The Threshold for Super-resolution via Extremal Functions

Ankur Moitra [*]

October 16, 2021

Abstract

Super-resolution is a natural mathematical abstraction for the problem of extracting fine-grained structure from coarse-grained measurements, and has received considerable attention following the pioneering works of Donoho [11] and Candes and Fernandez-Granda [5, 6]. Here we introduce new techniques based on *extremal functions* for studying this and related problems and we exactly resolve the threshold at which noisy super-resolution is possible. In particular, we establish a sharp phase transition for the relationship between the cutoff frequency (m) and the separation (Δ). If $m > 1/\Delta + 1$, our estimator converges to the true values at an inverse polynomial rate in terms of the magnitude of the noise. And when $m < (1-\epsilon)/\Delta$ no estimator can distinguish between a particular pair of Δ-separated signals even if the magnitude of the noise is exponentially small. Our results involve making novel connections between extremal functions and spectral properties of the Vandermonde matrix, such as bounding its condition number as well as constructing explicit preconditioners for it.

---

[*]Massachusetts Institute of Technology. Department of Mathematics and the Computer Science and Artificial Intelligence Lab. Email: `moitra@mit.edu`. This work is supported in part by a grant from the MIT NEC Corporation.

1   Introduction

1.1   Background

Super-resolution is a fundamental task in imaging, where the goal is to resolve features at a scale beyond the diffraction limit. Recently this problem and its extensions have garnered considerable attention in the statistics and machine learning communities [11, 5, 6, 3] where the mathematical model used to study this problem is described next. We are interested in a super-position of k point sources:

$$x(t) = \sum_{j=1}^{k} u_j \delta_{f_j}(t)$$

where $\delta_\tau(t)$ is the Dirac delta function at τ and we assume that each $f_j \in [0,1)$. Additionally each measurement v_ℓ of these point sources takes the form

$$v_\ell = \int_0^1 e^{i2\pi \ell t} x(t)\ dt + \eta_\ell = \sum_{j=1}^{k} u_j e^{i2\pi f_j \ell} + \eta_\ell$$

where η_ℓ is noise in the measurement. However due to physical limitations, we are only allowed to measure the data at low-frequencies — i.e. we can observe v_ℓ for $|\ell| \leq m$ where m is referred to as the *cutoff frequency*. The basic question is, how large does m need to be in order to recover the locations f_j and the coefficients u_j? The problem of recovering fine-grained structure from coarse-grained measurements has a broad set of applications including in domains such as medical imaging, microscopy, astronomy, radar, geophysics and spectroscopy (see [11, 26, 5] and references therein). This is also a fundamental algorithmic and statistical problem in its own right.

   This particular estimation problem has a long history of study, and there are a variety of methods that work in the noise-free case for $m = k$ such as Prony's method (1795) [13], Pisarenko's method (1973) [27], the matrix pencil method (1990) [19] and other linear prediction methods (see references in [35]). This is optimal since if we had any fewer than $2k + 1$ total measurements regardless of their frequency, the true signal would not necessarily be the sparsest solution to the inverse problem [15]. Here we study this problem in the presence of noise, and establish that it exhibits a sharp phase transition. Let Δ denote the minimum separation between any pair of the f_j's according to the wrap-around metric on the unit interval. Our main algorithmic result shows that if $m > 1/\Delta + 1$ then noisy super-resolution is possible and our estimates converge to the true values at an inverse polynomial rate in terms of the magnitude of the noise. This is tight in that if $m = (1 - \epsilon)/\Delta$ there is a pair of k point sources x and x' each with separation Δ where we would need the noise to be exponentially small to tell them apart.

   Along the way, we also prove a sharp phase transition for the condition number of the Vandermonde matrix based on connections to certain extremal functions studied in analytic number theory. To elaborate on this, one can think of a hypothesis testing version of various learning problems. Then in many settings we need a family of test functions that are strong enough to distinguish any pair of hypothesis with suitably different parameters. For mixtures of Gaussians, low degree polynomials work [20], [24], [1] but we can think of the present paper as fitting into a broader agenda where the goal is to find more creative choices for these test functions which in turn lead to better algorithms (see also [22] where a similar framework is used to design algorithms for the population recovery problem).

1.2   Previous Work and Connections

Recall that there are a variety of algorithms for super-resolution that work for $m = k$ in the noise-free case. We review one popular technique called the *matrix pencil method* in Section 2.1 where the idea is to express the unknown parameters as the solution to a generalized eigenvalue problem. There have been numerous attempts to analyze the noise tolerance of these methods, but all of them pass to limiting arguments as m tends to infinity and yield ineffective bounds. The main bottleneck is that this approach relies on solving linear systems in Vandermonde matrices, which can be very poorly conditioned. In fact, many other algorithms for inverse problems that use closely related techniques work in the exact case but are either hopeless or require very different settings of their parameters in the presence of noise. For example:

(a) Given the evaluation of a degree k polynomial $p(x)$ at $k+1$ distinct points we can solve for the coefficients of $p(x)$ and then we can evaluate $p(x)$ anywhere. But if we are given the values of $p(x)$ up to some additive noise η even for *all* values of $x \in [0,1]$ we can only hope to estimate $p(2)$ within an additive $c(2+\sqrt{3})^k\eta$. In fact this is tight by standard results in approximation theory [10].

(b) In sparse recovery, there is an unknown s-sparse signal x and we observe $Ax+\eta$. If $\eta = 0$ and A is chosen to be the first $2s$ rows of the discrete fourier transform matrix we can recover x exactly using Prony's method. But A must have at least $\Omega(s \log n/s)$ rows in order to recover x in the presence of some moderate amount of noise [14].

We could wonder whether super-resolution suffers from the same type of extreme noise-intolerance that makes it possible in the exact case, but requires (say) η_ℓ to be *exponentially* small in order to recover the u_j's and f_j's to any reasonable accuracy.

Prior to the more modern work on super-resolution, there was a considerable effort to develop empirical methods that work when there is some a priori information about the sparsity of the signal (see references in [11]). In a remarkable work, Donoho formulated the notion of the *Rayleigh index* which can be thought of as a continuous analogue to separation, and studied super-resolution where the f_j's are restricted to be on a grid [11]. Donoho used Beurling's theory of interpolation [2] to prove that any Δ-separated signal can be uniquely recovered from from its fourier transform on the continuous range $[-1/\Delta, 1/\Delta]$. However this result is information theoretic in the sense that there could be two distinct signals that are both Δ-separated, but whose fourier transforms restricted to $[-1/\Delta, 1/\Delta]$ differ in a measure zero set. (Hence they are not noticeably different). Donoho also gave results on the modulus of continuity for the recovery of Δ-separated signals, but these results lose an extra factor of two in the range of low-frequency measurements that are needed. The main question left open by this work was to prove recovery bounds when the signal is not restricted to be on a grid.

Subsequently, Candes and Fernandez-Granda [5, 6] proposed an exciting new method for solving this inverse problem based on convex programming. The authors proved a variety of results in the setting where (a) there is no noise or (b) there is noise, but the f_j's are restricted to be on a grid. In the noise-free setting, the authors showed their approach recovers the u_j's and f_j's exactly provided that $m \geq \frac{2}{\Delta}$ and $m \geq 128$ where Δ is the minimum separation. The authors further improved this bound when the u_j's are real to a separation condition of $m \geq \frac{1.87}{\Delta}$ (unpublished)[1]. In the grid setting, the authors defined a parameter C called the *super-resolution factor* and proved various bounds using it on the error of their estimator. However the approach used to measure the error is non-standard since it uses the Fejer kernel to 'mask' the error on the high-frequency parts of the signal. Finally, Ferndanez-Granda [16] returned to the setting where the f_j's can be at arbitrary locations, and showed that if the noise is small enough the estimate becomes localized around the true spikes.

In a related work, Bhaskar et al [3] defined a problem called *compressed sensing off the grid* where we are given a subset of the low frequency measurements. More precisely, we are given v_ℓ for all $\ell \in S \subset \{-m, -m+1, ..., m-1, m\}$ where $|S|$ is roughly $O(k \log m/k)$. The authors considered the noise-free setting and proved that if S is chosen uniformly at random and $m \geq 2/\Delta$, the convex program recovers the u_j's and f_j's exactly. Additionally, the authors needed a technical assumption that each $u_j \in \{\pm 1\}$ and is chosen uniformly at random.

