Distributed Noise-Shaping Quantization: II. Classical Frames

Evan Chou and C. Sinan Güntürk

Abstract This chapter constitutes the second part in a series of papers on distributed noise-shaping quantization. In the first part, the main concept of distributed noise shaping was introduced and the performance of distributed beta encoding coupled with reconstruction via beta duals was analyzed for random frames [6]. In this second part, the performance of the same method is analyzed for several classical examples of deterministic frames. Particular consideration is given to Fourier frames and frames used in analog-to-digital conversion. It is shown in all these examples that entropic rate-distortion performance is achievable.

Keywords Finite frames, quantization, A/D conversion, noise shaping, beta encoding, beta duals.

1 Introduction

The "analysis formulation" for the quantization problem (in short, the *analysis problem*) associated to any given frame seeks to find out how well signals can be approximated after quantizing signal measurements that are taken using this frame (see, e.g. [9]). More concretely, let $\Phi := (\varphi_\alpha)_{\alpha\in I}$ be a (finite) frame in a real or complex (finite dimensional) Hilbert space $\mathscr{H}$ with inner-product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$, and $L \geq 2$ be a given integer representing the number of quantization levels to be used. The *analysis distortion* $\mathscr{D}_\mathrm{a}(\Phi,L)$ (see [6]) is formally defined by the quantity

$$\inf\left\{\sup_{\|x\|\leq 1}\inf_{q\in\mathscr{A}^I}\left\|x-\sum_{\alpha\in I}q_\alpha\psi_\alpha\right\| : (\psi_\alpha) \text{ is any dual frame of } \Phi \text{ and } |\mathscr{A}| = L\right\}.$$

Evan Chou
Google New York, e-mail: chou@cims.nyu.edu

C. Sinan Güntürk
Courant Institute, NYU, 251 Mercer Street, New York, NY 10012, e-mail: gunturk@cims.nyu.edu

Here $\mathscr{A}$ stands for the quantization alphabet, i.e. any subset of the underlying field $\mathbb{F}$ (which equals $\mathbb{R}$ or $\mathbb{C}$) of L elements.

As it was described in [6] (albeit with slightly differing notation), the analysis distortion corresponds to a practical encoding-decoding scenario: The encoder chooses $\mathscr{A}$ and quantizes the signal measurements $(\langle x, \varphi_\alpha \rangle)_{\alpha \in I}$ to generate the discrete output $(q_\alpha)_{\alpha \in I}$ in $\mathscr{A}$, knowing that the decoder will produce the approximation $\sum q_\alpha \psi_\alpha$ where $\Psi := (\psi_\alpha)_{\alpha \in I}$ is some dual frame of Φ. In this sense, the quantization alphabet $\mathscr{A}$ and the dual frame Ψ are available to both the encoder and the decoder. $\mathscr{A}$ and Ψ should be seen as system parameters which can be optimized but must remain fixed for all signals x in the unit ball of $\mathscr{H}$. The analysis distortion then measures the best achievable reconstruction error bound (over all $\mathscr{A}$ and Ψ) that is valid uniformly for all x.

It is easy to see that the analysis distortion is invariant under scaling and unitary transformations. More precisely, given any frame $\Phi := (\varphi_\alpha)_{\alpha \in I}$ in $\mathscr{H}_1$, unitary transformation $U : \mathscr{H}_1 \to \mathscr{H}_2$, and nonzero scalar $c \in \mathbb{F}$, we have

$$\mathscr{D}_{\mathrm{a}}(cU\Phi, L) = \mathscr{D}_{\mathrm{a}}(\Phi, L)$$

where $cU\Phi$ stands for the frame $(cU\varphi_\alpha)_{\alpha \in I}$ in $\mathscr{H}_2$. Hence it is always possible to reduce the discussion of the analysis distortion of frames to that of matrices (finite or infinite) as it was done in [6] which focused on random matrices. In this paper it will be more convenient for us to maintain the general framework of Hilbert spaces to allow for the possibility of working with examples of frames that are not naturally presented as matrices.

The rate-distortion performance of any quantization method is constrained by universal entropic (or volumetric) bounds. For the analysis distortion, we have (see, e.g. [6])

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) \geq L^{-N/d} \tag{1}$$

for all frames Φ in $\mathbb{R}^d$ of size $|I| =: N$, and all L. One of the main results of [6] is that if the φ_α are chosen independently from the standard Gaussian distribution on $\mathbb{R}^d$, then for any $\eta > 0$, the event

$$\left\{ \mathscr{D}_{\mathrm{a}}(\Phi, L) \leq \sqrt{d} L^{-(1-\eta)N/d} \text{ for all } L \geq 2 \right\} \tag{2}$$

holds with probability at least $1 - \exp(-c\eta^2 N)$, provided d and N/d are sufficiently large (depending only on η). Of course, with the observation made in the previous paragraph concerning unitary invariance, (1) and (2) continue to hold in any d-dimensional real Hilbert space $\mathscr{H}$ where the standard Gaussian distribution may be defined by means of any orthonormal basis of $\mathscr{H}$.

Complex Hilbert spaces were not studied in [6] but can be handled with relatively straightforward modifications which we will introduce in this paper (see Section 2 and the Appendix). Note, in particular, that the universal lower bound (1) needs to be replaced by $L^{-N/2d}$ for the complex case; this can be seen by porting the Lebesgue measure on $\mathbb{R}^{2d}$ on to $\mathbb{C}^d$ and repeating the volume-covering argument given in [6].

Statement of the Main Results

This is the second part in an ongoing series of works on *distributed noise-shaping quantization*. In the first paper [6], the analysis distortion bound in (2) was achieved by means of a general algorithmic framework called *distributed noise-shaping*, and in particular, using the method of *distributed beta encoding* coupled with reconstruction via *beta duals*. In this second paper we will apply this method to some classical examples of deterministic frames.

The frames that we will consider in this paper fall into a general category we call *unitarily generated frames*. In essense, this means that the index set I can be chosen as $\mathbb{Z}_N$ or $\mathbb{Z}$ depending on the size of the frame, and there exists a unitary operator U on $\mathscr{H}$ such that

$$\varphi_n = U\varphi_{n-1} \tag{3}$$

for all $n \in I$. (See Section 4 for the technical definition.) Well-known examples that fall into this category include Fourier frames, real harmonic frames, and frames of (uniform) translates.

The main result of this paper in the case of unitarily generated frames of size N in d dimensions, assuming N is a multiple of d and a certain technical condition satisfied by Fourier frames, is that

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) \lesssim c(\varphi_0) N d^{-1} \cdot \begin{cases} L^{-N/d}, & \text{if } \mathbb{F} = \mathbb{R} \text{ and } L \geq 2, \\ \lfloor \sqrt{L} \rfloor^{-N/d}, & \text{if } \mathbb{F} = \mathbb{C} \text{ and } L \geq 4, \end{cases} \tag{4}$$

where $c(\varphi_0)$ is a constant that is independent of N and L (see Theorem 2). Generically, $c(\varphi_0)$ is of order $\sqrt{d}$. Note that the bound in (4) behaves better than the one in (2), and considering (1), it is essentially optimal.

