High-performance quantization for spectral super-resolution

C. Sinan Güntürk
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA
gunturk@cims.nyu.edu

Weilin Li
Courant Institute of Mathematical Sciences
New York University
New York, NY, USA
weilinli@cims.nyu.edu

Abstract—We show that the method of distributed noise-shaping beta-quantization offers superior performance for the problem of spectral super-resolution with quantization whenever there is redundancy in the number of measurements. More precisely, if the (integer) oversampling ratio λ is such that ⌈M/λ⌉−1 ≥ 4/Δ, where M denotes the number of Fourier measurements and Δ is the minimum separation distance associated with the atomic measure to be resolved, then for any number K ≥ 1 with the atomic measure to be resolved, then for any number K ≥ 1, the number of quantization levels available for the real and imaginary parts of the measurements, our quantization method guarantees reconstruction accuracy of order O(λ^3/2K^−λ/2), up to constants which are independent of K and λ. In contrast, memoryless scalar quantization offers superior performance for the spectral super-resolution problem. Super-resolution has received considerable attention in the past several years (e.g. [3], [4]). The goal is to accurately estimate an unknown discrete measure μ = \sum_{j=1}^{S} a_j δ_{t_j} \hspace{1cm} (1)

defined on \mathbb{T} := [0,1), from its noisy samples \tilde{y}_k = y_k + z_k, \quad k = 0, \ldots, M − 1,

where

\begin{align}
\tilde{y}_k := \hat{\mu}(k) := \int_0^1 e^{-2\pi i k t} \, d\mu(t) = \sum_{j=1}^{S} a_j e^{2\pi i k t_j}. \hspace{1cm} (2)
\end{align}

is the k-th Fourier coefficient of μ and the unknown noise vector \( z := (z_k)_{k=0}^{M-1} \) satisfies \( \|z\|_2 \leq \varepsilon \) for some known \( \varepsilon > 0 \). We emphasize that the total number of spikes S, the amplitudes \( a \in \mathbb{C}^S \), and the support set \( T = \{t_j\}_{j=1}^{S} \) are unknown.

This problem is ill-conditioned if there are points in T that are too close to one another (e.g. [10], [9], [13]). However, assuming a lower bound to their minimum separation, it has been shown that \( \mu \) can be recovered from its measurements in a robust way when M is sufficiently large, meaning that the reconstruction error, when measured in a suitable metric, is controlled by the noise energy in a graceful manner, and typically linearly (e.g. [12]).

Since no structure is assumed on the noise, robustness also becomes the key property that allows for quantization. In particular, it guarantees that sufficiently high-resolution quantizers will always produce sufficiently high quality approximations. However, the question of achievable limits of quantization accuracy is open. The answer depends on the interplay between the given (fixed) parameters, such as the number of measurements, the number of quantization levels per measurement, and the minimum separation distance of the measures of interest, as well as the quantization method which itself is a design parameter.

For simplicity, we will assume in this paper that quantization is the only source of perturbation. Other (generally uncontrolled) sources of perturbations can be incorporated into quantization as well; any such perturbations typically provide a noise floor. We assume that the real and imaginary parts...
of each Fourier measurement is replaced by an element of a (signal-independent) quantization alphabet $\mathcal{A}$ with $K$ levels.

The linear dependence of the reconstruction error on the noise energy implies that with simple rounding, i.e. MSQ, it is straightforward to achieve reconstruction accuracy of order $O(K^{-1})$. However, since the noise energy is measured in $\ell_2$, its bound $\varepsilon$ grows with the number of measurements, and therefore it is not even clear if there is any advantage of using any additional measurements. For example, the popular reconstruction method total-variation minimization (TV-min, see Section III) appears to be indifferent to oversampling in practice.

Our main result in this paper is that there is an alternative quantization method, called the distributed noise-shaping $\beta$-encoder (or $\beta$-quantization in short) which, together with an accompanying alternative recovery method derived from TV-min, is able to exploit any available redundancy. More precisely, if the (integer) oversampling ratio $\lambda$ is such that $[M/\lambda] - 1 \geq 4/\Delta$, where $M$ denotes the number of Fourier measurements and $\Delta$ is the minimum separation distance associated with $\mu$, then for any number $K$ of quantization levels, our quantization method guarantees reconstruction accuracy of order $O((\lambda/2)^2K^{-\lambda/2})$, up to constants which are independent of $K$ and $\lambda$. In principle our method can work with other robust recovery methods, too.

