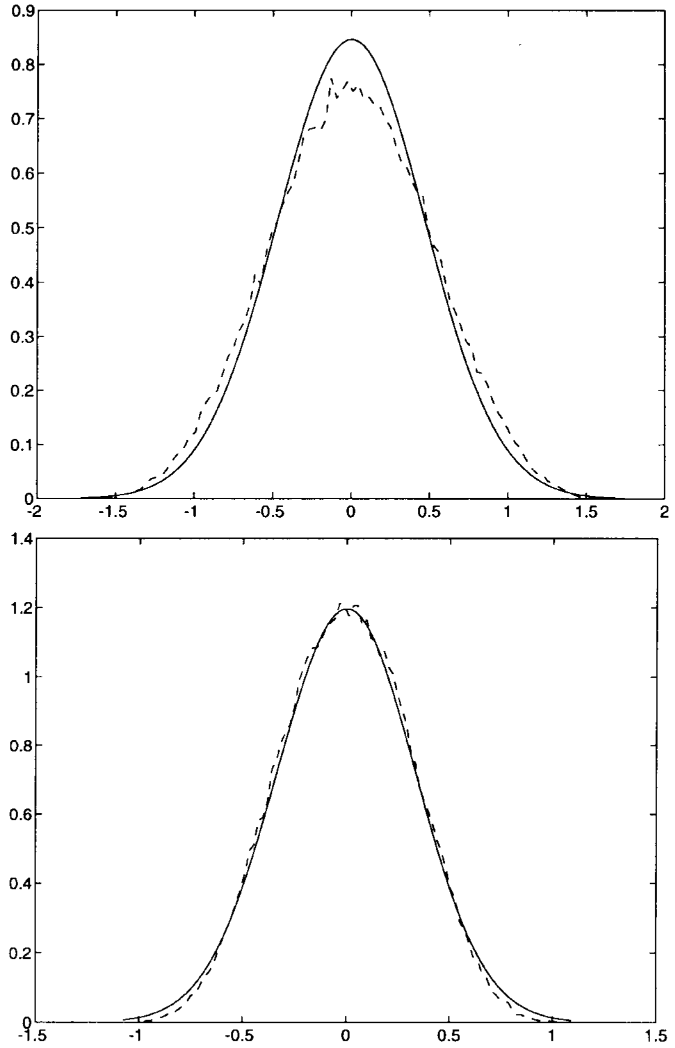

Fig. 8. Theoretical (solid line) and simulated (dashed line) densities of d_k in the second-order $\Sigma\Delta$ modulator for $\mu = 1$ and $\mu = 0.5$.

$Y_k = x_{k+1}, U_k = D_k$ and

$$G(w, y, U, \mu) = [\mathrm{sgn}(\mu y + U + h_1 w_1 + h_2 w_2 + \cdots + h_M w_M) - 0, \cdots, 0]^T.$$

Then

$$W_{k+1} = AW_k + \mu G(W_k, Y_k, U_{k+1}, \mu). \tag{8}$$

As mentioned in Section III, [13, Th. 2] is not directly applicable to (8). This is because of the μ dependence of G. However, [13, Th. 2] can be generalized, so that it applies to (8). This can be accomplished in a manner similar to how [11, Th. 1] was extended in [12] to handle μ-dependent G. We state without proof this modified version of [13, Th. 2] along with the technical assumptions.

Let $\mathcal{F}_k = \sigma((W_l, Y_l, U_l)_{l=0}^k)$ (the smallest sigma algebra with respect to which $(W_l, Y_l, U_l)_{l=0}^k$ are measurable) and define

$$\hat{G}(W_k, Y_k, \mu) = E[G(W_k, Y_k, U_{k+1}, \mu)|\mathcal{F}_k].$$

C1) $A = PJP^{-1}$ is the Jordan Decomposition of A. A has eigenvalues $\{\lambda_1, \lambda_2, \cdots, \lambda_d\}$ such that $|\lambda_i| < 1$ or $\lambda_i = 1$. Furthermore, assume that the Jordan blocks corresponding to the $\lambda_i = 1$ eigenvalue have size one.

C2) $\{Y_k\}_{k \geq 0}$ is a sequence of stationary ergodic random variables with distribution μ_Y.

C3) W_0 lies in the range space of $\lim_{k \to \infty} A^k$.

C4) $\lim_{\mu \to 0} \hat{G}(w, y, \mu) = \hat{G}(w, y, 0^+)$ uniformly for $(w, y) \in K_1 \times K_2 \subset \mathbb{R}^d \times \mathbb{R}$ where $\hat{G}(\cdot, \cdot, 0^+)$ is a continuous function and K_1 and K_2 are compact sets.

