BREAKING THE FEEDBACK LOOP OF $\Sigma\Delta$ MODULATORS

Nguyen T. Thao

Dept. of Electrical Engineering, City College, City University of New York, New York, NY 10031
e-mail: thao@ee.ccny.cuny.edu

ABSTRACT

Although the topic of analog-to-digital (A/D) conversion should fundamentally be part of the signal processing area, paradoxically, very little signal theory has been devoted to it. One reason is that the main technique currently used in data acquisition, called $\Sigma\Delta$ modulation, consists of coarse quantization with feedback. This prevents the normal resolution of the output of the system in terms of its input, contrary to the case of linear feedback systems. This paper introduces new tools to solve this problem and enable rigorous signal error analysis of $\Sigma\Delta$ modulators.

1. INTRODUCTION

The area of analog-to-digital (A/D) conversion has experienced tremendous progress since the introduction of oversampled quantization with feedback, known under the name of $\Sigma\Delta$ modulation [1] (see Figure 1(a)). The principle of the method is to achieve high resolutions of conversion by oversampling the input, rather than by increasing the resolution of the quantizer. The role of the feedback is to reduce the inband portion of the quantization error signal by taking advantage of the input redundancy. In spite of the simplicity and efficiency of this technique however, $\Sigma\Delta$ modulation has introduced a new system that is not accessible to standard signal processing analysis. This is because, contrary to linear systems, there does not exist a general technique to solve the output of a nonlinear feedback system in terms of its input. As a result, little rigorous signal theory has been available on $\Sigma\Delta$ modulation. In particular, the theoretical question of how fast the inband error decays with the oversampling has been not been rigorously solved yet in the general case. When the two-input linear part of a $\Sigma\Delta$ modulator (see Figure 1(a)) is characterized by the two transfer functions $G(z)$ and $H(z)$ as shown in Figure 1(b) and that $H(z) = B(z)/A(z)$ with

$$A(z) = 1 + a_1 z^{-1} + + a_n z^{-n} \text{ and } B(z) = (1 - z^{-1})^n, \quad (1)$$

it is commonly believed that the inband mean squared error (MSE) of the system decays with the oversampling ratio M in α/M^{2n+1}. It was found in [2] that this results only holds in the ideal case where $A(z) = 1$ and the quantizer is uniform and never overloaded. In this paper, we call this configuration of modulation the *ideal* $\Sigma\Delta$ modulators. In the general case, it was numerically observed that the asymptotic error decay is in β/M^{2n} instead. This is illustrated by the numerical results of Figure 2(a).

The only rigorous error analysis that exists until now applies to the ideal $\Sigma\Delta$ modulators [3, 4]. As a matter of fact, this analysis was possible because the ideal $\Sigma\Delta$ configuration was indeed found to yield an explicit output expression in terms of the input. This explicit expression is represented through the feedforward block diagram of Figure 3(b), which can be seen as a generalization of the block diagram of uniform quantization (see Figure 3(a)). The nonlinear part of this diagram is all reduced to the function $\mathrm{mod}_{[-\frac{1}{2},\frac{1}{2})}$ which we define to be the 1-periodic function that is invariant in $[-\frac{1}{2}, \frac{1}{2})$.

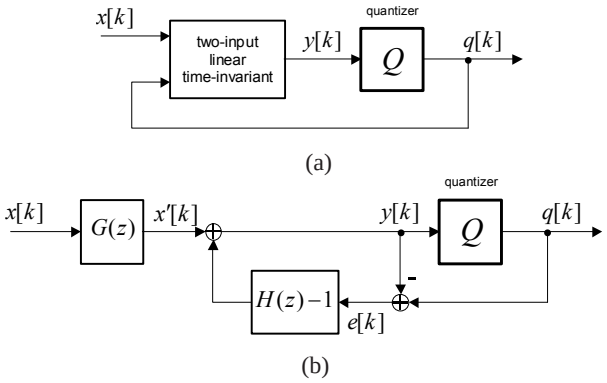

Fig. 1. General block diagrams of $\Sigma\Delta$ modulators: (a) original diagram; (b) equivalent diagram.

