Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Advances in Applied Mathematics 34 (2005) 523–560
ADVANCES IN Applied Mathematics
www.elsevier.com/locate/yaama

Ergodic dynamics in sigma–delta quantization: tiling invariant sets and spectral analysis of error ☆

C. Sinan Güntürk [a,*], Nguyen T. Thao [b]

[a] *Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA*
[b] *Department of Electrical Engineering, City College and Graduate School, City University of New York, Convent Avenue at 138th Street, New York, NY 10031, USA*

Received 5 September 2003; accepted 17 November 2003

Abstract

This paper has two themes that are intertwined. The first is the dynamics of certain piecewise affine maps on $\mathbb{R}^m$ that arise from a class of analog-to-digital conversion methods called $\Sigma\Delta$ (sigma–delta) quantization. The second is the analysis of reconstruction error associated with each such method.

$\Sigma\Delta$ quantization generates approximate representations of functions by sequences that lie in a restricted set of discrete values. These are special sequences in that their local averages track the function values closely, thus enabling simple convolutional reconstruction. In this paper, we are concerned with the approximation of constant functions only, a basic case that presents surprisingly complex behavior. An mth order $\Sigma\Delta$ scheme with input x can be translated into a dynamical system that produces a discrete-valued sequence (in particular, a 0–1 sequence) q as its output. When the schemes are stable, we show that the underlying piecewise affine maps possess invariant sets that tile $\mathbb{R}^m$ up to a finite multiplicity. When this multiplicity is one (the single-tile case), the dynamics within the tile is isomorphic to that of a generalized skew translation on $\mathbb{T}^m$.

The value of x can be approximated using any consecutive M elements in q with increasing accuracy in M. We show that the asymptotical behavior of reconstruction error depends on the regularity of the invariant sets, the order m, and some arithmetic properties of x. We determine the behavior

☆ This work has been supported in part by the National Science Foundation Grants DMS-0219072, DMS-0219053 and CCR-0209431.

* Corresponding author.
E-mail addresses: gunturk@cims.nyu.edu (C.S. Güntürk), thao@ee-mail.engr.ccny.cuny.edu (N.T. Thao).

0196-8858/$ – see front matter
doi:10.1016/j.aam.2003.11.004

in a number of cases of practical interest and provide good upper bounds in some other cases when exact analysis is not yet available.

1. Introduction

This paper is motivated by the mathematical problems exhibited in and suggested by a class of real-world practical algorithms that are used to perform analog-to-digital conversion of signals. There will be two themes in our study of these mathematical problems. The first theme is the dynamics of certain piecewise affine maps on $\mathbb{R}^m$ that are associated with these algorithms. The second theme is the analysis of the reconstruction error. While the first theme is somewhat independent of the second and is of great interest on its own, the second theme turns out to be crucially dependent on the first and is of interest for theoretical as well as practical reasons.

Let us start with the following abstract algorithm for analog-to-digital encoding. For each input real number x in some interval I, there is a map $\mathcal{T}_x$ on a space $\mathcal{S}$, and a finite partition $\Pi_x = \{\Omega_{x,1}, \ldots, \Omega_{x,K}\}$ of $\mathcal{S}$. For a fixed set of real numbers $d_1 < \cdots < d_K$, and a typically fixed (but arbitrary) initial point $u_0 \in \mathcal{S}$, we define a discrete-valued output sequence $q := q_x$ via

$$q[n] = d_i \quad \text{if } u[n-1] := \mathcal{T}_x^{n-1}(u_0) \in \Omega_{x,i}. \tag{1.1}$$

We would like the mapping $x \mapsto q$ to be invertible in a very special way: for an *input-independent* family of averaging kernels $\phi_M \in \ell^1(\mathbb{Z})$, $M = 1, 2, \ldots$, we require that for all $x \in I$, as $M \to \infty$,

$$(q * \phi_M)[n] := \sum_k \phi_M[k] q[n-k] \to x, \qquad \text{uniformly in } n. \tag{1.2}$$

For normalization, we ask that the size of the averaging window (i.e., the support of ϕ_M) grow linearly in M,[1] and the weights satisfy $\sum_n \phi_M[n] = 1$.

Note that such an encoding of real numbers is inherently different from binary-expansion (or any other expansion in a number system) in that, due to (1.2), equal length segments of the sequence q are required to be equally good in approximating the value of x. Hence, there is a "translation-invariance" property in the representation.

This setting is a special case of a more general one in which $x = (x[n])_{n\in\mathbb{Z}}$ is a bounded sequence taking values in I and

$$q[n] = d_i \quad \text{if } u[n-1] \in \Omega_{x[n],i}, \tag{1.3}$$

[1] It will be of interest to use infinitely supported kernels as well. We will define the necessary modifications to handle this situation later.

where we now define $u[n] := \mathcal{T}_{x[n]}(u[n-1])$, and require that

$$(q - x) * \phi_M \to 0 \quad \text{uniformly.} \tag{1.4}$$

The basic motivation behind this type of encoding is the following intuitive idea. Let the elements $x[n]$ be closely and regularly spaced samples of a smooth function $X : \mathbb{R} \to I$. Since local averages of these samples around any point k would approximate $x[k]$, i.e., $x * \phi_M \approx x$ for suitable ϕ_M, (1.4) would then imply that the sequence x (and therefore the function X) can be approximated by the convolution $q * \phi_M$.

Such analog-to-digital encoding algorithms have been developed and used in electrical engineering for a few decades now. Most notable examples are the $\Sigma\Delta$ *quantization* (also called $\Sigma\Delta$ *modulation*) of audio signals and the closely related *error-diffusion* in digital halftoning of images. There are several sources in the electrical engineering literature on the theoretical and practical aspects of $\Sigma\Delta$ quantization [6,10,22]. Digital halftoning and its connections to $\Sigma\Delta$ quantization can be found in [1,2,4,20,26]. Recently, $\Sigma\Delta$ quantization has also received interest in the mathematical community, especially in approximation theory and information theory, since a very important question is the rate of convergence in (1.4) [5,9,13,14,16].

We give in Section 2 the original description of an mth order $\Sigma\Delta$ modulation scheme in terms of difference equations. The underlying specific map $\mathcal{T}_x$, which we then refer to as $\mathcal{M}_x$ (the "modulator map"), is described in Section 4; $\mathcal{M}_x$ is the piecewise affine transformation on $\mathcal{S} = \mathbb{R}^m$ defined by

$$\mathcal{M}_x(\mathbf{v}) = \mathbf{L}\mathbf{v} + (x - d_i)\mathbf{1} \quad \text{if } \mathbf{v} \in \Omega_{x,i}, \tag{1.5}$$

where $\mathbf{L} := \mathbf{L}_m$ is the $m \times m$ lower triangular matrix of 1's and $\mathbf{1} := \mathbf{1}_m := (1, \ldots, 1)^\top \in \mathbb{Z}^m$. Each $\Sigma\Delta$ scheme is therefore characterized by its order m, the partition Π_x, and the numbers $\{d_i\}$. A scheme is called k-bit if the size K of the partition Π_x satisfies $2^{k-1} < K \leqslant 2^k$. If the numbers $\{d_i\}$ are in an arithmetic progression, this is referred to as *uniform quantization*. As a consequence of the normalization $\sum_n \phi_M[n] = 1$, the input numbers x are chosen in $I \subset [d_1, d_K]$. A scheme is said to be *stable* if for each x, forward trajectories under the action of $\mathcal{M}_x$ are bounded in $\mathbb{R}^m$. (More refined definitions of stability will be given in Section 4.) The partition Π_x is an essential part of the algorithm for its central role in stability.

It is natural to measure the accuracy of a scheme by how fast the worst case error $\|(x - q) * \phi_M\|_\infty$ converges to zero. It is known that for an mth order stable scheme, and an appropriate choice for the family $\mathcal{F} = \{\phi_M\}$ of filters,[2] this quantity is $O(M^{-m})$ [9]. The hidden constant depends on the scheme as well as the input sequence x. Here, the exponent m is not sharp; in fact, for $m = 1$ and $m = 2$, improvements have been given for various schemes [13,15]. We will review the basic approximation properties of $\Sigma\Delta$ quantization in Section 2.

[2] We shall adopt the electrical engineering terminology "filter" to refer to a sequence (or function) that acts convolutionally.

In applications, it is also common to measure the error in the root mean square norm due its more robust nature (this norm is defined in Section 3). It is known for a small class of schemes we call *ideal*, and a small class of sequences (basically, constants and pure sinusoids) that this norm, when averaged over a smooth distribution of values of x, has the asymptotic behavior $O(M^{-m-1/2})$ [8,11,17]. The analyses employed in obtaining these results rely on very special properties of these ideal schemes, such as employing an (effective) m-bit uniform quantizer for the mth order scheme. It was not known how to extend these results to low-bit schemes (in particular, 1-bit schemes) of high order for which experimental results and simulation suggested similar asymptotical behavior for the root mean square error.

It is the topic of this paper to provide a general framework and methodology to analyze $\Sigma\Delta$ quantization in an arbitrary setup (in terms of partition and number of bits) when inputs are constant sequences. With regard to the first theme of this paper, we prove in Section 5 that the maps $\mathcal{M}_x$ have an outstanding property of yielding *tiling invariant sets*, up to a multiplicity that is determined by the map. In the particular case of single tiles being invariant under $\mathcal{M}_x$ (which also appears to be systematically satisfied by all practical $\Sigma\Delta$ quantization schemes), we develop a spectral theory of $\Sigma\Delta$ quantization. This constitutes the second theme of the paper. The particular consequence of tiling that enables our spectral analysis is presented in Section 6. The resulting new error analysis for general and particular cases is presented in the remainder of the paper.

Some notation

The symbols $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ denote the set of real numbers, the set of integers and the set of natural numbers, respectively. $\mathbb{T}$ denotes the set of real numbers modulo 1, i.e., $\mathbb{T} = \mathbb{R}/\mathbb{Z}$. Functions on $\mathbb{R}^m$ that are 1-periodic in each dimension are assumed to be defined on $\mathbb{T}^m$ via the identification $\mathbb{T} = [0, 1)$, and functions defined on $\mathbb{T}^m$ are extended to $\mathbb{R}^m$ by periodization.

Vectors and matrices are denoted in boldface letters. Transpose is denoted by an upperscript $\top$. The jth coordinate of a vector $\mathbf{v}$ is denoted by v_j, unless otherwise specified. Sequence elements are denoted using brackets, such as in $\omega = (\omega[n])_{n\in\mathbb{Z}}$. The sequence $\tilde{\omega}$ denotes time reversal of ω defined by $\tilde{\omega}[n] := \omega[-n]$, and the symbol $*$ is used to denote the convolution operation.

We define two types of autocorrelation. For a square integrable real-valued function f, we define

$$A_f(t) := \big(f * \tilde{f}\big)(t) = \int f(\xi) f(\xi + t)\, \mathrm{d}\xi.$$

On the other hand, we define the autocorrelation ρ_ω for a bounded (real-valued) sequence ω by the formula

$$\rho_\omega[k] := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \omega[n]\omega[n+k],$$

provided the limit exists.

The Fourier series coefficients of a measure μ on $\mathbb{T}$ are given by

$$\hat{\mu}[n] := \int_{\mathbb{T}} e^{-2\pi i n\xi}\, \mathrm{d}\mu(\xi),$$

and the Fourier transform of a sequence $h = (h[n])_{n\in\mathbb{Z}}$ is denoted by the capital letter H, i.e.,

$$H(\xi) := \sum_{n\in\mathbb{Z}} h[n] e^{2\pi i n\xi}.$$

Hence, when Fourier inversion holds, we have $\widehat{H}[n] = h[n]$.

The "big oh" $f = O(g)$ and the "small oh" $f = o(g)$ notations will have their usual meanings. When it matters, we also use the notation $f \lesssim_\alpha g$ to denote that there exists a constant C that possibly depends on the parameter (or set of parameters) α such that $f \leqslant Cg$. We write $f \asymp g$ if $f \lesssim g$ and $g \lesssim f$, which is the same as $f = \Theta(g)$.

2. Basic theory of $\Sigma\Delta$ quantization

In this section, we describe the principles of $\Sigma\Delta$ quantization (modulation) via a set of defining difference equations. The description in terms of piecewise affine maps on $\mathbb{R}^m$ will be given in Section 4. Although the schemes representable by these difference equations do not constitute the whole collection of algorithms called by the name $\Sigma\Delta$ modulation, they are sufficiently general to cover a large class of algorithms that are used in practice and many more to be investigated.

Let m be the order of the scheme, and $x = (x[n])_{n\in\mathbb{Z}}$ be the input sequence. Then a sequence of state-vectors, denoted

$$\mathbf{u}[n] = \big(u_1[n], \ldots, u_m[n]\big)^\top, \quad n = 0, 1, \ldots$$

and a sequence of output quantized values (or symbols), denoted $q[n]$, $n = 1, 2, \ldots$, are defined recursively via the set of equations

$$\left\{\begin{array}{rcl} q[n] &=& Q(x[n], \mathbf{u}[n-1]), \\ u_1[n] &=& u_1[n-1] + x[n] - q[n], \\ u_2[n] &=& u_2[n-1] + u_1[n], \\ \vdots & = & \vdots \\ u_m[n] &=& u_m[n-1] + u_{m-1}[n], \end{array}\right\} \tag{2.1}$$

where the mapping $Q : \mathbb{R}^{m+1} \to \{d_1, \ldots, d_K\}$, called the *quantization rule*, or simply the *quantizer* of the $\Sigma\Delta$ modulator, is specific to the scheme. In circuit theory, these equations are represented as a feedback-loop system via the block diagram given in Fig. 1.

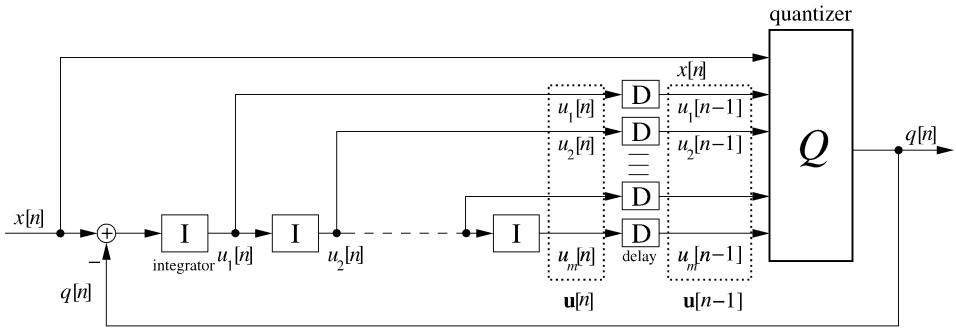


Fig. 1. Block diagram of an mth order $\Sigma\Delta$ modulator.

In addition to producing the output sequence q, the role of the quantizer Q of a $\Sigma\Delta$ modulator is to keep the variables u_j bounded. A more precise definition of this notion of stability will be given later. Let us see how boundedness of u_j results in a simple reconstruction algorithm. It can be seen directly from (2.1) that for each $j = 1, \dots, m$, the state variable u_j satisfies

$$x - q = \Delta^j u_j, \tag{2.2}$$

where Δ is the difference operator defined by $(\Delta v)[n] = v[n] - v[n-1]$. Consider $j = 1$, and assume that x is constant. From this, it follows that

$$\begin{aligned}\left| x - \frac{1}{M} \sum_{k=n+1}^{n+M} q[k] \right| &= \frac{1}{M} \left| \sum_{k=n+1}^{n+M} \big(x - q[k]\big) \right| = \frac{1}{M} \left| \sum_{k=n+1}^{n+M} \big(u_1[k] - u_1[k-1]\big) \right| \\ &= \frac{1}{M} \big| u_1[n+M] - u_1[n] \big| \leqslant \frac{2}{M} \|u_1\|_\infty. \end{aligned} \tag{2.3}$$

This means that simple averaging of *any* M consecutive output values $q[k]$ yields a reconstruction within $O(M^{-1})$.