The paper is organized as follows. Section II reviews the TV-min method and discusses MSQ for spectral super-resolution. In Section II, we introduce the proposed quantization method and the main ingredients needed for its error performance analysis for spectral super-resolution, which is done in Section IV. Finally, we provide a sample numerical experiment to demonstrate the practical performance of our proposed quantization method in Section V.

II. A REVIEW OF TV-MIN FOR SUPER-RESOLUTION AND MSQ

There are a number of robust recovery algorithms for super-resolution. For convenience and concreteness, we will focus on TV-min, also known under the name BLASSO (Beurling Lasso).

Let us denote by $F_M$ the operator which maps the measure $\mu$ to $(\tilde{\mu}(k))_0^{M-1}$, i.e. with our notation of (2), we have $y = F_M \mu$, where $y := (y_k)_0^{M-1}$. Given noisy data $\tilde{y} \in \mathbb{C}^M$ and a bound $\varepsilon > 0$ on the noise, the TV-min algorithm outputs an estimate $\tilde{\mu}$ of $\mu$ given by

$$\tilde{\mu} := \arg \min \{ \|\nu\|_{TV} : \|F_M \nu - \tilde{y}\|_2 \leq \varepsilon \}. \quad (3)$$

This is a convex program whose feasibility is guaranteed by the assumption $\|y - \tilde{y}\|_2 \leq \varepsilon$. The solution may not be unique, but it is known that there is at least one minimizer which is a discrete measure which we will identify with $\tilde{\mu}$. (See [4], [3], [11], [2] for this and other results.)

The performance of TV-min depends on the minimum separation of the measure, defined as

$$\Delta(\mu) := \min_{s \neq t, s,t \in \text{supp}(\mu)} |s - t|_T. \quad (4)$$

Here, $|s - t|_T := \min_{n \in \mathbb{Z}} |s - t - n|$ is the “wrap-around” metric on $\mathbb{T}$. If the number of samples $M$ is sufficiently large so that

$$\Delta(\mu) \geq \frac{4}{M - 1}, \quad (5)$$

then $\tilde{\mu}$ provides an accurate estimate of $\mu$ in the following sense: any spurious spikes in $\tilde{\mu}$ are smaller than $O(\varepsilon)$, the remaining spikes in $\tilde{\mu}$ are within $O(1/M)$ of the true spikes, and the recovered amplitudes are within $O(\varepsilon)$ of the true ones. More precisely, given the representation

$$\tilde{\mu} = \sum_{k=1}^S \tilde{a}_k \delta_{\tilde{t}_k}, \quad (6)$$

along with the “neighborhood” index sets

$$\mathcal{I}_j^M := \{ k : |\tilde{t}_k - t_j|_T \leq 2 \cdot 0.1649(M - 1)^{-1}, j = 1, \ldots, S \}, \quad (7)$$

and the residual index set

$$\mathcal{I}_0^M := \{1, \ldots, S\} \setminus \bigcup_{j=1}^S \mathcal{I}_j^M, \quad (8)$$

the following bounds are guaranteed:

$$\left| a_j - \sum_{k \in \mathcal{I}_j^M} \tilde{a}_k \right| \leq C_1 \varepsilon, \quad j = 1, \ldots, S, \quad (9)$$

$$\sum_{k \in \mathcal{I}_j^M} |\tilde{a}_k|^2 |t_j - \tilde{t}_k|_T^2 \leq C_2 M^{-2} \varepsilon, \quad j = 1, \ldots, S, \quad (10)$$

$$\sum_{k \in \mathcal{I}_0^M} |\tilde{a}_k| \leq C_3 \varepsilon. \quad (11)$$

Additional details can be found in [12, Theorem 1.2]; see also [1, Theorems 2.1 and 2.2] for related results.

This result provides an immediate error bound for MSQ. For each integer $K \geq 2$, let $Z_K$ denote the $K$-term origin-symmetric arithmetic progression of integers with spacing 2 and $A_K := K^{-1}Z_K \subset (-1, 1)$. It is then clear that for all $u \in [-1, 1]$ there exists $q \in A_K$ such that $|u - q| \leq 1/K$. Assuming that $\|\mu\|_{TV} \leq 1$ so that $|y|_{\infty} \leq 1$, it follows that for each complex measurement $y_k$, there is an element $q_k \in A_K + iA_K$ (found by separately rounding the real and the imaginary parts of $y_k$ to elements of $A_K$) such that $|y_k - q_k| \leq \sqrt{2}/K$. Consequently, we have $\|y - q\|_2 \leq \sqrt{2M}/K$. Setting $\bar{y} = q$ and $\varepsilon = \sqrt{2M}/K$ in (3) guarantees, in view of (9)-(11), an overall reconstruction accuracy of $O(\sqrt{M}/K)$.

