Refined Error Analysis in Second-Order $\Sigma\Delta$ Modulation With Constant Inputs

C. Sinan Güntürk and Nguyen T. Thao, *Member, IEEE*

***Abstract*—Although the technique of sigma–delta ($\Sigma\Delta$) modulation is well established in practice for performing high-resolution analog-to-digital (A/D) conversion, theoretical analysis of the error between the input signal and the reconstructed signal has remained partial. For modulators of order higher than 1, the only rigorous error analysis currently available that matches practical and numerical simulation results is only applicable to a very special configuration, namely, the standard and ideal k-bit k-loop $\Sigma\Delta$ modulator. Moreover, the error measure involves averaging over time as well as possibly over the input value. At the second order, it is known in practice that the mean-squared error decays with the oversampling ratio λ at the rate $O(\lambda^{-5})$. In this paper, we introduce two new fundamental results in the case of constant input signals. We first establish a framework of analysis that is applicable to all second-order modulators provided that the built-in quantizer has uniformly spaced output levels, and that the noise transfer function has its two zeros at the zero frequency. In particular, this includes the one-bit case, a rigorous and deterministic analysis of which is still not available. This generalization has been possible thanks to the discovery of the mathematical tiling property of the state variables of such modulators. The second aspect of our contribution is to perform an instantaneous error analysis that avoids infinite time averaging. Until now, only an $O(\lambda^{-4})$ type error bound was known to hold in this setting. Under our generalized framework, we provide two types of squared-error estimates; one that is statistically averaged over the input and another that is valid for almost every input (in the sense of Lebesgue measure). In both cases, we improve the error bound to $O(\lambda^{-4.5})$, up to a logarithmic factor, for a general class of modulators including some specific ones that are covered in this paper in detail. In the particular case of the standard and ideal two-bit double-loop configuration, our methods provide a (previously unavailable) instantaneous error bound of $O(\lambda^{-5})$, again up to a logarithmic factor.**

***Index Terms*—Analog-to-digital (A/D) conversion, discrepancy, exponential sums, piecewise affine transformation, quantization, sigma–delta ($\Sigma\Delta$) modulation, tiling, uniform distribution.**

I. Introduction

In current state-of-the-art circuit design, high-resolution analog-to-digital (A/D) and digital-to-analog (D/A) conversion is achieved by oversampling the input signal and transforming it into a sequence of coarsely quantized values which are selected from a small alphabet consisting of as few as two symbols. An approximation of the input is then obtained by extracting the in-band content of the quantized signal via appropriate filtering. Sigma–delta ($\Sigma\Delta$) modulation is a widely used method for this purpose (see [1], [19], [14]), owing its success largely to its robustness against circuit imperfections and ease of implementation.

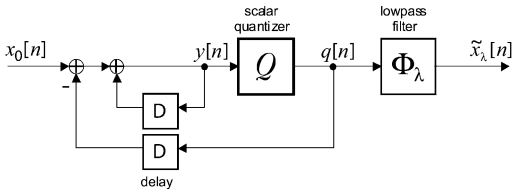


Fig. 1. Block diagram of classical first-order $\Sigma\Delta$ modulation.

The simplest version of $\Sigma\Delta$ modulation is the single-loop (first-order) version originally introduced in [15], which involves an integrator, a single-bit quantizer, and a negative feedback from the quantizer output into the integrator input. The system equations are

$$\begin{cases} y[n] = y[n-1] + x_0[n] - q[n-1] \\ q[n] = Q(y[n]). \end{cases} \tag{1}$$

The block diagram of Fig. 1 symbolizes this system. Here, $x_0[n] = X_0(n/\lambda)$, $n \in \mathbb{Z}$, denotes the sequence of samples of the continuous-time input signal $X_0(t)$ sampled at λ times per time unit. We shall normalize the time unit so that the spacing between the Nyquist-rate samples is equal to one time unit; thus, λ is also equal to the oversampling rate. Throughout the paper, λ will be assumed to take integer values. The signals $y[n]$ and $q[n] = Q(y[n])$ denote the quantizer input and the output, respectively. In this system, the quantizer Q is one-bit; i.e., it outputs values from a discrete set consisting of two values, although multibit quantizers are also used in practice, especially for higher order systems. Unless otherwise specified, the input is assumed to be in $\left[-\frac{1}{2}, \frac{1}{2}\right]$ when the quantization step size is normalized to 1.

It is one of the primary objectives of the theory to understand, as a function of λ, the behavior of the error between the input signal and the approximations given by

$$\tilde{x}_\lambda[n] := (\phi_\lambda * q)[n] := \sum_k \phi_\lambda[k] q[n-k] \tag{2}$$

for suitable low-pass filters ϕ_λ whose number of taps typically grows linearly in λ, thus spanning a uniformly bounded duration of real implementation time. Various norms can be considered for measuring the error signal

$$e_\lambda[n] := x_0[n] - \tilde{x}_\lambda[n] \tag{3}$$

Manuscript received August 23, 2001; revised August 26, 2003. This work was supported in part by the National Science Foundation under Grants DMS-9729992, DMS-0219053, DMS-0219072, CCR-0209431, and by the Francis Robbins Upton Fellowship at Princeton University.

C. S. Güntürk is with the Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA (e-mail: gunturk@cims.nyu.edu).

N. T. Thao is with the Department of Electrical Engineering, City College and Graduate School, City University of New York, New York, NY 10031 USA (e-mail: thao@ee-mail.engr.ccny.cuny.edu).

Communicated by R. Zamir, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2004.826635

such as the supremum norm, defined by

$$\|e_\lambda\|_\infty := \sup_n \left|e_\lambda[n]\right| \tag{4}$$

or an (infinite) time-averaged squared norm, defined by

$$\|e_\lambda\|_{\rm av}^2 := \limsup_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} \left|e_\lambda[n]\right|^2. \tag{5}$$

For continuous-time approximations of the input, one considers smooth interpolations $\tilde{X}_\lambda(t)$ of the sequence $\tilde{x}_\lambda[n]$ at the original time scale, i.e., in the sense that $\tilde{X}_\lambda(n/\lambda) = \tilde{x}_\lambda[n]$. Analogous norms can be defined for the corresponding continuous-time error signal $E_\lambda(t) := X_0(t) - \tilde{X}_\lambda(t)$. Note that in this case one would also have $e_\lambda[n] = E_\lambda(n/\lambda)$.

A priori, the error decay in λ depends on the reconstruction filter, the error measure, and the input signal. The effect of each of these factors is important and interesting in its own right; however, one may safely claim that the effect of the reconstruction filter is understood better than the other two. For the subsequent discussion, the reader may assume that the filters are ideal low-pass; what follows is usually valid for a wide range of filters, though sometimes small modifications may also be necessary.

Still in the simple case of first-order $\Sigma\Delta$ modulation, let us consider the sup-norm first. In the case of constant inputs $x_0[n] = x$, the sup-norm of the error $e_\lambda := e_{\lambda,x}$ has been known for a long time to be bounded by $C\lambda^{-1}$ (see, e.g., [6]) where C does not depend on x; this bound in the same form has been extended to the case of arbitrary band-limited functions as well [4]. Neither of these bounds is sharp, however. For constant inputs, one in fact has $\|e_{\lambda,x}\|_\infty \leq C(x)\lambda^{-2+\epsilon}$ for almost every x (in the sense of Lebesgue measure) where $\epsilon > 0$ may be arbitrarily small [2], [8], [10]. Here, the constant $C(x)$ depends on some fine arithmetical properties of x (in the sense of Diophantine approximations) and is quite irregular (for instance, it is not square integrable in x on any nonzero interval). For arbitrary band-limited functions, a corresponding improvement in the exponent of λ has been found only for the instantaneous error; for each $\epsilon > 0$ and each time instant t, one has $|E_\lambda(t)| \leq C(X_0'(t), \epsilon)\lambda^{-4/3+\epsilon}$ [8], [10].

It is clear that $\|e_\lambda\|_{\rm av}^2 \leq \|e_\lambda\|_\infty^2$; therefore, all upper bounds for the squared sup-norm apply to the time-averaged squared norm as well. In the first-order case with constant inputs, it turns out that these two norms behave somewhat similarly in the sense that time-averaging does not yield any significant gain in the exponent of λ and that $\|e_{\lambda,x}\|_{\rm av}^2 \geq c(x)\lambda^{-4}$ for infinitely many λ [24]. This is not the case for the higher order schemes (which will be defined shortly) and there still remains a large discrepancy—natural or artificial—between the best known exponents of λ in the bounds for these two error norms.

To provide more insight on the size of the error signal, let us now look at the effect of statistical averaging over the values of the constant input. Various mixed-type error norms can be considered depending on how one incorporates the mathematical expectation (taken over the input space) into the norm definition. It is known in the case of uniformly distributed constant inputs x that, the mean (expected) time-averaged squared error (or equivalently, mean-squared time-averaged error) defined by

$$\boldsymbol{E}\left(\|e_{\lambda,x}\|_{\rm av}^2\right) := \int_{-1/2}^{1/2} \|e_{\lambda,x}\|_{\rm av}^2 {\rm d}x \tag{6}$$

is bounded by $C\lambda^{-3}$ both from above and from below [6]. Note that the sup-norm estimate $C\lambda^{-1}$, which is uniform for all values of x, would yield the suboptimal estimate $C\lambda^{-2}$ for this quantity; on the other hand, we also see that the constant $C(x)$ in the improved sup-norm estimate $C(x)\lambda^{-2+\epsilon}$ cannot be square integrable with respect to x for otherwise it would imply an impossible $C\lambda^{-4+\epsilon}$ type upper bound for (6). In fact, a more refined analysis reveals that $C(x)$ is a highly singular function of x which fails to be square integrable on every subinterval of the input range. Note that this does not mean that the error is very large when $C(x)$ is large because the simple sup-norm bound $C\lambda^{-1}$ with the uniform constant C is always valid. It rather means that for each x, the $\lambda^{-2+\epsilon}$ type asymptotics kicks in at a different value of λ. Similar remarks can be made in the case of the mean-squared sup-norm

$$\boldsymbol{E}\left(\|e_{\lambda,x}\|_\infty^2\right) := \int_{-1/2}^{1/2} \|e_{\lambda,x}\|_\infty^2 {\rm d}x \tag{7}$$

as well. This norm is stronger (i.e., larger) than $\boldsymbol{E}\left(\|e_{\lambda,x}\|_{\rm av}^2\right)$, but it turns out that it obeys the only slightly worse upper bound $C\lambda^{-3}\log^2\lambda$ [9]. These results are summarized in the "$k = 1$" (first-order) column of Table I in terms of squared norms.

More complicated multiloop systems incorporate multiple number of integrators and feedbacks, and achieve better system performance as λ is increased [3], [12], [13]. For a kth-order system, the corresponding system equations involve a kth-order difference equation, which may be presented in the prototypical form

$$\Delta^k y[n] = x_0[n] - q[n] \tag{8}$$

where Δ^k denotes the k-fold composition of the standard difference operator Δ defined by $\Delta y[n] = y[n] - y[n-1]$, and $q[n]$ again represents the $\Sigma\Delta$ quantized output signal (possibly up to a shift in time). In the case of a kth-order *stable* $\Sigma\Delta$ modulator (one for which $y[n]$ is bounded), it is proved in [4] that $\|e_\lambda\|_\infty^2 \leq C\lambda^{-2k}$ where the constant C is uniform over the input. They also give the first infinite family of arbitrarily high-order single-bit schemes that are unconditionally stable for arbitrary bounded inputs. Since the sup-norm is the strongest norm among all the norms we consider, an $O(\lambda^{-2k})$-type estimate applies, in particular, to all the mean-squared norms considered above; however, similar to the first-order case, this does not necessarily reflect the optimal behavior of error in these norms. Indeed, it is known for the multibit multiloop configuration with constant inputs that $\|e_\lambda\|_{\rm av}^2$ has an $O(\lambda^{-2k-1})$-type decay [13]. However, the analysis of [13] is restricted to only a special case of $\Sigma\Delta$ modulation with a fixed uniform k-bit quantizer[1] for a kth-order scheme. While using this multibit quan-

[1]By this, we mean that the quantizer has uniformly spaced 2^k output values and each threshold level is the midpoint of an interval defined by these output values.

TABLE I
A COMPARISON OF PREVIOUSLY KNOWN ERROR ESTIMATES FOR $\Sigma\Delta$ MODULATION AND CONTRIBUTION OF THIS PAPER

order		$k=1$	$k=2$	$k \geq 3$	$k \geq 2$, k-loop
# bits		1 bit	1 bit	1 bit	k bits
time-varying inputs $x_0[\cdot]$	squared sup-norm [4]: $\|e_\lambda\|_\infty^2$	$C\lambda^{-2}$	$C\lambda^{-4}$	$C\lambda^{-2k}$	$C\lambda^{-2k}$
	squared instantaneous [9][10]: $\|e_\lambda(t)\|^2$	$C(x_0'(t),\epsilon)\lambda^{-\frac{8}{3}+\epsilon}$			
constant inputs x	squared sup-norm [2][9][10]: $\|e_{\lambda,x}\|_\infty^2$	$C(x)\lambda^{-4+\epsilon}$ a.e. x			
	input-averaged squared sup-norm [9]: $\int_{-\frac{1}{2}}^{\frac{1}{2}} \|e_{\lambda,x}\|_\infty^2 \, dx$	$C\lambda^{-3}\log^2\lambda$			
	time-averaged squared error [12][13][3]: $\|e_{\lambda,x}\|_{av}^2$	$C(x)\lambda^{-4+\epsilon}$ a.e. x	$C(x,\epsilon)\lambda^{-5+\epsilon}$ a.e. x [24]		$C\lambda^{-2k-1}$
	input and time-averaged squared error [6]: $\int_{-\frac{1}{2}}^{\frac{1}{2}} \|e_{\lambda,x}\|_{av}^2 \, dx$	$C\lambda^{-3}$			
	input-averaged instantaneous squared [this paper]: $\int_{-a}^{a} \left\|e_{\lambda,x}[n]\right\|^2 dx$, $a < \frac{1}{2}$		$C\lambda^{-4.5}\log^2\lambda$		
	instantaneous squared [this paper]: $\left\|e_{\lambda,x}[n]\right\|^2$ $\|x\| \leq a < \frac{1}{2}$		$C(x,n)$ $\lambda^{-4.5}\log^{\frac{7}{2}+\delta}\lambda$ a.e. x		

tizer avoids overloading and eases the analysis of the quantization error significantly, it is clearly a nonideal setup for large k, since one of the appeals of $\Sigma\Delta$ modulation is its capability of working with single-bit quantizers, producing one-bit-per-input sample.

In this paper, we allow more general quantizers, including single-bit quantizers. As one important contribution, we analyze $\Sigma\Delta$ schemes for which no better estimate than the one provided by [4] could previously be given. This includes the remaining "1-bit" columns in Table I. Under our generalized framework, we provide two types of squared instantaneous error estimates. The first one involves statistical averaging over the input and is uniform in time, measured by the sup-norm (in time) of the mean-squared instantaneous error

$$\left\|\boldsymbol{E}\left(e_{\lambda,x}^2[\cdot]\right)\right\|_\infty := \sup_n \int_{-a}^{a} \left|e_{\lambda,x}[n]\right|^2 \mathrm{d}x \tag{9}$$

where $(-a,a)$ is an interval of input values that may be restricted by the $\Sigma\Delta$ scheme. For the convenience and simplicity of notation, we shall use the shortcut $\mathrm{MSE}(\lambda)$ (for the generic "mean-squared error") for this particular error measure. Note that this measure satisfies the "sandwich" inequality

$$\int_{-a}^{a} \|e_{\lambda,x}\|_{\mathrm{av}}^2 \mathrm{d}x \leq \mathrm{MSE}(\lambda) \leq \int_{-a}^{a} \|e_{\lambda,x}\|_\infty^2 \mathrm{d}x. \tag{10}$$

For the (single-bit) schemes that we shall consider, the best available bound on $\mathrm{MSE}(\lambda)$ is $O(\lambda^{-4})$, which is provided by the bound in [4] for the much stronger squared sup-norm. We will show in this paper that $\mathrm{MSE}(\lambda)$ obeys an $O(\lambda^{-4.5})$-type bound, up to a logarithmic factor. The second bound we shall provide will be directly on the instantaneous error. We will show that for almost every input x, $\left|e_{\lambda,x}[n]\right|^2$ also obeys the upper bound $O(\lambda^{-4.5})$ up to a logarithmic factor. To the best of our knowledge, there has been no other improved estimates for these schemes yet (however, also see [24]).

For the two-bit configuration, our methods will produce an $O(\lambda^{-5})$ type bound, again up to a logarithmic factor. This result provides us with a rate estimate that matches the estimate for the time-averaged squared error norm; however, note that due to (10), neither of these results imply (i.e., is stronger than) the other one.

Due to the increased complexity of the analysis, we shall restrict this paper to second-order systems with constant inputs. Our methods, however, are not limited to second-order schemes only, but to a large class of arbitrary order modulators [24]. We also believe that the new techniques we introduce will prove to be very useful for time-varying inputs.

The paper is organized as follows. In Section II, we derive the basic equations and formulas for the time evolution of signals in second-order $\Sigma\Delta$ modulators. In particular, we express the reconstruction error in terms of the state vector of the system. At this point, the main obstacle against pushing the derivations further is the absence of an explicit expression of any of the node signals of the $\Sigma\Delta$ modulator, including its output and its state

vector, which is basically due to the nonlinear recursion embedded in the $\Sigma\Delta$ modulator. A first contribution of this paper is the introduction and the exploitation of a new remarkable property of $\Sigma\Delta$ modulators which, in principle, enables an explicit derivation of its output and its state-vector sequence. This property, which we call the *tiling property*, refers to the fact that the state vector remains in a set Γ_x that *tiles* the space by $\mathbb{Z}^2$ translations. We give the exact definition of this property in Section III, show experimental evidence of it, give mathematical justifications, and derive from the knowledge of Γ_x an explicit expression of the state vector in terms of x and n. While the existence of the tile Γ_x is clearly demonstrated by experiment and proved mathematically (up to finite multiplicity of the tile), detailed parametrizations of these tiles are not known in general. Further knowledge on these parameters requires explicit analyses of given configurations. We study in Section IV three different configurations for which a thorough analysis has been feasible. Part of this analysis involves the study of geometric regularity of the invariant sets as carried out in Section IV-D, which will turn out to be important for the improved error estimates mentioned above that we derive in Section V. These estimates depend heavily on the general machinery of the theory of uniform distribution [5], [16]. Appendix, part A is specially dedicated to the basic elements of this theory as utilized in this paper. We conclude the paper with further remarks and future research directions (Section VI).

II. Equations of the Second-Order Modulator

A. Feedback Equations and an Equivalent System

The generic architecture of a classical second-order $\Sigma\Delta$ modulator is shown in Fig. 2.[2] It can be easily derived from the block diagram that this system satisfies the second-order difference equation

$$\begin{aligned}\Delta^2 y[n] &= x_0[n] - (\alpha+\beta)q[n-1] + \beta q[n-2] \\ &= x_0[n] - (\alpha+\beta\Delta)q[n-1]. \end{aligned} \tag{11}$$

In the standard case of the double-loop configuration studied in [13] where $\alpha = \beta = 1$, this equation can be rewritten as a difference equation for the quantizer error $y[n] - q[n]$

$$\Delta^2(y[n] - q[n]) = x_0[n] - q[n]. \tag{12}$$

However, in the general case, a direct signal analysis of the system of Fig. 2 is difficult due to the complicated action of the feedback. We shall first derive an equivalent diagram of this system which yields simpler feedback mechanisms. Consider the change of input variable defined by the difference equation

$$x_0[n] = \gamma\Delta^2 x_1[n] + (\alpha+\beta\Delta)x_1[n-1] \tag{13}$$

where γ is a parameter to be chosen at our disposal. Next, define the auxiliary variables $u_1[n]$ and $u_2[n]$ to satisfy the difference equations

$$\Delta u_2[n] = u_1[n]; \quad \Delta u_1[n] = x_1[n] - q[n]. \tag{14}$$

[2]There are more general configurations that contain extra feedbacks from the quantizer input as well [19].

