SIGMA-DELTA QUANTIZATION ERRORS AND THE TRAVELING SALESMAN PROBLEM

YANG WANG

ABSTRACT. In transmission, storaging and coding of digital signals we frequently perform A/D conversion using quantization. In this paper we study the maxiaml and mean square errors as a result of quantization. We focus on the sigma-delta modulation quantization scheme in the finite frame expansion setting. We show that this problem is related to the classical Traveling Salesman Problem (TSP) in the Euclidean space. It is known ([3]) that the error bounds from the sigma-delta scheme depends on the ordering of the frame elements. By examining *a priori* bounds for the Euclidean TSP we show that error bounds in the sigma-delta scheme is superior to those from the pulse code mudulation (PCM) scheme in general. We also give a recursive algorithm for findng an ordering of the frame elements that will lead to good maximal error and mean square error.

1. INTRODUCTION

In signal processing, coding and many other practical applications, a signal is first decomposed using some suitably chosen atoms ("basis"). In the digital domain, a quantization is performed for transmission, storage, coding and other purposes. But quantizations inevitably induce errors. It is thus important that we understand how the errors behave and have a good estimate of these errors.

There are two commonly used ways to measure quantization errors. They are the worst case error (*maximal error*) and the "average" error (*mean square error, or MSE*). In this paper we study both measurements of quantization errors for the sigma-delta modulation scheme, a scheme that is used in, among others, audio applications. Particularly we show that in general, as far as errors are concerned, the sigma-delta modulation offers a definitive edge over the more straightforward pulse code modulation (PCM) scheme.

We first introduce some of the mathematical background involved in our study. In signal decomposition we start off with a "basis" $\{\mathbf{v}_j\}_{j\in\Lambda}$. It can take on a variety of forms, such

1991 *Mathematics Subject Classification.* Primary 42C15.

Key words and phrases. Sigma-delta modulation, PCM, quantization, frames, traveling salesman problem, Peano space-filling curve.

Supported in part by the National Science Foundation grant DMS-0139261.

as a set of functions or a set of vectors. Here the index set Λ is countable or finite. A signal $\mathbf{x}$ is represented as a linear combination

$$\mathbf{x} = \sum_{j\in\Lambda} c_j \mathbf{v}_j. \tag{1.1}$$

Note that in practice $\{\mathbf{v}_j\}$ is often not a *bona fide* basis, as redundancy is built into $\{\mathbf{v}_j\}$ for the purpose of error correction, recovery from channel erasures, denoising and general robustness. With redundancy the coefficients $\{c_j\}$ in (1.1) are no longer unique. In this paper we shall only consider the case where $\Lambda = \{1, 2, \ldots, N\}$ is a finite set and $\{c_j\}$ are real. Without loss of generality we may set up our problem by assuming that $\mathbf{x}$ and $\{\mathbf{v}_j\}_{j\in\Lambda}$ are all in $\mathbb{R}^d$, $d \geq 1$.

Given $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ we let $F = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$ be the corresponding matrix whose columns are $\{\mathbf{v}_j\}_{j=1}^N$. We say $\mathcal{F} = \{\mathbf{v}_j\}$ is a *frame* for $\mathbb{R}^d$ if the matrix F has rank d, and it is a *tight frame* for $\mathbb{R}^d$ if $FF^T = \lambda I_d$ where $\lambda > 0$. The value λ is called the *tight frame bound* for the tight frame. Note that for tight frames we have

$$\lambda d = \mathrm{tr}(FF^T) = \mathrm{tr}(F^T F) = \sum_{j=1}^N \|\mathbf{v}_j\|^2. \tag{1.2}$$

In particular if we consider *unit norm* tight frames in which $\|\mathbf{v}_j\| = 1$ for all j then $\lambda = N/d$. Our focus in this paper is on tight frames, with special attention given to unit norm tight frames.

With a frame $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ any signal $\mathbf{x}$ can be reconstructed from the data $\{\langle \mathbf{x}, \mathbf{v}_j\rangle\}_{j=1}^N$. Let $G = [\mathbf{u}_1, \cdots, \mathbf{u}_N]$ be a $d \times N$ matrix such that $GF^T = I_d$. Observe that

$$[\langle \mathbf{x}, \mathbf{v}_1\rangle, \langle \mathbf{x}, \mathbf{v}_2\rangle, \ldots, \langle \mathbf{x}, \mathbf{v}_N\rangle]^T = F^T\mathbf{x}.$$

Thus $\mathbf{x}$ can be reconstructed using G by $\mathbf{x} = G(F^T\mathbf{x})$. The columns of G, $\mathcal{G} = \{\mathbf{u}_j\}_{j=1}^N$, is called a *dual frame* of the frame $\mathcal{F}$. With a dual frame $\mathcal{G}$ we have

$$\mathbf{x} = GF^T\mathbf{x} = G(F^T\mathbf{x}) = \sum_{j=1}^N \langle \mathbf{x}, \mathbf{v}_j\rangle\, \mathbf{u}_j. \tag{1.3}$$

One such dual frame is to take $G = (FF^T)^{-1}F$, which clearly satisfies $GF^T = I_d$. The columns of this G form the so-called *canonical dual frame* of $\mathcal{F}$. If $\{\mathbf{v}_j\}_{j=1}^N$ is a tight frame with tight frame bound λ then the canonical dual frame corresponds to the matrix $G = \lambda^{-1}F$, whose columns are simply $\lambda^{-1}\mathcal{F}$. In this case the reconstruction formula (1.3)

becomes

$$\mathbf{x} = \frac{1}{\lambda} \sum_{j=1}^{N} \langle \mathbf{x}, \mathbf{v}_j \rangle \, \mathbf{v}_j. \tag{1.4}$$

In digital applications real numbers must be quantized. In other words, a real number t must be represented by a value in a quantization alphabet $\mathcal{A}$. Here we assume that $\mathcal{A} = \{0, \pm\Delta, \pm 2\Delta, \dots\} = \Delta\mathbb{Z}$, where $\Delta > 0$ is the quantization step. A standard way for quantizing a real number is *linear quantization*, in which a real value t is quantize to the value in $\mathcal{A} = \Delta\mathbb{Z}$ that is the closest to t, so

$$Q_\Delta(t) = \arg\min_{r \in \mathcal{A}} |t - r| = \Delta \left\lfloor \frac{t}{\Delta} + \frac{1}{2} \right\rfloor.$$

It is possible to use nonlinear quantization schemes to achieve better results in some cases, see Goyal, Vetterli and Thao [15], but we shall not discuss them here. Also, another popular alphabet $\mathcal{A}$ is the *midrise* alphabet $\mathcal{A} = \Delta\mathbb{Z} + \frac{\Delta}{2}$. The mathematical analysis in this paper makes no distinction, and all our results hold for this alphabet.

We now consider the reconstruction problem by a tight frame $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ with frame bound λ. Instead of the reconstruction formula (1.4) we now have an imperfect reconstruction

$$\hat{\mathbf{x}} = \frac{1}{\lambda} \sum_{j=1}^{N} q_j \, \mathbf{v}_j, \tag{1.5}$$

where $q_j \in \mathcal{A} = \Delta\mathbb{Z}$. The error from this construction will depend on the frame $\mathcal{F}$, Δ and how $\{q_j\}$ are chosen. Currently there are two commonly used schemes for choosing $\{q_j\}$: The PCM quantization scheme and the sigma-delta modulation quantization scheme. In the PCM scheme, we simply let

$$q_j = Q_\Delta(\langle \mathbf{x}, \mathbf{v}_j \rangle) = \left\lfloor \frac{\langle \mathbf{x}, \mathbf{v}_j \rangle}{\Delta} + \frac{1}{2} \right\rfloor \Delta,$$

which is the element in $\Delta\mathbb{Z}$ closest to $\langle \mathbf{x}, \mathbf{v}_j \rangle$. This is the most direct scheme. Let $\tau_\Delta(t) = t - Q_\Delta(t)$. Then the reconstruction error is

$$\mathbf{x} - \hat{\mathbf{x}} = \frac{1}{\lambda} \sum_{j=1}^{N} \tau_\Delta(\langle \mathbf{x}, \mathbf{v}_j \rangle) \, \mathbf{v}_j. \tag{1.6}$$

Note that $|\tau_\Delta(\langle \mathbf{x}, \mathbf{v}_j \rangle)| \le \Delta/2$. It is known that the worst case error is bounded by

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \le \sqrt{\frac{N}{\lambda}} \frac{\Delta}{2}, \tag{1.7}$$

see e.g. [18]. Under the so-called White Noise Hypothesis (WNH), in which $\{\tau_\Delta(\langle \mathbf{x}, \mathbf{v}_N\rangle)\}_{j=1}^N$ are assumed to be independent and uniformly distributed in $[-\Delta/2, \Delta/2)$, the mean square error (MSE) is

$$\mathcal{E}(\|\mathbf{x} - \hat{\mathbf{x}}\|^2) = \frac{\Delta^2}{12\lambda}, \tag{1.8}$$

see e.g. [15]. The WNH is shown to be realistic for small Δ under fairly general settings in [18].

