THE PERFORMANCE OF PCM QUANTIZATION UNDER TIGHT FRAME REPRESENTATIONS

YANG WANG AND ZHIQIANG XU

Abstract. In this paper, we study the performance of the PCM scheme with linear quantization rule for quantizing finite unit-norm tight frame expansions for $\mathbb{R}^d$ and derive the PCM quantization error *without White Noise Hypothesis*. Using the machine developed by Güntürk, we show that the upper bound of PCM quantization error is $\delta^{3/2}$ when N is big enough, where δ is the step size and N is the frame size. When $d = 2$, using tools of harmonic analysis, we prove that the bound $\delta^{3/2}$ is sharp on average for any unit-norm tight frame. We extend the result to high dimension and show that the upper bound of PCM quantization error is $\delta^{(d+1)/2}$ for the equidistributed unit-norm tight frame of $\mathbb{R}^d$. The results of this paper imply that the performance of PCM scheme depends heavily on the choose of finite unit-norm tight frames.

1. Introduction

In signal processing, coding and many other practical applications it is important to find a suitable representation for a given signal. In general, the first step towards this objective is finding an atomic decomposition of the signal using a given set of *atoms*, or *dictionary*. In this approach, we assume that the signal x is an element of a finite-dimensional Hilbert space $H = \mathbb{R}^d$ and x is represented as a linear combination of $\{e_j\}_{j=1}^N$, i.e.,

$$x = \sum_{j=1}^{N} c_j e_j, \tag{1}$$

where c_j are real numbers. In practical application, instead of a true basis, $\{e_j\}_{j=1}^N$ is chosen to be a *frame*. Given $\mathcal{F} = \{e_j\}_{j=1}^N$, we let $F = [e_1, \ldots, e_N]$ be the corresponding matrix whose columns are $\{e_j\}_{j=1}^N$. We say $\mathcal{F}$ is a *frame* of $\mathbb{R}^d$ if the matrix F has rank d. The frame $\mathcal{F}$ is *tight* with frame constant λ if $FF^* = \lambda I_d$. The matrix F^T as an operator $F^* : \mathbb{R}^d \to \mathbb{R}^N$ is often known as the *analysis operator* with respect to the frame $\mathcal{F}$, where $(F^*x)_j = \langle x, e_j\rangle$. The adjoint operator given by $F : \mathbb{R}^N \to \mathbb{R}^d$, $Fy = \sum_{j=1}^N y_j e_j$ is known as the *synthesis operator* with respect to $\mathcal{F}$. We call the operator $S := FF^*$ as the *frame operator*. Then $\{S^{-1}e_j\}_{j=1}^N$ is called the *canonical dual frame* of the frame $\mathcal{F}$. It is easy to

Yang Wang was supported in part by the National Science Foundation grant DMS-0456538. Zhiqiang Xu was supported by NSFC grant 10871196 and by the Funds for Creative Research Groups of China (Grant No. 11021101).

see that for any $x \in \mathbb{R}^d$ we have the reconstruction formula

$$x = \sum_{j=1}^{N} \langle x, e_j \rangle \, (S^{-1} e_j) = \sum_{j=1}^{N} \left\langle x, S^{-1} e_j \right\rangle e_j. \tag{2}$$

If $\mathcal{F}$ is a tight frame with frame bound λ then clearly $S^{-1} e_j = \lambda^{-1} e_j$. In particular, for the important case of *finite unit-norm tight frames* in which $\|e_j\| = 1$ for all j we have $\lambda = N/d$ and (2) is reduced to

$$x = \frac{d}{N} \sum_{j=1}^{N} \langle x, e_j \rangle \, e_j \qquad \text{for all } x \in \mathbb{R}^d.$$

In the digital domain the representation must be quantized. In other words the coefficients $\langle x, e_j \rangle$ from the analysis operator must be mapped to a discrete set of values $\mathcal{A}$ called the *quantization alphabet*. The simplest way for such a mapping is the *Pulse Code Modulation (PCM)* quantization scheme, which has $\mathcal{A} = \delta\mathbb{Z}$ with $\delta > 0$ and maps a value t the value in $\mathcal{A}$ that is the closest to t. More precisely, the mapping is done by the function

$$Q_\delta(t) \;:=\; \text{argmin}_{r \in \mathcal{A}} |t - r| \;=\; \delta \left\lfloor \frac{t}{\delta} + \frac{1}{2} \right\rfloor$$

and the quantization function Q_δ is called the *quantizer*. Thus in practical applications we in fact have only a quantized representation through the quantized analysis operator $\tilde{y}_j := Q_\delta((F^* x)_j) = Q_\delta(\langle x, e_j \rangle), j = 1, \ldots, N$ for each $x \in \mathbb{R}^d$. The reconstruction through the frame operator yields

$$\tilde{x} \;=\; \sum_{j=1}^{N} Q_\delta(\langle x, e_j \rangle)(S^{-1} e_j) \qquad \text{for all } x \in \mathbb{R}^d.$$

Naturally we may want to ask about the error for this reconstruction.

An important class of frames is the unit-norm tight frames. This paper shall focus on this class of frames, although the questions we raise and the techniques we use can be applied to other frames. Let $\mathcal{F} = \{e_j\}_{j=1}^N$ be a unit-norm tight frame in $\mathbb{R}^d$. For each $x \in \mathbb{R}^d$ we have

$$x = \frac{d}{N} \sum_{j=1}^{N} c_j e_j, \qquad \text{where} \qquad c_j = \langle x, e_j \rangle \, . \tag{3}$$

With PCM quantization and quantization alphabet $\mathcal{A} = \delta\mathbb{Z}$ the reconstruction becomes

$$\tilde{x}_{\mathcal{F}} \;=\; \frac{d}{N} \sum_{j=1}^{N} q_j e_j, \qquad \text{where} \qquad q_j = Q_\delta(c_j) \in \mathcal{A}. \tag{4}$$

Under this quantization we denote the reconstruction error by

$$E_\delta(x, \mathcal{F}) \;:=\; \|x - \tilde{x}_{\mathcal{F}}\|$$

where $\|\cdot\|$ is ℓ_2 norm. An important question is how $E_\delta(x, \mathcal{F})$ behaves for a given frame $\mathcal{F}$ and either for a given x or for a given distribution of x. To simplify the problem, the so-called *White Noise Hypothesis* (WNH) is employed by engineers and mathematicians in

this area (see [7, 1, 4, 5, 6, 11]). The WNH asserts that the quantization error sequence $\{x_j - q_j\}_{j=1}^N$ can be modeled as an independent sequence of i.i.d. random variables that are uniformly distributed on the interval $(-\delta/2, \delta/2)$. With the WNH, one can obtain the mean square error

$$MSE \;=\; \mathcal{E}(\|x - \tilde{x}_{\mathcal{F}}\|^2) \;=\; \frac{d^2\delta^2}{12N}.$$

It has been shown that the WNH is asymptotically correct for fine quantizations (i.e. as δ tends to 0) under rather general conditions, see [6, 11]. Although the result implies that the MSE decreases on the order of $1/N$, this is in fact quite misleading because the WNH holds only asymptotically when the frame $\mathcal{F}$ (and hence N) is fixed while δ decreases to 0, and with a fixed δ WNH cannot hold whenever $N > d$ [6]. Furthermore, the MSE only gives information about the average behavior of quantization errors. There has not been an in-depth study on the behavior of the error $E_\delta(x, \mathcal{F})$ for a given x and as one fixes δ. This contrast sharply with the study on the quantization error from the Sigma-Delta quantization schemes, where the quantization step δ is typically assumed to be fixed and rather coarse, see e.g. [4, 5]. One of the objectives of this paper is to study the behavior of $E_\delta(x, \mathcal{F})$ as we choose different unit-norm tight frames $\mathcal{F}$.

It is well known that with the Sigma-Delta quantization schemes the reconstruction error will diminish to 0 as we increase the redundancy of the frame $\mathcal{F}$. For unit-norm tight frames it means that for fixed δ by letting $N \to \infty$ the reconstruction error tends to 0, even when the quantization is coarse in the sense that $\delta \gg 0$. One naturally asks whether similar phenomenon also occurs with PCM quantizations, i.e. how much can we mitigate the reconstruction error $E_\delta(x, \mathcal{F})$ if we increase the redundancy of the frame $\mathcal{F}$, and is it possible that by increasing redundancy in a suitable way the reconstruction error $E_\delta(x, \mathcal{F})$ be made arbitrarily small for all x?

In this paper we attempt a more in-depth study of the PCM quantization error $E_\delta(x, \mathcal{F})$ with respect to unit-norm tight frames $\mathcal{F}$. In particular we study the asymptotic behavior of $E_\delta(x, \mathcal{F})$ as we increase the redundancy of the unit-norm tight frame. A surprising result (at least to us) is that *in general* the quantization error $E_\delta(x, \mathcal{F})$ does not diminish to 0 no matter how much one increases the redundancy of the frame $\mathcal{F}$. The following example is a good illustration.