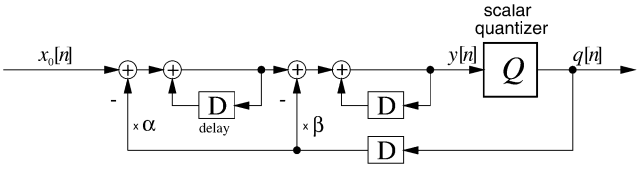


Fig. 2. Block diagram of classical second-order $\Sigma\Delta$ modulation.

Then, by subsequently applying (14), (11), and (13), it follows that

$$\begin{aligned}&\Delta^2(\alpha u_2[n-1] + \beta u_1[n-1] + \gamma x_1[n]) \qquad (15)\\ &\quad = (\alpha+\beta\Delta)(x_1[n-1] - q[n-1]) + \gamma\Delta^2 x_1[n] \\ &\quad = (\alpha+\beta\Delta)x_1[n-1] + \Delta^2 y[n] - x_0[n] + \gamma\Delta^2 x_1[n] \\ &\quad = \Delta^2 y[n]. \qquad (16)\end{aligned}$$

Assuming that the initial conditions for $x_1[n]$ have already been picked (arbitrarily, or by some criterion), the initial conditions for the sequences $u_1[n]$ and $u_2[n]$ can now be chosen so that (16) implies

$$\begin{aligned} y[n] &= \beta u_1[n-1] + \alpha u_2[n-1] + \gamma x_1[n] \\ &= T(u_1[n-1], u_2[n-1], x_1[n]) \end{aligned} \tag{17}$$

where

$$T(u_1, u_2, x) := \beta u_1 + \alpha u_2 + \gamma x. \tag{18}$$

Since

$$q[n] = Q(y[n]) \tag{19}$$

the signal $y[n]$ can now be thought of resulting from $x_1[n]$ through a new dynamical system shown in Fig. 3. In this system, the feedback loop simply carries the input–output difference $x_1[n] - q[n]$ at every instant and the remainder of the system uses this value to produce the next signal value $y[n+1]$ to be quantized.

Given the construction of this dynamical system, the complete $\Sigma\Delta$ modulation process can then be equivalently described as the transformation of $x_0[n]$ into $\tilde{x}_\lambda[n]$ through the sequence of (2), (13), (17), and (19). Thus, the signal processing of $\Sigma\Delta$ modulation based on the architecture in Fig. 2 can be represented by the block diagram of Fig. 4, where the block labeled "dynamical system" symbolizes the system of Fig. 3. The operator G is basically a recursive filter that transforms $x_0[n]$ into $x_1[n]$ through the difference (13). Note that in this setup, the operator G and the signal $x_1[n]$ appear by mathematical construction and do not necessarily exist physically in an actual implementation given by Fig. 2. We also deduce that the effect of γ as a parameter in G is cancelled out in the system of Fig. 3. For the realizability of this equivalence, we will assume that the parameters α and β have been chosen such that the operator G is stable for some γ.

Let us note at this stage that while the $\Sigma\Delta$ modulator described by Fig. 2 is favorable for the efficiency of its circuit implementation, it is also a legitimate option to switch to the slightly less efficient $\Sigma\Delta$ modulator scheme described solely

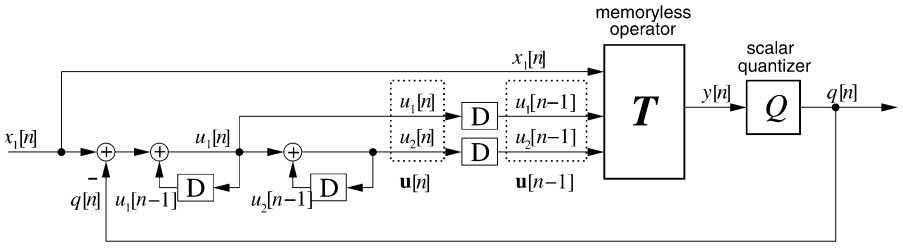


Fig. 3. Alternative representation of the pure feedback process of second-order ΣΔ modulation.

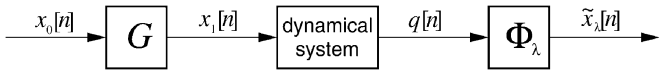


Fig. 4. Global signal processing chain of ΣΔ modulation: the first block is characterized by the difference equation of (13), the second block represents the system of Fig. 3, and the third block represents the convolution operation of (2). The first two blocks combined together generalizes the original second-order ΣΔ system of Fig. 2.

by Fig. 3 (i.e., without the prefilter G) when circuit implementation is not the primary concern. In this case, γ would be an additional parameter of design. In fact, there would be a whole range of flexibility in the choice of T if nonlinear functions are also allowed. We shall return to this issue in Section II-E.

B. The Quantizer

We assume in this paper that the quantizer Q is uniform of step size 1, in the sense that its output values are of the type $i-\frac{1}{2}$ where $i=i_0, i_0+1, \ldots, i_1$. The quantization intervals I^i which satisfy $Q(I^i)=i-\frac{1}{2}$ are defined by

$$I^i := \begin{cases} (-\infty, i_0), & i=i_0 \\ [i-1, i), & i_0 < i < i_1 \\ [i_1-1, +\infty), & i=i_1. \end{cases} \tag{20}$$

We call the quantizer k-bit if $i_1-i_0+1=2^k$. In the particular one-bit case, we assume that $i_0=0$, so that the quantizer mapping reduces to

$$Q(y)=\begin{cases} -\frac{1}{2}, & \text{if } y<0 \\ +\frac{1}{2}, & \text{if } y\geq 0. \end{cases} \tag{21}$$

We call the quantizer infinite if $i_0=-\infty$ and $i_1=+\infty$.

We say that the quantizer is overloaded if $|y-Q(y)|>\frac{1}{2}$. Note that the infinite quantizer is never overloaded.

C. State-Space Equations

At every instant, $u_1[n]$ and $u_2[n]$ constitute the state variables of the system. We will use the shorthand notation

$$\boldsymbol{u}[n]=\begin{pmatrix} u_1[n] \\ u_2[n] \end{pmatrix} \tag{22}$$

to represent the vector state of the system. The full recursive system equations of the block diagram of Fig. 3 is then

$$\begin{cases} q[n]=Q(T(\boldsymbol{u}[n-1], x_1[n])) \\ \boldsymbol{u}[n]=\boldsymbol{A}\boldsymbol{u}[n-1]+(x_1[n]-q[n])\boldsymbol{e} \end{cases} \tag{23}$$

where

$$\boldsymbol{A}=\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{e}=\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{24}$$

For each real number ξ, we define a partition $\{\Omega_\xi^i, i=i_0, \ldots, i_1\}$ of $\mathbb{R}^2$ by setting $\Omega_\xi^i=\{\boldsymbol{u} : T(\boldsymbol{u}, \xi)\in I^i\}$. Let $\boldsymbol{M}_\xi^i : \mathbb{R}^2 \to \mathbb{R}^2$ denote the affine transformation defined by

$$\boldsymbol{M}_\xi^i(\boldsymbol{u}) := \boldsymbol{A}\boldsymbol{u}+\left(\xi-i+\frac{1}{2}\right)\boldsymbol{e} \tag{25}$$

and $\boldsymbol{M}_\xi : \mathbb{R}^2 \to \mathbb{R}^2$ denote the *piecewise* affine transformation defined by

$$\boldsymbol{M}_\xi(\boldsymbol{u})=\boldsymbol{M}_\xi^i(\boldsymbol{u}), \qquad \text{if } \boldsymbol{u}\in\Omega_\xi^i. \tag{26}$$

With this notation, the recursive equations of (23) can be rewritten as

$$\boldsymbol{u}[n]=\boldsymbol{M}_{x_1[n]}(\boldsymbol{u}[n-1]). \tag{27}$$

D. Basic Error Analysis

Suppose we use the system of Fig. 2 and we would like to compute an approximation of the input $x_0[n]$ via the convolution $\tilde{x}_\lambda=\phi_\lambda * q$. Since the overall result is equivalently described by the signal processing chain of Fig. 4, it is natural to consider Φ_λ of the form $\Phi_\lambda=G^{-1}H_\lambda$ to remove the prefiltering effect of G. With this choice, all we need to satisfy is that H_λ is a suitable reconstruction filter for the system of Fig. 3 for the input $x_1[n]$. Indeed, if we know that $x_1-h_\lambda * q$ is small, then

$$\begin{aligned} x_0-\phi_\lambda * q &= g^{-1} * x_1 - g^{-1} * h_\lambda * q \\ &= g^{-1} * (x_1 - h_\lambda * q) \end{aligned} \tag{28}$$

will also be small since g^{-1} is a (causal) final inpulse response (FIR) filter of at most three taps as defined in (13). For the error analysis, it therefore suffices to consider the system of Fig. 3 only.

Now the error signal $x_1-h_\lambda * q$ for the system of Fig. 3 can be written as

$$\begin{aligned} x_1-h_\lambda * q &= x_1 - h_\lambda * x_1 + h_\lambda * (x_1-q) \\ &= e_\lambda^{[1]} + e_\lambda^{[2]} \end{aligned} \tag{29}$$

where the first error component

$$e_\lambda^{[1]} := x_1 - h_\lambda * x_1 \tag{30}$$

is a signal that does not depend on the quantization procedure and can be made arbitrarily small (in fact, even zero) by chosing h_λ suitably, and the second error component

$$e_\lambda^{[2]} := h_\lambda * (x_1 - q) \tag{31}$$

corresponds to the in-band portion of the "quantization error" signal $x_1[n] - q[n]$. It is this second error component that constitutes the center of interest of $\Sigma\Delta$ error analysis since, unlike the first one, it is highly nonlinear in the input.

In the particular case when the input is a constant signal $x_1[n] = x$ (as will be the case for the rest of the paper) we do not even have to worry about the first error component $e_\lambda^{[1]}$ since we can eliminate it completely by restricting h_λ to filters that satisfy $\sum_n h_\lambda[n] = 1$. It therefore causes no ambiguity to denote $e_\lambda^{[2]}$ by e_λ. Substituting $\Delta^2 u_2[n] = x_1[n] - q[n]$ from (14) and changing the order of convolution and differentiation yields the formula

$$e_\lambda = \Delta^2 h_\lambda * u_2. \tag{32}$$

When the dynamical system associated with the map $\boldsymbol{M}_x$ is stable (see Section IV), a basic decay estimate immediately follows. Indeed, stability implies that u_2 is bounded, which leads to the simple error bound

$$\|e_\lambda\|_\infty \le \|\Delta^2 h_\lambda\|_1 \|u_2\|_\infty. \tag{33}$$

This basic estimate cannot provide any decay rate better than $O(\lambda^{-2})$. To see this, recall first that the number of taps of h_λ was required to grow linearly in λ; this implies that $\|\Delta^2 h_\lambda\|_1 \ge c/\lambda^2$ for some constant $c > 0$. Indeed, let h be an L-tap filter with $h[n] = 0$ for $n < 0$ and $n \ge L$. Define

$$s[n] = \frac{1}{2}\left(n - \frac{L-3}{2}\right)^2.$$

Using $\Delta^2 s[n] = 1$, and applying summation by parts twice, one obtains

$$\begin{aligned} 1 &= \sum h[n]\Delta^2 s[n] \\ &= \sum s[n-2]\Delta^2 h[n] \\ &\le \left(\max_{0\le n\le L+1} |s[n-2]|\right) \sum_{n=0}^{L+1} |\Delta^2 h[n]| \\ &\le \frac{(L+1)^2}{8} \|\Delta^2 h\|_1 \end{aligned}$$

hence, the proof of the claim.

On the other hand, the upper bound $O(\lambda^{-2})$ is easily achieved by imposing some smoothness on h_λ. Let ρ_λ be the rectangular filter of length λ given by $\rho_\lambda[n] = \frac{1}{\lambda}$ if and only if $0 \le n < \lambda$. Consider

$$h_\lambda = \rho_\lambda * \rho_\lambda * \bar{h}_\lambda \tag{34}$$

where $\bar{h}_\lambda$ is any filter with linearly growing number of taps in λ and that satisfies $\sum \bar{h}_\lambda[n] = 1$ with $\|\bar{h}_\lambda\|_1 \le C$ for some absolute constant C. Clearly, we have $\sum h_\lambda[n] = 1$ as well. Note that $\Delta\rho_\lambda = \frac{1}{\lambda}(\delta_0 - \delta_\lambda)$ where δ_a denotes the sequence defined by $\delta_a[n] = \delta[n-a]$. This implies

$$\Delta^2 h_\lambda = \frac{1}{\lambda^2}(\delta_0 - 2\delta_\lambda + \delta_{2\lambda}) * \bar{h}_\lambda$$

and we obtain

$$\|\Delta^2 h_\lambda\|_1 \le \frac{1}{\lambda^2}\|\delta_0 - 2\delta_\lambda + \delta_{2\lambda}\|_1 \|\bar{h}_\lambda\|_1 \le \frac{4C}{\lambda^2}; \tag{35}$$

therefore, (33) implies $\|e_\lambda\|_\infty = O(\lambda^{-2})$.

The simplest choice for $\bar{h}_\lambda$ would be δ_0. In this case, h_λ is the triangular filter, i.e., the second-order discrete B-spline more commonly known in the circuit community as the $sinc^2$ filter due to its frequency-domain representation. It has been found, however, that the error decays faster than $O(\lambda^{-2})$ with the choice $\bar{h}_\lambda = \rho_\lambda$, in which case h_λ is the $sinc^3$ filter. To explain this phenomenon, we return to the exact error expression (32) which now yields the decomposition

$$\begin{aligned} e_\lambda &= \frac{1}{\lambda^2}(\delta_0 - 2\delta_\lambda + \delta_{2\lambda}) * (\rho_\lambda * u_2) \\ &= \frac{1}{\lambda^2}(\delta_0 - 2\delta_\lambda + \delta_{2\lambda}) * \bar{u}_\lambda \end{aligned} \tag{36}$$

where $\bar{u}_\lambda$ is the sequence of running averages defined by

$$\bar{u}_\lambda[n] := \rho_\lambda * u_2[n] = \frac{1}{\lambda}\sum_{m=0}^{\lambda-1} u_2[n-m]. \tag{37}$$

When λ is large, it is expected that the signal $\bar{u}_\lambda[n]$ will vary more slowly than $u_2[n]$ due to the long time averaging. In fact, if some form of "central limit theorem" could be shown to hold for $u_2[n]$, this would force $\bar{u}_\lambda[n]$ to be mostly concentrated about a mean value. Note that $(\delta_0 - 2\delta_\lambda + \delta_{2\lambda})$ is a difference operator which would bring out the residual value around this mean when convolved with $\bar{u}_\lambda$. Therefore, e_λ would at most vary as this residual value (up to the multiplicative factor $\frac{1}{\lambda^2}$). This additional cancellation provides an intuitive justification for using the $sinc^3$ filter. The quantification of this idea, which will be essential in the derivation of our improved estimates mentioned in Section I, is contained in much of the rest of this paper. Qualitatively, these results may be viewed as originating from the ergodicity of the mappings $\boldsymbol{M}_x$ with respect to the Lebesgue measure on certain invariant sets. On the other hand, quantitative results will depend heavily on fine analytic and algebraic properties of these invariant sets, which will be stated in Section IV.

It is known that the $sinc^3$ filter can be further improved by some coefficient modifications, yielding smaller multiplicative coefficients in the error [12]. However, in this paper, we shall stick to the $sinc^3$ filter, as it is simple to implement, and it captures the essence of our methods best. Let us briefly mention here that nonlinear reconstruction has been demonstrated to yield faster error decays [22], [25] but is not used in practice for complexity reasons.

E. Nonlinear Functions T

The error decay guaranteed by the inequality of (33) actually gives us more information than just an error bound. Note that it assumes no condition on the nature of the functions T and Q except that they have been designed to ensure that $\boldsymbol{u}$ is bounded. Theoretically, the constraint that T be a linear function is artificial as a design criterion since the composed operator $Q \circ T$ is in *any case* nonlinear because of Q. Moreover, note that it is globally the composed operator $Q \circ T$ that differentiates the stability properties of one modulator from another. We will see in Section III that relaxing the linearity of the function T will enable us to uncover general properties of the dynamical systems that are analytically unreachable with linear functions T. Also, although the feasibility of nonlinear functions T with regard to analog circuit implementation is still an unanswered question, a prototype of one-bit second-order modulator with a quadratic

function T was numerically demonstrated in [23] to have superior performances to the one-bit linear-T modulators. This prototype will be introduced in Section III.

III. Invariant Tiles Under Constant Inputs

The error relations (36) and (37) require a refined analysis of the state vector $\boldsymbol{u}[n]$. The fundamental difficulty is that the sequence $\boldsymbol{u}[n]$ is not known explicitly in terms of $x_1[n]$. As can be seen in (27), $\boldsymbol{u}[n]$ is only recursively determined in terms of $\boldsymbol{u}[n-1]$ with $x_1[n]$ as a varying parameter. The scope of this paper is the error analysis under constant inputs $x_1[n] = x, \forall\, n$. In this situation, $\boldsymbol{u}[n]$ recursively depends on $\boldsymbol{u}[n-1]$ through the fixed mapping $\boldsymbol{M}_x$, i.e.,

$$\boldsymbol{u}[n] = \boldsymbol{M}_x(\boldsymbol{u}[n-1]). \tag{38}$$

The key to the analysis lies in the study of the map $\boldsymbol{M}_x$.

A. Experimental Observation of Tiling

We start with the description of a particular experiment that led to the discovery of a remarkable property of the maps $\boldsymbol{M}_x$. For various second-order $\Sigma\Delta$ modulators, we plot in black in Fig. 5 several consecutive iterates $\boldsymbol{u}[n]$ of a fixed initial condition $\boldsymbol{u}[0]$ under the map $\boldsymbol{M}_x$, where x is a fixed constant input. In these plots, x is chosen to be irrational; we will return to this issue later. For each modulator, one can observe in this plot that the state points remain in (and fill out) a certain deterministic set $\Gamma := \Gamma_x$. However, there is more to this set in that in every case its integer ($\mathbb{Z}^2$) translations appear to *tile* the plane. We highlight this fact in the figure by representing the translates of the points $\boldsymbol{u}[n]$ by $(1,0)$ and $(1,1)$ in two gray tones, respectively. Formally, we say that a set Γ is a tile when for each point $\boldsymbol{v} \in \mathbb{R}^2$, there is a unique point $\boldsymbol{v}' \in \Gamma$ such that $\boldsymbol{v} - \boldsymbol{v}' \in \mathbb{Z}^2$. This is equivalent to the fact that the family $\{\Gamma + \boldsymbol{k}\}_{\boldsymbol{k}\in\mathbb{Z}^2}$ forms a partition of $\mathbb{R}^2$.

