Compressed Sensing Off the Grid

Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht
University of Wisconsin-Madison

July 26, 2012

Abstract

We consider the problem of estimating the frequency components of a mixture of s complex sinusoids from a random subset of n regularly spaced samples. Unlike previous work in compressed sensing, the frequencies are not assumed to lie on a grid, but can assume any values in the normalized frequency domain $[0, 1]$. We propose an atomic norm minimization approach to exactly recover the unobserved samples. We reformulate this atomic norm minimization as an exact semidefinite program. Even with this continuous dictionary, we show that most sampling sets of size $O(s \log s \log n)$ are sufficient to guarantee the exact frequency estimation with high probability, provided the frequencies are well separated. Extensive numerical experiments are performed to illustrate the effectiveness of the proposed method.

Keywords. atomic norm, basis mismatch, compressed sensing, continuous dictionary, line spectral estimation, nuclear norm relaxation, Prony's method, sparsity

1 Introduction

Compressed sensing has demonstrated that data acquisition and compression can often be combined, dramatically reducing the time and space needed to acquire many signals of interest [2, 10, 11, 19]. Despite the tremendous impact of compressed sensing on signal processing theory and practice, its development thus far has focused on signals with sparse representation in finite discrete dictionaries. However, signals encountered in applications such as radar, sonar, sensor array, communication, seismology, and remote sensing are usually specified by parameters in a *continuous* domain [23, 37, 47]. In order to apply the theory of compressed sensing to such applications, researchers typically adopt a discretization procedure to reduce the continuous parameter space to a finite set of grid points [1, 3, 21, 24, 28, 34, 38]. While this simple strategy yields state-of-the-art performance for problems where the true parameters lie on the grid, discretization has several significant drawbacks.

One major weakness of discretization is *basis mismatch*, where the true signal cannot even be sparsely represented by the discrete dictionary [16, 21, 29]. One might attempt to remedy basis mismatch by using a finer discretization or adding extra basis elements. However, increasing the size of the dictionary will also increase the correlation between the basis elements. Common wisdom in compressed sensing suggests that low-correlation[1] between dictionary elements is necessary for high fidelity signal recovery, casting doubt as to whether additional discretization is beneficial. Finer gridding also results in higher computational complexity and numerical instability, further diminishing any advantage it might have in sparse recovery applications.

[1] In compressed sensing, the maximum correlation between columns is called the *coherence* of the dictionary

We overcome the issues arising from discretization by working directly on the continuous parameter space for estimating the continuous frequencies and amplitudes of a mixture of complex sinusoids from partially observed time samples. In particular, the frequencies are not assumed to lie on a grid, and can instead take arbitrary values across the bandwidth of the signal. With a time-frequency exchange, our model is exactly the same as the one in Candès, Romberg, and Tao's foundational work on compressed sensing [10], except that we do not assume the frequencies lie on an equispaced grid. This major difference presents a significant technical challenge as the resulting dictionary is no longer an orthonormal Fourier basis, but is an infinite dictionary with continuously many atoms and arbitrarily high correlation between candidate models. We demonstrate that a sparse sum of complex sinusoids can be reconstructed exactly from a small sampling of its time samples provided the frequencies are sufficiently far apart from one another.

Our computational method and theoretical analysis is based upon the *atomic norm* induced by samples of complex exponentials [15]. Chandrasekaran *et al* showed the atomic norm is in some sense the best convex heuristic for underdetermined, structured linear inverse problems, and it generalizes the ℓ_1 norm for sparse recovery and the nuclear norm for low-rank matrix completion. The norm is a convex function, and, in the case of complex exponentials, can be computed via semidefinite programming. Below, we show how the atomic norm for moment sequences can be derived either from the perspective of sparse approximation or rank minimization [40], illuminating new ties between these related areas of study. Much as was the case in other problems where the atomic norm has been studied, we prove that atomic norm minimization achieves nearly optimal recovery bounds for reconstructing sums of sinusoids from incomplete data.

To be precise, we consider signals whose spectra consist of spike trains with unknown locations in a normalized interval $[0, 1]$. Rather than sampling the signal at all times $t = 0, \ldots, n-1$ we sample the signal at a subset of times $t_1, \ldots t_m$ where each $t_j \in \{0, \ldots, n-1\}$. Our main contribution is summarized by the following theorem.

Theorem 1.1. *Suppose we observe the time samples of the signal*

$$x_j^\star = \sum_{k=1}^{s} c_k e^{i2\pi f_k j} \tag{1.1}$$

with unknown frequencies $\{f_1, \ldots, f_s\} \subset [0, 1]$ *on an index set* $T \subset \{0, \ldots, n-1\}$ *of size* m *selected uniformly at random. Additionally, assume* $\operatorname{sign}(c_k)$ *are drawn i.i.d. from the uniform distribution on the complex unit circle and*

$$\Delta_f = \min_{k \neq j} |f_k - f_j|$$

where the distance $|f_k - f_j|$ *is understood as the wrap-around distance on the unit circle. If* $\Delta_f \geq \frac{1}{\lfloor (n-1)/4 \rfloor}$*, then there exists a numerical constant* C *such that*

$$m \geq C \max \left\{ \log^2 \frac{n}{\delta}, s \log \frac{s}{\delta} \log \frac{n}{\delta} \right\},$$

is sufficient to guarantee that we can recover $x^\star$ *via a semidefinite programming problem with probability at least* $1 - \delta$*.*

Once the missing entries are recovered exactly, the frequencies can be identified by Prony's method [17], a matrix pencil approach [31], or other linear prediction methods [45]. After identifying the frequencies, the coefficients $\{c_k\}_{k=1}^s$ can be obtained by solving a linear system.

Remark 1.2. **(Resolution)** An interesting artifact of using convex optimization methods is the necessity of a particular resolution condition on the spectrum of the underlying signal. For the signal to be recoverable via our methods using $O(s \log s \log n)$ random time samples from the set $\{0, 1, \ldots, n-1\}$, the spikes in the spectrum need to be separated by roughly $\frac{4}{n}$. In contrast, if one chose to acquire $O(s \log s \log n)$ *consecutive* samples from this set (equispaced sampling), the required minimum separation would be $\frac{4}{s \log s \log n}$; this sampling regime was studied by Candés and Fernandez-Granda [7]. Therefore, by using random sampling, we increase the resolution from $\frac{4}{s \log s \log n}$, which is what we get using equispaced sampling [7], to $\frac{4}{n}$, i.e., an exponential improvement. We comment that numerical simulations in Section 5 suggest that the critical separation is actually $\frac{1}{n}$. We leave tightening our bounds by the extra constant of 4 to future work.

Remark 1.3. **(Random Signs)** The randomness of the signs of the coefficients essentially assumes that the sinusoids have random phases. Such a model is practical in many spectrum sensing applications as argued in [47, Chapter 4.1]. Our proof will reveal that the phases can obey any symmetric distribution on the unit circle, not simply the uniform distribution.

Remark 1.4. **(Band-limited Signal Models)** Note that any mixture of sinusoids with frequencies bandlimited to $[-W, W]$, after appropriate normalization, can be assumed to have frequencies in $[0, 1]$. Consequently, a bandlimited signal of such a form leads to samples of the form (1.1). More precisely, suppose the frequencies lie in $[-W, W]$, and $x^\star(t)$ is a continuous signal of the form:

$$x^\star(t) = \sum_{k=1}^{s} c_k e^{i2\pi w_k t}.$$

By taking regularly spaced Nyquist samples at $t \in \{0/2W, 1/2W, \ldots, (n-1)/2W\}$, we observe

$$\begin{aligned} x_j^\star := x^\star(j/2W) &= \sum_{k=1}^{s} c_k e^{i2\pi \frac{w_k}{2W} j} \\ &= \sum_{k=1}^{s} c_k e^{i2\pi f_k j} \text{ with } f_k = \frac{w_k}{2W} \in \left[-\frac{1}{2}, \frac{1}{2}\right], \end{aligned}$$

exactly the same as our model (1.1) after a trivial translation of the frequency domain.

Remark 1.5. **(Basis Mismatch)** Finally, we note that our result completely obviates the basis mismatch conundrum of discretization methods, where the frequencies might well fall off the grid. Since our continuous dictionary is globally coherent, Theorem 1.1 shows that the global coherence of the frame is not an obstacle to recovery. What matters more is the local coherence characterized by the separation between frequencies in the true signal.

This paper is organized as follows. First, we specify our reconstruction algorithm as the solution to an atomic norm minimization problem in Section 2. We show that this convex optimization problem can be exactly reformulated as a semidefinite program and that our methodology is thus computationally tractable. We outline connections to prior art and the foundations that we build upon in Section 3. We then proceed to prove Theorem 1.1 in Section 4. Our proof requires the construction of an explicit certificate that satisfies certain interpolation conditions. The production of this certificate requires us to consider certain random polynomial kernels, and derive concentration inequalities for these kernels that may be of independent interest to the reader. In Section 5, we validate our theory by extensive numerical experiments, confirming that random under-sampling as a means of compression coupled with atomic norm minimization as a means of recovery are a viable, superior alternative to discretization techniques.

2 The Atomic Norm and Semidefinite Characterizations

Our signal model is a positive combination of complex sinuoids with arbitrary phases. As motivated in [15], a natural regularizer that encourages a sparse combination of such sinusoids is the *atomic norm* induced by these signals. Precisely, define atoms $a(f,\phi) \in \mathbb{C}^{|J|}$, $f \in [0,1]$ and $\phi \in [0, 2\pi)$ as

$$[a(f,\phi)]_j = \frac{1}{\sqrt{|J|}} e^{i(2\pi f j + \phi)},\ j \in J$$

and rewrite the signal model (1.1) in matrix-vector form

$$x^\star = \sum_{k=1}^{s} |c_k| a(f_k, \phi_k) \tag{2.1}$$

where J is an index set with values being either $\{0,\ldots,n-1\}$ or $\{-2M,\ldots,2M\}$ for some positive integer n and M, and ϕ_k is the phase of the complex number c_k. In the rest of the paper, we use $\Omega = \{f_1,\ldots,f_s\} \subset [0,1]$ to denote the unknown set of frequencies. In the representation (2.1), we could also choose to absorb the phase ϕ_k into the coefficient $|c_k|$ as we did in (1.1). We will use both representations in following and explicitly specify that the coefficient c_k is positive when the phase term ϕ_k is in the atom $a(f_k,\phi_k)$.

The set of atoms $\mathcal{A} = \{a(f,\phi) \ : \ f \in [0,1], \phi \in [0,2\pi)\}$ are building blocks of the signal $x^\star$, the same way that canonical basis vectors are building blocks for sparse signals, and unit-norm rank one matrices are building blocks for low-rank matrices. In sparsity recovery and matrix completion, the unit balls of the sparsity-enforcing norms, e.g., the ℓ_1 norm and the nuclear norm, are exactly the convex hulls of their corresponding building blocks. In a similar spirit, we define an atomic norm $\|\cdot\|_{\mathcal{A}}$ by identifying its unit ball with the convex hull of $\mathcal{A}$

$$\begin{aligned}\|x\|_{\mathcal{A}} &= \inf\{t > 0 : x \in t\,\mathrm{conv}(\mathcal{A})\} \\ &= \inf_{\substack{c_k \geq 0,\ \phi_k \in [0,2\pi) \\ f_k \in [0,1]}} \Big\{ \sum_k c_k : x = \sum_k c_k a(f_k,\phi_k) \Big\}.\end{aligned}$$

Roughly speaking, the atomic norm $\|\cdot\|_{\mathcal{A}}$ can enforce sparsity in $\mathcal{A}$ because low-dimensional faces of $\mathrm{conv}(\mathcal{A})$ correspond to signals involving only a few atoms. The idea of using atomic norms to enforce sparsity for a general set of atoms was first proposed and analyzed in [15].

When the phases ϕ are all 0, the set $\mathcal{A}_0 = \{a(f,0) \ : \ f \in [0,1]\}$ is called the *moment curve* which traces out a one-dimensional variety in $\mathbb{R}^{2|J|}$. It is well known that the convex hull of this curve is characterizable in terms of Linear Matrix Inequalities, and membership in the convex hull can thus be computed in polynomial time (see [42] for a proof of this result and a discussion of many other algebraic varieties whose convex hulls are characterized by semidefinite programming). When the phases are allowed to range in $[0,2\pi)$, a similar semidefinite characterization holds.

Proposition 2.1. *For $x \in \mathbb{C}^{|J|}$ with $J = \{0,\ldots,n-1\}$ or $\{-2M,\ldots,2M\}$,*

$$\|x\|_{\mathcal{A}} = \inf\left\{ \tfrac{1}{2}\,\mathrm{trace}(\mathrm{Toep}(u)) + \tfrac{1}{2}t \ : \ \begin{bmatrix} \mathrm{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0 \right\}.$$

In the proposition, we used the superscript * to denote conjugate transpose and $\mathrm{Toep}(u)$ to denote the Toeplitz matrix whose first column is equal to u. The proof of this proposition relies on the following classical Vandermonde decomposition Lemma for positive semidefinite Toeplitz matrices

Lemma 2.2 (Caratheodory-Toeplitz, [12,13,48])**.** *Any positive semidefinite Toeplitz matrix P can be represented as follows*

$$P = VDV^*,$$

where

$$V = \left[a\left(f_1, 0\right) \cdots a\left(f_r, 0\right)\right],$$
$$D = \mathrm{diag}\left(\left[d_1 \cdots d_r\right]\right),$$

d_k are real positive numbers, and $r = \mathrm{rank}(P)$.

The Vandermonde decomposition can be computed efficiently via root finding or by solving a generalized eigenvalue problem [31]. Indeed, in the experiments, we compute the Vandermonde decomposition of the solution of our semidefinite program to estimate the frequencies

Proof of Proposition 2.1. Denote the value of the right hand side by $\mathrm{SDP}(x)$. Suppose $x = \sum_k c_k a(f_k, \phi_k)$ with $c_k > 0$. Then observe that

$$\sum_k c_k \begin{bmatrix} a(f_k,\phi_k) \\ 1 \end{bmatrix} \begin{bmatrix} a(f_k,\phi_k) \\ 1 \end{bmatrix}^* = \sum_k c_k \begin{bmatrix} a(f_k,\phi_k) \\ 1 \end{bmatrix} \begin{bmatrix} a(f_k,\phi_k) \\ 1 \end{bmatrix}^* = \begin{bmatrix} \mathrm{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0 \tag{2.2}$$

with $\mathrm{trace}(\mathrm{Toep}(u)) = t = \sum_k c_k$. Since this holds for any decomposition of x, we conclude that $\|x\|_{\mathcal{A}} \geq \mathrm{SDP}(x)$.

Conversely, suppose for some u and x,

$$\begin{bmatrix} \mathrm{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0\,. \tag{2.3}$$

In particular, $\mathrm{Toep}(u) \succeq 0$. Form a Vandermonde decomposition

$$\mathrm{Toep}(u) = VDV^*$$

as promised by Lemma 2.2. Since $VDV^* = \sum_k d_k a(f_k, 0) a(f_k, 0)^*$ and $\|a(f_k, 0)\|_2 = 1$, $\mathrm{trace}(\mathrm{Toep}(u)) = \mathrm{trace}(D)$.

Using this Vandermonde decomposition and the matrix inequality (2.3), it follows that x is in the range of V, and hence

$$x = \sum_k w_k a(f_k, 0) = Vw$$

for some complex coefficient vector $w = [\cdots, w_k, \cdots]^T$. Finally, by the Schur Complement Lemma, we have

$$VDV^* \succeq t^{-1} V w w^* V^*$$

Let q be any vector such that $V^* q = \mathrm{sign}(w)$. Such a vector exists because V is full rank. Then

$$\mathrm{trace}(D) = q^* V D V^* q \succeq t^{-1} q^* V w w^* V^* q = t^{-1} \left(\sum_k |w_k| \right)^2 .$$

implying that $\operatorname{trace}(D)t \geq (\sum_k |w_k|)^2$. By the arithmetic geometric mean inequality,

$$\tfrac{1}{2}\operatorname{trace}(\operatorname{Toep}(u)) + \tfrac{1}{2}t = \tfrac{1}{2}\operatorname{trace}(D) + \tfrac{1}{2}t \geq \sqrt{\operatorname{trace}(D)t} \geq \sum_k |w_k|$$

implying that $\mathrm{SDP}(x) \geq \|x\|_{\mathcal{A}}$. □

There are several other approaches to proving the semidefinite programming characterization of the atomic norm. As we will see below, the dual norm of the atomic norm is related to the maximum modulus of trigonometric polynomials (see equation (4.1)). Thus, proofs based on Bochner's Theorem [35], the bounded real lemma [4, 22], or spectral factorization [41] would also provide a tight characterization.

The semidefinite programming characterization of the atomic norm also allows us to draw connections to the study of rank minimization [8, 26, 39, 40]. A direct way to exploit sparsity in frequency domain is via minimization of the following "ℓ_0-norm" type quantity

$$\|x\|_{\mathcal{A},0} = \min_{\substack{c_k \geq 0,\ \phi_k \in [0,2\pi) \\ f_k \in [0,1]}} \left\{ s : x = \sum_{k=1}^{s} c_k a(f_k, \phi_k) \right\}$$

This penalty function chooses the *sparsest* representation of a vector in terms of complex exponentials. This combinatorial quantity is closely related to the rank of positive definite Toeplitz matrices as delineated by the following Proposition:

Proposition 2.3. *The quantity $\|x\|_{\mathcal{A},0}$ is equal to the optimal value of the following rank minimization problem:*

$$\begin{array}{ll} \text{minimize}_{u,t} & \operatorname{rank}(\operatorname{Toep}(u)) \\ \text{subject to} & \begin{bmatrix} \operatorname{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0. \end{array} \tag{2.4}$$

Proof. The case for $x = 0$ is trivial. For $x \neq 0$, denote by $r^\star$ the optimal value of (2.4). We first show $r^\star \leq \|x\|_{\mathcal{A},0}$. Suppose $\|x\|_{\mathcal{A},0} = s < n$. Assume the decomposition $x = \sum_{k=1}^{s} c_k a(f_k, \phi_k)$ with $c_k > 0$ achieves $\|x\|_{\mathcal{A},0}$, and set $\operatorname{Toep}(u) = \sum_k c_k a(f_k, \phi_k) a(f_k, \phi_k)^* \succeq 0, t = \sum_k c_k > 0$. Then, as we saw in (2.2),

$$\begin{bmatrix} \operatorname{Toep}(u) & x \\ x^* & t \end{bmatrix} = \sum_{k=1}^{s} c_k \begin{bmatrix} a(f_k, \phi_k) \\ 1 \end{bmatrix} \begin{bmatrix} a(f_k, \phi_k) \\ 1 \end{bmatrix}^* \succeq 0\,.$$

This implies that $r^\star = \operatorname{rank}(\operatorname{Toep}(u)) \leq s$.