Example 1.1. We choose $x_0 = (\pi, e)^T$, $\delta = 1/16$ and $\mathcal{F} = \{e_j\}_{j=1}^N$ with $e_j = (\cos(\frac{2j\pi}{N}), \sin(\frac{2j\pi}{N}))^T$, $j = 0, \ldots, N-1$, which is a 2-dimensional unit-norm tight frame. Then we compute $E_\delta(x_0, \mathcal{F})$ for $N = 10, \ldots, 2000$ and show the result in Figure 1. From the figure one can see that although as we increase N the quantization error $E_\delta(x_0, \mathcal{F})$ decreases initially, it settles down to around positive value no matter how much redundancy is increased. Thus PCM quantization fails to take advantage of redundance.

Of particular interest to this study is a very popular class of unit-norm tight frames known as the *harmonic frames*. For any $N \geq d$ the harmonic frame $\mathcal{H}_N^d = \{h_j^N\}_{j=0}^{N-1}$ is

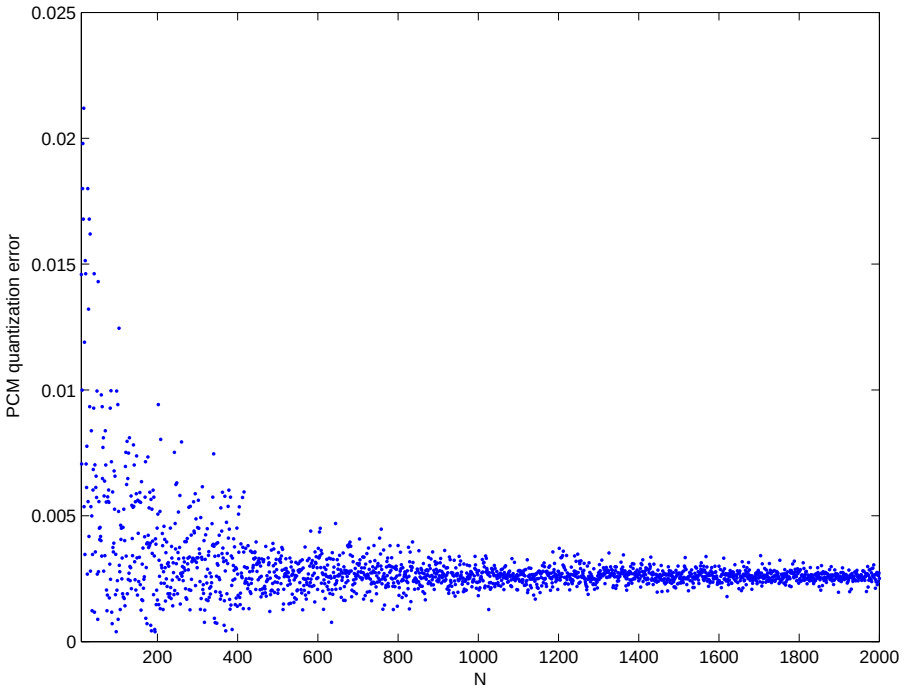

FIGURE 1. The frame expansion of $(\pi, e)^T \in \mathbb{R}^2$ with respect to the frame $(\cos(2\pi j/N), \sin(2\pi j/N))^T$ are quantized using the PCM scheme with $\delta = 1/16$. The figure shows the PCM error against the frame size N.

given by

$$h_j^N = \sqrt{\frac{2}{d}} \left[\cos \frac{2\pi j}{N}, \sin \frac{2\pi j}{N}, \cos \frac{2\pi 2j}{N}, \sin \frac{2\pi 2j}{N}, \ldots, \cos \frac{2\pi \tilde{d} j}{N}, \sin \frac{2\pi \tilde{d} j}{N} \right]^T,$$

if $d = 2\tilde{d}$ is even or

$$h_j^N = \sqrt{\frac{2}{d}} \left[\frac{1}{\sqrt{2}}, \cos \frac{2\pi j}{N}, \sin \frac{2\pi j}{N}, \ldots, \cos \frac{2\pi \tilde{d} j}{N}, \sin \frac{2\pi \tilde{d} j}{N} \right]^T,$$

if $d = 2\tilde{d} + 1$ is odd. Harmonic frames themselves are a special case of unit-norm tight frames obtained from uniform frame paths introduced in [2]. Let $f : [0, 1] \longrightarrow \mathbb{R}^d$ be a continuous function with $\|f(t)\| = 1$ for all t. It is called a *uniform frame path* if for any $N \geq d$ the set of vectors $\{f(\frac{j-1}{N})\}_{j=1}^N$ is a unit-norm tight frame in $\mathbb{R}^d$. So the harmonic frame $\mathcal{H}_N^d$ is obtained simply by taking N samples of the frame path

$$h(t) = \sqrt{\frac{2}{d}} \Big[\cos(t), \sin(t), \cos(2t), \sin(2t), \ldots, \cos(\tilde{d}t), \sin(\tilde{d}t) \Big]^T$$

if $d = 2\tilde{d}$ is even or

$$h(t) = \sqrt{\frac{2}{d}}\Big[\frac{1}{\sqrt{2}}, \cos(t), \sin(t), \cos(2t), \sin(2t), \ldots, \cos(\tilde{d}t), \sin(\tilde{d}t)\Big]^T$$

if $d = 2\tilde{d} + 1$ is odd. We examine the limitations of uniform frame paths in terms of its ability to mitigate quantization errors with increasing redundancies.

Throughout the paper we shall use the notation $X \ll_{a,b,\ldots} Y$ to refer to the inequality $X \leq C \cdot Y$, where the constant C may depend on $a, b, \ldots$, but no other variable. We now state our main results in this paper.

Theorem 1.2. *Let $f : [0,1] \longrightarrow \mathbb{R}^d$ be a uniform frame path with bounded f' and set $\mathcal{F} := \{f(\frac{j-1}{N})\}_{j=1}^N$. Suppose that $x \in \mathbb{R}^d$ and $h(t) := \langle x, f(t)\rangle$ is an entire function. Assume that $h''(t)$ has finitely many zeros in $[0,1]$ and on which $h'''(t)$ does not vanish. Then*

$$E_\delta(x, \mathcal{F}) \ \ll_x \ \sqrt{\frac{\delta}{N}}$$

for $N \leq 1/\delta^2$ and

$$E_\delta(x, \mathcal{F}) \ \ll_x \ \delta^{3/2}$$

for $N > 1/\delta^2$.

The above theorem shows that $\liminf_{\#\mathcal{F}\to\infty} E_\delta(x, \mathcal{F}) \ll_x \delta^{3/2}$ for unit-norm tight frames obtained through frame paths. The question is whether $O(\delta^{3/2})$ is sharp. Our next result shows that in $\mathbb{R}^2$ the bound $O(\delta^{3/2})$ is sharp for *all* unit-norm tight frames. As a result PCM can only partially take advantage of the redundancies in PCM quantization. We prove the result by showing that the average quantization error for any unit-norm tight frame in $\mathbb{R}^2$ is bounded from below by $O(\delta^{3/2})$. Set

$$\mathbb{E}_\delta(r, \mathcal{F}) \ := \ \left(\int_0^{2\pi} |E_\delta(x_\psi, \mathcal{F})|^2 d\psi\right)^{1/2},$$

where $x_\psi := r(\cos\psi, \sin\psi)^T$ and $r > 0$. Then

Theorem 1.3. *Set $R := r/\delta$ and $\varepsilon := R + 1/2 - \lfloor R + 1/2\rfloor$.*

(i) *Suppose $\mathcal{F}$ is an unit-norm tight frame in $\mathbb{R}^2$. If $\varepsilon = 0$, then*

$$\mathbb{E}_\delta(r, \mathcal{F}) \geq \frac{32}{3\pi^{5/2}} \frac{\delta^{3/2}}{\sqrt{r}}. \tag{5}$$

(ii) *Suppose $\mathcal{F}$ is an unit-norm tight frame in $\mathbb{R}^2$. There exists a $\varepsilon_0 > 0$ such that, when $\varepsilon \in [0, \varepsilon_0]$,*

$$\mathbb{E}_\delta(r, \mathcal{F}) \geq C \frac{\delta^{3/2}}{\sqrt{r}}$$

provided R is big enough, where C is a fixed constant.

(iii) *Suppose that* $\tilde{\mathcal{F}} = \{e_j\}_{j=1}^N$ *with* $e_j = [\cos(2j\pi/N), \sin(2j\pi/N)]^T$. *If we choose*

$$\delta = \frac{r\pi}{\sqrt{8 - 2\sqrt{16 - \pi^2}}},$$

then

$$\mathbb{E}_\delta(r, \tilde{\mathcal{F}}) = O(\frac{1}{N})$$

with $N = \#\tilde{\mathcal{F}}$.