1.3   Our Results and Techniques

Here we show the first noise tolerance bounds for the *matrix pencil method* (see Section 2.1) and we use this to give algorithms for noisy super-resolution with the *optimal* relationship between the number of measurements and the separation. Our key technical ingredient is giving a tight characterization of when the Vandermonde matrix $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ (see Section 2.1) is well conditioned. Set $\alpha_j = e^{i2\pi f_j}$. Recall that Δ is the minimum separation of the f_j's according to the wrap-around metric on the unit interval.

Theorem 1.1. *The condition number κ of $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ satisfies $\kappa^2 \leq \frac{m+1/\Delta-1}{m-1/\Delta-1}$ provided that $m > \frac{1}{\Delta} + 1$.*

---

[1]One should note though, that the methods we described above work in the noise-free setting for $m = k$ without any separation condition. So the crucial issue really is in understanding when super-resolution is possible in the presence of noise.

This theorem is easy to prove, given the right extremal functions. The key is the Beurling-Selberg majorant and minorant[2] (see Section 2.2) which majorize and minorize the sign function on the real line but have a compactly supported fourier transform [28]. In fact, these functions are the solutions to their respective extremal problems and have led to a sharp forms of various inequalities in analytic number theory. Most notably these functions can be used to give simple, direct proofs of the large sieve inequality (see [38]) and extensions of Hilbert's inequality [25]. These inequalities yield strong upper and lower bounds for integrals of sums of exponentials, and here we adapt the techniques to our setting to reason about evaluating sums of exponentials at discrete sets of points. This readily yields our upper bound on the condition number.

Indeed it is the condition number of V_m^k that governs whether or not the matrix pencil method is stable. Despite its widespread use in signal processing, this is the first *effective* bound on its noise tolerance and we get immediate consequences for super-resolution. Throughout this paper we will quantify the error between our estimates $\{(\widehat{u}_j, \widehat{f}_j)\}$ and the true values $\{(u_j, f_j)\}$ based on the best matching between them:

$$\min_\pi \max_j \max \left(\|\widehat{u}_j - u_{\pi(j)}\|, \|\widehat{f}_j - f_{\pi(j)}\| \right)$$

where π is a permutation. In particular, we refer to our estimates as being ϵ-close to the true parameters if there is a matching of the k spikes in our estimate to the true k spikes so that across the matching we get both the coefficients and the locations correct within an additive ϵ.

Theorem 1.2. *If the cut-off frequency satisfies $m > 1/\Delta + 1$ where Δ is the minimum separation (according to the wrap-around metric), then there is a polynomial time algorithm to recover the u_j's and f_j's whose estimates converge to the true values at an inverse polynomial rate in terms of the magnitude of the noise.*

We will state the precise quantitative bounds later (see Theorem 2.8). These results are tight (as we will describe next) and settle the question of what is the best relationship between the number of measurements and the minimum separation that we can achieve for super-resolution. Notice that this theorem has none of the restrictions that were needed in some of the previous work; there are no restrictions on the f_j's being on a grid, or on the coefficients.

As we noted earlier, if $k = 1/\Delta$ and $m < 1/\Delta$ then the true signal x is not necessarily the sparsest solution to the inverse problem. This is related to the fact that V_m^k is not full rank. However it is easy to give more compelling lower bounds where k is itself much smaller than m. Candes and Fernandez-Granda [5] proved almost what we need, and gave an x that is Δ-separated where $\int_E \|\widehat{x}(t)\|^2 \, dt < 2^{-\Omega(\epsilon k)}$ for $E = [-m/2, m/2]$ and $m = (1 - \epsilon)/\Delta$. Their approach was based on the asymptotics of Slepian's discrete prolate spheroidal wave functions which were developed in a sequence of papers [31, 32, 21, 29, 30]. Here we give a discrete version of the above result, and give an elementary and self-contained proof. First, we prove:

Theorem 1.3. *If $m < (1 - \epsilon)/\Delta$ and $k = \Omega(\log m)$, the condition number of $V_m^k(\alpha_{k/2}, \alpha_{k/2+1}, ..., \alpha_{3k/2-1})$ is at least $2^{\Omega(\epsilon k)}$.*

It turns out that this gives us a pair of point sources that are each 2Δ-separated, where setting the cut-off frequency $m = (1 - \epsilon)/2\Delta$ we would need the magnitude of the noise to be exponentially small in order to tell them apart. Hence we cannot even solve the following hypothesis testing variant in the presence of noise:

Corollary 1.4 (Informal). *If the cut-off frequency satisfies $m < (1-\epsilon)/\Delta$ where Δ is the minimum separation (according to the wrap-around metric), then there is a pair x and x' of k point sources each with separation Δ where we would need $|\eta_\ell| \leq 2^{-\Omega(\epsilon k)}$ in order to tell x and x' apart.*

This proves the optimality of our algorithms in a strong sense. In fact, this impossibility result holds even in a stochastic model, where the noise η_ℓ has its real and complex part chosen as Gaussian independent random variables with variance $2^{-\Omega(\epsilon k)}$. We state the precise result later (see Corollary 3.2).

We can think about our use of majorants and minorants as being a "two-function method" where we upper and lower bound some function we are interested in — in this case $\|V_m^k u\|_2^2$. We develop another framework which we call a "one-function method" in which we prove upper bounds on the condition number

---

[2]In fact, the minorant is the more important of the two extremal functions for our purposes. It is the opposite for most of the standard applications, where the majorant plays the key role and the minorant is a footnote.

of V_m^k in an alternate manner. For example, the proof of the Salem inequality (see Section 4.2) relies on lower bounding some integral we are interested in as:

$$\int_{-\infty}^{\infty} \chi(t)\|f(t)\|^2 \, dt \leq \int_E \|f(t)\|^2 \, dt$$

where $\chi(t)$ is an appropriately chosen "test function" that is supported on the interval E. If additionally $\widehat{f}$ is well-spaced and $\widehat{\chi}$ decays quickly, then we can write the left hand side as a relatively simple sum-of-squares plus some correction terms.

Our key observation is that if we transfer these types of inequalities to the discrete setting, the test function itself can be thought of as a way to re-weight the rows of V_m^k to precondition it. This is a very different setting than usual because we do not know $V_m^k = V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ — the matrix we want to precondition — only a family that it belongs to. We call this *universal preconditioning*, which seems to be an interesting notion in its own right. The key point is that $\chi(t)$ is an universal preconditioner for any V_m^k where the f_j's are well-separated, and we can use it to design much faster algorithms for super-resolution. Here we restrict to the case where each $u_j = 1$ to keep things simple. Our final result is:

Theorem 1.5. *If the cut-off frequency satisfies $m > \Omega((1/\Delta)\log 1/\epsilon)$ where Δ is the minimum separation (according to the wrap-around metric) and each $u_j = 1$, there is an $\widetilde{O}(m^2)$ time algorithm to recover the f_j''s up to accuracy ϵ, provided that the noise satisfies $\|\eta_j\| \leq \epsilon^2/(4k)$ for each j.*

We prove this result by using $\chi(t)$ to construct a function $F(z)$ whose local minima are ϵ-close to the f_j's. Moreover we establish that $F(z)$ is approximately strongly convex in a neighborhood around each f_j, and we give a variant of gradient descent that finds all of them in parallel. This approach is much faster than trying to solve the convex program in [5] or finding *all* the generalized eigenvalues for a pair of $m \times m$ Toeplitz matrices as in the matrix pencil method. We remark that least-squares is a popular approach to solve super-resolution, but most analyses pass to a limiting argument and use the fact that as m goes to infinity, the (re-normalized) columns of V_m^k become orthogonal. Our approach instead is to precondition V_m^k and we believe that this idea may have other algorithmic applications in signal processing.

2 The Optimal Separation Criteria

2.1 The Matrix Pencil Method

Here we review the matrix pencil method, which is a basic tool in statistics, signal processing and learning. The basic idea is to set up a generalized eigenvalue problem whose solutions are the unknown parameters. In particular, in the setting of super-resolution we are given

$$v_\ell = \sum_j u_j e^{i2\pi f_j \ell} + \eta_\ell$$

for integers $|\ell| \leq m$ and where η_ℓ is noise. We will limit our discussion (for now) to the idealized case to explain the method, and later we will see that it is the condition number of the Vandermonde matrix V_m^k (defined below) that controls whether or not this method is noise tolerant. Let $\alpha_j = e^{-2\pi i f_j}$, then:

$$V_m^k(\alpha_1, \alpha_2, ..., \alpha_k) = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_k \\ \vdots & \vdots & \ldots & \vdots \\ \alpha_1^{m-1} & \alpha_2^{m-1} & \ldots & \alpha_k^{m-1} \end{bmatrix}$$

Moreover let $D_u = \text{diag}(\{u_j\})$ and let $D_\alpha = \text{diag}(\{\alpha_j\})$. Recall that in a generalized eigenvalue problem we are given a pair of matrices A and B and the goal is to find the generalized eigenvalues λ for which there is a vector x that satisfies $Ax = \lambda Bx$. In particular, the eigenvalues of A are the solution to the generalized eigenvalue problem where $B = I$.