We next show $\|x\|_{\mathcal{A},0} \leq r^\star$. The $r^\star = n$ case is trivial as we could always expand x on a Fourier basis, implying $\|x\|_{\mathcal{A},0} \leq n$. We focus on $r^\star < n$. Suppose u is an optimal solution of (2.4). Then if

$$\operatorname{Toep}(u) = VDV^*$$

is a Vandermonde decomposition, postive semidefiniteness implies that x is in the range of V which means that x can be expressed as a combination of at most $r^\star$ atoms, completing the proof. □

Hence, for this particular set of atoms, atomic norm minimization is a trace relaxation of a rank minimization problem. The trace relaxation has been proven to be a powerful relaxation for recovering low rank matrices subject to random linear equations [40], values at a specified set

of entries [8], Euclidean distance constraints [32], and partial quantum expectation values [27]. However, our sampling model is far more constrained and none of the existing theory applies to our problem. Indeed, typical results on trace-norm minimization demand that the number of measurements exceed the rank of the matrix times the number of rows in the matrix. In our case, this would amount to $O(sn)$ measurements for an s sparse signal. We prove in the sequel that only $O(s \operatorname{polylog}(n))$ samples are required, dramatically reducing the dependence on n.

2.1 Atomic Norm Minimization for Continuous Compressed Sensing

Recall that we observe only a subset of entries $T \subset J$. As prescribed in [15], a natural algorithm for estimating the missing samples of a sparse sum of complex exponentials is the atomic norm minimization problem

$$\begin{array}{ll} \text{minimize}_x & \|x\|_{\mathcal{A}} \\ \text{subject to} & x_j = x_j^\star, j \in T \end{array} \tag{2.5}$$

or, equivalently, the semidefinite program

$$\begin{array}{ll} \text{minimize}_{u,\ x,\ t} & \frac{1}{2}\operatorname{trace}(\operatorname{Toep}(u)) + \frac{1}{2}t \\ \text{subject to} & \begin{bmatrix} \operatorname{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0 \\ & x_j = x_j^\star, j \in T. \end{array} \tag{2.6}$$

The main result of this paper is that this semidefinite program almost always recovers the missing samples provided the number of measurements is large enough and the frequencies are reasonably well-separated. We formalize this statement in the following theorem.

Theorem 2.4. *Suppose we observe the time samples of*

$$x_j = \frac{1}{\sqrt{4M+1}} \sum_{k=1}^{s} c_k e^{i2\pi f_k j}$$

on the index set $T \subset J = \{-2M, \ldots, 2M\}$ of size m selected uniformly at random. Additionally, assume $\operatorname{sign}(c_k)$ are drawn i.i.d. from a symmetric distribution on the complex unit circle. If $\Delta_f \geq \frac{1}{M}$, then there exists a numerical constant C such that

$$m \geq C \max\left\{ \log^2 \frac{M}{\delta}, s \log \frac{s}{\delta} \log \frac{M}{\delta} \right\},$$

is sufficient to guarantee that with probability at least $1-\delta$, $x^\star$ is the unique optimizer to (2.5).

We prove this theorem in Section 4. Note that Theorem 1.1 is a corollary of Theorem 2.4 via a simple reformulation. We provide a proof of the equivalence in Appendix A.

2.2 The Power of Rank Minimization

The analog of ℓ_0 minimization for the continuous compressed sensing problem is the rank minimization problem

$$\begin{array}{ll} \text{minimize}_{u,x,t} & \operatorname{rank}(\operatorname{Toep}(u)) \\ \text{subject to} & \begin{bmatrix} \operatorname{Toep}(u) & x \\ x^* & t \end{bmatrix} \succeq 0 \\ & x_j = x_j^\star, j \in T. \end{array} \tag{2.7}$$

The following proposition reveals that rank minimization is strictly stronger than trace minimization for the continuous compressed sensing problem.

Proposition 2.5. *Suppose the sampling set T is such that the set of vectors $\{a_T(f_k, 0) = [e^{i2\pi f_k j}]_{j\in T}, k = 1, \ldots, 2s\}$ is linearly independent for any $2s$ distinct frequencies $\{f_k, k = 1, \cdots, 2s\}$, and $x^\star = \sum_{k=1}^{s} c_k a(f_k, 0)$ for some distinct frequencies $\{f_k, k = 1, \ldots, s\}$ with the phases absorbed into the complex coefficients c_k. We have the following:*

1. *the rank minimization problem (2.7) recovers the original $x^\star$;*
2. *if the trace minimization (2.6) returns a solution with $\mathrm{rank}(\mathrm{Toep}(\hat{u})) \leq s$, then it also recovers $x^\star$.*

Proof. For the first part, we use the equivalence of rank minimization and $\|\cdot\|_{\mathcal{A},0}$ minimization. Suppose the $\|\cdot\|_{\mathcal{A},0}$ minimization returns a solution $\hat{x} = \sum_j \hat{c}_j a(\hat{f}_j, 0)$ with $\|\hat{x}\|_{\mathcal{A},0} = \hat{s} \leq s$. The feasibility of $\hat{x}$ and $x^\star$ implies

$$x^\star_T = \sum_{k=1}^{s} c_k a_T(f_k, 0) = \sum_{j=1}^{\hat{s}} \hat{c}_j a(\ \hat{f}_j, 0) = \hat{x}_T, \tag{2.8}$$

contradicting with the linear independence of $\{a_T(f_k, 0), a_T(\hat{f}_j, 0), k = 1, \cdots, s, j = 1, \cdots, \hat{s}\}$.

For the second part, suppose $(\hat{u}, \hat{x})$ is an optimal solution to the trace minimization problem. Positive semidefiniteness again implies that $\hat{x} \in \mathrm{Range}(\mathrm{Toep}(\hat{u}))$. This together with the Vandermonde decomposition readily give $\hat{x} = \sum_{j=1}^{r} \hat{c}_j a(\hat{f}_j, 0)$ for some frequencies $\{\hat{f}_j\}$, where $r = \mathrm{rank}(\mathrm{Toep}(\hat{u})) \leq s$. A contradiction argument based on (2.8) proves the claim. □

As a particularly important example, if T is a set of consecutive indices with size greater than $2s$, then $\{a_T(f_k, 0)\}$ are columns of a Vandermonde matrix and hence are linearly independent as long as the frequencies, $\{f_k\}$, are distinct. Claim 1 of Proposition 2.5 then states that we could recover $x^\star$, no matter the dimension of the signal and the separation of the frequencies. In this sense, we get a separation condition because of the trace relaxation. The connection with rank minimization also suggests using surrogate functions of rank other than the trace function, e.g., the logdet$(\cdot)$ function, which might yield better model order selection [36].

We close this section by noting that a *positive* combination of complex sinusoids with zero phases observed at the first $2s$ samples can be recovered via the trace relaxation with no limitation on the resolution. Why is there change when we add phase to the picture? A partial answer is provided by Figure 1. Figure 1 (a) and (b) display the set of atoms with no phase (i.e., $\{a(f, 0)\}$) and phase either 0 or π respectively. That is, Figure 1 (a) plots the set

$$\mathcal{A}_1 = \left\{ \begin{bmatrix} \cos(2\pi f) & \cos(4\pi f) & \cos(6\pi f) \end{bmatrix} \ : \ f \in [0, 1] \right\},$$

while (b) displays the set

$$\mathcal{A}_2 = \left\{ \begin{bmatrix} \cos(2\pi f + \phi) & \cos(4\pi f + \phi) & \cos(6\pi f + \phi) \end{bmatrix} \ : \ f \in [0, 1],\, \phi \in \{0, \pi\} \right\}.$$

Note that $\mathcal{A}_2$ is simply $\mathcal{A}_1 \cup -\mathcal{A}_1$. Their convex hulls are displayed in Figure 1 (c) and (d) respectively. The convex hull of $\mathcal{A}_1$ is *neighborly* in the sense that every edge between every pair of atoms is an exposed face and every atom is an extreme point. On the other hand, the only secants

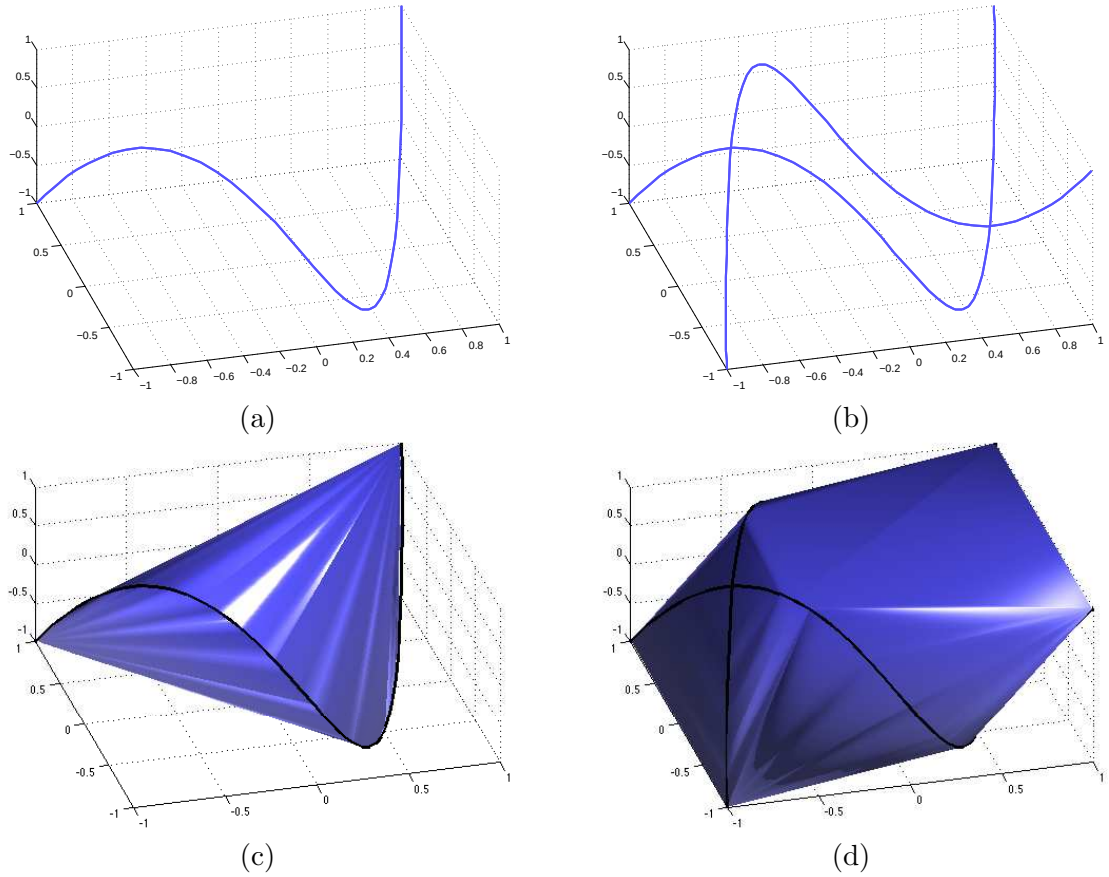


Figure 1: **Moments and their convex hulls.** (a) The real moment curve for the first three moments. (b) The moment curve for the same frequencies, but adding in phase. (c) The convex hull of (a). (d) The convex hull of (b). Whereas all of the secants of (a) are extreme in their convex hull (c), many segments between atoms of (b) lie inside the convex hull (d).

between atoms in $\mathcal{A}_2$ that are faces of the convex hull of $\mathcal{A}_2$ are those between atoms with far apart phase angles and frequencies. Problems only worsen if we let the phase range in $[0, 2\pi)$. Thus, our intuition from positive moment curves does not extend to the compressed sensing problem of sinusoids with complex phase. Nonetheless, we are able to demonstrate that under mild resolution assumptions, we can still recover sparse superpositions from very small sampling sets.

3 Prior Art and Inspirations

Frequency estimation is extensively studied and techniques for estimating sinusoidal frequencies from time samples dates back to the work of Prony [17]. Many linear prediction algorithms based on Prony's method were proposed to estimate the frequencies from *regularly spaced* time samples. A survey of these methods can be found in [5] and an extensive list of references is given in [45]. With equispaced samples, these root-finding based procedures deal with the problem directly on the continuous frequency domain, and can recover frequencies provided the number of samples is at least twice of the number of frequencies, regardless of how closely these frequencies are located

[5, 17, 45, 47].

In recent work [7], Candès and Fernandez-Granda studied this problem from the point of view of convex relaxations and proposed a total-variation norm minimization formulation that provably recovers the spectrum exactly. However, the convex relaxation requires the frequencies to be well separated by a the inverse of the number of samples. The proof techniques of this prior work forms the foundation of analysis in the sequel, but many major modifications are required to extend their results to the compressed sensing regime.

In [4], the authors proposed using atomic norm to denoise a line spectral signal corrupted with gaussian noise, and reformulated the resulting atomic norm minimization problem as a semidefinite program using the bounded real lemma [22]. Denosing is important to frequency estimation since the frequencies in a line spectral signal corrupted with moderate noise can be identified by linear prediction algorithms. Since the atomic norm framework in [4] is essentially the same as the total-variation norm framework of [7], the same semidefinite program can also be applied to total-variation norm minimization.

What is common to all aforementioned approaches, including linear prediction methods, is the reliance on observing uniform or equispaced time samples. In sharp contrast, we show that nonuniform sampling is not only a viable option, and that the original spectrum can be recovered exactly in the continuous domain, but in fact is a means of *compressive* or compressed sampling. Indeed non-uniform sampling allows us to effectively sample the signal at a sub-Nyquist rate. For array signal processing applications, this corresponds to a reduction in the number of sensors required for exact recovery, since each sensor obtains one spatial sample of the field. An extensive justification of the necessity of using randomly located sensor arrays can be found in [14]. To the best of our knowledge, little is known about exact line-spectrum recovery with non-uniform sampling using *parametric* methods, except sporadic work using ℓ_2-norm minimization to recover the missing samples [20], or based on nonlinear least square data fitting [46]. *Nonparametric* methods such as Periodogram and Correlogram for nonuniform sampling have gained popularity in recent years [44, 52], but their resolutions are usually low.

An interesting feature related to using convex optimization based methods for estimation such as [7] is a particular resolvability condition: the separation between frequencies is required to be greater than $\frac{4}{n}$ where n is the number of measurements. Linear prediction methods do not have a resolvability limitation, but it is known that in practice the numerical stability of root finding limits how close the frequencies can be. Theorem 2.4 can be viewed as an extension of the theory to nonuniform samples. Note that our approach gives an exact semidefinite characterization and is hence computationally tractable. We believe our results have potential impact on two related areas: extending compressed sensing to continuous dictionaries, and extending line spectral estimation to nonuniform sampling, thus providing new insight in sub-Nyquist sampling and super-resolution.

4 Proof of Theorem 2.4

The key to show that the optimization (2.5) succeeds is to construct a dual variable certifying the optimality of $x^\star$. In Section 4.1, we establish conditions the dual certificate should satisfy to guarantee unique optimality. Except for the optimality condition established in Section 4.1, which holds for both $J = \{0, \ldots, n-1\}$ and $J = \{-2M, \ldots, 2M\}$, the rest of the paper's proofs focus on the symmetric case $J = \{-2M, \ldots, 2M\}$.

We first show that the dual certificate can be interpreted as a polynomial with bounded modulus

on the unit circle. The polynomial will be constrained to have most of its coefficients equal to zero. In the case that all of the entries are observed, we show that the polynomial derived by Candès and Fernandez-Granda [7] suffices to guarantee optimality. Indeed they write the certificate polynomial via a kernel expansion and show that one can explicitly find appropriate kernel coefficients that certify optimality. We review this construction in Section 4.2. The requirements of the certificate polynomial in our case are far more stringent and require a non-trivial modification of their construction using a *random* kernel. This random kernel has nonzero coefficients only in the indices corresponding to observed locations (the randomness enters because the samples are observed at random). The expected value of our random kernel is a multiple of the kernel developed in [7].

Using a matrix Bernstein inequality, we show that we can find suitable coefficients to satisfy most of the optimality conditions. We then write our solution in terms of the deterministic kernel plus a random perturbation. The remainder of the proof is dedicated to showing that this random perturbation is small everywhere. First, we show that the perturbation is small on a fine gridding of the circle in Section 4.6. To do so, we emulate the proof of Candès and Romberg for reconstruction from incoherent bases [9]. Finally, in Section 4.7, we complete the proof by estimating the Lipschitz constant of the random polynomial, and, in turn, proving that the perturbations are small everywhere. Our proof is based on Bernstein's polynomial inequality which was used to estimate the noise performance of atomic norm de-noising by Bhaskar *et al* [4].

