Algorithmsic Foundations for the Diffraction Limit

Sitan Chen*  
EECS, MIT  
USA  
sitanc@mit.edu

Ankur Moitra†  
Math and EECS, MIT  
USA  
moitra@mit.edu

ABSTRACT
For more than a century and a half it has been widely-believed (but was never rigorously shown) that the physics of diffraction imposes certain fundamental limits on the resolution of an optical system. However our understanding of what exactly can and cannot be resolved has never risen above heuristic arguments which, even worse, appear contradictory. In this work we remedy this gap by showing that there is a phase transition where the sample complexity goes from polynomial to exponential.

INTRODUCTION
In this work we rigorously prove upper and lower bounds on the statistical inverse problem and, studying the diffraction limit as a statistical inverse problem and, observe that there is a phase transition where the sample complexity goes from polynomial to exponential. Surprisingly, we show that this does not occur at the Abbe limit, which has long been presumed to be the true diffraction limit.

CCS CONCEPTS
• Mathematics of computing → Probability and statistics; • Theory of computation → Unsupervised learning and clustering; • Applied computing → Physics.

KEYWORDS
Mixture models, Fourier optics, matrix pencil method, tensor decomposition, extremal functions

REFERENCE
Sitan Chen and Ankur Moitra. 2021. Algorithmic Foundations for the Diffraction Limit. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC ’21), June 21–25, 2021, Virtual, Italy. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3406325.3451078

1 INTRODUCTION
For more than a century and a half it has been widely believed (but was never rigorously shown) that the physics of diffraction imposes fundamental limits on the resolution of an optical system. In the standard physical setup, we observe incoherent illumination from far-away point sources through a perfectly circular aperture. Each point source produces a two-dimensional image, computed explicitly by Sir George Biddell Airy in 1835 [3] and now called an Airy disk. For a point source of light whose angular displacement from the optical axis is \( \mu \in \mathbb{R}^2 \), the normalized intensity at a point \( x \) on the observation plane is given by

\[
I(x) = \frac{1}{\pi \sigma^2} \left( \frac{2 J_1(||x - \mu||)/2/\sigma}{||x - \mu||/2/\sigma} \right)^2
\]

where \( J_1 \) is a Bessel function of the first kind. Under Feynman’s path integral formalism, \( I(x) \) is precisely the pdf of the distribution over where the photon is detected. The physical properties of the optical system, namely its numerical aperture and the wavelength of light being observed, determine \( \sigma \) which governs the amount by which each point source gets blurred.

Intuitively, when point sources are closer together it seems harder to resolve them. However, despite considerable interest over the years [1, 7, 28, 42, 44, 45], our understanding of what exactly can and cannot be resolved has never risen above heuristic arguments. For example, consider the Sparrow limit which is defined as the critical separation at which the average of two Airy disks goes from being bimodal to unimodal [45]. Sparrow argued that it is not possible to distinguish between one Airy disk and a mixture of two Airy disks when this happens. We refer to the appendix of the full version for some historical quotes by Sparrow and many other noted physicists, many of which actually contradict one another!

Our goal in this work is to remedy this gap in the literature, and draw new connections to recent work in theoretical computer science on provably learning mixture models, as we will describe next. First we remark that the way the diffraction limit is traditionally studied is in fact a mixture model. In particular we assume that, experimentally, we can measure photons that are sampled from the true diffraction image. However we only observe a finite number of them because our experiment has finite exposure time, and indeed as we will see in some settings the number of samples needed to resolve closely-spaced objects can explode and be essentially impossible just from statistical considerations. Moreover we may only be able to record the location of observed photons up to some finite accuracy, which can also be thought of as being related to sampling error. The main question we will be interested in is:

How many samples (i.e. photons) suffice to accurately estimate the centers and relative intensities of a mixture (i.e. superposition) of two or more Airy disks, as a function of their separation and the parameters of the optical system?