Then the following well-known observations are the key to the matrix pencil method:

4

Observation 1. *The non-zero generalized eigenvalues of VD_uV^H and $VD_uD_\alpha V^H$ are exactly $\{\alpha_j\}$.*

Observation 2. *The entry in row i, column j of VD_uV^H is v_{i-j} and the corresponding entry in $VD_uD_\alpha V^H$ is v_{i-j+1}*

Hence the entries in these matrices are precisely the low-frequency measurements we have access to. Now it is straightforward to find the unknown frequencies $\{f_j\}$ by setting up the above generalized eigenvalue problem and solving it. Moreover we can find the coefficients too by solving the following linear system:

$$v = V_n^k(\alpha_1, \alpha_2, ..., \alpha_k)u$$

where v and u are the vectors of $\{v_\ell\}$'s and $\{u_j\}$'s respectively.

This approach succeeds provided that $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ has full column rank, which is true if the f_j's are distinct and it has at least as many rows as columns. Thus as long as the cutoff frequency satisfies $m \geq k+1$ we can solve the super-resolution problem exactly in the noise-free setting. This algorithmic guarantee has been rediscovered numerous times. However this approach can sometimes be extremely intolerant to noise. What governs the noise-tolerance of the above generalized eigenvalue problem (and consequently the overall method) is the condition number of $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$. There are by now many known perturbation bounds for generalized eigenvalue problems (see [34]), however we must take extra care in that the rectangular Vandermonde matrix will have full column rank but will not be invertible in our setting. We defer proving perturbation bounds for the above problem until Section 2.4, but let us for now focus on understanding when the Vandermonde matrix is well-conditioned.

2.2 The Beurling-Selberg Majorant

The Beurling-Selberg majorant $B(t)$ plays a key role in analytic number theory, where it has been used to establish sharp forms of the large sieve inequality and also important extensions to the classic Hilbert inequality. In particular:

Definition 2.1. $B(t) = \left(\frac{\sin \pi t}{\pi}\right)^2 \left(\sum_{j=0}^{\infty}(t-j)^{-2} - \sum_{j=-\infty}^{-1}(t-j)^{-2} + 2t^{-1}\right)$

We will describe its key properties next but for now we remark that $B(t)$ is an entire function of exponential type 2π - i.e. its fourier transform $\widehat{B}$ is supported on $[-1, 1]$. Indeed, the definition above looks quite strange at first glance but any function of exponential type 2π satisfies a certain interpolation formula based on its values at the integers (see e.g. [37] or [39]) and $B(t)$ comes from plugging in the sign function instead (which is not of exponential type 2π).

The key properties of $B(t)$ are that it majorizes the sign function - i.e. $\text{sgn}(t) \leq B(t)$ - and that:

$$\int_{-\infty}^{\infty} B(t) - \text{sgn}(t) \ dt = 1$$

In fact, $B(t)$ is the extremal function that satisfies the above properties. Hence any function $F(t)$ of exponential type 2π that majorizes the sign function satisfies

$$\int_{-\infty}^{\infty} F(t) - \text{sgn}(t) \ dt \geq 1$$

and equality holds if and only if $F(t) = B(t)$. There is a similar construction that minorizes the sign function, and is also extremal. However these functions are usually used to majorize and minorize a given interval, and we will instead state the theorems we need restricted to that setting.

Let $E = [-n, +n]$ and let $I_E(t)$ be the indicator function for E. Then

Theorem 2.2. *There are entire functions $C_E(t)$ and $c_E(t)$ whose fourier transform is supported on $[-\Delta, \Delta]$ and that satisfy $c_E(t) \leq I_E(t) \leq C_E(t)$ and*

$$\int_{-\infty}^{\infty} C_E(t) - I_E(t) \ dt = \int_{-\infty}^{\infty} I_E(t) - c_E(t) \ dt = 1/\Delta$$

The functions $C_E(t)$ and $c_E(t)$ are called majorants and minorants respectively. In [25], Montgomery and Vaughan used these properties of $C_E(t)$ and $c_E(t)$ to give a strong generalization of Hilbert's inequality. Here, we will modify their approach and prove upper bounds on the condition number of the Vandermonde matrix.

2.3 Upper Bounding the Condition Number

Here we will adapt the approach of Montgomery and Vaughan [25] to give an upper bound on the condition number of V_m^k. Recall that $\alpha_j = e^{i2\pi f_j}$ and d_w is the wrap-around distance on the unit interval. Then suppose each $f_j \in [0,1)$ and the minimum separation is strictly larger than Δ - i.e. $d_w(f_j, f_{j'}) > \Delta$ for $j \neq j'$. We prove that $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ is well-conditioned provided that $m > \frac{1}{\Delta} + 1$. Moreover we will prove in Section 3 that if $m \leq \frac{1-\epsilon}{\Delta}$ then V_m^k can have condition number that is exponentially large in k, so there is a sharp transition.

Let $\widehat{C}_E$ denote the fourier transform of C_E and let $h(\ell) = \sum_{\ell=-\infty}^{\infty} \delta(\ell)$ denote the Dirac comb. Recall that $E = [-n, +n]$, $I_E(t)$ is the indicator function for E and

$$v_\ell = \sum_{j=1}^{k} u_j e^{i2\pi f_j \ell}$$

Then we can write

$$
\begin{aligned}
\sum_{\ell=-n}^{n} \|v_\ell\|^2 &= \int_{-\infty}^{\infty} h(\ell) I_E(\ell) \|v_\ell\|^2 \, d\ell \leq \int_{-\infty}^{\infty} h(\ell) C_E(\ell) \|v_\ell\|^2 \, d\ell \\
&= \sum_{j=1}^{k} \sum_{j'=1}^{k} u_j \bar{u}_{j'} \int_{-\infty}^{\infty} h(\ell) C_E(\ell) e^{i2\pi(f_j - f_{j'})\ell} \, d\ell \\
&= \sum_{j=1}^{k} \sum_{j'=1}^{k} \sum_{t=-\infty}^{\infty} u_j \bar{u}_{j'} \widehat{C}_E(f_j - f_{j'} + t) = \widehat{C}_E(0) \sum_{j=1}^{k} \|u_j\|^2
\end{aligned}
$$

where the last equality follows because $|f_j - f_{j'} + t| > \Delta$ for any integer t, and because $\widehat{C}_E$ is supported on $[-\Delta, \Delta]$. Moreover

$$\widehat{C}_E(0) = \int_{-\infty}^{\infty} C_E(\ell) \, d\ell = |E| + 1/\Delta = 2n + 1/\Delta$$

We can instead use the minorant $c_E(\ell)$ and combined with the above inequality we get

$$(2n - 1/\Delta) \sum_{j=1}^{k} \|u_j\|^2 \leq \sum_{\ell=-n}^{n} \|v_\ell\|^2 \leq (2m + 1/\Delta) \sum_{j=1}^{k} \|u_j\|^2$$

Now we can make the substitution $b_j = u_j e^{-i2\pi f_j n}$ in which case

$$V_{2n+1}^k(\alpha_1, \alpha_2, ..., \alpha_k) b = \sum_{\ell=-n}^{n} \|v_\ell\|^2 = (2n \pm 1/\Delta) \sum_{j=1}^{k} \|b_j\|^2$$

Recall that the condition number $\kappa(V) = \frac{\max_a \|Va\|}{\min_{a'} \|Va'\|}$ and hence we obtain the following theorem by setting $m = 2n + 1$:

Theorem 2.3. *The condition number κ of $V_m^k(\alpha_1, \alpha_2, ..., \alpha_k)$ satisfies $\kappa^2 \leq \frac{m + 1/\Delta - 1}{m - 1/\Delta - 1}$ provided that $m > \frac{1}{\Delta} + 1$.*

2.4 The Generalized Eigenvalue Problem

There are a number of well-known results that bound the noise-tolerance of a generalized eigenvalue problem (see [34] and references therein). However often these results assume that the matrices A and B are both full rank. We will be able to reduce to this case and appeal to their perturbation bounds as a black-box. For convenience, we will adopt the same notation as in [34] and let us first focus on settings where A and B are square matrices of order k. We are interested in the solutions to $\det(\alpha A - \beta B) = 0$.