An alternative quantization scheme is the sigma-delta modulation scheme. It is less direct but is known to offer some advatanges, such as robustness, over the PCM scheme. Here we only consider the first order sigma-delta scheme. Discussions on higher order sigma-delta schemes can be found in e.g. [7] and [4] and the references therein. The sigma-delta scheme utilizes summation by parts for quantization. Let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ be a frame and $\mathcal{G} = \{\mathbf{u}_j\}_{j=1}^N$ be a dual frame. In a nutshell, instead of quantizing $\{\langle \mathbf{x}, \mathbf{v}_j\rangle\}$ directly to obtain the reconstruction $\hat{\mathbf{x}} = \sum_{j=1}^N Q_\Delta(\langle \mathbf{x}, \mathbf{v}_j\rangle)\mathbf{u}_j$ we first rewrite the frame reconstruction formula using summation by parts:

$$\mathbf{x} = \sum_{j=1}^N \langle \mathbf{x}, \mathbf{v}_j\rangle\, \mathbf{u}_j = \sum_{k=1}^{N-1} S_k\, (\mathbf{u}_k - \mathbf{u}_{k+1}) + S_N \mathbf{u}_N, \tag{1.9}$$

where $S_0 := 0$ and $S_k := \sum_{j=1}^k \langle \mathbf{x}, \mathbf{v}_j\rangle$. Now we quantize $\{S_k\}$ to obtain a reconstruction

$$\tilde{\mathbf{x}} = \sum_{k=1}^{N-1} Q_\Delta(S_k)\, (\mathbf{u}_k - \mathbf{u}_{k+1}) + Q_\Delta(S_N)\mathbf{u}_N = \sum_{k=1}^N q_k\, \mathbf{u}_k,$$

where $q_k := Q_\Delta(S_k) - Q_\Delta(S_{k-1})$, $k \geq 1$. An alternative way, which is more practical for building circuits, is to describe the sigma-delta scheme by the sequences

$$\left\{ \begin{array}{rcl} r_k &=& \langle \mathbf{x}, \mathbf{v}_k\rangle + r_{k-1} - q_k, \\ q_k &=& Q_\Delta(\langle \mathbf{x}, \mathbf{v}_k\rangle + r_{k-1}), \end{array} \right.$$

with $r_0 := 0$ and $1 \leq k \leq N$. Observe that $r_k - r_{k-1} = \langle \mathbf{x}, \mathbf{v}_k\rangle - q_k$. Thus

$$r_k = r_k - r_0 = \sum_{j=1}^k (\langle \mathbf{x}, \mathbf{v}_j\rangle - q_j) = S_k - \sum_{j=1}^k q_j.$$

Since $r_k \in [-\Delta/2, \Delta/2]$ we have $\sum_{j=1}^{k} q_j = Q_\Delta(S_k)$, and hence $q_k = Q_\Delta(S_k) - Q_\Delta(S_{k-1})$. For tight frames with frame bound λ, the sigma-delta scheme yields the quantized reconstruction

$$\tilde{\mathbf{x}} = \frac{1}{\lambda} \sum_{j=1}^{N} q_j \mathbf{v}_j. \tag{1.10}$$

An important difference between the PCM scheme and the sigma-delta scheme is that while the PCM scheme is independent of the ordering of the frame elements $\{\mathbf{v}_j\}$, the sigma-delta scheme is *very sensitive* to it. If we do not choose the ordering carefully the sigma-delta scheme can perform very poorly, see [3] for a comprehensive discussion. For some particular tight frames such as the harmonic tight frames in $\mathbb{R}^d$, it is shown in [3] that with a suitable ordering the sigma-delta scheme is superior to the PCM scheme, yielding a maximal error bound of $O(\Delta/\lambda)$ and MSE bound of $O(\Delta^2/\lambda^2)$. However, the error comparison for general tight frames between the two schemes is far less clear. It is the goal of this paper to show that for tight frames under very general settings, when N is large and Δ is small, the sigma-delta scheme is superior to the PCM scheme, both in terms of maximal error bound and MSE. Key to our study is an interesting connection between the sigma-delta scheme and the Traveling Salesman Problem (TSP).

2. Maximal Error and MSE

Throughout this section we let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ and $\mathcal{G} = \{\mathbf{u}_j\}_{j=1}^N$ be a dual frame of $\mathcal{F}$. The quantization alphabet is $\mathcal{A} = \Delta\mathbb{Z}$. Recall from (1.9) the reconstruction formula for any $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{x} = \sum_{j=1}^{N} \langle \mathbf{x}, \mathbf{v}_j \rangle \, \mathbf{u}_j = \sum_{k=1}^{N-1} S_k \, (\mathbf{u}_k - \mathbf{u}_{k+1}) + S_N \mathbf{u}_N$$

where $S_k = S_k(x) := \sum_{j=1}^{k} \langle \mathbf{x}, \mathbf{v}_j \rangle = \langle \mathbf{x}, \sum_{j=1}^{k} \mathbf{v}_j \rangle$. The reconstruction after quantization is

$$\tilde{\mathbf{x}} = \tilde{\mathbf{x}}(\Delta) := \sum_{k=1}^{N-1} Q_\Delta(S_k) \, (\mathbf{u}_k - \mathbf{u}_{k+1}) + Q_\Delta(S_N) \mathbf{u}_N = \sum_{k=1}^{N} q_k \, \mathbf{u}_k,$$

where $q_k := Q_\Delta(S_k) - Q_\Delta(S_{k-1})$, $k \geq 1$ with $S_0 := 0$. Denote $\tau_\Delta(S_k) := S_k - Q_\Delta(S_k)$, the quantization round off error for S_k. Then the reconstruction error is given by

$$\mathbf{x} - \tilde{\mathbf{x}} = \sum_{k=1}^{N-1} \tau_\Delta(S_k)\,(\mathbf{u}_k - \mathbf{u}_{k+1}) + \tau_\Delta(S_N)\mathbf{u}_N. \tag{2.1}$$

This leads to the following *a priori* error bound, given in [3]:

Proposition 2.1 ([3])**.** *For any $\mathbf{x} \in \mathbb{R}^d$ we have*

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{\Delta}{2}\Big(\sum_{k=1}^{N-1} \|\mathbf{u}_k - \mathbf{u}_{k+1}\| + \|\mathbf{u}_N\|\Big). \tag{2.2}$$

As a corollary, if $\mathcal{F}$ is a tight frame with frame bound λ and $\mathcal{G} = \frac{1}{\lambda}\mathcal{F}$, then we have

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| \leq \frac{\Delta}{2\lambda}\Big(\sum_{k=1}^{N-1} \|\mathbf{v}_k - \mathbf{v}_{k+1}\| + \|\mathbf{v}_N\|\Big). \tag{2.3}$$

The quantity $\sum_{k=1}^{N-1} \|\mathbf{u}_k - \mathbf{u}_{k+1}\| + \|\mathbf{u}_N\|$ is called the *frame variation* of $\mathcal{G}$, following the terminology in [3].