C5) Define $P(U_{k+1} \in C|\mathcal{F}_k) = \eta(W_k, Y_k, C).\hat{G}(w, y, \mu)$ is differentiable with respect to w for all $\mu > 0$ and $\mu = 0^+$. G is square integrable with respect to $\eta(w, y, \cdot)$ for each pair $(w, y) \in \mathbb{R}^d \times \mathbb{R}$.

C6) Let

$$H(w, y, \mu) = E[[G(w, y, U_{k+1}, \mu) - \hat{G}(w, y, \mu)] \cdot [G(w, y, U_{k+1}, \mu) - \hat{G}(w, y, \mu)]^T]$$

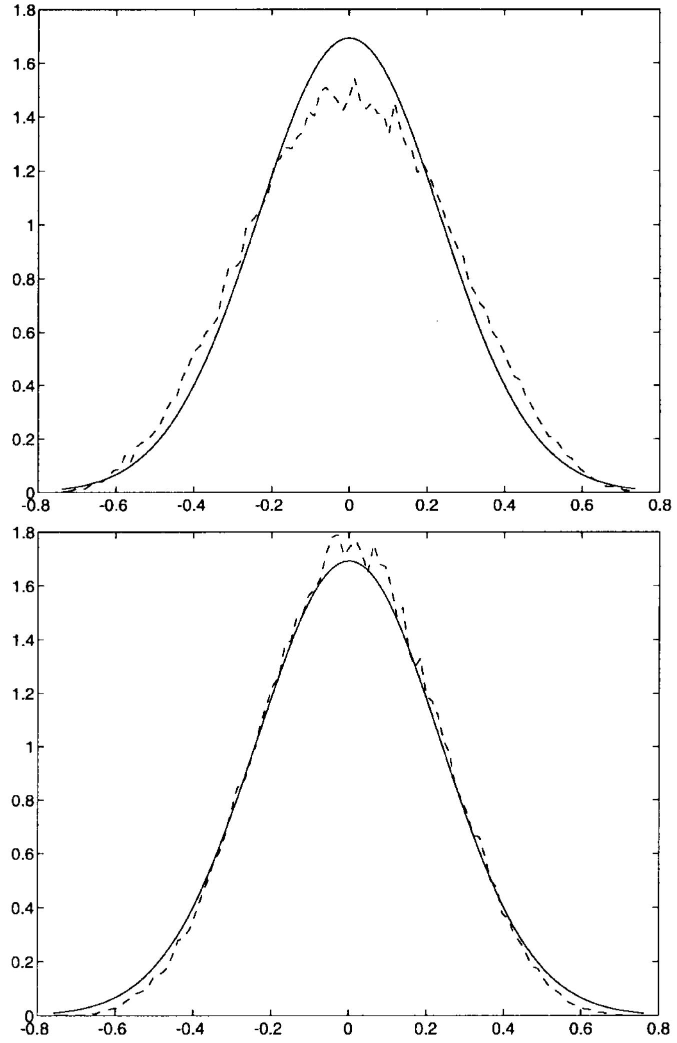

Fig. 9. Theoretical (solid line) and simulated (dashed line) densities of q_k and d_k in the second-order ΣΔ modulator with μ = 0.25.

and assume $H(w,y,\mu) \to H(w,y,0^+)$ uniformly for each $(w,y) \in K_1 \times K_2$ where $H(w,y,0^+)$ is a continuous function on $\mathbb{R}^d \times \mathbb{R}$ and K_1 and K_2 and compact sets.

C7) $\partial_w \hat{G}(w,y,\mu) \to \partial_w \hat{G}(w,y,0+)$ uniformly for each $(w,y) \in K_1 \times K_2$ as $\mu \to 0$.

C8)

$$E[\sup_{\mu>0, w\in\mathbb{R}^d} |G(w,Y_k,q(w,Y_k,\Psi_k),\mu)|^2] < \infty,$$

$$E[\sup_{\mu\geq 0, w\in\mathbb{R}^d} |\hat{G}(w,Y_k,\mu)|^2] < \infty.$$

$$E[\sup_{\mu\geq 0, w\in\mathbb{R}^d} |H(w,Y_k,\mu)|] < \infty$$

and

$$E[\sup_{\mu\geq 0, w\in\mathbb{R}^d} |\partial_w \hat{G}(w,Y_k,\mu)|] < \infty.$$

Now define

$$\tilde{M}_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (G(W_k,Y_k,U_{k+1},\mu) - \hat{G}(W_k,Y_k,\mu))\sqrt{\mu}$$

and

$$L_\mu(t) = \sum_{k=0}^{[t/\mu]-1} (\hat{G}(W(k\mu),Y_k,\mu) - \overline{G}(W(k\mu)))\sqrt{\mu}$$

where $W_\mu \Rightarrow W$ in probability as $\mu \to 0$ with $W(t)$ being the unique function satisfying $W(0) = W_0$ and $\dot{W}(t) = \overline{AG}(W(t))$.