Additionally let $\lambda = \alpha/\beta$. Then the pair (A, B) is called *regular* if there is any choice of α and β for which $\det(\alpha A - \beta B) \neq 0$. Of course, if either A or B is full rank then the pair is regular. The bounds in [34] are stated in terms of the chordal metric, which will be convenient for our purposes.

Definition 2.4. The chordal metric for $\lambda, \mu \in \mathbb{C}$ is defined as

$$\chi(\lambda, \mu) = \frac{|\lambda - \mu|}{\sqrt{1 + |\lambda|^2}\sqrt{1 + |\mu|^2}}$$

Equivalently if we take $s(\lambda)$ and $s(\mu)$ to be the points on $\mathbb{S}^2$ so that λ and μ are their stereographic projections, then $\chi(\lambda, \mu) = \frac{1}{2}\|s(\lambda) - s(\mu)\|_2$. This metric is well-suited for our applications since our eigenvalues λ will anyways be on the boundary of the complex disk.

Throughout this section let $\widehat{A} = A + E$ and let $\widehat{B} = B + F$. Let $\widehat{\lambda}$ be the generalized eigenvalues for the pair $\widehat{A}, \widehat{B}$. We will work with the following measure:

Definition 2.5. The matching distance with respect to χ is defined as

$$\mathrm{md}_\chi[(A, B), (\widehat{A}, \widehat{B})] = \min_\pi \max_i \chi(\lambda_i, \widehat{\lambda}_{\pi(i)})$$

where π is a permutation on $\{1, 2, ..., k\}$.

Note that we can define the matching distance analogously using another metric. In fact, we will state the guarantees of the matrix pencil method based on the matching distance of the output $\widehat{f}_j$ to the true f_j's according to the wrap-around metric.

We will make use of the following theorem, which combines Theorem 2.4 and Corollary 2.5 from [34]:

Theorem 2.6. *[34] Let (A, B) and $(\widehat{A}, \widehat{B})$ be regular pairs and further suppose that for some nonsingular X we have $(XAX^H, XBX^H) = (I, D)$ where D is diagonal. Also let e_i be the ith row of $X(A - \widehat{A})X^H$ and let f_i be the ith row of $X(B - \widehat{B})X^H$ and set $\rho = \max_i\{\|e_i\|_1 + \|f_i\|_1\}$. If the following regions*

$$\mathcal{G}_i = \left\{\mu \Big| \chi(D_{ii}, \mu) \leq \rho\right\}$$

are disjoint then the matching distance of the generalized eigenvalues of $(\widehat{A}, \widehat{B})$ to $\{D_{ii}\}$ is at most ρ with respect to the chordal metric.

2.5 A Modified Matrix Pencil Method

Now we are ready to prove error bounds for the matrix pencil method. Recall that we are interested in solving the generalized eigenvalue problem $(VD_uV^H, VD_uD_\alpha V^H)$ where V is short-hand for V_m^k, $D_u = \mathrm{diag}(u_j)$ and $D_\alpha = \mathrm{diag}(\alpha_j)$. Instead we are given $VD_uV^H + E$ and $VD_uD_\alpha V^H + F$. We will first reduce to a generalized eigenvalue problem on order k matrices so that the matrix pair will be regular. Ideally, we would project onto the column span of V to preserve the fact that the α_j's are the generalized eigenvalues. However we do not have access to this space, and instead we will project onto the top k singular vectors of $VD_uV^H + E$. Let $\widehat{U}$ be an orthonormal basis for this space.

It follows from Wedin's theorem (see e.g. [18]) that the column space of V and the span of $\widehat{U}$ have a small principal angle. In fact, an even stronger statement is true: we can align the basis of the column span of V to $\widehat{U}$. Throughout this section, let $u_{min} = \min_j \|u_j\|$ and $u_{max} = \max_j \|u_j\|$.

Lemma 2.7. *There is an orthonormal basis U for the column span of V that satisfies $\|U - \widehat{U}\|_2 \leq \frac{2\|E\|_2}{u_{min}}$.*

Proof. By Wedin's theorem and the definition of the principal angle we have that

$$\sin(\Theta(U,\widehat{U})) = \|(I - UU^H)\widehat{U}\|_2 \leq \frac{\|E\|_2}{u_{min}}$$

Then using Weyl's inequality (see e.g. [18]) we can write

$$\|U - \widehat{U}\|_2 \leq \|UU^H\widehat{U} - \widehat{U}\|_2 + \|U - UU^H\widehat{U}\|_2$$

The first term is just $\sin(\Theta(U,\widehat{U}))$. Also we can use the fact that $\cos(\Theta(U,\widehat{U})) = \sigma_{\min}(U^H\widehat{U})$. Then we can bound the second term as

$$\|U - UU^H\widehat{U}\|_2 \leq \|I - U^H\widehat{U}\|_2 \leq 1 - \cos(\Theta(U,\widehat{U})) \leq \sin(\Theta(U,\widehat{U})) \leq \frac{\|E\|_2}{u_{min}}$$

And this completes the proof. □

Now let $A = U^H V D_u V^H U$, let $B = U^H V D_u D_\alpha V^H U$ and let $X = D_u^{1/2} V^+ U$. It follows that $XAX^H = I$ and $XBX^H = D$. Also consider the perturbed generalized eigenvalue problem for $\widehat{A} = \widehat{U}^H(VD_uV^H + E)\widehat{U}$ and $\widehat{B} = \widehat{U}^H(VD_uD_\alpha V^H + F)\widehat{U}$. Let us set $\|U - \widehat{U}\|_2 = \tau$ for notational convenience.

Then we can write $\|X(A - \widehat{A})X^H\|_2 \leq \|X\|_2^2\|A - \widehat{A}\|_2$. And furthermore

$$\|A - \widehat{A}\|_2 \leq \|E\|_2 + \|U^H V D_u V^H U - \widehat{U}^H V D_u V^H \widehat{U}\|_2 \leq \|E\|_2 + \left(2\tau\|V\|_2^2 + \tau^2\|V\|_2^2\right)u_{max}$$

And combining these inequalities and $\tau = \frac{2\|E\|_2}{u_{min}}$ we have

$$\|X(A - \widehat{A})X^H\|_2 \leq \frac{\|E\|_2}{\sigma_{min}^2 u_{min}} + 4\kappa^2\left(\|E\|_2 + \|E\|_2^2\right)\frac{u_{max}}{u_{min}}$$

where κ and $\sigma_{\min}$ are the condition number and minimum singular value of V respectively. An identical bound holds for B and $\widehat{B}$ where we replace E by F. We remark that a bound on the spectral norm of $X(A - \widehat{A})X^H$ is also a bound on the ℓ_2-norm of any of its rows; and to get a bound on the ℓ_1-norm of any its rows, we can multiply the above bound by $k^{1/2}$.

Hence we can combine the above bounds and Theorem 2.6 to get that the generalized eigenvalues of $(\widehat{A}, \widehat{B})$ converge to the generalized eigenvalues of (A, B) at an inverse polynomial rate with the noise, with respect to the chordal metric. We remark that the generalized eigenvalues of (A, B) are on the complex disk. Hence if $(\widehat{\alpha}, \widehat{\beta})$ is a generalized eigenvalue of $(\widehat{A}, \widehat{B})$, we can find the closest point on the complex disk to $\widehat{\lambda} = \widehat{\alpha}/\widehat{\beta}$ which in turn will be have arc-length at most $O(\chi(\widehat{\lambda}, \lambda))$ where $\lambda = \alpha/\beta$ and (α, β) is the generalized eigenvalue of (A, B) that is matched to $(\widehat{\alpha}, \widehat{\beta})$.