The MSE of the quantized reconstruction, assuming $\mathbf{x}$ is chosen according to some probabilistic distribution, is not so easy to estimate. To circumvent the difficulties the so-called *White Noise Hypothesis* (WNH) is commonly employed by engineers and mathematicians working in the area, see e.g. [15]. The WNH asserts the following:

- *The round off error components $\{\tau_\Delta(S_k)\}_{k=1}^N$ are independent.*
- *Each $\tau_\Delta(S_k)$ is uniformly distributed in $[-\Delta/2, \Delta/2)$.*

Proposition 2.2. *Let $\mathbf{x}$ be an absolutely continuous random vector in $\mathbb{R}^d$. Under the WNH the MSE of the quantized reconstruction is given by*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) = \frac{\Delta^2}{12}\Big(\sum_{k=1}^{N-1} \|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2 + \|\mathbf{u}_N\|^2\Big). \tag{2.4}$$

In particular, if $\mathcal{F}$ is a tight frame with frame bound λ and $\mathcal{G} = \frac{1}{\lambda}\mathcal{F}$ then

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) = \frac{\Delta^2}{12\lambda^2}\Big(\sum_{k=1}^{N-1} \|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2 + \|\mathbf{v}_N\|^2\Big). \tag{2.5}$$

Proof. Let H be the $d \times N$ matrix $H = [\mathbf{u}_1 - \mathbf{u}_2, \cdots, \mathbf{u}_{N-1} - \mathbf{u}_N, \mathbf{u}_N]$. Then by (2.1) we have

$$\mathbf{x} - \tilde{\mathbf{x}} = H[\tau_\Delta(S_1), \cdots, \tau_\Delta(S_N)]^T.$$

Let $H^TH = [t_{ij}]_{i,j=1}^N$. Then

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) = \mathcal{E}\Big(\sum_{i,j=1}^N \tau_\Delta(S_i)\tau_\Delta(S_j)t_{ij}\Big) = \sum_{i,j=1}^N \mathcal{E}\Big(\tau_\Delta(S_i)\tau_\Delta(S_j)\Big)t_{ij}. \tag{2.6}$$

Now the WNH yields $\mathcal{E}(\tau_\Delta(S_i)\tau_\Delta(S_j)) = \frac{\Delta^2}{12}$ if $i=j$ and 0 otherwise. Thus

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) = \frac{\Delta^2}{12}\sum_i^N t_{ii} = \operatorname{tr}(H^TH) = \sum_{k=1}^{N-1}\|\mathbf{u}_k-\mathbf{u}_{k+1}\|^2 + \|\mathbf{u}_N\|^2.$$

This proves the proposition. ∎

Unfortunately the WNH is valid only under very stringent conditions, and when $N > d$ it is almost never true, see Jimenez, Wang and Wang [18] on the WNH for PCM quantization schemes. However, as Δ tends to zero $\{\tau_\Delta(S_k)\}_{k=1}^N$ become asymptotically uncorrelated and uniformly distributed in $[-\Delta/2, \Delta/2)$ under some minor assumptions. This weak version of the WNH is sufficient to yield an asymptotic version of Proposition 2.2.

To state our asymptotic result on the MSE in sigma-delta modulation scheme we first introduce some notations. We say $\mathbf{w}_1 \in \mathbb{R}^d$ is a $\frac{p}{q}$-*multiple* of $\mathbf{w}_2 \in \mathbb{R}^d$ if $\mathbf{w}_1, \mathbf{w}_2 \neq \mathbf{0}$ and $\mathbf{w}_1 = \frac{p}{q}\mathbf{w}_2$ where $p, q \in \mathbb{Z}$ and are coprime. For $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$ we define $r(\mathbf{w}_1, \mathbf{w}_2)$ as follows: $r(\mathbf{w}_1, \mathbf{w}_2) = \frac{1}{pq}$ if $\mathbf{w}_1$ is a $\frac{p}{q}$-multiple of $\mathbf{w}_2$ and $p+q$ is even; $r(\mathbf{w}_1, \mathbf{w}_2) = -\frac{1}{2pq}$ if $\mathbf{w}_1$ is a $\frac{p}{q}$-multiple of $\mathbf{w}_2$ and $p+q$ is odd; in all other cases we let $r(\mathbf{w}_1, \mathbf{w}_2) = 0$. For any $\mathbf{w} \in \mathbb{R}^d$ let $\delta(\mathbf{w}) = 1$ if $\mathbf{w} \neq \mathbf{0}$ and $\delta(\mathbf{w}) = 0$ if $\mathbf{w} = \mathbf{0}$.

Theorem 2.3. *Let $\mathbf{x}$ be an absolutely continuous random vector in $\mathbb{R}^d$. Then for small $\Delta > 0$ the MSE of the quantized reconstruction satisfies*

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) = \frac{\Delta^2}{12}\Big(\sum_{k=1}^N \|\mathbf{u}_k-\mathbf{u}_{k+1}\|^2\delta(\mathbf{w}_k) + 2\sum_{1\le k<l\le N} t_{kl}\, r(\mathbf{w}_k, \mathbf{w}_l)\Big) + o(\Delta^2) \tag{2.7}$$

where $\mathbf{u}_{N+1} := \mathbf{0}$, $\mathbf{w}_k := \sum_{j=1}^k \mathbf{v}_j$, and $t_{kl} := \langle \mathbf{u}_k - \mathbf{u}_{k+1}, \mathbf{u}_l - \mathbf{u}_{l+1}\rangle$.

Proof. Set $H = [\mathbf{u}_1 - \mathbf{u}_2, \cdots, \mathbf{u}_{N-1} - \mathbf{u}_N, \mathbf{u}_N]$. Then $H^TH = [t_{kl}]$ where $t_{kl} = \langle \mathbf{u}_k - \mathbf{u}_{k+1}, \mathbf{u}_l - \mathbf{u}_{l+1}\rangle$. By (2.6) we have

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) = \mathcal{E}\Big(\sum_{k,l=1}^N \tau_\Delta(S_k)\tau_\Delta(S_l)t_{kl}\Big) = \sum_{k,l=1}^N \mathcal{E}\Big(\tau_\Delta(S_k)\tau_\Delta(S_l)\Big)t_{kl}.$$

Observe that $\tau_\Delta(S_k) = \tau_\Delta(\langle \mathbf{x}, \mathbf{w}_k\rangle)$. Now it follows from Theorem 4.1 and Theorem 5.1 in [18] that

$$\mathcal{E}\Big(\tau_\Delta(S_k)\tau_\Delta(S_l)\Big) = \frac{r(\mathbf{w}_k, \mathbf{w}_l)}{12} + o(\Delta^2).$$

Note also that for $k = l$ we have $r(\mathbf{w}_k, \mathbf{w}_l) = \delta(\mathbf{w}_k)$. Thus

$$\begin{aligned}
\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) &= \sum_{k,l=1}^{N} \mathcal{E}\Big(\tau_\Delta(S_k)\tau_\Delta(S_l)\Big) t_{kl} \\
&= \frac{\Delta^2}{12}\Big(\sum_{k,l=1}^{N} \langle \mathbf{u}_k - \mathbf{u}_{k+1}, \mathbf{u}_l - \mathbf{u}_{l+1}\rangle r(\mathbf{w}_k, \mathbf{w}_l)\Big) + o(\Delta^2) \\
&= \frac{\Delta^2}{12}\Big(\sum_{k=1}^{N} \|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2 \delta(\mathbf{w}_k) + 2\sum_{1\le k<l\le N} t_{kl}\, r(\mathbf{w}_k, \mathbf{w}_l)\Big) + o(\Delta^2).
\end{aligned}$$

∎

Corollary 2.4. *Let $\mathbf{x}$ be an absolutely continuous random vector in $\mathbb{R}^d$. Assume that the vectors $\{\sum_{j=1}^{k} \mathbf{v}_j\}_{k=1}^{N}$ are pairwise linearly independent over $\mathbb{Q}$. Then the MSE of the quantized reconstruction satisfies*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^)) = \frac{\Delta^2}{12}\Big(\sum_{k=1}^{N} \|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2\Big) + o(\Delta^2), \tag{2.8}$$

where $\mathbf{u}_{N+1} := \mathbf{0}$.