4.1 Optimality Conditions

We start with examining the optimality conditions for (2.5). Define the inner product as $\langle q, x\rangle = x^* q$, and the real inner product as $\langle q, x\rangle_{\mathbb{R}} = \mathrm{Re}(\langle q, x\rangle)$. Then the dual norm of $\|\cdot\|_{\mathcal{A}}$ is

$$\|q\|_{\mathcal{A}}^* = \sup_{\|x\|_{\mathcal{A}}\leq 1} \langle q, x\rangle = \sup_{\phi\in[0,2\pi), f\in[0,1]} \langle q, e^{i\phi} a(f,0)\rangle = \sup_{f\in[0,1]} |\langle q, a(f,0)\rangle| \tag{4.1}$$

that is, it is the maximum modulus of the polynomial

$$q(z) = \sum_{j\in J} q_j z^{-j}$$

on the unit circle. The dual problem of (2.5) is thus

$$\begin{array}{ll} \text{maximize}_q & \langle q, x^\star\rangle_{\mathbb{R}} \\ \text{subject to} & \|q\|_{\mathcal{A}}^* \leq 1 \\ & q_{T^c} = 0 \end{array} \tag{4.2}$$

which follows from a standard Lagrangian analysis [15].

The following proposition provides a sufficient condition for exact completion using dual certificate, whose proof is given in Appendix B.

Proposition 4.1. *Suppose the atom is defined by* $[a(f,0)]_j = \frac{1}{\sqrt{|J|}} e^{i2\pi f j}, j \in J$ *with* J *being either* $\{-2M, \cdots, 2M\}$ *or* $\{0, \cdots, n-1\}$. *Then* $\hat{x} = x^\star$ *is the unique optimizer to (2.5) if there exists a dual polynomial*

$$Q(f) = \frac{1}{\sqrt{|J|}} \sum_{j\in J} q_j e^{-i2\pi j f} \tag{4.3}$$

satisfying

$$Q(f_k) = \operatorname{sign}(c_k), \forall f_k \in \Omega \tag{4.4}$$
$$|Q(f)| < 1, \forall f \notin \Omega \tag{4.5}$$
$$q_j = 0, \forall j \notin T. \tag{4.6}$$

The polynomial $Q(f)$ works as a dual certificate to certify that $x^\star$ is the primal optimizer. The conditions on $Q(f)$ are imposed on the values of the dual polynomial (condition (4.4) and (4.5)) and on the coefficient vector q (condition (4.6)).

4.2 A Detour: When All Entries are Observed

Before we consider the random observation model, we explain how to construct a dual polynomial when all entries in $J = \{-2M, \ldots, 2M\}$ are observed, i.e., $T = J$. The kernel-based construction method, which was first proposed in [7], inspires our random kernel based construction in Section 4.4. The results presented in this subsection are also necessary for our later proofs.

When all entries are observed, the optimization problem (2.5) has a trivial solution, but we can still apply duality to certify the optimality of a particular decomposition. Indeed, a dual polynomial satisfying the conditions given in Proposition 4.1 with $T^c = \emptyset$ means that $\|x^\star\|_{\mathcal{A}} = \sum_k |c_k|$, namely, the decomposition $x^\star = \sum_k c_k a(f_k, 0)$ achieves the atomic norm. To construct such a dual polynomial, Candés and Fernandez-Granda suggested considering a polynomial $\bar{Q}$ of the following form [7] :

$$\bar{Q}(f) = \sum_{k=1}^{s} \alpha_k \bar{K}_M(f - f_k) + \sum_{k=1}^{s} \beta_k \bar{K}'_M(f - f_k). \tag{4.7}$$

Here $\bar{K}_M(f)$ is the squared Fejer kernel

$$\bar{K}_M(f) = \left[\frac{\sin(\pi M f)}{M \sin(\pi f)}\right]^4 \tag{4.8}$$
$$= \frac{1}{M} \sum_{j=-2M}^{2M} g_M(j) e^{-i2\pi f j} \tag{4.9}$$

with $g_M(j) = \frac{1}{M} \sum_{k=\max(j-M,-M)}^{\min(j+M,M)} \left(1 - \left|\frac{k}{M}\right|\right) \left(1 - \left|\frac{j}{M} - \frac{k}{M}\right|\right)$ the discrete convolution of two triangular functions. The squared Fejer kernel is a good candidate kernel because it attains the value of 1 at its peak, and rapidly decays to zero. Provided a separation condition is satisfied by the original signal, a suitable set of coefficients α and β can always be found.

We use $\bar{K}'_M, \bar{K}''_M, \bar{K}'''_M$ to denote the first three derivatives of $\bar{K}_M$. We list some useful facts about the kernel $\bar{K}_M(f)$:

$$\bar{K}_M(0) = 1$$
$$\bar{K}'_M(0) = \bar{K}'''_M(0) = 0$$
$$\bar{K}''_M(0) = -\frac{4\pi^2(M^2 - 1)}{3}$$

For the weighting function $g_M(\cdot)$, we have

$$\|g_M\|_\infty = \sup_j |g_M(j)| \leq 1. \tag{4.10}$$

We require that the dual polynomial (4.7) satisfies

$$\bar{Q}(f_j) = \sum_{k=1}^{s} \alpha_k \bar{K}_M(f_j - f_k) + \sum_{k=1}^{s} \beta_k \bar{K}'_M(f_j - f_k) = \operatorname{sign}(c_j), \tag{4.11}$$

$$\bar{Q}'(f_j) = \sum_{k=1}^{s} \alpha_k \bar{K}'_M(f_j - f_k) + \sum_{k=1}^{s} \beta_k \bar{K}''_M(f_j - f_k) = 0, \tag{4.12}$$

for all $f_j \in \Omega$. The constraint (4.11) guarantees that $Q(f)$ satisfies the interpolation condition (4.4), and the constraint (4.12) is used to ensure that $|Q(f)|$ achieves its maximum at frequencies in Ω. Note that the condition (4.6) is absent in this section's setting since the set T^c is empty.

We rewrite these linear constraints in the matrix vector form:

$$\begin{bmatrix} \bar{D}_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1 \\ -\frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1 & -\frac{1}{|\bar{K}''(0)|}\bar{D}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \sqrt{|\bar{K}''(0)|}\beta \end{bmatrix} = \begin{bmatrix} u \\ 0 \end{bmatrix}$$

where $\left[\bar{D}_0\right]_{jk} = \bar{K}_M(f_j - f_k)$, $\left[\bar{D}_1\right]_{jk} = \bar{K}'_M(f_j - f_k)$, $\left[\bar{D}_2\right]_{jk} = \bar{K}''_M(f_j - f_k)$ and $u \in \mathbb{C}^s$ is the vector with $u_j = \operatorname{sign}(c_j)$. We have rescaled the system of linear equations such that the system matrix is symmetric, positive semidefinite, and very close to identity. Positive definiteness follows because the kernel is a positive combination of outer products. To get an idea of why the kernel is near the identity, observe that $\bar{D}_0$ is symmetric with diagonals one, $\bar{D}_1$ is antisymmetric, and $\bar{D}_2$ is symmetric with negative diagonals $K''(0)$. We define

$$\bar{D} = \begin{bmatrix} \bar{D}_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1 \\ -\frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1 & -\frac{1}{|\bar{K}''(0)|}\bar{D}_2 \end{bmatrix} = \begin{bmatrix} \bar{D}_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1 \\ \frac{1}{\sqrt{|\bar{K}''(0)|}}\bar{D}_1^* & -\frac{1}{|\bar{K}''[0)|}\bar{D}_2 \end{bmatrix} \tag{4.13}$$

and summarize properties of the system matrix $\bar{D}$ and its submatrices in the following proposition, whose proof is given in Appendix C.

Proposition 4.2. *Suppose $\Delta_f \geq \Delta_{\min} = \frac{1}{M}$. Then $\bar{D}$ is invertible and*

$$\|I - \bar{D}\| \leq 0.3623, \tag{4.14}$$

$$\|\bar{D}\| \leq 1.3623, \tag{4.15}$$

$$\|\bar{D}^{-1}\| \leq 1.568. \tag{4.16}$$

Where $\|\cdot\|$ denotes the matrix operator norm.

For notational simplicity, partition the inverse of $\bar{D}$ as

$$\bar{D}^{-1} = \begin{bmatrix} \bar{L} & \bar{R} \end{bmatrix}$$

where $\bar{L}$ and $\bar{R}$ are both $2s \times s$. Then, solving for α and $\sqrt{\left|\bar{K}''(0)\right|}\beta$ yields

$$\begin{bmatrix} \alpha \\ \sqrt{\left|\bar{K}''(0)\right|}\beta \end{bmatrix} = \bar{D}^{-1} \begin{bmatrix} u \\ 0 \end{bmatrix} = \bar{L}u. \tag{4.17}$$

Then the ℓth derivative of the dual polynomial (after normalization) is

$$\begin{aligned} \frac{1}{\sqrt{\left|\bar{K}_M''(0)\right|}^{\ell}} \bar{Q}^{(\ell)}(f) &= \sum_{k=1}^{s} \alpha_k \frac{1}{\sqrt{\left|\bar{K}_M''(0)\right|}^{\ell}} \bar{K}_M^{(\ell)}(f - f_k) + \sum_{k=1}^{s} \sqrt{\left|\bar{K}''(0)\right|}\beta_l \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^{\ell+1}} \bar{K}_M^{(\ell+1)}(f - f_k) \\ &= \bar{v}_\ell(f)^* \bar{L}u = \left\langle \bar{L}u, \bar{v}_\ell(f) \right\rangle. \end{aligned} \tag{4.18}$$

where we have defined

$$\bar{v}_\ell(f) = \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^{\ell}} \begin{bmatrix} \bar{K}_M^{(\ell)}(f - f_1)^* \\ \vdots \\ \bar{K}_M^{(\ell)}(f - f_s)^* \\ \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}} \bar{K}_M^{(\ell+1)}(f - f_1)^* \\ \vdots \\ \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}} \bar{K}_M^{(\ell+1)}(f - f_s)^* \end{bmatrix} \tag{4.19}$$

with $\bar{K}_M^{(\ell)}$ the ℓth derivative of $\bar{K}_M$.

To certify that the polynomial with the coefficients (4.17) are bounded uniformly on the unit circle, Candès and Fernandez-Granda divide the domain $[0, 1]$ into regions near to and far from the frequencies of $x^\star$. Define

$$\begin{aligned} \Omega_{\text{near}} &= \bigcup_{k=1}^{s} [f_k - f_{b,1}, f_k + f_{b,1}] \\ \Omega_{\text{far}} &= [0, 1] / \Omega_{\text{near}} \end{aligned}$$

with $f_{b,1} = 8.245 \times 10^{-2} \frac{1}{M}$. On Ω_{far}, $|Q(f)|$ was analyzed directly, while on Ω_{near} $|Q(f)|$ is bounded by showing its second order derivative is negative. The following results are derived in the proofs of Lemmas 2.3 and 2.4 in [7]:

Proposition 4.3. *Assume $\Delta_f \geq \Delta_{\min} = \frac{1}{M}$. Then we have*

$$\left|\bar{Q}(f)\right| < 0.99992, \text{for } f \in \Omega_{\text{far}} \tag{4.20}$$

and for $f \in \Omega_{\text{near}}$

$$\bar{Q}_R(f) \geq 0.9182 \tag{4.21}$$

$$\left|\bar{Q}_I(f)\right| \leq 3.61110^{-2} \tag{4.22}$$

$$\frac{1}{\left|\bar{K}''(0)\right|}\bar{Q}_R''(f) \leq -0.314 \tag{4.23}$$

$$\left|\frac{1}{\left|\bar{K}''(0)\right|}\bar{Q}_I''(f)\right| \leq 0.5755 \tag{4.24}$$

$$\left|\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}}\bar{Q}'(f)\right| \leq 0.4346. \tag{4.25}$$

and as a consequence,

$$\frac{1}{\left|\bar{K}''(0)\right|}\left(\bar{Q}_R(f)\,\bar{Q}_R''(f) + \left|\bar{Q}'(f)\right|^2 + \left|\bar{Q}_I(f)\right|\left|\bar{Q}_I''(f)\right|\right) \leq -7.86510^{-2}.$$

4.3 Bernoulli Observation Model

The uniform sampling model is difficult to analyze directly. However, the same argument used in [10] shows that the probability of recovery failure under the uniform model is at most twice of that under a Bernoulli model. Here by "recovery failure", we refer to that (2.5) would not recover the original signal $x^\star$. Therefore, without loss of generality, we focus on the following Bernoulli observation model in our proof.

We observe entries in J independently with probability p. Let $\delta_j = 1$ or 0 indicate whether we observe the jth entry. Then $\{\delta_j\}_{j\in J}$ are i.i.d. Bernoulli random variables such that

$$\mathbb{P}(\delta_j = 1) = p.$$

On average in this model, we will observe $p|J|$ entries. For $J = \{-2M, \ldots, 2M\}$, we use

$$p = \frac{m}{M} < 1.$$

4.4 Random Polynomial Kernels

We now turn to designing a dual certificate for the Bernoulli observation model. As for the case that all entires are observed, the challenge is to construct a dual polynomial satisfying

$$Q(f_k) = \operatorname{sign}(c_k), \forall f_k \in \Omega$$
$$|Q(f)| < 1, \forall f \notin \Omega,$$

as well as an additional constraint

$$q_j = 0, \forall j \in T^c. \tag{4.26}$$

The main difference in our random setting is that the demands of our polynomial $Q(f)$ are much stricter as manifested by (4.26). Our polynomial is required to have *most* of its coefficients be zero. Our approach will be to mimic the construction in the deterministic case, but using a *random kernel*

$K_M(\cdot)$, which has nonzero coefficients only on the random subset T and satisfies $\mathbb{E}K_M = p\bar{K}_M$. We will then prove that K_M concentrates tightly around $p\bar{K}_M$.

Our random kernel is simply the expansion (4.9), but with each term multiplied by a Bernoulli random variable corresponding to the observation of a component:

$$\begin{aligned} K_M(f) &= \frac{1}{M}\sum_{j\in T} g_M(j)\, e^{-i2\pi fj} \\ &= \frac{1}{M}\sum_{j=-2M}^{2M} \delta_j g_M(j)\, e^{-i2\pi fj}. \end{aligned}$$

As before

$$g_M(j) = \frac{1}{M}\sum_{k=\max(j-M,-M)}^{\min(j+M,M)} \left(1-\left|\frac{k}{M}\right|\right)\left(1-\left|\frac{j}{M}-\frac{k}{M}\right|\right)$$

is the convolution of two discrete triangular functions. The ℓth derivative of $K_M(f)$ is

$$K_M^{(\ell)}(f) = \frac{1}{M}\sum_{j=-2M}^{2M} (-i2\pi j)^\ell\, g_M(j)\, \delta_j e^{-i2\pi fj}.$$

Both $K_M(f-f_k)$ and $K'_M(f-f_k)$ are random trigonometric polynomials of degree at most $2M$. More importantly, they contain monomial $e^{-i2\pi fj}$ only if $\delta_j = 1$, or equivalently, $j \in T$. Hence $Q(f)$ is of the form (4.3) and satisfies $q_j = 0, j \in T^c$. It is easy to calculate the expected values of $K_M(f)$ and its ℓth derivatives:

$$\begin{aligned} \mathbb{E}K_M^{(\ell)}(f) &= \frac{1}{M}\sum_{j=-2M}^{2M} (-i2\pi j)^\ell\, g_M(j)\, \mathbb{E}\{\delta_j\} e^{-i2\pi fj} \\ &= p\frac{1}{M}\sum_{j=-2M}^{2M} (-i2\pi j)^\ell\, g_M(j)\, e^{-i2\pi fj} \\ &= p\bar{K}_M^{(\ell)}(f). \end{aligned} \tag{4.27}$$

In Figure 2, we plot $p^{-1}K_M(f)$ and $p^{-1}K'_M(f)$ laid over $\bar{K}_M(f)$ and $\bar{K}'_M(f)$, respectively. We see that far away from the peak, the random coefficients induce bounded oscillations to the kernel. Near 0, however, the random kernel remains sharply peaked.

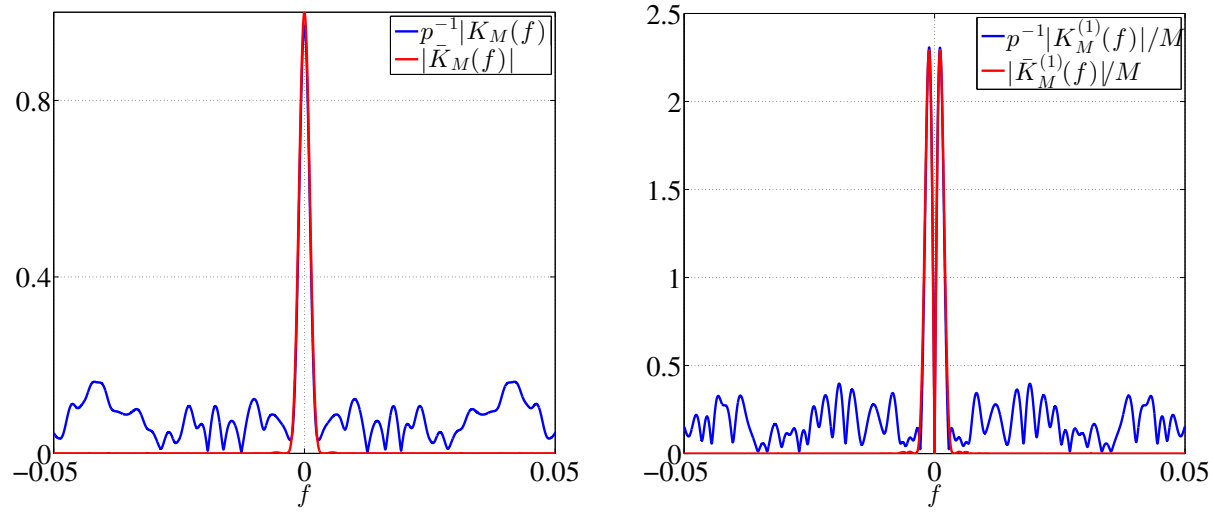
In order to satisfy the conditions (4.4) and (4.5), we require that the polynomial $Q(f)$ has the form

$$Q(f) = \sum_{k=1}^{s} \alpha_k K_M(f-f_k) + \sum_{k=1}^{s} \beta_k K'_M(f-f_k). \tag{4.28}$$

and satisfies

$$Q(f_j) = \sum_{k=1}^{s} \alpha_k K_M(f_j-f_k) + \sum_{k=1}^{s} \beta_k K'_M(f_j-f_k) = \operatorname{sign}(c_j), \tag{4.29}$$

$$Q'(f_j) = \sum_{k=1}^{s} \alpha_k K'_M(f_j-f_k) + \sum_{k=1}^{s} \beta_k K''_M(f_j-f_k) = 0 \tag{4.30}$$



(a) $p^{-1}|K_M(f)|$ and $|\bar{K}_M(f)|$

(b) $p^{-1}|K_M'(f)|/M$ and $|\bar{K}_M'(f)|/M$

Figure 2: Plots of the random kernel

for all $f_j \in \Omega$. As for $\bar{Q}(f)$, the constraint (4.29) guarantees that $Q(f)$ satisfies the interpolation condition (4.4), and the constraint (4.30) helps ensure that $|Q(f)|$ achieves its maximum at frequencies in Ω.