The case of infinite dimensional Hilbert spaces requires some modifications and we only consider the classical problem of analog-to-digital conversion of bandlimited functions via uniform sampling and reconstruction by interpolation. With the help of the beta dual machinery, first we establish a new sampling theorem, and then we show that for uniform sampling of real-valued bandlimited functions with oversampling ratio λ, the analysis distortion can be bounded by $C\lambda L^{-\lambda+1}$ which is the infinite dimensional analog of (4) (see Section 5).

2 Background and Review of Methodology

In this section we will review the general theory of noise-shaping quantizers as well as the particular method of distributed beta encoding and beta duals. Further details on the methodology can be found in [6].

2.1 Basics of Noise Shaping for Frames

The main principle of noise-shaping quantization is to arrange for the quantization error (the quantization "noise") to be close to the kernel of the reconstruction operator. For concreteness we assume here that I is a finite index set, but the principle extends to infinite dimensional cases with suitable modifications. Given the measurements $y_\alpha := \langle x, \varphi_\alpha \rangle$, $\alpha \in I$, of a signal $x \in \mathscr{H}$ using a frame $\Phi := (\varphi_\alpha)_{\alpha \in I}$, a noise-shaping quantizer seeks to find a solution (u,q) to the equation

$$y - q = Hu \tag{5}$$

where $y := (y_\alpha) \in \mathbb{F}^I$, $q := (q_\alpha) \in \mathscr{A}^I$, $H : \mathbb{F}^I \to \mathbb{F}^I$ is a linear operator called the "noise transfer operator" of the noise-shaping quantizer, and $u \in \mathbb{F}^I$ is an auxiliary variable, often called the "state vector". Sigma-delta ($\Sigma\Delta$) modulators constitute the most important example of traditional noise-shaping quantizers (see [10] for an engineering perspective, [7, 8] for mathematical expositions, and [3, 9] for applications to finite frames).

Given any dual frame $\Psi := (\psi_\alpha)$ of Φ, we then have

$$x - \sum_{\alpha \in I} q_\alpha \psi_\alpha = \sum_{\alpha \in I} (Hu)_\alpha \psi_\alpha = \sum_{\alpha' \in I} u_{\alpha'} \psi^H_{\alpha'} \tag{6}$$

where

$$\psi^H_{\alpha'} := \sum_{\alpha \in I} H_{\alpha,\alpha'} \psi_\alpha$$

and H has the matrix representation $(H_{\alpha,\alpha'})$. Noise-shaping quantizers are typically designed to keep $\|u\|_\infty$ small. Ideally $\|u\|_\infty$ should be controlled independently of $|I|$; such a scheme is called *stable*. With stability, the error representation (6) results in the effective bound

$$\left\| x - \sum_{\alpha \in I} q_\alpha \psi_\alpha \right\| \leq \|u\|_\infty \|\Psi^H\|_{\ell^\infty(I) \to \mathscr{H}} \tag{7}$$

where $\Psi^H : \ell^\infty(I) \to \mathscr{H}$ is the operator given by

$$\Psi^H(u) := \sum_{\alpha \in I} u_\alpha \psi^H_\alpha .$$

Picking an orthogonal basis for $\mathscr{H}$, we may identify the frame Ψ with a matrix (which we may also denote by Ψ) whose columns consist of the coefficients of ψ_α in this basis. Then we have $\Psi^H = \Psi H$ and the operator norm $\|\Psi^H\|_{\ell^\infty(I) \to \mathscr{H}}$ equals the matrix norm $\|\Psi H\|_{\infty \to 2}$.

With the objective of minimizing the error bound (7), the main question is then how to choose H and the dual frame Ψ while ensuring stability of u. In the next subsection we will review a particular choice of H and Ψ that was proposed in [6], namely the noise transfer operator of *distributed beta encoding* and the *beta dual* of Φ, respectively. To ensure stability, we will employ the common toolkit known as the *greedy quantizer*, which was also used in [6]. The small but necessary modifications for complex-valued measurements are explained in the Appendix where a general form of the greedy quantizer which results in some additional improvements is also given.

2.2 Distributed Beta Encoding and Beta Duals of Frames

For any given frame $\Phi := (\varphi_\alpha)_{\alpha\in I}$ in $\mathscr{H}$, pick a partition $\Pi := (I_0,\dots,I_{p-1})$ of I where $N := |I| \geq p \geq d := \dim(\mathscr{H})$, and for each j in

$$[p] := \{0,\dots,p-1\},$$

pick a scalar $\beta_j \in \mathbb{F}$ with magnitude at least 1 and a bijection $\sigma_j : [N_j] \to I_j$ where N_j denotes $|I_j|$. Define

$$\zeta_j := \sum_{n\in[N_j]} \bar{\beta}_j^{-n} \varphi_{\sigma_j(n)}; \;\; j \in [p].$$

Suppose $(\zeta_j)_0^{p-1}$ is itself a frame for $\mathscr{H}$. Let $(\eta_j)_0^{p-1}$ be any dual frame of $(\zeta_j)_0^{p-1}$ and define a new collection of vectors $\Psi := (\psi_\alpha)_{\alpha\in I}$ via

$$\psi_{\sigma_j(n)} := \beta_j^{-n}\eta_j; \quad n \in [N_j], \quad j \in [p].$$

Then Ψ is a dual of Φ because

$$\begin{aligned}
\sum_{\alpha\in I} \langle x, \varphi_\alpha\rangle \psi_\alpha &= \sum_{j\in[p]} \sum_{n\in[N_j]} \langle x, \varphi_{\sigma_j(n)}\rangle \psi_{\sigma_j(n)} \\
&= \sum_{j\in[p]} \sum_{n\in[N_j]} \langle x, \bar{\beta}_j^{-n} \varphi_{\sigma_j(n)}\rangle \eta_j \\
&= \sum_{j\in[p]} \langle x, \zeta_j\rangle \eta_j \\
&= x.
\end{aligned}$$

We assume that $(\eta_j)_0^{p-1}$ is chosen to be the canonical dual of $(\zeta_j)_0^{p-1}$, denoted by $(\widetilde{\zeta}_j)_0^{p-1}$. Supressing the underlying partition Π, the bijections (σ_j), and the values (β_j), we then call Ψ the *beta dual* of Φ. We also say that $(\zeta_j)_0^{p-1}$ is the *beta condensation* of Φ. (See [6] for a more general definition of condensation of frames.)

The concept of beta duals is inherently tied with what we call *distributed beta encoding*. This is a noise-shaping quantization method which is carried out via the system of difference equations

$$y_{\sigma_j(n)} - q_{\sigma_j(n)} = u_{\sigma_j(n)} - \beta_j u_{\sigma_j(n-1)}, \quad n \in [N_j], \;\; j \in [p], \tag{8}$$

where for notational convenience we set $u_{\sigma_j(-1)} := 0$. In other words, the noise-transfer operator H has a block-diagonal matrix representation (see [6]).