The above theorem shows that for any $\delta > 0$ there exists a $r_0 > 0$ such that $\mathbb{E}_\delta(r_0, \tilde{\mathcal{F}}) \gg \delta^{3/2}$, which implies that the bound $O(\delta^{3/2})$ is sharp. However, by (iii) in Theorem 1.3, for each $r > 0$, $\mathbb{E}_\delta(r, \tilde{\mathcal{F}})$ tends to 0 with the speed $1/N$ if $\delta = r\pi/\sqrt{8 - 2\sqrt{16 - \pi^2}}$. The result implies that, if the norm of $x \in \mathbb{R}^2$ is fixed, one can choose a δ such that PCM can take the advantage of the redundance. In fact, according to the proof of Theorem 1.3, if δ satisfies

$$\int_{-\pi}^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta = 0, \tag{6}$$

then $\mathbb{E}_\delta(r, \tilde{\mathcal{F}}) = O(\frac{1}{N})$, where $\Delta_\delta(t) := t - Q_\delta(t)$. It will be an interesting problem to give the sufficient and necessary condition for the valid of (6).

An immediate consequence is that the bound $O(\delta^{3/2})$ is sharp for harmonic frames in $\mathbb{R}^d$, as any harmonic frame in $\mathbb{R}^d$ has a 2-dimensional harmonic frame imbedded in it, and the lower bound applies to this imbedded 2-dimensional harmonic frame.

Given the limitation of harmonic frames in mitigating PCM quantization errors one naturally asks whether the error bound $O(\delta^{3/2})$ can be improved. It turns out that this is possible if we distribute the frame elements more evenly on the unit sphere $\mathbb{S}^{d-1}$. A sequence of finite sets $A_n \subset \mathbb{S}^{d-1}$ with cardinality $N_n = \#A_n$ is said to be *asymptotically equidistributed* on $\mathbb{S}^{d-1}$ if for any piecewise continuous function f on $\mathbb{S}^{d-1}$ we have

$$\lim_{n\to\infty} \frac{1}{N_n} \sum_{v\in A_n} f(v) \;=\; \int_{z\in\mathbb{S}^d} f(z)d\nu,$$

where f are piecewise continuous functions on $\mathbb{S}^d$ and $d\nu$ denotes the normalized Lebesgue measure on $\mathbb{S}^{d-1}$. We have

Theorem 1.4. *Let* $\mathcal{F}_n$ *be a unit-norm tight frame in* $\mathbb{R}^d$. *Assume that* $\mathcal{F}_n$ *are asymptotically equidistributed on* $\mathbb{S}^{d-1}$. *Then for any* $x \in \mathbb{R}^d$ *we have*

$$\lim_{n\to\infty} E_\delta(x, \mathcal{F}_n) \ll_d \delta^{(d+1)/2}.$$

Asymptomatically equidistributed unit-norm tight frames in $\mathbb{R}^d$ can be obtained via the spherical t-design [10] and other methods. We conjecture that the bound $\delta^{(d+1)/2}$ is sharp for any unit-norm tight frame in $\mathbb{R}^d$. If the conclusion holds, it implies that asymptotically equidistributed unit-norm tight frames in $\mathbb{R}^d$ are optimal unit-norm tight frame for PCM quantization.

Xu: I believe that the constant should be in the form of $2^{d/2}$ *in Theorem 1.4. However, it is non-trival to estimate the constant. Anyway, I suggest to remove*

the following paragraph though I believe that it is true: Interestingly, a corollary of Theorem 1.4 is that there indeed exist tight frames in $\mathbb{R}^d$ (but not unit-norm) such that by increasing redundancy the PCM quantization error will decay to 0. The idea is that by the above theorem we can take $\mathcal{F}_n$ in $\mathbb{R}^n$ such that

$$E_\delta(x, \mathcal{F}_n) \leq C\delta^{(n+1)/2}.$$

Now, project $\mathcal{F}_n$ onto $\mathbb{R}^d$ we still have a tight frame $\overline{\mathcal{F}}_n$ in $\mathbb{R}^d$ (but usually no longer unit-norm). Then for any $x \in \mathbb{R}^d$ we will have

$$E_\delta(x, \overline{\mathcal{F}}_n) \leq C\delta^{(n+1)/2} \longrightarrow 0.$$

The paper is organized as follows. After introducing some preliminaries in Section 2, we give an up bound of $E_\delta(x, \mathcal{F})$ under the WNH, which is valid with high probability, in Section 3. We present the proof of Theorem 1.2 in Section 4. The proof of Theorem 1.3 is given in Section 5. We finally give the proof of Theorem 1.4 in Section 6.

2. Preliminaries

Hoeffding's inequality [9]**.** Let $X_1, \ldots, X_N$ be independent random variables. Assume that for $1 \leq j \leq N$, $\Pr(X_j - \mathcal{E}(X_j) \in [a_j, b_j]) = 1$. Then for the sum of variables,

$$S \;=\; X_1 + \cdots + X_N$$

we have the inequality

$$\Pr(|S - \mathcal{E}(S)| \geq t) \;\leq\; 2\exp\left(-\frac{2t^2}{\sum_{j=1}^N (b_j - a_j)^2}\right),$$

which are valid for positive values of t.

Discrepancy and uniform distribution (see also [8]**).** Let $\{u_j\}_{j=1}^N$ be a set of points in $[-1/2, 1/2)$ identified with the 1-torus $\mathbb{T}$. The *discrepancy* of $\{u_j\}_{j=1}^N$ is defined by

$$\mathrm{Disc}(\{u_j\}_{j=1}^N) \;:=\; \sup_{I \subset \mathbb{T}} \left| \frac{\#(\{u_j\}_{j=1}^N \cap I)}{N} - |I| \right|$$

where the sup is taken over all subarcs I of $\mathbb{T}$.

We also need the following two well-known results:

Theorem 2.1. *(Koksma's inequality) For any sequence of points $u_1, \ldots, u_N$ in $[-1/2, 1/2)$ and any function $f : [-1/2, 1/2) \to \mathbb{R}$ of bounded variation,*

$$\left| \frac{1}{N} \sum_{j=1}^N f(u_j) - \int_{-1/2}^{1/2} f(t)dt \right| \leq \mathrm{Var}(f) \cdot \mathrm{Disc}(\{u_j\}_{j=1}^N), \tag{7}$$

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

Theorem 2.2. *(Erdös-Turán inequality) For any sequence of points $u_1, \dots, u_N$ in $[-1/2, 1/2)$, and any positive integer K,*

$$\mathrm{Disc}(\{u_n\}_{n=1}^N) \ll \frac{1}{K} + \sum_{k=1}^{K} \frac{1}{k} \left| \frac{1}{N} \sum_{j=1}^{N} e^{2\pi i k u_j} \right|.$$

Exponential sums. By Erdös-Turán inequality, to estimate the discrepancy, we need to compute the exponential sums

$$S = \sum_{m=1}^{n} e^{2\pi i f(m)},$$

where f is a real-valued function. We shall use the *truncated Poisson formula* and *van der Corput's Lemma* to estimate S.

Theorem 2.3. *(Truncated Poisson formula) Let f be a real-valued function and suppose that f' is continuous and increasing on $[a, b]$. Put $\alpha = f'(a)$ and $\beta = f'(b)$. Then*

$$\sum_{a \le m \le b} e^{2\pi i f(m)} = \sum_{\alpha - 1 \le v \le \beta + 1} \int_a^b e^{2\pi i (f(\tau) - v\tau)} d\tau + O(\log(2 + \beta - \alpha)).$$

Lemma 2.4. *(van der Corput) Suppose ϕ is real-valued and smooth in the interval (a, b) and that $|\phi^{(r)}(t)| \ge \mu$ for all $t \in (a, b)$ and for a positive integer r. If $r = 1$, suppose additionally that ϕ' is monotonic. Then*

$$\left| \int_a^b e^{i\phi(t)} dt \right| \le C_r \mu^{-1/r},$$

where C_r is a constant depending on r.