We now have $2s$ linear constraints (4.29), (4.30) on $2s$ unknown variables α, β. The remainder of the proof consists of three steps:

1. Show that the linear system (4.29), (4.30) is invertible with high probability using matrix Bernstein inequality [50];
2. Show $\left|Q^{(\ell)}(f) - \bar{Q}^{(\ell)}(f)\right|$, the random perturbations introduced by the random observation process, are small on a set of discrete points with high probability, implying the random dual polynomial satisfies the constraints in Proposition 4.1 on the grid; This step is proved using a modification of the idea in [9].
3. Extend the result to $[0,1]$ using Bernstein's polynomial inequality [43] and eventually show $|Q(f)| < 1$ for $f \notin \Omega$.

4.5 Invertibility

In this section we show the linear system (4.29) and (4.30) is invertible. Rewrite the linear system of equations (4.29) and (4.30) into the following matrix-vector form:

$$\begin{bmatrix} D_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}} D_1 \\ -\frac{1}{\sqrt{|\bar{K}''(0)|}} D_1 & -\frac{1}{|\bar{K}''(0)|} D_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \sqrt{|\bar{K}''(0)|}\beta \end{bmatrix} = \begin{bmatrix} u \\ 0 \end{bmatrix}, \tag{4.31}$$

where $[D_\ell]_{jk} = K_M^{(\ell)}(f_j - f_k)$, and $u = \mathrm{sign}(c)$. Note that we still rescale the derivatives using the deterministic quantity $\bar{K}''(0)$ rather than the random variable $K''(0)$.

The expectation computation (4.27) implies that $\mathbb{E}\left[D_\ell\right]_{jk} = \mathbb{E}K_M^{(\ell)}\left(f_j - f_k\right) = p\left[\bar{D}_\ell\right]_{jk}$, where $\left[\bar{D}_\ell\right]_{jk} = \bar{K}_M^{(\ell)}\left(f_j - f_k\right)$. Define

$$\begin{aligned} D &= \begin{bmatrix} D_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}} D_1 \\ -\frac{1}{\sqrt{|\bar{K}''(0)|}} D_1 & -\frac{1}{|\bar{K}''(0)|} D_2 \end{bmatrix} \\ &= \begin{bmatrix} D_0 & \frac{1}{\sqrt{|\bar{K}''(0)|}} D_1 \\ \frac{1}{\sqrt{|\bar{K}''(0)|}} D_1^* & -\frac{1}{|\bar{K}''(0)|} D_2 \end{bmatrix} \\ &= \frac{1}{M} \sum_{j=-2M}^{2M} g_M\left(j\right) \delta_j e\left(j\right) e\left(j\right)^* \end{aligned}$$

where

$$e\left(j\right) = \begin{bmatrix} e^{-i2\pi f_1 j} \\ \vdots \\ e^{-i2\pi f_s j} \\ \frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}} e^{-i2\pi f_1 j} \\ \vdots \\ \frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}} e^{-i2\pi f_s j} \end{bmatrix}. \tag{4.32}$$

Then we have

$$\begin{aligned} \mathbb{E}D &= \frac{1}{M} \sum_{j=-2M}^{2M} g_M\left(j\right) \mathbb{E}\{\delta_j\} e\left(j\right) e\left(j\right)^* \\ &= p\frac{1}{M} \sum_{j=-2M}^{2M} g_M\left(j\right) e\left(j\right) e\left(j\right)^* \\ &= p\bar{D}, \end{aligned}$$

with $\bar{D}$ defined in (4.13). As a consequence, we have

$$\begin{aligned} D - \mathbb{E}D &= D - p\bar{D} \\ &= \sum_{j=-2M}^{2M} \frac{1}{M} g_M\left(j\right)\left(\delta_j - p\right) e\left(j\right) e\left(j\right)^* \\ &= \sum_{j=-2M}^{2M} X_j. \end{aligned}$$

with $X_j = \frac{1}{M} g_M\left(j\right)\left(\delta_j - p\right) e\left(j\right) e\left(j\right)^*$ a zero mean random self-adjoint matrix. We will apply the noncommutative Bernstein inequality to show that D concentrates about its mean $p\bar{D}$ with high probability.

Lemma 4.4 (Noncommutative Bernstein Inequality, [50, Theorem 1.4]). *Let $\{X_j\}$ be a finite sequence of independent, random self-adjoint matrices of dimension d. Suppose that*

$$\begin{aligned}\mathbb{E}X_j &= 0\\ \|X_j\| &\le R, \text{almost surely}\\ \sigma^2 &= \left\| \sum_j \mathbb{E}\left(X_j^2\right) \right\|.\end{aligned}$$

Then for all $t \ge 0$,

$$\mathbb{P}\left\{ \left\| \sum_j X_j \right\| \ge t \right\} \le d \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right).$$

For $\tau > 0$, define the event

$$\mathcal{E}_{1,\tau} = \left\{ \left\| p^{-1}D - \bar{D} \right\| \le \tau \right\}. \tag{4.33}$$

The following lemma, proved in Appendix D, shows that $\mathcal{E}_{1,\tau}$ has a high probability if m is large enough.

Lemma 4.5. *If $\tau \in (0, 0.6377)$, then we have $\mathbb{P}\left(\mathcal{E}_{1,\tau}\right) \ge 1 - \delta$ provided*

$$m \ge \frac{50}{\tau^2} s \log \frac{2s}{\delta}.$$

Note that an immediate consequence of Lemma 4.5 is that D is invertible on $\mathcal{E}_{1,\tau}$. Additionally, Lemma 4.5 allows us to control the norms of the submatices of D^{-1}. For that purpose, we partition D^{-1} as

$$D^{-1} = \begin{bmatrix} L & R \end{bmatrix}$$

with L and R both $2s \times s$ and obtain:

Corollary 4.6. *On the event $\mathcal{E}_{1,\tau}$ with $\tau \in \left(0, \frac{1}{4}\right]$, we have*

$$\begin{aligned}\left\| L - p^{-1}\bar{L} \right\| &\le 2 \left\| \bar{D}^{-1} \right\|^2 p^{-1}\tau\\ \|L\| &\le 2 \left\| \bar{D}^{-1} \right\| p^{-1}.\end{aligned}$$

The proof of this corollary is elementary matrix analysis and can be found in Appendix E. Since on the event $\mathcal{E}_{1,\tau}$ with $\tau < 1/4$ the matrix $D = \begin{bmatrix} D_0 & D_1 \\ D_1 & D_2 \end{bmatrix}$ is invertible, we solve for α and $\sqrt{\left|\bar{K}''(0)\right|}\beta$ from (4.31):

$$\begin{aligned}\begin{bmatrix} \alpha \\ \sqrt{\left|\bar{K}''(0)\right|}\beta \end{bmatrix} &= D^{-1} \begin{bmatrix} u \\ 0 \end{bmatrix}\\ &= Lu.\end{aligned} \tag{4.34}$$

In the next section, we will plug (4.34) back into (4.28), and analyze the effect of random perturbations on the polynomial $Q(f)$.

4.6 Random Perturbations

In this section, we show that the dual polynomial $Q(f)$ concentrates around $\bar{Q}(f)$ on a discrete set Ω_{grid}.

We introduce a random analog of $\bar{v}_\ell$, defined by (4.19), as

$$\begin{aligned} v_\ell(f) &= \frac{1}{\sqrt{|\bar{K}''(0)|}^\ell} \begin{bmatrix} K_M^{(\ell)}(f-f_1)^* \\ \vdots \\ K_M^{(\ell)}(f-f_s)^* \\ \frac{1}{\sqrt{|\bar{K}''(0)|}} K_M^{(\ell+1)}(f-f_1)^* \\ \vdots \\ \frac{1}{\sqrt{|\bar{K}''(0)|}} K_M^{(\ell+1)}(f-f_s)^* \end{bmatrix} \\ &= \frac{1}{M} \sum_{j=-2M}^{2M} \left(\frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}} \right)^\ell g_M(j) \delta_j e^{i2\pi f j} e(j). \end{aligned} \tag{4.35}$$

with $K_M^{(\ell)}$ the ℓth derivative of K_M, and $e(j)$ defined in (4.32). Clearly we have the expectation of v_ℓ is equal to p times its deterministic counterpart defined by (4.19):

$$\mathbb{E} v_\ell(f) = p\bar{v}_\ell(f), \forall f \in [0,1]$$

Then, in a similar fashion to (4.18), we rewrite

$$\begin{aligned} \frac{1}{\sqrt{|\bar{K}_M''(0)|}^\ell} Q^{(\ell)}(f) &= \sum_{k=1}^s \alpha_k \frac{1}{\sqrt{|\bar{K}_M''(0)|}^\ell} K_M^{(\ell)}(f-f_k) \\ &\quad + \sum_{k=1}^s \sqrt{|\bar{K}''(0)|}\beta_l \frac{1}{\sqrt{|\bar{K}_M''(0)|}^{\ell+1}} K_M^{(\ell+1)}(f-f_k) \\ &= v_\ell(f)^* Lu = \langle Lu, v_\ell(f) \rangle = \langle u, L^* v_\ell(f) \rangle. \end{aligned}$$

We decompose $L^* v_\ell(f)$ into three parts:

$$\begin{aligned} L^* v_\ell(f) &= [(L - p^{-1}\bar{L}) + p^{-1}\bar{L}]^* [(v_\ell(f) - p\bar{v}_\ell(f)) + p\bar{v}_\ell(f)] \\ &= \bar{L}^* \bar{v}_\ell(f) + L^*(v_\ell(f) - p\bar{v}_\ell(f)) + (L - p^{-1}\bar{L})^* p\bar{v}_\ell(f), \end{aligned}$$

which induces a decomposition on $\frac{1}{\sqrt{|\bar{K}_M''(0)|}^\ell} Q^{(\ell)}(f)$

$$\begin{aligned} \frac{1}{\sqrt{|\bar{K}_M''(0)|}^\ell} Q^{(\ell)}(f) &= \langle u, L^* v_\ell(f) \rangle \\ &= \left\langle u, \bar{L}^* \bar{v}_\ell(f) \right\rangle + \left\langle u, L^*(v_\ell(f) - p\bar{v}_\ell(f)) \right\rangle + \left\langle u, (L - p^{-1}\bar{L})^* p\bar{v}_\ell(f) \right\rangle \\ &= \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \bar{Q}^{(\ell)}(f) + I_1^\ell(f) + I_2^\ell(f). \end{aligned} \tag{4.36}$$

Here $\frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}}\bar{Q}^{(\ell)}(f) = \langle u, \bar{L}^*\bar{v}_\ell(f)\rangle = \langle u\bar{L}, \bar{v}_\ell(f)\rangle$ as in (4.18) and we have defined

$$I_1^\ell(f) = \langle u, L^*(v_\ell(f) - p\bar{v}_\ell(f))\rangle$$

and

$$I_2^\ell(f) = \langle u, (L - p^{-1}\bar{L})^* p\bar{v}_\ell(f)\rangle .$$

The goal of the remainder of this section is to show, in Lemma 4.9 and 4.10, that $I_1^\ell(f)$ and $I_2^\ell(f)$ are small on a set of grid points $\Omega_{\text{grid}}^\ell$ with high probability. We use superscript ℓ on $\Omega_{\text{grid}}^\ell$ to emphasize that the set of grid points could change with ℓ.

The proof of Lemma 4.9, which shows $I_1^\ell(f)$ is small on $\Omega_{\text{grid}}^\ell$, essentially follows that of Candès and Romberg [9]. We include the proof details here for completeness, but very little changes in the argument. Since $I_1^\ell(f) = \langle u, L^*(v_\ell(f) - p\bar{v}_\ell(f))\rangle$ is a weighted sum of independent random variables following a symmetric distribution on the complex unit circle, for fixed $f \in [0,1]$, we apply Hoeffding's inequality to control its value. This in turn requires an estimate of $\|L^*(v_\ell(f) - p\bar{v}_\ell(f))\|_2$. In Lemma 4.7, we first use concentration of measure (Lemma F.1) to establish that $\|v_\ell(f) - p\bar{v}_\ell(f)\|_2$ is small with high probability. In Lemma 4.8, we then combine Lemma 4.7 and Lemma 4.5 to show $\|L^*(v_\ell(f) - p\bar{v}_\ell(f))\|_2$ is small. The extension from a fixed f to a finite set Ω_{grid} relies on union bound.

We start with bounding $\|v_\ell(f) - p\bar{v}_\ell(f)\|_2$ in the following lemma. The proof given in Appendix F is based on an inequality of Talagrand.

Lemma 4.7. *Fix $f \in [0,1]$. Let*

$$\bar{\sigma}_\ell^2 := 2^{4\ell+1}\frac{m}{M^2}\max\left\{1, 2^4\frac{s}{\sqrt{m}}\right\}$$

and fix a positive number

$$a \le \begin{cases} \sqrt{2}m^{1/4} & \text{if } 2^4\frac{s}{\sqrt{m}} \ge 1, \\ \frac{\sqrt{2}}{4}\sqrt{\frac{m}{s}} & \text{otherwise.} \end{cases}$$

Then we have

$$\mathbb{E}\,\|v_\ell(f) - p\bar{v}_\ell(f)\|_2 \le 2^{2\ell+3}\frac{\sqrt{ms}}{M}$$

$$\mathbb{P}\left(\|v_\ell(f) - p\bar{v}_\ell(f)\|_2 > 2^{2\ell+3}\frac{\sqrt{ms}}{M} + a\bar{\sigma}_\ell, \ell = 0,1,2,3\right) \le 64e^{-\gamma a^2}$$

for some $\gamma > 0$.

The following lemma combines Lemma 4.7 and Corollary 4.6 to show $\|L^*(v_\ell(f) - p\bar{v}_\ell(f)\|_2$ is small with high probability.

Lemma 4.8. *Let $\tau \in (0, 1/4]$. Consider a finite set $\Omega_{\text{grid}} = \{f_d\}$. With the same notation as last lemma, we have*

$$\mathbb{P}\left[\sup_{f_d \in \Omega_{\text{grid}}} \|L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\|_2 \geq 4\left(2^{2\ell+1}\sqrt{\frac{s}{m}} + \frac{M}{m}a\bar{\sigma}_\ell\right), \ell = 0,1,2,3\right] \leq 64\,|\Omega_{\text{grid}}|\, e^{-\gamma a^2} + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right).$$

Proof of Lemma 4.8. Conditioned on the event

$$\bigcap_{\ell, f_d \in \Omega_{\text{grid}}} \left\{\|v_\ell(f_d) - p\bar{v}_\ell(f_d)\|_2 \leq 2^{2\ell+1}\frac{\sqrt{ms}}{M} + a\bar{\sigma}_\ell\right\} \bigcap \mathcal{E}_{1,\tau}$$

we have

$$\begin{aligned}\|L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\|_2 &\leq \|L\|\left(2^{2\ell+1}\frac{\sqrt{ms}}{M} + a\bar{\sigma}_\ell\right) \\ &\leq 2\left\|\bar{D}^{-1}\right\| p^{-1}\left(2^{2\ell+1}\frac{\sqrt{ms}}{M} + a\bar{\sigma}_\ell\right) \\ &\leq 4\left(2^{2\ell+1}\sqrt{\frac{s}{m}} + \frac{M}{m}a\bar{\sigma}_\ell\right),\end{aligned}$$

where we have used Proposition 4.2 and Corollary 4.6, and plugged in $p = m/M$. The claim of the lemma then follows from union bound. □

Lemma 4.8 together with Hoeffding's inequality allow us to control the size of $\sup_{f_d \in \Omega_{\text{grid}}} I_1^\ell(f_d)$:

Lemma 4.9. *There exists a numerical constant C such that if*

$$m \geq C \max\left\{\frac{1}{\varepsilon^2}\max\left(s\log\frac{|\Omega_{\text{grid}}|}{\delta}, \log^2\frac{|\Omega_{\text{grid}}|}{\delta}\right), s\log\frac{s}{\delta}\right\},$$

then we have

$$\mathbb{P}\left\{\sup_{f_d \in \Omega_{\text{grid}}} \left|I_1^\ell(f_d)\right| \leq \varepsilon, \ell = 0,1,2,3\right\} \geq 1 - 12\delta$$

Next lemma controls $I_2^\ell(f)$. It's proof is similar to the proof of Lemma 4.9.

Lemma 4.10. *There exists a numerical constant C such that if*

$$m \geq C\frac{1}{\varepsilon^2}s\log\frac{s}{\delta}\log\frac{|\Omega_{\text{grid}}|}{\delta},$$

then we have

$$\mathbb{P}\left(\sup_{f_d \in \Omega_{\text{grid}}} \left|I_2^\ell(f_d)\right| < \varepsilon, \ell = 0,1,2,3\right) \leq 1 - 8\delta$$

Both Lemmas 4.9 and 4.10 are proven in the Appendix.