In this paper, we show that, in the case of constant inputs, all modulators characterized by (1) yield an equivalent feedforward diagram of the type of Figure 3(c). The only conditions required here are the stability of the modulators and the uniformity of the quantization levels. This implies in particular that the quantizer can be overloaded, have non-uniform quantization thresholds and $A(z)$ can be any polynomial or degree less or equal to n. We then show how the MSE behavior of β/M^{2n} versus α/M^{2n+1} can be explained with *time-varying* inputs, thanks to the equivalent block diagram for constant inputs.

2. THE TILING PROPERTY WITH CONSTANT INPUTS

The key to this research has been the recent observation of a remarkable property of $\Sigma\Delta$ modulators with constant inputs. We consider the general class of modulators mentioned in the introduction. Without loss of generality, we will however set $G(z)$ to 1. Mathematically, we simply assume that the quantizer function Q maps $\mathbb{R}$ into discrete values of $\mathbb{R}$ that lie on a uniform grid of period 1. Let $x[k]$ be equal to a constant x and $u[k]$ be the nth order integration of the output-input difference $x[k] - q[k]$ of the modulator. Formally, we have in the z-domain

$$B(z)U(z) = X(z) - Q(z). \quad (2)$$

In the case $n = 2$, we plot by black dots in Figure 4 the position of a large number of consecutive state vectors

$$\mathrm{U}[k] = [u[k] \;\; u[k-1] \;\; \cdots \;\; u[k-n+1]] \in \mathbb{R}^n. \quad (3)$$

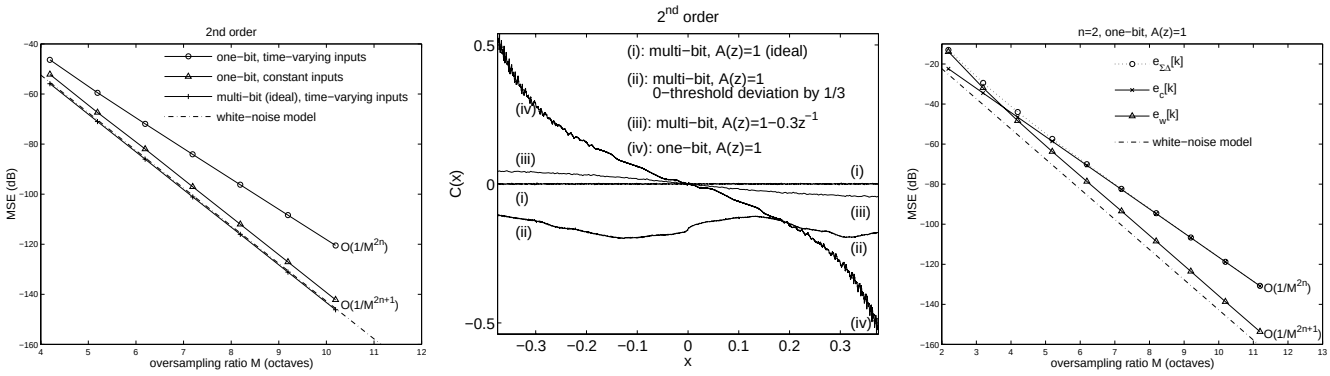

Fig. 2. Numerical experiments: (a) MSE results of second order ($n = 2$) $\Sigma\Delta$ modulators with $A(z) = 1$ and $\frac{\Delta^2}{12}$ as 0dB reference (where Δ is the quantization step size); (b) Centroid function $C(x)$ numerically obtained on various second order $\Sigma\Delta$ modulators; (c) MSE results of the sequences $e_{\Sigma\Delta}[k]$, $e_c[k]$ and $e_w[k]$ with the single-bit second order ($n = 2$) $\Sigma\Delta$ modulator with $A(z) = 1$ and time-varying inputs.

