Quantization of bandlimited functions using random samples

Rohan Joy
Department of Mathematics,
Indian Institute of Technology Madras
Chennai, India
rohanjoy96@gmail.com

Felix Krahmer
Department of Mathematics,
Technical University of Munich
and Munich Center for Machine Learning
Garching/Munich, Germany
felix.krahmer@tum.de

Alessandro Lupoli
Department of Mathematics,
Technical University of Munich
Garching/Munich, Germany
alessandrolupoli97@gmail.com

Radha Ramakrishnan
Department of Mathematics,
Indian Institute of Technology Madras
Chennai, India
radharam@iitm.ac.in

***Abstract*—We investigate the compatibility of distributed noise-shaping quantization with random samples of bandlimited functions. Let f be a real-valued π-bandlimited function. Suppose $R > 1$ is a real number, and assume that $\{x_i\}_{i=1}^m$ is a sequence of i.i.d random variables uniformly distributed on $[-\tilde{R}, \tilde{R}]$, where $\tilde{R} > R$ is appropriately chosen. We show that on using a distributed noise-shaping quantizer to quantize the values of f at $\{x_i\}_{i=1}^m$, a function $f^\sharp$ can be reconstructed from these quantized values such that $\|f - f^\sharp\|_{L^2[-R,R]}$ decays with high probability as m and $\tilde{R}$ increase.**

I. Introduction

In signal processing, one of the primary goals is to obtain a digital representation of a function in a signal space suitable for storage, transmission and recovery. This goal is usually attained through two steps, sampling and quantization. In sampling, we sample the function at appropriate data points such that the function can be stably reconstructed using those samples. Consider, for example, the space of the bandlimited functions. The Shannon sampling theorem tells us that any analog time π-bandlimited signal f can be reconstructed entirely by sampling it at the integer points $\mathbb{Z}$.

In the second step of quantization, we reduce these real or complex-valued function samples to a discrete finite set. The finiteness of the set allows for the digital storage and processing of its elements. More specifically, given a signal x and a vector y containing linear measurements (function samples, frame measurements etc.) Ax of x. The quantization process involves replacing the measurement vector y with a vector q from a finite set $\mathcal{A}$, called the quantization alphabet, such that accurate reconstruction of x from q is possible.

One of the most intuitive approaches to quantization is memoryless scalar quantization (MSQ), which simply replaces each coefficient of the measurement vector $y = Ax$ with its nearest element from $\mathcal{A}$. If the vector y consists of frame coefficients, a simple method for reconstruction is to fix a dual frame and linearly reconstruct with MSQ quantized coefficients. However, this is not an effective approach. In fact, the authors in [9] show that even when using an optimal reconstruction scheme to approximate x from its MSQ quantized coefficients, the mean squared error cannot be better than $\mathcal{O}(\lambda^{-1})$ with linear reconstruction methods. Here λ denotes the oversampling ratio.

One of the reasons why the MSQ approach falls short is because it naively quantizes each coefficient without regard for how the other coefficients in the vector are quantized. In order to solve this issue, quantization schemes such as $\Sigma\Delta$ quantization and distributed noise-shaping quantization were introduced. These schemes follow a recursive procedure to push the quantization error $y - q$ into an unoccupied portion of the signal spectrum.

$\Sigma\Delta$ quantization schemes were introduced in [14] for the quantization of oversampled bandlimited functions. Since then, they have been studied extensively [7], [8], [10] in this context. The use of $\Sigma\Delta$ quantization in the setting of finite frames has also been explored by various authors [1], [2], [11], [16]–[18].

In their paper [3], Chou and Güntürk introduced the concept of distributed noise-shaping quantization and were able to achieve an exponentially small error bound in the quantization of Gaussian random finite frame expansions. They extended their results to the setting of unitarily generated frames in [4]. Recently, this particular scheme has been applied to fast binary embeddings [13] and spectral super-resolution [12].

Although the quantization of oversampled bandlimited functions using uniform samples has been investigated significantly, to the best of our knowledge, only two papers [6], [19] are available in the literature that deal with the scenario of quantizing irregular samples of bandlimited functions. In [6], the authors first give a formula to reconstruct any bandlimited

FK and AL acknowledge support by the German Science Foundation (DFG) in the context of the collaborative research center TR 109, RJ and FK acknowledge support by the German Science Foundation (DFG) in the context of the Emmy Noether junior research group KR4512/1-1.

function f from its samples $\{f(t_n)\}_{n\in\mathbb{Z}}$, where $\{t_n\}_{n\in\mathbb{Z}}$ is a uniformly discrete sequence such that $\sup_{n\in\mathbb{Z}}|t_n - \frac{n}{\lambda}| < \infty$, and $\lambda > 1$. They then construct a dithered A/D converter and show that f can be accurately reconstructed from its quantized samples taken at $\{t_n\}_{n\in\mathbb{Z}}$. In [19], the authors show that if a bandlimited function f is sampled on an interleaved collection of N uniform grids $\{kT + T_n\}_{k\in\mathbb{Z}}$ with $\{T_n\}_{n=1}^N$ chosen independently from $[0, T]$ $(T < 1)$, and the samples are quantized with a first order $\Sigma\Delta$ scheme, then with high probability the error $\|f - \tilde{f}\|_{L^\infty(\mathbb{R})}$ turns out to be atmost of the order $\frac{\log N}{N}$. Here, $\tilde{f}$ represents the function reconstructed from the quantized values.