Denote

$$\mathcal{E}_2 = \left\{ \sup_{f_d \in \Omega_{\text{grid}}} \left| \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} Q^{(\ell)}(f_d) - \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \bar{Q}^{(\ell)}(f_d) \right| \le \frac{\varepsilon}{3}, \ell = 0,1,2,3 \right\}.$$

Combining the decomposition (4.36), Lemma 4.9, and Lemma 4.10 with suitable redefinition of ε and δ immediately yields the following proposition

Proposition 4.11. *Suppose $\Omega_{\text{grid}} \subset [0,1]$ is a finite set of points. There exists constant C such that*

$$m \ge C \frac{1}{\varepsilon^2} \max \left\{ \log^2 \frac{|\Omega_{\text{grid}}|}{\delta}, s \log \frac{s}{\delta} \log \frac{|\Omega_{\text{grid}}|}{\delta} \right\}, \tag{4.37}$$

is sufficient to guarantee

$$\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta.$$

4.7 Extension to Continuous Domain

We have proved that $\frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} Q^{(\ell)}(f)$ and $\frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \bar{Q}^{(\ell)}(f)$ are not far on a set of grid points. This section aims extending this statement to everywhere in $[0,1]$, and show $|Q(f)| < 1$ for $f \notin \Omega$ eventually. The key is the following Bernstein's polynomial inequality:

Lemma 4.12 (Bernstein's polynomial inequality, [43])**.** *Let p_N be any polynomial of degree N with complex coefficients. Then*

$$\sup_{|z| \le 1} |p'(z)| \le N \sup_{|z| \le 1} |p(z)|.$$

Our first proposition verifies that our random dual polynomial is close to the deterministic dual polynomial on all of $[0,1]$

Proposition 4.13. *Suppose $\Delta_f \ge \Delta_{\min} = \frac{1}{M}$ and*

$$m \ge C \max \left\{ \frac{1}{\varepsilon^2} \log^2 \frac{M}{\delta\varepsilon}, \frac{1}{\varepsilon^2} s \log \frac{s}{\delta} \log \frac{M}{\delta\varepsilon} \right\}.$$

Then with probability $1-\delta$, we have

$$\left| \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} Q^{(\ell)}(f) - \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \bar{Q}^{(\ell)}(f) \right| \le \varepsilon, \forall f \in [0,1], \ell = 0,1,2,3. \tag{4.38}$$

Proof. It suffices to prove (4.38) on $\mathcal{E}_{1,1/4}$ and $\mathcal{E}_2$ and then modify the lower bound (4.37). We first give a very rough estimate of $\sup_f \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \left| Q^{(\ell)}(f) \right|$ on the set $\mathcal{E}_{1,1/4}$:

$$\begin{aligned} \frac{1}{\sqrt{|\bar{K}_M''(0)|^\ell}} \left| Q^{(\ell)}(f) \right| &= |\langle u, L^* v_\ell(f) \rangle| \\ &\le \|u\|_2 \|L\| \|v_\ell(f)\|_2 \\ &\le C p^{-1} s \\ &\le C M^2 \end{aligned}$$

where we have used $\|u\|_2 \le \sqrt{s}$ and $\|v_\ell(f)\|_2 \le C\sqrt{s}$. To see the latter, we note

$$\begin{aligned}\|v_\ell(f)\|_2 &\le \sum_{j=-2M}^{2M} \left\| \frac{1}{M} \left(\frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}} \right)^\ell g_M(j)\, e(j) \right\|_2 \\ &\le (4M+1)\frac{1}{M} 4^{\ell+1} s^{1/2},\end{aligned}$$

where we have used

$$\begin{aligned}&\|g_M\|_\infty \le 1, \\ &\left| \frac{2\pi j}{\sqrt{\bar{K}''_M(0)}} \right| \le 4 \text{ when } M \ge 2, \\ &\|e(j)\|_2^2 \le s\left(1 + \max_{|j|\le 2M} \frac{(2\pi j)^2}{|K''(0)|}\right) \le 14s \text{ when } M \ge 4.\end{aligned}$$

Viewing $\frac{1}{\sqrt{|\bar{K}''_M(0)|}} Q^{(\ell)}(\cdot)$ as a trigonometric polynomial in $z = e^{-i2\pi f}$ of degree $2M$, according to Bernstein's polynomial inequality, we get

$$\begin{aligned}\left| \frac{1}{\sqrt{|\bar{K}''_M(0)|}} Q^{(\ell)}(f_a) - \frac{1}{\sqrt{|\bar{K}''_M(0)|}} Q^{(\ell)}(f_b) \right| &\le \left| e^{-i2\pi f_a} - e^{-i2\pi f_b} \right| \sup_z \left| \frac{d \frac{1}{\sqrt{|\bar{K}''_M(0)|}} Q^{(\ell)}(z)}{dz} \right| \\ &\le 4\pi |f_a - f_b| 2M \sup_f \left| \frac{1}{\sqrt{|\bar{K}''_M(0)|}} Q^{(\ell)}(f) \right| \\ &\le CM^3 |f_a - f_b|.\end{aligned}$$

We select $\Omega_{\text{grid}} \subset [0,1]$ such that for any $f \in [0,1]$, there exists a point $f_d \in \Omega_{\text{grid}}$ satisfying $|f - f_d| \le \frac{\varepsilon}{3CM^3}$. The size of Ω_{grid} is less than $3CM^3/\varepsilon$.

With this choice of Ω_{grid}, on the set $\mathcal{E}_{1,1/4} \bigcap \mathcal{E}_2$ we have

$$\begin{aligned}&\left| \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} Q^{(\ell)}(f) - \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} \bar{Q}^{(\ell)}(f) \right| \\ &\le \left| \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} Q^{(\ell)}(f) - \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} Q^{(\ell)}(f_d) \right| + \left| \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} Q^{(\ell)}(f_d) - \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} \bar{Q}^{(\ell)}(f_d) \right| \\ &+ \left| \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} \bar{Q}^{(\ell)}(f_d) - \frac{1}{\sqrt{|\bar{K}''_M(0)|^\ell}} \bar{Q}^{(\ell)}(f) \right| \\ &\le CM^3 |f - f_d| + \frac{\varepsilon}{3} + CM^3 |f - f_d| \\ &\le \varepsilon, \forall f \in [0,1].\end{aligned}$$

Finally, we modify the condition (4.37) according to our choice of Ω_{grid}:

$$m \geq C \max\left\{\frac{1}{\varepsilon^2}\log^2\frac{M}{\delta\varepsilon}, \frac{1}{\varepsilon^2}s\log\frac{s}{\delta}\log\frac{M}{\delta\varepsilon}\right\}$$

□

An immediate consequence of Proposition 4.13 and the bound (4.20) of Proposition 4.3 is the following estimate on $Q(f)$ for $f \in \Omega_{\text{far}} = [0,1]/\bigcup_k [f_k - f_{b,1}, f_k + f_{b,1}]$:

Lemma 4.14. *Suppose* $\Delta_f \geq \Delta_{\min} = \frac{1}{M}$ *and*

$$m \geq C \max\left\{\log^2\frac{M}{\delta}, s\log\frac{s}{\delta}\log\frac{M}{\delta}\right\}.$$

Then with probability $1-\delta$*, we have*

$$|Q(f)| < 1, \forall f \in \Omega_{\text{far}}.$$

Proof. It suffices to choose $\varepsilon = 10^{-5}$. The rest follows from (4.38), triangle inequality, and modification of the constant in (4.38). □

Similar statement holds for $f \in \Omega_{\text{near}} = \bigcup_k [f_k - f_{b,1}, f_k + f_{b,1}]$.

Lemma 4.15. *Suppose* $\Delta_f \geq \Delta_{\min} = \frac{1}{M}$ *and*

$$m \geq C \max\left\{\log^2\left(\frac{M}{\delta}\right), s\log\frac{s}{\delta}\log\frac{M}{\delta}\right\}.$$

Then we have $|Q(f)| < 1$ *for all* $f \in \Omega_{\text{near}}$.

Proof. Define $Q_R(f) = \text{Re}(Q(f))$ and $Q_I(f) = \text{Im}(Q(f))$. Since $|Q(f_k)| = 1$ and $Q'(f_k) = 0$ with the latter implying

$$\frac{d|Q|}{df}(f) = \frac{Q_R'(f)Q_R(f) + Q_I'(f)Q_I(f)}{|Q(f)|} = 0$$

we only need to show $\frac{d^2|Q(f)|}{df^2} < 0$ on Ω_{near}. Take the second order derivative of $|Q(f)|$:

$$\frac{d^2|Q|}{df^2}(f) = -\frac{\left(Q_R(f)Q_R'(f) + Q_I(f)Q_I'(f)\right)^2}{|Q(f)|^3} + \frac{|Q'(f)|^2 + Q_R(f)Q_R''(f) + Q_I(f)Q_I''(f)}{|Q(f)|}.$$

So it suffices to show that for $f \in \Omega_{\text{near}}$

$$Q_R(f)Q_R''(f) + \left|Q'(f)\right|^2 + |Q_I(f)|\left|Q_I''(f)\right| < 0.$$

As a consequence of (4.38) in Proposition 4.13, triangle inequality, and (4.21)-(4.25) of Proposition 4.3, we have on the set $\mathcal{E}_2$ for any $f \in \Omega_{\text{near}}$

$$\begin{aligned}
Q_R(f) &\geq \bar{Q}_R(f) - \varepsilon \geq 0.9182 - \varepsilon \\
|Q_I(f)| &\leq \left|\bar{Q}_I(f)\right| + \varepsilon \leq 3.611 \times 10^{-}2 + \varepsilon \\
\frac{1}{\left|\bar{K}''(0)\right|} Q_R''(f) &\leq \frac{1}{\left|\bar{K}''(0)\right|} \bar{Q}_R''(f) + \varepsilon \leq -0.314 + \varepsilon \\
\left|\frac{1}{\left|\bar{K}''(0)\right|} Q_I''(f)\right| &\leq \left|\frac{1}{\left|\bar{K}''(0)\right|} \bar{Q}_I''(f)\right| + \varepsilon \leq 0.5755 + \varepsilon \\
\left|\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}} Q_I'(f)\right| &\leq \left|\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}} \bar{Q}_I'(f)\right| + \varepsilon \leq 0.4346 + \varepsilon.
\end{aligned}$$

implying

$$\frac{1}{\left|\bar{K}''(0)\right|}\left(Q_R(f) Q_R''(f) + \left|Q'(f)\right|^2 + |Q_I(f)| \left|Q_I''(f)\right|\right) \leq -7.86510^{-2} + 2.714\varepsilon + \varepsilon^2 < 0$$

when ε assumes a sufficiently small numerical value. With this choice of ε, the condition of m becomes

$$m \geq C \max\left\{\log^2 \frac{M}{\delta}, s \log \frac{s}{\delta} \log \frac{M}{\delta}\right\}.$$

Therefore, $|Q(f)| < 1$ on Ω_{near} except for $f \in \Omega$. We actually proved a stronger result that with probability at least $1 - \delta$

$$|Q(f)| \leq 1 - 0.07\left|\bar{K}''(0)\right| (f - f_k)^2 \leq 1, \forall f \in [f_k - f_{b,1}, f_k + f_{b,1}].$$

□

Proof of Theorem 2.4. Finally, if $\Delta_{\min} \geq \frac{1}{M}$ and

$$m \geq C \max\left\{\log^2 \frac{M}{\delta}, s \log \frac{s}{\delta} \log \frac{M}{\delta}\right\},$$

combining Lemma 4.14 and 4.15, we have proved the claim of Theorem 2.4. □

5 Numerical Experiments

We conducted a series of numerical experiments to test the performance of (2.5) under various parameter settings (see Table 1). We use $J = \{0, \ldots, n-1\}$ for all numerical experiments.

We compared the performance of two algorithms: the semidefinite program (2.6) and the basis pursuit obtained through discretization:

$$\underset{c}{\text{minimize}} \ \|c\|_1 \text{ subject to } x_j^\star = (Fc)_j, j \in T. \tag{5.1}$$

Here F is a DFT matrix of appropriate dimension depending on the grid size. Note that since the components of c are complex, this is a second-order cone problem. In the following, we use SDP

and BP to label the semidefinite program algorithm and the basis pursuit algorithm, respectively. We solved the SDP with the SDPT3 solver [49] and the basis pursuit (5.1) with CVX [25] coupled with SDPT3. All parameters of the SDPT3 solver were set to default values and CVX precision was set to 'high'. For the BP, we used three levels of discretization at 4, 16, and 64 times the signal dimension.

To generate our instances of form (2.1), we sampled $s = \rho_s n$ normalized frequencies from $[0, 1]$, either *randomly*, or *equispaced*. Random frequencies are sampled randomly on $[0, 1]$ with an additional constraint on the minimal separation Δ_f. Given $s = \rho_s n$, s equispaced frequencies are generated with the same separation $1/s$ with an additional random shift. This random shift will ensure that in most cases, basis mismatch occurs for discretization method. The signal coefficient magnitudes $|c_1|, \cdots, |c_s|$ are either *unit*, i.e., equal to 1, or *fading*, i.e., equal to $.5 + w^2$ with w a zero mean unit variance Gaussian random variable. The signs $\{e^{i\phi_k}, k = 1, \cdots, s\}$ follow either Bernoulli ± 1 distribution, labeled as *real*, or uniform distribution on the complex unit circle, labeled as *complex*. A length n signal was then formed according to model (2.1). As a final step, we uniformly sample $\rho_m n$ entries of the resulting signal.

We tested the algorithms on four sets of experiments. In the first experiment, by running the algorithms on a randomly generated instance with $n = 256, s = 6$ and 40 samples selected uniformly at random, we compare SDP and BP's ability of frequency estimation and visually illustrate the the effect of discretization. We see from Figure 3 that SDP recovery followed by matrix pencil approach to retrieve the frequencies gives the most accurate result. We also observe that increasing the level of discretization can increase BP's accuracy in locating the frequencies.

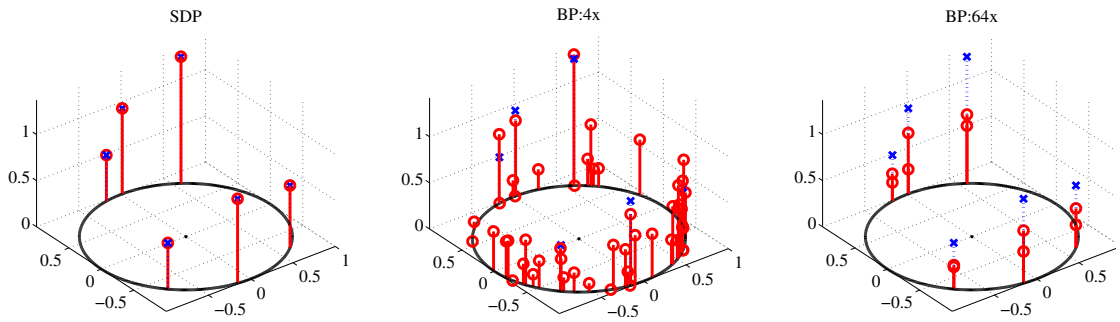


Figure 3: Frequency Estimation: Blue represents the true frequencies, while red represents the estimated ones.

In the second set of experiments, we compare the performance of SDP and BP with three levels of discretization in terms of solution accuracy and running time. The parameter configurations are summarized in Table 1. Each configuration was repeated 10 times, resulting a total of 1920 valid experiments excluding those with $\rho_m \geq 1$.

Table 1: Parameter configurations

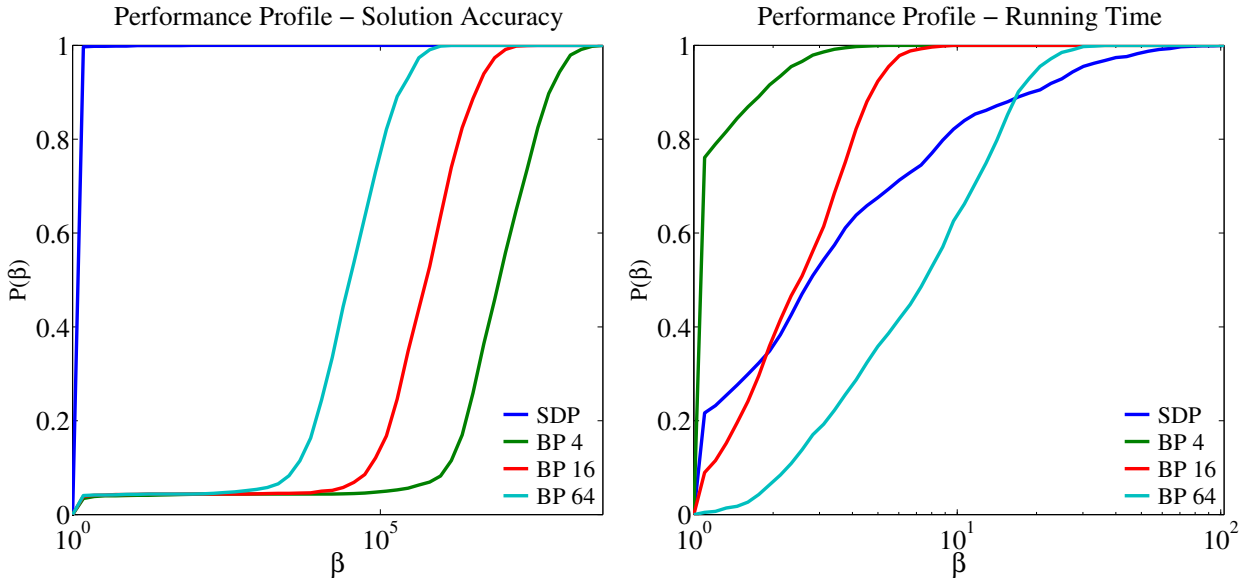
n	64, 128, 256
ρ_s	1/16, 1/32, 1/64
ρ_m/ρ_s	5, 10, 20
$\lvert c_k\rvert$	unit, fading
frequency	random, equispaced
sign	real, complex

We use the performance profile as a convenient way to compare the performance of different algorithms. The performance profile proposed in [18] visually presents the performance of a set of algorithms under a variety of experimental conditions. More specifically, let $\mathcal{P}$ be the set of experiments and $\mathcal{M}_a(p)$ specify the performance of algorithm a on experiment p for some metric $\mathcal{M}$ (the smaller the better), e.g., running time and solution accurary. Then the performance profile $\mathcal{P}_a(\beta)$ is defined as

$$\mathcal{P}_a(\beta) = \frac{\#\{p \in \mathcal{P} : \mathcal{M}_a(p) \leq \beta \min_a \mathcal{M}_a(p)}{\#(\mathcal{P})}, \beta \geq 1.$$

Roughly speaking, $\mathcal{P}_a(\beta)$ is the fraction of experiments such that the performance of algorithm a is within a factor β of that of the best performed one.