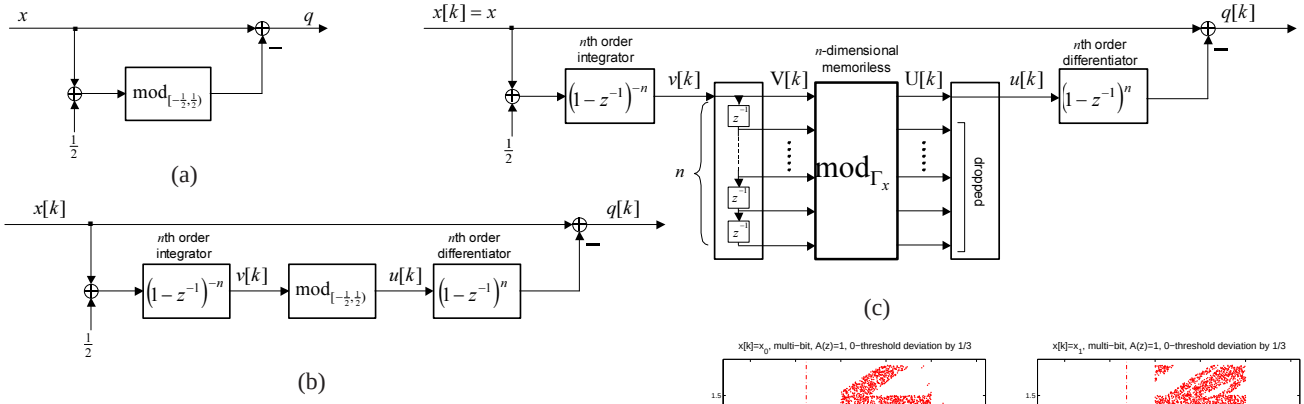

Fig. 3. Equivalent block diagrams to quantization systems: (a) infinite mid-riser uniform scalar quantizer of step size 1; (b) ideal nth order $\Sigma\Delta$ modulator; (c) non-ideal nth order $\Sigma\Delta$ modulator with a constant input $x[k] = x$.