Since the initial observation of this phenomenon [9], it has been systematically confirmed on any stable second-order modulator employing a quantizer with uniformly spaced output levels as assumed in this paper. In the cases of Fig. 5(a)–(d), the standard linear rule $T(u_1, u_2, x) = u_1 + u_2 + x$ is used with different versions of the quantizer. Fig. 5(a) is the case of the nonoverloaded ideal quantizer (infinite quantizer). In Fig. 5(b), we use the three-level quantizer introduced in [26] which employs $-1, 0, 1$ as the output levels and $\frac{1}{2}$, $-\frac{1}{2}$ as the threshold values. Fig. 5(c) is the standard one-bit quantizer case. In Fig. 5(d), we use an infinite quantizer whose threshold at 0 has been deviated by $+\frac{1}{3}$. Fig. 5(e) shows the case of a different linear rule $T(u_1, u_2, x) = u_1 + \frac{1}{2}u_2 + x$ with the regular infinite quantizer. Finally, Fig. 5(f) shows the case of the following "semilinear" rule introduced in [23]:

$$\begin{aligned} &T(u_1, u_2, x) \\ &\quad = (9 - 6|x|)u_1 + (6 - 12|x|)u_2 + (10 - 4|x|)x. \end{aligned} \tag{39}$$

Formally speaking, these experiments demonstrate the existence of a tile Γ_x that contains the forward trajectory $\mathcal{U} = \{\boldsymbol{u}[n]\}_{n\geq 0}$. Now, the recursive relation (38) implies that $\boldsymbol{M}_x(\mathcal{U}) \subset \mathcal{U}$. Since $\mathcal{U}$ appears to be dense in Γ_x, one is then tempted to conjecture that $\boldsymbol{M}_x(\Gamma_x) \subset \Gamma_x$. Once proved, this result implies that if the initial initial state $\boldsymbol{u}[0] \in \Gamma_x$, then all forward trajectories will be known to remain in Γ_x. However, the real significance of this result lies in the tiling property as will be explained in Section III-C.

B. Mathematical Justifications of Tiling

Some mathematical justifications of the above tiling conjecture have been recently provided in [24] under the assumption of stability. We call the dynamical system defined by a map $\boldsymbol{M}$ on $\mathbb{R}^m$ *positively stable* if there exists a bounded set Γ_0 satisfying $\boldsymbol{M}(\Gamma_0) \subset \Gamma_0$. We call such a set Γ_0 *positively invariant*. Certainly, the existence of such a set ensures stable operation of the modulator and there has been interest in finding such sets [21]. We are interested in invariant sets that are also tiles. It turns out that if a positively invariant set can be found for $\boldsymbol{M}$, then consecutive iterations of this set under $\boldsymbol{M}$ converge to an attractor set that is a tile, up to a multiplicity. In the following theorem, we summarize the results obtained in [24] in this direction.

Theorem 3.1 ([24]): Given a finite measurable partition $\{\Omega^i\}_i$ of $\mathbb{R}^m$, a collection of integer vectors $\boldsymbol{e}_i \in \mathbb{Z}^m$ and an irrational constant $x \in \mathbb{R}$, consider the piecewise affine map $\boldsymbol{M}$ on $\mathbb{R}^m$ defined by

$$\boldsymbol{M}(\boldsymbol{u}) := \boldsymbol{A}\boldsymbol{u} + \left(x + \frac{1}{2}\right)\boldsymbol{e} - \boldsymbol{e}_i, \qquad \text{if } \boldsymbol{u} \in \Omega^i \tag{40}$$

where $\boldsymbol{A}$ is the $m \times m$ lower triangular matrix of 1's and $\boldsymbol{e} = (1, \ldots, 1) \in \mathbb{Z}^m$. If there exists a bounded set Γ_0 of positive measure that is positively invariant under $\boldsymbol{M}_x$, then the set

$$\Gamma := \bigcap_{k\geq 0} \boldsymbol{M}^k(\Gamma_0) \tag{41}$$

is invariant by $\boldsymbol{M}$ (i.e., $\boldsymbol{M}(\Gamma) = \Gamma$) and is equal (up to a set of measure zero) to the disjoint union of a finite and nonempty collection of tiles.

A number of remarks is in order. First, note that the mapping defined in (24)–(26) is indeed of the form of (40). Second, not only does this theorem state the existence of an invariant set Γ, but (41) shows that Γ is an attractor of $\boldsymbol{M}$ within the region of stability Γ_0. Next, note that this theorem is valid in any dimensions m and under general conditions on the partition $\{\Omega^i\}_i$ as well as the integer translations $\boldsymbol{e}_i$, as long as the overall map $\boldsymbol{M}$ is positively stable. However, under these general assumptions, the conclusion is only that Γ is composed of one or more tiles, all up to a set of measure zero. Indeed, an example given in [24] shows that a map of the type (26) may yield an invariant set composed of two tiles. (Let us note that this example required the use of a particular nonlinear thresholding function T.) The exact conditions on T to yield a single tile are not currently known. From our experience (including, for example, the experiments of Fig. 5), we believe that all stable $\Sigma\Delta$ modulators using a linear thresholding function T yield a single invariant tile, at least at the second order and *including* the case of rational input constants x. However, care must be taken in the definition of the invariant set Γ when x is a rational number; the statement is on the existence of a tile Γ that is invariant under $\boldsymbol{M}$ and it may not necessarily be the case that Γ can be found as an attractor as in (41) or the closure of any trajectory. It remains a general conjecture that linear thresholding functions enjoy these properties.

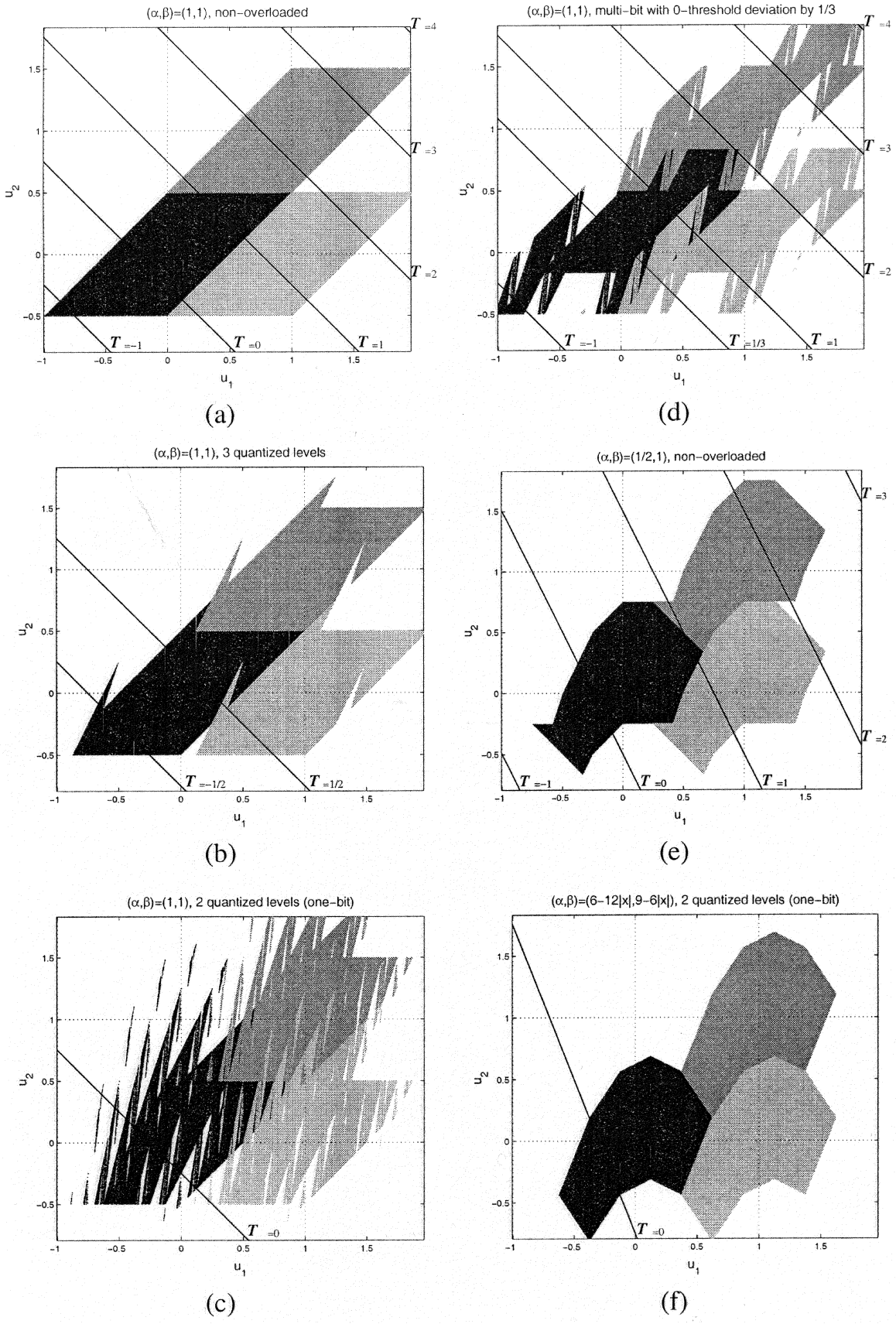


Fig. 5. Representation in black of several consecutive state points of various second-order $\Sigma\Delta$ modulators with an irrational constant input $x \simeq 1/4$. The copies in gray are the translated versions of the state points by $(1, 0)$ and $(1, 1)$, respectively.

In Section IV, we will give the proof of these properties on three particular configurations of second-order ΣΔ modulation. But before performing this analysis, we would like to show why the single-tile case is of crucial importance.

C. The Single Invariant Tile Case and its Fundamental Consequence

From now on, we only consider ΣΔ modulators for which the invariant set Γ_x is a single tile for each x. We shall see in this case that it is possible to find an explicit expression for $\boldsymbol{u}[n]$ in terms of n and Γ_x. To keep the discussion simple, we shall restrict ourselves to second-order modulators; the generalization to higher order modulators is routine.

We first introduce some notation. Let Γ be an arbitrary tile in $\mathbb{R}^2$. By definition, the collection of sets $\{\Gamma + \boldsymbol{k}\}_{\boldsymbol{k}\in\mathbb{Z}^2}$ form a partition of $\mathbb{R}^2$. This implies that for each $\boldsymbol{u} \in \mathbb{R}^2$, there exists a unique point in Γ, denoted $\langle \boldsymbol{u} \rangle_\Gamma$, such that

$$\langle \boldsymbol{u} \rangle_\Gamma - \boldsymbol{u} \in \mathbb{Z}^2. \tag{42}$$

In other words, $\boldsymbol{u} \mapsto \langle \boldsymbol{u} \rangle_\Gamma$ is the unique map from $\mathbb{R}^2$ to Γ that satisfies

$$\forall\, \boldsymbol{u} \in \Gamma, \qquad \langle \boldsymbol{u} \rangle_\Gamma = \boldsymbol{u} \tag{43}$$

and

$$\forall\, \boldsymbol{u} \in \mathbb{R}^2,\ \forall\, \boldsymbol{k} \in \mathbb{Z}^2, \qquad \langle \boldsymbol{u} + \boldsymbol{k} \rangle_\Gamma = \langle \boldsymbol{u} \rangle_\Gamma. \tag{44}$$

In the simple case where $\Gamma = [0,1)^2$, we will use the standard notation $\langle \boldsymbol{v} \rangle$ to denote $\langle \boldsymbol{v} \rangle_{[0,1)^2}$, where

$$\langle \boldsymbol{v} \rangle = \begin{pmatrix} \langle v_1 \rangle \\ \langle v_2 \rangle \end{pmatrix}. \tag{45}$$

Here $\langle v \rangle := v - \lfloor v \rfloor$ denotes the fractional part of a real number v, and $\lfloor v \rfloor$ denotes the greatest integer less than or equal to v.

We return to the sequence $\boldsymbol{u}[n]$ of the ΣΔ state vector which remains in Γ_x for all n. From (26) and (27), we can write

$$\boldsymbol{u}[n] = \left(\boldsymbol{A}\boldsymbol{u}[n-1] + \left(x + \frac{1}{2}\right)\boldsymbol{e} \right) - i\boldsymbol{e} \tag{46}$$

where $i \in \mathbb{Z}$. Since $-i\boldsymbol{e} \in \mathbb{Z}^2$, we obtain via (43) and (44) that

$$\boldsymbol{u}[n] = \left\langle \boldsymbol{A}\boldsymbol{u}[n-1] + \left(x + \frac{1}{2}\right)\boldsymbol{e} \right\rangle_{\Gamma_x}. \tag{47}$$

At the same time, let us artificially build a closely related sequence $\boldsymbol{v}[n]$, which we recursively define by

$$\boldsymbol{v}[n] = \boldsymbol{A}\boldsymbol{v}[n-1] + \left(x + \frac{1}{2}\right)\boldsymbol{e} \tag{48}$$

with the initial state $\boldsymbol{v}[0] = \boldsymbol{u}[0]$. We have the following property.

Proposition 3.2: For all n, we have

$$\boldsymbol{u}[n] = \langle \boldsymbol{v}[n] \rangle_{\Gamma_x}. \tag{49}$$

Proof: Since $\langle \boldsymbol{v} \rangle_{\Gamma_x} - \boldsymbol{v} \in \mathbb{Z}^2$ and $\boldsymbol{A}$ is a matrix with all integer coefficients, we have $\boldsymbol{A}\langle \boldsymbol{v} \rangle_{\Gamma_x} - \boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}(\langle \boldsymbol{v} \rangle_{\Gamma_x} - \boldsymbol{v}) \in \mathbb{Z}^2$. It follows from (44) that

$$\langle \boldsymbol{A}\langle \boldsymbol{v} \rangle_{\Gamma_x} + \boldsymbol{w} \rangle_{\Gamma_x} = \langle \boldsymbol{A}\boldsymbol{v} + \boldsymbol{w} \rangle_{\Gamma_x} \tag{50}$$

for any $\boldsymbol{w} \in \mathbb{R}^2$. The proposition is then proved by induction. For $n = 0$, we have $\boldsymbol{u}[0] = \langle \boldsymbol{u}[0] \rangle_{\Gamma_x} = \langle \boldsymbol{v}[0] \rangle_{\Gamma_x}$. Suppose (49) holds for $k = n-1$, i.e., $\boldsymbol{u}[n-1] = \langle \boldsymbol{v}[n-1] \rangle_{\Gamma_x}$. Then, by successively applying (47), (48), and (50), we obtain

$$\begin{aligned} \boldsymbol{u}[n] &= \left\langle \boldsymbol{A}\langle \boldsymbol{v}[n-1] \rangle_{\Gamma_x} + \left(x + \frac{1}{2}\right)\boldsymbol{e} \right\rangle_{\Gamma_x} \\ &= \left\langle \boldsymbol{A}\boldsymbol{v}[n-1] + \left(x + \frac{1}{2}\right)\boldsymbol{e} \right\rangle_{\Gamma_x} \\ &= \langle \boldsymbol{v}[n] \rangle_{\Gamma_x}. \end{aligned}$$

□

The power of this result lies in the fact that there is an explicit functional expression for $\boldsymbol{v}[n]$, which can be obtained by simply iterating (48) forward and backward. For all $n \in \mathbb{Z}$, we have

$$\begin{aligned} \boldsymbol{v}[n] &= \begin{pmatrix} v_1[n] \\ v_2[n] \end{pmatrix} \\ &= \begin{pmatrix} u_1[0] + n\left(x + \frac{1}{2}\right) \\ n u_1[0] + u_2[0] + \frac{1}{2}n(n+1)\left(x + \frac{1}{2}\right) \end{pmatrix}. \end{aligned} \tag{51}$$

Thus, under the assumption that the tile Γ_x is known, the combination of (49) and (51) provides an explicit expression of $\boldsymbol{u}[n]$ in terms of n.

We define $\boldsymbol{u}[n]$ for $n < 0$ by (49). It follows from the invariance of Γ_x under $\boldsymbol{M}_x$ that this definition is consistent with (38) for all $n \in \mathbb{Z}$.

Fig. 6(a) gives a graphical example of explicit determination of the sequence $\boldsymbol{u}[n]$ from the knowledge of the tile, via the preliminary calculation of the sequence $\boldsymbol{v}[n]$ from (51).

D. Further Developments on the Single-Tile Case

A remaining major difficulty of analysis is that the expression (49) for $\boldsymbol{u}[n]$ depends on the knowledge of the invariant set Γ_x. Not only that this set can be complex as in some of the examples in Fig. 5, but also explicit expressions are, in general, not easily obtainable. Nevertheless, an analysis of $\boldsymbol{u}[n]$ is still possible, thanks to a particular decomposition of $\langle \cdot \rangle_{\Gamma_x}$ into simpler components. This is based on the following lemma.

Lemma 3.3: Let Γ and Γ' be two sets that tile the plane with $\mathbb{Z}^2$ translations. For each $\boldsymbol{k} \in \mathbb{Z}^2$, let us define the set $\Pi_{\boldsymbol{k}} := \{\boldsymbol{u} : \langle \boldsymbol{u} \rangle_\Gamma \in \Gamma' - \boldsymbol{k}\}$. Then

i) the family $\{\Pi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathbb{Z}^2}$ forms a partition of $\mathbb{R}^2$;

ii) $\langle \boldsymbol{u} \rangle_{\Gamma'} - \langle \boldsymbol{u} \rangle_\Gamma = \boldsymbol{k}$ when $\boldsymbol{u} \in \Pi_{\boldsymbol{k}}$.