Corollary 2.5. *Let $\mathbf{x}$ be an absolutely continuous random vector in $\mathbb{R}^d$. Let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^{N}$ be a tight frame with frame bound λ and $\mathcal{G} = \frac{1}{\lambda}\mathcal{F}$. Assume that the vectors $\{\sum_{j=1}^{k} \mathbf{v}_j\}_{k=1}^{N}$ are pairwise linearly independent over $\mathbb{Q}$. Then the MSE of the quantized reconstruction satisfies*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) = \frac{\Delta^2}{12\lambda^2}\Big(\sum_{k=1}^{N-1} \|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2 + \|\mathbf{v}_N\|^2\Big) + o(\Delta^2). \tag{2.9}$$

It was shown in [3] that for harmonic frames in $\mathbb{R}^d$ the MSE is at least $O(\frac{\Delta^2}{N^2})$. With Theorem 2.3 we prove here that as $\Delta \to 0^+$ the MSE is in fact $O(\frac{\Delta^2}{N^3})$ for even d. Recall that the harmonic frame $\mathcal{H}_{d,N}$ in $\mathbb{R}^d$ is defined as follows: For $d = 2d'$ we have $\mathcal{H}_{d,N} = \{\mathbf{v}_j\}_{j=1}^{N}$ where

$$\mathbf{v}_j = \sqrt{\frac{2}{d}}\Big[\cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \cos\frac{4\pi j}{N}, \sin\frac{4\pi j}{N}, \ldots, \cos\frac{2d'\pi j}{N}, \sin\frac{2d'\pi j}{N}\Big]^T.$$

For $d = 2d' + 1$ we have $\mathcal{H}_{d,N} = \{\mathbf{v}_j\}_{j=1}^N$ where

$$\mathbf{v}_j = \sqrt{\frac{2}{d}}\Big[\frac{1}{\sqrt{2}}, \cos\frac{2\pi j}{N}, \sin\frac{2\pi j}{N}, \ldots, \cos\frac{2d'\pi j}{N}, \sin\frac{2d'\pi j}{N}\Big]^T.$$

It is well known that $\mathcal{H}_{d,N}$ is a unit norm tight frame with frame bound $\lambda = \frac{N}{d}$. Furthermore, for even d we have $\sum_{j=1}^N \mathbf{v}_j = \mathbf{0}$.

Theorem 2.6. *Let $\mathbf{x}$ be an absolutely continuous random vector. Let $\mathcal{F} = \mathcal{H}_{d,N}$ be the harmonic frame with $d \geq 2$ and $\mathcal{G} = \frac{d}{N}\mathcal{F}$ be the canonical dual frame. For any $k \geq 1$ let $A(k,N) := \sum_{j=1}^k \sin^2\big(\frac{\pi j}{N}\big)$.*

(i) *Suppose that $d = 2d'$. Then as $\Delta \to 0^+$,*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) = \frac{d(N-1)A(d',N)}{6N^2}\Delta^2 + o(\Delta^2).$$

(i) *Suppose that $d = 2d' + 1$. Then as $\Delta \to 0^+$,*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}\|^2) = \frac{d^2 + 2d(N-1)A(d',N)}{12N^2}\Delta^2 + o(\Delta^2).$$

Proof. We first claim that $\{\sum_{j=1}^k \mathbf{v}_j\}_{k=1}^{N-1}$ are pairwise linearly independent. Note that it suffices to prove it for $d = 2$, as $\{[\cos(\frac{2\pi j}{N}), \sin(\frac{2\pi j}{N})]^T\}_{j=1}^N$ are embedded in the vectors $\{\mathbf{v}_j\}$ (the first two entries for even d, and the second and third entries for odd d). Now, identifying $[\cos(\frac{2\pi ij}{N}), \sin(\frac{2\pi j}{N})]^T$ with ω_N^j, where $\omega_N := e^{\frac{2\pi i}{N}}$, we prove that for distinct $1 \leq k, l < N$ we must have

$$\sum_{j=1}^k \omega_N^j \neq c\sum_{j=1}^l \omega_N^j, \quad \text{or equivalently,} \quad \sum_{j=0}^{k-1} \omega_N^j \neq c\sum_{j=0}^{l-1} \omega_N^j$$

for all $c \in \mathbb{R}$. Assume it is false. Observe from simple geometry that $\sum_{j=0}^{k-1} \omega_N^j = a_k e^{\frac{\pi(k-1)}{N}i}$ for some $a_k \in \mathbb{R}$, $a_k \neq 0$. Thus

$$a_k e^{\frac{\pi(k-1)}{N}i} = c a_l e^{\frac{\pi(l-1)}{N}i}.$$

For this to happen we must have $k - l \equiv 0 \mod(N)$. This is not possible given our assumption. The claim is hence proved.

For $d = 2d'$ we note that $\sum_{j=1}^N \mathbf{v}_j = \mathbf{0}$. Thus by Theorem 2.3 we have

$$\begin{aligned}\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) &= \frac{\Delta^2}{12}\Big(\sum_{k=1}^{N-1}\|\mathbf{u}_k-\mathbf{u}_{k+1}\|^2\Big)+o(\Delta^2)\\ &= \frac{d^2\Delta^2}{12N^2}\Big(\sum_{k=1}^{N-1}\|\mathbf{v}_k-\mathbf{v}_{k+1}\|^2\Big)+o(\Delta^2).\end{aligned}$$

It is standard to check that

$$\|\mathbf{v}_k-\mathbf{v}_{k+1}\| = \frac{2}{d}\sum_{j=1}^{d'}\sin^2\Big(\frac{\pi j}{N}\Big) = \frac{2}{d}A(d',N).$$

Part (i) of the theorem now follows.

For $d = 2d'+1$ the property $\sum_{j=1}^N \mathbf{v}_j = \mathbf{0}$ no longer holds. We claim that $\{\sum_{j=1}^k \mathbf{v}_j\}_{k=1}^N$ are pairwise linearly independent. We have aleady shown that $\{\sum_{j=1}^k \mathbf{v}_j\}_{k=1}^{N-1}$ are pairwise linearly independent, so now we only need to show that $\sum_{j=1}^N \mathbf{v}_j$ is not parallel to any of the other vectors. Note that $\sum_{j=1}^N \mathbf{v}_j = [b,0,\ldots,0]^T$ with $b = N/\sqrt{2}$. However, all the other vectors have some nonzero second or third entries. The claim follows. Corollary 2.5 yields

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) = \frac{d^2\Delta^2}{12N^2}\Big(\sum_{k=1}^{N}\|\mathbf{v}_k-\mathbf{v}_{k+1}\|^2\Big)+o(\Delta^2)$$

with $\mathbf{v}_{N+1} := \mathbf{0}$. Again, it is easy to check that $\|\mathbf{v}_k-\mathbf{v}_{k+1}\|^2 = \frac{2}{d}A(d',N)$ if $k<N$ and $\|\mathbf{v}_N-\mathbf{v}_{N+1}\|^2 = 1$. ∎

Corollary 2.7. *Let $\mathbf{x}$ be an absolutely continuous random vector in $\mathbb{R}^d$. Let $\mathcal{F} = \mathcal{H}_{d,N}$ be the harmonic frame with $d \geq 2$ and $\mathcal{G} = \frac{d}{N}\mathcal{F}$ be the canonical dual frame. Suppose that $d = 2d'$. Then as $\Delta \to 0^+$,*

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) \leq \frac{d^2(d+1)(d+2)\pi^2}{144N^3}\Delta^2 + o(\Delta^2). \tag{2.10}$$

Proof. First we clearly have

$$\mathcal{E}(\|\mathbf{x}-\tilde{\mathbf{x}}\|^2) \leq \frac{dA(d',N)}{6N}\Delta^2 + o(\Delta^2).$$

Next we note that $A(d',N) \leq \sum_{j=1}^{d'} \frac{\pi^2 j^2}{N^2}$ and

$$\sum_{j=1}^{d'} j^2 = \frac{d'(d'+1)(2d'+1)}{6} = \frac{d(d+2)(d+1)}{24}.$$

The corollary follows. ∎

3. The Traveling Saleman Problem

The results in the previous section consider the sigma-delta modulation scheme using the given order of the frame. However, for a given frame $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ with dual frame $\mathcal{G} = \{\mathbf{u}_j\}_{j=1}^N$ and a permutation $p \in S_N$ of the indices $\{1, 2, \dots, N\}$, the set of vectors $\mathcal{F}_p := \{\mathbf{v}_{p(j)}\}_{j=1}^N$ with the indices permuted is again a frame with the same frame bounds and dual frame $\mathcal{G}_p = \{\mathbf{u}_{p(j)}\}_{j=1}^N$. If $\mathcal{F}$ is a tight frame with tight frame bound λ and dual frame $\frac{1}{\lambda}\mathcal{F}$, then $\mathcal{F}_p$ is again a tight frame with frame bound λ and dual frame $\frac{1}{\lambda}\mathcal{F}_p$.