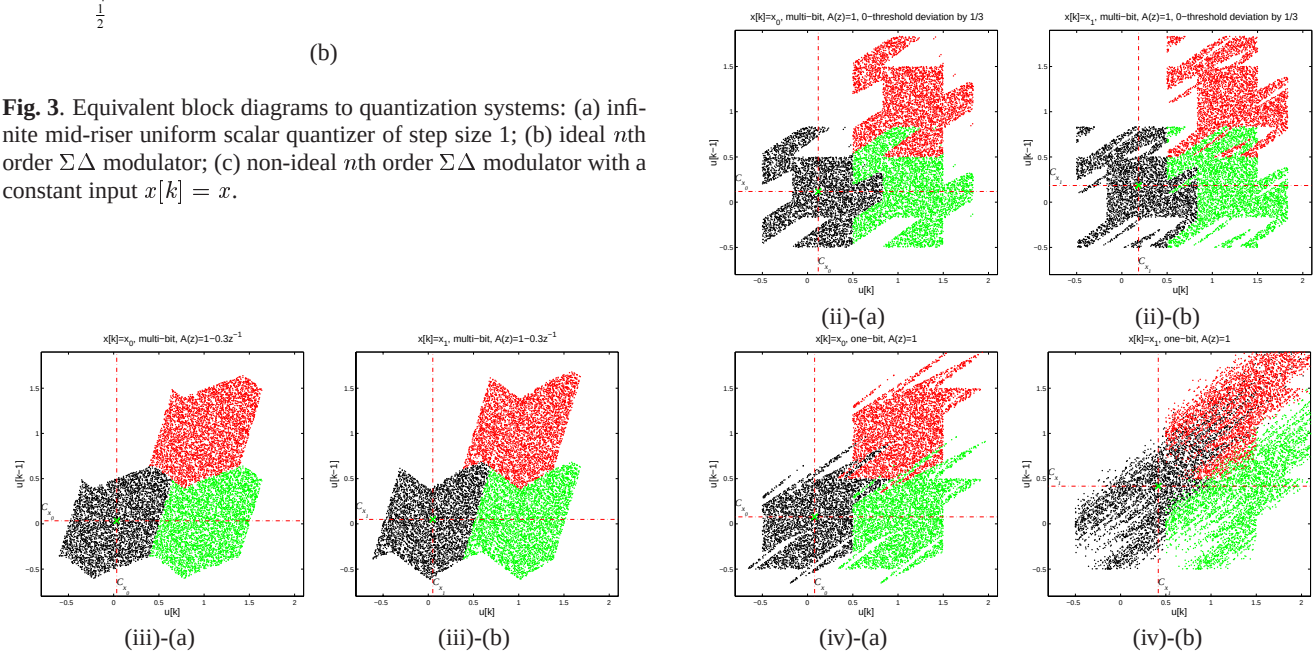

Fig. 4. Representation in black of 4,000 consecutive state points $\mathrm{U}[k]$ of various second order $\Sigma\Delta$ modulators with constant inputs $x[k]$: (a) $x[k] = x_0 = \sqrt{2}/11$; (b) $x[k] = x_1 = \sqrt{2}/4$; (ii) multi-bit configuration with $A(z) = 1$ and a deviation of the 0-threshold by $1/3$; (iii) multi-bit configuration with $A(z) = 1 - 0.3z^{-1}$; (iv) single-bit configuration with $A(z) = 1$. The points in gray are the translated versions of the state vector points by $[1\ 0]$ and $[1\ 1]$, respectively.

The figure includes three different modulators and two choices of value for x. The state points $\mathrm{U}[k]$ appear to remain in some deterministic set that we call Γ_x. The definition of this set depends on both the choice of modulator and the value of x. The remarkable phenomenon that can be observed about Γ_x is that it also appears to tile the space with its integer shifts in every dimension of $\mathbb{R}^n$. We say in this paper that Γ_x is a *tiling set*. We highlight this property in Figure 4 by representing in gray the shifted versions of the state points by $[1\ 0]$ and $[1\ 1]$ respectively. This phenomenon was first observed in [5, 6] with stable one-bit modulators and later experimentally confirmed on multi-bit configurations in [2].

Until now, there does not exist a full mathematical proof of this property. However, the characteristics of $\Sigma\Delta$ modulators that are responsible for it can be identified. For convenience, we will use the matrix notation

$$\mathrm{I}_{m_1,m_2} := \Big[\delta[i-j]\Big]_{\substack{1\le i\le m_1\\ 1\le j\le m_2}}$$

By following the node signal notation of Figure 1(b) and assuming $G(z)=1$, we have in the z-domain

$$Y(z) = X(z) + (H(z)-1)E(z) \quad \text{and} \quad Q(z) = Y(z) + E(z). \tag{4}$$

By eliminating $Q(z)$ and $E(z)$ between the three equations of (2) and (4), we obtain

$$\begin{aligned} Y(z) &= X(z) - (H(z)-1)H^{-1}(z)B(z)U(z) \\ &= X(z) - (B(z)-A(z))U(z). \end{aligned} \tag{5}$$

By expressing (2) and (5) in the time domain, we obtain respectively

$$\left\{\begin{array}{lcl} u[k] &=& -\sum_{i=1}^n b_i\cdot u[k-i] + (x[k]-q[k]) \\ y[k] &=& x[k] - \sum_{i=1}^n (b_i-a_i)u[k-i] \end{array}\right. \tag{6}$$

where b_i are the integer coefficients of the expansion $B(z) = (1-z^{-1})^n = \sum_{i=0}^n b_i z^{-i}$. With a matrix notation version of (6), we obtain the following dynamical system description of the $\Sigma\Delta$ modulator:

$$\left\{\begin{array}{lcl} \mathrm{U}[k] &=& \mathrm{U}[k-1]\cdot \mathrm{L} + (x[k]-q[k])\cdot \mathrm{I}_{1,n}, \\ q[k] &=& Q(y[k]), \\ y[k] &=& x[k] - \mathrm{U}[k-1]\cdot \mathrm{F}^\top, \end{array}\right. \tag{7}$$

$$\text{where} \qquad \mathrm{L} := \begin{bmatrix} -\mathrm{B}^\top & \mathrm{I}_{n,n-1} \end{bmatrix}, \quad \mathrm{B} := [b_1\ \cdots\ b_n] \tag{8}$$

and $\mathrm{F} := [(b_1-a_1)\ \cdots\ (b_n-a_n)]$. By eliminating $q[k]$ and $y[k]$ in (7), we obtain

$$\mathrm{U}[k] = \mathrm{U}[k-1]\cdot \mathrm{L} + \Big\{x[k] - Q(x[k] - \mathrm{U}[k-1]\cdot \mathrm{F}^\top)\Big\}\cdot \mathrm{I}_{1,n}. \tag{9}$$