In contrast to [6], where the particularly developed A/D converter needs a very specific sampling process, and [19], where the sample points are interleaved randomly shifted grids, we do not put such stringent limits on the structure of the sample points. Our work deals with the distributed noise-shaping quantization of bandlimited functions using random samples, where it is assumed that the samples $\{x_i\}_{i=1}^m$ are a sequence of i.i.d random variables uniformly distributed on a suitable interval, i.e. they are completely random and have no kind of structure. Let $R > 1$, and assume that $\{x_i\}_{i=1}^m$ is a sequence of i.i.d random variables uniformly distributed on $[-R-3m^{\frac{1}{16}}, R+3m^{\frac{1}{16}}]$. Given a real-valued π-bandlimited function f, in our main result Theorem 3.1 (stated in Section III), we prove that if a stable distributed noise-shaping quantization scheme is used, then the reconstruction error satisfies

$$\|f - f^\sharp\|_{L^2[-R,R]} \leq \frac{d_1 R}{m^{\frac{7}{16}}} \tag{1}$$

with probability greater than $1 - 17m^{\frac{15}{16}} \exp\left(-\frac{m^{\frac{3}{8}}}{d_2 R}\right)$. Here, the function $f^\sharp$ denotes the function reconstructed using the quantized values, and d_1 and d_1 are known positive constants. To illustrate our result, we discuss the following specific case. Select $m = R^{16}$ in the preceding configuration; here, it is assumed that R is such that R^{16} satisfies the condition on m in the theorem statement. Then, each x_i is distributed uniformly on $[-4R, 4R]$. Further, the bound in (1) simplifies to $\|f - f^\sharp\|_{L^2[-R,R]} \leq \frac{d_1}{R^6}$ and the probability bound changes to $1 - 17R^{15} \exp\left(-\frac{R^5}{d_2}\right)$. Hence, we obtain decay in R; this becomes particularly useful when R is large.

II. Preliminaries

A. Notations

- For any positive real number t, let

$$[t] := \{-\lfloor t\rfloor, -(\lfloor t\rfloor - 1), \cdots, 0, \cdots, \lfloor t\rfloor - 1, \lfloor t\rfloor\}.$$

  Here $\lfloor t\rfloor$ denotes the greatest integer less than or equal to t.
- Let $X = (x_1, \cdots, x_{n_1}) \in \mathbb{C}^{n_1}$ and $Y = (y_1, \cdots, y_{n_2}) \in \mathbb{C}^{n_2}$, then

$$X \frown Y := (x_1, \cdots, x_{n_1}, y_1, \cdots, y_{n_2}) \in \mathbb{C}^{n_1+n_2}.$$

- The Fourier transform of a function $f \in L^1(\mathbb{R})$, is defined as

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-it\xi} dt,$$

  with usual extension via a unitary operator to $L^2(\mathbb{R})$.
- Let $PW_{[-\pi,\pi]}$ denote the space of π-bandlimited functions i.e.

$$PW_{[-\pi,\pi]} := \{f \in L^2(\mathbb{R}) : \text{supp}(\hat{f}) \subset [-\pi, \pi]\}.$$

- $$C_{[-\pi,\pi]} := \{f \in PW_{[-\pi,\pi]} : \|f\|_{L^\infty(\mathbb{R})} \leq 1 \text{ and } f \text{ is real valued}\}.$$

B. Sampling of bandlimited functions

The celebrated Shannon sampling theorem says that for any $f \in PW_{[-\pi,\pi]}$, we have

$$f(t) = \sum_{n\in\mathbb{Z}} f(n)\,\text{sinc}(t - n) \qquad \forall\ t \in \mathbb{R} \tag{2}$$

where $\text{sinc}(t) := \frac{\sin \pi t}{\pi t}$. However, (2) is not useful in practice because $\text{sinc}(t)$ decays too slowly. To circumvent this issue, it is useful to introduce oversampling. This can easily be done as shown in [6]. We review the method over here.