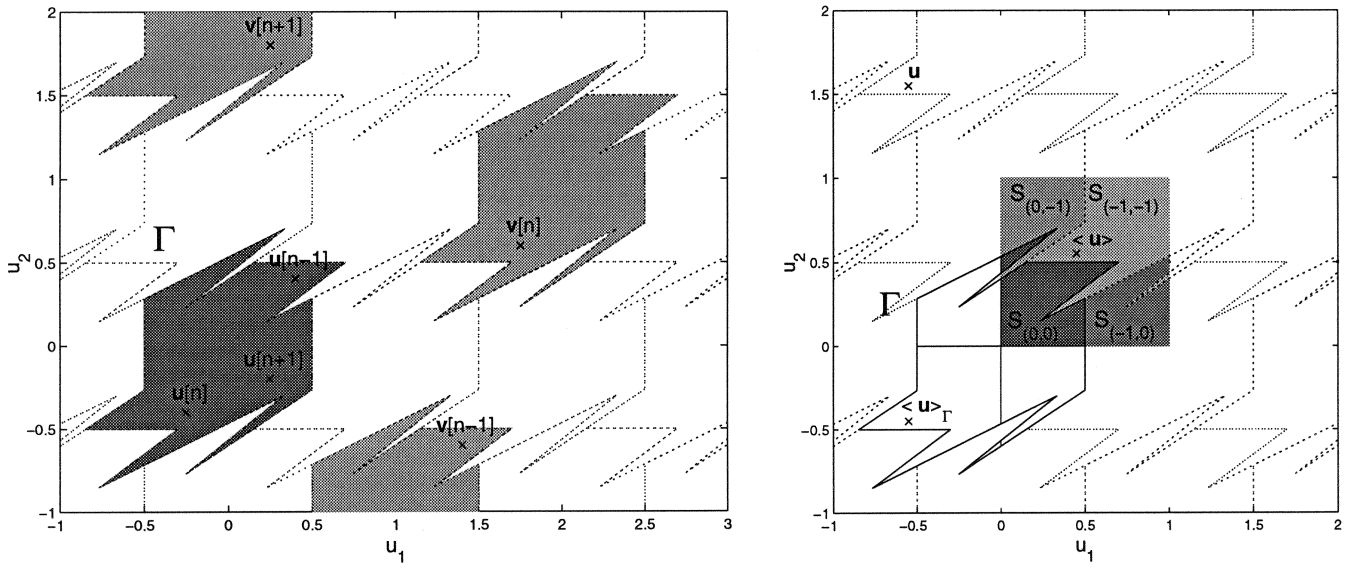
Proof: Since Γ' tiles the plane with $\mathbb{Z}^2$ translations, for any $\boldsymbol{u} \in \mathbb{R}^2$, there exists a unique $\boldsymbol{k} \in \mathbb{Z}^2$ such that $\langle \boldsymbol{u} \rangle_\Gamma \in \Gamma' - \boldsymbol{k}$. This proves part i). Now, consider any given $\boldsymbol{k} \in \mathbb{Z}^2$ and any $\boldsymbol{u} \in \Pi_{\boldsymbol{k}}$. By definition, $\langle \boldsymbol{u} \rangle_\Gamma \in \Gamma' - \boldsymbol{k}$, and we have $\langle \boldsymbol{u} \rangle_\Gamma + \boldsymbol{k} \in \Gamma'$. Since $\langle \boldsymbol{u} \rangle_\Gamma + \boldsymbol{k}$ differs from $\boldsymbol{u}$ by an element in $\mathbb{Z}^2$, and itself lies in Γ', it must indeed be equal to $\langle \boldsymbol{u} \rangle_{\Gamma'}$, i.e., $\langle \boldsymbol{u} \rangle_{\Gamma'} - \langle \boldsymbol{u} \rangle_\Gamma = \boldsymbol{k}$. □

Lemma 3.3 actually leads to the following explicit relation:

$$\langle \boldsymbol{u} \rangle_{\Gamma'} = \langle \boldsymbol{u} \rangle_\Gamma + \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \chi_{\Pi_{\boldsymbol{k}}}(\boldsymbol{u})\, \boldsymbol{k}$$

where χ_A stands for the characteristic function of the set A. Note that $\langle \boldsymbol{u} \rangle_\Gamma$ always belongs to Γ. Hence, $\boldsymbol{u} \in \Pi_{\boldsymbol{k}}$ if and only if $\langle \boldsymbol{u} \rangle_\Gamma \in \Gamma_{\boldsymbol{k}} := \Gamma \cap (\Gamma' - \boldsymbol{k})$. We can then also write

$$\langle \boldsymbol{u} \rangle_{\Gamma'} = \langle \boldsymbol{u} \rangle_\Gamma + \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \chi_{\Gamma_{\boldsymbol{k}}}(\langle \boldsymbol{u} \rangle_\Gamma)\, \boldsymbol{k}. \tag{52}$$



(a) (b)

Fig. 6. Modulo operations. (a) Illustration of $\boldsymbol{u}[n] = \langle \boldsymbol{v}[n] \rangle_\Gamma$ from (49). (b) Comparison between $\langle \boldsymbol{u} \rangle_\Gamma$ and $\langle \boldsymbol{u} \rangle$ (illustration of (53)). In both figures, two—dimensional points are marked using the symbol "×."

Of particular interest will be the case where $\Gamma' = \Gamma_x$ and $\Gamma = [0,1)^2$. This yields

$$\langle \boldsymbol{u} \rangle_{\Gamma_x} = \langle \boldsymbol{u} \rangle + \sum_{\boldsymbol{k} \in \mathbb{Z}^2} \chi_{S_{\boldsymbol{k}}}(\langle \boldsymbol{u} \rangle)\, \boldsymbol{k} \tag{53}$$

where $S_{\boldsymbol{k}} := [0,1)^2 \cap (\Gamma_x - \boldsymbol{k})$. This is the decomposition of the function $\langle \cdot \rangle_{\Gamma_x}$ as mentioned earlier.

Another useful property is the following.

Proposition 3.4: Let Γ and Γ' be two (Lebesgue) measurable sets that tile the plane with the $\mathbb{Z}^2$ lattice. Then $\langle \cdot \rangle_\Gamma$ as a mapping from Γ' to Γ is a (Lebesgue) measure preserving bijection whose inverse is given by $\langle \cdot \rangle_{\Gamma'}$. If $F : \mathbb{R}^2 \to \mathbb{R}$ is any $\mathbb{Z}^2$-periodic locally integrable function, then

$$\int_\Gamma F(\boldsymbol{u}) d\boldsymbol{u} = \int_{\Gamma'} F(\boldsymbol{u}) d\boldsymbol{u}. \tag{54}$$

Proof: Bijectivity is clear. On the other hand, if $\boldsymbol{u} \in \Gamma'$, then $\langle \langle \boldsymbol{u} \rangle_\Gamma \rangle_{\Gamma'} = \boldsymbol{u}$, since each of these mappings shifts its argument by an element of $\mathbb{Z}^2$, and the resulting point lies in Γ'. Hence, $\langle \cdot \rangle_{\Gamma'}$ inverts $\langle \cdot \rangle_\Gamma$. Now, for any $\boldsymbol{k} \in \mathbb{Z}^2$, let us define $\Gamma_{\boldsymbol{k}} := \Gamma \cap (\Gamma' - \boldsymbol{k})$ and $\Gamma'_{\boldsymbol{k}} := \Gamma' \cap (\Gamma - \boldsymbol{k})$. It follows easily from the tiling assumption that the families $\{\Gamma_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}^2}$ and $\{\Gamma'_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}^2}$ form partitions of Γ and Γ', respectively. It is also easy to see that $\Gamma'_{\boldsymbol{k}} = \Gamma_{-\boldsymbol{k}} - \boldsymbol{k}$. Now, for a measurable set $A \subset \Gamma$, (52) implies that

$$\langle A \rangle_{\Gamma'} = \bigcup_{\boldsymbol{k} \in \mathbb{Z}^2} ((A \cap \Gamma_k) + \boldsymbol{k}).$$

This is a disjoint union, and it follows that

$$\begin{aligned} |\langle A \rangle_{\Gamma'}| &= \sum_{\boldsymbol{k} \in \mathbb{Z}^2} |A \cap \Gamma_k + \boldsymbol{k}| \\ &= \sum_{\boldsymbol{k} \in \mathbb{Z}^2} |A \cap \Gamma_{\boldsymbol{k}}| \\ &= |A|. \end{aligned}$$

Hence $\langle \cdot \rangle_\Gamma$ preserves measure.

Since F is $\mathbb{Z}^2$-periodic, we have $F(\boldsymbol{u}) = F(\langle \boldsymbol{u} \rangle_{\Gamma'})$. Using this and the measure preserving property of $\langle \cdot \rangle_{\Gamma'}$, we get

$$\int_\Gamma F(\boldsymbol{u}) d\boldsymbol{u} = \int_\Gamma F(\langle \boldsymbol{u} \rangle_{\Gamma'}) d\boldsymbol{u} = \int_{\Gamma'} F(\boldsymbol{v}) d\boldsymbol{v}. \tag{55}$$

□

We conclude this section with a word on the dynamics of $\boldsymbol{M}_x$ on the invariant set Γ_x. Consider the mapping $\langle \boldsymbol{M}_x \rangle : [0,1)^2 \to [0,1)^2$ naturally defined by $\langle \boldsymbol{M}_x \rangle(\boldsymbol{u}) = \langle \boldsymbol{M}_x(\boldsymbol{u}) \rangle$. It can be easily checked that the mappings $\boldsymbol{M}_x|_{\Gamma_x}$ and $\langle \boldsymbol{M}_x \rangle$ are related to each other via

$$\boldsymbol{M}_x|_{\Gamma_x} = \langle \cdot \rangle_{\Gamma_x} \circ \langle \boldsymbol{M}_x \rangle \circ \langle \cdot \rangle.$$

It is well known that when x is irrational, $\langle \boldsymbol{M}_x \rangle$ is ergodic with respect to the Lebesgue measure (see, e.g., [20]). Since both $\langle \cdot \rangle_{\Gamma_x} : [0,1)^2 \to \Gamma_x$ and $\langle \cdot \rangle : \Gamma_x \to [0,1)^2$ are measure preserving, it follows that $\boldsymbol{M}_x$ (and also $\boldsymbol{M}_x^{-1}$) is ergodic on Γ_x with respect to the Lebesgue measure as well. This is the ergodicity property that was mentioned at the end of Section II-D.

IV. Thorough Study of Three Particular Configurations

The purpose of this section is three-fold. First, we would like to give some concrete examples of invariant tiles Γ_x in some practical configurations. Recall from Section III that a general criterion regarding when the invariant sets reduce to single tiles is not available yet and also that our signal analysis machinery is dependent on this condition. Second, we will see in these examples that the tiling phenomenon is not restricted to irrational inputs but that it applies to rational inputs as well. Third, we would like to extract some common analytical features of these invariant sets which will later be crucial in the error analysis of Section V.

A. Linear T and Two-Bit Quantizer: The $\mathcal{L}_2$ System

The $\Sigma\Delta$ configuration for which derivations are the easiest is the standard two-bit double-loop configuration previously

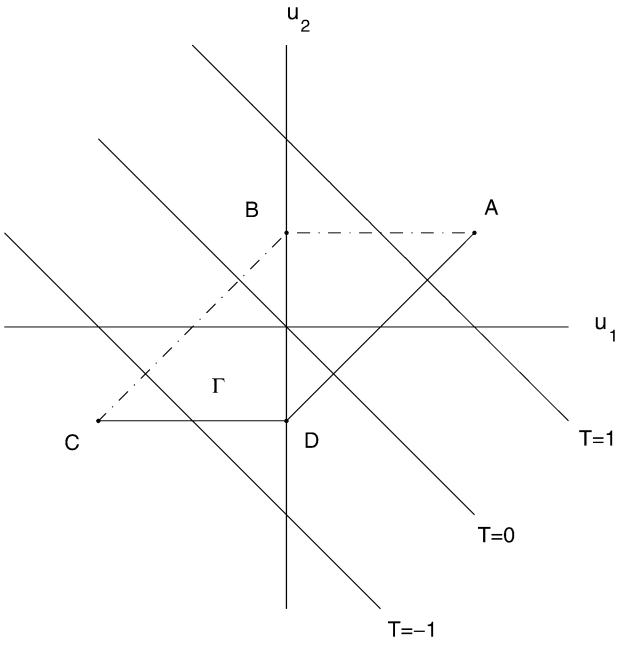


Fig. 7. The invariant set for the dynamical systems $\boldsymbol{M}_x$ considered in Section IV-A. The level sets of $T(\cdot)$ are drawn for $x = 0$.

studied in [12]. As explained in Section II-A, this corresponds to the case where $(\alpha, \beta) = (1, 1)$. We set $\gamma = 1$, so that $y[n]$ satisfies the relation

$$y[n] = u_1[n-1] + u_2[n-1] + x[n]. \tag{56}$$

With this choice of coefficients, the modulator satisfies a unique property that we describe here. One can easily derive from Fig. 3 that

$$u_2[n] = u_1[n-1] + u_2[n-1] + x[n] - q[n].$$

Because of (56), this implies that

$$u_2[n] = y[n] - q[n]$$

which is the quantizer error (up to the sign). First, assume that the quantizer Q is infinite as defined in Section II-B. This implies that $-\frac{1}{2} \le y[n] - q[n] < \frac{1}{2}$ for all n. Therefore, the point $(u_2[n-1], u_2[n])$ belongs to $[-\frac{1}{2}, \frac{1}{2})^2$ regardless of the input sequence $(x[n])_{n\in\mathbb{N}}$. Correspondingly, the couple $(u_1[n], u_2[n])$ belongs to the image Γ of the set $[-\frac{1}{2}, \frac{1}{2})^2$ under the bijection $\tau : (w_1, w_2) \mapsto (w_2 - w_1, w_2)$. We depict this set in Fig. 7. $\Gamma_x := \Gamma$ clearly remains invariant under $\boldsymbol{M}_x$ for all x, and its $\mathbb{Z}^2$ translations tile the plane; the latter follows easily from the observation that this is already true for the set $[-\frac{1}{2}, \frac{1}{2})^2$ and that the matrix representing τ is integer-valued with determinant ± 1.

Now, if $-\frac{1}{2} \le x[n] \le \frac{1}{2}$, then, as a consequence of (56), $y[n]$ always remains in the interval $(-2, 2)$. Hence the infinite quantizer can as well be replaced with the two-bit quantizer with output values $\{-1.5, -0.5, 0.5, 1.5\}$ to produce an equivalent system. For the corresponding threshold values $\{-1, 0, 1\}$ and for $x = 0$, the level sets of the function T are also drawn in Fig. 7. Note that there are four regions in Γ determined by these lines, and these are represented by the 2 bits of the quantizer output.

B. Linear T and 1-Bit Quantizer: The $\mathcal{L}_1$ System

To analyze properties that are likely to be representative of the general case of practical second-order ΣΔ modulators, it is important to consider at least one configuration where T is linear, but $(\alpha, \beta) \neq (1, 1)$ and the quantizer Q is only one-bit as defined in (21). Unfortunately, in this situation the invariant sets of $\boldsymbol{M}_x$ suddenly become complicated and very difficult to identify. Until now, this identification has been possible only in the particular case where $(\alpha, \beta, \gamma) = (\frac{1}{2}, 1, 0)$, the quantizer Q is one-bit, and x is limited to the interval $[-\frac{1}{6}, \frac{1}{6}]$. In this situation, the space is partitioned by the line $u_1 + \frac{1}{2}u_2 = 0$ into two half-spaces denoted by Ω_x^0 and Ω_x^1, and $\boldsymbol{M}_x$ is an affine transformation on each of these half-spaces. As x varies, these mappings exhibit invariant sets that depend on x in a nontrivial way. Consider the partition of the interval $(0, \frac{1}{6})$ as

$$\left(0, \frac{1}{6}\right) = \cdots \cup [\alpha_{k+1}, \alpha_k) \cup \cdots \cup [\alpha_2, \alpha_1)$$

where $\alpha_k = \frac{1}{2}(4k^2 - 1)^{-1}$, $k \ge 1$, and for each $x \in (0, \frac{1}{6})$, let $k = k_x$ be the unique integer such that $x \in [\alpha_{k+1}, \alpha_k)$. We show in the Appendix, Subsection B that the connected set Γ_x enclosed in the polygon shown in Fig. 8(a) (where the portion of the boundary represented in mixed line is excluded) is an invariant set for $\boldsymbol{M}_x$, and its $\mathbb{Z}^2$ translations tile the plane. The exact definition of the vertices of Γ_x is given in the Appendix, Subsection B, Table II. Note that the total number of vertices is equal to $4k+6$, which increases indefinitely as x approaches 0.

We add that Γ_0 and $\Gamma_{\frac{1}{6}}$ are obtained via the limits of $\mathcal{P}_x$, as $x \to 0$ and $x \to \frac{1}{6}$. Together with the symmetry $\Gamma_x = -\Gamma_{-x}$, which is a mere consequence of the relation

$$T(-\boldsymbol{u}, -x) = -T(\boldsymbol{u}, x) \tag{57}$$

we obtain the parameterization of all Γ_x in the range $[-\frac{1}{6}, \frac{1}{6}]$.

We also note that the polygonal boundary of Γ_x has bounded perimeter for all $x \in [-\frac{1}{6}, \frac{1}{6}]$.

C. A New Rule: Quadratic T and One-Bit Quantizer: The $\mathcal{Q}_1$ System

To extract further potential properties of the dynamical system of Fig. 3 in the one-bit case, it is interesting to explore the case where the linearity of T is relaxed, given the limited available results with linear T. Such a question was previously studied in [23] (see also [4] for a piecewise-linear choice of T).

It turns out that remarkably simple invariant sets are obtained (yet still using a one-bit quantizer) by considering the quadratic function T defined by

$$T(u_1, u_2, x) = C(x) + (6 - 4x)u_1 + (4 - 8x)u_2 + 4(x + u_1)^2 \tag{58}$$

for $x \ge 0$, where $C(x)$ is an arbitrary function of x. We present in the Appendix, Subsection C the reasoning behind this particular choice of T. For $x < 0$, T is defined by the symmetry relation (57). The invariant sets of the resulting dynamical systems have the property that the boundary of each of them is a piecewise-quadratic curve with four pieces. An example of

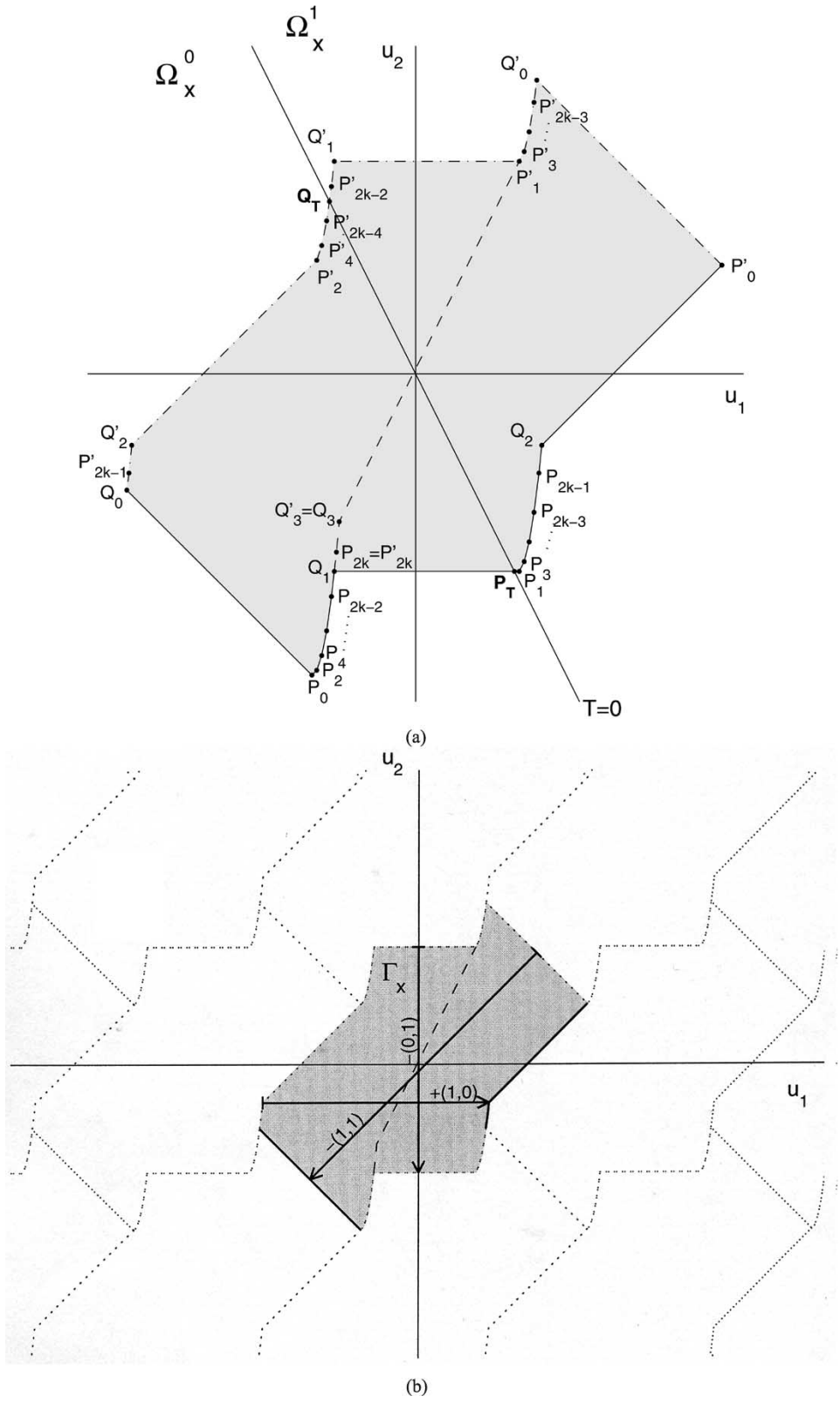


Fig. 8. The invariant set for the dynamical system given in Section IV-B for a generic k value (in this figure $k = 5$). (a) Detailed description. (b) Tiling demonstration.

these sets is depicted in Fig. 9 for $x = 0.24$, and for a particular choice of $C(x)$. The invariant set Γ_x is the region bounded by the quadratic curves that connect the points P_1, P_2, P_3, P_4, where the piece of curve that joins P_1 to P_2 is to be excluded. We show in the Appendix, Subsection C that Γ_x is an invariant set for $\boldsymbol{M}_x$, and its $\mathbb{Z}^2$ translations tile the plane.