When $x[k]=x$, this implies that $\mathrm{U}[k] = \mathrm{M}_x(\mathrm{U}[k-1])$ where

$$\mathrm{M}_x(\mathrm{U}) := \mathrm{U}\cdot \mathrm{L} + \Big\{x - Q(x - \mathrm{U}\cdot \mathrm{F}^\top)\Big\}\cdot \mathrm{I}_{1,n}. \tag{10}$$

The set Γ_x observed in the experiments of Figure 4 must then be an invariant set of M_x, i.e., a set such that $\mathrm{M}_x(\Gamma_x) = \Gamma_x$. The question is to know what characteristics of M_x are responsible for making this invariant set a tiling set. We have the following property.

Theorem 2.1 *Consider any mapping M_x of the type of (10) where L is a matrix with integer coefficients and a determinant equal to ± 1 and where Q has output values on a uniform grid of $\mathbb{R}$ of period 1. Then M_x transforms any tiling set of $\mathbb{R}^n$ into a tiling set of $\mathbb{R}^n$.*

This is proved in [7]. Because the polynomial $B(z) = (1-z^{-1})^n$ has by necessity integer coefficients with the constraint $b_n = (-1)^n$, one can easily verify that the matrix L resulting from (8) satisfies the conditions of the theorem. This theorem does not prove that the existence of an invariant set that is a tiling set, but it leads to the following intuition. Consider any arbitrary tiling set S_0 (for example, a unit hypercube of $\mathbb{R}^n$). We know from the theorem that the set $S_k := \mathrm{M}_x^k(S_0)$ must be a tiling set for all $k>0$. If S_k "converges" to a set Γ_x when k goes to $+\infty$, then Γ_x will also be a tiling set that satisfies the invariance relation $\Gamma_x = \mathrm{M}_x(\Gamma_x)$ as the limit of the recursive relation $S_k = \mathrm{M}_x(S_{k-1})$. However, this idea is only intuitive. A formalization of this idea is in fact a difficult problem. Some further mathematical constructions for the proof of existence of a tiling invariant set have been proposed in [7].

3. BREAKING THE FEEDBACK LOOP

Based on the fact that $\mathrm{U}[k]$ remains in a tiling set Γ_x when $x[k]=x$, we show in this section that the $\Sigma\Delta$ modulator yields the equivalent feedforward diagram of Figure 3(c). We explicitly assume here that the output values of the quantizer function Q are of the form $i-\frac{1}{2}$ where $i\in\mathbb{Z}$ (mid-riser quantization). If we denote by $b^{(-1)}[k]$ the causal inverse z-transform of $B(z)^{-1}$, then we have in the time domain

$$u[k] = b^{(-1)}[k] * (x[k]-q[k]). \tag{11}$$

Note that the sequence $b^{(-1)}[k]$ only contains integer values. As a result,

$$d[k] := b^{(-1)}[k] * (q[k]+\tfrac{1}{2}) \tag{12}$$

is a sequence that contains only integer values, since $q[k]+\frac{1}{2}$ is always an integer. Now, one can always decompose $u[k]$ as

$$u[k] = v[k] - d[k] \tag{13}$$

with

$$v[k] := b^{(-1)}[k] * (x[k]+\tfrac{1}{2}) \tag{14}$$

Similarly to (3), let us define

$$\mathrm{V}[k] := [v[k]\ \ v[k-1]\ \cdots\ v[k-m+1]], \tag{15}$$

$$\mathrm{D}[k] := [d[k]\ \ d[k-1]\ \cdots\ d[k-m+1]]. \tag{16}$$

where $v[k]$ and $d[k]$ were defined in (14) and (12) respectively. Because of (13), we obviously have $\mathrm{U}[k] = \mathrm{V}[k] - \mathrm{D}[k]$. Now, since $d[k]$ is always an integer, then $\mathrm{D}[k]\in\mathbb{Z}^m$. Since $\Gamma_x$is a tiling set, there exists a unique function mod_{Γ_x} of $\mathbb{R}^n$ that is invariant in Γ_x and is 1-periodic in each dimension. With the fact that $\mathrm{U}[k]\in\Gamma_x$, this implies that

$$\begin{aligned} \mathrm{U}[k] &= mod_{\Gamma_x}(\mathrm{U}[k]) \\ &= mod_{\Gamma_x}(\mathrm{V}[k]). \end{aligned} \tag{17}$$

One then obtains the diagram of Figure 3(c) by implementing (14), (15), (17) and the relation $q[k] = x[k] - b[k] * u[k]$ that results from (2).