Let $\lambda > 1$ be a fixed real number. Choose a function g such that

1) $\hat{g} \in C^\infty$.
2) $$\hat{g}(\xi) = \begin{cases} \frac{1}{\sqrt{2\pi\lambda}} & \xi \in [-\pi, \pi], \\ 0 & |\xi| > (2\lambda - 1)\pi. \end{cases} \tag{3}$$
3) $$\sum_{k\in\mathbb{Z}} \left|\hat{g}(\xi + 2k\lambda\pi)\right|^2 = \frac{1}{2\pi\lambda} \quad \forall\, \xi \in \mathbb{R}. \tag{4}$$

Using [5, Theorem 9.3] and (4), it can be seen that $\left\{g\left(\cdot - \frac{k}{\lambda}\right) : k \in \mathbb{Z}\right\}$ forms a orthonormal system. Now, consider any bandlimited function $f \in PW_{[-\pi,\pi]}$, then its Fourier transform can be viewed as an element in $L^2[-(2\lambda-1)\pi, (2\lambda-1)\pi]$. Using the Fourier series expansion of $\hat{f}$ on $[-(2\lambda-1)\pi, (2\lambda-1)\pi]$ and (3), it can be shown that

$$f(t) = \frac{1}{\sqrt{\lambda}} \sum_{n\in\mathbb{Z}} f\left(\frac{n}{\lambda}\right) g\left(t - \frac{n}{\lambda}\right) \quad \forall\, t \in \mathbb{R}. \tag{5}$$

In comparison to (2), the above reconstruction formula, although requiring more samples, has the advantage that each sample is weighted in a very localized way $\left(f\left(\frac{n}{\lambda}\right)\right.$ only contributes in a small neighbourhood of $\left.\frac{n}{\lambda}\right)$. This property will be exploited by us when we find a finite-dimensional approximation space for $PW_{[-\pi,\pi]}$.

C. Distributed noise-shaping quantization

First, we define the quantization alphabet $\mathcal{A}_L^\delta$ that we use in our paper.

Definition 2.1: For a positive integer L and a real number $\delta > 0$, let the quantization alphabet $\mathcal{A}_L^\delta$ be defined as

$$\mathcal{A}_L^\delta := \{\pm(2l-1)\delta : 1 \le l \le L, l \in \mathbb{Z}\}.$$

Let m and p be positive integers such that p divides m. Then for any fixed $\beta > 1$, the block diagonal distributed noise-shaping transfer operator H_β [3] is defined as

$$H_\beta = I_p \otimes \tilde{H}_\beta, \tag{6}$$

where I_p denotes the identity matrix of order p and $\tilde{H}_\beta$ is the following $\frac{m}{p} \times \frac{m}{p}$ matrix

$$(\tilde{H}_\beta)_{ij} := \begin{cases} 1 & \text{if } i = j, \\ -\beta & \text{if } i = j+1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Definition 2.2: Given a quantization alphabet $\mathcal{A}$, by a distributed noise-shaping quantizer we mean any map $Q : \mathbb{R}^m \to \mathcal{A}^m$ that satisfies

$$y - q = H_\beta u, \tag{8}$$

where $q := Q(y)$, H_β is the $m \times m$ matrix given in (6), and u is a vector such that $\|u\|_\infty \le c$ for some constant c which is independent of m. The existence of such schemes is given by the following lemma.

Lemma 2.3: [13, Lemma 4.2] Let $\mathcal{A} := \mathcal{A}_L^\delta$. Suppose that $\|y\|_\infty \le \mu$ and

$$\beta + \frac{\mu}{\delta} \le 2L.$$

For each $s \ge 1$, let $w_s := y_s + \sum_{j=1}^{s-1}(I - H_\beta)_{s,s-j} u_{s-j}$,

$$q_s := (\mathcal{Q}_\beta(y))_s = \arg\min_{r \in \mathcal{A}} |w_s - r|,$$

and

$$u_s := w_s - q_s.$$

Then the resulting q satisfies the noise-shaping relation (8) with $\|u\|_\infty \le \delta$.

Let the vector $\nu_\beta := [\beta^{-1}\beta^{-2}\cdots\beta^{-\frac{m}{p}}]$ and the β condensation operator [3] $\tilde{V}_\beta$ be defined as

$$\tilde{V}_\beta := I_p \otimes \frac{\nu_\beta}{\|\nu_\beta\|_1}. \tag{9}$$

Then we can easily determine the following bound, which we will utilize later.

$$\|\tilde{V}_\beta H_\beta\|_{\infty \to 2} \le \sqrt{p}\beta^{-\frac{m}{p}+1}. \tag{10}$$

III. Main result

A. Statement of the result

Theorem 3.1: Let $R > 1, \delta > 0$ be real numbers, L be a positive integer, and $f \in C_{[-\pi,\pi]}$. Fix $\beta \in (1, 2L - \frac{1}{\delta})$. Assume that $\{x_i\}_{i=1}^m$ is a sequence of i.i.d random variables that are uniformly distributed on $[-R - 3m^{\frac{1}{16}}, R + 3m^{\frac{1}{16}}]$. Let $\mathcal{Q}_\beta$ be the quantization scheme from Lemma 2.3 with alphabet $\mathcal{A}_L^\delta$. If $m^{\frac{15}{16}}$ is an integer such that it divides m and m is sufficiently large, then with probability greater than

$$1 - 17m^{\frac{15}{16}} \exp\left(-\frac{m^{\frac{3}{8}}}{d_2 R}\right)$$

we have

$$\|f - F_{W\tilde{V}_\beta} q\|_{L^2[-R,R]} \le \frac{d_1 R}{m^{\frac{7}{16}}}, \tag{11}$$

where $F_{W\tilde{V}_\beta}$ and y are as defined in (15) and (16) respectively, $q := Q_\beta(y)$, and d_1 and d_2 are known positive constants.