We show the performance profiles for numerical accuracy and running times in Figure 4a and 4b, respectively. We see that SDP significantly outperforms BP for all tested discretization levels in terms of numerical accuracy. When the discretization levels are higher, e.g., 64x, the running times of BP exceed that of SDP.



(a) Solution accuracy

(b) Running times

Figure 4: Performance profiles for solution accuracy and running times. Note the β-axes are in logarithm scale for both plots.

To give the reader a better idea of the numerical accuracy and the running times, in Table 2

we present their medians and median absolute deviation for the four algorithms. As one would expect, the running time increases as the discretization level increases. We also observe that SDP is very accurate, with an median error at the order of 10^{-9}. Increasing the level of discretization can increase the accuracy of BP. However, with discretization level $N = 64n$, we get a median accuracy at the order of 10^{-5}, but the median running time already exceeds that of SDP.

Table 2: Medians and median absolute deviation (MAD) for solution accuracy and running time

		SDP	BP: 4x	BP: 16x	BP: 64x
Solution Accuracy	Median	1.39e-09	1.23e-02	7.67e-04	4.65e-05
	MAD	1.26e-09	9.44e-03	6.05e-04	3.64e-05
Running Time (s)	Median	34.03	11.72	20.39	70.46
	MAD	27.32	4.83	12.19	55.37

In the third set of experiments, we compiled two phase transition plots. To prepare the Figure 5a, we pick $n = 128$ and vary $\rho_s = \frac{2}{n} : \frac{2}{n} : \frac{100}{n}$ and $\rho_m = \frac{2}{n} : \frac{2}{n} : \frac{126}{n}$. For each fixed (ρ_m, ρ_s), we randomly generate $s = n\rho_s$ frequencies while maintaining a frequency separation $\Delta_f \geq \frac{1}{n}$. The coefficients are generated with random magnitudes and random phases, and the entries are observed uniform randomly. We then run the SDPT3-SDP algorithm to recover the missing entries. The recovery is considered successful if the relative error $\|\hat{x} - x^\star\|_2 / \|x^\star\|_2 \leq 10^{-6}$. This process was repeated 10 times and the rate of success was recorded. Figure 5a shows the phase transition results. The x-axis indicates the fraction of observed entries ρ_m, while the y-axis is $\rho_s = \frac{s}{n}$. The color represents the rate of success with red corresponding to perfect recovery and blue corresponding to complete failure.

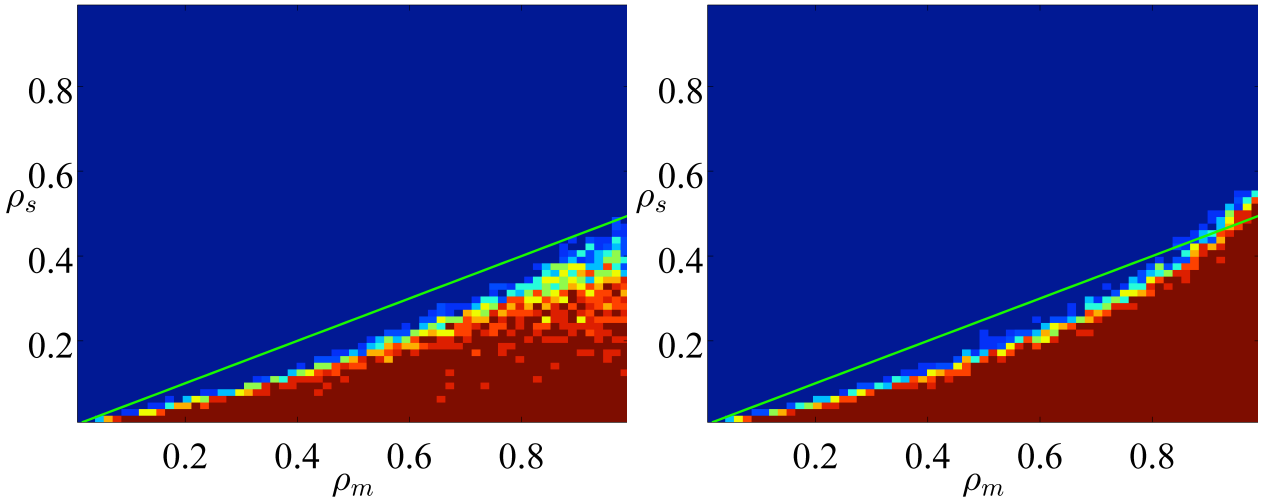
We also plot the line $\rho_s = \rho_m/2$. Since a signal of s frequencies has $2s$ degrees of freedom, including s frequency locations and s magnitudes, this line serves as the boundary above which any algorithm should have a chance to fail. In particular, Prony's method requires $2s$ consecutive samples in order to recover the frequencies and the magnitudes.

From Figure 5a, we see that there is a transition from perfect recovery to complete failure. However, the transition boundary is not very sharp. In particular, we notice failures below the boundary of the transition where complete success should happen. Examination of the failures show that they correspond to instances with minimal frequency separations marginally exceeding $\frac{1}{n}$. We expect to get cleaner phase transitions if the frequency separation is increased.

To prepare Figure 5b, we repeated the same process in preparing Figure 5a except that the frequency separation was increased from $\frac{1}{n}$ to $\frac{1.5}{n}$. In addition, to respect the minimal separation, we reduced the range of possible sparsity levels to $\{2, 4, \ldots, 70\}$. We now see a much sharper phase transition. The boundary is actually very close to the $\rho_s = \rho_m/2$ line. When ρ_m is close to 1, we even observe successful recovery above the line.

In the last set of experiments, we use a simple example to illustrate the noise robustness of the proposed method. The signal was generated with $n = 40$, $s = 3$, random frequencies, fading amplitudes, and random phases. A total number of 18 uniform samples indexed by T were taken. The noisy observations y was generated by adding complex noise w with bounded ℓ_2 norm $\varepsilon = 2$ to $x^\star_T$. We denoised and recovered the signal by solving the following optimization:

$$\underset{x}{\text{minimize}} \ \|x\|_{\mathcal{A}} \text{ subject to } \|y - x_T\|_2 \leq \varepsilon, \tag{5.2}$$



(a) Phase transition: $\Delta_f \geq \frac{1}{n}$

(b) Phase transition: $\Delta_f \geq \frac{1.5}{n}$

Figure 5: **Phase transition:** The phase transition plots were prepared with $n = 128$, and $\rho_m = 2/n : 2/n : 126/n$. The frequencies were generated randomly with minimal separation Δ_f. Both signs and magnitudes of the coefficients are random. In Figure 5a, the separation $\Delta_f \geq 1/n$ and $\rho_s = 2/n : 2/n : 100/n$, while in Figure 5b, the separation $\Delta_f \geq 1.5/n$ and $\rho_s = 2/n : 2/n : 70/n$.

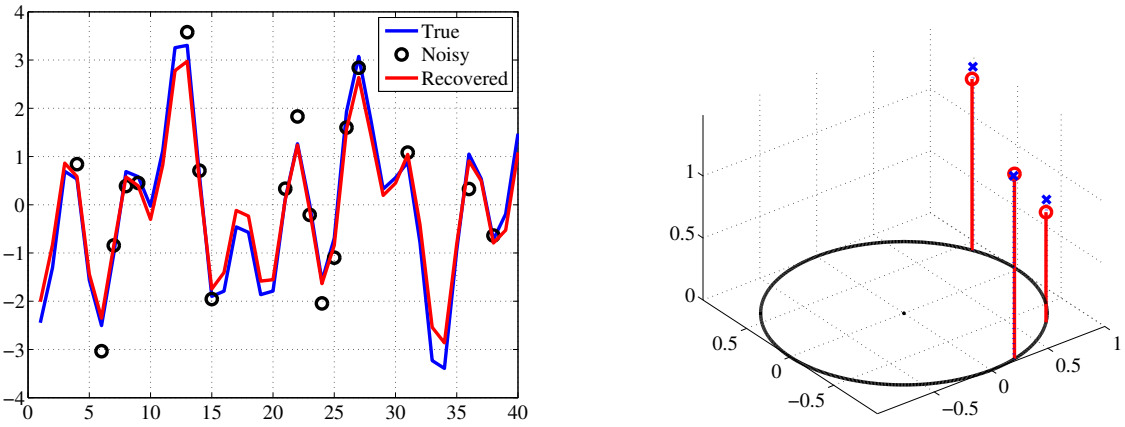
which clearly is equivalent to a semidefinite program. Matrix pencil approach was then applied to the recovered x to retrieve the frequencies. Figure 6 illustrates the approximate frequency recovery achieved by the optimization problem (5.2) in presence of noise.

6 Conclusion and Future Work

By leveraging the framework of atomic norm minimization, we were able to resolve the basis mismatch problem in compressed sensing of line spectra. For signals with well-separated frequencies, we show the number of samples needed is roughly propositional to the number of frequencies, up to polylogarithmic factors. This recovery is possible even though our continuous dictionary is not incoherent at all and does not satisfy any sort of restricted isometry conditions.

There are several interesting future directions to be explored to further expand the scope of this work. First, it would be useful to understand what happens in the presence of noise. We cannot expect exact support recovery in this case, as our dictionary is continuous and any noise will make the exact frequencies un-identifiable. In a similar vein, techniques like those used in [7] that still rely on discretization are not applicable for our current setting. However, since our numerical method is rather stable, we are encouraged that a theoretical stability result is possible.

Second, we saw in our numerical experiments that modest discretization introduces substantial error in signal reconstruction and fine discretization carries significant computational burdens. In this regard, it would be of great interest to speed up our semidefinite programming solvers so that we can scale our algorithms beyond the synthetic experiments of this paper. Our rudimentary experimentation with first-order methods developed in [4] did not suffice for this problem as they were unable to achieve the precision necessary for fine frequency localization. So, instead, it would be of interest to explore second order alternatives such as active set methods or the like to speed



(a) (b)

Figure 6: Noisy frequency recovery: (a) Real part of true, noisy, and recovered signals, (b)True frequencies (blue) and recovered frequencies (red)

up our computations.

Finally, we are interested in exploring the class of signals that are semidefinite characterizable in hopes of understanding which signals can be exactly recovered. Our continuous frequency model captures all of the essential ingredients of applying compressed sensing to problems with continuous dictionaries. It would be of great interest to see how our techniques may be extended to other continuously parametrized dictionaries. Models involving image manifolds may fall into this category [51]. Fully exploring the space of signals that can be acquired with just a few specially coded samples provides a fruitful and exciting program of future work.

Acknowledgements

The authors would like to thank Robert Nowak for many helpful discussions about this work. BR is generously supported by ONR award N00014-11-1-0723. BB, BR, and PS are generously supported by NSF award CCF-1139953. GT is generously supported by DARPA Grant Number N66001-11-1-4090.

References

[1] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak. Compressed channel sensing: A new approach to estimating sparse multipath channels. *Proc. IEEE*, 98(6):1058–1076, June 2010.

[2] R. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Process. Mag.*, 24(4):118–121, July 2007.

[3] R. Baraniuk and P. Steeghs. Compressive radar imaging. In *IEEE Radar Conf.*, pages 128–133, Waltham, MA, Apr. 2007.

[4] B. N. Bhaskar, G. Tang, and B. Recht. Atomic norm denoising with applications to line spectral estimation. *arXiv.org*, cs.IT, Apr. 2012.

[5] T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse sampling of signal innovations. *IEEE Signal Process. Mag.*, 25(2):31–40, Mar. 2008.

[6] S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ Pr, Mar. 2004.

[7] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *arXiv.org*, cs.IT, Mar. 2012.

[8] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, Dec. 2009.

[9] E. J. Candès and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969–985, Apr. 2007.

[10] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Thy.*, 52(2):489–509, Feb. 2006.

[11] E. J. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25(2):21–30, Mar. 2008.

[12] C. Carathéodory. Über den variabilitätsbereich der fourierschen konstanten von positiven harmonischen funktionen. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 32(1):193–217, 1911.

[13] C. Carathéodory and L. Fejér. Über den zusammenhang der extremen von harmonischen funktionen mit ihren koeffizienten und über den picard-landauschen satz. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 32(1):218–239, 1911.

[14] L. Carin, D. Liu, and B. Guo. Coherence, compressive sensing, and random sensor arrays. *IEEE Antennas Propag. Mag.*, 53(4):28–39, Aug. 2011.

[15] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *arXiv.org*, math.OC, Dec. 2010.

[16] Y. Chi, L. Scharf, A. Pezeshki, and A. Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Trans. Signal Process.*, 59(5):2182–2195, May 2011.

[17] B. G. R. de Prony. Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, à différentes températures. *Journal de l'école Polytechnique*, 1(22):24–76, 1795.

[18] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Prog., A*, 91(2):201–203, Mar. 2002.

[19] D. Donoho. Compressed sensing. *IEEE Trans. Inf. Thy.*, 52(4):1289–1306, Apr. 2006.

[20] E. Dowski, C. Whitmore, and S. Avery. Estimation of randomly sampled sinusoids in additive noise. *IEEE Trans. Acoust., Speech, Signal Process.*, 36(12):1906–1908, Dec. 1988.

[21] M. Duarte and R. Baraniuk. Spectral compressive sensing. *dsp.rice.edu*, July 2011.

[22] B. Dumitrescu. *Positive Trigonometric Polynomials and Signal Processing Applications*. Springer Verlag, Feb. 2007.

[23] C. Ekanadham, D. Tranchina, and E. P. Simoncelli. Neural spike identification with continuous basis pursuit. In *Computational and Systems Neuroscience (CoSyNe)*, Salt Lake City, Utah, Feb. 2011.

[24] A. C. Fannjiang, T. Strohmer, and P. Yan. Compressed remote sensing of sparse objects. *SIAM J. Imag. Sci.*, 3(3):595–618, Jan. 2010.

[25] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. `../../cvx`, Apr. 2011.

[26] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Thy.*, 57(3):1548–1566, Mar. 2009.

[27] D. Gross, Y. Liu, S. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Phys. Rev. Lett.*, 105(15):150401, Oct. 2010.

[28] M. A. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.*, 57(6):2275–2284, June 2009.

[29] M. A. Herman and T. Strohmer. General deviants: An analysis of perturbations in compressed sensing. *IEEE J. Sel. Topics Signal Process.*, 4(2):342–349, Apr. 2010.

[30] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ Pr, Cambridge, UK, Feb. 1990.

[31] Y. Hua and T. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. Acoust., Speech, Signal Process.*, 38(5):814–824, May 1990.

[32] A. Javanmard and A. Montanari. Localization from incomplete noisy distance measurements. In *Proc. IEEE Inter. Symp. Inf. Thy.*, pages 1584–1588, Saint Petersburg, Russia, July 2011.

[33] M. LeDoux. *The Concentration of Measure Phenomenon*. Amer Mathematical Society, 2001.

[34] D. Malioutov, M. Cetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Process.*, 53(8):3010–3022, Aug. 2005.

[35] A. Megretski. Positivity of trigonometric polynomials. In *Proc. 42nd IEEE Conf. Decision and Control*, volume 4, pages 3814–3817, Maui, HI, Dec. 2003.

[36] K. Mohan and M. Fazel. Reweighted nuclear norm minimization with application to system identification. In *American Contr. Conf. (ACC), 2010*, pages 2953–2959, Baltimore, MD, June 2010.

[37] C. E. Parrish and R. D. Nowak. Improved approach to lidar airport obstruction surveying using full-waveform data. *J. Surveying Eng.*, 135(2):72–82, May 2009.

[38] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Applied and Comput. Hamon. Analy.*, 22(1):16–42, Jan. 2007.

[39] B. Recht. A simpler approach to matrix completion. *J. Machine Learn.*, 12:3413–3430, Jan. 2011.

[40] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, Jan. 2010.

[41] T. Roh and L. Vandenberghe. Discrete transforms, semidefinite programming, and sum-of-squares representations of nonnegative polynomials. *SIAM J. Optim.*, 16(4):939–964, Jan. 2006.

[42] R. Sanyal, F. Sottile, and B. Sturmfels. Orbitopes. *Mathematika*, 57:275–314, 2011.

[43] A. Schaeffer. Inequalities of A. Markoff and S. Bernstein for polynomials and related functions. *Bull. Amer. Math. Soc.*, 47, Nov. 1941.

[44] M. Shaghaghi and S. A. Vorobyov. Spectral estimation from undersampled data: Correlogram and model-based least squares. *arXiv.org*, math.ST, Feb. 2012.

[45] P. Stoica. List of references on spectral line analysis. *Signal Process.*, 31(3):329–340, Apr. 1993.

[46] P. Stoica, J. Li, and H. He. Spectral analysis of nonuniformly sampled data: A new approach versus the periodogram. *IEEE Trans. Signal Process.*, 57(3):843–858, Mar. 2009.

[47] P. Stoica and R. L. Moses. *Spectral analysis of signals*. Prentice Hall, Upper Saddle River, New Jersey, 1 edition, 2005.

[48] O. Toeplitz. Zur theorie der quadratischen und bilinearen formen von unendlichvielen veränderlichen. *Mathematische Annalen*, 70(3):351–376, 1911.

[49] K. C. Toh, M. Todd, and R. H. Tütüncü. *SDPT3: A MATLAB software package for semidefinite-quadratic-linear programming*. Available from `http://www.math.nus.edu.sg/~mattohkc/sdpt3.html`.