The significance of distributed beta encoding coupled with beta duals for reconstruction lies in the following calculation:

$$\begin{aligned} x - \sum_{\alpha\in I} q_\alpha \psi_\alpha &= \sum_{j\in[p]} \sum_{n\in[N_j]} (y_{\sigma_j(n)} - q_{\sigma_j(n)}) \psi_{\sigma_j(n)} \\ &= \sum_{j\in[p]} \left(\sum_{n\in[N_j]} (u_{\sigma_j(n)} - \beta_j u_{\sigma_j(n-1)}) \beta_j^{-n} \right) \widetilde{\zeta}_j \\ &= \sum_{j\in[p]} u_{\sigma_j(N_j-1)} \beta_j^{-N_j+1} \widetilde{\zeta}_j. \end{aligned} \tag{9}$$

Let A_ζ to be the lower frame bound of $(\zeta_j)_0^{p-1}$, that is,

$$\sum_{j\in[p]} |\langle x, \zeta_j\rangle|^2 \geq A_\zeta \|x\|^2 \quad \text{for all } x \in \mathscr{H}.$$

Then, as is well-known in frame theory, we have

$$\left\| \sum_{j\in[p]} a_j \widetilde{\zeta}_j \right\| \leq \|a\|_2 / \sqrt{A_\zeta} \quad \text{for all } a \in \mathbb{F}^p. \tag{10}$$

(In frame theory terminology, this result is a consequence of the fact that if T, T^*, and $S := T^*T$ denote the analysis, the synthesis, and the frame operators for $(\zeta_j)_0^{p-1}$, respectively, then $S^{-1}T^*$ is the synthesis operator for $(\widetilde{\zeta}_j)_0^{p-1}$ with norm equal to $\|S^{-1/2}\| = 1/\sqrt{A_\zeta}$.) Combining (9) and (10), it follows that

$$\left\| x - \sum_{\alpha\in I} q_\alpha \psi_\alpha \right\| \leq \frac{1}{\sqrt{A_\zeta}} \left\| \left(u_{\sigma_j(N_j-1)} \beta_j^{-N_j+1} \right)_0^{p-1} \right\|_2 \leq \|u\|_\infty \beta_*^{-N_*+1} \sqrt{\frac{p}{A_\zeta}}, \tag{11}$$

where $\beta_* := \min_j |\beta_j|$, $N_* := \min_j N_j$. Note that $N_* \leq N/p$ but there always exists a partition Π that achieves $N_* = \lfloor N/p \rfloor$. As in [6] we will assume this is the case. In fact, in all of the examples considered in this paper p will divide N and all the β_j will be equal to a common positive real number that we will call β.

We show in the Appendix that

- for $\mathbb{F} = \mathbb{R}$, the condition $\beta + \|y\|_\infty/\delta \leq L$ is sufficient to guarantee that (8) is solvable with $\|u\|_\infty \leq \delta$ and $q \in \mathscr{A}^I$ for some $\mathscr{A} \subset \mathbb{R}$, $|\mathscr{A}| = L$, and
- for $\mathbb{F} = \mathbb{C}$, the condition $\beta + \|y\|_\infty/\delta \leq \lfloor \sqrt{L} \rfloor$ is sufficient to guarantee that (8) is solvable with $\|u\|_\infty \leq \sqrt{2}\delta$ and $q \in \mathscr{A}^I$ for some $\mathscr{A} \subset \mathbb{C}$, $|\mathscr{A}| = L$.

Since $\beta \geq 1$, the above sufficient condition for the complex case can only be invoked if $L \geq 4$. However, $L = 3$ can also be employed using a different quantizer.

We show in the Appendix that (8) is solvable for any $\beta < 4/3$. Note that $\beta \leq \sqrt{L}$ is a necessary condition for the complex case due to the entropic lower bound $L^{-N/2d}$ for the analysis distortion. Currently we do not know if the gap from $4/3$ to $\sqrt{3}$ can be closed for $L = 3$. Also see [1] where the case $L = 3$ appears for $\beta = 1$.

In order to bound $\mathscr{D}_{\mathrm{a}}(\Phi, L)$ via (11), a two-level strategy can be executed: At the basic level, the system parameters β and δ should be chosen optimally, i.e. so as to minimize $\delta\beta^{-N_*+1}$, subject to one of the sufficient stability conditions above. At the more advanced level, the partition Π and the bijections $(\sigma_j)_0^{p-1}$ should also be seen as system parameters that can be chosen optimally so as to minimize $1/\sqrt{A_\zeta}$. In other words, the beta condensation frame $(\zeta_j)_0^{p-1}$ should be made as tight as possible. This second stage of optimization was not invoked for random frames in [6] (except for the value of p) and it will not be invoked for the classical examples considered in this paper either because natural partition choices will work near optimally; however, in other specific examples there may be need to consider it. Here note that A_ζ implicitly depends on β too, but for the examples we will study in this paper this dependence will not play a critical role.

It is worth noting that the case $\beta = 1$ with $p = 1$ corresponds to first-order $\Sigma\Delta$ quantization which has been studied in depth for finite frames [3]. The second level of optimization that arises in this case has been found to relate to the traveling salesman problem [11]. Higher-order $\Sigma\Delta$ schemes perform better but they remain suboptimal in the rate-distortion sense.

3 Warm up: Beta Duals of Finite Fourier Frames

Let $\mathscr{H} := \mathbb{C}^d$ be equipped with the Euclidean inner-product. For any $N \geq d$, the standard finite Fourier frame $\mathscr{F}_{N,d} := (\varphi_n)_0^{N-1}$ of size N is given in Cartesian coordinates by

$$\varphi_{n,k} := \frac{1}{\sqrt{d}} \mathrm{e}^{2\pi \mathrm{i} nk/N}; \quad n \in [N]; \quad k \in [d].$$

For simplicity, we assume in this paper that N is a multiple of d. With this assumption, we set $p := d$, $N_j := N_* := N/d$ for all $j \in [d]$, and

$$\sigma_j(n) := jN_* + n; \quad j \in [d]; \quad n \in [N_*].$$

Also we set $\beta_j = \beta$ for all $j \in [p]$, where β is a real number greater than 1 to be determined later. Then the beta condensation of $\mathscr{F}_{N,d}$ is computed explicitly to be

$$\zeta_{j,k} = \sum_{n \in [N_*]} \beta^{-n} \varphi_{\sigma_j(n),k} = \frac{1}{\sqrt{d}} w_k \mathrm{e}^{2\pi \mathrm{i} jk/d}; \quad j \in [d]; \;\; k \in [d],$$

where

$$w_k := \sum_{n \in [N_*]} \left(\beta^{-1} \mathrm{e}^{2\pi \mathrm{i} k/N}\right)^n; \quad k \in [d].$$

This formula shows that the beta condensation of a finite Fourier frame (with the parameters we have used) is actually a weighted discrete Fourier system (which is a basis if and only if all w_k are nonzero). It is now straightforward to compute the frame bounds. Indeed $\langle x, \zeta_j \rangle$ can be seen as the jth Discrete Fourier Transform (DFT) coefficient of $(x_k \overline{w}_k)_0^{d-1}$ so that (either by Parseval's identity or by explicit calculation) we have

$$\sum_{j\in[d]} |\langle x, \zeta_j \rangle|^2 = \sum_{k\in[d]} |x_k|^2 |w_k|^2.$$

Note that for any complex number $|z| < 1$ and any $m \geq 1$, we have

$$|1+z+\cdots+z^{m-1}| = \left| \frac{1-z^m}{1-z} \right| \geq \frac{1-|z|}{1+|z|} \tag{12}$$

so that

$$\min_{k\in[d]} |w_k| \geq \frac{1-\beta^{-1}}{1+\beta^{-1}} =: C_\beta. \tag{13}$$