We can produce a lower bound on the worst-signal reconstruction error of MSQ as follows: Even if we knew the support $T$ of $\mu$, where $|T| = S$, memoryless scalar quantization of the $M$ linear measurements of all possible coefficient vectors $a \in \mathbb{C}^S$ chosen from any fixed ball results in a partition of this ball using at most $O(MK)$ hyperplanes, and therefore into at most $(eMK/S)^S$ cells. (Here $e$ is an absolute constant.) Consequently, there will always be a cell of diameter at least $O(S/MK)$ whose elements are all mapped to the same quantized vector.

Therefore, suppressing the dependence on $M$, it follows that MSQ cannot offer error performance better than $O(K^{-1})$. 
III. PROPOSED QUANTIZATION METHOD

Our proposed quantization approach in this paper is based on the general framework of distributed noise-shaping β-encoding developed in [6], [7] (see also [5] and [8] for prior versions). However, the specialization for the spectral super-resolution problem requires some new choices and adaptations.

Let λ ≥ 1 be an integer which should be thought of as a lower bound on the oversampling ratio. For the simplicity of discussion, we assume M is divisible by λ and set m := M/λ.

Let β > 1 be a parameter which shall be chosen later, and consider the m × M matrix

\[ V := \begin{bmatrix} I_m & \beta^{-1}I_m & \cdots & \beta^{-λ+1}I_m \end{bmatrix}, \]

where \( I_m \) denotes the m × m identity matrix. Observe that

\[ (Vy)_ℓ = \sum_{k=0}^{λ-1} \beta^{-k}y_{mk+ℓ} = \sum_{j=1}^{S} a_j w_j e^{-2πiℓt_j}, \]

for ℓ = 1, ..., m, where

\[ w_j := \frac{1 - β^{-λ}e^{-2πimλj}}{1 - β^{-1}e^{-2πimj}}, \quad j = 1, \ldots, S. \]  

In other words, we have the relation \( V_y = F_mμ_V \) where

\[ μ_V := \sum_{j=1}^{S} b_j δ_{t_j} \quad \text{and} \quad b_j := a_j w_j. \]

Observe that μ and μ_V have identical supports, but different amplitudes. However, the weights \( w_j \) satisfy

\[ \frac{1}{c_β} |w_j| ≤ c_β \quad \text{where} \quad c_β := \frac{1 + β^{-1}}{1 - β^{-1}}. \]  

Let us define \( H := H_β \) to be the \( M \times M \) matrix where

\[ H_{j,k} := \begin{cases} 1, & \text{if } j = k, \\ -β, & \text{if } j = k + m \text{ and } 1 ≤ k ≤ M - m. \end{cases} \]

The following is a special case of [7, Lemma 2]:

**Lemma 1.** Let \( K ≥ 2 \) be an integer, and suppose the parameters \( α, β, δ > 0 \) satisfy the inequality

\[ β + αδ^{-1} ≤ K, \]

and consider the quantization alphabet \( A := \{Z_K + iZ_K\} \).

Then, for any \( y ∈ C^K \) with \( \|y\|∞ ≤ α \), there exists \( q ∈ A^K \) and \( u ∈ C^K \) with \( \|u\|∞ ≤ \sqrt{2δ} \) satisfying the relationship

\[ y - q = Hu. \]

The mapping \( y → q \) implied by the above lemma can be implemented by means of a simple recursive algorithm. We omit the details and refer to [6], [7].