[50] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, Aug. 2011.

[51] M. B. Wakin. A manifold lifting algorithm for multi-view compressive imaging. In *2009 Picture Coding Symp. (PCS)*, pages 1–4, Chicago, IL, May 2009.

[52] Y. Wang, J. Li, and P. Stoica. *Spectral Analysis of Signals: The Missing Data Case*. Morgan & Claypool Publishers, San Rafael, CA, 1 edition, July 2005.

A Proof of Theorem 1.1

Proof. Assume $n = 4M + n_0$ with $M = \lfloor (n-1)/4 \rfloor$ and $n_0 = 1, 2, 3$ or 4. Suppose the signal $x^\star$ has decomposition

$$x^\star = \sum_{k=1}^{s} c_k \frac{1}{\sqrt{n}} \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(n-1)f_k} \end{bmatrix}$$

$$= \sum_{k=1}^{s} \underbrace{\frac{\sqrt{4M+1}}{\sqrt{n}} c_k e^{i2\pi f_k (2M)}}_{\tilde{c}_k} \frac{1}{\sqrt{4M+1}} \begin{bmatrix} e^{i2\pi f_k(-2M)} \\ e^{i2\pi f_k(-2M+1)} \\ \vdots \\ e^{i2\pi f_k(2M)} \\ \vdots \\ e^{i2\pi f_k(2M+n_0-1)} \end{bmatrix}$$

The rest of the proof argues that the dual polynomial constructed for the symmetric case can be modified to certify the optimality of $x^\star$ for the general case.

If the coefficients $\{c_k, k = 1, \ldots, s\}$ have uniform random complex signs, for fixed $\{f_k\}$, $\{\tilde{c}_k, k = 1, \ldots, s\}$ also have uniform random complex signs. In addition, the Bernoulli observation model $\{\delta_j\}_{j=0}^{n-1}$ on index set $\{0, \cdots, n-1\}$ naturally induces a Bernoulli observation model $\{\tilde{\delta}_j = \delta_{j+2M}\}_{j=-2M}^{2M}$ on $\{-2M, \cdots, 2M\}$ with $\mathbb{P}(\tilde{\delta}_j = 1) = m/n$. Denote $\tilde{T} = \{j : \tilde{\delta}_j = 1\} \subset \{-2M, \cdots, 2M\}$. Therefore, if $\Delta_f \geq \Delta_{\min} = 1/M$ and

$$m \geq C \max \left\{ \log^2 \frac{M}{\delta}, s \log \frac{s}{\delta} \log \frac{M}{\delta} \right\}, \tag{A.1}$$

according to the proof of Theorem 2.4, with probability great than $1 - \delta$, we could construct a dual polynomial

$$\tilde{Q}(f) = \frac{1}{\sqrt{4M+1}} \sum_{j=-2M}^{2M} \tilde{q}_j e^{-i2\pi j f}$$

satisfying

$$\begin{aligned}\tilde{Q}\left(f_k\right) &= \operatorname{sign}\left(\tilde{c}_k\right), \forall f_k \in \Omega \\ \left|\tilde{Q}\left(f\right)\right| &< 1, \forall f \notin \Omega \\ \tilde{q}_j &= 0, \forall j \notin \tilde{T}.\end{aligned}$$

Now define

$$q_j = \begin{cases} \tilde{q}_{j-2M} & j = 0, \cdots, 4M \\ 0 & \text{otherwise.} \end{cases}$$

and

$$\begin{aligned} Q(f) &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} q_j e^{-i2\pi fj} \\ &= \frac{1}{\sqrt{4M+1}} \sum_{j=0}^{n-1} \tilde{q}_{j-2M} e^{-i2\pi fj} \\ &= e^{-i2\pi f(2M)} \tilde{Q}(f). \end{aligned}$$

Clearly, the polynomial $Q(f)$ satisfies

$$\begin{aligned} Q\left(f_k\right) &= e^{-i2\pi f_k(2M)} \operatorname{sign}\left(\tilde{c}_k\right) = \operatorname{sign}(c_k), \forall f_k \in \Omega \\ \left|Q\left(f\right)\right| &= |\tilde{Q}(f)| < 1, \forall f \notin \Omega \\ q_j &= 0, \forall j \notin T, \end{aligned}$$

where $T = \{j : \delta_j = 1\} \subset \{0, \ldots, n-1\}$. The theorem then follows from rewriting (A.1) in terms of n and Proposition 4.1. □

B Proof of Proposition 4.1

Proof. Consider the primal optimization problem(2.5) and its dual (4.2). Let (x, q) be primal-dual feasible. Note that

$$\begin{aligned} \langle q, x\rangle_{\mathbb{R}} &= \langle q_T, x_T\rangle_{\mathbb{R}}\,, \text{ since } q_{T^c} = 0. \\ &= \langle q_T, x_T^\star\rangle_{\mathbb{R}}\,, \text{ since } x_T = x_T^\star \\ &= \langle q, x^\star\rangle_{\mathbb{R}}\,. \end{aligned}$$

Thus, we can use $\langle q, x\rangle_{\mathbb{R}}$ in place of the dual objective $\langle q, x^\star\rangle_{\mathbb{R}}$ whenever x is primal feasible.

Since the primal is only equality constrained, Slater's condition naturally holds , implying strong duality [6, Section 5.2.3]. According to the strong duality theory, we have

$$\langle q, x\rangle_{\mathbb{R}} = \langle q, x^\star\rangle_{\mathbb{R}} \leq \|x\|_{\mathcal{A}}$$

for any x primal feasible and any q dual feasible, and

$$\langle \hat{q}, \hat{x}\rangle_{\mathbb{R}} = \langle \hat{q}, x^\star\rangle_{\mathbb{R}} = \|\hat{x}\|_{\mathcal{A}}$$

if and only if $\hat{q}$ is dual optimal and $\hat{x}$ is primal optimal.

For the dual certificate q that satisfies the conditions in Proposition 4.1, which is clearly dual feasible, we have

$$\begin{aligned}
\langle q, x^\star \rangle_{\mathbb{R}} &= \Big\langle q, \sum_{k=1}^{s} c_k a\left(f_k, 0\right) \Big\rangle_{\mathbb{R}} \\
&= \sum_{k=1}^{s} \mathrm{Re}\left(c_k^* \left\langle q, a\left(f_k, 0\right)\right\rangle\right) \\
&= \sum_{k=1}^{s} \mathrm{Re}\left(c_k^* \,\mathrm{sign}\left(c_k\right)\right) \\
&= \sum_{k=1}^{s} |c_k| \\
&\geq \|x^\star\|_{\mathcal{A}}.
\end{aligned}$$

So we must have equality and $x^\star$ is an optimal solution.

For uniqueness, suppose $\hat{x} = \sum_k \hat{c}_k a(\hat{f}_k, 0)$ with $\|\hat{x}\|_{\mathcal{A}} = \sum_k |\hat{c}_k|$ is another optimal solution. We then have for the dual certicifate q:

$$\begin{aligned}
\langle q, \hat{x} \rangle_{\mathbb{R}} &= \Big\langle q, \sum_{k} \hat{c}_k a(\hat{f}_k, 0) \Big\rangle_{\mathbb{R}} \\
&= \sum_{f_k \in \Omega} \mathrm{Re}\left(\hat{c}_k^* \left\langle q, a\left(f_k, 0\right)\right\rangle\right) + \sum_{f_l \notin \Omega} \mathrm{Re}\left(\hat{c}_l^* \left\langle q, a\left(f_l, 0\right)\right\rangle\right) \\
&< \sum_{f_k \in \Omega} |\hat{c}_k| + \sum_{f_l \notin \Omega} |\hat{c}_l| \\
&\leq \|\hat{x}\|_{\mathcal{A}}
\end{aligned}$$

due to condition (4.5) if $\hat{x}$ is not solely supported on Ω. So all optimal solutions are supported on Ω. Since for both $J = \{-2M, \cdots, 2M\}$ and $\{0, \cdots, n-1\}$, the set of atoms with frequencies in Ω are linearly independent, the optimal solution is unique. □

C Proof of Proposition 4.2

Proof. Under the assumption that $\Delta_{\min} \geq \frac{1}{M}$, we cite the results of [7, Proof of Lemma 2.2] as follows:

$$\begin{aligned}
\left\| I - \bar{D}_0 \right\|_\infty &\leq 6.253 \times 10^{-3} \\
\left\| \frac{1}{\sqrt{\left|\bar{K}''\left(0\right)\right|}} \bar{D}_1 \right\|_\infty &\leq 4.212 \times 10^{-2} \\
\left\| I - \Big(- \frac{1}{\left|\bar{K}''\left(0\right)\right|} \bar{D}_2 \Big) \right\|_\infty &\leq 0.3201,
\end{aligned}$$

where $\|\cdot\|_\infty$ is the matrix infinity norm, namely, the maximum absolute row sum. Since $I-\bar{D}$ is symmetric and has zero diagonals, the Geršhgorin circle theorem [30] implies that

$$\begin{aligned}
&\left\|I-\bar{D}\right\| \leq \left\|I-\bar{D}\right\|_\infty \\
&\leq \max\Big\{ \left\|I-\bar{D}_0\right\|_\infty + \left\|\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}}\bar{D}_1\right\|_\infty, \\
&\qquad \left\|\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}}\bar{D}_1\right\|_\infty + \left\|I-\Big(-\frac{1}{\left|\bar{K}''(0)\right|}\bar{D}_2\Big)\right\|_\infty\Big\} \\
&= 0.3623.
\end{aligned}$$

As a consequence, $\bar{D}$ is invertible and

$$\begin{aligned}
\|\bar{D}\| &\leq 1+\left\|I-\bar{D}\right\| \leq 1.3623, \\
\left\|\bar{D}^{-1}\right\| &\leq \frac{1}{1-\left\|I-\bar{D}\right\|} \leq 1.568.
\end{aligned}$$

□

D Proof of Lemma 4.5

Proof of Lemma 4.5. We start with computing the quantities necessary to apply Lemma 4.4:

$$\begin{aligned}
&\mathbb{E}X_j = 0 \\
&\|X_j\| = \left\|\frac{1}{M}g_M(j)(\delta_j-p)\,e(j)\,e(j)^*\right\| \\
&\qquad \leq \frac{1}{M}\|g_M\|_\infty\, s\left(1+\max_{|j|\leq 2M}\frac{(2\pi j)^2}{\left|K''(0)\right|}\right) \\
&\qquad \leq R := 14\frac{s}{M} \text{ for } M\geq 4.
\end{aligned}$$

Here we have used

$$\begin{aligned}
&\|g_M\|_\infty \leq 1, \\
&\|e(j)\|_2^2 = s\left(1+\max_{|j|\leq 2M}\frac{(2\pi j)^2}{\left|K''(0)\right|}\right) \leq 14s, \text{ for } M\geq 4.
\end{aligned}$$

We continue with σ^2:

$$\begin{aligned}
\sigma^2 &= \left\|\sum_{j=-2M}^{2M}\mathbb{E}\left(\frac{1}{M^2}g_M^2(j)(\delta_j-p)^2\|e(j)\|_2^2 e(j)\,e^*(j)\right)\right\| \\
&\leq 14\frac{p(1-p)}{M}s\left\|\frac{1}{M}\sum_{j=-2M}^{2M}g_M^2(j)\,e(j)\,e^*(j)\right\|.
\end{aligned}$$

To further bound σ^2, we note

$$\begin{aligned}
&\frac{1}{M}\sum_{j=-2M}^{2M} g_M^2(j)\, e(j)\, e^*(j) \\
\preccurlyeq\ & \|g_M\|_\infty \left\{ \frac{1}{M}\sum_{j=-2M}^{2M} g_M(j)\, e(j)\, e^*(j) \right\} \\
=\ & \|g_M\|_\infty \bar{D},
\end{aligned}$$

which leads to

$$\begin{aligned}
&\left\| \frac{1}{M}\sum_j g_M^2(j)\, e(j)\, e^*\Big(j\Big) \right\| \\
=\ & \lambda_{\max}\Big(\frac{1}{M}\sum_j g_M^2(j)\, e(j)\, e^*(j) \Big) \\
\le\ & \lambda_{\max}\left(\|g_M\|_\infty \bar{D} \right) \\
=\ & \|g_M\|_\infty \left\|\bar{D}\right\| \\
\le\ & 1.3623 \|g_M\|_\infty \text{ by (4.15) and (4.10).}
\end{aligned}$$

Therefore, we have

$$\sigma^2 \le 20\frac{p}{M}s.$$

Invoking the non-commutative Bernstein's inequality and setting $t = p\tau$, we have

$$\begin{aligned}
\mathbb{P}\left(\left\|p^{-1}D - \bar{D}\right\| \ge \tau \right) &\le 2s \exp\left(\frac{-p^2\tau^2/2}{20\frac{p}{M}s + 14\frac{s}{M}p\tau/3} \right) \\
&\le 2s\exp\left(-\frac{1}{50}\tau^2\frac{m}{s} \right) (\text{used } \tau \le 1) \\
&\le \delta.
\end{aligned}$$

if

$$m \ge \frac{50}{\tau^2} s \log\frac{2s}{\delta}.$$

Consequently, when $\tau < 1 - 0.3623 \le 1 - \left\|I - \bar{D}\right\|$ according to (4.14), we have $\left\|I - p^{-1}D\right\| \le \left\|I - \bar{D}\right\| + \left\|p^{-1}D - \bar{D}\right\| < 1$, confirming the invertibility of $p^{-1}D$. □

E Proof of Corollary 4.6

Assuming B is invertible and $\|A - B\| \left\|B^{-1}\right\| \le \frac{1}{2}$, we have the following two inequalities:

$$\begin{aligned}
\left\|A^{-1}\right\| &\le \frac{\left\|B^{-1}\right\|}{1 - \|A - B\| \left\|B^{-1}\right\|} \le 2\left\|B^{-1}\right\| \\
\left\|A^{-1} - B^{-1}\right\| &\le \frac{\|A - B\| \left\|B^{-1}\right\|^2}{1 - \|A - B\| \left\|B^{-1}\right\|} \le 2\left\|B^{-1}\right\|^2 \|A - B\|,
\end{aligned}$$

which are rearrangements of

$$\begin{aligned}\left\|A^{-1}-B^{-1}\right\| &\leq \left\|A^{-1}\right\|\|A-B\|\left\|B^{-1}\right\| \\ \left\|A^{-1}\right\| &\leq \left\|A^{-1}-B^{-1}\right\|+\left\|B^{-1}\right\| \\ &\leq \left\|A^{-1}\right\|\|A-B\|\left\|B^{-1}\right\|+\left\|B^{-1}\right\|.\end{aligned}$$

Therefore, we establish that when $\tau \leq \frac{1}{4} < \frac{1}{2\left\|\bar{D}^{-1}\right\|}$ on the set $\mathcal{E}_{1,\tau}$:

$$\begin{aligned}\left\|D^{-1}-p^{-1}\bar{D}^{-1}\right\| &\leq 2\left\|p^{-1}\bar{D}^{-1}\right\|^2\left\|D-p\bar{D}\right\| = 2\left\|\bar{D}^{-1}\right\|^2 p^{-1}\tau \\ \left\|D^{-1}\right\| &\leq 2\left\|p^{-1}\bar{D}^{-1}\right\| = 2\left\|\bar{D}^{-1}\right\| p^{-1}.\end{aligned}$$

Since the operator norm of a matrix dominates that of all submatrices, this competes the proof.

F Proof of Lemma 4.7

The proof uses Talagrand's concentration of measure inequality:

Lemma F.1 ([33, Corollary 7.8]). *Let $\{Y_j\}$ be a finite sequence of independent random variables taking values in a Banach space and let V be defined as*

$$V = \sup_{h\in\mathcal{H}} \sum_j h\left(Y_j\right)$$

for a countable family of real valued functions $\mathcal{H}$. Assume that $|h| \leq B$ and $\mathbb{E}h\left(Y_j\right) = 0$ for all $h \in \mathcal{H}$ and every j. Then for all $t > 0$,

$$\mathbb{P}\left(|V-\mathbb{E}V| > t\right) \leq 16\exp\left(-\frac{t}{KB}\log\left(1+\frac{Bt}{\sigma^2+B\mathbb{E}\bar{V}}\right)\right),$$

where $\sigma^2 = \sup_{h\in\mathcal{H}}\sum_j \mathbb{E}h^2(Y_j)$, $\bar{V} = \sup_{h\in\mathcal{H}}\left|\sum_j h\left(Y_j\right)\right|$, and K is a numerical constant.

Proof of Lemma 4.7. Based on the definition of $v_\ell(f)$ in (4.35) and $\bar{v}_\ell(f)$ in (4.19), we explicitly write $v_\ell\left(f\right) - p\bar{v}_\ell\left(f\right) = v_\ell(f) - \mathbb{E}v_\ell(f)$ as

$$\begin{aligned}v_\ell\left(f\right) - p\bar{v}_\ell\left(f\right) &= \sum_{j=-2M}^{2M}\frac{1}{M}\left(\frac{i2\pi j}{\sqrt{\left|\bar{K}''\left(0\right)\right|}}\right)^\ell g_M\left(j\right)\left(\delta_j - p\right)e^{i2\pi f j}e(j) \\ &= \sum_{j=-2M}^{2M} Y_j^\ell,\end{aligned}$$

where $e(j)$ is defined in (4.32) and we have defined Y_j^ℓ as

$$Y_j^\ell = \frac{1}{M}\left(\frac{i2\pi j}{\sqrt{\left|\bar{K}''\left(0\right)\right|}}\right)^\ell g_M\left(j\right)\left(\delta_j - p\right)e^{i2\pi f j}e(j).$$

It is clear that $\{Y_j^\ell\}_{j=-2M}^{2M}$ are independent random vectors with zero mean.