Hence the lower frame bound A_ζ of $(\zeta_j)_0^{d-1}$ satisfies

$$A_\zeta \geq C_\beta^2. \tag{14}$$

In light of the discussion of the previous section, we can now proceed with the optimization of system parameters. For all $x \in \mathbb{C}^d$ such that $\|x\|_2 \leq 1$, we have $\|y\|_\infty \leq 1$ so that for any $L \geq 4$ we can employ a quantization alphabet $\mathscr{A} \subset \mathbb{C}$ with at most L elements, guaranteeing $\|u\|_\infty \leq \sqrt{2}\delta$, where β and δ must satisfy the condition $\beta + 1/\delta \leq \lfloor \sqrt{L} \rfloor$. For any such β and δ, it follows from (11) that

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d}, L) \leq \sqrt{2d} C_\beta^{-1} \delta \beta^{-\frac{N}{d}+1}. \tag{15}$$

In order to choose the special values of δ and β, we employ the following elementary lemma whose proof we leave as an exercise (for a nearly identical version, see [6, Lemma 3.2]):

Lemma 1. *For any $K \geq 2$ and $\alpha \geq 1$, let $\beta := K(\alpha+1)/(\alpha+2)$ and $\delta := (\alpha+2)/K$. Then $\beta \geq 4/3$, $\beta + 1/\delta = K$, and*

$$\delta \beta^{-\alpha+1} < \mathrm{e}(\alpha+1)K^{-\alpha}. \tag{16}$$

Furthermore, C_β as defined by (13) *satisfies $C_\beta^{-1} \leq 7$.*

We use this lemma for $K := \lfloor \sqrt{L} \rfloor$ and $\alpha := N/d$. Injecting the resulting bound (16) and the bound $C_\beta^{-1} \leq 7$ in (15), we arrive at the following near-optimal result:

Theorem 1. *Suppose N is a multiple of d. Then for any number of quantization levels $L \geq 4$, the analysis distortion of the finite Fourier frame of N elements in $\mathbb{C}^d$ satisfies*

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d},L) < 7\mathrm{e}\sqrt{2d}\left(\frac{N}{d}+1\right)\lfloor\sqrt{L}\rfloor^{-N/d}.$$

Remark 1. The above theorem is actually still valid for $L \leq 3$ but it does not offer a useful bound since then we have $\lfloor\sqrt{L}\rfloor = 1$. For $L = 3$ we may instead invoke the triangular alphabet $\mathscr{A}$ and the associated quantization rule described in the Appendix for which we may set $\beta := \left(\frac{4}{3}\right)^{1-\varepsilon}$ for any $\varepsilon \in (0,1)$. It can then be checked that

$$\mathscr{D}_{\mathrm{a}}(\mathscr{F}_{N,d},3) \lesssim_{\varepsilon} \sqrt{d}\left(\frac{4}{3}\right)^{-(1-\varepsilon)N/d}.$$

We omit the details. This is the only upper bound we know for $L = 3$ that is exponentially small in N/d; however it does not match the entropic lower bound of $3^{-N/2d}$.

4 Generalization: Unitarily Generated Frames

A general method of constructing uniform tight frames based on frame paths was introduced in [4] in connection with analyzing the performance of $\Sigma\Delta$ quantization for finite frames. In this section, first we will slightly extend this frame construction method to include a larger class of frames, and then bound the analysis distortion of these frames using distributed beta encoding.

4.1 *Unitary frame paths*

Let $\mathscr{H} := \mathbb{C}^d$ be equipped with the Euclidean inner-product and Ω be a $d \times d$ Hermitian matrix. Consider the 1-parameter group of unitary operators on $\mathscr{H}$ given by

$$U_t := \mathrm{e}^{2\pi\mathrm{i}\Omega t}; \;\; t \in \mathbb{R},$$

and for any $\varphi_0 \in \mathbb{C}^d$ of unit norm, let

$$\varphi_n := U_{\frac{n}{N}}\varphi_0; \quad n = 0,\ldots,N-1.$$

The curve $\{t \mapsto U_t\varphi_0 : t \in [0,1]\}$ is called a *unitary frame path* if $\Phi := (\varphi_n)_0^{N-1}$ yields a frame for infinitely many $N \geq d$. We also say that Φ is *unitarily generated.*

Assume Ω has d distinct integer eigenvalues $\lambda_0,\ldots,\lambda_{d-1}$ which are also distinct modulo N. Let us denote the corresponding normalized eigenvectors of Ω by $v_0,\ldots,v_{d-1}$. This collection gives us an orthogonal basis of $\mathscr{H}$. Now note that

$$\langle v_k,\varphi_n\rangle = \langle \mathrm{e}^{-2\pi\mathrm{i}\Omega n/N}v_k,\varphi_0\rangle = \mathrm{e}^{-2\pi\mathrm{i}\lambda_k n/N}\langle v_k,\varphi_0\rangle; \;\; n \in [N],\; k \in [d],$$

so that

$$\begin{aligned}\sum_{n\in[N]}|\langle x,\varphi_n\rangle|^2 &= \sum_{n\in[N]}\left|\sum_{k\in[d]}\langle x,v_k\rangle\langle v_k,\varphi_n\rangle\right|^2\\ &= \sum_{k\in[d]}\sum_{l\in[d]}\langle x,v_k\rangle\langle v_l,x\rangle\langle v_k,\varphi_0\rangle\langle\varphi_0,v_l\rangle\sum_{n\in[N]}\mathrm{e}^{2\pi\mathrm{i}(\lambda_l-\lambda_k)n/N}\\ &= N\sum_{k\in[d]}|\langle x,v_k\rangle|^2|\langle\varphi_0,v_k\rangle|^2,\end{aligned}$$

where in the last equality we have used the assumption that $\lambda_0,\dots,\lambda_{d-1}$ are distinct modulo N.

With this identity, it now follows that

$$N\left(\min_{k\in[d]}|\langle\varphi_0,v_k\rangle|^2\right)\|x\|^2 \le \sum_{n\in[N]}|\langle x,\varphi_n\rangle|^2 \le N\left(\max_{k\in[d]}|\langle\varphi_0,v_k\rangle|^2\right)\|x\|^2. \qquad (17)$$

Hence we see that $(\varphi_n)_0^{N-1}$ is a frame if and only if $\langle\varphi_0,v_k\rangle\neq 0$ for all $k\in[d]$. We also see (as in [4]) that $(\varphi_n)_0^{N-1}$ is a unit-norm tight frame if and only $|\langle\varphi_0,v_k\rangle| = 1/\sqrt{d}$ for all k. Note that the frame condition is generic, i.e. the set of φ_0 which yield a frame is an open dense subset of $\mathscr{H}$. In contrast, the condition for tightness of the frame is quite strict, corresponding to a nowhere dense set of φ_0.

Remark 2. The above argument continues to hold under the weaker assumption that all pairwise differences $\lambda_l-\lambda_k$ are integers and are nonzero modulo N if $l\neq k$. In other words, it is possible to shift all the eigenvalues by a common real value without changing the frame property. Note that (U_t) is 1-periodic in t if and only if all the eigenvalues are integers in which case the frame path is a closed curve.

Remark 3. Note that the finite Fourier frame of the previous section corresponds to the case when Ω is the diagonal matrix with the diagonal entries $0,\dots,d-1$ and $\varphi_0=(1,\dots,1)/\sqrt{d}$.