This approximation result can be generalized easily. For simplicity of the discussion, let us assume that the difference equation (2.2) is satisfied on the whole of $\mathbb{Z}$ (with some care, this can be achieved via backwards iteration of (2.1)). For a given averaging filter $\phi \in \ell^1(\mathbb{Z})$ with $\sum_n \phi[n] = 1$, let

$$e_{x,\phi} := x - q * \phi \tag{2.4}$$

be the error sequence. Since x is a constant sequence, we have $x = x * \phi$. Therefore

$$e_{x,\phi} = (x - q) * \phi = \big(\Delta^m u_m\big) * \phi = u_m * \big(\Delta^m \phi\big), \tag{2.5}$$

where at the last step we have used commutativity of convolutional operators. From this, we obtain

$$\|e_{x,\phi}\|_\infty \leqslant \|u_m\|_\infty \big\|\Delta^m \phi\big\|_1. \tag{2.6}$$

It is not hard to show that there is a family of averaging kernels $\phi_{M,m}$ (which can be, for instance, discrete B-splines of degree m) with support size growing linearly in M such that $\|\Delta^m \phi_{M,m}\|_1 \leqslant C_m M^{-m}$. Combined with (2.6), this yields the bound $O(M^{-m})$ on the uniform approximation error. A proof of this result in the more general setting of oversampling of bandlimited functions can be found in [9,12].

3. Mean square error and its spectral representation

For the rest of this paper, we shall be interested in the mean square error (also called, the *time-averaged square error*) of approximation defined by

$$\mathcal{E}(x,\phi) := \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \big|e_{x,\phi}[n]\big|^2, \tag{3.1}$$

provided the limit exists (otherwise the lim is replaced by a lim sup). The *root* mean square error is defined to be $\sqrt{\mathcal{E}(x,\phi)}$. For convenience in the notation, we shall work with $\mathcal{E}(x,\phi)$.

The mean square error enjoys properties that are desirable from an analytic point of view. The definition of autocorrelation sequence yields an alternative description given by

$$\mathcal{E}(x,\phi) = \rho_{e_{x,\phi}}[0]. \tag{3.2}$$

Using the formula (2.5) and the standard relation $\rho_{\omega * g} = \rho_\omega * g * \tilde{g}$ whenever ρ_ω exists and $g \in l^1$, we find that

$$\mathcal{E}(x,\phi) = \Big(\rho_{u_m} * \big(\Delta^m \phi\big) * \widetilde{\big(\Delta^m \phi\big)}\Big)[0]. \tag{3.3}$$

We shall abbreviate ρ_{u_m} by ρ_u.

The computation of $\mathcal{E}(x,\phi)$ can also be carried out in the spectral domain. Since ρ_u is positive-definite, it constitutes, by Herglotz' theorem [19, p. 38], the Fourier coefficients of a non-negative measure μ on $\mathbb{T}$ (the *power spectral measure*), i.e.,

$$\rho_u[k] = \int_{\mathbb{T}} e^{-2\pi i k\xi}\, \mathrm{d}\mu(\xi). \tag{3.4}$$

Combining this result with (3.3) and elementary Fourier analysis yields the spectral formula

$$\mathcal{E}(x,\phi) = \int_{\mathbb{T}} \left|2\sin(\pi\xi)\right|^{2m} \left|\Phi(\xi)\right|^2 \mathrm{d}\mu(\xi), \tag{3.5}$$

where Φ has the absolutely convergent Fourier series representation

$$\Phi(\xi) = \sum_n \phi[n] e^{2\pi i n\xi}.$$

This computational alternative is effective when the measure μ has a simple description. On the other hand, it can happen that this measure is too complex to compute with directly. In our case, as we shall demonstrate, μ will generally have a pure point (discrete) component μ_{pp}, and an absolutely continuous component μ_{ac}. (There will not be any continuous singular component.) We will denote the Radon–Nikodym derivative of μ_{ac} by Ψ (i.e., $\mathrm{d}\mu_{\mathrm{ac}}(\xi) = \Psi(\xi)\,\mathrm{d}\xi$, where $\Psi \in L^1(\mathbb{T})$), and call it the spectral density of μ_{ac}. We shall analyze these two components μ_{pp} and μ_{ac} via their Fourier series coefficients. Under certain conditions, we will be able to describe both of these components explicitly and compute the asymptotical behavior of $\mathcal{E}(x,\phi_M)$ as $M \to \infty$.

4. Piecewise affine maps of $\Sigma\Delta$ quantization

In this section, we study the difference equations of $\Sigma\Delta$ modulation as a dynamical system arising from the iteration of certain piecewise affine maps on $\mathbb{R}^m$. It easily follows from the equations in (2.1) that

$$u_j[n] = \sum_{i=1}^{j} u_i[n-1] + \big(x[n] - q[n]\big), \quad 1 \leqslant j \leqslant m, \tag{4.1}$$

or in short,

$$\mathbf{u}[n] = \mathbf{L}\mathbf{u}[n-1] + \big(x[n] - q[n]\big)\mathbf{1}, \tag{4.2}$$

where the matrix $\mathbf{L}$ and the vector $\mathbf{1}$ were defined in (1.5). Using the definition of $q[n]$ in (2.1), we introduce a one-parameter family of maps $\{\mathcal{M}_x\}_{x\in I}$ on $\mathbb{R}^m$ defined by

$$\mathcal{M}_x(\mathbf{v}) := \mathbf{L}\mathbf{v} + \big(x - Q(x,\mathbf{v})\big)\mathbf{1}. \tag{4.3}$$

Hence, the evolution of the state vector $\mathbf{u}[n]$ is given by

$$\mathbf{u}[n] = \mathcal{M}_{x[n]}\big(\mathbf{u}[n-1]\big). \tag{4.4}$$

According to the formulation presented in the introduction, the elements of the partition Π_x are then given by $\Omega_{x,i} = \{\mathbf{v} \in \mathbb{R}^m\colon\ Q(x,\mathbf{v}) = d_i\}$, and the expression (4.3) is equivalent to (1.5). For the rest of the paper, we shall assume that $x[n] = x$ is a constant sequence so that

$$\mathbf{u}[n] = \mathcal{M}_x^n\big(\mathbf{u}[0]\big), \tag{4.5}$$

and

$$q[n] = Q\big(x, \mathcal{M}_x^{n-1}\big(\mathbf{u}[0]\big)\big). \tag{4.6}$$

A variety of choices for the quantizer Q have been introduced in the practice of $\Sigma\Delta$ modulation. Most of these are designed with circuit implementation in mind, and therefore necessitate simple arithmetic operations, such as linear combinations and simple thresholding. A canonical example would be

$$Q_0(x,\mathbf{v}) = \lfloor \alpha_0 x + \alpha_1 v_1 + \cdots + \alpha_m v_m + \beta_0 \rfloor + \beta_1, \tag{4.7}$$

where the coefficients α_i and β_i are specific to each scheme. We will call these rules "linear", referring to the fact that the sets $\Omega_{x,i}$ are separated by translated hyperplanes in $\mathbb{R}^m$. There has also been recent research on more general quantization rules and their benefits [9,15,16].

Typically, an electrical circuit cannot handle arbitrarily large amplitudes, and clips off quantities that are beyond certain values. This is called *overloading*. In this case, the effective mapping Q is given by

$$Q(x,\mathbf{v}) = \begin{cases} Q_0(x,\mathbf{v}) & \text{if } Q_0(x,\mathbf{v}) \in \{d_1,\dots,d_K\}, \\ d_1 & \text{if } Q_0(x,\mathbf{v}) < d_1, \\ d_K & \text{if } Q_0(x,\mathbf{v}) > d_K. \end{cases} \tag{4.8}$$

For the rest of the paper, we assume that the d_i form a subset of an arithmetic progression of spacing 1, such as the case for the rule (4.7). Since we can always subtract a fixed constant from x and the d_i, we also assume, without loss of generality, that the d_i are simply integers. We shall be most interested in one-bit quantization rules, i.e., rules for which $\mathrm{Ran}(Q) = \{d_1, d_2\}$. Let us mention that one-bit $\Sigma\Delta$ modulators are usually overloaded by their nature.

Let us emphasize once again that the quantization rule is crucial in the stability of the system. For a given x, we call a $\Sigma\Delta$ scheme defined by the quantization rule $Q(x,\cdot)$ *orbit stable*, or simply *stable*, if for every initial condition $\mathbf{u}[0]$ in an open set, the forward trajectory under the map $\mathcal{M}_x$ is bounded in $\mathbb{R}^m$, and *positively stable*, if there exists a bounded set $\Gamma_0 \subset \mathbb{R}^m$ with non-empty interior that is positively invariant under $\mathcal{M}_x$, i.e., $\mathcal{M}_x(\Gamma_0) \subset \Gamma_0$. These two notions are closely related. Clearly, positive stability implies stability. On the other hand, in a stable scheme, if the forward trajectories of points in an open set are bounded with a uniform bound, then this would also imply the existence of a positively invariant bounded set. In practice, it is also desirable that stability holds uniformly in x. However, we shall not need this kind of uniformity in this paper.

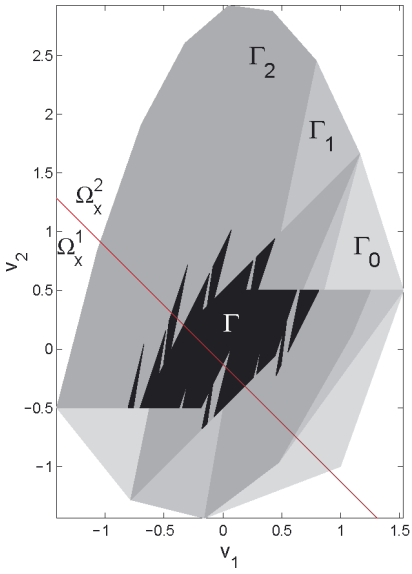


Fig. 2. The decreasing family of nested sets $\Gamma_k = \mathcal{M}_x^k(\Gamma_0)$ indicated by decreasing brightness. The limit set Γ is invariant (see Theorem 5.1).

In Fig. 2, we depict a positively invariant set Γ_0 under the map $\mathcal{M}_x$ which is defined by a one-bit linear rule in $\mathbb{R}^2$. The set Γ_0 was found by a computer algorithm. In general, constructing positively invariant sets for these maps is a non-trivial task [24,27]. Despite the presence of a vast collection of $\Sigma\Delta$ schemes that are used in hardware, only a small set of them are proved to be stable. Most of the engineering practice relies on extensive numerical simulation.

In Fig. 2, we also show in decreasing brightness the forward iterates of Γ_0 given by $\Gamma_k = \mathcal{M}_x^k(\Gamma_0)$. (In this picture, each set Γ_k is the union of the region in which the label "Γ_k" is placed and all the other regions that are shaded in darker colors.) These sets converge to a limit set Γ, or the *attractor*, which is shaded in black. These invariant sets are the topic of discussion of next section.

To avoid heavy and awkward notation, we shall drop the real parameter x from our notation except when we need it for a specific purpose or for emphasis. It must be understood, however, that unless noted otherwise, all objects that are derived from these dynamical systems generally depend on x.

5. Stability implies tiling invariant sets

In this section we prove a crucial property of the dynamics involved in positively stable $\Sigma\Delta$ schemes. This is called the *tiling property* and refers to the fact that there exist trapping invariant sets that are unions of a finite collection of disjoint tiles in $\mathbb{R}^m$. Here a tile, or a $\mathbb{Z}^m$-tile, means any subset S of $\mathbb{R}^m$ with the property that $\{S + \mathbf{k}\}_{\mathbf{k}\in\mathbb{Z}^m}$ is a partition of $\mathbb{R}^m$.

Later in the paper, this property will lead us to an exact spectral analysis of the mean square error when the multiplicity of tiling is one.

We consider a slightly more general class of piecewise affine maps $\mathcal{M} := \mathcal{M}_x$ on $\mathbb{R}^m$, which are defined by

$$\mathcal{M}(\mathbf{v}) = \mathcal{A}_{x,i}(\mathbf{v}) := \mathbf{L}\mathbf{v} + x\mathbf{1} + \mathbf{d}_i \quad \text{if } \mathbf{v} \in \Omega_{x,i}, \tag{5.1}$$

where $\mathbf{L}$ is the lower triangular matrix of all 1's, and $\{\Omega_{x,i}\}_{i=1}^K$ is a finite Lebesgue measurable partition of $\mathbb{R}^m$, and $\mathbf{d}_i \in \mathbb{Z}^m$ for all $i = 1, \ldots, K$. When $\mathbf{d}_i = -d_i\mathbf{1}$, these maps are the same as those that arise from $\Sigma\Delta$ quantization.

Theorem 5.1 [25]. *Assume that there exists a bounded set $\Gamma_0 \subset \mathbb{R}^m$ that is positively invariant under $\mathcal{M}$, i.e., $\mathcal{M}(\Gamma_0) \subset \Gamma_0$. Then, the set $\Gamma \subset \Gamma_0$ defined by*

$$\Gamma := \bigcap_{k \geqslant 0} \mathcal{M}^k(\Gamma_0) \tag{5.2}$$

satisfies the following properties:

(a) $\mathcal{M}(\Gamma) = \Gamma$,
(b) *if Γ_0 contains a tile, then so does Γ.*

Proof. This was previously proved in [25]. For completeness of the discussion, we include the proof here.

(a) Clearly, $\mathcal{M}(\Gamma) \subset \Gamma \subset \Gamma_0$ since Γ_0 is positively invariant. We need to show that $\Gamma \subset \mathcal{M}(\Gamma)$. Let $\mathbf{v} \in \Gamma$ be an arbitrary point. Define $\Gamma_k := \mathcal{M}^k(\Gamma_0)$, $k \geqslant 0$. The sets Γ_k form a decreasing sequence, and so is the case for the sets $F_k := \mathcal{M}^{-1}(\mathbf{v}) \cap \Gamma_k$. Note that $\mathcal{M}^{-1}(\mathbf{v})$ is always finite since there are only finitely many $\mathcal{A}_{x,i}$'s in the definition of $\mathcal{M}$, each of which is 1–1. (F_k would be finite even if there were infinitely many sets $\Omega_{x,i}$ because inverse images under $\mathcal{M}$ have to differ by points in $\mathbb{Z}^m$ and only finitely many of them can be present in Γ_k.) On the other hand $\mathbf{v} \in \Gamma_{k+1} = \mathcal{M}(\Gamma_k)$, therefore $\mathbf{v}$ has an inverse image in Γ_k, i.e., F_k is non-empty. Since F_k form a decreasing sequence of non-empty finite sets, it follows that $\mathcal{M}^{-1}(\mathbf{v}) \cap \Gamma = \bigcap_{k\geqslant 0} F_k \neq \emptyset$, i.e., $\mathbf{v} \in \mathcal{M}(\Gamma)$. Hence $\Gamma \subset \mathcal{M}(\Gamma)$.

(b) Let Γ_0 contain a tile G_0, and define $G_k = \mathcal{M}^k(G_0)$. Each G_k is a tile. To see this, note that for any given i, $\mathcal{A}_{x,i}$ maps tiles to tiles, and for all $\mathbf{v} \in \mathbb{R}^m$, $\mathcal{M}(\mathbf{v}) - \mathcal{A}_{x,i}(\mathbf{v}) \in \mathbb{Z}^m$ so that $\mathcal{M}$ maps tiles to tiles as well. For an arbitrary point $\mathbf{w} \in \mathbb{R}^m$, define the decreasing sequence of sets $H_k = (\mathbb{Z}^m + \mathbf{w}) \cap \Gamma_k$. Because Γ_0 is bounded, each H_k is finite. On the other hand, $\Gamma_k \supset G_k$ implies that each Γ_k contains a tile, yielding $H_k \neq \emptyset$. Hence $(\mathbb{Z}^m + \mathbf{w}) \cap \Gamma = \bigcap_{k\geqslant 0} H_k \neq \emptyset$. Since $\mathbf{w}$ is arbitrary, this means that Γ contains a tile. □

In what follows, measurable means Lebesgue measurable, and $\mathfrak{m}(S)$ denotes the Lebesgue measure of a set S.