4. ERGODICITY AND CONSEQUENCES

With the equivalent feedforward diagram of Figure 3(c), all the difficulty of analysis finds itself concentrated in the nonlinear function mod_{Γ_x}. Not only can this mapping be of extreme complexity, but we do not currently have tools to systematically derive the set Γ_x. However, some new global properties can still be extracted. It is easy to see from the diagram that the sequence $v[k]$

is basically a polynomial in k with leading coefficient equal to $\frac{1}{n!}(x+\frac{1}{2})$. If we assume for a moment that Γ_x is a unit hypercube of $\mathbb{R}^n$, it is know from the ergodic theory that the sequence of points $\mathrm{U}[k] = mod_{\Gamma_x}(\mathrm{V}[k])$ will have a uniform density in Γ_x when x is irrational [8]. As shown in [7], this property remains valid when Γ_x is more generally a measurable tiling set. Since $u[k] = \mathrm{U}[k] \cdot \mathrm{I}_{1,n}^{\top}$, the uniform density of $\mathrm{U}[k]$ in Γ_x implies that

$$\lim_{M \to +\infty} \frac{1}{M} \sum_{j=1}^{M} u[j] = C(x) \qquad (18)$$

where $C(x) := \int_{\Gamma_x} \mathrm{U} \cdot \mathrm{I}_{1,n}^{\top} d\mathrm{U}$. A qualitative interpretation of (18) is that $u[k]$ has a DC component equal to $C(x)$. Note that, although (18) is only guaranteed for x irrational, the function $C(x)$ is defined for all x as long as the condition that Γ_x is a measurable set is realized. While $C(x)$ cannot be analytically derived, since we currently do not have general tools to derive Γ_x, it is still possible to evaluate it numerically thanks to (18) at least on irrational values of x. We plot some numerical results in Figure 2(b). One can see that in the cases (ii-iii-iv) of Figure 2(b), $C(x)$ is a non-constant function of x. This is consistent with our observation of the dependence of Γ_x with x in Figure 4. Note however that $C(x)$ is constant in the case (i) corresponding to an ideal $\Sigma\Delta$ modulator. This is because the equivalent diagram of an ideal modulator of Figure 3(b) implies that of Figure 3(c) with $\Gamma_x = [-\frac{1}{2}, \frac{1}{2})^n$ which is independent of x. We believe that this independence only occurs with ideal modulators.

5. ERROR ANALYSIS

In this section, we qualitatively explain how the equivalent feedforward diagram of a $\Sigma\Delta$ modulator with constant inputs can be used to clarify its asymptotic MSE behavior with *time-varying* inputs with regard to the oversampling M. Details on the rigorous justification of the arguments that follow can be found in [5, 9]. The performance of a $\Sigma\Delta$ modulator is measured by evaluating the inband error remaining in the quantized signal $q[k]$. More precisely, if $x[k]$ is the sampled version of a bandlimited signal $x(t)$ at $t = k\frac{T}{M}$ and $f(t)$ is the ideal lowpass filter[1] that preserves $x(t)$, the inband signal error is equal to $e_{\Sigma\Delta}[k] := f_M[k] * q[k] - x[k]$ where $f_M[k]$ is the sampled version of a bandlimited signal $f(t)$ at $t = k\frac{T}{M}$. Using the relation (2), we have

$$e_{\Sigma\Delta}[k] = f_M[k] * (q[k] - x[k]) = -g_M[k] * u[k] \qquad (19)$$

where $g_M[k] := f_M[k] * b[k]$. In the general context of a time-varying oversampled signal $x[k]$, we have this intuitive idea that $x[k]$ is a slowly varying signal with k at high oversampling. With our knowledge of the constant input case, we qualitatively expect $u[k]$ to "locally" have a DC component equal to $C(x[k])$. Let us decompose $u[k] = C(x[k]) + w[k]$ where $w[k] = u[k] - C(x[k])$ is qualitatively the "AC" component of $u[k]$. Then (19) yields the decomposition $e_{\Sigma\Delta}[k] = e_c[k] + e_w[k]$ where

$$e_c[k] = -g_M[k] * C(x[k]) \quad \text{and} \quad e_w[k] = -g_M[k] * w[k]. \quad (20)$$