More generally, we may pick any $J\subset[N]$ of cardinality d to form the diagonal entries $\lambda_0,\dots,\lambda_{d-1}$ (in increasing order) of a diagonal matrix Ω. The resulting tight frame can be characterized equivalently as the restriction of the finite Fourier basis of $\mathbb{C}^N$ to the space of *timelimited* vectors $\mathscr{H}:=\{x\in\mathbb{C}^N:\mathrm{supp}(x)\subset J\}$.

By duality we can also consider the space of discrete *bandlimited* vectors $\mathscr{B}_J:=\{x\in L^2(\mathbb{Z}_N):\mathrm{supp}(\widehat{x})\subset J\}$. For any φ_0 such that $\mathrm{supp}(\widehat{\varphi_0})=J$, the system $(\varphi_n)_{n\in\mathbb{Z}_N}$ defined via translating φ_0, i.e. by setting $\varphi_{n,k}:=\varphi_0(k-n)$, $k,n\in\mathbb{Z}_N$, constitute a unitarily generated frame for $\mathscr{B}_J$.

Unitarily generated frames in $\mathbb{R}^d$. If the Hermitian Ω is such that all of its entries are purely imaginary, i.e., $\mathrm{i}\Omega$ is a real, skew-symmetric matrix, then (U_t) reduces to a group of real, orthogonal matrices. Then $(\varphi_n)_0^{N-1}$ is a unitarily generated frame in $\mathbb{R}^d$ provided $\varphi_0\in\mathbb{R}^d$ and $\langle\varphi_0,v_k\rangle\neq 0$ for all $k\in[d]$. Note that the eigenvectors (v_k) would still need to be considered as vectors in $\mathbb{C}^d$.

The simplest nontrivial example is in $\mathbb{R}^2$. Two examples are worth mentioning: First, we may consider $\Omega := B := \left(\begin{smallmatrix}0 & \mathrm{i}\\ -\mathrm{i} & 0\end{smallmatrix}\right)$. Here the eigenvalues 1 and -1 of Ω are distinct modulo N if only if $N \geq 3$. We may also consider $\Omega := B/2 = \begin{pmatrix} 0 & \mathrm{i}/2 \\ -\mathrm{i}/2 & 0 \end{pmatrix}$ for which the condition in Remark 2 is satisfied for all $N \geq 2$. This frame is actually the *semicircle frame* in $\mathbb{R}^2$ (see [4, 6]).

The *harmonic frames* in $\mathbb{R}^d$ for $d = 2m$ are obtained by setting Ω to be the block diagonal matrix with the blocks $B, 2B, \ldots, mB$, and eigenvalues $\{\pm 1, \ldots, \pm m\}$. Again we require $N \geq d+1$ as a frame condition. For $d = 2m+1$, a 1×1 "0 block" is added resulting in the eigenvalues $\{0, \pm 1, \ldots, \pm k\}$. See [4] for additional information.

4.2 Beta Duals of Unitarily Generated Frames

Let $(\varphi_n)_0^{N-1}$ be a unitarily generated frame in $\mathbb{F}^d$ as described in Section 4.1 where Ω, $(\lambda_k)_0^{d-1}$, and $(v_k)_0^{d-1}$ have the same meaning as before. Let $d \leq p \leq N$. For simplicity, we assume that N is a multiple of p, and set $N_* := N_j := N/p$ for all $j \in [p]$. As in Section 3, we set $\sigma_j(n) := jN_* + n$, $n \in [N_*]$, and $\beta_j = \beta > 1$ for all $j \in [p]$. Then the beta condensation of the frame $(\varphi_n)_0^{N-1}$ is given by

$$\zeta_j := \sum_{n \in [N_*]} \beta^{-n} \varphi_{\sigma_j(n)} = \sum_{k \in [d]} w_k \mathrm{e}^{2\pi \mathrm{i} j \lambda_k / p} \langle \varphi_0, v_k \rangle v_k; \quad j \in [p],$$

where

$$w_k := \sum_{n \in N_*} \left(\beta^{-1} \mathrm{e}^{2\pi \mathrm{i} \lambda_k / N} \right)^n; \;\; k \in [d].$$

Assuming the stronger hypothesis that $\lambda_0, \ldots, \lambda_{d-1}$ are distinct modulo p, or more generally, that

$$\lambda_l - \lambda_k \text{ are integers and nonzero modulo } p \text{ if } l \neq k, \tag{18}$$

we have

$$\begin{aligned} \sum_{j \in [p]} |\langle x, \zeta_j \rangle|^2 &= \sum_{j \in [p]} \left| \sum_{k \in [d]} w_k \mathrm{e}^{2\pi \mathrm{i} j \lambda_k / p} \langle \varphi_0, v_k \rangle \langle x, v_k \rangle \right|^2 \\ &= p \sum_{k \in [d]} |\langle x, v_k \rangle|^2 |\langle \varphi_0, v_k \rangle|^2 |w_k|^2. \end{aligned}$$

Using (12), we have $|w_k| \geq C_\beta = (1 - \beta^{-1})/(1 + \beta^{-1})$ as before, so we find that the lower frame bound A_ζ of $(\zeta_j)_0^{p-1}$ satisfies

$$A_\zeta \geq p C_\beta^2 \left(\min_{k \in [d]} |\langle \varphi_0, v_k \rangle|^2 \right).$$

The rest of the discussion where we bound the analysis distortion of Φ is the same as before. Namely, we invoke (11) and follow the same procedure as in the case of finite Fourier frames of Section 3, starting from (14). In addition, for $\mathbb{F} = \mathbb{R}$ we may employ a quantizer in $\mathbb{R}$ for all $L \geq 2$ where we can set $K = L$ and the state vector satisfies $\|u\|_\infty \leq \delta$. The result is summarized in the following theorem:

Theorem 2. *Suppose N is a multiple of p where $p \geq d$, and Φ is a unitarily generated frame in $\mathbb{F}^d$ such that* (18) *holds. Then we have*

$$\mathscr{D}_{\mathrm{a}}(\Phi, L) < 7\mathrm{e}\left(\frac{N}{p}+1\right) c(\varphi_0) \cdot \begin{cases} \sqrt{2}\lfloor\sqrt{L}\rfloor^{-N/p}, & \text{if } \mathbb{F} = \mathbb{C} \text{ and } L \geq 4, \\ L^{-N/p}, & \text{if } \mathbb{F} = \mathbb{R} \text{ and } L \geq 2, \end{cases}$$

where

$$c(\varphi_0) := \left(\min_{1 \leq k \leq d} |\langle \varphi_0, v_k \rangle| \right)^{-1}.$$

Of course, a bound for the case $L = 3$ can also be given as in the previous section.

5 An Infinite-Dimensional Case: Bandlimited Functions on $\mathbb{R}$

Discussing quantized frame representations in infinite dimensional Hilbert spaces requires special care due to the fact that the coefficient sequence $(q_\alpha)_{\alpha \in I}$ is not in $\ell^2(I)$ (and therefore the reconstruction is not guaranteed to be of finite norm) unless $q_\alpha = 0$ for all but finitely many α. Of course, for this to happen, 0 would need to be a permissible quantization level in $\mathscr{A}$ in the first place. Then the problem becomes similar to a finite dimensional one with one main difference: the finite dimensional subspace from which a quantized approximation is sought would need to be either specified *a priori*, or determined *a posteriori* by means of the quantization algorithm itself.