This is a central question in optics. Fortunately, there are many parallels between this question and that of provably learning mixture models that surprisingly seem to have gone undiscovered. In
particular, let us revisit Sparrow’s argument that resolution is impossible when the density function becomes unimodal. In fact there are already counter-examples to this claim, albeit not for mixtures of Airy disks. It is known that there are algorithms for learning mixtures of two Gaussians that take a polynomial number of samples and run in polynomial time. These algorithms work even when the density function is unimodal, and require just that the overlap between the components is bounded away from one. Moreover when there are \( k \) components it is known that there is a critical separation above which it is possible to learn the parameters accurately with polynomially many samples, and below which accurate learning requires superpolynomially many information-theoretically [43]. Thus a natural way to formulate what the diffraction limit is, so that it can be studied rigorously, is to ask:

**At what critical separation does the sample complexity of learning mixtures of \( k \) Airy disks go from polynomial to exponential?**

In this work we will give algorithms whose running time and sample complexity are polynomial in \( k \) above some critical separation, and prove that below some other critical separation the sample complexity is necessarily exponential in \( k \). These bounds will be within a universal constant, and thus we approximately locate the true diffraction limit. There will also be some surprises along the way, such as the fact that the Abbe limit, which has long been postulated to be the true diffraction limit, is not actually the correct answer!

Before we proceed, we also want to emphasize that there is an important conceptual message in our work. First, for mixtures of Gaussians the model was only ever supposed to be an approximation to the true data generating process. For example, Karl Pearson introduced mixtures of Gaussians in order to model various physical measurements of the Naples crabs. However mixtures of Gaussians always have some chance of producing samples with negative values, but Naples crabs certainly do not have negative forehead lengths! In contrast, for mixtures of Airy disks the model is an extremely accurate approximation to the observations in many experimental setups because it comes from first principles. It is particularly accurate in astronomy where for all intents and purposes the lens is spherical and the star is so far away that it is a point source, and the question itself is highly relevant because it arises when we want to locate double-stars [16].

Furthermore we believe that there ought to be many more examples of inverse problems in science and engineering where tools and ideas from the literature on provably learning mixture models ought to be useful. Indeed both mixtures of Gaussians and mixtures of Airy disks can be thought of as inverse problems with respect to simple differential equations, for the heat equation and a modified Bessel equation respectively. While this is a well-studied topic in applied mathematics, usually one makes some sort of smoothness assumption on the initial data. What is crucial to both the literature on learning mixtures of Gaussians and our work is that we have a parametric assumption that there are few components. Thus we ask: Are there provable algorithms for other inverse problems, coming from differential equations, under parametric assumptions? Even better: Could techniques inspired by the method of moments play a key role in such a development?

### 1.1 Overview of Results

It is often the case that heuristic arguments, despite being quite far from a rigorous proof, predict the correct thresholds for a wide range of statistical problems. However here there will be a surprise. In a seminal work in 1873, Ernest Abbe formulated what is now called the Abbe limit. Since then it has been widely accepted in the optics literature as the critical distance below which diffraction makes resolution impossible for classical optical systems. In the mixture model formalism outlined above, it corresponds to a separation of \( \pi \sigma \) between any pair of Airy disk centers \( \mu_1, \mu_2 \). This distance arises naturally because it corresponds to the radius of the support of the Fourier transform of the Airy disk kernel

\[
A_\sigma : x \mapsto \frac{1}{\pi \sigma} \left( \frac{\sin(\sigma x)}{\sigma x} \right)^2.
\]

One of the main results of this work is to show that resolution is statistically hard even above the Abbe limit! Specifically, we show that even for mixtures of Airy disks whose centers have a pairwise separation that is a constant factor larger than the Abbe limit, the problem of recovering their locations can require \( \exp(\Omega(\sqrt{\log k})) \) samples. The main challenge is that no configuration where the Airy disk centers are all on the same line can beat the Abbe limit. Instead we construct a new, natural lower bound instance.

**Theorem 1.1.** Let \( \gamma \doteq \sqrt{4/3} \approx 1.55. \) For any \( 0 < \epsilon < 1, \) there exist two superpositions of \( k \) Airy disks \( \rho, \rho' \) which are both \( \gamma \cdot (1 - \epsilon) \cdot \pi \sigma \)-separated and such that 1) \( \rho \) and \( \rho' \) have noticably different sets of centers, and yet 2) it would take at least \( \exp(\Omega(e \sqrt{\pi \sigma})) \) samples to distinguish whether the samples came from \( \rho \) or from \( \rho' \).