Theorem 5.2. *Under the condition of Theorem* 5.1, *assume moreover that x is irrational and that Γ_0 is measurable and $\mathfrak{m}(\Gamma_0) \neq 0$. Then, the set Γ defined in* (5.2) *differs from the*

union of a finite and non-empty collection of disjoint $\mathbb{Z}^m$-tiles at most by a set of measure zero.

Proof. Clearly, Γ is measurable since $\mathcal{M}$ is piecewise affine. Let us show that Lebesgue measure on Γ is invariant under $\mathcal{M}$. From now on, we identify $\mathcal{M}$ with its restriction on Γ. From Theorem 5.1, $\mathcal{M}(\Gamma) = \Gamma$ which implies $\mathcal{M}^{-1}(\Gamma) = \Gamma$ as well. Let us define A to be the set of points in Γ with more than one pre-image. A is measurable, simply because

$$A = \bigcup_{i \neq j} \mathcal{M}(\Gamma \cap \Omega_i) \cap \mathcal{M}(\Gamma \cap \Omega_j).$$

We claim that $\mathfrak{m}(A) = 0$. The definition of $\mathcal{M}$ implies that $\mathcal{M}$ preserves the measure of sets on which it is 1–1. Since $\mathcal{M}$ is 1–1 on $\mathcal{M}^{-1}(\Gamma \backslash A)$, we have $\mathfrak{m}(\Gamma \backslash A) = \mathfrak{m}(\mathcal{M}^{-1}(\Gamma \backslash A))$. On the other hand, since each point in A has at least 2 pre-images, we have $2\mathfrak{m}(A) \leqslant \mathfrak{m}(\mathcal{M}^{-1}(A))$. This implies

$$2\mathfrak{m}(A) \leqslant \mathfrak{m}\big(\mathcal{M}^{-1}(A)\big) = \mathfrak{m}\big(\mathcal{M}^{-1}(\Gamma)\big) - \mathfrak{m}\big(\mathcal{M}^{-1}(\Gamma \backslash A)\big) = \mathfrak{m}(\Gamma) - \mathfrak{m}(\Gamma \backslash A) = \mathfrak{m}(A).$$

Therefore $\mathfrak{m}(A) = \mathfrak{m}(\mathcal{M}^{-1}(A)) = 0$. Hence, for any measurable subset B of Γ, the disjoint union $B = (B \cap A) \cup (B \backslash A)$ yields

$$\mathfrak{m}\big(\mathcal{M}^{-1}(B)\big) = \mathfrak{m}\big(\mathcal{M}^{-1}(B \cap A)\big) + \mathfrak{m}\big(\mathcal{M}^{-1}(B \backslash A)\big) = \mathfrak{m}(B \backslash A) = \mathfrak{m}(B),$$

i.e., $\mathcal{M}$ preserves Lebesgue measure on Γ.

Let $\pi : \Gamma \to \mathbb{T}^m$ be the projection defined by $\pi(\mathbf{v}) = \langle \mathbf{v} \rangle$. Here we identify $[0,1)^m$ with $\mathbb{T}^m$. Let ν be the transformation of the measure $\mathfrak{m}|_\Gamma$ on $\mathbb{T}^m$ under the projection π, which is defined on the Lebesgue measurable subsets of $\mathbb{T}^m$ by $\nu(B) = \mathfrak{m}(\pi^{-1}(B))$. Let $\mathcal{L} = \mathcal{L}_x$ be the generalized skew translation on $\mathbb{T}^m$ defined by

$$\mathcal{L}\mathbf{v} := \mathbf{L}\mathbf{v} + x\mathbf{1} \pmod 1. \tag{5.3}$$

Note that $\pi\mathcal{M} = \mathcal{L}\pi$. Hence, for any measurable $B \subset \mathbb{T}^m$, we have

$$\nu\big(\mathcal{L}^{-1}(B)\big) = \mathfrak{m}\big(\pi^{-1}\mathcal{L}^{-1}(B)\big) = \mathfrak{m}\big(\mathcal{M}^{-1}\pi^{-1}(B)\big) = \mathfrak{m}\big(\pi^{-1}(B)\big) = \nu(B),$$

i.e., ν is invariant under $\mathcal{L}$.

At this point, we note that when x is irrational, $\mathcal{L}$ is *uniquely ergodic*, i.e., there is a unique normalized non-trivial measure invariant under $\mathcal{L}$, which, in this case, is the Lebesgue measure. (See, for example, [7], [23, p. 17] for $m = 2$, and [18, p. 159] for general m.[3]) Hence, $\nu = c\mathfrak{m}$ for some $c \geqslant 0$; this includes the possibility of the trivial invariant measure $\nu \equiv 0$.

[3] Here, unique ergodicity is stated for the map $(v_1, \ldots, v_m) \mapsto (v_1 + x, v_2 + v_1, \ldots, v_m + v_{m-1})$, which is easily shown to be isomorphic to $\mathcal{L}$.

For each $j = 0, 1, \ldots$, define

$$T_j = \{\mathbf{v} \in \mathbb{T}^m\colon \operatorname{card}(\pi^{-1}(\mathbf{v})) = j\}.$$

$\{T_j\}_{j\geqslant 0}$ is a finite measurable partition of $\mathbb{T}^m$. The finiteness is due to the fact that Γ is a bounded set and measurability is simply due to the relation

$$T_j = \left\{\mathbf{v} \in \mathbb{R}^m\colon \sum_{\mathbf{k}\in\mathbb{Z}^m} \chi_{\Gamma+\mathbf{k}}(\mathbf{v}) = j\right\}.$$

Note that

$$c\mathfrak{m}(T_j) = \nu(T_j) = \mathfrak{m}(\pi^{-1}(T_j)) = j\mathfrak{m}(T_j).$$

This shows that there cannot exist two such sets T_i and T_j both with non-zero measure. Hence, there exists a (unique) j, namely, $j = c$, such that $\mathfrak{m}(\mathbb{T}^m\backslash T_j) = 0$. This implies that Γ is the union of j copies of $\mathbb{T}^m$, possibly with the exception of a set of zero measure.

Let us now show that $j \geqslant 1$. Consider π on the whole $\mathbb{R}^m$ with the same definition. Note that the relation $\pi\mathcal{M} = \mathcal{L}\pi$ continues to hold. Let $\Sigma_0 := \pi(\Gamma_0) \subset \mathbb{T}^m$. Since Γ_0 is positively invariant, we find that $\mathcal{L}(\Sigma_0) = \pi\mathcal{M}(\Gamma_0) \subset \pi(\Gamma_0) = \Sigma_0$. Since $\mathcal{L}$ is 1–1, we have $\mathcal{L}^{-1}(\Sigma_0) \supset \Sigma_0$. Hence, $\mathcal{L}^{-1}(\Sigma_0) \triangle \Sigma_0 = \mathcal{L}^{-1}(\Sigma_0)\backslash\Sigma_0 = \mathcal{L}^{-1}(\Sigma_0\backslash\mathcal{L}(\Sigma_0))$. This implies, since $\mathcal{L}$ is measure-preserving,

$$\begin{aligned}\mathfrak{m}(\mathcal{L}^{-1}(\Sigma_0) \triangle \Sigma_0) &= \mathfrak{m}(\mathcal{L}^{-1}(\Sigma_0\backslash\mathcal{L}(\Sigma_0))) = \mathfrak{m}(\Sigma_0\backslash\mathcal{L}(\Sigma_0))\\ &= \mathfrak{m}(\Sigma_0) - \mathfrak{m}(\mathcal{L}(\Sigma_0)) = 0.\end{aligned}$$

Ergodicity of $\mathcal{L}$ implies that $\mathfrak{m}(\Sigma_0)$ is 0 or 1. The first case is not possible, since each point in Σ_0 has at most finitely many inverse images under π^{-1} and this would violate $\mathfrak{m}(\Gamma_0) > 0$. Therefore $\mathfrak{m}(\Sigma_0) = 1$, implying that $\mathfrak{m}(\Gamma_0) \geqslant 1$. Define $\Sigma_k := \pi(\Gamma_k)$. We then have $\mathcal{L}(\Sigma_k) = \mathcal{L}\pi(\Gamma_k) = \pi\mathcal{M}(\Gamma_k) = \pi(\Gamma_{k+1}) = \Sigma_{k+1}$, which implies that $\mathfrak{m}(\Sigma_k) = 1$ and $\mathfrak{m}(\Gamma_k) \geqslant 1$ for all $k \geqslant 0$. Hence $\mathfrak{m}(\Gamma) = \lim_{k\to\infty} \mathfrak{m}(\Gamma_k) \geqslant 1$. □

When x is irrational, Theorem 5.2 improves Theorem 5.1(b) in two respects. First, the outcome is that Γ not only contains a tile, but in fact is *composed* of disjoint tiles, up to a set of measure zero. Second, it suffices to check that Γ_0 has positive measure, instead of the stronger (though equivalent) requirement that Γ_0 contain a tile. On the other hand, Theorem 5.1(b) is still interesting due to its purely algebraic nature. It can be used to test if Γ contains an exact tile (i.e., $\pi(\Gamma) = \mathbb{T}^m$), and it remains valid even when x is rational.

Let us also note as an application of Theorem 5.2 that whenever a positively invariant set Γ_0 of $\mathcal{M}_x$ (for irrational x) can be found with $0 < \mathfrak{m}(\Gamma_0) < 2$, the invariant set Γ is a single tile.

In Fig. 3, we show an illustration of an invariant set which is composed of two tiles. In this example, the $\Sigma\Delta$ scheme is 1-bit 2nd order and the partition is determined by a cubic curve.

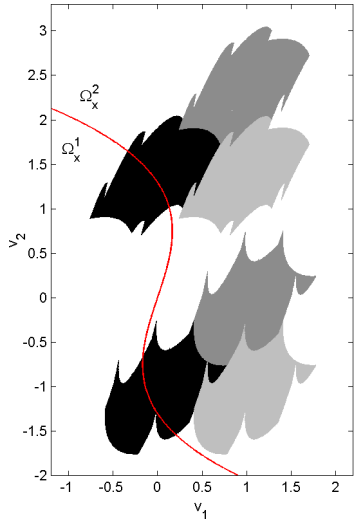


Fig. 3. Represented in black is the invariant set Γ of a 1-bit 2nd order scheme whose partition is determined by the cubic curve shown in the figure. The copies in gray are the translated versions of Γ by $(1,0)$ and $(1,1)$, respectively. In this example, each connected component of Γ is also invariant.

6. The single-tile case and its consequence

Since the initial experimental discovery of the tiling property in [12,15], we have observed that the invariant sets Γ resulting from stable second order $\Sigma\Delta$ schemes that are used in practice systematically appear to be single tiles. We show in Fig. 4 experimental examples of Γ on some of these second order schemes. In Fig. 5, we show the set Γ in three cases where an explicit analytical derivation has been possible [15]. (In these particular cases, Γ is actually proven to be an exact tile.) A fundamental question is to characterize maps $\mathcal{M}_x$ which yield a single invariant tile. For the rest of this paper, we will simply assume that this condition is realized. As will be seen, the analysis of the dynamics becomes particularly simplified. Furthermore, a whole new set of tools for error analysis becomes available.

A tile Γ intrinsically generates a unique projection $\langle\cdot\rangle_\Gamma : \mathbb{R}^m \to \Gamma$ such that $\mathbf{v} - \langle\mathbf{v}\rangle_\Gamma \in \mathbb{Z}^m$ for all $\mathbf{v} \in \mathbb{R}^m$. The restriction of this $\mathbb{Z}^m$-periodic projection to the unit cube $[0,1)^m$ (which we identify with $\mathbb{T}^m$) is a measure preserving bijection (note that the inverse of $\langle\cdot\rangle_\Gamma : \mathbb{T}^m \to \Gamma$ is the map π that was defined in the proof of Theorem 5.2). When Γ is invariant under $\mathcal{M}$, the map $\langle\cdot\rangle_\Gamma : \mathbb{T}^m \to \Gamma$ establishes an isomorphism between $\mathcal{M}$ on Γ and the affine transformation $\mathcal{L} := \mathcal{L}_x$ on $\mathbb{T}^m$ defined by (5.3). Indeed, the definition of $\mathcal{L}$ easily yields $\mathcal{L}(\mathbf{v}) - \mathcal{M}(\langle\mathbf{v}\rangle_\Gamma) \in \mathbb{Z}^m$. Hence,

$$\big\langle \mathcal{L}(\mathbf{v}) \big\rangle_\Gamma = \mathcal{M}\big(\langle\mathbf{v}\rangle_\Gamma\big),$$

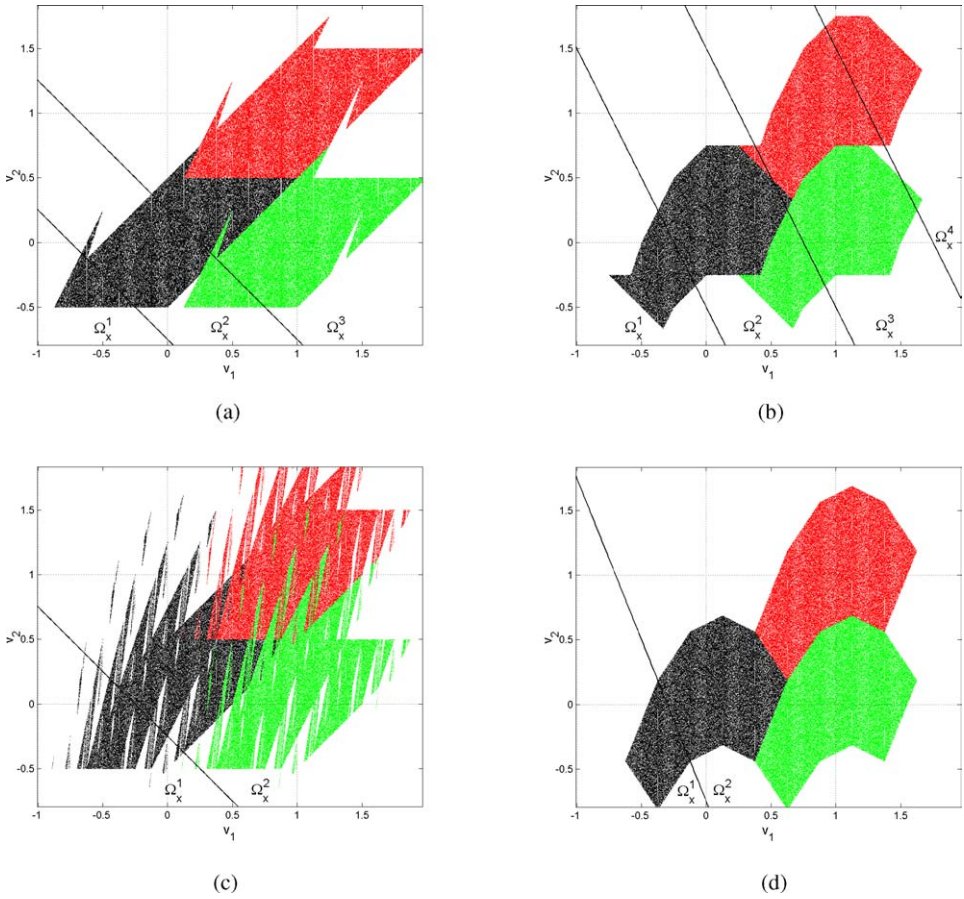


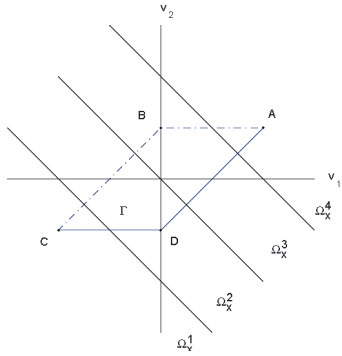
Fig. 4. Representation in black of several consecutive state points $\mathbf{u}[n]$ of various second order $\Sigma\Delta$ modulators with the irrational input $x \approx 3/4$. The copies in other colors are the translated versions of the state points by $(1, 0)$ and $(1, 1)$, respectively.