Another approach is to relax the Hilbertian frame setting and consider frame-like representations in other suitable normed spaces and with a different sense of convergence, as well as the possibility of approximation by quantized representations from outside these spaces. Indeed this is the sense in which the classical oversampled quantization problem of bandlimited functions on $\mathbb{R}$ has been studied mathematically [7, 8]. A general (and highly nontrivial) theory for quantization for frames in Banach spaces was also developed in [5]. In this short section we will only be concerned with the case of uniform sampling of bandlimited functions where it will be possible for us to work from scratch.

The analysis distortion of sampling. Let $\mathscr{B}_\Omega$ be the space of bounded continuous functions x on $\mathbb{R}$ for which the (distributional) Fourier transform $\widehat{x}$ is supported in $[-\Omega, \Omega]$. This space contains the classical Paley-Wiener space PW_Ω which comes with an additional square-integrability constraint. (PW_Ω is therefore a Hilbert space with respect to the standard inner-product on $L^2(\mathbb{R})$.) We equip $\mathscr{B}_\Omega$ with the L^∞-norm which is more suitable for quantization. The celebrated Shannon-Nyquist sampling theorem (in the context of $\mathscr{B}_\Omega$) says that any $x \in \mathscr{B}_\Omega$ can be recovered perfectly from its samples $(x(k\tau))_{k\in\mathbb{Z}}$ via a pointwise absolutely convergent expansion

$$x(t) = \tau \sum_{k\in\mathbb{Z}} x(k\tau)\psi(t-k\tau), \tag{19}$$

where $\tau < \tau_{\text{crit}} := \frac{1}{2\Omega}$ and ψ is any function of rapid decay on $\mathbb{R}$ such that

$$\widehat{\psi}(\xi) = \begin{cases} 1, \ |\xi| \le \Omega, \\ 0, \ |\xi| \ge \frac{1}{2\tau}. \end{cases} \tag{20}$$

We will say that such a ψ is (Ω,τ)-*admissible*. The value $\rho := 1/\tau$ is called the sampling rate, and $\rho_{\text{crit}} := 1/\tau_{\text{crit}} = 2\Omega$ is called the critical (or Nyquist) sampling rate. The *oversampling ratio* given by

$$\lambda := \frac{\rho}{\rho_{\text{crit}}} = \frac{\tau_{\text{crit}}}{\tau} \tag{21}$$

corresponds to the "redundancy" of the sampling operator $\Phi_\tau : \mathscr{B}_\Omega \to \ell^\infty(\mathbb{Z})$ where

$$(\Phi_\tau x)_k := x(k\tau), \quad k \in \mathbb{Z}.$$

Let us say that a collection of bounded continuous functions $\Psi := (\psi_k)_{k\in\mathbb{Z}}$ on $\mathbb{R}$ is *quantization admissible* if $\sum c_k \psi_k$ converges (pointwise absolutely) to a bounded function whenever $c \in \ell^\infty(\mathbb{Z})$. Let us also say that Ψ is *dual* to Φ_τ on $\mathscr{B}_\Omega$ if, in addition, we have

$$x = \sum_{k\in\mathbb{Z}} (\Phi_\tau x)_k \psi_k = \sum_{k\in\mathbb{Z}} x(k\tau)\psi_k \ \text{ for all } x \in \mathscr{B}_\Omega, \tag{22}$$

where again the convergence is understood to be pointwise and absolute. This equation generalizes the concept of frame and the classical sampling formula (19) where the ψ_k are $\tau\mathbb{Z}$-translations of a fixed function. The analog of analysis distortion associated to Φ_τ on $\mathscr{B}_\Omega$ for L levels of quantization, now denoted by $\mathscr{D}_{\text{a}}(\Phi_\tau|\mathscr{B}_\Omega, L)$ is then naturally defined to be

$$\inf\left\{ \sup_{\|x\|_\infty \le 1} \inf_{q\in\mathscr{A}^{\mathbb{Z}}} \left\| x - \sum_{k\in\mathbb{Z}} q_k\psi_k \right\|_\infty : \Psi \text{ is dual to } \Phi_\tau \text{ on } \mathscr{B}_\Omega \text{ and } |\mathscr{A}| = L \right\}.$$

Beta dual of the sampling operator. Note that in the context of PW_Ω this sampling operator can be realized in terms of the unitarily generated frame consisting of the $\tau\mathbb{Z}$-translations of a fixed sinc kernel. The following construction mimicks the beta dual machinery of Sections 2 and 4. Since our setup is not Hilbertian, we will take a direct approach in our construction.

Given $\tau < \tau_{\text{crit}}$, let $\lambda_* := \lceil\lambda\rceil - 1 = \lceil\tau_{\text{crit}}/\tau\rceil - 1$ and $\tau_* := \lambda_*\tau$. Note that $\lambda > \lambda_* \ge 1$ and $\tau \le \tau_* < \tau_{\text{crit}}$. For any given $\beta > 1$, consider the operators

$$\begin{aligned} Tf &:= \sum_{n\in[\lambda_*]} \beta^{-n} f(\cdot + n\tau), \\ Sf &:= f - \beta^{-1} f(\cdot + \tau), \quad \text{and} \\ Rf &:= \sum_{n=0}^{\infty} \beta^{-\lambda_* n} f(\cdot + n\tau_*) \end{aligned}$$

on $L^\infty(\mathbb{R})$ where they are also clearly bounded. All three operators represent convolution operators with distributional kernels and it is evident after inspecting their Fourier multipliers that RS inverts T. Avoiding distributions, we can check this fact directly. Indeed, for any $f \in L^\infty(\mathbb{R})$, we have

$$RSTf = R(f - \beta^{-\lambda_*} f(\cdot + \tau_*)) = f.$$

All three operators enjoy a crucial property which is stronger than their continuity on $L^\infty(\mathbb{R})$: Whenever a function series $\sum f_k$ converges pointwise absolutely (but not necessarily uniformly) to a bounded function, we have

$$R\sum f_k = \sum R f_k \quad \text{(similarly for } S \text{ and } T\text{)}, \tag{23}$$

where the latter series also converges pointwise absolutely. To see this, simply note that the iterated series

$$\sum_{n=0}^{\infty} \beta^{-\lambda_* n} \sum_k |f_k(t + n\tau_*)|$$

is convergent for all t; hence it is justified to change the order of summation that is required to prove (23).

Let ψ_* be (Ω, τ_*)-admissible. Given any $x \in \mathscr{B}_\Omega$, it is clear that $Tx \in \mathscr{B}_\Omega$ as well, and we can apply Shannon's sampling theorem to Tx with the reconstruction filter ψ_* on the sampling grid $\tau_*\mathbb{Z}$ to obtain

$$Tx = \tau_* \sum_{j\in\mathbb{Z}} \left(\sum_{n\in[\lambda_*]} \beta^{-n} x(j\tau_* + n\tau) \right) \psi_*(\cdot - j\tau_*).$$

We apply RS to both sides of this equation. Using (23) and noting translation invariance of RS, we obtain

$$x = \sum_{j\in\mathbb{Z}} \sum_{n\in[\lambda_*]} x\left((\lambda_* j + n)\tau\right) \tau_* \beta^{-n} (RS\psi_*)(\cdot - j\tau_*). \tag{24}$$

We now set $\psi := RS\psi_*$, and define $\Psi := (\psi_k)_{k\in\mathbb{Z}}$ by

$$\psi_{\lambda_* j + n} := \tau_* \beta^{-n} \psi(\cdot - j\tau_*); \quad j \in \mathbb{Z},\ n \in [\lambda_*].$$

Then (24) says nothing but that Ψ is dual to Φ_τ on $\mathscr{B}_\Omega$. It is easy to see that ψ also has rapid decay so that Ψ is a quantization-admissible dual.