On the other hand, we also show that when the Airy disks have separation that is a small constant factor larger than this critical distance, there is an algorithm for recovering the centers that takes a polynomial number of samples and runs in polynomial time.

**Theorem 1.2 (see Theorem 4.2).** Define the absolute constant \( \overline{\gamma} \doteq \frac{2}{\pi} \approx 0.6366197724 \) as the first positive zero of the Bessel function \( J_0 \). Let \( \rho \) be a \( \overline{\gamma} \cdot \pi \sigma \)-separated superposition of \( k \) Airy disks where every disk has relative intensity at least \( \lambda \). Then for any target error \( \epsilon < 0 \), there is an algorithm with time and sample complexity \( N = \text{poly}(k, 1/\triangle, 1/\lambda, 1/\epsilon) \) which outputs an estimate for the centers and relative intensities of \( \rho \) which incurs error \( \epsilon \) with probability at least \( 9/10 \). Furthermore, this holds even when there is granularity in the photon detector, as long as it is at most some inverse polynomial in all parameters.

The main open question of our work is to prove matching upper and lower bounds that pin down the true diffraction limit. However, as we will discuss, this is a challenging problem in harmonic analysis, despite being connected to areas where there has been considerable recent progress. Moreover this phase transition for resolution is actually more dramatic than what happens for mixtures of Gaussians [43]. Even ignoring the issue of computational complexity, for spherical Gaussian mixtures it is known that at separation of \( \sqrt{\log k} \), super-polynomially many samples are needed, while at separation \( \Omega(\sqrt{\log k}) \), polynomially many suffice.

We now say a word about the techniques that go into proving Theorem 1.1 and Theorem 1.2. It turns out that both are closely related to proving a modified version of an Ingham-type estimate [35]:
Question 1.1. What is the smallest \( \Delta \) for which the quantity
\[
\int_{B} \left| \sum_{j=1}^{k} \lambda_j e^{-2\pi i (\mu_j, \omega)} \right|^2 \, d\omega \geq \frac{1}{\text{poly}(k)} \|\lambda\|_2^2
\]
for all vectors \( \lambda \in \mathbb{R}^k \) and all sets of centers \( \{\mu_j\} \) for which \( \|\mu_j - \mu_j'\|_2 > \Delta \) for all \( j \neq j' \), where the integration is over the origin-centered ball \( B \subset \mathbb{R}^2 \) of radius 1/2?

In particular, the main technical step for showing Theorem 1.2 is to show that the critical \( \Delta \) in Question 1.1 is at most 2\( \sqrt{\lambda} \). This can be obtained via the following extremal function. A ball minorant is, in general, a function \( F \) satisfying the properties that
\[
(1) \quad F(x) \leq 1 \{x \in B\} \quad \text{and} \quad (2) \quad \hat{F} \text{ is supported on the ball of radius } \Delta
\]

In [10, 23, 26] it was shown that such a ball minorant exists for \( \Delta = 2\sqrt{\lambda} \) (interestingly, this paved the way to some recent progress on Montgomery’s famous pair correlation conjecture for the Riemann zeta function [10]). One can use property (1) to pass from integrating against the function \( 1 \{x \in B\} \) to integrating against \( F \). And because by property (2) \( F \) is localized in the frequency domain, the latter integral is large. In fact the one-dimensional analogue of Question 1.1 was resolved in [39] using the univariate analogue of \( F \), namely the Beurling-Selberg minorant. However the algorithmic approach only made sense in one-dimension. In our case, we employ the tensor generalization of the matrix pencil method, originally introduced in [30]. We defer the details of this to Section 4.3.

For Theorem 1.1, we need to answer a variant of Question 1.1.

Question 1.2. What is the smallest \( \Delta \) for which
\[
\int_{B} \left| \sum_{j=1}^{k} \lambda_j e^{-2\pi i (\mu_j, \omega)} \right|^2 \, d\omega \geq \frac{1}{\text{poly}(k)} \quad (1)
\]
for all \( \lambda, \lambda' \in \mathbb{R}^k \) and \( \{\mu_j\}, \{\mu_j'\} \subset \mathbb{R}^2 \) for which \( \|\mu_j - \mu_j'\|_2 > \Delta \) and \( \|\mu_j - \mu_j'\|_2 > \Delta \) for \( j \neq j' \), where \( B \) is the unit ball?