Euler-Maclaurin formula. Suppose ϕ is smooth in the interval $[a, b]$, where a and b are integers. Then

$$\sum_{j=a}^{b} \phi(j) = \int_a^b \phi(x) dx + (\phi(a) + \phi(b))/2 + \sum_{j=2}^{p} (B_j/j!) \left(\phi^{(j-1)}(a) - \phi^{(j-1)}(b) \right) + E_p,$$

where B_j are the Bernoulli numbers and

$$|E_p| \le \frac{2}{(2\pi)^p} \int_a^b |\phi^{(p)}(x)| dx.$$

3. The error bound under the WNH

In this section, given $x \in \mathbb{R}^d$, we derive a bound for $E_\delta(x, \mathcal{F})$, which is valid with high probability, under the WNH. As a conclusion, $E_\delta(x, \mathcal{F})$ tends to 0 with probability 1 when

$\#\mathcal{F} \to \infty$. Recall that $\mathcal{F} = \{e_j\}_{j=1}^N$ is a finite tight frame in $\mathbb{R}^d$ and $F = [e_1, \ldots, e_N]$ be the corresponding matrix whose columns are $\{e_j\}_{j=1}^N$. We define the variation of F as

$$\sigma(F) := \min_p \sum_{j=1}^{N-1} \|e_{p(j)} - e_{p(j+1)}\|.$$

Then we have

Theorem 3.1. *Under the WNH, for each fixed $x \in \mathbb{R}^d$ and $\varepsilon \in (0, 1/2)$, we have*

$$\Pr\left(\|x - \tilde{x}_N\| \le \frac{d\delta}{N^{1/2-\varepsilon}}(\sigma(F) + 1)\right) \ge 1 - 2N\exp(-2N^{2\varepsilon}).$$

Proof. The WNH implies that $x_j - q_j \in [-\delta/2, \delta/2)$ and $\mathcal{E}(x_j - q_j) = 0$. To this end, we set

$$u_j := \sum_{k=1}^{j}(x_k - q_k), \qquad u_0 := 0.$$

Then, by Hoeffding's inequality, we have

$$\Pr(|u_j| \le N^{1/2+\varepsilon}\delta) \;\ge\; 1 - 2\exp(-2N^{1+2\varepsilon}/j) \;\ge\; 1 - 2\exp(-2N^{2\varepsilon}),$$

for $j = 1, \ldots, N$. We obtain that

$$\Pr\left(\bigcap_{j=1}^{N}(|u_j| \le N^{1/2+\varepsilon}\delta)\right) \;\ge\; (1 - 2\exp(-2N^{2\varepsilon}))^N \;\ge\; 1 - 2N\exp(-2N^{2\varepsilon}).$$

Noting that

$$\begin{aligned} E_\delta(x, \mathcal{F}) &= \frac{d}{N}\sum_{j=1}^{N}(x_j - q_j)e_j \;=\; \frac{d}{N}\sum_{j=1}^{N}(u_j - u_{j-1})e_j \\ &= \frac{d}{N}\left(\sum_{j=1}^{N} u_j e_j - \sum_{j=1}^{N-1} u_j e_{j+1}\right) \\ &= \frac{d}{N}\left(\sum_{j=1}^{N-1} u_j(e_j - e_{j+1}) + u_N e_N\right), \end{aligned}$$

we have

$$E_\delta(x, \mathcal{F}) \le \frac{d \cdot N^{1/2+\varepsilon}\delta}{N}(\sigma(F) + 1) = \frac{d \cdot \delta}{N^{1/2-\varepsilon}}(\sigma(F) + 1),$$

with probability $1 - 2N\exp(-2N^{2\varepsilon})$. □

4. The Proof of Theorem 1.2

Let $f : [0,1] \to \mathbb{R}^d$ be a uniform frame path and set $\mathcal{F} := \{e_j\}_{j=1}^N$ with $e_j = f(\frac{j-1}{N})$. For $x \in \mathbb{R}^d$, we use $\{c_j\}_{j=1}^N$ to denote the corresponding sequence of frame coefficients with respect to $\mathcal{F}$, i.e. $c_j = \left\langle x, f(\frac{j-1}{N}) \right\rangle$. Let $\{q_j\}_{j=1}^N$ be the quantized sequence which is obtained by PCM scheme, i.e. $q_j = Q_\delta(x_j)$. The resulting quantized expansion is

$$\tilde{x}_{\mathcal{F}} \ = \ \frac{d}{N} \sum_{j=1}^{N} q_j e_j^N.$$

We set

$$u_j := \sum_{k=1}^{j} (c_k - q_k), \ j = 1, \ldots, N, \text{ and } u_0 := 0.$$

Then we have

$$\begin{aligned} x - \tilde{x}_{\mathcal{F}} \ &= \ \frac{d}{N} \sum_{j=1}^{N} (c_j - q_j) e_j = \frac{d}{N} \sum_{j=1}^{N} (u_j - u_{j-1}) e_j \\ &= \ \frac{d}{N} \left(\sum_{j=1}^{N} u_j e_j - \sum_{j=1}^{N-1} u_j e_{j+1} \right) \\ &= \ \frac{d}{N} \left(\sum_{j=1}^{N-1} u_j (e_j - e_{j+1}) + u_N e_N \right), \end{aligned}$$

which implies that

$$\|x - \tilde{x}_{\mathcal{F}}\| \ = \ \frac{d}{N} \| \sum_{j=1}^{N-1} u_j (e_j - e_{j+1}) + u_N e_N \|. \tag{8}$$

Hence, when working with the approximation error written as (8), the main step is to find a good estimate for u_j.

Lemma 4.1. *Suppose that there exists an entire function h, such that*

$$\text{for any } d \le N \text{ and } 1 \le j \le N, \ \ c_j = h((j-1)/N).$$

Suppose furthermore that $h''(t)$ has finitely many zeros in $[0,1]$ and on which $h'''(t)$ does not vanish. Then

$$\max_{1 \le j \le N} |u_j| \ \ll_h \ \sqrt{N} \log N\delta + \sqrt{N\delta} + N\delta^{3/2}.$$

Proof. Set $y_n := x_n - q_n$ and $\tilde{y}_n := y_n/\delta = (x_n - q_n)/\delta$. Recall that we use $\mathrm{Disc}(\cdot)$ to denote the discrepancy of a sequence. Koksma's inequality implies that

$$\begin{aligned} |u_j| &= \delta \left| \sum_{n=1}^{j} \tilde{y}_n \right| = j\delta \left| \frac{1}{j} \sum_{n=1}^{j} \tilde{y}_n - \int_{-1/2}^{1/2} y dy \right| \\ &\leq j\,\delta\,\mathrm{Disc}(\{\tilde{y}_n\}_{n=1}^{j}). \end{aligned}$$

Using Erdös-Túran inequality, one has

$$\text{for any } K \in \mathbb{N},\ \mathrm{Disc}(\{\tilde{y}_n\}_{n=1}^{j}) \leq \frac{1}{K} + \frac{1}{j} \sum_{k=1}^{K} \frac{1}{k} \left| \sum_{n=1}^{j} e^{2\pi i k \tilde{y}_n} \right|.$$

Now we need to estimate

$$\left| \sum_{n=1}^{j} e^{2\pi i k \tilde{y}_n} \right|.$$

Set

$$X_N(\cdot) := h(\cdot/N). \tag{9}$$

Then we have

$$y_n = X_N(n) \text{ modulo } [-\delta/2,\, \delta/2).$$

Let $\{z_t\}_{t=1}^{n^*}$ be the set of zeros of h'' in $[0,1]$, and let $0 < \alpha < 1$ be a fixed constant to be specified later. Without loss of generality, we suppose $z_t < z_{t+1}, t = 1, \ldots, n^* - 1$. Define the intervals I_t and J_t by

$$\begin{aligned} \text{for } t = 1, \ldots, n^*, \quad I_t &= [Nz_t - N^\alpha,\ Nz_t + N^\alpha], \\ \text{for } t = 1, \ldots, n^* - 1, \quad J_t &= [Nz_t + N^\alpha,\ Nz_{t+1} - N^\alpha], \end{aligned}$$

and

$$J_0 = [1, Nz_1 - N^\alpha] \text{ and } J_{n^*} = [Nz_{n^*} + N^\alpha, N].$$

If $z_1 = 0$, we modify I_1 as $[1, N^\alpha]$ and no longer need J_0. Similarly, if $z_{n^*} = 1$, we change I_{n^*} as $[N - N^\alpha, N]$ and remove J_{n^*}. Note that

$$[1, N] \subset J_0 \cup I_1 \cup J_1 \cup \cdots \cup I_{n^*} \cup J_{n^*}.$$

Noting $h'''(z_t) \neq 0$, by Taylor expansion, we have

$$\text{for } n \in \mathbb{N} \cap J_t, \qquad \frac{1}{N^{1-\alpha}} = \frac{N^\alpha}{N} \ll_h \left| h''(\frac{n}{N}) \right|$$

provided N is large enough, which implies that

$$\text{for } n \in \mathbb{N} \cap J_t, \quad \frac{k}{N^{3-\alpha} \cdot \delta} \ll_h |\frac{k}{\delta} X_N''(n)|.$$

Since $h \in L^\infty(\mathbb{R})$, by (9), we have

$$\text{for } n \in \mathbb{N} \cap J_t, \quad |\frac{k}{\delta} X_N'(n)| \ll_h \frac{k}{N \cdot \delta}.$$

Note than X'_N is a monotonic function in J_t and set

$$\alpha_t := \min_{n\in\mathbb{N}\cap J_t} \frac{k}{\delta} X'_N(n), \qquad \beta_t := \max_{n\in\mathbb{N}\cap J_t} \frac{k}{\delta} X'_N(n).$$

Then, a simple observation is that $\beta_t - \alpha_t \ll_h \frac{2k}{N\cdot\delta}$.