Because $g_M[k]$ is nothing but the nth order differentiation of $f_M[k]$, it can be shown at high oversampling that $g_M[k]$ is asymptotically equivalent to the sampled version of $(\frac{T}{M})^n f^{(n)}(t)$ at $t = k\frac{T}{M}$, where $f^{(n)}(t)$ designates the nth derivative of $f(t)$. With some assumptions on the variations of the function $C(x)$ with x [9], it can be shown that asymptotically high oversampling M

$$e_c[k] \sim \frac{1}{M^n} e_c(k\tfrac{T}{M})$$

where $e_c(t) := -T^n f^{(n)}(t) * C(x(t))$ and $*$ designates here the continuous-time convolution, unless $e_c(t)$ is itself a zero signal. As $e_c(t)$ is a fixed and deterministic signal that does not depend on M, this qualitatively shows that the MSE of $e_c[k]$ decays with M in β/M^{2n}. Based on the numerical extraction of $C(x)$ from Figure 3(b), the numerical evaluation of $e_c[k]$ plotted in Figure 3(c) confirms this result on a one-bit second order modulator, while it points out the MSE behavior of $e_w[k]$ in α/M^{2n+1}. This implies that the asymptotic MSE behavior of $e_{\Sigma\Delta}[k]$ must be in β/M^{2n}. Now, this result is conditioned on the assumption that $e_c(t)$ is a non-zero signal. With ideal modulators, we saw that $C(x)$ is a constant function of x. In this case, $e_c[k]$ is identically zero because $g_M[k]$ cancels constant inputs. This explains why the global MSE decay rate is exceptionally in α/M^{2n+1} with ideal modulators.

[1] For rigorous convergence of the derivations, it is however necessary to assume that the lowpass filter has at least some ϵ transition width [5].

6. REFERENCES

[1] S.R.Norsworthy, R.Schreier, and G.C.Temes, eds., *Delta-sigma data converters: theory, design and simulation*. IEEE Press, 1996.

[2] N.T.Thao, "MSE behavior and centroid function of mth order asymptotic $\Sigma\Delta$ modulators," *IEEE Trans. Circuits and Systems II*, vol. 49, pp. 86–100, Feb. 2002.

[3] R.M.Gray, W.Chou, and P.-W.Wong, "Quantization noise in single-loop sigma-delta modulation with sinusoidal input," *IEEE Trans. Commun.*, vol. COM-37, pp. 956–968, Sept. 1989.

[4] N.He, F.Kuhlmann, and A.Buzo, "Multi-loop sigma-delta quantization," *IEEE Trans. Information Theory*, vol. IT-38, pp. 1015–1028, May 1992.

[5] I.Daubechies and R.DeVore, "Reconstructing a bandlimited function from very coarsely quantized data. I. a family of stable sigma-delta modulators of arbitrary order," *Annals of Mathematics*. Accepted for publication.

[6] S.Gunturk, "Harmonic analysis of two problems in signal quantization and compression," Oct. 2000. PhD dissertation, Program in Applied and Computational Mathematics, Princeton University, http://www.math.ias.edu/ğunturk/research.html.

[7] N.T.Thao, "Breaking the feedback loop of $\Sigma\Delta$ modulators," *IEEE Trans. on Signal Proc.* Submitted, http://www-ee.engr.ccny.cuny.edu/www/web/thao/nguyen.html.

[8] K.Petersen, *Ergodic Theory*. Cambridge England: Cambridge University Press, 1983.

[9] N.T.Thao, "Asymptotic MSE law of nth order $\Sigma\Delta$ modulators," *IEEE Trans. Circuits and Systems II*. Submitted, http://www-ee.engr.ccny.cuny.edu/www/web/thao/nguyen.html.