Although the relevance of such a system to analog circuit implementation is still to be evaluated, this quadratic function T

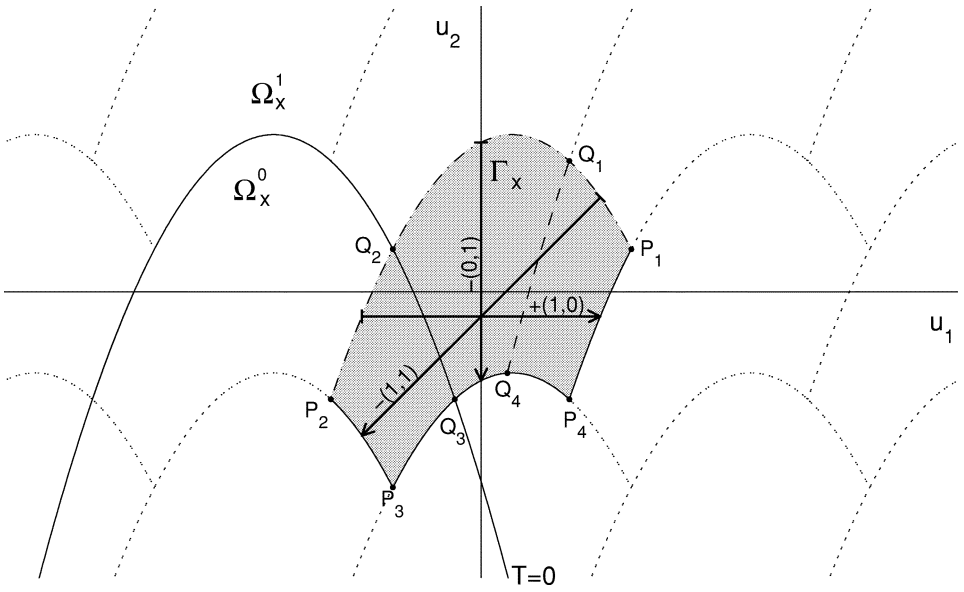


Fig. 9. The invariant set for the dynamical system given in Section IV-C. ($x = 0.24$).

is still interesting to be considered as it gives us a first situation where the invariant set of the dynamical system of Fig. 3 is entirely available analytically. It also gives us what we believe to be the simplest configuration of invariant sets theoretically achievable by one-bit second-order modulators.

One remark is on the robustness of the implementation. It was shown in the work of Yılmaz in [26] that second-order ΣΔ modulation is robust against small functional perturbations of (what we refer here as) $Q \circ T$ in the sense that stability of the state variable $\boldsymbol{u}$ is ensured as long as the boundary separating the partition $\{\Omega_x^i\}_{i=0,1}$ stays in a particular region. It also follows from this work that the quadratic function T presented in this paper is robust in this sense, at least in a range of inputs x. This increases the chances of the implementability of this quadratic function in real circuitry.

D. Boundedness, Regularity, and Tiling

The invariant sets of the dynamical systems given above possess three properties which turn out to be crucial in the estimates we shall prove in this paper. These properties are uniform boundedness in x, regularity of the boundary, and *tiling*, as summarized in Proposition 4.1. Before we proceed to the statement of the proposition, let us define the regularity class of sets $\mathcal{M}_b$ to be the collection of sets $H \subset [0,1)^d$ for which

$$|\{\boldsymbol{u} \in H^c : \operatorname{dist}(\boldsymbol{u}, H) < \epsilon\}| \leq b(\epsilon) \quad \text{and}$$
$$|\{\boldsymbol{u} \in H : \operatorname{dist}(\boldsymbol{u}, H^c) < \epsilon\}| \leq b(\epsilon)$$

for every $\epsilon > 0$, where $b : (0,\infty) \to (0,\infty)$ is a monotonically increasing function such that $\lim_{\epsilon \to 0^+} b(\epsilon) = 0$. Here $|A|$ denotes the (Lebesgue) measure of the set A, and A^c denotes the complement of A. Every Jordan measurable set (i.e., a set whose boundary has Lebesgue measure zero) belongs to such a class $\mathcal{M}_b$ for some b.

Proposition 4.1: For each of the one-parameter family of dynamical systems $\mathcal{D} = (\boldsymbol{M}_x, x \in [-\frac{1}{2}, \frac{1}{2}])$ given in Sections IV-A–C, there exists a subinterval $I = I(\mathcal{D})$ of $[-\frac{1}{2}, \frac{1}{2}]$ such that for each $x \in I$, the map $\boldsymbol{M}_x$ possesses an invariant set Γ_x with the following properties.

1) *Uniform boundedness in x*: There exists a positive constant M_0 such that
$$\sup_{x \in I} \sup_{\boldsymbol{u} \in \Gamma_x} |\boldsymbol{u}| \leq M_0.$$
2) *Regularity of the boundary*: There exists a positive constant C_0 such that $\Gamma_x \in \mathcal{M}_b$ for all $x \in I$, where $b(\epsilon) = C_0 \epsilon$.
3) *Tiling*: For each $x \in I$, the set Γ_x is a tile congruent to $[0,1)^2$ modulo translations by vectors in $\mathbb{Z}^2$. That is, the translates of Γ_x by the integer lattice tile the plane
$$\Gamma_x + \mathbb{Z}^2 = \mathbb{R}^2 \quad \text{and}$$
$$(\Gamma_x + \boldsymbol{k}) \cap \Gamma_x = \emptyset \text{ if } \boldsymbol{k} \in \mathbb{Z}^2 \backslash \{\boldsymbol{0}\}.$$
We say that Γ_x is a tiling-invariant set, or equivalently, an *invariant tile*.

The proof of this proposition is given in the Appendix, part D.

V. The Main Theorem

We shall continue to use the notation $\mathcal{L}_2$, $\mathcal{L}_1$, and $\mathcal{Q}_1$ to denote the one-parameter family of dynamical systems

$$\mathcal{D} = \left(\boldsymbol{M}_x, x \in \left[-\frac{1}{2}, \frac{1}{2}\right]\right)$$

for the two-bit linear, one-bit linear, and the one-bit quadratic schemes given in Sections IV-A–C, respectively. For each of these second-order ΣΔ schemes, we assume that for each $x \in I(\mathcal{D})$, the initial condition $\boldsymbol{u}[0]$ is chosen from the invariant set Γ_x of the associated dynamical system $\boldsymbol{M}_x$, and the sequence $\boldsymbol{u}[n]$ is defined for all $n \in \mathbb{Z}$ as described in Section III-C. Recall that the subscript x in $e_{\lambda,x}[n]$ denotes the dependence of the error $e_\lambda[n] = h_\lambda * (x_1[n] - q[n])$ on the value x of the

constant input signal $x_1[n]$, and h_λ is the $sinc^3$ filter defined in Section II-D. For each family $\mathcal{D} \in \{\mathcal{L}_2, \mathcal{L}_1, \mathcal{Q}_1\}$, we set, as in Section I

$$\mathrm{MSE}\,(\lambda;\mathcal{D}) = \sup_n \int_I \left|e_{\lambda,x}[n]\right|^2 \mathrm{d}x \tag{59}$$

where $I = I(\mathcal{D})$ is as defined in Proposition 4.1. The following theorem lists our improved estimates for these second order schemes.

Theorem 5.1: Let $\mathcal{D}$ be a second-order $\Sigma\Delta$ modulation scheme that satisfies the properties listed in Proposition 4.1, in particular any of the schemes $\mathcal{L}_2$, $\mathcal{L}_1$, or $\mathcal{Q}_1$. Then the following estimates hold.

a) The mean-square error defined by (59) satisfies

$$\mathrm{MSE}\,(\lambda;\mathcal{D}) \le C\lambda^{-9/2}\log^2\lambda \tag{60}$$

for all $\lambda > 0$, where $C = C(\mathcal{D})$ is a constant that depends only on the scheme $\mathcal{D}$.

b) For almost every $x \in I$, and all $n \in \mathbb{Z}$

$$\left|e_{\lambda,x}[n]\right|^2 \le C\lambda^{-9/2}\log^4\lambda \tag{61}$$

for all $\lambda > 0$, where $C = C(\mathcal{D}, x, n)$ does not depend on λ, but otherwise may depend on the scheme $\mathcal{D}$, input x, and the time point n.

c) For $\mathcal{D} = \mathcal{L}_2$, the same estimates in a) and b) hold with $9/2$ replaced by 5.

Before we proceed onto the proof of the theorem, let us list some further qualitative observations. Equation (36) states that

$$e_{\lambda,x}[n] = \frac{1}{\lambda^2}\left(\bar{u}_\lambda[n] - 2\bar{u}_\lambda[n-\lambda] + \bar{u}_\lambda[n-2\lambda]\right). \tag{62}$$

As a reminder from (37), $\bar{u}_\lambda[n]$ is qualitatively the average of the discrete sequence $u_2[m]$ over the time interval of $(n-\lambda, n]$. This expression immediately suggests that $e_{\lambda,x}[n] = o(\lambda^{-2})$. To see this, denote by F_2 the mapping that takes $\boldsymbol{u} = (u_1, u_2)$ to u_2. When x is irrational, for almost all $\boldsymbol{u}[0]$ and for all n, the ergodic theorem yields

$$\begin{aligned}\lim_{\lambda\to\infty}\bar{u}_\lambda[n] &= \lim_{\lambda\to\infty}\frac{1}{\lambda}\sum_{j=0}^{\lambda-1} F_2(\boldsymbol{M}_x^{-j}\boldsymbol{u}[n]) \\ &= \int_{\Gamma_x} F_2(\boldsymbol{w})d\boldsymbol{w} \\ &= \int_{\Gamma_x} w_2 dw_1 dw_2. \end{aligned}\tag{63}$$

Now, it is a simple exercise to show that

$$\begin{aligned}&\bar{u}_\lambda[n] - 2\bar{u}_\lambda[n-\lambda] + \bar{u}_\lambda[n-2\lambda] \\ &\qquad = 3\Big(\bar{u}_\lambda[n] - 2\bar{u}_{2\lambda}[n] + \bar{u}_{3\lambda}[n]\Big);\end{aligned}\tag{64}$$

therefore, (62) and (63) together imply that $e_{\lambda,x}[n] = o(\lambda^{-2})$. Note that this argument does not provide us with any information about the improvement on the exponent of λ. The proof of Theorem 5.1 will heavily use techniques from the theory of uniform distribution. Subsection A of the Appendix contains the definitions and the tools that we shall employ in the proof.

Proof of Theorem 5.1: Let us define a residual sequence r_λ by

$$\begin{aligned}r_\lambda[n] &:= \bar{u}_\lambda[n] - \int_{\Gamma_x} w_2 dw_1 dw_2 \\ &= \frac{1}{\lambda}\sum_{n-\lambda<m\le n} F_2(\boldsymbol{u}[m]) - \int_{\Gamma_x} F_2(\boldsymbol{w})d\boldsymbol{w}.\end{aligned}\tag{65}$$

Since $\bar{u}_\lambda[n]$ and $r_\lambda[n]$ differ by an absolute constant, we can replace $\bar{u}_\lambda[n]$ in (64) by $r_\lambda[n]$. When combined with (62), this yields

$$\begin{aligned}\left|e_{\lambda,x}[n]\right| &= \frac{3}{\lambda^2}\left|r_\lambda[n] - 2r_{2\lambda}[n] + r_{3\lambda}[n]\right| \\ &\le \frac{3}{\lambda^2}\Big(\left|r_\lambda[n]\right| + \left|2r_{2\lambda}[n]\right| + \left|r_{3\lambda}[n]\right|\Big)\end{aligned}\tag{66}$$

and by Cauchy–Schwarz

$$\left|e_{\lambda,x}[n]\right|^2 \le \frac{C}{\lambda^4}\left(\left|r_\lambda[n]\right|^2 + \left|r_{2\lambda}[n]\right|^2 + \left|r_{3\lambda}[n]\right|^2\right); \tag{67}$$

therefore, it suffices, for each time point n, to estimate $\left|r_\lambda[n]\right|$ for general λ.

Let us first consider the rather simple case $\mathcal{D} = \mathcal{L}_2$. Note that the invariant set Γ_x given in Fig. 7 is such that the ordinate of any point in Γ_x always lies in $[-\frac{1}{2}, \frac{1}{2})$. Also, the sequence $\boldsymbol{v}[n]$ defined in Section III-C satisfies $u_2[n] - v_2[n] \in \mathbb{Z}$. Therefore,

$$u_2[n] = \langle v_2[n]\rangle_{[-\frac{1}{2},\frac{1}{2})} = \left\langle v_2[n] + \frac{1}{2}\right\rangle - \frac{1}{2}. \tag{68}$$

Since in this case $\int_{\Gamma_x} w_2 dw_1 dw_2 = 0$, (65) becomes

$$\begin{aligned}r_\lambda[n] &= \frac{1}{\lambda}\sum_{n-\lambda<m\le n}\left(\left\langle v_2[m] + \frac{1}{2}\right\rangle - \frac{1}{2}\right) \\ &= \frac{1}{\lambda}\sum_{n-\lambda<m\le n}\left\langle v_2[m] + \frac{1}{2}\right\rangle - \int_0^1 w dw.\end{aligned}\tag{69}$$

Let $D_{(n-\lambda,n]}(\langle v_2\rangle)$ denote the *discrepancy* (see Subsection A of the Appendix) of the λ consecutive sequence elements $\{\langle v_2[m]\rangle;\ n-\lambda < m \le n\}$. *Koksma's inequality* (Appendix, Subsection A) can be used to bound $\left|r_\lambda[n]\right|$

$$\left|r_\lambda[n]\right| \le D_{(n-\lambda,n]}(\langle v_2\rangle) \tag{70}$$

where we have used the invariance of discrepancy under translations of the torus $\mathbb{T} = [0, 1)$ in the equality

$$D_{(n-\lambda,n]}\left(\left\langle v_2 + \frac{1}{2}\right\rangle\right) = D_{(n-\lambda,n]}(\langle v_2\rangle).$$

The estimate (70) therefore reduces the problem for the case $\mathcal{D} = \mathcal{L}_2$ to estimating the λ-term discrepancy of the sequence $\langle v_2\rangle$.

The general case for $\mathcal{D}$ which includes $\mathcal{D} \in \{\mathcal{L}_1, \mathcal{Q}_1\}$ is more difficult because it is no longer possible to obtain an expression

of $u_2[n]$ as simple as in (68). Initially, the expression (65) suggests the need for some two-dimensional version of Koksma's inequality, defined on an arbitrary set (in our case Γ_x); however, the setup for the so-called *Koksma–Hlawka inequality* [5, Theorem 1.14]) is the unit cube $[0,1)^d$. Using (49), the $\mathbb{Z}^2$-periodicity of $\langle\cdot\rangle_{\Gamma_x}$, and Proposition 3.4, we can transform the expression (65) into

$$r_\lambda[n] = \frac{1}{\lambda} \sum_{n-\lambda<m\le n} F_2\big(\,\langle\langle \boldsymbol{v}[m]\rangle\rangle_{\Gamma_x}\,\big) \int\limits_{[0,1)^2} F_2\left(\langle \boldsymbol{w}\rangle_{\Gamma_x}\right) d\boldsymbol{w} \tag{71}$$

and attempt to use the Koksma–Hlawka inequality for the sequence $\langle \boldsymbol{v}[m]\rangle$ and the function $f = F_2 \circ \langle\cdot\rangle_{\Gamma_x}$. At first, this attempt also appears to be defeated because Koksma–Hlawka inequality holds for functions that are of bounded variation in the sense of Hardy and Krause (see [5, p. 10] for the definition), which is a more restrictive class than the usual functional class $BV([0,1)^d)$ when $d \ge 2$, and which does not necessarily contain $F_2 \circ \langle\cdot\rangle_{\Gamma_x}$ due to the geometry of Γ_x.

We overcome this difficulty with the following procedure. By applying (53) on both $\big\langle\langle \boldsymbol{v}[m]\rangle\big\rangle_{\Gamma_x}$ and $\langle \boldsymbol{w}\rangle_{\Gamma_x}$, and by using the linearity of F_2, we first obtain

$$r_\lambda[n] = A_\lambda[n] + \sum_{\boldsymbol{k}\in\mathbb{Z}^2} B_\lambda^{\boldsymbol{k}}[n] F_2(\boldsymbol{k}) \tag{72}$$

where

$$A_\lambda[n] := \frac{1}{\lambda} \sum_{n-\lambda<m\le n} F_2(\langle \boldsymbol{v}[m]\rangle) - \int\limits_{[0,1)^2} F_2(\boldsymbol{w}) d\boldsymbol{w}$$

and

$$B_\lambda^{\boldsymbol{k}}[n] := \frac{1}{\lambda} \sum_{n-\lambda<m\le n} \chi_{S_{\boldsymbol{k}}}(\langle \boldsymbol{v}[m]\rangle) - \int\limits_{[0,1)^2} \chi_{S_{\boldsymbol{k}}}(\boldsymbol{w}) d\boldsymbol{w}.$$

(Note that we have replaced $\langle \boldsymbol{w}\rangle$ with $\boldsymbol{w}$ since $\boldsymbol{w} \in [0,1)^2$.) It is now possible to apply the Koksma–Hlawka inequality to the first term $A_\lambda[n]$. This gives

$$\big|A_\lambda[n]\big| \le C_0 D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle) \tag{73}$$

where $C_0 = \mathrm{Var}_{\mathrm{HK}}(F_2)$ is the variation of F_2 in the sense of Hardy and Krause, and $D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)$ is the two-dimensional discrepancy (Appendix, Subsection A) of

$$\{\langle \boldsymbol{v}[m]\rangle : n-\lambda < m \le n\}.$$

On the other hand, it is still true that the functions $\chi_{S_{\boldsymbol{k}}}$ are not necessarily of bounded variation in the sense of Hardy and Krause. However, now the notion of discrepancy with respect to a given subset can be invoked (Appendix, Subsection A). Indeed, by definition, we have

$$\left|B_\lambda^{\boldsymbol{k}}[n]\right| = D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle, S_{\boldsymbol{k}}). \tag{74}$$

We now make use of the regularity of the sets Γ_x in order to estimate these quantities. From Proposition 4.1 (Property 2) and Theorem A.4, it follows that

$$\sup_{x\in I} \sup_{\boldsymbol{k}\in\mathbb{Z}^2} D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle, S_{\boldsymbol{k}}) \le C_1 D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)^{1/2} \tag{75}$$

for some constant C_1, where $I = I(\mathcal{D})$, as defined in Proposition 4.1. Note also that $S_{\boldsymbol{k}} = \emptyset$ as soon as $|\boldsymbol{k}| > M_0$, where M_0 is some absolute constant that only depends on the system $\mathcal{D}$, and the range I. We can therefore limit the summation over $\boldsymbol{k}$ in (72) to the set $\mathcal{K} = \{\boldsymbol{k} \in \mathbb{Z}^2 : |\boldsymbol{k}| \le M_0\}$, whose cardinality $\#\mathcal{K}$ does not exceed M_0^2. We have also $|F_2(\boldsymbol{k})| \le |\boldsymbol{k}| \le M_0$ for all $\boldsymbol{k} \in \mathcal{K}$. With (73) and (75), we finally obtain the analogous bound for $|r_\lambda[n]|$

$$\begin{aligned}\big|r_\lambda[n]\big| &\le C_0 D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle) + M_0 C_1 \sum_{\boldsymbol{k}\in\mathcal{K}} D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)^{1/2} \\ &\le (C_0 + C_1 M_0^3) D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)^{1/2}\end{aligned} \tag{76}$$

where we have used the fact that discrepancy is always between 0 and 1 when merging the terms with different powers. Therefore, the problem in the general case is also reduced to estimating the λ-term discrepancy, but this time of the two-dimensional point sequence $\langle \boldsymbol{v}\rangle$. The following lemma addresses this issue.