Hence the output of MODIFIEDMPM is a set of locations $\{\widehat{f}_j\}$ whose matching distance to $\{f_j\}$ with respect to d_w converges at an inverse polynomial rate to zero with the magnitude of the error. All that remains is to recover the coefficients too. To this end, we write $\widehat{V}$ to denote the Vandermonde matrix with $\widehat{f}_j$'s instead of f_j's. Then we are given

$$v = Vu + \eta = \widehat{V}u + (V - \widehat{V})u + \eta$$

Then if we solve the linear system $v = \widehat{V}\widehat{u}$ we get $\widehat{u}$ such that

$$\|u - \widehat{u}\|_2 \leq \frac{\|V - \widehat{V}\|_2\|u\|_2}{\|\eta\|_2}$$

Let the matching distance between $\{f_j\}$ and $\{\widehat{f}_j\}$ be ρ. Then it is easy to see that each entry in $V - \widehat{V}$ is bounded by $O(\rho m)$ and hence

$$\|u - \widehat{u}\|_2 \leq O\left(\frac{\rho m^{3/2}k u_{max} + \|\eta\|_2}{m - 1 - \frac{1}{\Delta - 2\rho}}\right)$$

since the minimum separation of $\{\widehat{f}_j\}$ is at least $\Delta - 2\rho$.

We can now state our main theorem, which gives a quantitative rate for how quickly our estimates converge to the true values:

8

Algorithm 1 MODIFIEDMPM, **Input:** k and estimates v_ℓ for $|\ell| \le m$
Output: $\{(\widehat{u}_j, \widehat{f}_j)\}$

---

1. Let $\widetilde{A}$ and $\widetilde{B}$ be matrices where $\widetilde{A}_{i,j} = v_{i-j}$ and $\widetilde{B}_{i,j} = v_{i-j+1}$
2. Let $\widehat{U}$ be the top k singular vectors of $\widetilde{A}$
3. Set $\widehat{A} = \widehat{U}^H \widetilde{A} \widehat{U}$ and $\widehat{B} = \widehat{U}^H \widetilde{B} \widehat{U}$
4. Solve for the generalized eigenvalues of $(\widehat{A}, \widehat{B})$
5. For each generalized eigenvalue $\widehat{\lambda} = \widehat{\alpha}/\widehat{\beta}$, let $\widehat{f}_j$ be the argument of the projection of $\widehat{\lambda}$ onto the complex disk
6. Set $\widehat{V}$ to be the Vandermonde matrix with $\{\widehat{f}_j\}$
7. Solve for $\widehat{u}$ such that $\widehat{V}\widehat{u} = v$
8. Output $\{(\widehat{u}_j, \widehat{f}_j)\}$

Theorem 2.8. *Let η be the vector with entries $\{\eta_\ell\}$ and set*

$$\gamma = \frac{k\|\eta\|_2}{\sigma_{min}^2 u_{min}} + 4\kappa^2\Big(k\|\eta\|_2 + k^{3/2}\|\eta\|_2^2\Big)\frac{u_{max}}{u_{min}}$$

and also

$$\zeta = O\Big(\frac{\gamma m^{3/2} k u_{max} + \|\eta\|_2}{m - 1 - \frac{1}{\Delta - 2\gamma}}\Big)$$

Further suppose that $\|E\|_2 + \|F\|_2 < \sigma_{min}^2 u_{min}$ and $\gamma < \Delta/4$. Then the output of MODIFIEDMPM satisfies

$$d_w(f_j, \widehat{f}_{\pi(j)}) \le \gamma \text{ and } \|u_j - \widehat{u}_{\pi(j)}\| \le \zeta$$

for all j, for some permutation π.

The important point is that the estimates converge at an inverse polynomial rate to the true values as a function of $\|\eta\|_2$ provided that $m > (1 + \epsilon)/\Delta$ for any $\epsilon > 0$.

3 The Need for Separation

In the previous section, we gave an upper bound for the condition number of V_m^k based on m and Δ. Here we establish that there is a sharp phase transition, and if $m = (1 - \epsilon)/\Delta$ then V_m^k has an exponentially large condition number. From this it follows that if $m = (1 - \epsilon)/\Delta$ there is a pair of of k point sources x and x' each with separation Δ which we would need exponentially small noise to tell apart.

3.1 Impossibility Results

In order to make the connection to super-resolution it will actually be more convenient to work with the shifted Vandermonde matrix (below), but all the results in this section carry-over the other case as well.

$$V_{-n/2,n/2}^k(\alpha_1, \alpha_2, ..., \alpha_k) = \begin{bmatrix} \alpha_1^{-n/2} & \alpha_2^{-n/2} & \ldots & \alpha_k^{-n/2} \\ \alpha_1^{-n/2+1} & \alpha_2^{-n/2+1} & \ldots & \alpha_k^{-n/2+1} \\ \vdots & \vdots & \ldots & \vdots \\ \alpha_1^{n/2} & \alpha_2^{n/2} & \ldots & \alpha_k^{n/2} \end{bmatrix}$$

Throughout this section we will use slightly different notation where we will be interested in delta functions spaced at intervals of length $1/m$ and we will use n to denote the number of measurements. Let $n = (1-\epsilon)m$.

Theorem 3.1. *Let $k = \Omega(\log m)$. Then the matrix $V_{-n/2,n/2}^k$ where $\alpha_j = e^{i2\pi j/m}$ has condition number at least $2^{\Omega(\epsilon k)}$.*

Throughout this section let V be short-hand for $V^k_{-n/2,n/2}$. It is easy to upper bound the maximum singular value of V and hence the culprit is the smallest singular value. We conclude that there is a vector u with $\|u\|_2 = 1$ and $\|Vu\|_2 \leq 2^{-\Omega(\epsilon k)}$. Then:

$$\sum_{k=-n/2}^{n/2} \left(\sum_{j \text{ is odd}} u_j e^{i2\pi jk/m} + \sum_{j \text{ is even}} u_j e^{i2\pi jk/m} \right)^2 = \|Vu\|_2^2 \leq 2^{-\Omega(\epsilon k)}$$

And hence we have two sets of point sources corresponding to the odd and even indices respectively, with separation $\Delta = 2/m$ but where measuring them up to the cutoff frequency $n/2 = (1-\epsilon)m/2 = (1-\epsilon)/\Delta$ would require exponentially small error to tell them apart.

Corollary 3.2. *There is a pair of k point sources x and x' each with separation Δ where if $m = (1-\epsilon)/\Delta$ and the noise η_ℓ has its real and complex part chosen as Gaussian independent random variables with variance $2^{-\Omega(\epsilon k)}$, then for any integer ℓ with $|\ell| \leq m$, $d_{TV}(v_\ell, v'_\ell) \leq 2^{-\Omega(\epsilon k)}$ where v_ℓ and v'_ℓ are the random variables associated with measuring x and x' at frequency ℓ respectively.*

Recall that in the noise-free case, the separation is irrelevant. It is possible to reconstruct the point sources with $m = k$. But in the noisy version of the problem, k is irrelevant and the minimum number of measurements essentially only depends on the minimum separation. Let us now prove the main theorem in this section. We will make use of the Fejer kernel here (and later), although many other kernels could be used in its place.

3.2 The Fejer Kernel I: Basic Properties

Here we introduce the Fejer kernel, which we will use in the next subsection to construct a unit vector u where $\|Vu\|_2$ is exponentially small.

Definition 3.3. The Fejer kernel[3] is

$$K_\ell(x) = \frac{1}{\ell^2} \sum_{j=-\ell}^{\ell} (\ell - |j|) e^{i2\pi jx} = \frac{1}{\ell^2} \left(\frac{\sin \ell\pi x}{\sin \pi x} \right)^2$$

Here we have given both the fourier representation and the closed form for the Fejer kernel. We will use $\widehat{K}_\ell$ to denote its fourier transform. The key is that $\widehat{K}_\ell$ is supported on $\{-\ell, -\ell+1, ...\ell-1, \ell\}$ and yet $K_\ell(x)$ decays quickly. We will also make use of powers of the Fejer kernel which we denote by $K^r_\ell(x)$ and note that its fourier transform is supported on $\{-r\ell, -r\ell+1, ..., r\ell-1, r\ell\}$.

We make use of the following well-known property of the Fejer kernel:

Fact 3.4. $K_\ell(x) \leq \frac{1}{4\ell^2 x^2}$ for $x \in [-1/2, 1/2]$ and $K^r_\ell(x) \leq \frac{1}{4^r \ell^{2r} x^{2r}}$

Proof. It is well known that $\sin \pi x \geq 2x$ for $x \in [0, 1/2]$ and substituting this bound into the denominator and using $(\sin \ell\pi x)^2 \leq 1$ implies the bound on $K_\ell(x)$ and similarly for $K^r_\ell(x)$. $\square$

3.3 Lower Bounding the Condition Number

We are now ready to prove Theorem 3.1.