The connection to Theorem 1.1 is standard: by Plancherel’s and smoothness of \( \mathcal{A}_\sigma \), one can upper bound the \( L_1 \) distance between the mixture of Airy disks given by parameters \( \{\lambda_j\}, \{\mu_j\} \) and the mixture given by \( \{\lambda'_j\}, \{\mu'_j\} \) in terms of the left-hand side of \( (1) \). So if one can construct a set of \( \Delta \)-separated centers \( \{\mu_j\}, \{\mu'_j\} \) for which \( (1) \) fails to hold but for which the collection \( \{\mu_j\} \) is separated from \( \{\mu'_j\} \) but the resulting mixtures of Airy disks are \( o(1/\text{poly}(k)) \)-close in total variation distance it implies that resolution is statistically impossible with polynomially many samples. This is the recipe used in known lower bounds [24, 40, 43] for learning Gaussian mixtures.

For our purposes, it turns out that “tensoring” one-dimensional lower bounds does not work because it would not beat the Abbe limit [39]. Morally, this is because tensoring the unit interval with itself would give us the unit square, which corresponds to separation in the \( L_\infty \) distance rather than the \( L_2 \) distance, and the \( L_2 \) distance is the right distance in optics because it is rotationally invariant. The main technical contribution in our lower bound is to give a more sophisticated construction given by interleaving two triangular lattices and placing the centers at points on these lattices (see Figure 3). The analysis is rather delicate, and we defer the details to Section 2 and the full version.

To complete the picture, we show that there is no diffraction limit when the number of Airy disks is a constant. In particular we show that for any constant number of Airy disks there is an algorithm that takes a polynomial number of samples and runs in polynomial time that learns the parameters to any desired accuracy regardless of the separation.

Theorem 1.3 (see Theorem 4.1). Let \( \rho \) be a \( \Delta \)-separated superposition of \( k \) Airy disks where every disk has relative intensity at least \( \lambda \). Then for any \( \epsilon, \delta > 0 \), there is an algorithm which draws \( N = \text{poly} \left( (k\sigma/\Delta)^k, 1/\lambda, 1/\epsilon, \log(1/\delta) \right) \) samples from \( \rho \), runs in time \( O(N) \), and outputs an estimate for the centers and relative intensities of \( \rho \) which incurs error \( \epsilon \) with probability at least \( 1 - \delta \). Furthermore, this holds even when there is granularity in the photon detector, as long as it is at most some inverse polynomial in \( N \).

This result turns out to be simple in retrospect, and comes from assembling a few standard tools from the literature on provably learning mixture models. Nevertheless it underscores an important point that existing tools can already have important implications for inverse problems in the sciences. We first estimate the Fourier transform \( \hat{\rho} \) from samples and then pointwise divide by \( \hat{\Delta}_\sigma \), thereby simulating noise access to the Fourier transform of the mixture of delta functions at \( \mu_1, \ldots, \mu_k \). However \( \hat{\Delta}_\sigma \) has compact support, so we can only access frequencies with bounded \( L_2 \) norm. Now we can reduce to the one-dimensional case [39] by projecting \( \rho \) along two nearby directions, solving each resulting univariate problem, and then solving an appropriate linear system to recover the centers and relative intensities. This method is reminiscent of [33, 40], which gives algorithms for learning high-dimensional mixtures of Gaussians based on reducing to a series of one-dimensional problems and stitching together these estimates carefully.

1.2 Related Work

We have already mentioned that our work is closely related to the vast literature on learning mixture models and, in particular, on learning mixtures of Gaussians [2, 4–6, 11–13, 19, 24, 27, 29, 33, 36, 40, 43, 48]. Here we mention some other connections to work on recovering spike trains from noisy, band-limited Fourier measurements.