Define

$$V^\ell := \|v_\ell(f) - p\bar{v}_\ell(f)\|_2 = \sup_{h:\|h\|_2=1} \langle v_\ell(f) - p\bar{v}_\ell(f), h\rangle_{\mathbb{R}} = \sup_{h\in\mathbb{C}^{2s}:\|h\|_2=1} \sum_{j=-2M}^{2M} \langle Y_j^\ell, h\rangle_{\mathbb{R}}$$

and

$$h(Y_j^\ell) = \langle Y_j^\ell, h\rangle_{\mathbb{R}} = \mathrm{Re}\Big(\sum_{k=1}^{s} h_k^* Y_{j,k}^\ell\Big).$$

To compute the quantities necessary to apply Lemma F.1, we will extensively use the following elementary bounds:

$$\begin{aligned} &\|g_M\|_\infty \le 1, \\ &\Big|\frac{2\pi j}{\sqrt{\bar{K}_M''(0)}}\Big| \le 4 \text{ when } M \ge 2, \\ &\|e(j)\|_2^2 \le s\left(1 + \max_{|j|\le 2M} \frac{(2\pi j)^2}{|K''(0)|}\right) \le 14s \text{ when } M \ge 4. \end{aligned}$$

First, we obtain an upper bound on $|h|$:

$$\begin{aligned} |h(Y_j^\ell)| &= \left|\left\langle \frac{1}{M}\left(\frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}}\right)^\ell g_M(j) e^{i2\pi f j} e(j)(\delta_j - p), h\right\rangle_{\mathbb{R}}\right| \\ &\le \frac{1}{M}\left|\frac{i2\pi j}{\sqrt{|\bar{K}''(0)|}}\right|^\ell \|g_M\|_\infty \|e(j)\|_2 \\ &\le B_\ell := 4^{\ell+1}\frac{\sqrt{s}}{M}. \end{aligned}$$

The expected value of $\|v_\ell(f) - p\bar{v}_\ell(f)\|_2^2$ is upper bounded as follows:

$$\begin{aligned} \mathbb{E}\|v_\ell(f) - p\bar{v}_\ell(f)\|_2^2 &= \sum_{j=-2M}^{2M} \mathbb{E}\left\langle Y_j^\ell, Y_j^\ell\right\rangle_{\mathbb{R}} + \sum_{j\ne k} \mathbb{E}\left\langle Y_j^\ell, Y_k^\ell\right\rangle_{\mathbb{R}} \\ &= \sum_{j=-2M}^{2M} \mathbb{E}\left\langle Y_j^\ell, Y_j^\ell\right\rangle_{\mathbb{R}} \\ &\le \sum_{j=-2M}^{2M} \frac{1}{M^2}\left|\frac{2\pi j}{\sqrt{|\bar{K}''(0)|}}\right|^{2\ell} g_M^2(j) p(1-p)\|e(j)\|_2^2 \\ &\le 4^{2\ell+3}\frac{ps}{M} \text{ when } M \ge 4. \end{aligned} \tag{F.1}$$

Observe that $\bar{V}^\ell = V^\ell = \|v_\ell(f) - p\bar{v}_\ell(f)\|_2$. We apply Jensen's inequality and combine with (F.1) to get

$$\begin{aligned}\mathbb{E}\bar{V}^\ell = \mathbb{E}V^\ell &\le \sqrt{\mathbb{E}V^{\ell 2}} \le \sqrt{4^{2\ell+3}\frac{ps}{M}} \\ &\le 2^{2\ell+3}\frac{\sqrt{ms}}{M}.\end{aligned}$$

Next, we upper bound σ^2:

$$\begin{aligned}&\mathbb{E}h^2(Y_j^\ell) = \langle Y_j^\ell, h\rangle_{\mathbb{R}}^2 \\ &\le \frac{1}{M^2}4^{2\ell}\|g_M\|_\infty \mathbb{E}(\delta_j - p)^2 \left|\left\langle \sqrt{g_M(j)}e(j), h\right\rangle\right|^2\end{aligned}$$

implying

$$\begin{aligned}\sum_j \mathbb{E}h^2(Y_j^\ell) &\le \frac{1}{M^2}4^{2\ell}p\sum_{j=-2M}^{2M}\left|h^*\sqrt{g_M(j)}e(j)\right|^2 \\ &= \frac{1}{M^2}4^{2\ell}p\|h^*P\|_2^2 \\ &\le 4^{2\ell}\frac{p\|P\|^2}{M^2}\end{aligned}$$

where P is a matrix in $\mathbb{C}^{2s\times(4M+1)}$ whose jth column is $\sqrt{g_M(j)}e(j)$. Note that

$$\frac{PP^*}{M} = \frac{1}{M}\sum_{j=-2M}^{2M} g_M(j)e(j)e(j)^* = \bar{D}.$$

Therefore, we have

$$\begin{aligned}\sigma_\ell^2 = \sum_j \mathbb{E}h^2(Y_j) &\le 4^{2\ell}\frac{p}{M^2}\|P\|^2 \\ &\le 4^{2\ell}\frac{1}{M^2}pM\left\|\bar{D}\right\| \\ &\le 2^{4\ell+1}\frac{m}{M^2}\left(\text{used } \left\|\bar{D}\right\| \le 2 \text{ from } (4.15)\right)\end{aligned}$$

In conclusion, Lemma F.1 shows that

$$\begin{aligned}&\mathbb{P}\left(\left|\|v_\ell(f) - p\bar{v}_\ell(f)\|_2 - \mathbb{E}\|v_\ell(f) - p\bar{v}_\ell(f)\|_2\right| > t\right) \\ &\le 16\exp\left(-\frac{t}{KB_\ell}\log\left(1 + \frac{B_\ell t}{\sigma_\ell^2 + B_\ell\mathbb{E}\bar{V}^\ell}\right)\right) \\ &\le 16\exp\left(-\frac{t}{KB_\ell}\log\left(1 + \frac{B_\ell t}{2^{4\ell+1}\frac{m}{M^2} + B_\ell 2^{2\ell+3}\frac{\sqrt{ms}}{M}}\right)\right)\end{aligned}$$

Suppose now $\bar{\sigma}_\ell^2 = B_\ell 2^{2\ell+3}\frac{\sqrt{ms}}{M} \ge 2^{4\ell+1}\frac{m}{M^2}$, and fix $t = a\bar{\sigma}_\ell$. Then it follows that

$$\mathbb{P}\left(\left|\|v_\ell(f) - p\bar{v}_\ell(f)\|_2 - \mathbb{E}\|v_\ell(f) - p\bar{v}_\ell(f)\|_2\right| > a\bar{\sigma}_\ell\right) \le 16e^{-\gamma a^2},$$

for some $\gamma > 0$ provided $B_\ell t \le \bar{\sigma}_\ell^2$. The same is true if $\bar{\sigma}_\ell^2 = 2^{4\ell+1}\frac{m}{M^2} \ge B_\ell 2^{2\ell+3}\frac{\sqrt{ms}}{M}$ and $B_\ell t \le 2^{4\ell+1}\frac{m}{M^2}$. Therefore, let

$$\begin{aligned}\bar{\sigma}_\ell^2 &= \max\left\{2^{4\ell+1}\frac{m}{M^2}, B_\ell 2^{2\ell+3}\frac{\sqrt{ms}}{M}\right\} \\ &= 2^{4\ell+1}\frac{m}{M^2}\max\left\{1, 2^4\frac{s}{\sqrt{m}}\right\},\end{aligned}$$

and fix $a > 0$ obeying

$$a \le \begin{cases}\sqrt{2}m^{1/4} & \text{if } 2^4 s/\sqrt{m} \ge 1 \\ \frac{\sqrt{2}}{4}\sqrt{\frac{m}{s}} & \text{otherwise.}\end{cases}$$

Then we have

$$\mathbb{P}\left(\left\|v_\ell(f) - p\bar{v}_\ell(f)\right\|_2 > 2^{2\ell+1}\frac{\sqrt{ms}}{M} + a\bar{\sigma}_\ell\right) \le 16e^{-\gamma a^2}$$

for some $\gamma > 0$. Application of union bound proves the lemma. □

G Proof of Lemma 4.9

The proof of Lemma 4.9 is based on Hoeffding's inequality presented below:

Lemma G.1 (Hoeffding's inequality)**.** *Let the components of $u \in \mathbb{C}^n$ be sampled i.i.d. from a symmetric distribution on the complex unit circle, $w \in \mathbb{C}^n$, and t be a positive real number. Then*

$$\mathbb{P}\left(|\langle u, w\rangle| \ge t\right) \le 4e^{-\frac{t^2}{4\|w\|_2^2}}.$$

Proof of Lemma 4.9. Consider the random inner product $\langle u, L^*(v_\ell(f) - p\bar{v}_\ell(f))\rangle$ where $\{u_j\}$ are i.i.d. symmetric random variables with values on the complex unit circle. Conditioned on a particular realization

$$\omega \in \mathcal{E} := \left\{\omega : \sup_{f_d \in \Omega_{\text{grid}}} \left\|L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\right\|_2 < \lambda_\ell, \ell = 0,1,2,3\right\},$$

Hoeffding's inequality and union bound then imply

$$\mathbb{P}\Big(\sup_{f_d \in \Omega_{\text{grid}}} |\langle u, L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\rangle| > \varepsilon \Big| \omega\Big) \le 4\left|\Omega_{\text{grid}}\right| e^{-\frac{\varepsilon^2}{4\lambda_\ell^2}}.$$

Elementary probability calculation shows

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f_d \in \Omega_{\text{grid}}} |\langle u, L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\rangle| > \varepsilon\Big) \\ &= 4|\Omega_{\text{grid}}|e^{-\frac{\varepsilon^2}{4\lambda^2}} + \mathbb{P}\left(\mathcal{E}^c\right).\end{aligned}$$

Setting

$$\lambda_\ell = 4\left(2^{2\ell+1}\sqrt{\frac{s}{m}} + \frac{M}{m}a\bar{\sigma}_\ell\right)$$

in $\mathcal{E}$ and applying Lemma 4.8 yield,

$$\begin{aligned}&\mathbb{P}\Big(\sup_{f_d\in\Omega_{\text{grid}}} |\langle u, L^*(v_\ell(f_d) - p\bar{v}_\ell(f_d))\rangle| > \varepsilon \Big| \omega\Big)\\ &\le 4\left|\Omega_{\text{grid}}\right| e^{-\frac{\varepsilon^2}{4\lambda^2}} + 64\left|\Omega_{\text{grid}}\right| e^{-\gamma a^2} + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right)\end{aligned}$$

For the second term to be less than δ, we choose a such that

$$a^2 = \gamma^{-1}\log\frac{64\left|\Omega_{\text{grid}}\right|}{\delta},$$

and assume this value from now on. The first term is less than δ if

$$\frac{1}{\lambda^2} \ge \frac{4}{\varepsilon^2}\log\frac{4\left|\Omega_{\text{grid}}\right|}{\delta}. \tag{G.1}$$

First assume that $2^4 s/\sqrt{m} \ge 1$. The condition in Lemma 4.7 is $a \le \sqrt{2}m^{1/4}$ or equivalently

$$m \ge \frac{1}{4}\gamma^{-2}\log^2\frac{64\left|\Omega_{\text{grid}}\right|}{\delta}. \tag{G.2}$$

In this case, we have $a\bar{\sigma}_\ell \le 2^{2\ell+3}\frac{\sqrt{ms}}{M}$, leading to

$$\frac{1}{\lambda^2} = \frac{1}{16\left(2^{2\ell+1}\sqrt{\frac{s}{m}} + \frac{M}{m}a\bar{\sigma}_1\right)^2} \ge \frac{1}{2^{4\ell+6}25}\frac{m}{s}.$$

Now suppose that $2^4 s/\sqrt{m} \le 1$. If $32s \ge a^2$, then $a\bar{\sigma}_\ell \le 2^{2\ell+3}\frac{\sqrt{ms}}{M}$ which again gives the above lower bound on $1/\lambda^2$. On the other hand if $32s \le a^2$, then $\lambda \le 5\sqrt{2}2^{2\ell-2}\frac{a}{\sqrt{m}}$ and

$$\frac{1}{\lambda^2} \ge \frac{1}{2^{4\ell-3}25}\frac{m}{a^2}$$

Therefore, to verify (G.1) it suffices to take m obeying (G.2) and

$$m\min\left(\frac{1}{2^{4\ell+6}25}\frac{1}{s}, \frac{1}{2^{4\ell-3}25}\frac{1}{a^2}\right) \ge \frac{4}{\varepsilon^2}\log\frac{4\left|\Omega_{\text{grid}}\right|}{\delta}$$

This analysis shows that the first term is less than δ if

$$\begin{aligned}m \ge \max\Big\{&\frac{4}{\varepsilon^2}2^{4\ell+6}25s\log\frac{4|\Omega_{\text{grid}}|}{\delta}, \frac{4}{\varepsilon^2}2^{4\ell-3}25\gamma^{-1}\log\frac{64|\Omega_{\text{grid}}|}{\delta}\log\frac{4|\Omega_{\text{grid}}|}{\delta},\\ &\frac{1}{4}\gamma^{-2}\log^2\frac{64\left|\Omega_{\text{grid}}\right|}{\delta}\Big\}.\end{aligned}$$

According to Lemma 4.5, the last term is less than δ if

$$m \ge \frac{50}{\tau^2}s\log\frac{2s}{\delta}.$$

Setting $\tau = 1/4$, combining all lower bounds on m together, and absorbing all constants into one, we get

$$m \geq C \max\left\{\frac{1}{\varepsilon^2}\max\left(s\log\frac{|\Omega_{\text{grid}}|}{\delta}, \log^2\frac{|\Omega_{\text{grid}}|}{\delta}\right), s\log\frac{s}{\delta}\right\},$$

is sufficient to guarantee

$$\sup_{f_d\in\Omega_{\text{grid}}}\left|I_1^\ell(f_d)\right| \leq \varepsilon$$

with probability at least $1-3\delta$. Union bound then proves the lemma. □

H Proof of Lemma 4.10

Proof of Lemma 4.10. Recall that

$$I_2^\ell(f) = \left\langle u, (L - p^{-1}\bar{L})^* p\bar{v}_\ell(f)\right\rangle$$

On the set $\mathcal{E}_{1,\tau}$ defined in (4.33), we established in Corollary 4.6 that

$$\left\|L - p^{-1}\bar{L}\right\| \leq 2\left\|\bar{D}^{-1}\right\|^2 p^{-1}\tau.$$

We use the ℓ_1 norm to bound the ℓ_2 norm of $p\bar{v}_\ell(f)$:

$$\begin{aligned}\left\|p\bar{v}_\ell(f)\right\|_2 &\leq \left\|p\bar{v}_\ell(f)\right\|_1 \\ &= p\left(\sum_{k=1}^s \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^\ell}\left|\bar{K}_M^{(\ell)}(f-f_k)\right| + \sum_{k=1}^s \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^{(\ell+1)}}\left|\bar{K}_M^{(\ell+1)}(f-f_k)\right|\right).\end{aligned}$$

To get a uniform bound on $\sum_{k=1}^s \frac{1}{\sqrt{|\bar{K}''(0)|}^\ell}\left|\bar{K}_M^{(\ell)}(f-f_k)\right|$, we need the following bound:

$$\frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^\ell}\left|\bar{K}_M^{(\ell)}(f)\right| \leq \begin{cases} C_1 & \forall f\in[-\frac{1}{2},\frac{1}{2}], \\ \frac{C_2}{M^4|f|^4} & \text{if } \frac{1}{4M}\leq|f|\leq\frac{1}{2}.\end{cases}$$

for suitably chosen numerical constant C_1 and C_2. The bound over the region $[\frac{1}{4M},\frac{1}{2}]$ is a consequence of the more accurate bound established in [7, Lemma 2.6], while the uniform bound C_1 can

be obtained by checking the expression of $\bar{K}_M^{(\ell)}(f)$. Consequently, we have

$$\begin{aligned}
&\sum_{k=1}^{s} \frac{1}{\sqrt{\left|\bar{K}''(0)\right|}^{\ell}} \left|\bar{K}_M^{(\ell)}(f-f_k)\right| \\
&\leq \sum_{k:|f-f_k|<\frac{2}{M}} C_1 + \sum_{k:\frac{2}{M}\leq|f-f_k|\leq\frac{1}{2}} \frac{C_2}{M^4|f-f_k|^4} \\
&\leq 4C_1 + C_2 \sum_{k=1}^{\infty} \frac{1}{M^4(k\Delta_{\min})^4} \\
&\leq 4C_1 + C_2 \sum_{k=1}^{\infty} \frac{1}{k^4} \\
&= C := 4C_1 + \frac{\pi^4}{90} C_2.
\end{aligned}$$

We conclude that on the set $\mathcal{E}_{1,\tau}$

$$\|(L - p^{-1}\bar{L})^* p \bar{v}_\ell(f)\|_2 \leq C\tau.$$

Again, application of Hoeffding's inequality and union bound gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{f_d \in \Omega_{\text{grid}}} \left|I_2^{\ell}(f_d)\right| > \varepsilon\right) \\
&\quad \leq 4\left|\Omega_{\text{grid}}\right| \exp\left(-\frac{\varepsilon^2}{4C\tau^2}\right) + \mathbb{P}\left(\mathcal{E}_{1,\tau}^c\right).
\end{aligned}$$

To make the first term less than δ, it suffices to take

$$\tau^2 = \frac{\varepsilon^2}{4C \log \frac{4\left|\Omega_{\text{grid}}\right|}{\delta}}.$$

To have the second term less than δ, we require

$$\begin{aligned}
m &\geq \frac{C}{\tau^2} s \log \frac{2s}{\delta} \\
&= \frac{C}{\frac{\varepsilon^2}{\log \frac{4\left|\Omega_{\text{grid}}\right|}{\delta}}} s \log \frac{2s}{\delta} \\
&= C \frac{1}{\varepsilon^2} s \log \frac{2s}{\delta} \log \frac{4\left|\Omega_{\text{grid}}\right|}{\delta}.
\end{aligned}$$

Another application of union bound with respect to $\ell = 0, 1, 2, 3$ proves the lemma. □