Proof. Consider $H(x) = K^r_\ell(x/m)e^{i\pi x}$. We will set $\ell = 4/\epsilon$ and $r = (k-1)/2\ell$. We will establish some elementary properties of $H(x)$ and its fourier transform $\widehat{H}$:

Claim 3.5. $\widehat{H}(t) = \sum_{j=k/2}^{3k/2-1} h_j e^{i2\pi jt/m}$ *and* $\sum_j |h_j| = 1$.

---

[3]We use a slightly non-standard definition for the Fejer kernel, where ours will be normalized so that $K_\ell(0) = 1$ and periodic on $[0, 1)$ instead of $[0, 2\pi)$ since this suits our applications more conveniently.

Proof. Let $\widehat{K}_\ell$ denote the fourier transform of $K_\ell(x/m)$. Then we can write $\widehat{H}$ as the convolution of $\widehat{K}_\ell$ with itself r times, then convolved with $\delta(1/2)$. This directly implies the first part of the claim. Furthermore, the coefficients of $\widehat{K}_\ell$ are nonnegative and sum to one, so we can interpret both $\widehat{K}_\ell$ and $\widehat{H}$ as distributions on a discrete set of points, and this yields the second part of the claim. $\qquad\square$

We will let u denote the vector whose entries are the h_j's. Then $\|u\|_2 \geq \|u\|_1/\sqrt{k} = 1/\sqrt{k}$ because u has k non-zeros. Now we give an upper bound for $\|Vu\|_\infty$.

Claim 3.6. $\|Vu\|_\infty \leq \max_{|x|\leq n/2} |H(x)|$

Proof. The claim is immediate since each coordinate in Vu is equal to $H(j)$ for some $j \in \{-n/2, -n/2 + 1, ...n/2\}$. $\qquad\square$

Now all that remains is to bound the right hand side above. We can use Fact 3.4 and $K_\ell^r(x) \leq \frac{1}{(8x/\epsilon)^{2r}}$. Moreover $H(x) = K_\ell^r(x/m - 1/2)$ and so

$$\max_{|x|\leq n/2} |H(x)| \leq \frac{1}{4^{2r}} \leq 2^{-\Omega(\epsilon k)}$$

where we have used the fact that $K_\ell^r(x)$ has period one and that $r = \Omega(\epsilon k)$. This completes the proof of the theorem. $\qquad\square$

4 Universal Preconditioning

In the previous sections, we analyzed the condition number of V_m^k based on m and the minimum separation Δ according to the wrap-around distance, and this gave us algorithms for super-resolution that are *optimal* in the relationship between the number of measurements and the minimum separation. Here we pursue a different direction - namely, is there a way to precondition V_m^k to make it almost orthogonal? Just like the Beurling-Selberg majorant played the key role in the previous section, here the Fejer kernel will itself be a good preconditioner for a family of Vandermonde matrices.

4.1 The Fejer Kernel II: Local Approximation

In Section 3.2 we introduced the Fejer kernel and stated some of its basic properties. Here we will also need need upper and lower bounds for $K_\ell(x)$ in a neighborhood of zero, that we establish here for later reference. Recall that

$$K_\ell(x) = \frac{1}{\ell^2}\left(\frac{\sin \ell\pi x}{\sin \pi x}\right)^2 \text{ and } K_\ell^r(x) = \left(K_\ell(x)\right)^r$$

Lemma 4.1. *For $\ell \geq 4$ and $|x| \leq 1/\ell$ and any integer $r \geq 1$ we have*

$$1 - Cr\ell^2 x^2 \leq K_\ell^r(x) \leq 1 - c\ell^2 x^2$$

where we can take $C = 12$ and $c = 1/3$.

Proof. In order to analyze the behavior of $K_\ell^r(x)$ in a neighborhood of zero, we need upper and lower bounds for $\sin x$ that behave like $x - \Theta(x^3)$. We will make crucial use of the following facts, which are easy to check:

Fact 4.2. $x(1 - \frac{x^2}{2}) \leq \sin x$ *for all $x \geq 0$*

Fact 4.3. $\sin x \leq x(1 - \frac{x^2}{10})$ *for all $x \in [0, \pi)$*

Then substituting these inequalities we obtain

$$K_\ell^r(x) \leq \frac{1}{\ell^{2r}}\left(\frac{\ell\pi x(1 - (\ell\pi x)^2/10)}{\pi x(1 - (\pi x)^2/2)}\right)^{2r} = \left(\frac{1 - (\ell\pi x)^2/10}{1 - (\pi x)^2/2}\right)^{2r} \leq \left(\frac{(1 - (\ell\pi x)^2/10)}{(1 - (\pi x)^2/2)}\right)$$

where we have used the facts above, and we require $|x| \leq 1/\ell$ in order to invoke Fact 4.3. Finally we have

$$K_\ell^r(x) \leq 1 - (\ell\pi x)^2/10 + (\pi x)^2 \leq 1 - \ell^2 x^2/3$$

11

where we have used that $(1 - (\pi x)^2/2)^{-1} = 1 + (\pi x)^2/2 + ... \leq 1 + (\pi x)^2$ which holds for $x \leq 1/\pi$. Finally we can prove our lower bound

$$K_\ell^r(x) \geq \frac{1}{\ell^{2r}} \left(\frac{\ell\pi x(1 - (\ell\pi x)^2/2)}{\pi x(1 - (\pi x)^2/10)} \right)^{2r} = \left(\frac{1 - (\ell\pi x)^2/2}{1 - (\pi x)^2/10} \right)^{2r} \geq \left(1 - (\ell\pi x)^2/2 \right)^{2r} \geq 1 - r \left(\ell\pi x \right)^2$$

where again we have used the facts above, and we require that $|x| \leq 1/\pi$ in order to invoke Fact 4.3.   □

We remark that Candes and Fernandez-Granda [5] gave numerical evidence for particular values of m and k that super-resolution is difficult without a separation condition.

4.2   From Uncertainty Principles to Preconditioners

Let us contrast our approach here with the way in which we used majorants in the previous section: Recall our goal was to upper and lower bound the sum $\sum_{t=-m}^{m} \|f(t)\|^2$. Once we plugged in majorants and minorants, both the upper and lower bounds became simple expressions but the bounds themselves are not that close to each other unless m is much larger than $1/\Delta$ and this is roughly tight. Instead, we could ask: are there coefficients c_t so that the sum $\sum_{t=-m}^{m} c_t \|f(t)\|^2$ itself has a simple expression that we can get tighter upper and lower bounds for (see e.g. Claim 4.5).

The point is that many inequalities in harmonic analysis are established this way. One classic example is the Salem inequality (see e.g. [39]). Again let $f(t) = \sum_{j=1}^{k} u_j e^{i2\pi f_j t}$ and suppose that $|f_j - f_{j'}| \geq \Delta$ for all $j \neq j'$. Then

Lemma 4.4. *[39] Let E be any interval of length $(1 + \epsilon)/\Delta$. Then*

$$\sum_j \|a_j\|^2 \leq \frac{C_\epsilon}{|E|} \int_E \|f(t)\|^2 \, dt$$

where $C_\epsilon = \frac{2\pi(1+\epsilon)^2}{4\epsilon(2+\epsilon)} = O(1 + 1/\epsilon)$.

In fact the proof of this inequality follows by choosing an appropriate test function $\chi(t)$ that is supported on E, and analyzing

$$\int_{-\infty}^{\infty} \chi(t) \|f(t)\|^2 \, dt$$

directly and showing that it equals (up to scaling) $\sum_j \|u_j\|^2$ plus cross terms that decay as the separation increases. So when we use a single test function (as opposed a majorant and a minorant) it can really be thought of as a preconditioner that makes the right hand side close to the sum-of-squares of the coefficients!

We will pursue this direction, but will make use of different test functions than in [39] that are supported on a discrete set of points, namely the fourier transform of the Fejer kernel and its powers. This has the effect of transforming the right hand side into a sum instead of an integral.