Lemma 5.2: The following estimates hold.

a) For all $\lambda > 0$ and $n \in \mathbb{Z}$

$$\int_{-1/2}^{1/2} \left[D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)\right]^2 \mathrm{d}x \le C\lambda^{-1} \log^4 \lambda \tag{77}$$

where C is an absolute constant.

b) For almost every $x \in I$, all $\lambda > 0$, and $n \in \mathbb{Z}$

$$D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle) \le C\lambda^{-1/2} \log^{7/2+\delta} \lambda \tag{78}$$

where $C = C(x,n)$ does not depend on λ, but otherwise may depend on the input x and the time point n and δ is some fixed small positive number.

c) If $\boldsymbol{v}$ is replaced by v_2, then $\log^4 \lambda$ can be replaced by $\log^2 \lambda$ in a) and $\log^{7/2+\delta} \lambda$ can be replaced by $\log^{5/2+\delta} \lambda$ in b).

The proof of this lemma is independent of the rest of the proof of Theorem 5.1 and is presented separately at the end of this section.

Lemma 5.2 is essentially all that was needed to complete the proof of Theorem 5.1.

a) We square and integrate both sides of the inequality (76) and apply Cauchy–Schwarz followed by Lemma 5.2, part a) to obtain

$$\begin{aligned}\int_I \left|r_\lambda[n]\right|^2 \mathrm{d}x &\le C \int_I D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle) \mathrm{d}x \\ &\le C|I|^{1/2} \left(\int_I \left[D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)\right]^2 \mathrm{d}x\right)^{1/2} \\ &\le C\lambda^{-1/2} \log^2 \lambda.\end{aligned} \tag{79}$$

Note that C does not depend on n. Therefore, this result together with (67) implies (60).

b) In this case, we simply apply (67), (76), and Lemma 5.2, part b) to obtain (61).

c) For the mean-square error, we square and integrate both sides of (70) and apply (67) and Lemma 5.2, part c). For

the instantaneous error, we simply apply (67), (70) and Lemma 5.2, part c) to obtain the desired estimate. □

Proof of Lemma 5.2:

a) Define, for $\boldsymbol{k} = (k_1, k_2) \in \mathbb{Z}^2$

$$\mathcal{S}_{(a,b]}(\boldsymbol{k}, x) := \frac{1}{b-a} \sum_{a<m\leq b} e^{2\pi i \boldsymbol{k}\cdot\langle \boldsymbol{v}[m]\rangle} \tag{80}$$

where the dependence on x becomes explicit when (51) is inserted in this expression. Using the periodicity of the exponential function, one can rewrite $\mathcal{S}_{(a,b]}$ as

$$\mathcal{S}_{(a,b]}(\boldsymbol{k}, x) = \frac{1}{b-a} \sum_{a<m\leq b} c_m e^{2\pi i d_m x} \tag{81}$$

where

$$c_m = e^{2\pi i[u_1[0]k_1 + (u_2[0]+mu_1[0])k_2]}$$

and

$$d_m = mk_1 + \frac{1}{2}m(m+1)k_2.$$

Note that $|c_m| = 1$ and $d_m \in \mathbb{Z}$ for all m. Since d_m is a quadratic polynomial in m, it can attain any given value at most twice. Hence, if $\mathcal{S}_{(a,b]}(\boldsymbol{k}, x)$ is rewritten as a trigonometric polynomial in x with distinct frequencies, the amplitude of each frequency will be bounded by $2/(b-a)$, since $\max_{m,l} |c_m + c_l| \leq 2$. Also, there will be at most $b-a$ distinct frequencies. Thus, using Parseval's theorem, one easily bounds $\|\mathcal{S}_{(a,b]}(\boldsymbol{k}, \cdot)\|_{L^2(\mathbb{T})}$ by

$$\|\mathcal{S}_{(a,b]}(\boldsymbol{k}, \cdot)\|_{L^2(\mathbb{T})} \leq \frac{2}{\sqrt{b-a}} \tag{82}$$

uniformly in $\boldsymbol{k}$. Now, for any positive integer K, *Erdős–Turán–Koksma inequality* (Theorem A.5) yields the estimate

$$D_{(a,b]}(\langle \boldsymbol{v}\rangle) \leq C\left(\frac{1}{K} + \sum_{0<\|\boldsymbol{k}\|_\infty \leq K} \frac{\left|\mathcal{S}_{(a,b]}(\boldsymbol{k}, x)\right|}{r(\boldsymbol{k})}\right) \tag{83}$$

which, upon taking the square, using the (Cauchy–Schwarz) inequality $(y+z)^2 \leq 2(y^2+z^2)$, and integrating gives

$$\int_{-1/2}^{1/2} D^2_{(a,b]}(\langle \boldsymbol{v}\rangle) dx \leq 2C\left(\frac{1}{K^2} + \sum_{0<\|\boldsymbol{k}\|_\infty, \|\boldsymbol{l}\|_\infty \leq K} \frac{\mathcal{T}_{(a,b]}(\boldsymbol{k}, \boldsymbol{l})}{r(\boldsymbol{k})r(\boldsymbol{l})}\right) \tag{84}$$

where

$$\mathcal{T}_{(a,b]}(\boldsymbol{k}, \boldsymbol{l}) := \int_{-1/2}^{1/2} |\mathcal{S}_{(a,b]}(\boldsymbol{k}, x)||\mathcal{S}_{(a,b]}(\boldsymbol{l}, x)|\, dx. \tag{85}$$

This last quantity can be bounded by $4/(b-a)$ using Cauchy–Schwarz inequality and (82). On the other hand, one has

$$\sum_{0<\|\boldsymbol{k}\|_\infty \leq K} \frac{1}{r(\boldsymbol{k})} = 4\left(\sum_{k_1=1}^{K}\sum_{k_2=1}^{K} \frac{1}{k_1 k_2} + \sum_{k_1=1}^{K} \frac{1}{k_1}\right) \leq C' \log^2 K \tag{86}$$

so that (84) reduces to, for $a = n-\lambda$ and $b = n$

$$\int_{-1/2}^{1/2} \left[D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)\right]^2 \mathrm{d}x \leq C'' \inf_{K\geq 1}\left(\frac{1}{K^2} + \frac{1}{\lambda}\log^4 K\right) \leq C\lambda^{-1}\log^4 \lambda \tag{87}$$

where it suffices to choose $K \sim \lambda^{1/2}$ at the last step.

b) The proof of this result may seem somewhat unexpected since it is actually derived from the input-averaged estimate. However, the technique we shall use in our proof is well known in the metric theory of discrepancy [5, Sec. 1.6.1].

Let D_Λ denote the discrepancy of a given sequence $\boldsymbol{w}_m$ over the set of indexes $m \in \Lambda$ and $\#\Lambda$ denote the number of integers in Λ. (For an arbitrary finite set J, $\#J$ will simply denote the cardinality of J.) A crucial aspect of the method is that the function $\Lambda \mapsto (\#\Lambda)D_\Lambda$ is subadditive, i.e., for $\Lambda_1 \cap \Lambda_2 = \emptyset$, we have

$$(\#\Lambda_1 \cup \Lambda_2)D_{\Lambda_1\cup\Lambda_2} \leq (\#\Lambda_1)D_{\Lambda_1} + (\#\Lambda_2)D_{\Lambda_2} \tag{88}$$

which follows straightforwardly from the definition of discrepancy given by (A9).

Denote by Λ^j_m the collection of all dyadic subintervals $\Lambda \subset [0, 2^m)$ of length 2^j. For example, $\Lambda^2_3 = \{[0,4), [4,8)\}$. Note that $\#\Lambda^j_m = 2^{m-j}$.

It is clear by considering the binary expansion of any given $\lambda \in [0, 2^m)$ that one can write $[0, \lambda)$ as a disjoint union of at most m dyadic intervals. Let us call the collection of these intervals J_λ. Hence, we have $J_\lambda \subset \bigcup_{j=0}^{m-1} \Lambda^j_m$, $\#J_\lambda \leq m$, and $[0,\lambda) = \bigcup_{\Lambda\in J_\lambda} \Lambda$.

Fix n. Since $(n-\lambda, n] = n - [0, \lambda)$, we have

$$\lambda D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle) \leq \sum_{\Lambda\in J_\lambda} (\#\Lambda) D_{n-\Lambda}(\langle \boldsymbol{v}\rangle) \tag{89}$$

so that by Cauchy–Schwarz, we get

$$\lambda^2 \left[D_{(n-\lambda,n]}(\langle \boldsymbol{v}\rangle)\right]^2 \leq (\#J_\lambda) \sum_{\Lambda\in J_\lambda} (\#\Lambda)^2 \left[D_{n-\Lambda}(\langle \boldsymbol{v}\rangle)\right]^2 \leq m\Psi_m(x) \tag{90}$$

where we define $\Psi_m(x)$ to be the function

$$\Psi_m(x) := \sum_{j=0}^{m-1} \sum_{\Lambda\in\Lambda^j_m} (\#\Lambda)^2 \left[D_{n-\Lambda}(\langle \boldsymbol{v}\rangle)\right]^2, \quad m \geq 1. \tag{91}$$

Now, note that Lemma 5.2, part a) implies

$$\int_{-1/2}^{1/2} \Psi_m(x)\mathrm{d}x = \sum_{j=0}^{m-1} \sum_{\Lambda\in\Lambda^j_m} (\#\Lambda)^2 \int_{-1/2}^{1/2} \left[D_{n-\Lambda}(\langle \boldsymbol{v}\rangle)\right]^2 \mathrm{d}x \leq C_1 \sum_{j=0}^{m-1} \sum_{\Lambda\in\Lambda^j_m} (\#\Lambda)\log^4(\#\Lambda) \leq C_2 2^m m^5. \tag{92}$$

Therefore, for an arbitrary positive number $\delta > 0$, we obtain

$$\sum_{m=1}^{\infty} \int_{-1/2}^{1/2} \frac{\Psi_m(x)}{2^m m^{6+\delta}} \mathrm{d}x < \infty. \tag{93}$$

Now, as we show next, a standard Borel–Cantelli argument yields the bound

$$\Psi_m(x) \leq C(x) 2^m m^{6+\delta} \tag{94}$$

for all m and almost every x. To see this, let

$$E_m := \left\{ x \in \left[-\frac{1}{2}, \frac{1}{2}\right] : \Psi_m(x) \geq 2^m m^{6+\delta} \right\}$$

with Lebesgue measure $|E_m|$. Since we have

$$|E_m| = \int_{E_m} 1 \, \mathrm{d}x \leq \int_{-1/2}^{1/2} \frac{\Psi_m(x)}{2^m m^{6+\delta}} \mathrm{d}x \tag{95}$$

it follows that $\sum_m |E_m| < \infty$. Hence, the set

$$\bigcap_{l=1}^{\infty} \bigcup_{m \geq \boldsymbol{l}} E_m$$

(i.e., the collection of points $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ for which $\Psi_m(x) \geq 2^m m^{6+\delta}$ for infinitely many m) has measure zero. This means that for almost every $x \in \left[-\frac{1}{2}, \frac{1}{2}\right]$, one has $\Psi_m(x) \leq 2^m m^{6+\delta}$ for all but finitely many m. For each x, we remove this finite set of unwanted values of m by multiplying the upper bound by a suitable constant $C(x)$. This proves (94). (One can extend this argument to show that $\Psi_m(x) = o(2^m m^{6+\delta})$ almost everywhere; see [5, p. 154].)

Now, for each $\lambda > 0$, there exists a unique m such that $2^{m-1} \leq \lambda < 2^m$. Then (90) implies together with $\Psi_m(x) \leq C(x) 2^m m^{6+\delta}$ almost everywhere that

$$\left[D_{(n-\lambda,n]}(\langle \boldsymbol{v} \rangle)\right]^2 \leq C(x,n) \lambda^{-1} \log^{7+\delta} \lambda, \quad \text{a.e. } x \tag{96}$$

where we have also restored the possible dependence of the constant $C(x)$ on n which was fixed at the beginning of the proof.

c) These inequalities are proved exactly in the same manner as in a) and b), however, using the one-dimensional *Erdős–Turán inequality* (Theorem A.3) instead. □

VI. Discussion and Further Remarks

What has fundamentally enabled our analysis of the ΣΔ modulators in this paper is the tiling property of the invariant sets of the associated dynamical systems. The tiling property allowed us to find an explicit expression of the error signal for constant inputs. In this paper, we have concentrated on upper bounds for the instantaneous error of the modulator in two cases: in the mean and the almost surely, when the constant input comes from a uniform distribution. In both cases, we have derived bounds in the form of $\lambda^{-4.5}$ (modulo logarithmic factors) under the general regularity conditions of Proposition 4.1. Apart from the $\mathcal{L}_2$ case, what kept us from achieving the experimentally observed generic decay rate λ^{-5} was the lack of a more customized discrepancy estimate than what is implied by Theorem A.4. It would be interesting to improve this machinery and further close this gap.

The constants appearing in the error bounds that we have derived in this paper are unfortunately only implicit. While it is very desirable for practical implementations to know explicit (and perhaps tight) constants, at this stage we do not know if the functional forms of these error bounds reflect the accurate order of magnitude of the norms we have considered. Therefore, we have not focused on the values of constants in this paper.

It turns out [24] that some of these problems are eliminated if the time-averaged square error measure is used instead; it then becomes possible via the tools of ergodic theory to extract a more refined form of the error decay rate in λ. This constitutes a generalization of the work in [7], [12] to a much more general setup of ΣΔ quantization schemes.

We note that the analysis of this paper can be straightforwardly generalized to higher order ΣΔ modulators with constant input once the tiling property (with single invariant tiles) is established and these tiles satisfy the properties listed in Proposition 4.1 (in fact, it is possible to relax the regularity conditions stated in there via the weaker general conditions of Theorem A.4). We leave the details of this generalized analysis to the reader.

In parallel, a substantial topic of investigation is a better understanding of the tiling phenomenon, and in particular, how the constant input theory can be generalized to time-varying inputs. This is not easy, however, since there is yet no scheme apart from $\mathcal{L}_2$ in which the invariant sets Γ_x do not vary with x. Understanding this dependence will prove to be crucial in improving the error estimates for second- and higher order ΣΔ modulators for time-varying inputs.

Appendix

A. Tools From the Theory of Uniform Distribution

Let $\{w_n\}_{n=1}^{\infty}$ be a sequence of points in $[0,1)$ identified with the 1-torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. The sequence $\{w_n\}$ is said to be *uniformly distributed* (in short, *u.d.*) if

$$\lim_{N \to \infty} \frac{\#\{1 \leq n \leq N : w_n \in I\}}{N} = |I| \tag{A1}$$

for every arc I in $\mathbb{T}$. Define the N-term *discrepancy* of the sequence $\{w_n\}$ as

$$D_N := D_N(w) := \sup_{I \in \mathcal{I}} \left| \frac{\#\{1 \leq n \leq N : w_n \in I\}}{N} - |I| \right| \tag{A2}$$

where $\mathcal{I}$ denotes the set of all intervals in $[0,1)$ considered as the 1-torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. It is an elementary result that $\{w_n\}$ is u.d. if and only if $D_N(w) \to 0$ as $N \to \infty$. Equivalent characterizations of uniform distribution are given by *Weyl's criterion*.

Theorem A.1 (Weyl):

$$\{w_n\} \text{ is u.d.} \iff \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k w_n} \to 0, \ \forall \, k \in \mathbb{Z} \setminus \{0\} \tag{A3}$$

$$\iff \frac{1}{N} \sum_{n=1}^{N} f(w_n) \to \int_{\mathbb{T}} f(w) dw \tag{A4}$$

for every Riemann-integrable (or, equivalently, continuous) function f on $\mathbb{T}$.

These are "qualitative" statements. The relation between how good the distribution of a sequence is and how fast (A3) and (A4) converge are studied in the "quantitative" theory. The second Weyl criterion is especially relevant to numerical integration. Fundamental quantitative measures in the theory are the following.

Theorem A.2 (Koksma's Inequality [16]): Given any function $f : [0,1] \to \mathbb{R}$ that is of bounded variation and a finite sequence of points $w_1, \ldots, w_N$ in $[0,1)$

$$\left| \frac{1}{N} \sum_{n=1}^{N} f(w_n) - \int_0^1 f(w) dw \right| \leq \mathrm{Var}(f) D_N(w) \tag{A5}$$

where $\mathrm{Var}(f)$ denotes the total variation of f.

Theorem A.3 (Erdős–Turán Inequality [16]):

$$D_N(w) \leq C \inf_{K \geq 1} \left(\frac{1}{K} + \sum_{k=1}^{K} \frac{1}{k} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i k w_n} \right| \right) \tag{A6}$$

for some absolute constant C.

The theory of uniform distribution generalizes naturally to higher dimensions, however, with some added complexity. Let $\{\boldsymbol{w}_n\}$ be a sequence in $[0,1)^d$ identified with $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$. For a measurable subset H of $[0,1)^d$, define

$$D_N(\boldsymbol{w}; H) := \left| \frac{\#\{1 \leq n \leq N : \boldsymbol{w}_n \in H\}}{N} - |H| \right| \tag{A7}$$

where $|H|$ denotes the d-dimensional Lebesgue measure of H. Let $\mathcal{I}^d$ denote the set of all intervals (i.e., the set of all rectangles whose sides are parallel to the axes) in $\mathbb{T}^d$. The discrepancy D_N is by definition

$$D_N(\boldsymbol{w}) = \sup_{H \in \mathcal{I}^d} D_N(\boldsymbol{w}; H). \tag{A8}$$

The sequence $\{\boldsymbol{w}_n\}$ is said to be u.d. if the condition $\lim_{N\to\infty} D_N(\boldsymbol{w}; H) = 0$ holds for every $H \in \mathcal{I}^d$. Again, this is equivalent to $\lim_{N\to\infty} D_N = 0$. Weyl's criterion naturally extends using multidimensional versions of (A3) and (A4).