While the quantization error in the PCM scheme does not depend on the order of the frame elements, it is very sensitive to the order in the sigma-delta modulation scheme, as pointed out in [3].

For any set of vectors $\mathcal{S} = \{\mathbf{w}_j\}_{j=1}^N$ in $\mathbb{R}^d$ and any $p \in S_N$ denote

$$\sigma_r(\mathcal{S}, p) = \sum_{j=1}^{N-1} \|\mathbf{w}_p(j) - \mathbf{w}_{p(j+1)}\|^r + \|\mathbf{w}_{p(N)}\|^r,$$

with $r > 0$. For $r = 1$ this quantity is called the *frame variation* (if $\mathcal{S}$ is a frame) in [3]. The frame variation plays an important role in the estimation of maximal quantization error in the sigma-delta modulation scheme. As we shall see, $\sigma_2(\mathcal{S}, p)$ plays an important role in the estimation of the MSE in the sigma-delta modulation scheme.

As before, throughout this section we let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ be a frame in $\mathbb{R}^d$ and $\mathcal{G} = \{\mathbf{u}_j\}_{j=1}^N$ be a dual frame of $\mathcal{F}$. The quantization alphabet is $\mathcal{A} = \Delta\mathbb{Z}$. Let $p \in S_N$. Using the permuted frames $\mathcal{F}_p$ and $\mathcal{G}_p$ we obtain a reconstruction

$$\mathbf{x} = \sum_{k=1}^{N-1} S_k^p \left(\mathbf{u}_{p(k)} - \mathbf{u}_{p(k+1)}\right) + S_N^p \, \mathbf{u}_{p(N)},$$

where $S_k^p := \langle \mathbf{x}, \sum_{j=1}^k \mathbf{v}_{p(j)} \rangle$. The reconstruction after quantization is

$$\tilde{\mathbf{x}}_p := \sum_{k=1}^{N-1} Q_\Delta(S_k^p) \left(\mathbf{u}_{p(k)} - \mathbf{u}_{p(k+1)}\right) + Q_\Delta(S_N^p) \, \mathbf{u}_{p(N)}.$$

By (2.1) the reconstruction error after quantization using sigma-delta scheme for any $\mathbf{x} \in \mathbb{R}^d$ is

$$\mathbf{x} - \tilde{\mathbf{x}}_p := \sum_{k=1}^{N-1} \tau_\Delta(S_k^p) \left(\mathbf{u}_{p(k)} - \mathbf{u}_{p(k+1)}\right) + \tau_\Delta(S_N^p) \, \mathbf{u}_{p(N)}. \tag{3.1}$$

The following estimate, established in [3], is a direct consequence of (3.1).

Proposition 3.1 ([3]). *Let $p \in S_N$. For any $\mathbf{x} \in \mathbb{R}^d$ we have*

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{\Delta}{2}\sigma_1(\mathcal{G}, p). \tag{3.2}$$

For the MSE we can apply results from the previous section.

Proposition 3.2. *Let $\mathbf{x} \in \mathbb{R}^d$ be an absolutely continuous random vector. Let $p \in S_N$. Assume that $\{\sum_{j=1}^k \mathbf{v}_{p(j)}\}_{j=1}^N$ are pairwise linearly independent over $\mathbb{Q}$. Then as $\Delta \to 0^+$ we have*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}_p\|^2) = \frac{\Delta^2}{12}\sigma_2(\mathcal{G}, p) + o(\Delta^2). \tag{3.3}$$

We now focus on the following question: How small can $\sigma_1(\mathcal{G}, p)$ and $\sigma_2(\mathcal{G}, p)$ be? For $\sigma_1(\mathcal{G}, p)$ this is essentially the classical Traveling Salesman Problem (TSP) in the Euclidean space $\mathbb{R}^d$. For a give set $\mathcal{S} = \{\mathbf{w}_j\}_{j=1}^N$ in $\mathbb{R}^d$, the TSP considers

$$\mathrm{TSP}(\mathcal{S}) := \min_{p \in S_N} \sum_{j=1}^{N-1} \|\mathbf{w}_{p(j)} - \mathbf{w}_{p(j+1)}\|.$$

Since we usually consider only bounded set of vectors, the problem of finding an *a priori* bound for $\min_{p \in S_N} \sigma_1(\mathcal{S}, p)$ is, aside from an additive constant, is the same as finding an *a priori* bound for $\mathrm{TSP}(\mathcal{S})$. The *a priori* bound for $\mathrm{TSP}(\mathcal{S})$ has been extensively studied, highlighted by the classical Bearwood-Halton-Hammersley Theorem [1], see M. Steele's excellent book [24] for a comprehensive discussion. It is well known that if $\mathcal{S} \subset [-\frac{1}{2}, \frac{1}{2}]^d$ with $|\mathcal{S}| = N$ then

$$\mathrm{TSP}(\mathcal{S}) \leq C\, N^{1-\frac{1}{d}} \tag{3.4}$$

for some *a priori* constant $C > 0$ depending only on d. Few [11] proved that in dimension $d = 2$ one has $\mathrm{TSP}(\mathcal{S}) \leq \sqrt{2N} + 7/4$. For general d it is known that the worst case tour length is asymptotically

$$\sup_{|\mathcal{S}|=N} \mathrm{TSP}(\mathcal{S}) \approx \alpha_d \sqrt{d} N^{1-\frac{1}{d}} \tag{3.5}$$

for large d, where α_d is a constant depending only on d, and is bounded for all d. Few [11] proved that for large d one has $0.2419 \leq \alpha_d \leq 0.7071$. The estimate was improved to $\alpha_d \leq 0.6136$ by Moran [20], and to $\alpha_d \leq 0.4051$ by Goddyn [13]. There is an elementary argument in Newman ([21], Problem 73) that easily establishes the bound

$$\sup_{|\mathcal{S}|=N} \mathrm{TSP}(\mathcal{S}) \leq (1 + \sqrt{d})\big(1 + N^{\frac{1}{d}}\big)^{d-1}. \tag{3.6}$$

Moreover, it is not hard to show that the condition $\mathcal{S} \subset [-\frac{1}{2}, \frac{1}{2}]^d$ can be replaced with $\mathcal{S} \subset \Omega$, where Ω is a bounded subset of $\mathbb{R}^d$ lying in a submanifold of dimension $d' \leq d$. In this setting,

$$\mathrm{TSP}(\mathcal{S}) \leq C(\Omega)\, N^{1-\frac{1}{d'}}. \tag{3.7}$$

Theorem 3.3. *Let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ be a tight frame in $\mathbb{R}^d$ with frame bound λ and dual frame $\mathcal{G} = \frac{1}{\lambda}\mathcal{F}$, $d \geq 2$.*

(1) *Assume that $0 < a \leq \|\mathbf{v}_j\| \leq b$ for all j. Then there exists a permutation $p \in S_N$ and a constant C depending only on d such that*

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{Cb}{a^2 N^{\frac{1}{d}}}\Delta.$$

(2) *Assume that $\mathcal{F}$ is a unit norm tight frame. Then there exists a permutation $p \in S_N$ and a constant C depending only on d such that*

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{C}{N^{\frac{1}{d-1}}}\Delta.$$

Proof. (1) By Proposition 3.1 we have

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{\Delta}{2}\sigma_1(\mathcal{G}, p) = \frac{\Delta}{2\lambda}\sigma_1(\mathcal{F}, p).$$

Since $(2b)^{-1}\mathcal{F} \subset [-\frac{1}{2}, \frac{1}{2}]^d$, by (3.4) there exists a $p \in S_N$ such that $\sigma_1((2b)^{-1}\mathcal{F}, p) \leq C_0 N^{1-\frac{1}{d}}$ where C_0 depends only on d. Thus $\sigma_1(\mathcal{F}, p) \leq 2C_0 b N^{1-\frac{1}{d}}$. Furthermore,

$$d\lambda = \sum_{j=1}^N \|\mathbf{v}_j\|^2 \geq a^2 N.$$

Hence $\frac{N}{\lambda} \leq \frac{d}{a^2}$. It follows that

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{2C_0 bd}{2a^2 N^{\frac{1}{d}}}\Delta = \frac{Cb}{a^2 N^{\frac{1}{d}}}\Delta,$$

with $C = dC_0$.