The significance of \( V \) and \( H \) is that \( VH \) is very small when \( β \) or \( λ \) is large. Indeed, as shown in [6], we have \( \|VH\|∞ → 2 = \sqrt{mβ^{-λ+1}} \). The immediate consequence is that

\[ \|Vy - Vq\|₂ ≤ \|VH\|∞ → 2 \|u\|∞ ≤ \sqrt{2mβ^{-λ+1}}δ. \]

The above findings provide the core strategy of our proposed quantization and recovery method. First, we note that for any \( 1 < β < K \) and any \( 0 ≤ α < ∞ \), there exists \( δ > 0 \) such that the condition (17) is satisfied. Hence the existence of the mapping \( y → q \) (with a fixed quantization alphabet \( A \)) is guaranteed over any bounded set of inputs \( y \). Next, recall that with \( μ = F_mμ \), we have \( Vq = F_mμ \). Since \( Vq \) is now a small perturbation of \( Vq \), we can obtain a close approximation \( \tilde{μ}_V \) of \( μ_V \) by means of any robust super-resolution recovery method, such as the TV-min algorithm. Then, since \( μ \) and \( μ_V \) have identical supports, we can define an approximate recovery \( \tilde{μ} \) by means of approximate weights \( \tilde{w}_j \) derived from the approximate support \( T \) of \( \tilde{μ}_V \).

Let us summarize the proposed quantization method.

**System parameters and assumptions:**

- \( M \) (number of Fourier measurements),
- \( α \) (upper bound on TV-norm of the input measures),
- \( Δ \) (lower bound on minimum separation distance),
- \( M = λm, λ ≥ 1, m - 1 ≥ 4/Δ, \)
- \( K \) (number of quantization levels),
- \( β \) and \( δ \) such that (17) holds.

**Encoding (quantization) stage:**

- **Input to quantizer:** \( y \) such that \( \|y\|∞ ≤ α \),
- \( V \) and \( H \) defined via (12) and (16),
- Quantization alphabet: \( A := \{Z_K + iZ_K\} \),
- Output of quantizer: \( q ∈ A^K \) such that \( \|Vq - Vq\|₂ ≤ \sqrt{2mβ^{-λ+1}}δ \).

**Decoding (recovery) stage:**

- **Input to decoder:** \( q \),
- Compute a minimum TV-norm measure \( \tilde{μ}_V \) of the form \( \sum_{k=1}^{S} \tilde{a}_k δ_{tk} \) satisfying \( \|F_mμ_V - Vq\|₂ ≤ ε_V \) where
  \[ ε_V := \sqrt{2mβ^{-λ+1}}δ; \]
- abort if it cannot be found (e.g. invalid measurements),
- Set \( \tilde{w}_k := \frac{1 - β^{-λ}e^{-2πimλk}}{1 - β^{-1}e^{-2πimk}}, \quad k = 1, \ldots, S; \)
- Output of decoder: \( \tilde{μ} := \sum_{k=1}^{S} \tilde{a}_k δ_{tk} \) with \( \tilde{a}_k := \tilde{w}_k/\tilde{w}_k \).

IV. ERROR ANALYSIS

Let us start by noting that when the input to the quantizer \( y \) equals \( F_mμ \) for some \( μ \) of the form (1) with \( Δ(μ) ≥ Δ \) and \( \|μ\|TV ≤ α \), then the decoder will always output a measure \( \tilde{μ} \), thanks to the fact that \( Vq = F_mμ \) where \( μ_V \) is defined by (14) which guarantees that \( μ_V \) is a feasible measure for the TV-min program.

Let us now proceed to find an error bound for \( \tilde{μ} \). We start by comparing \( μ_V \) to \( μ_V \). With the error bounds of the general TV-min method reviewed in Section II, we have

\[ |b_j - \sum_{k ∈ Z_m^n} \tilde{b}_k| ≤ C_1 ε_V, \quad j = 1, \ldots, S, \]

\[ \sum_{k ∈ Z_m^n} \|b_k|t_j - \tilde{t}_k\|₂^2 ≤ C_2 m^{-2} ε_V, \quad j = 1, \ldots, S, \]

\[ \sum_{k ∈ Z_m^n} \|\tilde{b}_k| ≤ C_3 ε_V. \]
where the index sets $\mathcal{I}_j^m$, $j = 0, \ldots, S$ are as in (7) and (8), only for $m$ measurements. Note that for all $j \in \{1, \ldots, S\}$,
\[ |a_j - \sum_{k \in \mathcal{I}_j^m} \tilde{a}_k| \leq \frac{1}{|w_j|} |b_j - \sum_{k \in \mathcal{I}_j^m} \tilde{b}_k| + \sum_{k \in \mathcal{I}_j^m} |\tilde{b}_k| \frac{1}{|w_j|} - \frac{1}{|\tilde{w}_k|}. \]  
(23)