If a finite index set Λ replaces the set of indexes $1, \ldots, N$, then we shall use the notation $D_\Lambda(\boldsymbol{w})$ to denote the discrepancy of the points $\boldsymbol{w}_n$, $n \in \Lambda$, i.e.,

$$D_\Lambda := D_\Lambda(\boldsymbol{w}) := \sup_{I \in \mathcal{I}} \left| \frac{\#\{n \in \Lambda : \boldsymbol{w}_n \in I\}}{\#\Lambda} - |I| \right|. \tag{A9}$$

A definition of discrepancy exists also for arbitrary nonnegative Borel measures μ on $[0,1)^d$. The discrepancy of μ with respect to the set $H \in [0,1)^d$, denoted by $D(\mu; H)$, is defined to be $|\mu(H) - |H||$. Similarly, one has the definition

$$D(\mu) := \sup_{H \in \mathcal{I}^d} D(\mu; H) \tag{A10}$$

for the discrepancy of μ. By definition, $D_N(\boldsymbol{w}; H) = D(\mu_N; H)$, where the measure μ_N is defined by

$$\mu_N(A) := \frac{1}{N} \sum_{n=1}^{N} \chi_A(\boldsymbol{w}_n)$$

for $A \subset \mathbb{T}^d$.

If the supremum in (A10) is taken instead over all *convex* subsets of $\mathbb{T}^d$, then this quantity defines the *isotropic* discrepancy $J(\mu)$. Clearly, one has $D(\mu) \leq J(\mu)$; on the other hand, an inequality in the reverse direction exists only in a weaker sense: $J(\mu) \leq C_d D(\mu)^{1/d}$, where C_d is a constant that depends only on the dimension d. The following theorem [18, p. 173] (see also [17]), gives a discrepancy estimate for sets in the larger family of Jordan-measurable sets. Let $\mathcal{M}_b$ denote the class of sets defined in Section IV-D.

Theorem A.4 (Niederreiter, Wills): Let $b : (0,\infty) \to (0,\infty)$ be monotonically increasing such that $b(\epsilon) \geq \epsilon$ for all $\epsilon > 0$, and $\lim_{\epsilon \to 0^+} b(\epsilon) = 0$. Then, for every $H \in \mathcal{M}_b$, one has

$$D(\mu; H) \leq 4\, b\left(2\sqrt{d} D(\mu)^{1/d}\right). \tag{A11}$$

A multidimensional version of Koksma's inequality (called the Koksma–Hlawka inequality) holds for functions of bounded variation in the sense of Hardy and Krause. We will not go into the details but refer to [5], [16] only. On the other hand, a generalization of Erdős–Turán inequality is simpler to state and is given by the following.

Theorem A.5 (Erdős–Turán–Koksma Inequality [16]):

$$D_N(\boldsymbol{w}) \leq C_d \inf_{K \geq 1} \left(\frac{1}{K} + \sum_{0 < \|\boldsymbol{k}\|_\infty \leq K} \frac{1}{r(\boldsymbol{k})} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i \boldsymbol{k} \cdot \boldsymbol{w}_n} \right| \right) \tag{A12}$$

where

$$r(\boldsymbol{k}) := \prod_{i=1}^{d} \max\{1, |k_i|\}, \qquad \text{for } \boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d$$

and C_d is a constant that only depends on the dimension d.

B. Invariant Set Γ_x for the $\mathcal{L}_1$ System

In this subsection, we show how the set Γ_x described in Section IV-B can be shown by inspection to be invariant by $\boldsymbol{M}_x$. We will not mention, however, by what process the invariant set can be found initially, as our current method is limited to practical inspection. Consider an integer $k \geq 1$ and $x \in [\alpha_{k+1}, \alpha_k)$. Consider then the two points Q_T and P_T of the thresholding line $u_1 + \frac{1}{2} u_2 = 0$, defined in the first part of Table II. Let us define the following points:

$$\begin{aligned} Q_0 &:= \boldsymbol{M}_x^1(Q_T) = \boldsymbol{M}_x(Q_T) \\ Q_0' &:= \boldsymbol{M}_x^0(Q_T) \neq \boldsymbol{M}_x(Q_T) \\ P_0 &:= \boldsymbol{M}_x^1(Q_T) = \boldsymbol{M}_x(P_T) \\ P_0' &:= \boldsymbol{M}_x^0(P_T) \neq \boldsymbol{M}_x(P_T). \end{aligned} \tag{B1}$$

Then, for $j \geq 0$, let us recursively define

$$\begin{aligned} Q_{j+1} &:= \boldsymbol{M}_x(Q_j), \quad Q_{j+1}' := \boldsymbol{M}_x(Q_j') \\ P_{j+1} &:= \boldsymbol{M}_x(P_j), \quad P_{j+1}' := \boldsymbol{M}_x(P_j'). \end{aligned} \tag{B2}$$

Thanks to (B1), (B2), and the definition of $\boldsymbol{M}_x$ in (26), one can easy establish the second part of Table II. To derive the third part of the table, one first has to note the following properties. Because of (25), it is trivial that $\boldsymbol{M}_x^0(P) - \boldsymbol{M}_x^1(P) = (1,1)$ for any point P. It then results from (B1) that $Q_0' - Q_0 = (1,1)$ and $P_0' - P_0 = (1,1)$. Next, if two points P and P' are such that $P' - P \in \mathbb{Z}^2$, it is easy to see from (26) and (25) that $\boldsymbol{M}_x(P') - \boldsymbol{M}_x(P) \in \mathbb{Z}^2$. Thus, one recursively obtains from (B2) that $Q_j' - Q_j \in \mathbb{Z}^2$ and $P_j' - P_j \in \mathbb{Z}^2$. We derive these integer difference vectors for a certain number of indexes j and show the results in the third part of the table.

Let us denote by $R_1 R_2 \cdots R_n$ the set enclosed into the polygon of vertices $R_1, R_2, \ldots, R_n$, and by $[R_1, R_2, \ldots, R_n]$ the union of the segments $[R_{i-1}, R_i]$ for $i = 2, \ldots, n$. We define the set

$$\Gamma_x := \left(S_x^0 \setminus B_x^0\right) \cup \left(S_x^1 \setminus B_x^1\right)$$

TABLE II
CHARACTERIZATION OF THE VERTICES OF THE INVARIANT SET OF SYSTEM $\mathcal{L}_1$

Point	Domain	Abscissa	Ordinate	Range
Q_T	$\Omega_x^1 \cap \overline{\Omega_x^0}$	$-\frac{1}{4}+\frac{1}{4(2k-1)}+x(k-3)$	$\frac{1}{2}-\frac{1}{2(2k-1)}-2x(k-3)$	
P_T	$\Omega_x^1 \cap \overline{\Omega_x^0}$	$-\frac{3}{2}x+\frac{1}{4}$	$3x-\frac{1}{2}$	
Q_0	Ω_x^0	$-\frac{3}{4}+\frac{1}{4(2k-1)}+x(k-2)$	$-\frac{1}{4}-\frac{1}{4(2k-1)}-x(k-4)$	
Q_1	Ω_x^0	$-\frac{1}{4}+\frac{1}{4(2k-1)}+x(k-1)$	$3x-\frac{1}{2}$	
Q_2	Ω_x^1	$\frac{1}{4}+\frac{1}{4(2k-1)}+xk$	$-\frac{1}{4}+\frac{1}{4(2k-1)}+x(k+3)$	
Q_3		$-\frac{1}{4}+\frac{1}{4(2k-1)}+x(k+1)$	$-\frac{1}{2}+\frac{1}{2(2k-1)}+2x(k+2)$	
P_{2j}	Ω_x^0	$-\frac{1}{4}+x(2j-\frac{1}{2})$	$-\frac{3}{4}+x(2j^2+\frac{5}{2})$	$j=0,\ldots,k$
P_{2j-1}	Ω_x^1	$\frac{1}{4}+x(2j-\frac{3}{2})$	$-\frac{1}{2}+x(2j^2-2j+3)$	$j=1,\ldots,k$
Q'_0	Ω_x^1	$Q_0+(1,1)$		
Q'_1	Ω_x^1	$Q_1+(0,1)$		
Q'_2	Ω_x^0	$Q_2-(1,0)$		
Q'_3		Q_3		
P'_0	Ω_x^0	$P_0+(1,1)$		
P'_{2j}	Ω_x^0	$P_{2j}+(0,1)$		$j=0,\ldots,k-1$
P'_{2j-1}	Ω_x^1	$P_{2j-1}+(0,1)$		$j=1,\ldots,k-1$
P'_{2k-1}	Ω_x^0	$P_{2k-1}-(1,0)$		
P'_{2k}		P_{2k}		

where

$$S_x^0 := Q_T P'_{2k-4}\cdots P'_4 P'_2 Q'_2 P'_{2k-1} Q_0 P_0 P_2 \cdots P_{2k-2} Q_1 P_T$$
$$S_x^1 := Q_T P'_{2k-2} Q'_1 P'_1 P'_3 \cdots P'_{2k-3} Q'_0 P'_0 Q_2 P_{2k-1} \cdots P_3 P_1 P_T$$
$$B_x^0 := \left[Q_0, P'_{2k-1}, Q'_2, P'_2, P'_4, \ldots, P'_{2k-4}, Q_T, P_T\right]$$
$$B_x^1 := \left[Q_T, P'_{2k-2}, Q'_1, P'_1, P'_3, \ldots, P'_{2k-3}, Q'_0, P'_0\right].$$

The above sets can be recognized in Fig. 8(a). They are also highlighted in Fig. 10(a) where S_x^0 and S_x^1 are represented by shaded areas, and B_x^0 and B_x^1 are represented by a dashed line and a mixed line, respectively. The set Γ_x is basically formed by removing from $S_x^0 \cup S_x^1$ the upper boundary shown in mixed line in Fig. 8(a). Note from the definition of B_x^0 that we are also removing the inner segment $[Q_T, P_T]$ from S_x^0. However, this inner segment still remains in Γ_x because it is part of S_x^1. Now, note that $S_x^0 \setminus B_x^0 \subset \Omega_x^0$ and $S_x^1 \setminus B_x^1 \subset \Omega_x^1$. Therefore,

$$\begin{aligned}\boldsymbol{M}_x(\Gamma_x) &= \boldsymbol{M}_x^0\left(S_x^0 \backslash B_x^0\right) \cup \boldsymbol{M}_x^1\left(S_x^1 \backslash B_x^1\right) \\ &= \left(\boldsymbol{M}_x^0\left(S_x^0\right) \backslash \boldsymbol{M}_x^0\left(B_x^0\right)\right) \cup \left(\boldsymbol{M}_x^1\left(S_x^1\right) \backslash \boldsymbol{M}_x^1\left(B_x^1\right)\right).\end{aligned} \tag{B3}$$

In the last equality, we have used the fact that $\boldsymbol{M}_x^0$ and $\boldsymbol{M}_x^1$ are injective. Let us derive $\boldsymbol{M}_x^0(S_x^0)$. Since $\boldsymbol{M}_x^0$ is affine, $\boldsymbol{M}_x^0(S_x^0)$ is simply the polygonal set whose vertices are obtained by transforming those of S_x^0 through $\boldsymbol{M}_x^0$. Now, except for Q_T and P_T, all the vertices of S_x^0 belong to Ω_x^0. Their images by $\boldsymbol{M}_x^0$ and by $\boldsymbol{M}_x$ are therefore the same. Their images through $\boldsymbol{M}_x$ are then trivially obtained from (B2). Meanwhile, the explicit transformation of Q_T and P_T through $\boldsymbol{M}_x^0$ is obtained from (B1). By applying similar reasonings to $\boldsymbol{M}_x^1(S_x^1)$, $\boldsymbol{M}_x^0(B_x^0)$, and $\boldsymbol{M}_x^1(B_x^1)$ we then find

$$\boldsymbol{M}_x^0(S_x^0) = Q'_0 P'_{2k-3} \cdots P'_5 P'_3 Q'_3 P'_{2k} Q_1 P_1 P_3 \cdots P_{2k-1} Q_2 P'_0$$
$$\boldsymbol{M}_x^1(S_x^1) = Q_0 P'_{2k-1} Q'_2 P'_2 P'_4 \cdots P'_{2k-2} Q'_1 P'_1 Q_3 P_{2k} \cdots P_2 P_0$$
$$\boldsymbol{M}_x^0(B_x^0) = \left[Q_1, P'_{2k}, Q'_3, P'_3, P'_5, \ldots, P'_{2k-3}, Q'_0, P'_0\right]$$
$$\boldsymbol{M}_x^1(B_x^1) = \left[Q_0, P'_{2k-1}, Q'_2, P'_2, P'_4, \ldots, P'_{2k-2}, Q'_1, P'_1\right].$$

These sets can also be recognized in Fig. 8(a) and are highlighted in Fig. 10(b). By using (B3), one can see by inspection that $\boldsymbol{M}_x(\Gamma_x) = \Gamma_x$.

C. Invariant Set Γ_x for the $\mathcal{Q}_1$ System

We briefly describe here the construction principle of the function of (58). The basic idea is to find a change of coordinates such that in the new coordinate system the dynamical system becomes somewhat "simpler." Denote by Φ_x the bijection defining the change of coordinates, and $\tilde{\boldsymbol{M}}_x$ the transformation in the new coordinate system, given by

$$\tilde{\boldsymbol{M}}_x = \Phi_x \boldsymbol{M}_x \Phi_x^{-1}.$$

Let the pieces of $\tilde{\boldsymbol{M}}_x$ on $\tilde{\Omega}_x^0 := \Phi_x(\Omega_x^0)$ and $\tilde{\Omega}_x^1 := \Phi_x(\Omega_x^1)$ be denoted by $\tilde{\boldsymbol{M}}_x^0$ and $\tilde{\boldsymbol{M}}_x^1$, respectively. It turns out that it is possible to find Φ_x which reduces one of $\tilde{\boldsymbol{M}}_x^0$ or $\tilde{\boldsymbol{M}}_x^1$ to a pure translation, while keeping the other one still affine. Assuming $x > 0$, it is actually interesting (and more intuitive) to reduce $\tilde{\boldsymbol{M}}_x^1$ to a translation, since in this case the state variable $\boldsymbol{u}[n]$ stays in Ω_x^1 more frequently than Ω_x^0. This can be realized by setting

$$\Phi_x(u_1, u_2) := \left(u_1, u_2 + \frac{1}{2a_x}\left(u_1 - \frac{a_x}{2}\right)^2 - c_x\right) \tag{C1}$$

where

$$a_x := \frac{1}{2} - x > 0 \tag{C2}$$

and c_x is an arbitrary constant that may depend on x. We denote $\Phi_x(\boldsymbol{u})$ also by $\tilde{\boldsymbol{u}} := (\tilde{u}_1, \tilde{u}_2)$. Then $\tilde{\boldsymbol{M}}_x^1$ is given by

$$\tilde{\boldsymbol{M}}_x^1 \tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}} + \left(x - \frac{1}{2}\right)\boldsymbol{f}$$

and $\tilde{\boldsymbol{M}}_x^0$ by

$$\tilde{\boldsymbol{M}}_x^0 \tilde{\boldsymbol{u}} = \boldsymbol{A}_x \tilde{\boldsymbol{u}} + \left(x + \frac{1}{2}\right)\boldsymbol{g}_x$$

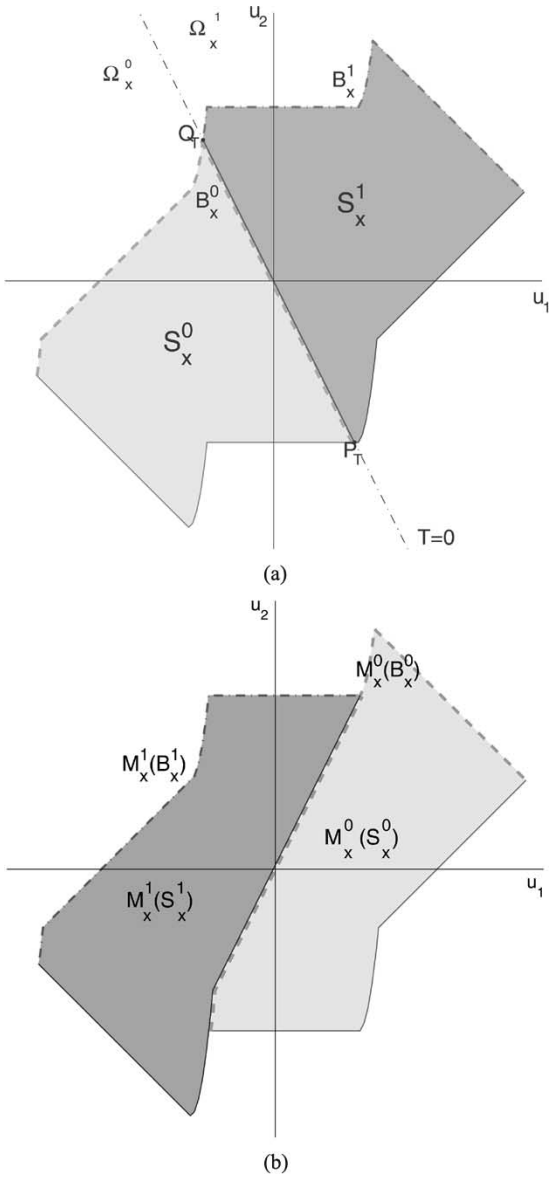


Fig. 10. A schematic diagram of the action of $\boldsymbol{M}_x$ on Γ_x for the $\mathcal{L}_1$ system. (a) Before the mapping. (b) After the mapping.

where

$$\boldsymbol{f} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{A}_x = \begin{pmatrix} 1 & 0 \\ \frac{1}{a_x} & 1 \end{pmatrix}, \text{ and } \boldsymbol{g}_x = \begin{pmatrix} 1 \\ \frac{1}{2a_x} \end{pmatrix}.$$

Note that the description of $\tilde{\boldsymbol{M}}_x$ is independent of the constant c_x and that $\tilde{\boldsymbol{M}}_x^1$ is simply the translation along the $\tilde{u}_1$-axis by the negative constant $-a_x$.