Using the *truncated Poisson formula* and *van der Corput's Lemma*, we obtain that

$$\begin{aligned}
\left|\sum_{n\in\mathbb{N}\cap J_t} e^{2\pi i k\tilde{y}_n}\right| &= \left|\sum_{n\in\mathbb{N}\cap J_t} e^{2\pi i k X_N(n)/\delta}\right| \\
&\leq \sum_{\alpha_t-1\leq v\leq\beta_t+1} \left|\int_{J_t} e^{2\pi i(\frac{kX_N(\tau)}{\delta}-v\tau)}d\tau\right| + O(\log(2+\beta_t-\alpha_t)) \\
&\ll_h (\frac{2k}{N\delta}+2)\left|\int_{J_t} e^{2\pi i(\frac{kX_N(\tau)}{\delta}-v\tau)}d\tau\right| + O(\log(2+\frac{2k}{N\delta})) \\
&\ll_h \sqrt{\frac{k}{\delta}}N^{(1-\alpha)/2} + \sqrt{\frac{\delta}{k}}N^{(3-\alpha)/2} + O(\log(2+\frac{2k}{N\delta})).
\end{aligned}$$

The estimate above is also valid if we restrict $n \in [1, j]$, i.e.,

$$\left|\sum_{n\in\mathbb{N}\cap J_t\cap[1,j]} e^{2\pi i k\tilde{y}_n}\right| \ll_h \sqrt{\frac{k}{\delta}}N^{(1-\alpha)/2} + \sqrt{\frac{\delta}{k}}N^{(3-\alpha)/2} + O(\log(2+\frac{2k}{N\delta})).$$

We also have the trivial estimate:

$$\left|\sum_{n\in\mathbb{N}\cap I_t} e^{2\pi i k\tilde{y}_n}\right| \leq 2N^{\alpha}.$$

Hence, we have

$$\left|\sum_{n=1}^{j} e^{2\pi i k\tilde{y}_n}\right| \ll_h 2N^{\alpha} + \sqrt{\frac{k}{\delta}}N^{(1-\alpha)/2} + \sqrt{\frac{\delta}{k}}N^{(3-\alpha)/2} + O(\log(2+\frac{2k}{N\delta})).$$

Now, we can estimate $\mathrm{Disc}(\{\tilde{y}_n\}_{n=1}^j)$ as follows:

$$\begin{aligned}
&\text{for any } K\in\mathbb{N},\ \mathrm{Disc}(\{\tilde{y}_n\}_{n=1}^j) \ll \frac{1}{K} + \frac{1}{j}\sum_{k=1}^{K}\frac{1}{k}\left|\sum_{n=1}^{j} e^{2\pi i k\tilde{y}_n}\right| \\
\ll_h\ & \frac{1}{K} + \frac{1}{j}\sum_{k=1}^{K}\frac{1}{k}\left(2N^{\alpha} + 2\sqrt{\frac{k}{\delta}}N^{\frac{1-\alpha}{2}} + \sqrt{\frac{\delta}{k}}N^{\frac{3-\alpha}{2}} + O(\log(2+\frac{2k}{N\delta}))\right) \\
\ll_h\ & \frac{1}{K} + \frac{1}{j}\left(2N^{\alpha}\log K + 2\sqrt{\frac{K}{\delta}}N^{\frac{1-\alpha}{2}} + \sqrt{\frac{\delta}{K}}N^{\frac{3-\alpha}{2}} + O(\sum_{k=1}^{K}\frac{1}{k}\log(2+\frac{2k}{N\delta}))\right).
\end{aligned}$$

We choose $K = \lfloor \sqrt{N} \rfloor$ and $\alpha = 1/2$. Then

$$\begin{aligned} |u_j| &\leq j\,\delta\,\mathrm{Disc}(\{\tilde{y}_n\}_{n=1}^{j} \\ &\ll_h \frac{j\delta}{\sqrt{N}} + \left(\sqrt{N}\log N\delta + 2\sqrt{N\delta} + N\delta^{3/2} + O(\delta \log N \log(2+\frac{2}{\sqrt{N\delta}}))\right) \\ &\ll \frac{j\delta}{\sqrt{N}} + \left(\sqrt{N}\log N\delta + 2\sqrt{N\delta} + N\delta^{3/2}\right), \end{aligned}$$

which follows

$$\max_{1\leq j\leq N} |u_j| \ \ll_h \ \sqrt{N}\log N\delta + \sqrt{N\delta} + N\delta^{3/2}.$$

□

Proof of Theorem 1.2. Set $y_j := x_j - q_j$. We consider

$$\begin{aligned} \|x - \tilde{x}_N\| &= \|\frac{d}{N}\sum_{j=1}^{N} y_j e_j\| \\ &= \frac{d}{N}\|\sum_{j=1}^{N-1} u_j(e_j - e_{j+1}) + u_N e_N\| \ll \frac{d}{N}\max_{1\leq j\leq N}|u_j|, \end{aligned} \tag{10}$$

where the last inequality follows by $\|e_j - e_{j+1}\| = \|f(\frac{j-1}{N}) - f(\frac{j}{N})\| \ll \frac{1}{N}$ with $\|f'\|$ being bounded. Lemma 4.1 implies that

$$\frac{1}{N}\max_{1\leq j\leq N}|u_j| \ \ll_x \ \sqrt{\frac{\delta}{N}} + \frac{(\log N)\cdot\delta}{\sqrt{N}} + \delta^{3/2}. \tag{11}$$

A simple observation is that

$$\max\{\sqrt{\frac{\delta}{N}}, \frac{\delta\log N}{\sqrt{N}}, \delta^{3/2}\} \ \leq \ \sqrt{\frac{\delta}{N}} \tag{12}$$

when $N \leq \frac{1}{\delta^2}$. Combing (11) and (12), we have

$$\frac{1}{N}\max_{1\leq j\leq N}|u_j| \ \ll_x \ \sqrt{\frac{\delta}{N}}, \text{ when } N \leq \frac{1}{\delta^2},$$

which implies that

$$\|x - \tilde{x}_{\mathcal{F}}\| \ \ll_x \ \sqrt{\frac{\delta}{N}}$$

provided $N \leq \frac{1}{\delta^2}$.

We now turn to the case where $N \geq \frac{1}{\delta^2}$. To this end, in the basis of (11), we only need to prove that

$$\max\{\sqrt{\frac{\delta}{N}}, \frac{\delta\log N}{\sqrt{N}}, \delta^{3/2}\} \ \leq \ \delta^{3/2}.$$

Indeed, $N \geq 1/\delta^2$ implies that $\sqrt{\frac{\delta}{N}} \leq \delta^{3/2}$. Moreover, we have

$$\frac{\delta \log N}{\sqrt{N}} = \frac{\delta}{N^{1/4}} \frac{\log N}{N^{1/4}} \leq \frac{\delta}{N^{1/4}} \leq \delta^{3/2}$$

provided N is big enough. The claim follows. □

5. The proof of Theorem 1.3

Throughout the rest of this paper, we set

$$\Delta_\delta(t) := t - Q_\delta(t).$$

Then we have

Lemma 5.1. *Suppose that $r > 0, \delta > 0$. Set $R := r/\delta$ and $\varepsilon := R + 1/2 - \lfloor R + 1/2 \rfloor$.*

(i) *If $\varepsilon = 0$, then*

$$\frac{16\sqrt{2}}{3\pi^2} \frac{\delta^{3/2}}{\sqrt{r}} \leq \int_{-\pi}^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta.$$

(ii) *There exists a $\varepsilon_0 > 0$ such that, when $\varepsilon \in [0, \varepsilon_0]$,*

$$C_1 \frac{\delta^{3/2}}{\sqrt{r}} \leq \int_{-\pi}^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta,$$

provided R is big enough, where C_1 is a fixed constant.