or in other words, the following diagram commutes:

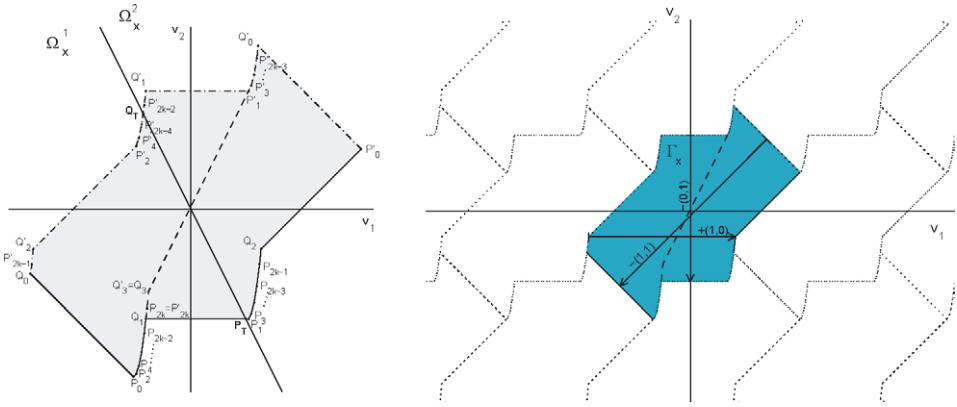
$$\begin{array}{ccc} \mathbb{T}^m & \xrightarrow{\mathcal{L}} & \mathbb{T}^m \\ {\scriptstyle \langle\cdot\rangle_\Gamma}\downarrow & & \downarrow{\scriptstyle \langle\cdot\rangle_\Gamma} \\ \Gamma & \xrightarrow[\mathcal{M}]{} & \Gamma \end{array}$$

The first important consequence of single invariant tiles is that it reduces the dynamical system $\mathcal{M}$ to the much simpler $\mathcal{L}$ whose n-fold composition can be computed explicitly. It follows that if $\mathbf{u}[0] \in \Gamma$, then

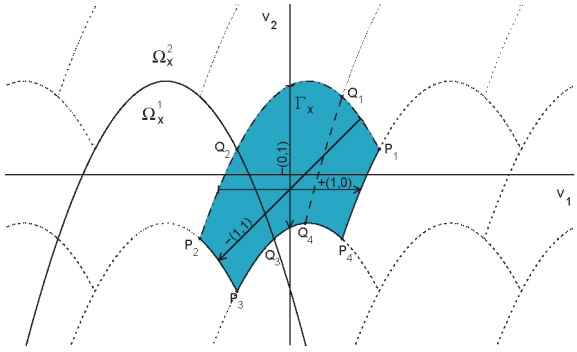
$$\mathbf{u}[n] = \mathcal{M}^n\big(\mathbf{u}[0]\big) = \big\langle \mathcal{L}^n\big(\mathbf{u}[0]\big)\big\rangle_\Gamma = \big\langle \mathbf{L}^n\mathbf{u}[0] + x\mathbf{s}[n]\big\rangle_\Gamma, \tag{6.1}$$



(a) 2-bit "linear" with $(d_1, d_2, d_3, d_4) = (-1, 0, 1, 2)$ $(x = 0.5)$.



(b) 1-bit "linear" with $(d_1, d_2) = (0, 1)$ $(x \approx 0.52)$.



(c) 1-bit "quadratic" with $(d_1, d_2) = (0, 1)$ $(x = 0.74)$.

Fig. 5. Three families of quantization rules for which the tiling property was proven in [15] with parametric explicit expressions for the corresponding invariant sets.

where $\mathbf{s}[n] := \mathbf{s}_m[n] := (s_1[n], \ldots, s_m[n])^\top$ is defined by

$$\mathbf{s}[n] = \left(\sum_{k=0}^{n-1} \mathbf{L}^k \right) \mathbf{1}. \tag{6.2}$$

It is an easy computation to show that $s_j[n] = \binom{j+n-1}{j}$.

The second important consequence is that for irrational values of x the map $\mathcal{M}_x$ on Γ inherits the ergodicity of $\mathcal{L}_x$ via the isomorphism generated by $\langle\cdot\rangle_\Gamma$. Since $\langle\cdot\rangle_\Gamma : \mathbb{T}^m \to \Gamma$ preserves Lebesgue measure, $\mathcal{M}_x$ is then ergodic with respect to the restriction of Lebesgue measure on Γ. Hence the Birkhoff Ergodic Theorem yields

Proposition 6.1. *Let x be an irrational number and Γ be a Lebesgue measurable $\mathbb{Z}^m$-tile (up to a set of measure zero) that is invariant under $\mathcal{M}$. Then for any function $F \in L^1(\Gamma)$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} F\big(\mathbf{u}[n]\big) = \int_\Gamma F(\mathbf{v})\,d\mathbf{v} = \int_{\mathbb{T}^m} F\big(\langle\mathbf{v}\rangle_\Gamma\big)\,d\mathbf{v} \tag{6.3}$$

for almost every initial condition $\mathbf{u}[0] \in \Gamma$.

This formula will be the fundamental computational tool for the analysis of the autocorrelation sequence ρ_u. For the remainder of this paper, we shall assume that we are working with quantization rules for which the invariant sets are composed of single tiles. This will save us from repetition in the assumptions of our results. However, it will also be important to know certain geometric features of these invariant tiles. We will state these explicitly when needed.

7. Analysis of the autocorrelation sequence ρ_u

Let $\mathsf{P}(\mathbf{v}) = v_m$ be the projection of a vector $\mathbf{v} \in \mathbb{R}^m$ onto its mth coordinate. If we define the function

$$F_k(\mathbf{v}) := \mathsf{P}(\mathbf{v})\mathsf{P}\big(\mathcal{M}^k(\mathbf{v})\big), \tag{7.1}$$

then it follows that

$$u_m[n]u_m[n+k] = \mathsf{P}\big(\mathbf{u}[n]\big)\mathsf{P}\big(\mathcal{M}^k\big(\mathbf{u}[n]\big)\big) = F_k\big(\mathbf{u}[n]\big),$$

and therefore Proposition 6.1 gives an expression for the value of $\rho_u[k]$:

$$\rho_u[k] = \int_\Gamma F_k(\mathbf{v})\,d\mathbf{v} = \int_{\mathbb{T}^m} F_k\big(\langle\mathbf{v}\rangle_\Gamma\big)\,d\mathbf{v}. \tag{7.2}$$

A direct evaluation of $\rho_u[k]$ in either of these forms is not easy, because the k-fold iterated map $\mathcal{M}^k$ as well as the invariant set Γ are implicitly-defined and complex objects. The problem can be somewhat simplified via the conjugate map $\mathcal{L}^k$. Indeed, one has

$$F_k \circ \langle\cdot\rangle_\Gamma = \big(\mathsf{P}\circ\langle\cdot\rangle_\Gamma\big)\big(\mathsf{P}\circ\mathcal{M}^k\circ\langle\cdot\rangle_\Gamma\big) = \big(\mathsf{P}\circ\langle\cdot\rangle_\Gamma\big)\big(\mathsf{P}\circ\langle\cdot\rangle_\Gamma\circ\mathcal{L}^k\big),$$

so that if we define

$$G_\Gamma = \mathsf{P} \circ \langle \cdot \rangle_\Gamma,$$

then via (6.1), we obtain the formula

$$\rho_u[k] = \int_{\mathbb{T}^m} G_\Gamma(\mathbf{v}) G_\Gamma\big(\mathcal{L}^k(\mathbf{v})\big)\,\mathrm{d}\mathbf{v} = \int_{\mathbb{T}^m} G_\Gamma(\mathbf{v}) G_\Gamma\big(\mathbf{L}^k\mathbf{v} + x\mathbf{s}[k]\big)\,\mathrm{d}\mathbf{v}, \tag{7.3}$$

which now only depends on Γ.

As it is standard in the spectral theory of dynamical systems (see, e.g., [23]), let $\mathcal{U} := \mathcal{U}_\mathcal{L}$ be the unitary operator on $L^2(\mathbb{T}^m)$ defined by $(\mathcal{U}f)(\mathbf{v}) = f(\mathcal{L}(\mathbf{v}))$. Then (7.3) reduces to

$$\rho_u[k] = \big(G_\Gamma, \mathcal{U}^k G_\Gamma\big)_{L^2(\mathbb{T}^m)}. \tag{7.4}$$

For any $f \in L^2(\mathbb{T}^m)$, the inner products $(f, \mathcal{U}^k f)_{L^2(\mathbb{T}^m)}$, $k \in \mathbb{Z}$, define a positive-definite sequence so that there exists a unique non-negative measure ν_f on $\mathbb{T}$ with Fourier coefficients

$$\hat{\nu}_f[k] = \big(f, \mathcal{U}^k f\big)_{L^2(\mathbb{T}^m)} \tag{7.5}$$

for all $k \in \mathbb{Z}$. Note that when $f = G_\Gamma$, it follows from (7.4) that the corresponding measure $\nu_{G_\Gamma} = \mu$, where μ is the spectral measure that was mentioned in Section 3, with $\hat{\mu} = \rho_u$.

7.1. Decomposition of the mixed spectrum: general results

We shall separate the autocorrelation sequence ρ_u into two additive components that result from two different types of spectral behavior. Using the spectral theorem for unitary operators, we decompose $L^2(\mathbb{T}^m)$ into two $\mathcal{U}$-invariant, orthogonal subspaces as $L^2(\mathbb{T}^m) = \mathcal{H}_{\mathrm{pp}} \oplus \mathcal{H}_{\mathrm{c}}$, where

$$\mathcal{H}_{\mathrm{pp}} = \big\{ f \in L^2\big(\mathbb{T}^m\big) \colon\ \nu_f \text{ is purely atomic} \big\},$$

which is also equal to the closed linear span of the set of all eigenfunctions of $\mathcal{U}$, and

$$\mathcal{H}_{\mathrm{c}} = \mathcal{H}_{\mathrm{pp}}^\perp = \big\{ f \in L^2\big(\mathbb{T}^m\big) \colon\ \nu_f \text{ is non-atomic (continuous)} \big\}.$$

In the particular case of the transformation $\mathcal{L}$, it turns out that every spectrum on $\mathcal{H}_{\mathrm{pp}}^\perp$ is absolutely continuous (see Appendix A). Therefore we denote $\mathcal{H}_{\mathrm{c}}$ by $\mathcal{H}_{\mathrm{ac}}$. Any $f \in L^2(\mathbb{T}^m)$ can now be uniquely decomposed as

$$f = f_{\mathrm{pp}} + f_{\mathrm{ac}},$$

where $f_{\mathrm{pp}} \in \mathcal{H}_{\mathrm{pp}}$ and $f_{\mathrm{ac}} \in \mathcal{H}_{\mathrm{ac}}$. For $\mathcal{L}$, it is known (and as we also show in Appendix A), that

$$\mathcal{H}_{\mathrm{pp}} = \big\{ f \in L^2\big(\mathbb{T}^m\big) \colon\ f(\mathbf{v}) \text{ only depends on } v_1 \big\},$$

and the orthogonal projection of f onto $\mathcal{H}_{\mathrm{pp}}$ is given by

$$f_{\mathrm{pp}}(\mathbf{v}) = \int_{\mathbb{T}^{m-1}} f\big(v_1, \mathbf{v}'\big)\, \mathrm{d}\mathbf{v}'. \tag{7.6}$$

In order to avoid double subscripts (e.g., when $f = G_\Gamma$), we will use the alternative notation $\mathring{f} := f_{\mathrm{pp}}$ and $\breve{f} := f_{\mathrm{ac}}$ whenever it will be convenient.

We now consider the decomposition

$$G_\Gamma = \mathring{G}_\Gamma + \breve{G}_\Gamma. \tag{7.7}$$

Because of orthogonality and $\mathcal{U}$-invariance of $\mathcal{H}_{\mathrm{pp}}$ and $\mathcal{H}_{\mathrm{ac}}$, (7.4) implies that

$$\rho_u[k] = \big(\mathring{G}_\Gamma, \mathcal{U}^k \mathring{G}_\Gamma\big)_{L^2(\mathbb{T}^m)} + \big(\breve{G}_\Gamma, \mathcal{U}^k \breve{G}_\Gamma\big)_{L^2(\mathbb{T}^m)}, \tag{7.8}$$

providing the decomposition

$$\rho_u = \mathring{\rho}_u + \breve{\rho}_u.$$

Here, using formula (6.1) and the fact that functions in the subspace $\mathcal{H}_{\mathrm{pp}}$ depend only on the first variable, we obtain

$$\mathring{\rho}_u[k] = \big(\mathring{G}_\Gamma, \mathcal{U}^k \mathring{G}_\Gamma\big)_{L^2(\mathbb{T}^m)} = \int_{\mathbb{T}} \mathring{G}_\Gamma(v_1) \mathring{G}_\Gamma(v_1 + kx)\, \mathrm{d}v_1 \tag{7.9}$$

and

$$\breve{\rho}_u[k] = \big(\breve{G}_\Gamma, \mathcal{U}^k \breve{G}_\Gamma\big)_{L^2(\mathbb{T}^m)} = \int_{\mathbb{T}^m} \breve{G}_\Gamma(\mathbf{v}) \breve{G}_\Gamma\big(\mathbf{L}^k \mathbf{v} + x\mathbf{s}[k]\big)\, \mathrm{d}\mathbf{v}. \tag{7.10}$$

This decomposition provides the Fourier coefficients of the pure-point μ_{pp} and the absolutely continuous μ_{ac} components of the spectral measure, respectively. It also yields an explicit simple formula for μ_{pp} in terms of the Fourier coefficients of $\mathring{G}_\Gamma$. We have

Theorem 7.1.

$$\mu_{\mathrm{pp}} = \sum_{n \in \mathbb{Z}} \Big| \widehat{\mathring{G}_\Gamma}[n] \Big|^2 \delta_{nx}, \tag{7.11}$$

where δ_a denotes the unit Dirac mass at $a \in \mathbb{T}$.

Proof. Let ν denote the measure given on the right-hand side of (7.11). It suffices to verify that $\hat{\nu}[k] = \mathring{\rho}_u[k]$ for all $k \in \mathbb{Z}$. We find by direct evaluation that

$$\hat{\nu}[k] = \sum_{n\in\mathbb{Z}} \left|\widehat{\mathring{G}_\Gamma}[n]\right|^2 e^{-2\pi i k n x} = \int_{\mathbb{T}} \mathring{G}_\Gamma(v)\mathring{G}_\Gamma(v+kx)\,\mathrm{d}v = \mathring{\rho}_u[k];$$

hence the result follows. □

Note. It is easy to see that this result holds for any function $f \in L^2(\mathbb{T}^m)$ in the sense that

$$(\nu_f)_{\mathrm{pp}} = \sum_{n\in\mathbb{Z}} \left|\widehat{\mathring{f}}[n]\right|^2 \delta_{nx}. \tag{7.12}$$

On the other hand, the computation of μ_{ac} is not easy. Since absolute continuity of μ_{ac} results in an integrable density Ψ, where $\mathrm{d}\mu_{\mathrm{ac}}(\xi) = \Psi(\xi)\,\mathrm{d}\xi$, the Riemann–Lebesgue lemma implies that the Fourier coefficients $\breve{\rho}_u[k] \to 0$ as $|k| \to \infty$. However, the rate of decay is determined by further properties of this measure, which turn out to be intrinsically related to the geometry of Γ.