C9) There are a variety of different conditions that imply L_μ converges weakly to a Brownian motion. We simply assume this convergence. $L_\mu \Rightarrow L$.

Theorem 2: Suppose C1–C9 hold. Then $\tilde{M}_\mu \Rightarrow \tilde{M}$ where $\tilde{M}$ is a mean zero Brownian motion independent of L with

$$E[\tilde{M}(t)\tilde{M}^t(t)] = \int_0^t \overline{H}(W(s))\,ds$$

where $\overline{H}(w) = \int_{\mathbb{R}} H(w,y,0^+)\mu_Y(dy)$, and $V_\mu \Rightarrow V$ where V satisfies

$$V(t) = \overline{A}\tilde{M}(t) + \overline{A}L(t) + \overline{A}\int_0^t \partial_w \overline{G}(W(s))V(s)\,ds.$$

To apply Theorem 2 we need to verify that A1–A4 imply C1–C9. It is easy to verify that C1–C8 are satisfied if A1-A4 hold. To see that C9 also holds, recall that $L_\mu(t) = [l_\mu(t), 0, \cdots, 0]^T$ where

$$l_\mu(t) = \sum_{k=0}^{[t/\mu]-1} f(\mu Y_k)\sqrt{\mu} - \sum_{k=0}^{[t/\mu]-1} Y_k\sqrt{\mu}.$$

We need the following result to compute the limit of l_μ.

Lemma 1: Let $\{Y_k\}_{k=0}^\infty$ be $\mathbb{R}$-valued stationary random variables and assume $f: \mathbb{R} \to \mathbb{R}$ satisfies $|f(x)| \leq K|x|$ for all $x \in \mathbb{R}$. Then

$$\lim_{\mu\to 0} \sum_{k=0}^{[t/\mu]-1} f(\mu Y_k)\sqrt{\mu} = 0$$

where the above limit is in L^1.

Proof: Omitted. □

Hence, using Lemma 1 and H4 it follows (see [16, ch. 4]) that $L(t) = [l(t), 0, \cdots, 0]^T$ where $l(t)$ is Brownian motion with variance σ^2 and

$$\sigma^2 = E[Y_0^2] + \sum_{k=1}^{\infty} E[Y_0 Y_k].$$

Hence C9 holds and thus Theorem 2 applies.

The following are easy computations: $\overline{A} = [a_1\alpha, a_2\alpha, \cdots, a_M\alpha]$ where

$$a_k = \frac{1 + \sum_{j=1}^{k-1} h_j}{M + \sum_{j=1}^{M-1} (M-j)h_j} \tag{9}$$

$$\partial_w \overline{G}(w) = \begin{pmatrix} f'(h^t w)h_1 & f'(h^t w)h_2 & f'(h^t w)h_2 & \cdots & f'(h^t w)h_M \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}. \tag{12}$$

and $\alpha = [1, 1, \cdots, 1]^T$.

$$\begin{aligned} \hat{G}(w, y, \mu) &= [f(\mu y + h^t w) - \mu y, 0, \cdots, 0]^t] \\ \hat{G}(w, y, 0^+) &= [f(h^t w) - y, 0, \cdots, 0]^t \\ \overline{G} &= [f(h^t w), 0, \cdots, 0]^t \\ H(w, y, \mu) &= \begin{pmatrix} [f(y + h^t w)]^2 + 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \end{aligned} \tag{10}$$

$$H(w, y, 0^+) = \begin{pmatrix} [f(h^t w)]^2 + 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix} \tag{11}$$

(see (12) at the top of the page). Let $w_\mu(t) = q_{[t/\mu]}$ and $v_\mu(t) = w_\mu(t)/\sqrt{\mu}$, then it follows

$$w_\mu \Rightarrow 0$$

and

$$v_\mu \Rightarrow v$$

where

$$v(t) = a_1(\tilde{m}(t) + l(t)) - a_1 f'(0) \int_0^t v(s)\, ds,$$

and where $\tilde{m} + l$ is Brownian motion with variance $1 + \sigma^2$ and

$$\sigma^2 = E[Y_0^2] + \sum_{k=}^{\infty} E[Y_0 Y_k].$$

If $a_1 f'(0) > 0$, then $v(t)$ is an asymptotically stationary Gaussian random process with mean zero, variance $a_1(1 + \sigma^2)/2f'(0)$ and autocorrelation function

$$r_v(\tau) = E[v(t)v(t + \tau)] = \frac{(\sigma^2 + 1)a_1}{2f'(0)} e^{-a_1 f'(0)\tau}. \tag{13}$$

Then the conclusions of Theorem 1 follow.