Proof. We first consider the case with $\varepsilon = 0$. Note that

$$\begin{aligned} \int_{-\pi}^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta &= 2\int_0^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta \\ &= 2\int_0^{\pi} (r\cos\theta - \delta\lfloor R\cos\theta + 1/2\rfloor)\cos\theta d\theta \\ &= \pi r - 2\delta \int_0^{\pi} \lfloor R\cos\theta + 1/2\rfloor \cos\theta d\theta \\ &= \delta\left(\pi R - 2\int_0^{\pi} \lfloor R\cos\theta + 1/2\rfloor \cos\theta d\theta\right). \end{aligned} \tag{13}$$

If $R = 1/2$, then a simple calculation shows that

$$\int_{-\pi}^{\pi} \Delta_\delta(r\cos\theta)\cos\theta d\theta = \pi\delta/2$$

which implies the conclusion. To this end, we only need investigate the case $R \in \mathbb{Z}_{\geq 1} + 1/2$. We set $L := -R + 1/2 \in \mathbb{Z}$ and $U := R + 1/2 \in \mathbb{Z}$. We now consider

$$\begin{aligned}
& \int_0^\pi \lfloor R\cos\theta + 1/2 \rfloor \cos\theta d\theta \\
= & \sum_{j=L}^{U-1} j \int_{\arccos((j+1/2)/R)}^{\arccos((j-1/2)/R)} \cos\theta d\theta \\
= & \sum_{j=L}^{U-1} j(\sqrt{1-((j-1/2)/R)^2} - \sqrt{1-((j+1/2)/R)^2}) \\
= & \sum_{j=L-1}^{U-2} (j+1)\sqrt{1-((j+1/2)/R)^2} - \sum_{j=L}^{U-1} j\sqrt{1-((j+1/2)/R)^2} \\
= & \sum_{j=L}^{U-2} \sqrt{1-((j+1/2)/R)^2} = (\sum_{j=L}^{U-2} \sqrt{R^2-(j+1/2)^2})/R \\
= & \sum_{j=0}^{R-3/2} \left(2\sqrt{R^2-(j+1/2)^2}\right)/R. \qquad (14)
\end{aligned}$$

Set $\psi_j := \arccos((j+1/2)/R) - \arccos((j+3/2)/R)$. We claim that

$$\int_0^\pi \lfloor R\cos\theta + \frac{1}{2} \rfloor \cos\theta d\theta = R \sum_{j=0}^{R-3/2} \sin\psi_j + \frac{1}{2}\sqrt{1-\frac{1}{4R^2}}. \qquad (15)$$

Indeed, the claim can follow from the following calculation

$$\begin{aligned}
& R \sum_{j=0}^{R-3/2} \sin\psi_j \\
= & \frac{1}{R} \sum_{j=0}^{R-3/2} \left((j+\frac{3}{2})\sqrt{R^2-(j+\frac{1}{2})^2} - (j+\frac{1}{2})\sqrt{R^2-(j+\frac{3}{2})^2}\right) \\
= & \frac{1}{R}\left(\frac{3}{2}\sqrt{R^2-\frac{1}{4}} + \sum_{j=0}^{R-5/2} 2\sqrt{R^2-(j+\frac{3}{2})^2}\right) \\
= & \int_0^\pi \lfloor R\cos\theta + \frac{1}{2} \rfloor \cos\theta d\theta - \frac{1}{2}\sqrt{1-\frac{1}{4R^2}},
\end{aligned}$$

where the last equation follows from (14). Moreover, by Taylor expansion, we have

$$\sum_{j=0}^{R-3/2} \psi_j = \arccos(\frac{1}{2R}) \leq \frac{\pi}{2} - \frac{1}{2R},$$

which implies that

$$2R\sum_{j=0}^{R-3/2}\psi_j+1\le\pi R. \tag{16}$$

Then, combining (15) and (16), we obtain that

$$2R\sum_{j=0}^{R-3/2}(\psi_j-\sin\psi_j)\le\pi R-2\int_0^{\pi}\lfloor R\cos\theta+\frac{1}{2}\rfloor\cos\theta d\theta \tag{17}$$

Noting that $\psi_j-\sin\psi_j>0$, we have

$$\begin{aligned}&\sum_{j=0}^{R-3/2}(\psi_j-\sin\psi_j)\ge\psi_{R-3/2}-\sin\psi_{R-3/2}\\ \ge\ &\frac{4}{3\pi^2}\psi_{R-3/2}^3=\frac{4}{3\pi^2}(\arccos(1-1/R))^3\\ \ge\ &\frac{8\sqrt{2}}{3\pi^2}R^{-3/2}.\end{aligned} \tag{18}$$

Combining (13), (17) and (18), we arrive at

$$\begin{aligned}&\int_{-\pi}^{\pi}\Delta_\delta(r\cos\theta)\cos\theta d\theta\\ =\ &\delta\left(\pi R-2\int_0^{\pi}\lfloor R\cos\theta+\frac{1}{2}\rfloor\cos\theta d\theta\right)\\ \ge\ &\frac{16\sqrt{2}}{3\pi^2}\frac{\delta^{3/2}}{\sqrt{r}}.\end{aligned}$$

We next consider the case $\varepsilon\neq 0$. Using the similar method as before, we have

$$\int_0^{\pi}\lfloor R\cos\theta+\frac{1}{2}\rfloor\cos\theta d\theta=R\sum_{j=0}^{R-1/2-\varepsilon}\sin\psi_{j-1}-\frac{1}{2}\sqrt{1-\frac{1}{4R^2}}+(R+1-\varepsilon)\sqrt{\frac{2\varepsilon}{R}-(\frac{\varepsilon}{R})^2}$$

and

$$\pi R=2R\sum_{j=0}^{R-1/2-\varepsilon}\psi_{j-1}+2\sqrt{2\varepsilon R}-1+\frac{\sqrt{2}}{6}\frac{\varepsilon^{3/2}}{R^{1/2}}+O(\frac{1}{R^{3/2}}).$$

Then

$$\begin{aligned}
& \pi R - 2\int_0^{\pi} \lfloor R\cos\theta + 1/2\rfloor \cos\theta d\theta \\
= \;& 2R \sum_{j=0}^{R-1/2-\varepsilon} (\psi_{j-1} - \sin\psi_{j-1}) - 2(R+1-\varepsilon)\sqrt{\frac{2\varepsilon}{R} - (\frac{\varepsilon}{R})^2} \\
& + 2\sqrt{2\varepsilon R} + \frac{\sqrt{2}}{6}\frac{\varepsilon^{3/2}}{\sqrt{R}} + O(\frac{1}{R^{3/2}}) \\
\geq \;& 2R(\psi_{R-3/2-\varepsilon} - \sin\psi_{R-3/2-\varepsilon}) + (\frac{2\sqrt{2}}{3}\varepsilon^{3/2} - 2(1-\varepsilon)\sqrt{2\varepsilon})\frac{1}{\sqrt{R}} + O(\frac{1}{R^{3/2}}) \\
= \;& \left(\frac{(\sqrt{2+2\varepsilon} - \sqrt{2\varepsilon})^3}{3} + \frac{2\sqrt{2}}{3}\varepsilon^{3/2} - 2(1-\varepsilon)\sqrt{2\varepsilon}\right)\frac{1}{\sqrt{R}} + O(\frac{1}{R^{3/2}}). \qquad (19)
\end{aligned}$$

Note that there exists a $\varepsilon_0 > 0$ such that

$$\frac{(\sqrt{2+2\varepsilon} - \sqrt{2\varepsilon})^3}{3} + \frac{2\sqrt{2}}{3}\varepsilon^{3/2} - 2(1-\varepsilon)\sqrt{2\varepsilon}$$

is positive when $\varepsilon \in [0, \varepsilon_0]$, which implies that

$$\pi R - 2\int_0^{\pi} \lfloor R\cos\theta + 1/2\rfloor \cos\theta d\theta \geq \frac{C_1}{\sqrt{R}}$$

when R is big enough, where C_1 is a positive constant. The conclusion follows. □

Proof of Theorem 1.3. We suppose $x_{\psi_0} = r[\cos\psi_0, \sin\psi_0]^T$ and $\mathcal{F} = \{e_j\}_{j=1}^N$ with $e_j = [\cos\theta_j, \sin\theta_j]^T, \theta_j \in [0, 2\pi)$. Set $R := r/\delta$. Then

$$\begin{aligned}
E_\delta(x_{\psi_0}, \mathcal{F}) &= \frac{2}{N}\|\sum_{j=1}^N \Delta_\delta(r\cos(\theta_j - \psi_0))e_j\| \\
&= \frac{2\delta}{N}\|\sum_{j=1}^N \Delta_1(R\cos(\theta_j - \psi_0))e_j\|.
\end{aligned}$$

Let

$$P_{\psi_0} := \begin{bmatrix} \cos\psi_0 & \sin\psi_0 \\ -\sin\psi_0 & \cos\psi_0 \end{bmatrix}$$

Then a simple observation is that

$$E_\delta(x_{\psi_0}, \mathcal{F}) = \frac{2\delta}{N}\|\sum_{j=1}^N \Delta_1(R\cos(\theta_j - \psi_0))P_{\psi_0}e_j\|. \qquad (20)$$