7.2. Properties of $\breve{\rho}_u$ for the class of v_m-connected invariant tiles

In this section, we derive explicit formulae for $\breve{\rho}_u[k]$ when the invariant tile Γ has a certain type of geometric regularity. For a given tile Γ for $\mathbb{R}^m$, let us define

$$\Lambda_\Gamma := \bigcup_{\mathbf{k}'\in\mathbb{Z}^{m-1}} \Gamma + (\mathbf{k}', 0), \tag{7.13}$$

and for any $\mathbf{v}' \in \mathbb{R}^{m-1}$,

$$\Lambda_\Gamma(\mathbf{v}') := \mathsf{P}\big(\Lambda_\Gamma \cap \{\mathbf{v}'\}\times\mathbb{R}\big) = \big\{v_m \in \mathbb{R}\colon\ (\mathbf{v}', v_m) \in \Lambda_\Gamma\big\}. \tag{7.14}$$

Proposition 7.2. *For each $\mathbf{v}' \in \mathbb{R}^{m-1}$, the set $\Lambda_\Gamma(\mathbf{v}')$ is a tile in $\mathbb{R}$ with respect to $\mathbb{Z}$-translations, and*

$$G_\Gamma(\mathbf{v}', v_m) = \langle v_m\rangle_{\Lambda_\Gamma(\mathbf{v}')}. \tag{7.15}$$

Proof. Since Γ is a tile, the collection of sets $\{\Lambda_\Gamma + (\mathbf{0}, k)\colon\ k \in \mathbb{Z}\}$ forms a partition of $\mathbb{R}^m$. Therefore for any $\mathbf{v}' \in \mathbb{R}^{m-1}$, the v_m-section of this collection given by $\{\Lambda_\Gamma(\mathbf{v}') + k\colon\ k \in \mathbb{Z}\}$, is a partition of $\mathbb{R}$. This shows that $\Lambda_\Gamma(\mathbf{v}')$ is a tile. For the second part of the claim, let $\mathbf{v} = (\mathbf{v}', v_m)$. The definition of P immediately yields

$$\big(\mathbf{v}', G_\Gamma(\mathbf{v})\big) = \big(\mathbf{v}', \mathsf{P}\big(\langle\mathbf{v}\rangle_\Gamma\big)\big) = \langle\mathbf{v}\rangle_\Gamma + (\mathbf{k}', 0)$$

for some $\mathbf{k}' \in \mathbb{Z}^{m-1}$. This says that $(\mathbf{v}', G_\Gamma(\mathbf{v})) \in \Lambda_\Gamma$ and therefore $G_\Gamma(\mathbf{v}) \in \Lambda_\Gamma(\mathbf{v}')$. The result follows since $G_\Gamma(\mathbf{v}', v_m) - v_m \in \mathbb{Z}$. □

Definition 7.3. We say that a tile $\Gamma \subset \mathbb{R}^m$ is v_m-connected if for each $\mathbf{v}' \in \mathbb{R}^{m-1}$, the one-dimensional tile $\Lambda_\Gamma(\mathbf{v}')$ is a connected set, i.e., a unit-length interval. In this case, we denote by $\lambda_\Gamma(\mathbf{v}')$ the midpoint of $\Lambda_\Gamma(\mathbf{v}')$ and call λ_Γ the midpoint function.

In Fig. 6, we display examples of the function Λ_Γ for various schemes. The tiles in (a), (c) and (d) are v_2-connected whereas the tile in (b) is not. Note that v_m-connectedness of a tile is different from its v_m-sections being connected.

Let us use the shorthand notation $\langle\alpha\rangle_0 := \langle\alpha\rangle_{[-\frac{1}{2},\frac{1}{2})} = \langle\alpha+\frac{1}{2}\rangle - \frac{1}{2}$. For a v_m-connected tile, we have the following simple observation:

Corollary 7.4. *If the tile Γ is v_m-connected, then for any $\mathbf{v}' \in \mathbb{R}^{m-1}$*

$$G_\Gamma(\mathbf{v}', v_m) = \langle v_m - \lambda_\Gamma(\mathbf{v}')\rangle_0 + \lambda_\Gamma(\mathbf{v}') \tag{7.16}$$

and

$$\mathring{G}_\Gamma = \mathring{\lambda}_\Gamma. \tag{7.17}$$

Proof. If Γ is v_m-connected, then $\Lambda_\Gamma(\mathbf{v}') = [\lambda_\Gamma(\mathbf{v}') - \frac{1}{2}, \lambda_\Gamma(\mathbf{v}') + \frac{1}{2})$. Now, (7.16) follows from Proposition 7.2 and the identity

$$\langle\beta\rangle_{[\alpha-\frac{1}{2},\alpha+\frac{1}{2})} = \langle\beta-\alpha\rangle_0 + \alpha$$

which holds for any α and β. Next, (7.17) is a simple consequence of the fact that the first term in (7.16) integrates to zero over v_m. □

Before we state the following proposition, let $\mathbf{J} := \mathbf{J}_m$ be the "backward identity" permutation matrix defined by $(\mathbf{J}_m)_{ij} = \delta_{i,m+1-j}$, for $1 \leqslant i, j \leqslant m$. Note that the matrix $\mathbf{L}^k := \mathbf{L}_m^k$ can now be decomposed as

$$\mathbf{L}_m^k = \begin{pmatrix} \mathbf{L}_{m-1}^k & \boxed{\mathbf{0}} \\ \boxed{\mathbf{s}_{m-1}^\top[k]\mathbf{J}_{m-1}} & 1 \end{pmatrix}, \tag{7.18}$$

which easily follows from (6.2); note that $\mathbf{s}[k] = \mathbf{L}\,\mathbf{s}[k-1] + \mathbf{1}$ with $\mathbf{s}[0] = \mathbf{0}$.

Proposition 7.5. *Let the invariant tile Γ be v_m-connected. Define for each $k \in \mathbb{Z}$, and $\mathbf{v}' \in \mathbb{R}^{m-1}$,*

$$g_k(\mathbf{v}') = \mathbf{s}_{m-1}^\top[k]\mathbf{J}_{m-1}\mathbf{v}' + x s_m[k] - \lambda_\Gamma\big(\mathbf{L}_{m-1}^k\mathbf{v}' + x\mathbf{s}_{m-1}[k]\big) + \lambda_\Gamma(\mathbf{v}').$$

Then

$$\breve{\rho}_u[k] = \int_{\mathbb{T}^{m-1}} A_{\langle\cdot\rangle_0}\big(g_k(\mathbf{v}')\big)\,\mathrm{d}\mathbf{v}' + \big(\breve{\lambda}_\Gamma, \mathcal{U}^k\breve{\lambda}_\Gamma\big)_{L^2(\mathbb{T}^{m-1})}. \tag{7.19}$$

In particular, if $m = 2$ or if $\mathsf{P}(\Gamma)$ is an interval of unit length, then the second term drops.

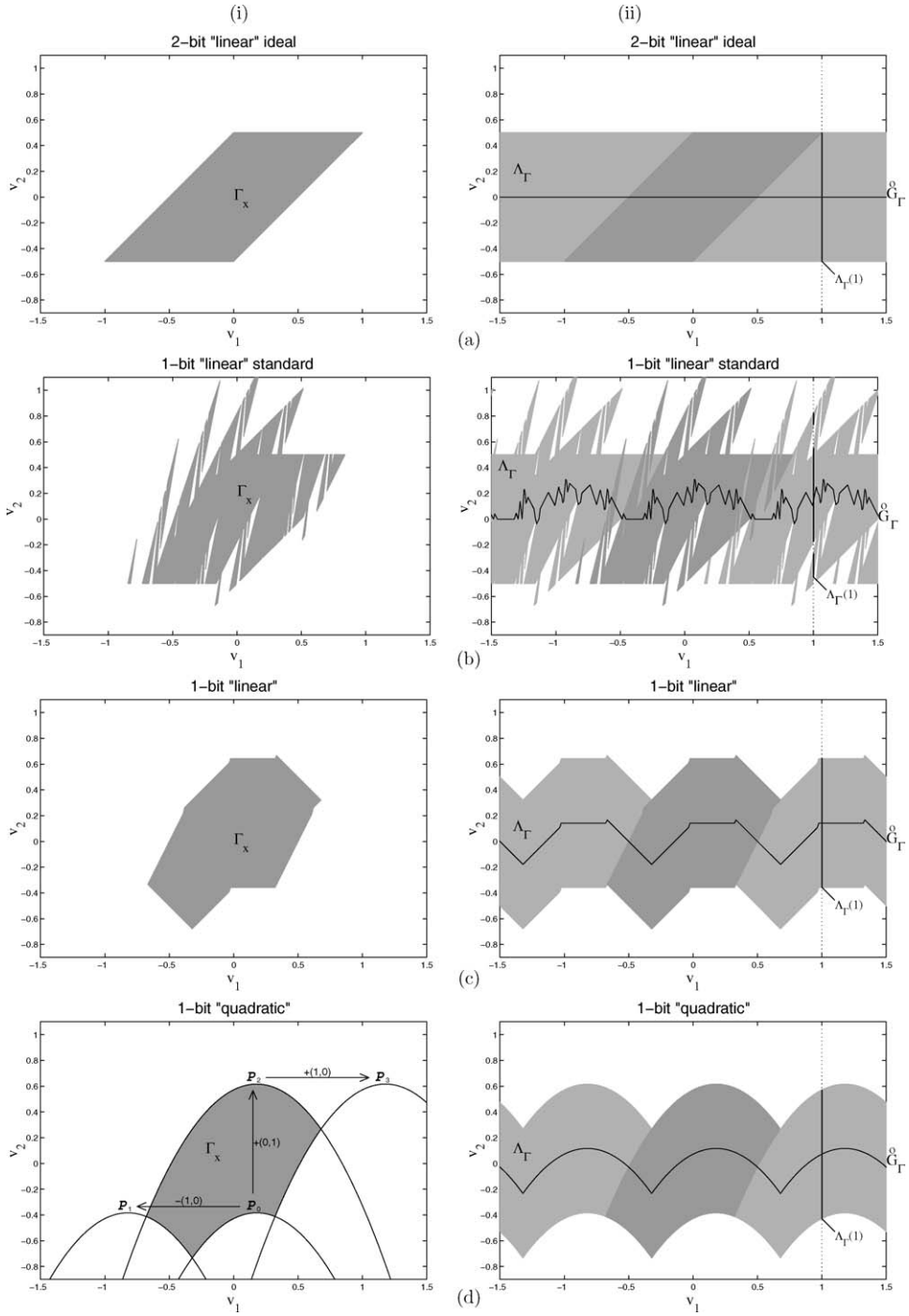


Fig. 6. Invariant tiles of various second order modulators: (i) invariant tile Γ_x, (ii) corresponding set Λ_Γ.

Proof. We employ Corollary 7.4 for the evaluation of $G_\Gamma(\mathbf{v})$ and $G_\Gamma(\mathbf{L}^k\mathbf{v}+x\mathbf{s}[k])$. Let us again write $\mathbf{v}=(\mathbf{v}', v_m)$. Note first that from (7.18) we obtain

$$\mathbf{L}^k\mathbf{v}+x\mathbf{s}[k]=\big(\mathbf{L}_{m-1}^k\mathbf{v}'+x\mathbf{s}_{m-1}[k],\, v_m+\mathbf{s}_{m-1}^\top[k]\mathbf{J}_{m-1}\mathbf{v}'+xs_m[k]\big).$$

It follows that

$$\begin{aligned}
&\int_{\mathbb{T}} G_\Gamma(\mathbf{v})G_\Gamma\big(\mathbf{L}^k\mathbf{v}+x\mathbf{s}[k]\big)\,\mathrm{d}v_m\\
&\quad=\int_{\mathbb{T}}\big\langle v_m-\lambda_\Gamma(\mathbf{v}')\big\rangle_0\big\langle v_m+\mathbf{s}_{m-1}^\top[k]\mathbf{J}_{m-1}\mathbf{v}'+xs_m[k]-\lambda_\Gamma\big(\mathbf{L}_{m-1}^k\mathbf{v}'+x\mathbf{s}_{m-1}[k]\big)\big\rangle_0\,\mathrm{d}v_m\\
&\qquad+\lambda_\Gamma(\mathbf{v}')\lambda_\Gamma\big(\mathbf{L}_{m-1}^k\mathbf{v}'+x\mathbf{s}_{m-1}[k]\big),
\end{aligned}$$

where the cross terms have dropped because $\int_{\mathbb{T}}\langle v_m+\varphi(\mathbf{v}')\rangle_0\,\mathrm{d}v_m=0$ for any function φ. The first term above is equal to $A_{\langle\cdot\rangle_0}(g_k(\mathbf{v}'))$, whereas if the second term is integrated over $\mathbb{T}^{m-1}$ we find $(\lambda_\Gamma,\mathcal{U}^k\lambda_\Gamma)_{L^2(\mathbb{T}^{m-1})}$. The result follows since $\mathring{\lambda}_\Gamma=\mathring{G}_\Gamma$.

If $m=2$, then moreover $\lambda_\Gamma=\mathring{G}_\Gamma$, so that we have $\breve{\lambda}_\Gamma=0$. If $\mathsf{P}(\Gamma)$ is an interval of unit length, then it is necessarily the case that $\Lambda_\Gamma=\mathbb{R}^{m-1}\times\mathsf{P}(\Gamma)$. In this case, λ_Γ is a constant function so that $\mathring{\lambda}_\Gamma=\lambda_\Gamma$ and hence $\breve{\lambda}_\Gamma=0$. Hence the second term drops in both cases. □

7.3. *Special case when* $\mathsf{P}(\Gamma)=[-\frac{1}{2},\frac{1}{2})$

There is a class of quantization rules [8,11,17], for which $u_m[n]\in[-\frac{1}{2},\frac{1}{2})$ for all n (for all x), so that the invariant tile Γ satisfies $\mathsf{P}(\Gamma)=[-\frac{1}{2},\frac{1}{2})$. These are the "ideal" rules that were mentioned in Section 1, and represent essentially the simplest possible situation. It turns out that the spectral measure μ is quite different in its nature for $m=1$ and $m\geqslant 2$.

For $m=1$, by definition we have $\mathring{G}_\Gamma=G_\Gamma=\langle\cdot\rangle_0$. Hence μ is pure-point, and Theorem 7.1 yields

$$\mu=\sum_{n\neq 0}\frac{1}{4\pi^2n^2}\delta_{nx}.$$

For $m\geqslant 2$, we simply note that $\lambda_\Gamma\equiv 0$, so that $G_\Gamma(\mathbf{v})=\langle v_m\rangle_0$ by (7.16). The fact that $\int_{\mathbb{T}}\langle v_m\rangle_0\,\mathrm{d}v_m=0$ implies $\mathring{G}_\Gamma\equiv 0$. Hence $\mu_{\mathrm{pp}}=0$, i.e., μ is absolutely continuous. In addition, Proposition 7.5 yields

$$\rho_u[k]=\int_{\mathbb{T}^{m-1}}A_{\langle\cdot\rangle_0}\big(\mathbf{s}_{m-1}^\top[k]\mathbf{J}_{m-1}\mathbf{v}'+xs_m[k]\big)\,\mathrm{d}\mathbf{v}'.$$

For $k = 0$, the argument of the integrand is identically zero, so we obtain $\rho_u[0] = A_{\langle\cdot\rangle_0}(0) = \frac{1}{12}$. On the other hand, for all $k \neq 0$, we find that $\rho_u[k] = 0$ since the integrand is of the form $A_{\langle\cdot\rangle_0}(kv_{m-1} + \alpha)$ which integrates to zero over the variable v_{m-1}. Therefore,

$$\rho_u[k] = \begin{cases} \frac{1}{12} & \text{if } k = 0, \\ 0 & \text{if } k \neq 0, \end{cases}$$

and consequently μ is flat, and equal to $\frac{1}{12}$ times Lebesgue measure on $\mathbb{T}$, and the spectral density Ψ is the constant function $\Psi(\xi) \equiv \frac{1}{12}$.

These two results were previously obtained, in the case $m = 1$ in [11], and in the case $m \geqslant 2$ in [8,17].