(2) We again use the estimate $\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{\Delta}{2\lambda}\sigma_1(\mathcal{F}, p)$. Now $\mathcal{F}$ is a subset of the $(d-1)$-sphere in $\mathbb{R}^d$. By (3.7) there exist a $p \in S_N$ and a C_1 depending only on d such that $\sigma_1(\mathcal{F}, p) \leq C_1 N^{1-\frac{1}{d-1}}$. It follows from $d\lambda = N$ that for any $\mathbf{x} \in \mathbb{R}^d$ we have

$$\|\mathbf{x} - \tilde{\mathbf{x}}_p\| \leq \frac{C_1 d}{2N^{\frac{1}{d-1}}}\Delta = \frac{C}{N^{\frac{1}{d-1}}}\Delta,$$

with $C = dC_1/2$. ∎

Remark 3.1. Theorem 3.3 shows that under the modest assumption $0 < a \leq \|\mathbf{v}_j\| \leq b$ in the tight frame case, the sigma-delta modulation scheme, with a suitable permutation of the order of the elements in the frame, yields better bound for the maximal error than the PCM scheme when N is large (The PCM scheme has a maximal error bound of $\sqrt{\frac{N}{2\lambda}}\Delta \geq \sqrt{\frac{d}{2b^2}}\Delta$). This extends the result in [3] for harmonic frames in $\mathbb{R}^d$ and all unit norm tight frames in $\mathbb{R}^2$.

We now focus on the MSE from the sigma-delta quantization. To do so we need to estimate $\min_p \sigma_2(\mathcal{G}, p)$. For any $\mathcal{S} = \{\mathbf{w}_j\}_{j=1}^N$ in $\mathbb{R}^d$ define

$$\mathrm{TSP}_2(\mathcal{S}) := \min_{p \in S_N} \sum_{j=1}^{N-1} \|\mathbf{w}_{p(j)} - \mathbf{w}_{p(j+1)}\|^2.$$

Clearly

$$\mathrm{TSP}_2(\mathcal{S}) \leq \min_{p \in S_N} \sigma_2(\mathcal{S}, p) \leq \mathrm{TSP}_2(\mathcal{S}) + \max_j \|\mathbf{w}_j\|^2.$$

Thus it suffices to estimate $\mathrm{TSP}_2(\mathcal{S})$. To this end it is known that for $\mathcal{S} \subset [-\frac{1}{2}, \frac{1}{2}]^d$ with $d \geq 2$ and $|\mathcal{S}| = N$ one has

$$\mathrm{TSP}_2(\mathcal{S}) \leq C N^{1-\frac{2}{d}} \tag{3.8}$$

where C is a constant depending only on d. Of particular note is that if $d = 2$ then $\mathrm{TSP}_2(\mathcal{S})$ is bounded by a constant C independent of $|\mathcal{S}|$. The simplest way to prove (3.8) is to use space-filling Peano curves, see [24]. Here we shall use this technique to prove our next lemma. Interestingly, to make both $\sigma_1(\mathcal{S}, p)$ and $\sigma_2(\mathcal{S}, p)$ small one should not be too greedy with either. It is known that if the permutation p is chosen to minimize $\sigma_1(\mathcal{S}, p)$ then typically one does not get a good estimate for $\sigma_2(\mathcal{S}, p)$. For $d = 2$, such a p would only yield $\sigma_2(\mathcal{S}, p) \leq C \log N$ instead of a constant bound, see [12] and [23]. Nevertheless, a permutation that yields good bounds for both $\sigma_1(\mathcal{S}, p)$ and $\sigma_2(\mathcal{S}, p)$ exists.

Lemma 3.4. *Let $\mathcal{S} \subset [-\frac{1}{2}, \frac{1}{2}]^d$ with $d \geq 2$ and $|\mathcal{S}| = N$. There exists a $p \in S_N$ such that*

$$\sigma_1(\mathcal{S}, p) \leq 2\sqrt{d+3}\, N^{1-\frac{1}{d}} + \frac{\sqrt{d}}{2}, \qquad \sigma_2(\mathcal{S}, p) \leq 4(d+3)\, N^{1-\frac{2}{d}} + \frac{d}{4}. \tag{3.9}$$

Proof. We apply the space-filling curve technique here. It is shown in Milne [19] that there exists a map $\phi : [0,1] \longrightarrow [-\frac{1}{2}, \frac{1}{2}]^d$ with the following properties: (i) ϕ is surjective; (ii) ϕ is Lipschitz-$\frac{1}{d}$ with constant $2\sqrt{d+3}$, i.e. for any $t, s \in [0,1]$ we have

$$|\phi(t) - \phi(s)| \leq 2\sqrt{d+3}\, |t-s|^{\frac{1}{d}}.$$

Note that if $r \le 1$ and $s_1 + s_2 + \cdots + s_k = a$ with $s_j \ge 0$ then

$$\sum_{j=1}^{k} s_j^r \le k\Big(\frac{a}{k}\Big)^r = a^r k^{1-r}.$$

This fact is an easy exercise in calculus. We apply this fact here. Let $\mathcal{S} = \{x_j\}_{j=1}^N$. Assume that $x_j = \phi(t_j)$. Choose $p \in S_N$ so that $t_{p(1)} < t_{p(2)} < \cdots < t_{p(N)}$. We have

$$\begin{aligned}
\sigma_1(\mathcal{S}, p) &= \sum_{j=1}^{N-1} |\phi\big(t_{p(j+1)}\big) - \phi\big(t_{p(j)}\big)| + |\phi\big(t_{p(N)}\big)| \\
&\le 2\sqrt{d+3} \sum_{j=1}^{N-1} \big|t_{p(j+1)} - t_{p(j)}\big|^{\frac{1}{d}} + \frac{\sqrt{d}}{2} \\
&\le 2\sqrt{d+3}\,(N-1)^{1-\frac{1}{d}} + \frac{\sqrt{d}}{2}.
\end{aligned}$$

Similarly we obtain the bound on $\sigma_2(\mathcal{S}, p)$. ∎

We can apply this result to the estimation of MSE from the sigma-delta quantization.

Theorem 3.5. *Let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^N$ be a tight frame in $\mathbb{R}^d$ with frame bound λ and dual frame $\mathcal{G} = \frac{1}{\lambda}\mathcal{F}$, $d \ge 2$. Assume that $0 < a \le \|\mathbf{v}_j\| \le b$ for all j.*

(1) *There exists a permutation $p \in S_N$ and constants C_1, C_2 and B_1, B_2 depending only on d and b such that*

$$\sigma_1(\mathcal{F}, p) \le C_1\, N^{1-\frac{1}{d}} + B_1, \qquad \sigma_2(\mathcal{F}, p) \le C_2\, N^{1-\frac{2}{d}} + B_2.$$

Furthermore, we may choose $C_1 = 4\sqrt{d+3}\,b$, $C_2 = C_1^2 = 16(d+3)b^2$, $B_1 = \sqrt{d}\,b$ and $B_2 = B_1^2 = d\,b^2$.

(2) *Assume that for the above p the nonzero elements in $\{\sum_{j=1}^k \mathbf{v}_{p(j)}\}_{k=1}^N$ are pairwise linearly independent over $\mathbb{Q}$. Then as $\Delta \to 0^+$,*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}_p\|^2) \le \frac{Cb^2}{a^4 N^{1+\frac{2}{d}}}\Delta^2 + o(\Delta^2).$$

where C depends on d only.

Proof. (1) This follows directly from Lemma 3.4, with the observation that $\mathcal{F} \subset [-b, b]^d$. By rescaling the box to $[-\frac{1}{2}, \frac{1}{2}]^2$ we prove part (1) immediately.