B. Sketch of proof of Theorem 3.1

Our proof combines the theory of quantization of bandlimited functions with the theory of frames. It is divided into three steps. In the following, we will outline the ideas of these steps; for the complete proof, we refer the reader to the journal version of this paper [15].

1) A suitable finite dimensional approximation space: As the first step, we project the infinite-dimensional space of bandlimited functions onto a finite-dimensional space. This enables us to work with a finite number of samples, which ensures that the sequence of quantized frame measurements $(q_\lambda)_{\lambda \in I}$ belongs to $l^2(I)$. The chosen finite-dimensional space is such that the orthogonal projection Pf of any bandlimited function f onto it approximates f well in a neighbourhood of $[-R, R]$.

To make this notion of neighbourhood mathematically precise, fix an $\epsilon > 0$ such that $\epsilon R \ge 1$ and consider the interval $[-(1+3\epsilon)R, (1+3\epsilon)R]$. It can be partitioned into $I_{1\epsilon} := (-(1+\epsilon)R, (1+\epsilon)R)$, $I_{2\epsilon} := (-(1+2\epsilon)R, -(1+\epsilon)R] \cup [(1+\epsilon)R, (1+2\epsilon)R)$ and $I_{3\epsilon} := [-(1+3\epsilon)R, -(1+2\epsilon)R] \cup [(1+2\epsilon)R, (1+3\epsilon)R]$. Let g and λ be as defined in the Subsection II-B.

Definition 3.2: Define,

$$V^{(1+\frac{5}{2}\epsilon)R}(g) := span\left\{g\left(\cdot - \frac{k}{\lambda}\right) : k \in \left[\lambda\left(1 + \frac{5}{2}\epsilon\right)R\right]\right\}. \tag{12}$$

Let P denote the orthogonal projection from $L^2(\mathbb{R})$ onto $V^{(1+\frac{5}{2}\epsilon)R}(g)$. Then, for any $f = \frac{1}{\sqrt{\lambda}}\sum_{k\in\mathbb{Z}} f(\frac{k}{\lambda}) g(\cdot - \frac{k}{\lambda}) \in PW_{[-\pi,\pi]}$, we have

$$P(f) = \frac{1}{\sqrt{\lambda}} \sum_{k \in [\lambda(1+\frac{5}{2}\epsilon)R]} f\left(\frac{k}{\lambda}\right) g\left(\cdot - \frac{k}{\lambda}\right). \tag{13}$$

Given an $f \in PW_{[-\pi,\pi]}$, we approximate it with the function Pf from the finite dimensional approximation space $V^{(1+\frac{5}{2}\epsilon)R}(g)$. From the formula for Pf, it is clear that it has been calculated by simply replacing the samples of f outside

$\left[-\left(1+\frac{5}{2}\epsilon\right)R, \left(1+\frac{5}{2}\epsilon\right)R\right]$ with 0. However, as the function samples are weighted locally because of the decay of g, this replacement has a minimal effect on the function f in the region $I_{1\epsilon} \cup I_{2\epsilon}$ if ϵR is large enough. Hence the projected function Pf will be a good approximation of f on $I_{1\epsilon}$ and $I_{2\epsilon}$, but it need not be on $I_{3\epsilon}$.

2) A random frame for the approximation space: In the next step, we find a random frame for the finite-dimensional approximation space $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$, such that the frame measurements for any $h \in V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$ can be calculated using its function samples.

If we want to work with a reasonable number of samples, then any frame satisfying the above condition will require us to sample in a region larger than $[-\left(1+\frac{5}{2}\epsilon\right)R, \left(1+\frac{5}{2}\epsilon\right)R]$. We give a simple heuristic argument for this. Suppose we have no sample points in an interval $[a,b] \subset [-\left(1+\frac{5}{2}\epsilon\right)R, \left(1+\frac{5}{2}\epsilon\right)R]$. Then it cannot be expected that a function h, which is concentrated on $[a,b]$, can be reconstructed using a feasible number of its frame measurements, as the frame measurements do not use the samples of h on the region where it is concentrated.

Subsequently, assume that $\{x_i\}_{i=1}^m$ is a sequence of i.i.d random variables that are uniformly distributed on $[-(1+3\epsilon)R, (1+3\epsilon)R]$. Since we work in the finite-dimensional space $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$, the projected function's samples $\{Pf(x_i)\}_{i=1}^m$ would be needed to compute the frame measurements of Pf. However, the samples available are of the original function f's and not the projected function Pf's. Since by construction, Pf approximates f well in $I_{1\epsilon} \cup I_{2\epsilon}$, we use the function samples $\{f(x_i)\}_{i=1}^m$ instead of $\{Pf(x_i)\}_{i=1}^m$ to calculate frame measurements. Here, we must remember that we also have sampling points in the region $I_{3\epsilon}$, where the Pf does not approximate f well.