8. Analysis of the mean square error

We are interested in the asymptotical behavior of $\mathcal{E}(x, \phi)$ for a given $\Sigma\Delta$ modulation scheme of order m as the support of ϕ increases and its Fourier transform Φ localizes around zero frequency. There will be two standard choices for Φ:

(1) The ideal low-pass filter given by

$$\Phi_M^{\mathrm{id}}(\xi) := \chi_{[-\frac{1}{M}, \frac{1}{M}]}(\xi).$$

(2) The "sinc" family[4] given by

$$\Phi_{M,p}^{\mathrm{sinc}}(\xi) := \left(\frac{1}{M} \sum_{n=0}^{M-1} e^{2\pi i n \xi} \right)^p = \left(\frac{\sin(\pi M \xi)}{M \sin(\pi \xi)} e^{i\pi(M-1)\xi} \right)^p.$$

Note that $\Phi_{M,p}^{\mathrm{sinc}}$ has Fourier coefficients given by

$$\phi_{M,p}^{\mathrm{sinc}}[n] := r_M^{(p)}[n] := (\underbrace{r_M * r_M * \cdots * r_M}_{p \text{ times}})[n],$$

where r_M denotes the rectangular sequence

$$r_M[n] = \begin{cases} 1/M, & 0 \leqslant n < M, \\ 0, & \text{otherwise.} \end{cases}$$

It is a standard fact that $\phi_{M,p}^{\mathrm{sinc}}$ is a discrete B-spline of degree $p - 1$.

[4] The terminology for this filter is derived from its continuous analog which is related to $\mathrm{sinc}(x) := \sin(x)/x$.

For any filter ϕ, we decompose the mean square error $\mathcal{E}(x,\phi)$ as

$$\mathcal{E}(x,\phi) = \mathcal{E}_{\mathrm{pp}}(x,\phi) + \mathcal{E}_{\mathrm{ac}}(x,\phi)$$

which correspond to the additive contributions of μ_{pp} and μ_{ac}, respectively, in the formula (3.5). Note that both terms are non-negative. Note also that for the above two filter choices we have

$$\left|\Phi_M^{\mathrm{id}}(\xi)\right| \lesssim_p \left|\Phi_{M/2,p}^{\mathrm{sinc}}(\xi)\right|, \quad \forall \xi \in \mathbb{T}; \tag{8.1}$$

hence it suffices to prove lower bounds for the ideal low-pass filter and upper bounds for sinc filters.

8.1. *The pure-point contribution* $\mathcal{E}_{\mathrm{pp}}(x,\phi)$

Our first formula follows directly from plugging the expression for μ_{pp} given by Theorem 7.1 in (3.5):

$$\mathcal{E}_{\mathrm{pp}}(x,\phi) = \sum_{n\in\mathbb{Z}} \left|2\sin(\pi n x)\right|^{2m} \left|\Phi(nx)\right|^2 \left|\widehat{\mathring{G}_\Gamma}[n]\right|^2. \tag{8.2}$$

Before we carry out our analysis of this expression, let us recall some elementary facts about Diophantine approximation. For $\alpha \in \mathbb{R}$, let $\|\alpha\|$ denote the distance of α to the nearest integer, that is $\|\alpha\| := \min(\langle\alpha\rangle, \langle -\alpha\rangle)$. We say that α is (Diophantine) of type η if η is the infimum of all numbers σ for which

$$\|n\alpha\| \gtrsim_{\sigma,\alpha} |n|^{-\sigma}, \quad \forall n \in \mathbb{Z}\backslash\{0\}.$$

Almost every real number (in the sense of Lebesgue measure) is of type 1, the smallest attainable type.

The following theorem shows that for almost every x, if the function G_Γ has a sufficiently regular projection $\mathring{G}_\Gamma$, then the pure-point part of the mean square error after filtering with $\phi_{M,m+1}^{\mathrm{sinc}}$ decays faster than M^{-2m-1}.

Theorem 8.1. *Let x be Diophantine of type η. If for some $\beta > \eta/2$ the invariant tile $\Gamma = \Gamma_x$ of an mth order $\Sigma\Delta$ modulator with input x satisfies*

$$\left|\widehat{\mathring{G}_\Gamma}[n]\right| \lesssim |n|^{-\beta}$$

for all $n \in \mathbb{Z}\backslash\{0\}$, then

$$\mathcal{E}_{\mathrm{pp}}\big(x, \phi_{M,m+1}^{\mathrm{sinc}}\big) \lesssim_{x,m,\alpha,\beta} M^{-2m-1-\alpha} \tag{8.3}$$

for all M, where α is any number satisfying $0 \leqslant \alpha < \min(1, \frac{2\beta}{\eta} - 1)$.

Proof. Formula (8.2) reads

$$\mathcal{E}_{\mathrm{pp}}\big(x,\phi^{\mathrm{sinc}}_{M,m+1}\big)=\frac{2^{2m}}{M^{2m+2}}\sum_{n\in\mathbb{Z}\setminus\{0\}}\frac{\sin^{2m+2}(\pi Mnx)}{\sin^2(\pi nx)}\Big|\widehat{\mathring{G}_\Gamma}[n]\Big|^2. \tag{8.4}$$

Since $|\sin(\pi\theta)|\asymp\|\theta\|$, $1-\alpha\leqslant 2m+2$, and $\|Mnx\|\leqslant\min(1,M\|nx\|)$, we have

$$\frac{\sin^{2m+2}(\pi Mnx)}{\sin^2(\pi nx)}\lesssim_m\frac{\|Mnx\|^{2m+2}}{\|nx\|^2}\leqslant\frac{\|Mnx\|^{1-\alpha}}{\|nx\|^2}\leqslant\frac{M^{1-\alpha}}{\|nx\|^{1+\alpha}}.$$

Given the decay of $|\widehat{\mathring{G}_\Gamma}[n]|$, we then obtain

$$\mathcal{E}_{\mathrm{pp}}\big(x,\phi^{\mathrm{sinc}}_{M,m+1}\big)\lesssim_m\frac{1}{M^{2m+1+\alpha}}\sum_{n=1}^{\infty}\frac{1}{n^{2\beta}\|nx\|^{1+\alpha}}. \tag{8.5}$$

Since $1+\alpha\geqslant 1$, it suffices to show the convergence of the sum

$$\sum_{n=1}^{\infty}\frac{1}{n^{2\beta/(1+\alpha)}\|nx\|}.$$

Let $\lambda:=2\beta/(1+\alpha)$. Since $1+\alpha<2\beta/\eta$, we have $\lambda>\eta$. Now, summation by parts shows that

$$\sum_{n=1}^{\infty}\frac{1}{n^{\lambda}\|nx\|}\lesssim_\lambda\sum_{n=1}^{\infty}\frac{1}{n^{\lambda+1}}\left(\sum_{k=1}^{n}\frac{1}{\|kx\|}\right), \tag{8.6}$$

and furthermore it is well known [21, Ex. 3.11] that

$$\sum_{k=1}^{n}\frac{1}{\|kx\|}\lesssim_{x,\sigma}n^{\sigma}$$

for any $\sigma>\eta$. Choosing σ such that $\lambda>\sigma>\eta$, we obtain the convergence of (8.6) with a sum depending on x and λ. Combining this result with (8.5) the result follows. □

Note. The Diophantine condition on x can be removed if $\mathring{G}_\Gamma$ is a trigonometric polynomial. In this case, (8.4) reduces to a finite sum, and therefore it is always convergent.

On the other hand, our next result shows that if $\mathring{G}_\Gamma$ does not have enough regularity in a certain sense as specified in the following theorem, then the result of Theorem 8.1 is the best one can get in the following sense: there is a dense set of exceptional values of x for which the exponent of the error decay rate is never better than $2m$, even for the ideal low pass filter.

Theorem 8.2. *Given a $\Sigma\Delta$ modulator of order m, let ϕ_M, $M = 1, 2, \ldots$, be a sequence of averaging filters such that $|\Phi_M(\xi)| \geqslant c_1$ on the interval $|\xi| \leqslant c_2/M$, where c_1 and c_2 are positive constants that do not depend on M. There exists a dense set E of irrational numbers with the following property: for any $x \in E$, if we can find positive constants β_x and C_x such that the invariant tile $\Gamma = \Gamma_x$ satisfies*

$$\left|\widehat{\mathring{G}_\Gamma}[n]\right| \geqslant C_x|n|^{-\beta_x}$$

for all but finitely many $n \in \mathbb{Z}$, then for all $\delta > 0$,

$$\limsup_{M\to\infty} \mathcal{E}_{\mathrm{pp}}(x, \phi_M) M^{2m+\delta} = \infty. \tag{8.7}$$

Proof. It suffices to find, for any open interval J, a point $x \in J$ with the property (8.7) for all $\delta > 0$. Given an open interval J, let $x_0 \in J$ be a dyadic rational. Let $l = \max(b, d)$ for the minimum b and d such that $b! - 1$ is an upper bound for the length of the binary expansion of x_0 and $2^{-d!+1}$ is a lower bound for the distance of x_0 to the boundary of J. Set

$$x = x_0 + \sum_{k\geqslant l} 2^{-k!}.$$

Then clearly $x \in J$. It is also a standard fact that x is irrational, in fact x is a Liouville number.

Note that for $q \geqslant l$, we have

$$\left\langle 2^{q!}x\right\rangle = \sum_{k=q+1}^{\infty} 2^{-k!+q!} = 2^{-q\cdot q!} + \sum_{k=q+2}^{\infty} 2^{-k!+q!}.$$

For $q = 1, 2, \ldots$, let $n_q = 2^{q!}$ and $M_q = 2^{q\cdot q!-r}$ where r is a fixed integer such that $2^{-r+1} \leqslant c_2$. Note that $M_q = 2^{-r}n_q^q$ is integer valued for all sufficiently large values of q. We also have

$$\frac{2^{-r}}{M_q} = 2^{-q\cdot q!} < \|n_q x\| < 2^{-q\cdot q!+1} \leqslant \frac{c_2}{M_q}. \tag{8.8}$$

The right side of this chain of inequalities implies $|\Phi_{M_q}(n_q x)| \geqslant c_1$ by our assumption on $\{\phi_M\}$. On the other hand, the left side implies $|2\sin(\pi n_q x)| \geqslant 4\|n_q x\| > 2^{-r+2}/M_q$. Therefore

$$\begin{aligned}\mathcal{E}_{\mathrm{pp}}(x, \phi_{M_q}) &\geqslant \left|2\sin(\pi n_q x)\right|^{2m}\left|\Phi_{M_q}(n_q x)\right|^2\left|\widehat{\mathring{G}_\Gamma}[n_q]\right|^2 \\ &\geqslant C_x^2 c_1^2 2^{2m(-r+2)} 2^{-2r\beta_x/q} M_q^{-2m} M_q^{-2\beta_x/q}.\end{aligned} \tag{8.9}$$

The result of the theorem follows by letting $q \to \infty$ and therefore exhibiting the subsequence $M_q \to \infty$ for which (8.7) holds for any $\delta > 0$. □

8.2. The absolutely continuous contribution $\mathcal{E}_{\mathrm{ac}}(x,\phi)$

Let us denote by Ψ the Radon–Nikodym derivative of the absolutely continuous spectral measure μ_{ac}, i.e., $\mathrm{d}\mu_{\mathrm{ac}} = \Psi(\xi)\,\mathrm{d}\xi$. A priori, we only know that $\Psi \in L^1(\mathbb{T})$, which is somewhat weak for what we would like to achieve in terms of understanding the decay of $\mathcal{E}_{\mathrm{ac}}(x,\phi)$. Our first theorem concerns the decay rate of

$$\mathcal{E}_{\mathrm{ac}}\bigl(x,\phi^{\mathrm{sinc}}_{M,m+1}\bigr) = \int_{\mathbb{T}} \bigl|2\sin(\pi\xi)\bigr|^{2m}\bigl|\Phi^{\mathrm{sinc}}_{M,m+1}(\xi)\bigr|^2\Psi(\xi)\,\mathrm{d}\xi$$

when it is known that Ψ belongs to a smaller L^p space.

Theorem 8.3. *If the measure μ_{ac} has density $\Psi \in L^p(\mathbb{T})$ for some $1 \leqslant p \leqslant \infty$, then*

$$\mathcal{E}_{\mathrm{ac}}\bigl(x,\phi^{\mathrm{sinc}}_{M,m+1}\bigr) \lesssim_{m,p} \|\Psi\|_{L^p(\mathbb{T})} M^{-2m-1+1/p}. \tag{8.10}$$

Proof. Let p' be the dual index of p, i.e., $1/p + 1/p' = 1$. Note that

$$\bigl|2\sin(\pi\xi)\bigr|^{2m}\bigl|\Phi^{\mathrm{sinc}}_{M,m+1}(\xi)\bigr|^2 = \bigl|2\sin(\pi M\xi)\bigr|^{2m}\bigl|\Phi^{\mathrm{sinc}}_{M,2}(\xi)\bigr|M^{-2m} \tag{8.11}$$

$$\lesssim_m \bigl|\Phi^{\mathrm{sinc}}_{M,2}(\xi)\bigr|M^{-2m}, \tag{8.12}$$

so that Hölder's inequality yields

$$\mathcal{E}_{\mathrm{ac}}\bigl(x,\phi^{\mathrm{sinc}}_{M,m+1}\bigr) \lesssim_m \|\Psi\|_{L^p(\mathbb{T})}\bigl\|\Phi^{\mathrm{sinc}}_{M,2}\bigr\|_{L^{p'}(\mathbb{T})} M^{-2m}.$$

Furthermore, the simple bound $|\Phi^{\mathrm{sinc}}_{M,1}(\xi)| \leqslant \min(1, (2M|\xi|)^{-1})$ implies

$$\bigl\|\Phi^{\mathrm{sinc}}_{M,2}\bigr\|_{L^{p'}(\mathbb{T})} \lesssim_{p'} M^{-1/p'}, \tag{8.13}$$

hence the theorem follows. □

On the other hand, it turns out that if Ψ is continuous at 0, then one can calculate the exact asymptotics of $\mathcal{E}_{\mathrm{ac}}(x,\phi^{\mathrm{sinc}}_{M,m+1})$ without additional assumptions.

Theorem 8.4. *If the spectral density Ψ is continuous at* 0*, then*

$$\mathcal{E}_{\mathrm{ac}}\bigl(x,\phi^{\mathrm{sinc}}_{M,m+1}\bigr) = \binom{2m}{m}\Psi(0)M^{-2m-1} + o\bigl(M^{-2m-1}\bigr). \tag{8.14}$$

Proof. The proof has two parts. First part is the easy calculation

$$\int_{\mathbb{T}} \bigl|2\sin(\pi\xi)\bigr|^{2m}\bigl|\Phi^{\mathrm{sinc}}_{M,m+1}(\xi)\bigr|^2\,\mathrm{d}\xi = \binom{2m}{m}M^{-2m-1}. \tag{8.15}$$

To see this, note that (8.11) and the definition of $\Phi^{\mathrm{sinc}}_{M,m+1}$ imply

$$\left|2\sin(\pi\xi)\right|^{2m}\left|\Phi^{\mathrm{sinc}}_{M,m+1}(\xi)\right|^2=\left(\frac{e^{i\pi M\xi}-e^{-i\pi M\xi}}{i}\right)^{2m}\sum_{k=0}^{M-1}\sum_{j=0}^{M-1}e^{2\pi i(k-j)\xi}M^{-2m-2}.$$

The right-hand side is the product of two trigonometric polynomials; the first polynomial has frequencies only at integer multiples of $2\pi M$ and the second polynomial has frequencies between $-2\pi(M-1)$ and $2\pi(M-1)$. The zero frequency term of the product is therefore given only by the product of the zero frequency terms of each factor, which is equal to

$$\binom{2m}{m}(-1)^m i^{-2m}\left(\sum_{k=0}^{M-1}1\right)M^{-2m-2}=\binom{2m}{m}M^{-2m-1},$$

hence the result.