References

[1] J. C. Candy and O. J. Benjamin, "The structure of quantization noise from sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-29, pp. 1316–1323, Sept. 1981.

[2] J. C. Candy and J. Temes, Eds, *Oversampling Delta-Sigma Data Converters*. New York: IEEE Press, 1988.

[3] R. Khoini-Poorfard and D. A. Johns, "Analysis of $\Sigma\Delta$ modulators with zero mean stochastic inputs," *IEEE Trans. Circuits Syst. II*, vol. 42, Mar. 1995.

[4] W. Chou, "Dithering and its effect on Sigma-Delta and multi Sigma-Delta modulation," *IEEE Trans. Inform. Theory*, vol. 37, May 1991.

[5] R. M. Gray, "Oversampled sigma-delta modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481–489, May 1987.

[6] ——, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, June 1989.

[7] ——, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.

[8] O. Hermann and H. W. Schussler, "Design of nonrecursive digital filters with minimum phase," *Electron. Lett.*, vol. 6, pp. 329–330, 1970.

[9] H. Inose and Y. Yasuda, "A unity bit coding method by negative feedback," *Proc. IEEE*, vol. 51, pp. 1524–1535, Nov. 1963.

[10] N. S. Jayant and P. Knoll, *Digital Coding of Analog Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[11] J. A. Bucklew, T. G. Kurtz, and W. A. Sethares, "Weak convergence and local stability properties of fixed step size recursive algorithms," *IEEE Trans. Inform. Theory*, vol. 39, pp. 966–978, May 1993.

[12] J. A. Bucklew and W. A. Sethares, "The covering problem and μ-dependent adaptive algorithms," *IEEE Trans. Signal Processing*, vol. 42, pp. 2616–2627, Oct. 1994.

[13] R. Sharma, W. A. Sethares, and J. A. Bucklew, "Analysis of momentum adaptive filtering algorithms," *IEEE Trans. Signal Processing*, to be published.

[14] R. Steele, *Delta Modulation Systems*. New York: Wiley, 1975.

[15] R. Durrett, *Probability: Theory and Examples*. Cole, CA: Wadsworth and Brooks, 1991.

[16] P. Billingsley, *Convergence of Probability Measures*. New York: Wiley, 1968.

[17] S. Ethier and T. Kurtz, *Markov Processes—Characterization and Convergence*. New York: Wiley-Interscience, 1986.

Rajesh Sharma (S'92–M'95) was born in Srinagar, India, on September 20, 1970. He received the B.S. degree (with distinction) and the M.S. and Ph.D. degrees all in electrical engineering from the University of Wisconsin, Madison, in 1991, 1992, and 1995, respectively.

Since 1995, he has been with the Advanced Information Systems Group of the Environmental Research Institute of Michigan, Ann Arbor. His research interest include adaptive signal processing, SAR image formation and segmentation, computer vision and applied probability.

Dr. Sharma is a member of Eta Kappa Nu.

James A. Bucklew (S'75–M'79–SM'96) received the Ph.D. degree from Purdue University, West Lafayette, IN, in 1979.

He is currently a Professor with the Department of Electrical and Computer Engineering and the Department of Mathematics, University of Wisconsin, Madison. His research interests are in the application of probability to signal processing and communication problems.

Dr. Bucklew is the recipient of a Presidential Young Investigator Award (1984). He has served as the Associate Editor at Large (1989–1992) and the Associate Editor for Detection (1992) for the IEEE TRANSACTIONS ON INFORMATION THEORY.

William A. Sethares (S'84–M'86) received the B.A. degree in mathematics from Brandeis University, Waltham, MA, and the M.S. and Ph.D. degrees in electrical engineering from Cornell University, Ithaca, NY.

He has been with the Raytheon Company as a Systems Engineer and is currently on the faculty of the Department of Electrical and Computer Engineering, University of Wisconsin, Madison. His research interests include adaptive systems in signal processing, communications and electronic music.