The goal is to devise a reconstruction process where the error generated by this approximation of samples can be controlled and minimized. It turns out that the main issue is the completely random ordering of the sample points $\{x_i\}_{i=1}^m$, which leads to a mix-up of accurately approximated sample values $\{f(x_i)\}_{x_i \in I_{1\epsilon} \cup I_{2\epsilon}}$ and potentially inaccurately approximated sample values $\{f(x_i)\}_{x_i \in I_{3\epsilon}}$ during reconstruction. In order to solve this, we design a procedure to partition the random variables $\{x_i\}_{i=1}^m$ into three disjoint collections. As the random variables are i.i.d uniformly distributed, also the random variables in each of these three collections are i.i.d uniformly distributed, albeit on different intervals. More precisely, in any realization of the random variables $\{x_i\}_{i=1}^m$, there will be some m_1 points in $I_{1\epsilon}$, m_2 points in $I_{2\epsilon}$ and m_3 points in $I_{3\epsilon}$. These m_1, m_2 and m_3 will be random variables and can take the values $\{0, 1, \cdots, m\}$.

Definition 3.3: Let p be a positive integer that divides m. We define the following new random variables.

1) $p_i := \sum_{j=1}^{i} \left\lfloor \frac{m_j p}{m} \right\rfloor, \tilde{m}_i := \left\lfloor \frac{m_i p}{m} \right\rfloor \frac{m}{p} \;\; \forall\, i \in \{1,2,3\}$.
2) $\tilde{m} := \frac{m}{p} p_3$.
3) For all $i \in \{1, \cdots, \tilde{m}_1\}$, define $y_i^1 = x_n$ where n is the i-th number such that $x_n \in I_{1\epsilon}$. Conditionally on m_1, m_2 and m_3, each of the random variables $\{y_i^1\}_{i=1}^{\tilde{m}_1}$ will be i.i.d uniformly distributed on $I_{1\epsilon}$.
4) For all $i \in \{1, \cdots, \tilde{m}_2\}$, define $y_i^2 = x_n$ where n is the i-th number such that $x_n \in I_{2\epsilon}$. Conditionally on m_1, m_2 and m_3, each of the random variables $\{y_i^2\}_{i=1}^{\tilde{m}_2}$ will be i.i.d uniformly distributed on $I_{2\epsilon}$.
5) For all $i \in \{1, \cdots, \tilde{m}_3\}$, define $y_i^3 = x_n$ where n is the i-th number such that $x_n \in I_{3\epsilon}$. Conditionally on m_1, m_2 and m_3, each of the random variables $\{y_i^3\}_{i=1}^{\tilde{m}_3}$ will be i.i.d uniformly distributed on $I_{3\epsilon}$.
6) Define $\{\epsilon_i^1\}_{i=1}^{\tilde{m}_1}$,$\{\epsilon_i^2\}_{i=1}^{\tilde{m}_2}$ and $\{\epsilon_i^3\}_{i=1}^{\tilde{m}_3}$ to be sequences of ± 1 Bernoulli independent random variables that are also independent from all the above defined random variables.

Define the operator E from $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$ to $\mathbb{C}^{\tilde{m}}$ as

$$\begin{aligned} E(f) &= \{f(y_i^1)\}_{i=1}^{\tilde{m}_1} \frown \{f(y_i^2)\}_{i=1}^{\tilde{m}_2} \frown \{f(y_i^3)\}_{i=1}^{\tilde{m}_3} \\ &= \frown_{j=1}^{3} \{f(y_i^j)\}_{i=1}^{\tilde{m}_j}. \end{aligned}$$

Let $\tilde{V}_\beta$ be the β condensation operator (see (9)) with p_3 rows and $\tilde{m}$ columns. Define the following two matrices.

$$(W)_{ij} := \begin{cases} \sqrt{\frac{2(1+\epsilon)R}{p_1}} & i=j, i \in \{1,2,\cdots,p_1\}, \\ \sqrt{\frac{2\epsilon R}{p_2-p_1}} & i=j, i \in \{p_1+1,\cdots,p_2\}, \\ \sqrt{\frac{2\epsilon R}{p_3-p_2}} & i=j, i \in \{p_2+1,\cdots,p_3\}, \\ 0 & i \neq j. \end{cases}$$

$$(\Phi)_{ij} = \begin{cases} \epsilon_i^1 & i=j, i \in \{1,\cdots,\tilde{m}_1\}, \\ \epsilon_{i-\tilde{m}_1}^2 & i=j, i \in \{\tilde{m}_1+1,\cdots,\tilde{m}_1+\tilde{m}_2\}, \\ \epsilon_{i-\tilde{m}_1-\tilde{m}_2}^3 & i=j, i \in \{\tilde{m}_1+\tilde{m}_2+1,\cdots,\tilde{m}\}, \\ 0 & i \neq j. \end{cases}$$

And, let $h_j := \sum_{i=1}^{\tilde{m}_1} (W\tilde{V}_\beta)_{ji} \epsilon_i^1 k_{y_i^1} + \sum_{i=1}^{\tilde{m}_2} (W\tilde{V}_\beta)_{j(\tilde{m}_1+i)} \epsilon_i^2 k_{y_i^2} + \sum_{i=1}^{\tilde{m}_3} (W\tilde{V}_\beta)_{j(\tilde{m}_1+\tilde{m}_2+i)} \epsilon_i^3 k_{y_i^3}$ for all $j \in \{1, \cdots p_3\}$, then we prove the following lemma.