For the quantization process, we employ the same distributed beta encoding approach as before, and this time, set $y_k := x(k\tau)$, $\sigma_j(n) := \lambda_* j + n$,

$$y_{\sigma_j(n)} - q_{\sigma_j(n)} = u_{\sigma_j(n)} - \beta u_{\sigma_j(n-1)}; \quad j \in \mathbb{Z},\ n \in [\lambda_*],$$

with $\sigma_j(-1) := 0$ so that

$$x - \sum_{j\in\mathbb{Z}} \sum_{n\in[\lambda_*]} q_{\lambda_* j+n} \tau_* \beta^{-n} \psi(t - j\tau_*) = \beta^{-\lambda_*+1} \sum_{j\in\mathbb{Z}} u_{\sigma_j(\lambda_*-1)} \tau_* \psi(t - j\tau_*),$$

and therefore

$$\left\| x - \sum_{k\in\mathbb{Z}} q_k \psi_k \right\|_\infty \leq \beta^{-\lambda_*+1} \|u\|_\infty C(\psi), \tag{25}$$

where

$$\begin{aligned} C(\psi) &:= \left\| \sum_{j\in\mathbb{Z}} \tau_* |\psi(\cdot - j\tau_*)| \right\|_\infty \\ &\leq \left\| \sum_{j\in\mathbb{Z}} \tau_* RS |\psi_*(\cdot - j\tau_*)| \right\|_\infty \\ &\leq \|RS\|_{\infty\to\infty} \left\| \sum_{j\in\mathbb{Z}} \tau_* |\psi_*(\cdot - j\tau_*)| \right\|_\infty \\ &\leq \frac{\beta+1}{\beta-1} C(\psi_*). \end{aligned} \tag{26}$$

Note that τ_* is near τ_{crit} in a uniform manner; for example, it is easy to show that we have $\tau_* \in [\tau_{\text{crit}}/2, \tau_{\text{crit}})$. This allows us to choose ψ_* purely as a function of Ω via $\psi_*(t) := \Omega \psi_{*,0}(\Omega t)$ for a fixed $\psi_{*,0}$. Consequently, we may replace $C(\psi_*)$ by a universal constant independent of Ω.

Assuming we are only concerned with real-valued functions and employing a quantization alphabet of L levels requiring $\beta + 1/\delta \leq L$, we may set β and δ in (25) as indicated by Lemma 1. The end result now reads

$$\mathscr{D}_{\text{a}}(\Phi_\tau | \mathscr{B}_\Omega, L) \lesssim \lambda L^{-\lceil \lambda \rceil + 1}.$$

The modifications for complex-valued bandlimited functions would be the same as before.

Concluding Remarks

We have not touched upon many classical frames that are popular in theory and practice, such as non-harmonic Fourier frames, frames of irregular sampling and interleaved sampling, Gabor frames, and filter bank frames. Gabor frames are generated by two unitary transformations, modulation and translation, which do not commute. Sub-optimal results can be obtained in a straightforward manner by focusing on only one of the generators and applying the basic beta dual machinery. However, additional work (e.g. on the noise transfer operator) may be necessary in order to exploit all of the redundancy present in a Gabor frame. Similar comments are applicable for filter bank frames as well.

Appendix: Greedy Quantizer for Complex Measurements

In this section we will provide a generalization of the complex-valued $\Sigma\Delta$ quantization algorithm given in [2, Proposition 3.1] and the greedy noise-shaping quantization algorithm given in [6, Theorem 2.1]. The result, which is applicable to both real and complex quantization alphabets, offers nontrivial improvements in the complex case thanks to the use of general semi-norms to measure closeness.

Lemma 2. *Let $\mathscr{A}$ be a quantization alphabet in $\mathbb{C}$, B_* be the closed unit ball of a semi-norm $|\cdot|_*$ on $\mathbb{C}$ treated as a vector space over $\mathbb{R}$, and $H := (H_{n,m})_{n,m\in[N]}$ be an $N \times N$ real-valued lower-triangular matrix with unit diagonal. Suppose there exist positive real numbers μ, δ, γ such that*

$$\delta B_* + \mathscr{A} \supset \gamma B_* \tag{27}$$

and

$$\mu + \delta \max_{n\in[N]} \sum_{m<n} |H_{n,m}| \leq \gamma. \tag{28}$$

Then for any $y \in \mathbb{C}^N$ such that $|y_n|_* \leq \mu$ for all $n \in [N]$, there exist $q \in \mathscr{A}^N$ and $u \in \mathbb{C}^N$ such that

$$y - q = Hu$$

where $|u_n|_* \leq \delta$ for all $n \in [N]$.

Proof. The proof of this result is yet another adaptation of a well-known induction argument. By our assumption on H, we are seeking to satisfy the equations

$$u_n = \left(y_n - \sum_{m<n} H_{n,m} u_m \right) - q_n \tag{29}$$

for all $n \in [N]$.

Since $|y_0|_* \leq \mu \leq \gamma$, (27) implies that there exist $q_0 \in \mathscr{A}$ and $u_0 \in \delta B_*$ such that $u_0 + q_0 = y_0$. Hence (29) is satisfied for $n = 0$ and $|u_0|_* \leq \delta$.

For the induction step, assume that $|u_m|_* \leq \delta$ for all $m < n$, and let

$$w_n := y_n - \sum_{m<n} H_{n,m} u_m.$$

Using sub-additivity and homogeneity of $|\cdot|_*$ followed by the condition given in (28), we get

$$|w_n|_* \leq \mu + \delta \sum_{m<n} |H_{n,m}| \leq \gamma;$$

hence, because of (27) again, there exist $q_n \in \mathscr{A}$ and $u_n \in \delta B_*$ such that $u_n + q_n = w_n$, i.e. (29) holds. □

Special known cases. There are certainly many ways to choose $\mathscr{A}$ and $|\cdot|_*$. We first note two important special cases of practical importance. Here L denotes $|\mathscr{A}|$.

($\mathbb{R}$) *Real arithmetic progression*
This quantizer uses $\mathscr{A} := \mathscr{A}_{L,\delta} := \{(-L+2l-1)\delta : 1 \leq l \leq L\} \subset \mathbb{R}$, i.e. the origin-symmetric arithmetic progression of length L and spacing 2δ along with $|z|_* := |\Re(z)|$. Then B_* is the infinite vertical strip $\{z : |\Re(z)| \leq 1\}$ and (27) holds for $\gamma := L\delta$. Using the algorithm in Lemma 2, $y \in \mathbb{R}^N$ results in $u \in \mathbb{R}^N$, and $\|y\|_\infty \leq \mu$ implies $\|u\|_\infty \leq \delta$ so that the setup becomes identical to that of [6].