The second part of the proof concerns the residual term

$$\left|\int_{\mathbb{T}}\left|2\sin(\pi\xi)\right|^{2m}\left|\Phi^{\mathrm{sinc}}_{M,m+1}(\xi)\right|^2\big(\Psi(\xi)-\Psi(0)\big)\,\mathrm{d}\xi\right|,$$

which is bounded, using (8.11), by

$$2^{2m}M^{-2m}\int_{\mathbb{T}}\Phi^{\mathrm{sinc}}_{M,2}(\xi)\big|\Psi(\xi)-\Psi(0)\big|\,\mathrm{d}\xi=2^{2m}M^{-2m-1}\int_{\mathbb{T}}K_{M-1}(\xi)\big|\Psi(\xi)-\Psi(0)\big|\,\mathrm{d}\xi,$$

where

$$K_{M-1}(\xi)=\frac{1}{M}\left(\frac{\sin(\pi M\xi)}{\sin(\pi\xi)}\right)^2$$

is the Fejér kernel. The limit

$$\lim_{M\to\infty}\int_{\mathbb{T}}K_{M-1}(\xi)\big|\Psi(\xi)-\Psi(0)\big|\,\mathrm{d}\xi$$

is the Cesàro sum of the Fourier series of the function $f(t)=|\Psi(-t)-\Psi(0)|$ evaluated at $t=0$. Since f is continuous at 0, the Cesàro sum converges to $f(0)=0$, and therefore the limit is 0. This concludes the proof. □

Notes. (1) A similar calculation shows that for the ideal filter ϕ^{id}_M, the error has the asymptotics given by

$$\mathcal{E}_{\mathrm{ac}}\big(x,\phi^{\mathrm{id}}_M\big)=\frac{(2\pi)^{2m+1}}{m+1/2}\Psi(0)M^{-2m-1}+O\big(M^{-2m-3}\big)\tag{8.16}$$

again assuming that Ψ is continuous at 0.

(2) The value of $\Psi(0)$ is equal to the sum of its Fourier coefficients $\breve{\rho}_u[k]$.

9. Estimates for second order schemes with v_2-connected invariant tiles

Second order $\Sigma\Delta$ modulators with v_2-connected invariant tiles are interesting because the value of x and the midpoint function $\lambda_\Gamma = \mathring{G}_\Gamma$ completely describe the MSE behavior via the theorems we have stated in the previous sections. In particular, Proposition 7.5 provides us with the formula

$$\begin{aligned}\breve{\rho}_u[k] &= \int_{\mathbb{T}} A_{\langle\cdot\rangle_0}\left(kv_1 + \frac{k(k+1)}{2}x - \lambda_\Gamma(v_1+kx) + \lambda_\Gamma(v_1)\right)\mathrm{d}v_1 \\ &= \int_{\mathbb{T}} A_{\langle\cdot\rangle_0}\left(kv - \lambda_\Gamma\left(v - \frac{x}{2} + k\frac{x}{2}\right) + \lambda_\Gamma\left(v - \frac{x}{2} - k\frac{x}{2}\right)\right)\mathrm{d}v, \qquad (9.1)\end{aligned}$$

where we have used the change of variable $v = v_1 + (k+1)x/2$ to obtain the second representation.

By Riemann–Lebesgue lemma, we already know that $\breve{\rho}_u[k]$ must converge to zero as $|k| \to \infty$ since $\breve{\rho}_u[k] = \hat{\Psi}[k]$, where $\Psi \in L^1(\mathbb{T})$ is the spectral density. However, we would like to quantify the rate of decay in $|k|$ as this would then allow us to draw conclusions about Ψ. Intuitively speaking, it is not hard to see from this formula that the smoother λ_Γ is, the faster $\breve{\rho}_u[k]$ must decay in $|k|$ as $|k| \to \infty$, since $A_{\langle\cdot\rangle_0}$ is a zero mean function on $\mathbb{T}$. Our objective in this section is to study this relation rigorously.

Let $\mathrm{BV}(\mathbb{T})$ denote the space of functions on $\mathbb{T}$ that have bounded variation, where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation semi-norm, and let $\mathrm{A}(\mathbb{T})$ denote the space of functions on $\mathbb{T}$ with absolutely convergent Fourier series with the norm $\|f\|_{\mathrm{A}(\mathbb{T})}$ given by $\sum|\hat{f}[n]|$. We have the following lemma, whose proof is given in the appendix.

Lemma 9.1. *Let $f \in \mathrm{A}(\mathbb{T})$ and φ be two real valued functions on $\mathbb{T}$, where f has zero mean. Consider the integrals*

$$c[k] = \int_{\mathbb{T}} f\big(kv + \varphi(v)\big)\,\mathrm{d}v. \qquad (9.2)$$

The following bounds hold:

(1) *If $\varphi \in \mathrm{BV}(\mathbb{T})$, then for all $k \in \mathbb{Z}\backslash\{0\}$,*

$$\big|c[k]\big| \leqslant \frac{1}{|k|}\|f\|_{\mathrm{A}(\mathbb{T})}\|\varphi\|_{\mathrm{TV}}. \qquad (9.3)$$

(2) *If φ is differentiable almost everywhere and $\varphi' \in \mathrm{BV}(\mathbb{T})$, then for all $k \in \mathbb{Z}\backslash\{0\}$,*

$$\big|c[k]\big| \leqslant \frac{1}{k^2}\left(\frac{1}{\sqrt{12}}\|f\|_{L^2(\mathbb{T})}\big\|\varphi'\big\|_{\mathrm{TV}} + \|f\|_{L^\infty(\mathbb{T})}\big\|\varphi'\big\|^2_{L^2(\mathbb{T})}\right). \tag{9.4}$$

Theorem 9.2. *Let x be given and Γ be the invariant tile corresponding to a second order $\Sigma\Delta$ modulator. Then we have the following*:

(1) *If the midpoint function λ_Γ has bounded variation on $\mathbb{T}$, then*

$$\big|\breve{\rho}_u[k]\big| \leqslant \frac{1}{6|k|}\|\lambda_\Gamma\|_{\mathrm{TV}}. \tag{9.5}$$

Consequently, one has

$$\mathcal{E}_{\mathrm{ac}}\big(x, \phi^{\mathrm{sinc}}_{M,3}\big) \lesssim_{x,\varepsilon} M^{-5+\varepsilon} \tag{9.6}$$

for any $\varepsilon > 0$. If the type η of x is strictly less than 2, then

$$\mathcal{E}_{\mathrm{pp}}\big(x, \phi^{\mathrm{sinc}}_{M,3}\big) \lesssim_{x,\delta} M^{-5-\delta} \tag{9.7}$$

for any $0 \leqslant \delta < (2-\eta)/\eta$.

(2) *If the midpoint function λ_Γ has a derivative that has bounded variation on $\mathbb{T}$, then*

$$\big|\breve{\rho}_u[k]\big| \leqslant \frac{1}{k^2}\left(\frac{1}{12\sqrt{15}}\big\|\lambda'_\Gamma\big\|_{\mathrm{TV}} + \frac{1}{3}\big\|\lambda'_\Gamma\big\|^2_{L^2(\mathbb{T})}\right). \tag{9.8}$$

In particular, the spectral density Ψ is continuous. Consequently, one has

$$\mathcal{E}_{\mathrm{ac}}\big(x, \phi^{\mathrm{sinc}}_{M,3}\big) = 6\Psi(0)M^{-5} + o\big(M^{-5}\big), \tag{9.9}$$

where

$$\Psi(0) \leqslant \frac{1}{12} + \frac{\pi^2}{3}\left(\frac{1}{12\sqrt{15}}\big\|\lambda'_\Gamma\big\|_{\mathrm{TV}} + \frac{1}{3}\big\|\lambda'_\Gamma\big\|^2_{L^2(\mathbb{T})}\right). \tag{9.10}$$

If the type η of x is strictly less than 4, then

$$\mathcal{E}_{\mathrm{pp}}\big(x, \phi^{\mathrm{sinc}}_{M,3}\big) \lesssim_{x,\delta} M^{-5-\delta} \tag{9.11}$$

for any $0 \leqslant \delta < \min(1, (4-\eta)/\eta)$.

Proof. Let

$$f(v) := A_{\langle\cdot\rangle_0}(v) = \sum_{n\neq 0} \frac{1}{4\pi^2 n^2} e^{2\pi i n v}.$$

For each k, define

$$\varphi_k(v) := -\lambda_\Gamma\left(v - \frac{x}{2} + k\frac{x}{2}\right) + \lambda_\Gamma\left(v - \frac{x}{2} - k\frac{x}{2}\right).$$

For these functions, we have the following exact formulas and bounds:

$$\|f\|_{\mathrm{A}(\mathbb{T})} = \frac{1}{12}, \tag{9.12}$$

$$\|f\|_{L^\infty(\mathbb{T})} = \frac{1}{12}, \tag{9.13}$$

$$\|f\|_{L^2(\mathbb{T})} = \frac{1}{12\sqrt{5}}, \tag{9.14}$$

$$\|\varphi_k\|_{\mathrm{TV}} \leqslant 2\|\lambda_\Gamma\|_{\mathrm{TV}}, \tag{9.15}$$

$$\|\varphi_k'\|_{\mathrm{TV}} \leqslant 2\|\lambda_\Gamma'\|_{\mathrm{TV}}, \tag{9.16}$$

$$\|\varphi_k'\|_{L^2(\mathbb{T})}^2 \leqslant 4\|\lambda_\Gamma'\|_{L^2(\mathbb{T})}^2. \tag{9.17}$$

(1) In this case we only know that λ_Γ is of bounded variation.

The decay estimate (9.5) simply follows from the bound (9.3) coupled with (9.12) and (9.15).

Given that the Fourier coefficients $\breve{\rho}_u[k] = \hat{\Psi}[k]$ decay like $1/k$, it follows from Riesz–Thorin interpolation theorem that the spectral density $\Psi \in L^p(\mathbb{T})$ for any $p < \infty$. Therefore Theorem 8.3 implies, with $m = 2$ and $\varepsilon = 1/p$, the bound (9.6).

For the pure-point estimate, we use Theorem 8.1 with $\beta = 1$ and $m = 2$. If we define $\delta = \alpha$, where α is as defined in Theorem 8.1, then the result follows as stated.

(2) In this case we know that λ_Γ has a derivative that is of bounded variation.

The decay estimate (9.8) follows from the bound (9.4) coupled with (9.13), (9.14), (9.16) and (9.17).

Since $\breve{\rho}_u$ is summable, it follows that Ψ is continuous. We therefore apply Theorem 8.4 to compute the exact asymptotics of $\mathcal{E}_{\mathrm{ac}}(x, \phi_{M,3}^{\mathrm{sinc}})$. In this case, the non-negative number $\Psi(0)$ will be bounded by $\sum |\breve{\rho}_u[k]|$. We simply, add up the bounds given by (9.8), including the trivial case $|\breve{\rho}_u[0]| \leqslant \|f\|_{L^\infty(\mathbb{T})}$. This computation yields the bound (9.10).

For the pure-point estimate, we again use Theorem 8.1, but now with $\beta = 2$. We define $\delta = \alpha$, where α is as defined in Theorem 8.1, and note that the condition $\alpha \leqslant 1$ must be imposed, which was automatically satisfied in case (1). Then the result follows as stated. □

10. Further remarks

In this paper, we have covered only a portion of the mathematical problems that concern $\Sigma\Delta$ quantization. We believe that the following currently unresolved problems are interesting both from the dynamical systems standpoint and the engineering perspective.

1. *Which maps $\mathcal{M}$ are stable?* Satisfactory answers of this question would include non-trivial sufficient conditions in terms of the quantization rule Q, or in terms of the partition Π_x and the quantization levels $\{d_i\}$.

2. *Which stable maps $\mathcal{M}$ yield single invariant tiles?* One can include in this the case when Γ is composed of tiles each of which is invariant under $\mathcal{M}$. In principle, each of these invariant tiles would represent a different "mode of operation".

3. *What is an appropriate generalization of our spectral analysis of mean square error when Γ is composed of more than one tile?*

4. *Given the quantization rule, what can be said about the geometric regularity of Γ?* We used two types of geometric information about Γ in deriving our analytical results on the mean square error asymptotics. The first type concerned "shape" (such as v_m-connectedness), and the second concerned "regularity" (such as the decay of Fourier coefficients of $\mathring{G}_\Gamma$). At this stage, the relation between the quantization rule and these two issues is highly unclear, although we have partial understanding in some cases. Even for "linear" rules, there seems to be a wide range of possibilities.

5. *What are the universal principles behind tiling?* Tiling invariant sets are found even when x is rational. In addition, trajectories seem to remain within exact tiles, and not just tiles "up to sets of measure zero".

Acknowledgments

The authors thank Ingrid Daubechies, Ron DeVore, Özgür Yılmaz and Yang Wang for conversations on the topic of $\Sigma\Delta$ quantization, tiling, and related issues.

Appendix A. On the spectral theory of the map $\mathcal{L}$

In this section, we will review some basic facts about the spectral theory of the map $\mathcal{L} = \mathcal{L}_x$ on $\mathbb{T}^m$, where $\mathcal{L}_x\mathbf{v} = \mathbf{L}\mathbf{v} + x\mathbf{1}$, and x is an irrational number. Most of what follows below can be derived or generalized from Anzai's work on ergodic skew product transformation [3].

The eigenfunctions of $\mathcal{U}_\mathcal{L}$

We start by showing that the set of all eigenfunctions of $\mathcal{U} = \mathcal{U}_\mathcal{L}$, where $\mathcal{U}_\mathcal{L} f := f \circ \mathcal{U}_\mathcal{L}$, is precisely given by the collection of complex exponentials f_n, where

$$f_n(\mathbf{v}) = e^{2\pi i n v_1}, \quad n \in \mathbb{Z}.$$

To see this, let $f \in L^2(\mathbb{T}^m)$ be an eigenfunction of $\mathcal{U}$ with eigenvalue λ. Since $\mathcal{U}$ is unitary, $|\lambda| = 1$. Consider the Fourier series expansion of f given by

$$f(\mathbf{v}) = \sum_{\mathbf{n}\in\mathbb{Z}^m} c[\mathbf{n}] e^{2\pi i \mathbf{n}\cdot\mathbf{v}}.$$

Since $f = \frac{1}{\lambda}(\mathcal{U}f)$, we have the relation

$$\sum_{\mathbf{n}\in\mathbb{Z}^m} c[\mathbf{n}]e^{2\pi i\mathbf{n}\cdot\mathbf{v}} = \frac{1}{\lambda}\sum_{\mathbf{n}\in\mathbb{Z}^m} c[\mathbf{n}]e^{2\pi i x\mathbf{n}\cdot\mathbf{1}}e^{2\pi i\mathbf{n}\cdot(\mathbf{L}\mathbf{v})} = \frac{1}{\lambda}\sum_{\mathbf{n}\in\mathbb{Z}^m} c[\mathbf{K}\mathbf{n}]e^{2\pi i x(\mathbf{K}\mathbf{n})\cdot\mathbf{1}}e^{2\pi i\mathbf{n}\cdot\mathbf{v}},$$

where $\mathbf{K} = (\mathbf{L}^{-1})^{\top}$. Comparing the coefficients, we obtain the equality

$$\big|c[\mathbf{n}]\big| = \big|c[\mathbf{K}\mathbf{n}]\big|, \quad \forall \mathbf{n}\in\mathbb{Z}^m.$$

Since $f \in L^2(\mathbb{T}^m)$, we can conclude that $c[\mathbf{n}] = 0$ for any $\mathbf{n}$ that is not preserved under $\mathbf{K}^j$ for some positive integer j, for otherwise we would have the infinite sequence of coefficients $c[\mathbf{n}], c[\mathbf{K}\mathbf{n}], c[\mathbf{K}^2\mathbf{n}], \ldots$ of equal and strictly positive magnitude.

On the other hand, it is a simple exercise to show that the only vectors that satisfy $\mathbf{n} = \mathbf{K}^j\mathbf{n}$ for some power $j \geqslant 1$ are those of the form $\mathbf{n} = (n_1, 0, \ldots, 0)$. Hence, any eigenfunction of $\mathcal{U}$ depends only on the first variable v_1. On the first coordinate v_1 of $\mathbf{v}$, $\mathcal{L}$ reduces to the irrational rotation by x, and hence as it is well known, these eigenfunctions are nothing but the given complex exponentials $\{f_n\}_{n\in\mathbb{Z}}$. These eigenfunctions span the subspace $\mathcal{H}_{\mathrm{pp}}$ of $L^2(\mathbb{T}^m)$.