Lemma 3.4: Let $\gamma, t \in (0,1)$ be such that $1-\gamma-3t > 0$ and ϵR be large enough, then

$$\begin{aligned} \frac{\|\nu_\beta\|_2^2}{\|\nu_\beta\|_1^2}(1-\gamma-3t)\|f\|_{L^2(\mathbb{R})}^2 &\leq \sum_{j=1}^{p_3} |\langle f, h_j \rangle|^2 \\ &\leq \frac{\|\nu_\beta\|_2^2}{\|\nu_\beta\|_1^2}(1+3t)\|f\|_{L^2(\mathbb{R})}^2 \quad \forall f \in V^{\left(1+\frac{5}{2}\epsilon\right)R}(g) \end{aligned} \tag{14}$$

holds with probability greater than $1 - 6\exp\left(-\frac{m\epsilon}{12(1+3\epsilon)}\right) - 5p\exp\left(-\frac{t^2(\beta-1)}{42(2\lambda-1)(\beta+1)R}\left(\frac{p}{2(1+3\epsilon)} - \frac{1}{\epsilon R}\right)\right)$.

Using Lemma 3.4, we conclude that with probability greater than $1 - 6\exp\left(-\frac{m\epsilon}{12(1+3\epsilon)}\right) - 5p\exp\left(-\frac{t^2(\beta-1)}{42(2\lambda-1)(\beta+1)R}\left(\frac{p}{2(1+3\epsilon)} - \frac{1}{\epsilon R}\right)\right)$, the following hold.

- The collection $\{h_j\}_{j=1}^{p_3}$ forms a frame [5] for $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$.

- The frame operator S [5] corresponding to $\{h_j\}_{j=1}^{p_3}$, is invertible. Let $F_{W\tilde{V}_\beta}$ be the operator from $\mathbb{C}^{\tilde{m}}$ to $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$ defined as

$$F_{W\tilde{V}_\beta}(c) = \sum_{j=1}^{p_3} \left(W\tilde{V}_\beta c\right)_j S^{-1}h_j. \tag{15}$$

Then $F_{W\tilde{V}_\beta}$ satisfies $F_{W\tilde{V}_\beta}\Phi E = I$, where I is the identity operator on $V^{\left(1+\frac{5}{2}\epsilon\right)R}(g)$.

3) Quantization and error bound computation: Let $f \in C_{[-\pi,\pi]}$ and $\tilde{f} := Pf$. Define,

1) $\tilde{y} := \Phi E\tilde{f} = \frown_{j=1}^{3} \{\epsilon_i^j \tilde{f}(y_i^j)\}_{i=1}^{\tilde{m}_j}$,
2)
$$y := \frown_{j=1}^{3} \{\epsilon_i^j f(y_i^j)\}_{i=1}^{\tilde{m}_j}, \tag{16}$$
3) $e := y - \tilde{y}$.

We quantize y using the distributed noise shaping quantizer $\mathcal{Q}_\beta$. Therefore, $y - q = H_\beta u$. Consequently, it can be shown that with probability greater than $1 - 6\exp\left(-\frac{m\epsilon}{12(1+3\epsilon)}\right) - 5p\exp\left(-\frac{t^2(\beta-1)}{42(2\lambda-1)(\beta+1)R}\left(\frac{p}{2(1+3\epsilon)} - \frac{1}{\epsilon R}\right)\right)$ we have

$$\begin{aligned}
&\left\|\tilde{f} - F_{W\tilde{V}_\beta}q\right\|_{L^2[-R,R]} \\
&= \left\|F_{W\tilde{V}_\beta}\Phi E\tilde{f} - F_{W\tilde{V}_\beta}q\right\|_{L^2[-R,R]} \\
&= \|F_{W\tilde{V}_\beta}(\tilde{y} - q)\|_{L^2[-R,R]} \\
&= \|F_{W\tilde{V}_\beta}(y - e - q)\|_{L^2[-R,R]} \\
&= \|F_{W\tilde{V}_\beta}(H_\beta u - e)\|_{L^2[-R,R]} \\
&= \|F_{W\tilde{V}_\beta}H_\beta u\|_{L^2[-R,R]} + \|F_{W\tilde{V}_\beta}e\|_{L^2[-R,R]} \\
&\leq \frac{\|\nu_\beta\|_1}{\|\nu_\beta\|_2\sqrt{1-\gamma-3t}}\sqrt{\frac{2}{\frac{p}{2(1+\epsilon)R} - \frac{1}{\epsilon R}}}\|\tilde{V}_\beta H_\beta\|_{\infty\to 2}\|u\|_\infty \\
&+ \|F_{W\tilde{V}_\beta}e\|_{L^2[-R,R]}.
\end{aligned} \tag{17}$$