(2) By Proposition 3.2 we have

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}_p\|^2) = \frac{\Delta^2}{12\lambda^2}\sigma_2(\mathcal{F}, p) + o(\Delta^2).$$

Now the proof in Theorem 3.3 already establishes the inequality $\lambda \geq \frac{a^2 N}{d}$. Thus

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}_p\|^2) \leq \frac{d^2}{12N^2a^4}\sigma_2(\mathcal{F}, p)\Delta^2 + o(\Delta^2).$$

Part (2) of the theorem now follows from part (1). Note that part (1) also allows us to write down a specific C. ∎

Remark 3.2. The condition that the nonzero elements in $\{\sum_{j=1}^{k} \mathbf{v}_{p(j)}\}_{k=1}^{N}$ are pairwise linearly independent over $\mathbb{Q}$ is not a strong condition. It is satisfied by a generic frame $\mathcal{F}$. Even when the condition is violated, it can often be restored with a simple cyclic permutation, or switching the order of two adjacent elements. The estimates are maintained with these operations.

Remark 3.3. Theorem 3.5 shows that in general sigma-delta scheme is superior to the PCM scheme, at least from the MSE point of view, when N is large and Δ is small. For the PCM scheme, under the same assumptions the MSE is in the order of $C'\,N^{-1}\Delta^2 + o(\Delta^2)$ where C' depends on a, b, d, see [18].

Theorem 3.6. *Let $\mathcal{F} = \{\mathbf{v}_j\}_{j=1}^{N}$ be a unit norm tight frame in $\mathbb{R}^d$ with dual frame $\mathcal{G} = \frac{1}{\lambda}\mathcal{F} = \frac{d}{N}\mathcal{F}$, $d \geq 2$.*

(1) *There exists a permutation $p \in S_N$ and constants C_1, C_2 and B_1, B_2 depending only on d such that*

$$\sigma_1(\mathcal{F}, p) \leq C_1\, N^{1-\frac{1}{d-1}} + B_1, \qquad \sigma_2(\mathcal{F}, p) \leq C_2\, N^{1-\frac{2}{d-1}} + B_2.$$

(2) *Assume that for the above p the nonzero elements in $\{\sum_{j=1}^{k} \mathbf{v}_{p(j)}\}_{k=1}^{N}$ are pairwise linearly independent over $\mathbb{Q}$. Then as $\Delta \to 0^+$,*

$$\mathcal{E}(\|\mathbf{x} - \tilde{\mathbf{x}}_p\|^2) \leq \frac{C}{N^{1+\frac{2}{d-1}}}\Delta^2 + o(\Delta^2).$$

where C depends on d only.

Proof. (1) Recall that part (1) of Theorem 3.5 follows from Lemma 3.4, which follows from the space-filling Peano curve argument. To prove part (1) here we need to show that there exists a curve $\phi : [0, 1] \longrightarrow S^{d-1}$ that is surjective and Lipschitz-$\frac{1}{d-1}$. This is well known and quite easily seen. Since S^{d-1} is a $d-1$ dimensional manifold we can cut it up into finitely many pieces $P_1, P_2, \ldots, P_m$ such that for each j there exists a map $\psi_j : [-\frac{1}{2}, \frac{1}{2}]^{d-1} \longrightarrow P_j$ such that $|\psi_j(x) - \psi_j(y)| \leq |x - y|$. Let ϕ_0 be any Lipschitz-$\frac{1}{d-1}$ Peano curve from $[0, 1]$ to

$[-\frac{1}{2},\frac{1}{2}]^{d-1}$. Then $\psi_j \circ \phi_0$ is a space-filling curve from $[0,1]$ to P_j, and it is Lipschitz-$\frac{1}{d-1}$. Now a space-filling curve from $[0,1]$ to S^{d-1} can be constructed by joining together the curves $\psi_j \circ \phi_0$ (we can use any smooth curve to join together $\psi_j \circ \phi_0$ to the next one).

With this space-filling curve the argument used in Theorem 3.5 can now be used to prove (1).

To prove (2) we apply (1) and the same argument used to prove the second part of Theorem 3.5. ∎

4. Implementation

The space-filling curve heuristics used in the previous section for the traveling salesman problem, while concise and elegant mathematically, yields no clue as to how the permutations in Lemma 3.4 and Theorem 3.5 can be found. In fairness, given the very construction of Peano curves (see [19]) it is entirely possible to make the space-filling curve heuristics constructive. Nevertheless, we shall not attempt such a feat here in this paper. Instead, we outline a new method to find "good" permutations p of S_N based on recursion that will lead to small $\sigma_1(\mathcal{S},p)$ and $\sigma_2(\mathcal{S},p)$. In fact, the results even slightly improves those in Lemma 3.4, and the method is very easy to implement.

Lemma 4.1. *Let $k_1+k_2+\cdots+k_m=N$ and $k_j\geq 0$. Let $0<s\leq 1$. Then*

$$k_1^s+k_2^s+\cdots+k_m^s\leq m^{1-s}N^s.$$

Proof. It is a simple exercise in Lagrange multipliers that the maximum of $k_1^s+k_2^s+\cdots+k_m^s$ given $k_1+k_2+\cdots+k_m=N$ is achieved when all k_j's are equal, $k_j=N/m$. The lemma immediately follows. ∎

Theorem 4.2. *Let $\mathcal{S}\subset[-\frac{1}{2},\frac{1}{2}]^d$ with $d\geq 3$ and $|\mathcal{S}|=N$. There exists a $p\in S_N$ such that*

$$\sigma_1(\mathcal{S},p)\leq 2\sqrt{d+3}\,N^{1-\frac{1}{d}}-2\sqrt{d+3},\qquad \sigma_2(\mathcal{S},p)\leq 4(d+3)\,N^{1-\frac{2}{d}}-4(d+3). \tag{4.1}$$

Proof. we give a constructive proof using induction on $|\mathcal{S}|$. This proof easily leads to a recursive algorithm for finding such a permutation p.

Clearly, if $|\mathcal{S}| = 1$ then both $\sigma_1(\mathcal{S}, p) = 0$ and $\sigma_2(\mathcal{S}, p) = 0$, and the theorem holds. We also have obvious upper bounds

$$\sigma_1(\mathcal{S}, p) \leq (|\mathcal{S}| - 1)\sqrt{d}, \qquad \sigma_2(\mathcal{S}, p) \leq (|\mathcal{S}| - 1)d.$$

One can easily check that the theorem holds for all $\mathcal{S}$ with $|\mathcal{S}| \leq 5$. Assume that the theorem holds when $|\mathcal{S}| < N$, where $N \geq 6$. We prove the theorem when $|\mathcal{S}| = N$.

Divide the unit cube $[-\frac{1}{2}, \frac{1}{2}]^d$ into 2^d disjoint cubes of size $\frac{1}{2}$ and call them $F_1, F_2, \cdots, F_{2^d}$. Let $\mathcal{S}_j = F_j \cap \mathcal{S}$ and $k_j = |\mathcal{S}_j|$. Without loss of generality we may assume that $\mathcal{S}_j \neq \emptyset$ for $1 \leq j \leq m$ while the rest are empty. So we have $k_j \geq 1$ for $1 \leq j \leq m$ and $\sum_{j=1}^m k_j = N$.

We first assume $m > 1$. So each $k_j < N$. By the induction hypothesis there exists a permutation p_j of the elements in $\mathcal{S}_j$, $1 \leq j \leq m$, such that

$$\sigma_1(2\mathcal{S}_j, p_j) \leq 2\sqrt{d+3}\, k_j^{1-\frac{1}{d}} - 2\sqrt{d+3}, \qquad \sigma_2(2\mathcal{S}_j, p_j) \leq 4(d+3)\, k_j^{1-\frac{2}{d}} - 4(d+3). \tag{4.2}$$

Now we order the elements of $\mathcal{S}$ as follows: The first k_1 elements are those in $\mathcal{S}_1$ in the order defined by the permutation p_1; the next k_2 elements are those in $\mathcal{S}_2$ in the order defined by the permutation p_2, and so on. Let p be the permutation defining this ordering of $\mathcal{S}$. We now have

$$\sigma_1(\mathcal{S}, p) \leq \frac{1}{2}\sum_{j=1}^m \sigma_1(2\mathcal{S}_j, p_j) + (m-1)\sqrt{d} \tag{4.3}$$

$$\sigma_2(\mathcal{S}, p) \leq \frac{1}{4}\sum_{j=1}^m \sigma_2(2\mathcal{S}_j, p_j) + (m-1)d. \tag{4.4}$$