Claim 4.5. *Let $F(x)$ be a kernel where $\widehat{F}(t) = \sum_{j=-m}^{m} c_j \delta(t - j)$. Then*

$$\sum_{j=-m}^{m} c_j \|f(j)\|^2 = \sum_{j=1}^{k} \sum_{j'=1}^{k} u_j \bar{u}_{j'} F(f_j - f_{j'})$$

Proof.

$$\sum_{j=-m}^{m} c_j \|f(j)\|^2 = \sum_{j=1}^{k} \sum_{j'=1}^{k} u_j \bar{u}_{j'} \int_{-\infty}^{\infty} \widehat{F}(t) e^{i2\pi(f_j - f_{j'})t} \, dt = \sum_{j=1}^{k} \sum_{j'=1}^{k} u_j \bar{u}_{j'} F(f_j - f_{j'})$$

□

Now we can set F to be the Fejer kernel or its powers. Using the above claim we will show that this choice of c_j's is a good preconditioner. Suppose that $d_w(f_j, f_{j'}) \geq \Delta$.

Lemma 4.6. *If c_j are the coefficients of $\widehat{K}_\ell$, then*

$$\left| \sum_{j=-\ell}^{\ell} c_j \|f(j)\|^2 - \sum_{j=1}^{k} \|u_j\|^2 \right| \leq \sum_{j=1}^{k} \|u_j\|^2 \frac{\pi^2}{24\ell^2\Delta^2}$$

Proof. We can use Claim 4.5 to obtain

$$\left| \sum_{j=-\ell}^{\ell} c_j \|f(j)\|^2 - \sum_{j=1}^{k} \|u_j\|^2 \right| \leq \sum_{j \neq j'} \|u_j\| \|u_{j'}\| |K_\ell(f_j - f_{j'})|$$

where we have used the fact that $K_\ell(0) = 1$ (recall that we chose a non-standard normalization for the Fejer kernel). Moreover rearranging we get:

$$\sum_{j \neq j'} \|u_j\| \|u_{j'}\| |K_\ell(f_j - f_{j'})| \leq \sum_{j \neq j'} \frac{1}{2}(\|u_j\|^2 + \|u_{j'}\|^2) |K_\ell(f_j - f_{j'})|$$

$$\leq \sum_{j=1}^{k} \|u_j\|^2 \sum_{j'=1}^{\infty} \frac{1}{4\ell^2(j')^2\Delta^2} = \sum_{j=1}^{k} \|u_j\|^2 \frac{\pi^2}{24\ell^2\Delta^2}$$

where have used Fact 3.4 in the last inequality. This implies the lemma. $\qquad\square$

Corollary 4.7. *If c'_j are the coefficients of $\widehat{K}_\ell^r$ then*

$$\left| \sum_{j=-r\ell}^{r\ell} c'_j \|f(j)\|^2 - \sum_{j=1}^{k} \|u_j\|^2 \right| \leq \sum_{j=1}^{k} \|u_j\|^2 \frac{\pi^2}{4^r \ell^{2r}\Delta^{2r}}$$

Proof. We can follow the same proof as above, but for powers of the Fejer kernel. The only difference is that we use fact that $K_\ell^r(\Delta) \leq \frac{1}{4^r \ell^{2r}\Delta^{2r}}$ and that $\sum_{j=1}^{\infty} \frac{1}{j^{2r}} \leq \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$. This implies the corollary. $\qquad\square$

4.3 Approximate Strong Convexity

In the previous subsection, we proved upper and lower bounds for $\sum_{j=-r\ell}^{r\ell} c'_j \|f(j)\|^2$, and eventually we want to use this potential function to guide our algorithm. Let us focus on the case where all the coefficients $a_j = 1$ for simplicity. What we want to prove are soundness and completeness results that

$$\sum_{j=-r\ell}^{r\ell} c'_j \|f(j) - e^{i2\pi jx}\|^2 \approx k - 1$$

if and only if x is sufficiently close to some f_j. So the challenge is that we cannot directly apply Lemma 4.6 or Corollary 4.7 because the set $\{x\} \cup \{f_j\}_j$ will not necessarily be well separated (and we hope it won't be).

The starting point is combining Claim 4.5 and Corollary 4.7 to obtain:

$$\sum_{j=-r\ell}^{r\ell} c'_j \|f(j) - e^{i2\pi jx}\|^2 = T - 2\sum_{j=1}^{k} K_\ell^r(x - f_j)$$

where $T = \sum K_\ell^r(f_j - f_{j'}) + K_\ell^r(0) \approx k + 1$ is a constant that does not depend on x, and we have used the fact that K_ℓ^r is symmetric about the origin. Let $\gamma = d_w(x, \{f_j\})$.

Lemma 4.8. *If $\gamma < \min(\Delta/2, 1/\ell)$ and $\ell \geq 4$ then*

$$1 - 12r\ell^2\gamma^2 - \frac{\pi^2}{4^r \ell^{2r}\Delta^{2r}} \leq \sum_{j=1}^{k} K_\ell^r(x - f_j) \leq 1 - \frac{\ell^2\gamma^2}{3} + \frac{\pi^2}{4^r \ell^{2r}\Delta^{2r}}$$

And if instead $\gamma \geq 1/\ell$ then

$$\sum_{j=1}^{k} K_\ell^r(x - f_j) \leq \frac{1}{4^r} + \frac{\pi^2}{4^r \ell^{2r}\Delta^{2r}}$$

13

Algorithm 2 ITERATIVEREFINEMENT, **Input:** Oracle F and k and constants C, c, ϵ
Output: $\{\widehat{f}_j\}$ with matching distance at most ϵ to $\{f_j\}$

---

1. Set $\delta_j = \frac{1}{4\ell}(\frac{c}{Cr})^j$ and $\gamma_j = (\frac{Cr}{c})^{1/2}\delta_j + \epsilon/2$

2. Set U to be the grid of $[0, 1)$ with length δ_1

3. For $i = 1$ to k

4.   Find z in U that minimizes $F(z)$; Set $\widehat{f}_i$ to z

5.   Remove all x' with $d_w(\widehat{f}_i, x') \leq 1/(2\ell)$

6. End

7. Set $j = 2$

8. While $\gamma_j > \epsilon$

9.   For $i = 1$ to k

10.   Let U_i be the grid of points x' with $d_w(\widehat{f}_i, x') \leq \gamma_{j-1}$ with length δ_j.

11.   Find z in U_i that minimizes $F(z)$; Set $\widehat{f}_i$ to z

12.   End

13. End

14. Output $\{\widehat{f}_i\}$

Proof. Let $f_1, f_2, ..., f_k$ be sorted so that $d_w(x, f_j)$ is monotonically decreasing. We can write

$$\sum_{j=1}^k K_\ell^r(x - f_j) = K_\ell^r(\gamma) + \sum_{j=2}^k K_\ell^r(x - f_j) \leq 1 - \frac{r\ell^2\gamma^2}{5} + \sum_{j=2}^k K_\ell^r(x - f_j)$$

where we have used Lemma 4.1 to give upper bounds for $K_\ell^r(\gamma)$. Moreover we have that and $d_w(x, f_j) \geq \Delta(j-1) - \gamma \geq \Delta(j-1)/2$ for all j and so we can use Fact 3.4 as we did in the proof of the modified Salem inequality (Lemma 4.6) and this implies the upper bound. An identical argument implies the lower bound too. The second part of the lemma follows by instead invoking Fact 3.4. $\square$

Notice that if we set $r = \Theta(\log n)$ and $\ell > \frac{1}{\Delta}$ then the terms that depend on Δ can be bounded by any inverse polynomial and we have proven that our potential function is additively close to being strongly convex. The intuition is that if we have any two solutions x and x' at distance γ and γ' respectively from f_1 if $\gamma < \frac{\gamma'}{C\ell}$ our potential function must be smaller for x than it is for x', and so we iteratively discard points that are too far from f_1 and make geometric progress in each step.

4.4   The Algorithm

Here we consider the following abstraction: Suppose there are unknown f_j's which are Δ separated according to the wrap-around metric and set $\ell = 1/\Delta$. Furthermore we are given an oracle F so that if $d_w(z, \{f_j\}) = \gamma$, then either

$$A + c\ell^2\gamma^2 - \epsilon^2/4 \leq F(z) \leq A + Cr\ell^2\gamma^2 + \epsilon^2/4 \tag{$*$}$$

if $\gamma \leq 1/\ell$ and otherwise $F(z) \leq 1/4$. Our goal is to find an approximation $\widehat{f}_j$ to the f_j's using this oracle, that have matching distance at most ϵ. This will immediately yield a fast, greedy algorithm for the super-resolution problem.