The first term in (17) is bounded using (10). It gives us exponential decay in $\frac{m}{p}$. Further, the second term can be bounded by using the localization property of the frame $\{h_j\}_{j=1}^{p_3}$, and the fact that by construction, the approximation error $|f(y_i^j) - \tilde{f}(y_i^j)|$ is small if $y_i^j \in [-R, R]$. However, showing this involves a significant amount of calculation. Next, use the triangle inequality to bound $\|f - F_{W\tilde{V}_\beta}q\|_{L^2[-R,R]}$. In the end, choose $\epsilon = \frac{m^{\frac{1}{16}}}{R}, \lambda = 2, t = \frac{1}{12m^{\frac{1}{4}}}$, $\gamma = \frac{1}{8m^{\frac{1}{4}}}$ and $p = \frac{m}{m^{\frac{1}{16}}} = m^{\frac{15}{16}}$ to prove the result.

IV. Numerical experiments

In order to test the accuracy of Theorem 3.1, we run numerical experiments. Let w be defined as

$$w(\xi) = \begin{cases} e^{-\frac{1}{\xi}} & \xi > 0, \\ 0 & \xi \leq 0. \end{cases}$$

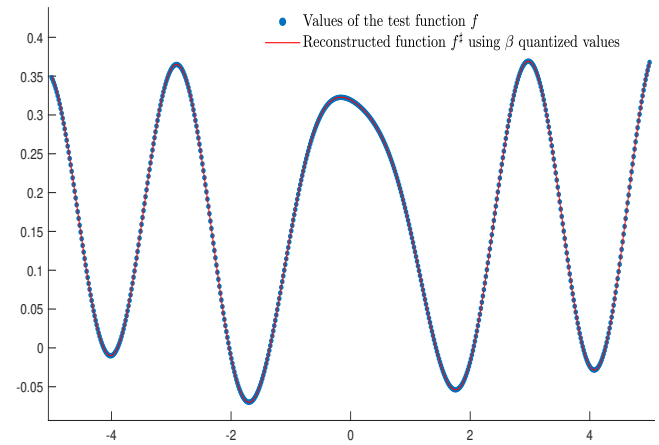

Fig. 1. Plot of f along with $f^\sharp$ in the interval $[-5, 5]$. The greedy quantizer defined in Lemma 2.3 is used to quantize the samples, with the noise transfer operator taken as $H_\beta = H_5$, i.e. $\beta = 5$ and the quantization alphabet as $\mathcal{A}_{10}^{0.1}$.The value of p is fixed as 100.

For g, we chose the following function defined via the Fourier transform

$$\hat{g}(\xi) = \begin{cases} \frac{1}{\sqrt{2\lambda\pi}} & |\xi| \leq \pi, \\ \frac{1}{\sqrt{2\lambda\pi}}\cos\left(\frac{\pi}{2}\nu\left(\frac{|\xi|-\pi}{(2\lambda-2)\pi}\right)\right) & \pi < |\xi| \leq (2\lambda-1)\pi, \\ 0 & (2\lambda-1)\pi < |\xi|. \end{cases}$$

Here $\nu(\xi) := \frac{w(\xi)}{w(\xi)+w(1-\xi)}$ $\forall\, \xi \in \mathbb{R}$ and $\lambda = 2$. We run our experiment in two parts. In the first part, we show visually that we achieve good reconstruction using our method. Taking 1200 random samples from the interval $\left[-\frac{25}{2}, \frac{25}{2}\right]$, in Fig 1, we plot f along with the reconstructed function $f^\sharp := F_{W\tilde{V}_\beta}$ in the interval $[-5, 5]$. It is easy to see from this graph that the β quantization scheme is very effective.

Working with the same function f, in Fig 2, we plot average error $\left\|\left\{f(t_i) - f^\sharp(t_i)\right\}_{i=1}^{200}\right\|_\infty$ after ten iterations, where $\{t_i\}_{i=1}^{200}$ are evenly spaced points from $[-5, 5]$, along with the number of random samples m used to calculate $f^\sharp$. Here, like in the first part, each sample x_i is sampled according to the uniform distribution on $\left[-\frac{25}{2}, \frac{25}{2}\right]$. From the plot, we can see that the error decays initially with the increase in the number of samples; however, after a certain stage, it stagnates. This may be because the projection error $\|f - Pf\|_{L^2[-5,5]}$ does not decrease with increasing samples.