Note the terms $(m-1)\sqrt{d}$ and $(m-1)d$ in (4.3) and (4.4) represent the jumps from $\mathcal{S}_j$ to $\mathcal{S}_{j+1}$. Combining (4.2) and (4.3) we obtain

$$\begin{aligned}\sigma_1(\mathcal{S}, p) &\leq \sqrt{d+3}\sum_{j=1}^m k_j^{1-\frac{1}{2}} - m\sqrt{d+3} + (m-1)\sqrt{d} \\ &\leq \sqrt{d+3}\, m^{\frac{1}{d}} N^{1-\frac{1}{d}} - m\sqrt{d+3} + (m-1)\sqrt{d},\end{aligned}$$

where the second inequality is a result of Lemma 4.1. Now if $m \geq d+2$ then one can easily check that $-m\sqrt{d+3} + (m-1)\sqrt{d} \leq -2\sqrt{d+3}$. Hence in this case, combining the fact that $m \leq 2^d$ so $m^{\frac{1}{d}} \leq 2$, we have

$$\sigma_1(\mathcal{S}, p) \leq 2\sqrt{d+3} N^{1-\frac{1}{2}} - 2\sqrt{d+3}.$$

If on the other hand $m \leq d+1$ then

$$\sigma_1(\mathcal{S}, p) \leq 2\sqrt{d+3}N^{1-\frac{1}{2}} - 2\sqrt{d+3} - C_1$$

where

$$\begin{aligned} C_1 &= (2 - m^{\frac{1}{d}})\sqrt{d+3}N^{1-\frac{1}{d}} + (m-2)\sqrt{d+3} - (m-1)\sqrt{d} \\ &\geq \Big((2 - m^{\frac{1}{d}})N^{1-\frac{1}{d}} - 1\Big)\sqrt{d+3} \\ &\geq \Big((2 - \sqrt[d]{d+1})N^{1-\frac{1}{d}} - 1\Big)\sqrt{d+3} \\ &\geq \Big((2 - \sqrt[3]{4})6^{1-\frac{1}{3}} - 1\Big)\sqrt{d+3} \\ &> 0. \end{aligned}$$

To prove the bound for $\sigma_2(\mathcal{S}, p)$ we use (4.4). By Lemma 4.1 we have

$$\begin{aligned} \sigma_2(\mathcal{S}, p) &\leq \frac{1}{4}\sum_{j=1}^{m} \sigma_2(2\mathcal{S}_j, p_j) + (m-1)d \\ &\leq (d+3)\sum_{j=1}^{m} k_j^{1-\frac{2}{d}} - m(d+3) + (m-1)d \\ &\leq (d+3)m^{\frac{2}{d}}N^{1-\frac{2}{d}} - m(d+3) + (m-1)d \\ &= 4(d+3)N^{1-\frac{2}{d}} - 4(d+3) - C_2, \end{aligned}$$

where

$$C_2 = (4 - m^{\frac{2}{d}})(d+3)N^{1-\frac{2}{d}} + (m-4)(d+3) - (m-1)d.$$

Note that $m \leq 2^d$ and $(m-4)(d+3) - (m-1)d \geq 0$ if $m \geq d+4$, so $C_2 \geq 0$ whenever $m \geq d+4$. Also, we can rewrite C_2 as

$$C_2 = (4 - m^{\frac{2}{d}})(d+3)N^{1-\frac{2}{d}} - 3(d+3) + 3(m-1).$$

If $m \leq d+3$ and $d \geq 4$ then $m^{\frac{2}{d}} \leq (d+3)^{\frac{2}{d}} \leq 7^{\frac{2}{4}}$. Thus for $N \geq 6$ we have

$$C_2 \geq (4 - \sqrt{7})(d+3)6^{1-\frac{2}{4}} - 3(d+3) + 3(m-1) > 0.$$

Finally, for $d = 3$ and $m \leq d+3$ one can check in all cases we have $C_2 > 0$. Hence

$$\sigma_2(\mathcal{S}, p) \leq 4(d+3)N^{1-\frac{2}{d}} - 4(d+3).$$

It remains to complete the induction for the case $m = 1$. If so, it means all points in $\mathcal{S}$ are concentrated in the cube $\mathcal{S}_1$, which has size $\frac{1}{2}$. Thus

$$\sigma_2(\mathcal{S}, p) = \frac{1}{4}\sigma_2(2\mathcal{S}_1, p) < \sigma_2(2\mathcal{S}_1, p).$$

Since $2\mathcal{S}_1$ is contained in a unit cube, we may now simply repeat the proof for $2\mathcal{S}_1$. ∎

Remark. The proof does not work in $d = 2$. While the estimate for $\sigma_1(\mathcal{S}, p)$ can still be made in this case, the estimate for $\sigma_2(\mathcal{S}, p)$ fails.

References

[1] J. Beardwood, J. Halton, J. M. Hammersley, The shortest path through many points, *Proc. Cambridge Philos. Soc.* **55** (1959), 299–327.

[2] J. Benedetto and M. Fickus, Finite normalized tight frames, *Advances in Computational Mathematics*, **18**, (2003) 357–385. 2003.

[3] J. Benedetto, A. M. Powell, and Ö Yılmaz, Sigma-Delta ($\Sigma\Delta$) quantization and finite frames, *Preprint*

[4] J. Benedetto, A. M. Powell, and Ö Yılmaz, Second order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions, *Preprint*

[5] W. Bennett, Spectra of quantized signals, *Bell Syst.Tech.J* **27** (1948) 446–472.

[6] P. G. Casazza and J. Kovačević, Uniform tight frames with erasures, *Advances in Computational Mathematics*, **18**, (2003) 387–430.

[7] CI. Daubechies and R. DeVore, Reconstructing a bandlimited function from very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order, *Annals of Math.*, **158** (2003), no. 2, 679–710.

[8] R. J. Duffin and A. C. Schaeffer, A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.* **72** (1952), 341–366.

[9] Y. Elder and G. D. Forney, Optimal tight frames and quantum measurement, Preprint.

[10] D. J. Feng, L. Wang, and Y. Wang, Generation of finite tight frames by Householder transformations, *Advances in Computational Mathematics*, to appear.

[11] L. Few, The shortest path and the shortest road through n points, *Mathematika* **2** (1955), 141–144.

[12] J. Gao and J. M. Steele, Sums of squares of edge lengths and spacefilling curve heuristics for the traveling salesman problem, *SIAM J. Discrete Math.*, **7** (1994), no. 2, 314–324.

[13] L. Goddyn, Quantizers and the worst-case Euclidean traveling salesman problem, *J. Combin. Theory Ser. B*, **50** (1990), no. 1, 65–81.

[14] V. K. Goyal, J. Kovačcević, and J. Kelner, Quantized frame expansions with erasures, *Appl. Comput. Harmon. Anal.*, **10**, (2001) 203–233.

[15] V. K. Goyal, M. Vetterli and N. T. Thao, Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms, *IEEE Trans. Inform. Theory*, **44** (1998), 16–31.

[16] R. Gray, Quantized noise spectra, *IEEE Trans. Inform. Theory*, **36** (1990), no. 6, 1220–1244.

[17] S. Güntürk, Approximating a bandlimited function using very coarsely quantized data, *J. Amer. Math. Soc.*, to appear.

[18] D. Jimenez, L. Wang and Y. Wang, PCM quantization errors and the white noise hypothesis, submitted to *SIAM J. Math. Analysis*.

[19] S. Milne, Peano curves and smoothness of functions, *Adv. in Math.*, **35** (1980), no. 2, 129–157.

[20] S. Moran, On the length of optimal TSP circuits in sets of bounded diameter, *J. Combin. Theory Ser. B*, **37** (1984), no. 2, 113–141.

[21] D. J. Newman, *A Problem Seminar*, Springer-Verlag, New York-Berlin, 1982

[22] G. Rath and C. Guillemot, Recent advances in DFT codes based on quantized finite frame expansions with erasure channels, preprint.

[23] T. Snyder and J. M. Steele, A priori bounds on the Euclidean traveling salesman, *SIAM J. Comput.*, **24** (1995), no. 3, 665–671.

[24] Steele, J. M., *Probabilistic Theory and Combinatorial Optimization*, SIAM, Philadelphia 1997.

[25] N. Thao and M. Vetterli, Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimate, *IEEE Trans. Signal Proc.*, **42** (1994), no. 3, 519–531.

School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.
E-mail address: wang@math.gatech.edu