Denote $\mu_{\mathcal{F}} = \frac{1}{N}\sum_{j=1}^N \delta_{\theta_j}$ where δ_{θ_j} is the Dirac measure at θ_j. Then we can rewrite (20) as

$$E_\delta(x_{\psi_0}, \mathcal{F}) = 2\delta\|(H_R * \mu_{\mathcal{F}})(\psi_0)\|,$$

where $*$ is the component-wise convolution operator and

$$H_R(t) := \Delta_1(R\cos t)\begin{bmatrix}\cos t\\ -\sin t\end{bmatrix}.$$

Note that

$$\begin{aligned}\mathbb{E}_\delta(r,\mathcal{F}) &= \left(\int_0^{2\pi}(E(x_\psi,\mathcal{F}))^2 d\psi\right)^{1/2}\\ &= 2\delta\|H_R*\mu_{\mathcal{F}}\|_{L^2} = \frac{\sqrt{2}\delta}{\sqrt{\pi}}(\sum_{k\in\mathbb{Z}}\|\widehat{H_R*\mu}_{\mathcal{F}}(k)\|^2)^{1/2}\\ &= \frac{\sqrt{2}\delta}{\sqrt{\pi}}(\sum_{k\in\mathbb{Z}}\|\hat{H}_R(k)\|^2|\hat{\mu}_{\mathcal{F}}(k)|^2)^{1/2},\end{aligned}$$

where $\|\cdot\|_{L^2}$ and $\|\cdot\|$ denote the L^2 norm of vector functions and ℓ^2 norm of $\mathbb{R}^2$, respectively. Since $\hat{\mu}_{\mathcal{F}}(0)=1$, it follows that

$$\begin{aligned}\mathbb{E}_\delta(r,\mathcal{F}) &= \frac{\sqrt{2}\delta}{\sqrt{\pi}}(\|\hat{H}_R(0)\|^2+\sum_{k\in\mathbb{Z}\setminus\{0\}}\|\hat{H}_R(k)\|^2|\hat{\mu}_{\mathcal{F}}(k)|^2)^{1/2}\\ &\geq \frac{\sqrt{2}\delta}{\sqrt{\pi}}\|\hat{H}_R(0)\|.\end{aligned}$$

We still need to estimate $\hat{H}_R(0)$. Note

$$\begin{aligned}\hat{H}_R(0) &= \int_0^{2\pi}H_R(t)dt = \int_0^{2\pi}\Delta_1(R\cos t)\begin{bmatrix}\cos t\\ -\sin t\end{bmatrix}dt\\ &= \begin{bmatrix}\int_0^{2\pi}\Delta_1(R\cos t)\cos t dt\\ 0\end{bmatrix}.\end{aligned}$$

By (i) in Lemma 5.1, when δ is small enough,

$$\int_0^{2\pi}\Delta_1(R\cos t)\cos t dt = \frac{1}{\delta}\int_0^{2\pi}\Delta_\delta(r\cos t)\cos t dt \geq \frac{16\sqrt{2}}{3\pi^2}\sqrt{\frac{\delta}{r}},$$

which implies that

$$\mathbb{E}_\delta(r,\mathcal{F}) \geq \frac{32}{3\pi^{5/2}}\frac{\delta^{3/2}}{\sqrt{r}}.$$

Similarly, (ii) can be proved by (ii) in Lemma 5.1.

We now turn to (iii). A simple calculation shows that

$$\int_{-\pi}^{\pi}\Delta_\delta(r\cos\theta)\cos\theta d\theta = \pi r - 4\delta\sqrt{1-\frac{\delta^2}{4r^2}}$$

provided $r\leq\delta\leq 2r$. The result implies that

$$\hat{H}_R(0)=0$$

if

$$\delta = \frac{r\pi}{\sqrt{8-2\sqrt{16-\pi^2}}}.$$

Also, note that $\hat{\mu}_{\tilde{\mathcal{F}}}(k) = \delta_{k-N}$ with $N = \#\tilde{\mathcal{F}}$. Hence,

$$\mathbb{E}_\delta(r,\mathcal{F}) = \frac{\sqrt{2}\delta}{\sqrt{\pi}} \left(\sum_{k\in\mathbb{Z}\setminus\{0\}} \|\hat{H}_R(kN)\|^2 \right)^{1/2}.$$

According to Riemann-Lebesgue Lemma,

$$\|\hat{H}_R(N)\| = O(\frac{1}{N}).$$

The conclusion follows. □

6. The proof of Theorem 1.4

To prove Theorem 1.4, we first introduce a lemma.

Lemma 6.1.

$$\left| \int_0^\pi \Delta_\delta(r\cos\theta)\cos\theta(\sin\theta)^{d-1}d\theta \right| \ll_d \ \delta^{(d+2)/2}.$$

Proof. Similar with Lemma 5.1, we set $R := r/\delta$. To state conveniently, we only consider the case $R + 1/2 \in \mathbb{Z}$. The other case can be proved by a similar method. A simple observation is that

$$\begin{aligned} &\int_0^\pi \Delta_\delta(r\cos\theta)\cos\theta(\sin\theta)^{d-1}d\theta = \\ &\delta\left(R\int_0^\pi (\cos\theta)^2(\sin\theta)^{d-1}d\theta - \int_0^\pi \lfloor R\cos\theta + \frac{1}{2}\rfloor \cos\theta(\sin\theta)^{d-1}d\theta \right). \end{aligned} \tag{21}$$

Then, we consider

$$\begin{aligned} &\int_0^\pi \lfloor R\cos\theta + \frac{1}{2}\rfloor \cos\theta(\sin\theta)^{d-1}d\theta \\ =& \sum_{j=-R+1/2}^{R-1/2} j\int_{\arccos((j+1/2)/R)}^{\arccos((j-1/2)/R)} \cos\theta(\sin\theta)^{d-1}d\theta \\ =& \frac{1}{d}\sum_{j=-R+1/2}^{R-1/2} j((1-((j-1/2)/R)^2)^{d/2} - (1-((j+1/2)/R)^2)^{d/2}) \\ =& \frac{1}{d}\sum_{j=-R+1/2}^{R-1/2} (1-((j-1/2)/R)^2)^{d/2}. \end{aligned} \tag{22}$$

As we shall see later,

$$\begin{aligned} &\frac{1}{d}\sum_{j=-R+3/2}^{R-1/2}(1-((j-1/2)/R)^2)^{d/2} \\ =\ & R\int_0^{\pi}(\cos\theta)^2(\sin\theta)^{d-1}d\theta+O(1/R^{d/2}). \end{aligned} \tag{23}$$

Then combining (21), (22) and (23), we reach the conclusion.

We remain to argue (23). We set

$$f(x)\ :=\ \frac{1}{d}\left(1-((2x-1)/(2R))^2\right)^{d/2}.$$

Then the left side of (23) equals to $\sum_{j=-R+3/2}^{R-1/2} f(j)$. Recall that Euler-Maclaurin formula

$$\begin{aligned} \sum_{j=-R+3/2}^{R-1/2} f(j) \quad = \quad & \int_{-R+3/2}^{R-1/2} f(x)dx+(f(-R+3/2)+f(R-1/2))/2 \\ & +\sum_{j=2}^{p}(B_j/j!)\left(f^{(j-1)}(-R+3/2)-f^{(j-1)}(R-1/2)\right)+E_p, \end{aligned}$$

where B_j are the Bernoulli numbers and

$$|E_p|\ \le\ \frac{2}{(2\pi)^p}\int_{-R+3/2}^{R-1/2}|f^{(p)}(x)|dx.$$

Note that

$$\int_{-R+3/2}^{R-1/2} f(x)dx=\int_{-R+1/2}^{R+1/2} f(x)dx+O(\frac{1}{R^{d/2}}).$$

Let $\theta=\arccos((2x-1)/2R)$. We have

$$\int_{-R+1/2}^{R+1/2} f(x)dx\ =\ \frac{R}{d}\int_0^{\pi}(\sin\theta)^{d+1}d\theta\ =\ R\int_0^{\pi}(\cos\theta)^2(\sin\theta)^{d-1}d\theta,$$

where the last equality follows from the integration by parts. Then we arrive at

$$\int_{-R+3/2}^{R-1/2} f(x)dx=R\int_0^{\pi}(\cos\theta)^2(\sin\theta)^{d-1}d\theta+O(\frac{1}{R^{d/2}}).$$

Note that

$$f(-R+3/2)=f(R-1/2)\ll\frac{1}{R^{d/2}}$$

and

$$|f^{(j-1)}(-R+3/2)-f^{(j-1)}(R-1/2)|\ll_j\frac{1}{R^{d/2}},\qquad j\ge 2.$$

Hence, to prove (23), we just need estimate the error term in Euler-Maclaurin formula. We first consider the case where d is an even number. We take $p=d+1$ in Euler-Maclaurin

formula, and then $E_p = 0$ with $f^{(j)} \equiv 0$ provided $j \geq d+1$. Then (23) follows when d is even.