With (15) and (20), the first term is bounded by $c_\beta C_1 \varepsilon_V$. For the second term, we note that
\[ |w_j^{-1} - \tilde{w}_k^{-1}| \leq m C_{\beta,\lambda} |t_j - \tilde{t}_k| \]  
(24)

where $C_{\beta,\lambda}$ stands for the Lipschitz constant of the map
\[ t \mapsto \frac{1 - \beta^{-1} e^{-2\pi i t}}{1 - \beta^{-1} e^{-2\pi i t}}, \quad t \in \mathbb{T}. \]

It can be shown $C_{\beta,\lambda} \leq 4\pi \lambda \beta (\beta - 1)^{-2}$.

Using (24) and Cauchy-Schwarz, we see that the second term in (23) is bounded by
\[ m C_{\beta,\lambda} \left( \sum_{k \in \mathcal{I}_j^m} |\tilde{b}_k| |t_j - \tilde{t}_k|^2 \right)^{1/2} \left( \sum_{k \in \mathcal{I}_j^m} |\tilde{b}_k|^2 \right)^{1/2} \]

Note that $\|\tilde{b}\| \leq c_\beta \alpha$ since $\|\tilde{b}\| = \|\tilde{\mu}_V\|_{TV}$ and
\[ \|\tilde{\mu}_V\|_{TV} \leq \|\mu_V\|_{TV} \leq \|w\|_{\infty} \|a\|_1 \leq c_\beta \|\mu\|_{TV} \leq c_\beta \alpha. \]

Hence, with (21) we deduce
\[ \sum_{k \in \mathcal{I}_j^m} |\tilde{b}_k| \frac{1}{|w_j|} - \frac{1}{|\tilde{w}_k|} \leq C_{\beta,\lambda} \sqrt{c_\beta \alpha} \sqrt{C_{2\varepsilon_V}}. \]  
(25)

Injecting (25) into (23) we have
\[ |a_j - \sum_{k \in \mathcal{I}_j^m} \tilde{a}_k| \leq c_\beta C_1 \varepsilon_V + C_{\beta,\lambda} \sqrt{c_\beta \alpha} \sqrt{C_{2\varepsilon_V}}. \]  
(26)

Finally, we also have
\[ \sum_{k \in \mathcal{I}_j^m} |\tilde{a}_k| |t_j - \tilde{t}_k|^2 \leq c_\beta C_2 m^{-2} \varepsilon_V, \]  
(27)

\[ \sum_{k \in \mathcal{I}_j^m} |\tilde{a}_k| \leq c_\beta C_3 \varepsilon_V. \]  
(28)

which follow readily from (15), (21) and (22).

At this point, we note the following elementary fact: For any $K \geq 2$, setting $\beta := K(\lambda + 1)/(\lambda + 2)$ and $\delta := (\lambda + 2)\alpha/K$ results in $\beta + \alpha \delta^{-1} = K$ and $\delta \beta^{-\lambda+1} < c_\alpha (\lambda + 1) K^{-\lambda}$. (See, e.g. [6, Lemma 3.2] and [7, Lemma 1.]) This choice of parameters results in $\varepsilon_V \leq c_\alpha \sqrt{2m(\lambda + 1) K^{-\lambda}}$. Furthermore it is readily seen that $\beta \geq 4/3$, $c_\beta \leq 7$ and $C_{\beta,\lambda} \leq 12\pi \lambda$. Hence it follows from (26), (27) and (28) that $\mu$ is approximated by $\tilde{\mu}$ up to resolution $O(\sqrt{\lambda \lambda^3/2 K^{-\lambda}/2})$.

V. NUMERICAL RESULTS

We compare the reconstruction error, quantified by the term on the left hand side of (26), when the Fourier samples are quantized using our proposed beta-quantization versus MSQ. More specifically, we set $\Delta = 1/10$ and we randomly select a measure $\mu$ such that $\Delta(\mu) \geq \Delta$; the amplitudes are chosen uniformly at random and normalized to have unit $\ell^1$ norm.

For various choices of $\lambda$ and $K$, we quantize the Fourier measurements using both MSQ and $\beta$-quantization. Figure 1 displays the reconstruction error as a function of $\lambda$, averaged over 110 trials. The experiment validates our theoretical results and also shows that performance of MSQ is suboptimal in the over-sampling regime.

REFERENCES