References

[1] J. J. Benedetto, A. M. Powell, and Ö Yilmaz, *Sigma-Delta* ($\Sigma\Delta$) *quantization and finite frames*, IEEE Trans. Inform. Theory **52** (2006), no. 5, 1990–2005. MR 2234460

[2] J. Blum, M. Lammers, A. M. Powell, and Ö. Yilmaz, *Sobolev duals in frame theory and sigma-delta quantization*, J. Fourier Anal. Appl. **16** (2010), no. 3, 365–381. MR 2643587

[3] E. Chou and C. S. Güntürk, *Distributed noise-shaping quantization: I. Beta duals of finite frames and near-optimal quantization of random measurements*, Constr. Approx. **44** (2016), no. 1, 1–22. MR 3514402

[4] ———, *Distributed noise-shaping quantization: II. Classical frames*, Excursions in harmonic analysis. Vol. 5, Appl. Numer. Harmon. Anal., Birkhäuser/Springer, Cham, 2017, pp. 179–198. MR 3699683

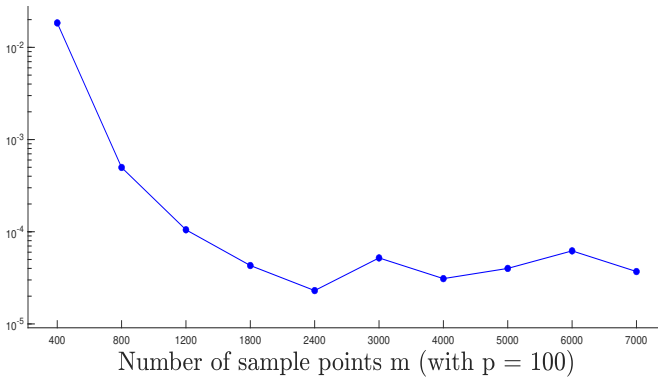

Fig. 2. The test signal f, the quantizer and the quantization alphabet are the same as in Fig 1. We plot the reconstruction error along with the sample size m.

[5] O. Christensen, *An introduction to frames and Riesz bases*, second ed., Applied and Numerical Harmonic Analysis, Birkhäuser/Springer, [Cham], 2016. MR 3495345

[6] Z. Cvetković, I. Daubechies, and B. F. Logan, Jr., *Single-bit oversampled A/D conversion with exponential accuracy in the bit rate*, IEEE Trans. Inform. Theory **53** (2007), no. 11, 3979–3989. MR 2446550

[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*, Ann. of Math. (2) **158** (2003), no. 2, 679–710. MR 2018933

[8] P. Deift, C. S. Güntürk, and F. Krahmer, *An optimal family of exponentially accurate one-bit sigma-delta quantization schemes*, Comm. Pure Appl. Math. **64** (2011), no. 7, 883–919. MR 2828585

[9] V. K. Goyal, M. Vetterli, and N. T. Thao, *Quantized overcomplete expansions in* $\mathbf{R}^N$*: analysis, synthesis, and algorithms*, IEEE Trans. Inform. Theory **44** (1998), no. 1, 16–31. MR 1486646

[10] C. S. Güntürk, *One-bit sigma-delta quantization with exponential accuracy*, Comm. Pure Appl. Math. **56** (2003), no. 11, 1608–1630. MR 1995871

[11] C. S. Güntürk, M. Lammers, A. M. Powell, R. Saab, and Ö. Yilmaz, *Sobolev duals for random frames and* $\Sigma\Delta$ *quantization of compressed sensing measurements*, Found. Comput. Math. **13** (2013), no. 1, 1–36. MR 3009528

[12] C. S. Güntürk and W. Li, *Quantization for spectral super-resolution*, Constr. Approx. **56** (2022), no. 3, 619–648. MR 4519592

[13] T. Huynh and R. Saab, *Fast binary embeddings and quantized compressed sensing with structured matrices*, Comm. Pure Appl. Math. **73** (2020), no. 1, 110–149. MR 4033891

[14] H. Inose and Y. Yasuda, *A unity bit coding method by negative feedback*, Proceedings of the IEEE **51** (1963), no. 11, 1524–1535.

[15] R. Joy, F. Krahmer, A. Lupoli, and R. Ramakrishnan, *On the reconstruction of bandlimited signals from random samples quantized via noise-shaping*, 2023, https://doi.org/10.48550/arXiv.2306.15758.

[16] F. Krahmer, R. Saab, and R. Ward, *Root-exponential accuracy for coarse quantization of finite frame expansions*, IEEE Trans. Inform. Theory **58** (2012), no. 2, 1069–1079. MR 2918010

[17] F. Krahmer, R. Saab, and Ö. Yilmaz, *Sigma-Delta quantization of sub-Gaussian frame expansions and its application to compressed sensing*, Inf. Inference **3** (2014), no. 1, 40–58. MR 3311448

[18] M. Lammers, A. M. Powell, and Ö. Yilmaz, *Alternative dual frames for digital-to-analog conversion in sigma-delta quantization*, Adv. Comput. Math. **32** (2010), no. 1, 73–102. MR 2574568

[19] A. M. Powell, J. Tanner, Y. Wang, and Ö. Yilmaz, *Coarse quantization for random interleaved sampling of bandlimited signals*, ESAIM Math. Model. Numer. Anal. **46** (2012), no. 3, 605–618. MR 2877367