We turn to the case where d is odd. We consider the jth derivative of f. Recall that $\theta(x) = \arccos((2x-1)/2R)$ is a function about x. Then, using the new variable θ, $f(x) = (\sin\theta)^d/d$ and $\theta'(x) = -1/(R\sin\theta)$. A simple calculation shows that

$$\begin{aligned} f'(x) &= f'(\theta)\theta'(x) = -\frac{1}{R}(\sin\theta)^{d-2}\cos\theta, \\ f''(x) &= \frac{1}{R^2}\left((d-2)(\sin\theta)^{d-4} - (d-1)(\sin\theta)^{d-2}\right). \end{aligned}$$

Then, by induction, $f^{(2j)}(x)$ is in the form of

$$\frac{1}{R^{2j}} \sum_{j \leq k \leq 2j} C_{k,d}(\sin\theta)^{d-2k},\ C_{k,d} \in \mathbb{R},$$

while $f^{(2j+1)}(x)$ is in the form of

$$\frac{1}{R^{2j+1}} \sum_{j+1 \leq k \leq 2j+1} C'_{k,d}(\sin\theta)^{d-2k}\cos\theta,\ C'_{k,d} \in \mathbb{R}.$$

We take $p = (d+3)/2$ in Euler-Maclaurin formula. Then

$$|E_p| \ll_p \frac{1}{R^{(d+1)/2}} \int_{\arccos(1-1/R)}^{\arccos(-1+1/R)} \frac{1}{\sin^2\theta} d\theta \ll \frac{1}{R^{d/2}}.$$

when p is even. Similar argument also holds when p is odd. □

Proof of Theorem 1.4. We denote the number of the non-zero entries in x by $\|x\|_0$, i.e.,

$$\|x\|_0 := \#\{j : x_j \neq 0\}.$$

The proof is by induction on $\|x\|_0$. Note that

$$\begin{aligned} \lim_{n\to\infty} E_\delta(x, \mathcal{F}_n) &= \lim_{n\to\infty} \| \frac{d+1}{N_n} \sum_{j=1}^{N_n} \Delta_\delta(x \cdot e_j) e_j \| \\ &= (d+1) \| \int_{\mathbf{z}\in\mathbb{S}^d} \Delta_\delta(x \cdot \mathbf{z}) \mathbf{z} d\omega \|. \end{aligned}$$

We begin with $\|x\|_0 = 1$. Without loss of generality, we suppose $x = [x_1, 0, \ldots, 0]^T \in \mathbb{R}^{d+1}$ and consider $\lim_{n\to\infty} E_\delta(x, \mathcal{F}_n)$. By the sphere coordinate system, each $\mathbf{z} = [z_1, \ldots, z_d] \in \mathbb{S}^d$ can be written in the form of

$$[\cos\theta_1, \sin\theta_1\cos\theta_2, \sin\theta_1\sin\theta_2\cos\theta_3, \ldots, \sin\theta_1\cdots\sin\theta_d]^T,$$

where $\theta_1 \in [0, \pi)$ and $\theta_j \in [-\pi, \pi), 2 \leq j \leq d$. To state conveniently, we set

$$\Theta := [0,\pi) \times \underbrace{[-\pi,\pi) \times \cdots \times [-\pi,\pi)}_{d-1}, \qquad S_m(\theta) := \prod_{j=1}^{m} \sin\theta_j$$

and

$$J_d(\theta) := \left|(\sin\theta_1)^{d-1}(\sin\theta_2)^{d-2}\cdots(\sin\theta_{d-1})\right|.$$

Noting that

$$d\omega = J_d(\theta)d\theta_1\cdots d\theta_d \quad \text{and} \quad \int_{\mathbf{z}\in\mathbb{S}^d}\Delta_\delta(x_1z_1)z_j d\omega = 0, \qquad j\geq 2$$

we have

$$\begin{aligned}
&\lim_{n\to\infty} E_\delta(x,\mathcal{F}_n) = (d+1)\|\int_{\mathbf{z}\in\mathbb{S}^d}\Delta_\delta(x\cdot\mathbf{z})\mathbf{z}d\omega\| \\
= \quad & (d+1)\left|\int_{\mathbf{z}\in\mathbb{S}^d}\Delta_\delta(x_1z_1)z_1 d\omega\right| \\
= \quad & (d+1)\left|\int_{\theta\in\Theta}\Delta_\delta(x_1\cos\theta_1)\cos\theta_1(\sin\theta_1)^{d-1}|(\sin\theta_2)^{d-2}\cdots(\sin\theta_{d-1})|d\theta_1\cdots d\theta_d\right| \\
\ll_d \quad & \delta^{(d+2)/2}
\end{aligned}$$

where the last inequality follows from Lemma 6.1.

For the induction step, we suppose that the conclusion holds for the case where $\|x\|_0\leq k$. We now consider $\|x\|_0\leq k+1$. Without loss of generality, we suppose x is in the form of $[0,\ldots,0,x_{d-k+1},\ldots,x_{d+1}]\in\mathbb{R}^{d+1}$. We can write $[x_d,x_{d+1}]$ in the form of $(r\cos\varphi_0, r\sin\varphi_0)$, where $r\in\mathbb{R}_+$ and $\varphi_0\in[0,2\pi)$. Then

$$x\cdot\mathbf{z} = \sum_{m=d-k+1}^{d-1} x_m S_m(\theta)\cos\theta_m + r\sin\theta_1\cdots\sin\theta_{d-1}\cos(\theta_d-\varphi_0) =: T(\varphi_0).$$

A simple observation is

$$\begin{aligned}
&\left(\int_{\theta\in\Theta}\Delta_\delta(T(\varphi_0))S_{d-1}(\theta)J_d(\theta)\cos\theta_d d\theta\right)^2 + \left(\int_{\theta\in\Theta}\Delta_\delta(T(\varphi_0))S_{d-1}(\theta)J_d(\theta)\sin\theta_d d\theta\right)^2 \\
= \quad & \left(\int_{\theta\in\Theta}\Delta_\delta(T(0))S_{d-1}(\theta)J_d(\theta)\cos\theta_d d\theta\right)^2 + \left(\int_{\theta\in\Theta}\Delta_\delta(T(0))S_{d-1}(\theta)J_d(\theta)\sin\theta_d d\theta\right)^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\lim_{n\to\infty} E_\delta(x,\mathcal{F}_n) = (d+1)\|\int_{\mathbf{z}\in\mathbb{S}^d}\Delta_\delta(x\cdot\mathbf{z})\mathbf{z}d\omega\| \\
= \quad & (d+1)\left(\sum_{m=d-k+1}^{d}(\int_{\theta\in\Theta}\Delta_\delta(T(\varphi_0))S_{m-1}(\theta)J_d(\theta)\cos\theta_m d\theta)^2 + (\int_{\theta\in\Theta}\Delta_\delta(T(\varphi_0))S_d(\theta)J_d(\theta)d\theta)^2\right)^{1/2} \\
= \quad & (d+1)\left(\sum_{m=d-k+1}^{d}(\int_{\theta\in\Theta}\Delta_\delta(T(0))S_{m-1}(\theta)J_d(\theta)\cos\theta_m d\theta)^2 + (\int_{\theta\in\Theta}\Delta_\delta(T(0))S_d(\theta)J_d(\theta)d\theta)^2\right)^{1/2} \\
\ll_d \quad & \delta^{(d+2)/2}
\end{aligned}$$

where the last inequality follows from the fact $\|x\|_0\leq k$ provided $\varphi_0 = 0$.

□

References

1. W. Bennett, Spectra of quantized signals, Bell Syst. Tech. J. 27 (1948)446-472.
2. B. Bodmann and V. Paulsen, Frame paths and error bounds for sigma-delta quantization, Appl. Comput. Harmon. Anal. 22 (2007), 176-197.
3. S. Borodachov and Y. Wang On the distribution of uniform quantization errors, Applied and Computational Harmonic Analysis 27 (2009), no. 3, 334–341.
4. J. Benedetto, A. M. Powell and Ö. Yılmaz, Sigma-delta quantization and finite frames, IEEE Trans. Inform. Theory, 52(2006), 1990-2005.
5. J. Benedetto, A. M. Powell and Ö. Yılmaz, Second order sigmal-delta quantization of finite frame expansions, Appl. Comput. Harmon. Anal., 20 (2006), 126-248.
6. S. Borodachov and Y. Wang, Lattice quantization error for redundant representations, Appl. Comput. Harmon. Anal. 27 (2009) 334-341.
7. V. Goyal, M. Vetterli, N. Thao, Quantized overcomplete expansions in $\mathbb{R}^n$: Analysis, synthesis, and algorithms, IEEE Trans. Inform. Theory 44 (1) (1998)16-31.
8. S. Güntürk, Approximating a bandlimited function using very coarsely quantized data: improved error estimates in sigma-delta modulation, J. Amer. Math. Soc., Vol.17, No.1 (2003), 229-242.
9. W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Stat. Asso. 58 (301): 13-30, March 1963.
10. R. Holmes and V. Paulsen, Optimal frames for erasures, Linear Algebra Appl., Vol. 377, 2004, 31-51.
11. D. Jimenez, L. Wang and Y. Wang, White noise hypothesis for uniform quantization errors, SIAM J. Math. Anal. Vol.38, No.6,2007, 2042-2056.

Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

E-mail address: ywang@math.msu.edu

LSEC, Inst. Comp. Math., Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, 100091, China

E-mail address: xuzq@lsec.cc.ac.cn