The absolutely continuous spectrum

We shall next show that continuous part of the spectrum is in fact absolutely continuous. This is in fact a consequence of the fact that there exists an orthonormal basis $\{\psi_{j,k}\colon\ j\in\mathbb{Z},\ k\in\mathbb{N}\}$ of $\mathcal{H}_{\mathrm{pp}}^{\perp}$ with the property that $\mathcal{U}\psi_{j,k} = \psi_{j+1,k}$ for all j and k. (I.e., $\mathcal{L}$ has *countable Lebesgue spectrum* on $\mathcal{H}_{\mathrm{pp}}^{\perp}$.) First we will construct such a basis, and then we shall prove the statement on the absolute continuity.

From the discussion above on the eigenfunctions of $\mathcal{U}$, we know that the complex exponentials

$$f_{\mathbf{n}}(\mathbf{v}) = e^{2\pi i\mathbf{n}\cdot\mathbf{v}}, \quad \mathbf{n}\in\mathbb{Z}^m\backslash\big(\mathbb{Z}\times\{0\}^{m-1}\big),$$

form an orthonormal complete set in $\mathcal{H}_{\mathrm{pp}}^{\perp}$. Note also that

$$\mathcal{U}f_{\mathbf{n}} = e^{2\pi i x\mathbf{n}\cdot\mathbf{1}}f_{\mathbf{L}^{\top}\mathbf{n}}.$$

Therefore we consider the orbit of each $\mathbf{n}\in\mathbb{Z}^m$ under $\mathbf{L}^{\top}$, given by

$$\mathcal{O}(\mathbf{n}) = \big\{\big(\mathbf{L}^{\top}\big)^j\mathbf{n}\big\}_{j\in\mathbb{Z}}.$$

It is easy to see that each $\mathbf{n}\in\mathbb{Z}\times\{0\}^{m-1}$ is a fixed point of $\mathbf{L}^{\top}$ and every other $\mathbf{n}$ is such that the orbit is an infinite sequence of distinct points in $\mathbb{Z}^m\backslash(\mathbb{Z}\times\{0\}^{m-1})$. Since $\mathbf{L}^{\top}$ is

invertible, orbits do not intersect. Hence we can divide $\mathbb{Z}^m\backslash(\mathbb{Z}\times\{0\}^{m-1})$ into equivalence classes of orbits $\mathcal{O}(\mathbf{n}_k)$, $k\in\mathbb{N}$, and define

$$\psi_{0,k}=f_{\mathbf{n}_k},\quad \psi_{j,k}=\mathcal{U}^j\psi_{0,k},\quad j\in\mathbb{Z},\ k\in\mathbb{N}.$$

Each $\psi_{j,k}$ is equal to some complex exponential $f_{\mathbf{n}}$ multiplied by a complex number of unit magnitude. The collection of $\psi_{j,k}$ is distinct, and all frequencies $\mathbf{n}\in\mathbb{Z}^m\backslash(\mathbb{Z}\times\{0\}^{m-1})$ appear, hence $\{\psi_{j,k}\}_{j\in\mathbb{Z},k\in\mathbb{N}}$ form an orthonormal basis of $\mathcal{H}_{\mathrm{pp}}^{\perp}$ with the property that $\mathcal{U}\psi_{j,k}=\psi_{j+1,k}$.

Let us show that every spectral measure is absolutely continuous on $\mathcal{H}_{\mathrm{pp}}^{\perp}$. Let g and h be arbitrary functions in $L^2(\mathbb{T}^m)$ with representations $g=\sum a_{j,k}\psi_{j,k}$ and $h=\sum b_{j,k}\psi_{j,k}$. Let the functions A_k and B_k be defined on $\mathbb{T}$ for each k with Fourier coefficients $(a_{j,k})_{j\in\mathbb{Z}}$ and $(b_{j,k})_{j\in\mathbb{Z}}$, respectively. From orthogonality, we have

$$\|g\|^2=\sum_k\int_{\mathbb{T}}\bigl|A_k(\xi)\bigr|^2\,\mathrm{d}\xi<\infty,$$

and similarly for h and B_k.

Now, we have $\mathcal{U}^n h=\sum b_{j,k}\psi_{j+n,k}$, so that

$$\begin{aligned}\bigl(g,\mathcal{U}^n h\bigr)_{L^2(\mathbb{T}^m)}&=\sum_k\sum_j a_{j+n,k}\overline{b_{j,k}}=\sum_k\int_{\mathbb{T}}\sum_j a_{j+n,k}e^{2\pi ij\xi}\,\overline{B_k(\xi)}\,\mathrm{d}\xi\\&=\int_{\mathbb{T}}e^{-2\pi in\xi}\biggl(\sum_k A_k(\xi)\overline{B_k(\xi)}\biggr)\mathrm{d}\xi,\end{aligned}$$

the nth Fourier coefficient of the function $\sum_k A_k(\xi)\overline{B_k(\xi)}$. On the other hand, applying Cauchy–Schwarz inequality twice, we have

$$\begin{aligned}\int_{\mathbb{T}}\biggl|\sum_k A_k(\xi)\overline{B_k(\xi)}\biggr|\,\mathrm{d}\xi&\leqslant\int_{\mathbb{T}}\biggl(\sum_k\bigl|A_k(\xi)\bigr|^2\biggr)^{1/2}\biggl(\sum_k\bigl|B_k(\xi)\bigr|^2\biggr)^{1/2}\mathrm{d}\xi\\&\leqslant\biggl(\int_{\mathbb{T}}\sum_k\bigl|A_k(\xi)\bigr|^2\,\mathrm{d}\xi\biggr)^{1/2}\biggl(\int_{\mathbb{T}}\sum_k\bigl|B_k(\xi)\bigr|^2\,\mathrm{d}\xi\biggr)^{1/2}\\&=\|g\|_{L^2}\|h\|_{L^2}<\infty.\end{aligned}$$

Hence the inner products $(g,\mathcal{U}^n h)_{L^2(\mathbb{T}^m)}$ are in fact the Fourier coefficients of an L^1 function. This shows that the measure associated to these inner products is absolutely continuous.

Appendix B. Proof of Lemma 9.1

Let us start by writing $f(t) = \sum_{n\neq 0} \hat{f}[n] e^{2\pi i n t}$ so that we have

$$c[k] = \sum_{n\neq 0} \hat{f}[n] \int_{\mathbb{T}} e^{2\pi i n(kv+\varphi(v))}\,\mathrm{d}v, \tag{B.1}$$

where we have changed the order of summation and integration. Applying integration by parts we obtain

$$\begin{aligned}\int_{\mathbb{T}} e^{2\pi i n\varphi(v)} e^{2\pi i n k v}\,\mathrm{d}v &= -\frac{1}{2\pi i n k}\int_{\mathbb{T}} e^{2\pi i n k v}\,\mathrm{d}\big[e^{2\pi i n\varphi(v)}\big] \\ &= -\frac{1}{k}\int_{\mathbb{T}} e^{2\pi i n k v} e^{2\pi i n\varphi(v)}\,\mathrm{d}\varphi(v). \end{aligned}\tag{B.2}$$

Part (1). For the integral in (B.2), we use the bound

$$\left|-\frac{1}{k}\int_{\mathbb{T}} e^{2\pi i n k v} e^{2\pi i n\varphi(v)}\,\mathrm{d}\varphi(v)\right| \leqslant \frac{1}{|k|}\int_{\mathbb{T}} \big|\mathrm{d}\varphi(v)\big| = \frac{1}{|k|}\|\varphi\|_{\mathrm{TV}},$$

and we simply get

$$\big|c[k]\big| \leqslant \sum_{n\neq 0} \frac{1}{|k|}\|\varphi\|_{\mathrm{TV}}\big|\hat{f}[n]\big| \leqslant \frac{1}{|k|}\|\varphi\|_{\mathrm{TV}}\|f\|_{\mathrm{A}(\mathbb{T})}.$$

Part (2). Let φ be differentiable and $\varphi' \in \mathrm{BV}(\mathbb{T})$. Substitute $\mathrm{d}\varphi(v) = \varphi'(v)\,\mathrm{d}v$ and apply another integration by parts to (B.2) obtain

$$-\frac{1}{k}\int_{\mathbb{T}} \varphi'(v) e^{2\pi i n\varphi(v)} e^{2\pi i n k v}\,\mathrm{d}v = \frac{1}{k(2\pi i n k)}\int_{\mathbb{T}} e^{2\pi i n k v}\,\mathrm{d}\big[\varphi'(v) e^{2\pi i n\varphi(v)}\big].$$

Now,

$$\mathrm{d}\big[\varphi'(v) e^{2\pi i n\varphi(v)}\big] = e^{2\pi i n\varphi(v)}\,\mathrm{d}\varphi'(v) + \big(\varphi'(v)\big)^2 (2\pi i n) e^{2\pi i n\varphi(v)}\,\mathrm{d}v,$$

so that substituting the above two formulas together with (B.2) in (B.1), we get

$$c[k] = \frac{1}{k^2}\left(\sum_{n\neq 0} \frac{\hat{f}[n]}{2\pi i n}\int_{\mathbb{T}} e^{2\pi i n(kv+\varphi(v))}\,\mathrm{d}\varphi'(v) + \sum_{n\neq 0} \hat{f}[n]\int_{\mathbb{T}} \big(\varphi'(v)\big)^2 e^{2\pi i n(kv+\varphi(v))}\,\mathrm{d}v\right). \tag{B.3}$$

For the first part of this sum we use

$$\left| \int_{\mathbb{T}} e^{2\pi i n(kv+\varphi(v))} \, \mathrm{d}\varphi'(v) \right| \leqslant \int_{\mathbb{T}} \left| \mathrm{d}\varphi'(v) \right| = \left\| \varphi' \right\|_{\mathrm{TV}},$$

and

$$\sum_{n\neq 0} \frac{|\hat{f}[n]|}{2\pi |n|} \leqslant \left(\sum_{n\neq 0} \frac{1}{(2\pi n)^2} \right)^{1/2} \left(\sum_n |\hat{f}[n]|^2 \right)^{1/2} = \frac{1}{\sqrt{12}} \|f\|_{L^2(\mathbb{T})},$$

so that

$$\left| \sum_{n\neq 0} \frac{\hat{f}[n]}{2\pi i n} \int_{\mathbb{T}} e^{2\pi i n(kv+\varphi(v))} \, \mathrm{d}\varphi'(v) \right| \leqslant \frac{1}{\sqrt{12}} \|f\|_{L^2(\mathbb{T})} \left\| \varphi' \right\|_{\mathrm{TV}}.$$

On the other hand, the second term reduces to

$$\begin{aligned} \sum_{n\neq 0} \hat{f}[n] \int_{\mathbb{T}} \big(\varphi'(v)\big)^2 e^{2\pi i n(kv+\varphi(v))} \, \mathrm{d}v &= \int_{\mathbb{T}} \big(\varphi'(v)\big)^2 \sum_{n\neq 0} \hat{f}[n] e^{2\pi i n(kv+\varphi(v))} \, \mathrm{d}v \\ &= \int_{\mathbb{T}} \big(\varphi'(v)\big)^2 f\big(kv+\varphi(v)\big) \, \mathrm{d}v. \end{aligned}$$

We bound this integral by $\|f\|_{L^\infty(\mathbb{T})} \|\varphi'\|^2_{L^2(\mathbb{T})}$. Combining these, the expression of (B.3) can now be bounded from above in absolute value as

$$\left|c[k]\right| \leqslant \frac{1}{k^2} \left(\frac{1}{\sqrt{12}} \|f\|_{L^2(\mathbb{T})} \left\| \varphi' \right\|_{\mathrm{TV}} + \|f\|_{L^\infty(\mathbb{T})} \left\| \varphi' \right\|^2_{L^2(\mathbb{T})} \right),$$

concluding the proof.

References

[1] R.L. Adler, B.P. Kitchens, M. Martens, C.P. Tresser, C.W. Wu, The mathematics of halftoning, IBM J. Res. Dev. 47 (1) (2003).

[2] D. Anastassiou, Error diffusion coding for A/D conversion, IEEE Trans. Circuits Systems 36 (3) (1989) 1175–1186.

[3] H. Anzai, Ergodic skew product transformation on the torus, Osaka Math. J. 3 (1951) 83–99.

[4] T. Bernard, From Σ–Δ modulation to digital halftoning of images, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc., Toronto, 1991, pp. 2805–2808.

[5] R. Calderbank, I. Daubechies, The pros and cons of democracy, IEEE Trans. Inform. Theory 48 (2002) 1721–1725.

[6] J.C. Candy, G.C. Temes (Eds.), Oversampling Delta–Sigma Data Converters: Theory, Design and Simulation, IEEE Press, 1992.

[7] H. Furstenberg, Strict ergodicity and transformations of the torus, Amer. J. Math. 83 (1961) 573–601.

[8] W. Chou, P.W. Wong, R.M. Gray, Multistage $\Sigma\Delta$ modulation, IEEE Trans. Inform. Theory 35 (1989) 784–796.
[9] I. Daubechies, R. DeVore, Reconstructing a bandlimited function from very coarsely quantized data: a family of stable sigma–delta modulators of arbitrary order, Ann. of Math. 158 (2) (2003) 679–710.
[10] R.M. Gray, Oversampled sigma–delta modulation, IEEE Trans. Comm. COM-35 (1987) 481–489.
[11] R.M. Gray, Spectral analysis of quantization noise in a single-loop sigma–delta modulator with dc input, IEEE Trans. Comm. COM-37 (1989) 588–599.
[12] C.S. Güntürk, Harmonic analysis of two problems in signal quantization and compression, PhD thesis, Princeton University, 2000.
[13] C.S. Güntürk, Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma–delta modulation, J. Amer. Math. Soc. 17 (1) (2004) 229–242.
[14] C.S. Güntürk, J.C. Lagarias, V.A. Vaishampayan, On the robustness of single loop sigma–delta modulation, IEEE Trans. Inform. Theory 47 (5) (2001) 1735–1744.
[15] C.S. Güntürk, N.T. Thao, Refined error analysis in second order $\Sigma\Delta$ modulation with constant inputs, IEEE Trans. Inform. Theory 50 (5) (2004) 839–860.
[16] C.S. Güntürk, One-bit sigma–delta quantization with exponential accuracy, Comm. Pure Appl. Math. 56 (11) (2003) 1608–1630.
[17] N. He, F. Kuhlmann, A. Buzo, Multi-loop $\Sigma\Delta$ quantization, IEEE Trans. Inform. Theory 38 (1992) 1015–1028.
[18] A. Katok, B. Hasselblatt, Introduction to the Modern Theory of Dynamical Systems, Cambridge University Press, 1995.
[19] Y. Katznelson, An Introduction to Harmonic Analysis, John Wiley & Sons, 1968 (reprint: Dover Publs).
[20] T.D. Kite, B.L. Evans, A.C. Bovik, T.L. Sculley, Digital image halftoning as 2-D delta–sigma modulation, in: Proc. IEEE Int. Conf. on Image Proc., Santa Barbara, CA, vol. I, 1997, pp. 799–802.
[21] L. Kuipers, H. Niederreiter, Uniform Distribution of Sequences, Wiley, 1974.
[22] S.R. Norsworthy, R. Schreier, G.C. Temes (Eds.), Delta–Sigma Data Converters: Theory, Design and Simulation, IEEE Press, 1996.
[23] W. Parry, Topics in Ergodic Theory, Cambridge University Press, 1981.
[24] R. Schreier, M.V. Goodson, B. Zhang, An algorithm for computing convex positively invariant sets for delta–sigma modulators, IEEE Trans. Circuits Systems, I 44 (1997) 38–44.
[25] N.T. Thao, Breaking the feedback loop of a class of sigma–delta A/D converters, IEEE Trans. Signal Process. 52 (12) (2004), in press.
[26] R. Ulichney, Digital Halftoning, MIT Press, Cambridge, 1987.
[27] Ö. Yılmaz, Stability analysis for several second-order sigma–delta methods of coarse quantization of bandlimited functions, Constr. Approx. 18 (4) (2002) 599–623.