Let C, c be constants, then:

Theorem 4.9. *The output of* ITERATIVEREFINE *is a set* $\{\widehat{f}_i\}$ *that has matching distance at most* ϵ *to* $\{f_i\}$ *with respect to* d_w. *If R is the running time to call the oracle for $F(z)$, then* ITERATIVEREFINE *runs in time* $O(R\ell r + Rkr^{1/2}\log(1/\epsilon))$.

14

Proof. We will prove that IterativeRefinement succeeds by proving a set of invariants for the algorithm. We will refer to Steps 2-6 as the initialization phase.

Claim 4.10. *After initialization, the matching between $\{\widehat{f}_i\}$ and $\{f_i\}$ according to d_w is at most $\gamma_1 \le 1/(4\ell)$.*

Proof. Consider the first iteration where $i = 1$. Since the grid length is δ_1, there must be an x where $A + Cr\ell^2\delta_1^2 - \epsilon^2/4 \le F(x)$. Hence z must satisfy

$$A + Cr\ell^2\delta_1^2 - \epsilon^2/4 \le F(z) \le A + c\ell^2\gamma^2 + \epsilon^2/4$$

where $d_w(\{f_i\}, z) = \gamma$. This implies that $\gamma \le \gamma_1 = (\frac{Cr}{c})^{1/2}\delta_1 + \epsilon/2 \le \frac{1}{4\ell}$. Hence if f_1 is the closest to z according to d_w, we necessarily remove all points x' with $d_w(x', f_1) \le 1/(4\ell)$. At the start of the next iteration, there still remains a point z that is within δ_1 of some f_j (even after we delete points close to $\widehat{f}_1$) and so the next point z will be close to a new f_j. This establishes the claim by induction. $\qquad\square$

Claim 4.11. *After iteration j, the matching distance between $\{\widehat{f}_i\}$ and $\{f_i\}$ is at most γ_j.*

Proof. By induction, at the start of iteration j the matching distance between $\{\widehat{f}_j\}$ and $\{f_j\}$ is at most γ_{j-1}. Thus in iteration j, for each i there is a point that is within distance δ_j of f_i. Moreover the grids U_i are disjoint. Again using the properties of the oracle, this implies that for each i the z we find satisfies $d_w(f_i, z) = \gamma$ with

$$\gamma \le \left(\frac{Cr}{c}\right)^{1/2}\delta_{j-1} + \epsilon/2 = \gamma_j$$

and this completes the proof of the claim. $\qquad\square$

These invariants immediately yield our main theorem. $\qquad\square$

We are now ready to prove Theorem 1.5.

Proof. We set $\ell = 1/\Delta$ and $r = \Theta(\log 1/\epsilon)$. Recall that $m = r\ell$ is the cut-off frequency. Furthermore let

$$F(z) = \sum_{j=-r\ell}^{r\ell} c'_j \|f(j) - e^{i2\pi jz}\|^2$$

in which case $A = T - 2$ and using Lemma 4.8 we conclude that $F(z)$ has the desired form in $(*)$. Moreover we can evaluate $F(z)$ in time $O(m)$ by directly computing it. Finally it is easy to see that if there is noise in our measurements we can account for it expanding $F(z)$ as follows:

$$\left| \sum_{j=-r\ell}^{r\ell} c'_j \|f(j) - e^{i2\pi jz} + \eta_j\|^2 - F(z) \right| \le \sum_{j=-r\ell}^{r\ell} c'_j \left(2(k+1)\|\eta_j\| + \|\eta_j\|^2 \right)$$

Note that $\sum_j c'_j = 1$ because it is the convolution of discrete probability distributions. Hence if the noise satisfies $\|\eta_j\| \le \epsilon^2/(4k)$ for each j, we can group the error terms due to the noise into the terms in $(*)$ that depend on ϵ, and the theorem now follows by invoking Theorem 4.9. $\qquad\square$

Acknowledgements

References

[1] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

[2] A. Beurling. Interpolation for an interval on $\mathbb{R}^1$. (Mittag-Leffler Lectures on Harmonic Analysis) Collected Works of Arne Buerling: Volume 2, Harmonic Analysis, pages 351–359, 1989.

[3] B. Bhaskar, G. Tang and B. Recht. Compressed sensing off the grid. *arXiv:1207.6053*, 2012.

[4] A. Bhaskara, M. Charikar, A. Moitra and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *STOC*, 2014.

[5] E. Candes and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Comm. of Pure and Applied Math.*, to appear.

[6] E. Candes and C. Fernandez-Granda. Super-resolution from noisy data. *J. of Fourier Analysis and Applications*, pages 1229–1254, 2013.

[7] E. Candes, J. Romberg and T. Tao. Robust uncertainty principles: exat signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, pages 489–509, 2006.

[8] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51(12):4203–4215, 2005.

[9] V. Chandrasekaran, B. Recht, P. Parrilo and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, pages 805–849, 2012.

[10] E. Cheney. *Introduction to Approximation Theory*. American Mathematical Society, 2000.

[11] D. Donoho. Superresolution via sparsity constraints. *SIAM Journal on Mathematical Analysis*, pages 1309–1331, 1992.

[12] D. Donoho and B. Logan. Signal recovery and the large sieve inequality. *SIAM Journal of Applied Math*, pass 577–591, 1992.

[13] B. G. R. de Prony. Essay exprimental et analytique: sur les lois de la dilatabilit de fluides lastique et sur celles de la force expansive de la vapeur de l'alcool, a diffrentes tempratures. *Journal de l'cole Polytechnique*, pages 24–76, 1795.

[14] K. Do Ba, P. Indyk, E. Price and D. Woodruff. Lower bounds for sparse recovery. *SODA*, pages 1190–1197, 2010.

[15] M. Elad. *Sparse and Redundant Representations*. Springer, 2010.

[16] C. Fernandez-Granda. Support detection in super-resolution. *International Conference on Sampling Theory and Appl.*, pages 145–148, 2013.

[17] N. Goyal, S. Vempala and Y. Xiao. Fourier PCA. *STOC*, 2014, to appear.

[18] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[19] Y. Hua and T. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages 814–824, 1990.

[20] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, pages 553-562, 2010.

[21] D. Slepian and H. Pollack. Prolate spheroidal wave functions, fourier analysis and uncertainty – III *Bell System Technical Journal*, pages 1295–1336, 1962.

[22] A. Moitra and M. Saks. A polynomial time algorithm for lossy population recovery. In *FOCS*, pages 110–116, 2013.

[23] H. Montgomery. *Ten Lectures on the Interface Between Analytic Number Theory and Harmonic Analysis*. American Mathematical Society, 1994.

[24] A. Moitra and G. Valiant. Setting the polynomial learnability of mixtures of gaussians. In *FOCS*, pages 93–102, 2010.

[25] H. Montgomery and R. Vaughan. Hilbert's inequality. *Journal of the London Math. Society*, pages 43–81, 1974.

[26] S. Park, M. Park and M. Kang. Super-resolution image reconstruction: a technical overview. *Signal Processing Magazine*, pages 21–36, 2003.

[27] V. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomical Society*, pages 347–366, 1973.

[28] A. Selberg. *Collected Papers*. Springer-Verlag, 1991.

[29] D. Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty – IV *Bell System Technical Journal*, pages 3009–3058, 1964.

[30] D. Slepian. Prolate spheroidal wave functions, fourier analysis and uncertainty – V *Bell System Technical Journal*, pages 1371–1430, 1978.

[31] D. Slepian and H. Pollack. Prolate spheroidal wave functions, fourier analysis and uncertainty – I *Bell System Technical Journal*, pages 43–64, 1961.

[32] D. Slepian and H. Pollack. Prolate spheroidal wave functions, fourier analysis and uncertainty – II *Bell System Technical Journal*, pages 65–84, 1961.

[33] G. Stewart. Gershgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$. *Mathematics of Computation*, pages 600–606, 1975.

[34] G. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press Inc., 1990.

[35] P. Stoica. List of references on spectral line analysis. *Signal Processing*, pages 329–340, 1993.

[36] G. Tang , B. Bhaskar, and B. Recht. Sparse recovery over continuous dictionaries: Just discretize. *Asilomar Conference on Signals, Systems and Computers*, 2013.

[37] A. Timan. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press-Macmillan, 1963.

[38] J. Vaaler. Some extremal functions in fourier analysis. *Bulletin of the American Mathematical Soc.*, pages 183–216, 1985.

[39] A. Zygmund. *Trigonometric Series*. Cambridge University Press, 1968.