($\mathbb{C}$) *Complex square lattice quantizer*
This quantizer assumes $L = K^2$ for some positive integer K and sets $\mathscr{A} := \mathscr{A}_{K,\delta} + i\mathscr{A}_{K,\delta} \subset \mathbb{C}$ along with $|z|_* := \max(|\Re(z)|, |\Im(z)|)$. B_* can be identified with $[-1,1]^2$ (as a subset $\mathbb{R}^2$) so that (27) is valid for $\gamma := K\delta$. Since $|z|_* \leq |z| \leq \sqrt{2}|z|_*$ for any $z \in \mathbb{C}$, $\|y\|_\infty \leq \mu$ implies $|y_n|_* \leq \mu$ for all n and Lemma 2 then yields $\|u\|_\infty \leq \sqrt{2}\delta$.
When K is even, the resulting $\mathscr{A}$ has no real points and it may be desirable to require that $y \in \mathbb{R}^N$ always yields $q \in \mathbb{R}^N$. In this case, we may instead use the slightly larger alphabet $\mathscr{A} := \mathscr{A}_{K,\delta} + i\mathscr{A}_{K+1,\delta}$ for which $L = K(K+1)$. This choice indeed corresponds to the one made in [2]. Another natural possibility in this case is to use the 1-norm in $\mathbb{R}^2$ coupled with the diamond lattice as shown in Fig. 1 for $K = 2$.
Note that for the square (or diamond) lattice quantizer of K^2 levels, using the Euclidean norm $|\cdot|$ on $\mathbb{C}$ would be sub-optimal. Indeed, the largest value of γ that can be used in (27) is $\gamma = \frac{K}{\sqrt{2}}\delta$.

Hexagonal norm for a tri-level complex alphabet. It is natural to ask if a complex quantization alphabet $\mathscr{A}$ with fewer than 4 levels can be used in connection with the noise-shaping quantization algorithm of Lemma 2. For $L = 3$, we may set $\mathscr{A}$ to be the vertices of an equilateral triangle in $\mathbb{C}$ centered at the origin. If the Euclidean norm is used, then it is not difficult to prove that the largest value of γ that can be used in (27) is $\gamma = \frac{2}{\sqrt{3}}\delta$ (see Fig. 2 for a demonstration of this covering). In this case, $\|y\|_\infty \leq \mu$ yields $\|u\|_\infty \leq \delta$.

An alternative we have found useful is to employ the norm $|\cdot|_*$ induced by a regular hexagonal body whose sides are aligned with the sides of the triangle. Then, as shown in Fig. 2, we can attain $\gamma = \frac{4}{3}\delta$. By choosing the scale of the hexagonal body suitably, we can ensure $|z|_* \leq |z| \leq \frac{2}{\sqrt{3}}|z|_*$ so that $\|y\|_\infty \leq \mu$ implies $|y_n|_* \leq \mu$ for all n, and therefore Lemma 2 yields $\|u\|_\infty \leq \frac{2}{\sqrt{3}}\delta$. Despite the increase in the bound for $\|u\|_\infty$, there is a sizable gain in the "expansion factor" γ/δ from $\frac{2}{\sqrt{3}}$ to $\frac{4}{3}$. This gain is crucial for beta encoding because any β up to this expansion factor is admissible for stability via Lemma 2 provided $\mathscr{A}$, γ, and δ are suitably scaled to meet (27) and (28) simultaneously.

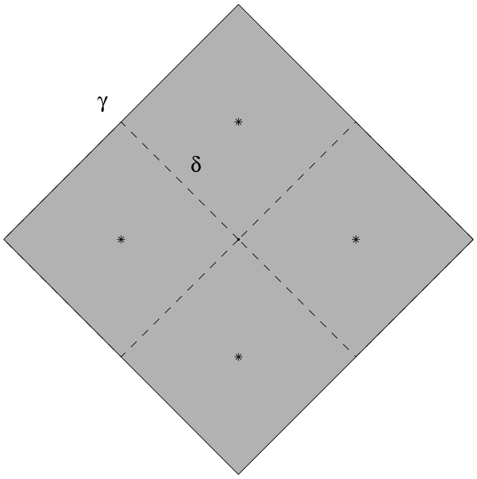

Fig. 1 Lattice covering for the 1-norm in $\mathbb{R}^2$ where $L = 4$ and $\gamma = 2\delta$.

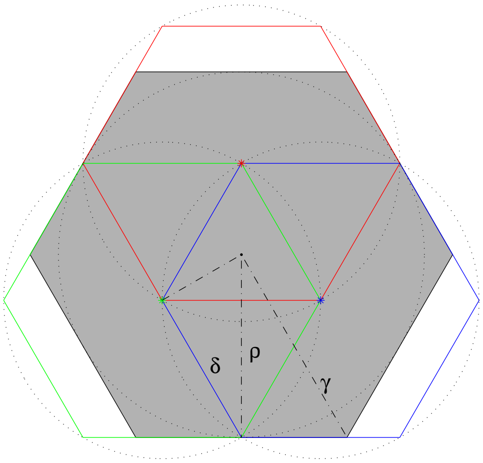

Fig. 2 Three identical hexagons at scale δ covering a larger hexagon at scale $\gamma = \frac{4}{3}\delta$ compared to three identical circular discs at scale δ covering a larger circular disc at scale $\rho = \frac{2}{\sqrt{3}}\delta$.

References

1. R. Adler, T. Nowicki, G. Świrszcz, and C. Tresser. Convex dynamics with constant input. *Ergodic Theory and Dynamical Systems*, 30:957–972, 8 2010.
2. J.J. Benedetto, O. Oktay, and A. Tangboondouangjit. Complex sigma-delta quantization algorithms for finite frames. In *Radon transforms, geometry, and wavelets*, volume 464 of *Contemp. Math.*, pages 27–49. Amer. Math. Soc., Providence, RI, 2008.
3. J.J. Benedetto, A.M. Powell, and Ö. Yılmaz. Sigma-delta quantization and finite frames. *Information Theory, IEEE Transactions on*, 52(5):1990–2005, 2006.
4. B.G. Bodmann and V.I. Paulsen. Frame paths and error bounds for sigma-delta quantization. *Appl. Comput. Harmon. Anal.*, 22(2):176–197, 2007.
5. P.G. Casazza, S.J. Dilworth, E. Odell, Th. Schlumprecht, and A. Zsk. Coefficient quantization for frames in Banach spaces. *Journal of Mathematical Analysis and Applications*, 348(1):66 – 86, 2008.
6. E. Chou and C.S. Güntürk. Distributed noise-shaping quantization: I. Beta duals of finite frames and near-optimal quantization of random measurements. *Constructive Approximation*, 44(1):1–22, 2016.
7. I. Daubechies and R. DeVore. Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbitrary order. *The Annals of Mathematics*, 158(2):679–710, 2003.
8. C.S. Güntürk. Mathematics of analog-to-digital conversion. *Comm. Pure Appl. Math.*, 65(12):1671–1696, 2012.
9. A.M. Powell, R. Saab, and Ö. Yılmaz. Quantization and finite frames. In G. Peter Casazza and Gitta Kutyniok, editors, *Finite Frames: Theory and Applications*, Appl. Numer. Harmon. Anal., pages 267–302. Birkhäuser/Springer, New York, 2013.
10. R. Schreier and G. C. Temes. *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.
11. Y. Wang. Sigma-delta quantization errors and the traveling salesman problem. *Adv. Comput. Math.*, 28(2):101–118, 2008.