The final ingredient is the specification of the partition $\{\Omega_x^0, \Omega_x^1\}$, or equivalently, the partition $\{\tilde{\Omega}_x^0, \tilde{\Omega}_x^1\}$. This is done with the help of eight characteristic points of the mappings $\tilde{\boldsymbol{M}}_x^0$ and $\tilde{\boldsymbol{M}}_x^1$, denoted by $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3, \tilde{P}_4$ and $\tilde{Q}_1, \tilde{Q}_2, \tilde{Q}_3, \tilde{Q}_4$ (see Fig. 11). These points are defined by

$$\begin{aligned} \tilde{P}_1 &= \left(\frac{1}{2} + \frac{1}{2}a_x, 1\right), \qquad \tilde{P}_2 = \left(-\frac{1}{2} - \frac{1}{2}a_x, 1\right) \\ \tilde{P}_3 &= \left(-\frac{1}{2} + \frac{1}{2}a_x, 0\right), \quad \tilde{P}_4 = \left(\frac{1}{2} - \frac{1}{2}a_x, 0\right) \end{aligned} \tag{C3}$$

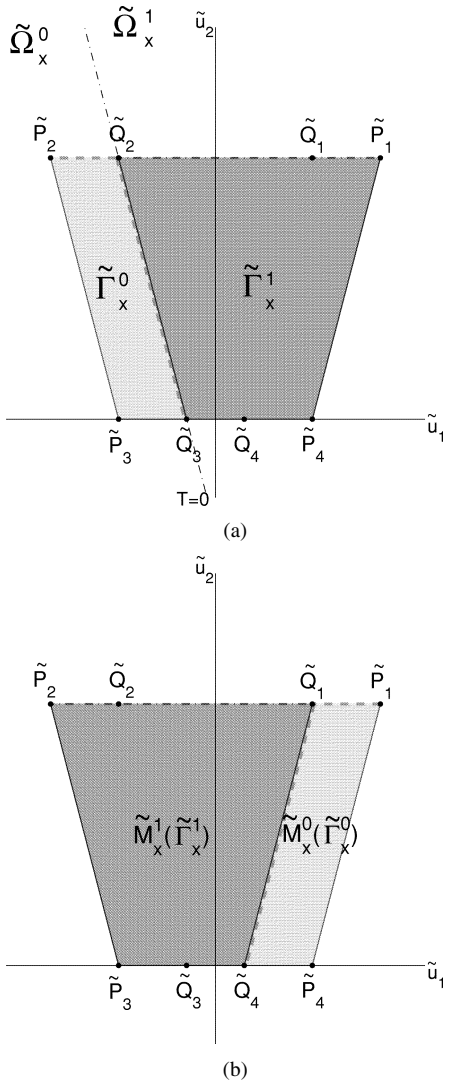


Fig. 11. The invariant set for the $\mathcal{Q}_1$ system in the $(\bar{u}_1, \bar{u}_2)$ domain ($x = 0.24$).

and

$$\begin{aligned} \tilde{Q}_1 &= \tilde{P}_1 - (a_x, 0), \quad \tilde{Q}_2 = \tilde{P}_2 + (a_x, 0) \\ \tilde{Q}_3 &= \tilde{P}_3 + (a_x, 0), \quad \tilde{Q}_4 = \tilde{P}_4 - (a_x, 0). \end{aligned} \tag{C4}$$

Consider the set $\tilde{\Gamma}_x := \tilde{\Gamma}_x^0 \cup \tilde{\Gamma}_x^1$ where

$$\begin{aligned} \tilde{\Gamma}_x^0 &:= \tilde{Q}_3\tilde{Q}_2\tilde{P}_2\tilde{P}_3 \setminus \left[\tilde{Q}_3, \tilde{Q}_2, \tilde{P}_2\right] \quad \text{and} \\ \tilde{\Gamma}_x^1 &:= \tilde{P}_4\tilde{P}_1\tilde{Q}_2\tilde{Q}_3 \setminus \left[\tilde{P}_1, \tilde{Q}_2\right]. \end{aligned}$$

It is implied in these definitions that $\tilde{\Gamma}_x^0$ is formed by taking the parallelograms $\tilde{Q}_3\tilde{Q}_2\tilde{P}_2\tilde{P}_3$ without the two boundary segments $[\tilde{Q}_3, \tilde{Q}_2]$ and $[\tilde{Q}_2, \tilde{P}_2]$, and $\tilde{\Gamma}_x^1$ is formed in a similar manner. These two sets are illustrated in Fig. 11(a). Note from the figure that $\tilde{\Gamma}_x$ is then simply the trapezoid $\tilde{P}_4\tilde{P}_1\tilde{P}_2\tilde{P}_3$ from which the upper boundary segment $[\tilde{P}_1, \tilde{P}_2]$ has been removed. Now, one can easily check that

$$\begin{aligned} \tilde{\boldsymbol{M}}_x^0(\tilde{\Gamma}_x^0) &= \tilde{P}_1\tilde{Q}_1\tilde{Q}_4\tilde{P}_4 \setminus \left[\tilde{P}_1, \tilde{Q}_1, \tilde{Q}_4\right] \quad \text{and} \\ \tilde{\boldsymbol{M}}_x^1(\tilde{\Gamma}_x^1) &= \tilde{Q}_4\tilde{Q}_1\tilde{P}_2\tilde{P}_3 \setminus \left[\tilde{Q}_1\tilde{P}_2\right]. \end{aligned}$$

These sets are illustrated in Fig. 11(b). One can easily see that $\tilde{\boldsymbol{M}}_x^0(\tilde{\Gamma}_x^0) \cup \tilde{\boldsymbol{M}}_x^1(\tilde{\Gamma}_x^1) = \tilde{\Gamma}_x$. If we choose the straight line passing through $\tilde{Q}_2$ and $\tilde{Q}_3$ as the boundary between $\tilde{\Omega}_x^0$ and $\tilde{\Omega}_x^1$, i.e.,

$$\tilde{\Omega}_x^1 = \left\{ (\tilde{u}_1, \tilde{u}_2) : \tilde{u}_1 + a_x \left(\tilde{u}_2 - \frac{3}{2} \right) + \frac{1}{2} \geq 0 \right\}$$

then we ensure that $\tilde{\Gamma}_x^0 \subset \tilde{\Omega}_x^0$ and $\tilde{\Gamma}_x^1 \subset \tilde{\Omega}_x^1$. In this situation, we have $\tilde{\boldsymbol{M}}_x(\tilde{\Gamma}_x) = \tilde{\boldsymbol{M}}_x^0(\tilde{\Gamma}_x^0) \cup \tilde{\boldsymbol{M}}_x^1(\tilde{\Gamma}_x^1) = \tilde{\Gamma}_x$. Back to the original space, the set $\Gamma_x := \Phi_x^{-1}(\tilde{\Gamma}_x)$ then satisfies $\boldsymbol{M}_x(\Gamma_x) = \Gamma_x$. Because of the quadratic nature of Φ_x, it is clear that Γ_x has a boundary composed of four parabolic pieces. This is illustrated in Fig. 9. Because the boundary segment $[\tilde{P}_1, \tilde{P}_2]$ is excluded from $\tilde{\Gamma}_x$, the boundary parabola passing through P_1 and P_2 is excluded from Γ_x.

To find the expression for T back in the original system of coordinates, we substitute the expressions for $\tilde{u}_1$, $\tilde{u}_2$ and a_x from (C1) and (C2). Then the resulting function T, up to a scaling factor, is given by (58) with

$$C(x) = 8 \left(x - \frac{1}{2} \right) c_x - 3x^2 + 11x - \frac{7}{4}. \qquad \text{(C5)}$$

In this paper, the choice of $C(x)$ does not matter. However, when dealing with time-varying inputs, it is shown in [23] that it is interesting to choose $C(x)$ so that the centroid of Γ_x is located at $(0,0)$ regardless of x. In this situation, it is indeed numerically shown that the resulting modulator becomes superior in performance to the one-bit linear-T second-order modulators. Since Γ_x is here entirely known analytically, such a value of $C(x)$ is easy to derive. We show in the Appendix, Subsection E that this is achieved when $C(x) = (7 - 3x)x - \frac{1}{12}$ for $x \in [0, \frac{1}{2})$.

D. Proof of Proposition 4.1

Property 1 can be checked in a straightforward manner using the explicit parametric descriptions of the invariant sets given in the respective sections and in the Appendix. Property 2 is a consequence of the fact that each of the invariant sets possesses a boundary that is composed of a finite number of smooth curves, totaling a finite perimeter. Hence, the ϵ-neighborhood of each $\partial\Gamma_x$ cover an area that decreases as $O(\epsilon)$ as $\epsilon \to 0$. The uniformity of the constants M_0 and C_0 are guaranteed by the choice of the intervals $I(\mathcal{D})$. In particular, these intervals can be chosen to be $\left[-\frac{1}{2}, \frac{1}{2}\right]$, $\left[-\frac{1}{6}, \frac{1}{6}\right]$ and an arbitrary closed subinterval of $\left(-\frac{1}{2}, \frac{1}{2}\right)$, respectively. Let us prove Property 3.

The $\mathcal{L}_2$ system: This was shown in Section IV-A.

The $\mathcal{Q}_1$ system: The invariant set is $\Gamma_x = \Phi_x^{-1}(\tilde{P}_1\tilde{P}_2\tilde{P}_3\tilde{P}_4)$ where the points $\tilde{P}_i$ are given in (C3) and Φ_x is given in (C1). From (C3), we derive the equations of the four linear boundaries of the parallelograms $\tilde{P}_1\tilde{P}_2\tilde{P}_3\tilde{P}_4$ and obtain

$$\begin{aligned}
\left(\tilde{P}_1\tilde{P}_2\right) &: \tilde{u}_2 = 1 \\
\left(\tilde{P}_2\tilde{P}_3\right) &: \tilde{u}_2 = -\frac{1}{2a_x}\tilde{u}_1 - \frac{1}{2a_x} + \frac{1}{2} \\
\left(\tilde{P}_3\tilde{P}_4\right) &: \tilde{u}_2 = 0 \\
(\tilde{P}_1\tilde{P}_4) &: \tilde{u}_2 = +\frac{1}{2a_x}\tilde{u}_1 - \frac{1}{2a_x} + \frac{1}{2}.
\end{aligned}$$

By writing $(\tilde{u}_1, \tilde{u}_2) = \Phi_x(u_1, u_2)$ and applying (C1), we derive the equations of the four parabolic boundaries of $\Phi_x^{-1}(\tilde{P}_1\tilde{P}_2\tilde{P}_3\tilde{P}_4)$ and obtain

$$\begin{aligned}
\left(\widehat{P_1P_2}\right) &: u_2 = -\frac{1}{2a_x}\left(u_1 - \frac{a_x}{2}\right)^2 + c_x + 1 \\
\left(\widehat{P_2P_3}\right) &: u_2 = -\frac{1}{2a_x}\left(u_1 - \frac{a_x}{2}\right)^2 + c_x - \frac{1}{2a_x}u_1 - \frac{1}{2a_x} + \frac{1}{2} \\
\left(\widehat{P_3P_4}\right) &: u_2 = -\frac{1}{2a_x}\left(u_1 - \frac{a_x}{2}\right)^2 + c_x \\
\left(\widehat{P_1P_4}\right) &: u_2 = -\frac{1}{2a_x}\left(u_1 - \frac{a_x}{2}\right)^2 + c_x + \frac{1}{2a_x}u_1 - \frac{1}{2a_x} + \frac{1}{2}.
\end{aligned}$$

One can then easily check that the four above parabolas satisfy the following relations:

$$\begin{aligned}
\left(\widehat{P_1P_4}\right) &= \left(\widehat{P_1P_2}\right) + (1,0) \\
\left(\widehat{P_3P_4}\right) &= \left(\widehat{P_1P_2}\right) - (0,1) \\
\left(\widehat{P_2P_3}\right) &= \left(\widehat{P_1P_2}\right) - (1,1). \qquad \text{(D1)}
\end{aligned}$$

This can be also graphically seen in Fig. 9. This is sufficient to prove Property 3 for the $\mathcal{Q}_1$ system. A graphical representation of the tiling property is shown in the same figure.

The $\mathcal{L}_1$ system: Proving the tiling property of the invariant set Γ_x described in Section IV-B is a tedious process. Here, we will only point out boundary relations similar to (D1). Given n vertices $R_1, R_2, \ldots, R_n$, let us use the notation $[R_1, R_2, \ldots, R_n]$ to designate the union of the segments $[R_{i-1}, R_i]$ for $i = 2, \ldots, n$. By using Table II, one can see that

$$\begin{aligned}
[P_{2k-1}, Q_2] &= [P'_{2k-1}, Q'_2] + (1,0) \\
[Q_0, P_0] &= [Q'_0, P'_0] - (1,1)
\end{aligned}$$

and

$$\begin{aligned}
&[P_2, P_4, \ldots, P_{2k-2}, Q_1, P_1, P_3, \ldots, P_{2k-3}] \\
&\quad = [P'_2, P'_4, \ldots, P'_{2k-2}, Q'_1, P'_1, P'_3, \ldots, P'_{2k-3}] - (0,1).
\end{aligned}$$

We illustrate these three relations by the three arrows in Fig. 8(b). □

E. On the Analysis of the Quadratic Scheme: Zero-Centroid Setting of $C(x)$

Let us call $G_x = (u_{1,x}, u_{2,x})$ the centroid point of Γ_x and write $\tilde{\boldsymbol{u}} = (\tilde{u}_1, \tilde{u}_2)$ in general. We have

$$G_x = \int_{\boldsymbol{u} \in \Gamma_x} \boldsymbol{u}\, d\boldsymbol{u} = \int_{\tilde{\boldsymbol{u}} \in \Phi_x(\Gamma_x)} \Phi_x^{-1}(\tilde{\boldsymbol{u}})\, d\tilde{\boldsymbol{u}}.$$

In the last equality, we have used the fact that $d\boldsymbol{u} = d\tilde{\boldsymbol{u}}$ since the transformation Φ_x from (C1) conserves measure. We know from Section IV-C that $\Phi_x(\Gamma_x)$ is the trapezoid $\tilde{P}_1\tilde{P}_2\tilde{P}_3\tilde{P}_4$. From (C2) and the explicit coordinates of its vertices given in (C3) and (C4), we have

$$\Phi_x(\Gamma_x) = \left\{ (\tilde{u}_1, \tilde{u}_2) : 0 \leq \tilde{u}_2 \leq 1 \text{ and } |\tilde{u}_1| \leq \frac{1}{2} + a_x \left(\tilde{u}_2 - \frac{1}{2} \right) \right\}.$$

From (C1), one easily derives that

$$\Phi_x^{-1}(\tilde{\boldsymbol{u}}) = \left(\tilde{u}_1, \tilde{u}_2 - \frac{1}{2a_x}\left(\tilde{u}_1 - \frac{a_x}{2} \right)^2 + c_x \right).$$

Consequently, the first component of G_x is

$$u_{1,x} = \int_{\tilde{\boldsymbol{u}} \in \Phi_x(\Gamma_x)} \tilde{u}_1 \, d\tilde{\boldsymbol{u}}.$$

Since $\Phi_x(\Gamma_x)$ is clearly symmetrical with respect to the $\tilde{u}_2$-axis, we already have $u_{1,x} = 0$. The second component of G_x is equal to

$$\begin{aligned} u_{2,x} &= \int_0^1 \int_{-\frac{1}{2}-a_x(u-\frac{1}{2})}^{\frac{1}{2}+a_x(u-\frac{1}{2})} \left(\tilde{u}_2 - \frac{1}{2a_x} \left(\tilde{u}_1 - \frac{a_x}{2} \right)^2 + c_x \right) d\tilde{u}_1 d\tilde{u}_2 \\ &= \frac{1}{2} - \frac{1}{24a_x} + c_x. \end{aligned}$$

The component u_x will then be systematically equal to 0 by choosing $c_x = \frac{1}{24a_x} - \frac{1}{2}$. Using (C2) and (C5), this implies that $C(x) = (7 - 3x)x - \frac{1}{12}$.

ACKNOWLEDGMENT

The authors would like to thank Ingrid Daubechies, Ron DeVore, Özgür Yılmaz, and Jade Vinson for interesting discussions on the topic of this paper, and the anonymous referees for their helpful comments and suggestions on the presentation. C. Sinan Güntürk would like to thank the Institute for Advanced Study and Courant Institute for their hospitality during the writing stage of this work.

REFERENCES

[1] J. C. Candy and G. C. Temes, Eds., *Oversampling Delta-Sigma Data Converters: Theory Design and Simulation*. Piscataway, NJ: IEEE Press, 1992.

[2] W. Chou, T. H. Meng, and R. M. Gray, "Time domain analysis of sigma delta modulation," in *Proc. ICASSP-90 , Int. Conf. Acoustics, Speech and Signal Processing*, vol. 3, Albuquerque, NM, Apr. 1990, pp. 1751–1754.

[3] W. Chou, P. W. Wong, and R. M. Gray, "Multistage sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 35, pp. 784–796, July 1989.

[4] I. Daubechies and R. A. DeVore, "Reconstructing a bandlimited function from very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order," *Ann. Math.*, vol. 158, no. 2, pp. 679–710, Sept. 2003.

[5] M. Drmota and R. F. Tichy, *Sequences, Discrepancies and Applications*. New York: Springer-Verlag, 1997.

[6] R. M. Gray, "Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input," *IEEE Trans. Commun.*, vol. 37, pp. 588–599, June 1989.

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

[8] C. S. Güntürk, "Improved error estimates for first order sigma-delta modulation," in *Proc. Int. Workshop Sampling Theory and Applications, SampTA'99*, Loen, Norway, Aug. 1999, pp. 171–176.

[9] ——, "Harmonic analysis of two problems in signal quantization and compression," Ph.D. dissertation, Princeton Univ., Princeton, NJ, 2000.

[10] ——, "Approximating a bandlimited function using very coarsely quantized data: Improved error estimates in sigma-delta modulation," *J. Amer. Math. Soc.*, vol. 17, pp. 229–242, Jan. 2004.

[11] C. S. Güntürk, J. Lagarias, and V. Vaishampayan, "On the robustness of single loop sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1735–1744, July 2001.

[12] N. He, F. Kuhlmann, and A. Buzo, "Double-loop sigma-delta modulation with dc input," *IEEE Trans. Commun.*, vol. 38, pp. 487–495, Apr. 1990.

[13] ——, "Multi-loop sigma-delta quantization," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1015–1028, May 1992.

[14] D. F. Hoschele Jr, *Analog-to-Digital and Digital-to-Analog Conversion Techniques*. New York: Wiley, 1994.

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

[16] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*. New York: Wiley, 1974.

[17] M. Laczkovich, "Discrepancy estimates for sets with small boundary," *Stud. Sci. Math. Hung.*, vol. 30, pp. 105–109, 1995.

[18] H. Niederreiter, "Application of diophantine approximations to numerical integration," in *Diophantine Approximation and its Applications*, C. F. Osgood, Ed. New York: Academic, 1973, pp. 129–199.

[19] S. R. Norsworthy, R. Schreier, and G. C. Temes, *Delta-Sigma Data Converters: Theory, Design and Simulation*. Piscataway, NJ: IEEE Press, 1996.

[20] W. Parry, *Topics in Ergodic Theory*. Cambridge, U.K.: Cambridge Univ. Press, 1981.

[21] R. Schreier, M. V. Goodson, and B. Zhang, "An algorithm for computing convex positively invariant sets for delta-sigma modulators," *IEEE Trans. Circuits Syst., I*, vol. 44, pp. 38–44, Jan. 1997.

[22] N. T. Thao, "Vector quantization analysis of $\Sigma\Delta$ modulation," *IEEE Trans. Signal Processing*, vol. 44, pp. 808–817, Apr. 1996.

[23] ——, "MSE behavior and centroid function of m th order asymptotic $\Sigma\Delta$ modulators," *IEEE Trans. Circuits Syst., II*, vol. 49, pp. 86–100, Feb. 2002.

[24] C. S. Güntürk and N. T. Thao, "Ergodic dynamics in $\Sigma\Delta$ quantization: Tiling invariant sets and spectral analysis of error," *Adv. Appl. Math.*, to be published.

[25] N. T. Thao and M. Vetterli, "Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates," *IEEE Trans. Signal Processing*, vol. 42, pp. 519–531, Mar. 1994.

[26] Ö Yılmaz, "Stability analysis for several sigma-delta methods of coarse quantization of bandlimited functions," *Constructive Approx.*, vol. 18, no. 4, pp. 599–623, 2002.