Superresolution. The seminal work of [14, 15] was one of the first to put this question on rigorous footing. Donoho studied the modulus of continuity for this problem on a grid as the grid width goes to zero. Later Candes and Fernandez-Granda [9] gave a practical algorithm based on \( L_1 \) minimization over a continuous domain. There has been a long line of work on this problem which it would also be impossible to survey fully, so we refer the reader to [8, 17, 18, 37–39, 41, 44] and references therein.

We remark that essentially all works on super-resolution in high dimensions focus on the case where measurements are \( L_\infty \)-band-limited rather than \( L_2 \)-band-limited. Given the prevalence of Airy disks and circular apertures in statistical optics, one upshot of our work is that, technical issues related to the so-called box (aka \( L_\infty \) ball) minorant problem notwithstanding, the \( L_2 \) setting may be the more practically relevant one to consider anyways.
Sparse Fourier Transform. There are also connections to the extensive literature on the sparse Fourier transform, which can be interpreted in some sense as the “agnostic” version of the super-resolution problem where the goal is to compete with the error of the best \( k \)-sparse approximation to the discrete Fourier transform, even in the presence of noise, using few measurements \([20–22, 25, 31, 34]\). When the \( k \) spikes need not be at discrete locations and the low-frequency measurements are randomly chosen, this is the problem of compressed sensing off the grid introduced by [46], for which recovery is possible with far fewer measurements. This can be thought of as the one-dimensional case of the setting of [30]. To our knowledge, the only work that addresses the continuous, high-dimensional version of the sparse Fourier transform is the very recent work of [32]. The emphasis in this literature is primarily on obtaining sample complexity near-linear in \( k \), whereas our guarantees are only polynomial in \( k \). Consequently the results in the sparse Fourier transform literature lose log factors in the level of separation they require, whereas in our setting the emphasis is primarily on the level of separation needed to get polynomial-time and -sample algorithms.

### 1.3 Visualizing the Diffraction Limit

In this section we present a striking visual representation of the statistical barrier imposed by the diffraction limit when the number of components is large. Recall that the upshot of Theorems 1.1 and 1.2 is that \( k \) plays a leading role in determining when resolution is and is not feasible: slightly above the Abbe limit, the sample (and computational) complexity is polynomial in \( k \), and anywhere beneath the Abbe limit, the sample complexity becomes exponential in \( k \). This helps clarify why in some domains like astronomy, where there are only ever a few tightly spaced point sources, there is evidently no diffraction limit. Yet in other domains like microscopy where there are a large number of tightly spaced objects, the diffraction limit is indeed a fundamental barrier, at least in the classical physical setup. This helps explain why different communities have settled on different beliefs about whether or not there is a diffraction limit.

In Figure 1 we experimentally investigate this phenomenon and illustrate how the total variation distance scales as we vary the number of disks and the separation in our earlier constructions. It is evident from this plot that for any superposition of a few Airy disks, there is no sharp dividing line between what is and is not possible to resolve. But when the number of Airy disks becomes large, with any reasonable number of samples, it is feasible to resolve the superposition if and only if their separation is at least as large as the Abbe limit.

We emphasize that in the instance constructed for Figure 1 (as well as the instance we construct and analyze for Theorem 1.1), the centers are plotted on a line. For such instances, by projecting in the direction of the line and using our deconvolution techniques, one can actually reduce to the problem of one-dimensional super-resolution, for which polynomial-time algorithms exist for any separation strictly greater than the diffraction limit [39], and by adapting the lower bound in [39] to this specific instance, one can see that this is tight. In contrast, if the centers can be placed anywhere in \( \mathbb{R}^2 \), there is a constant factor gap (\( \sqrt{4/3} \) versus \( \frac{2\pi}{\sqrt{3}} \)) between the lower bound in Theorem 1.1 and the upper bound in Theorem 1.2.

### 1.4 Roadmap

In Section 2 we give a preview of our lower bound proof by providing a self-contained answer to Question 1.2, deferring the details to the full version. In Section 3, we give an overview of our probabilistic model, some notation, and other mathematical preliminaries. In Section 4, we prove the algorithmic results in Theorems 1.3 and 1.2. In Section 5 we conclude with some directions for future work. We refer the reader to the full version of this work, available at https://arxiv.org/abs/2004.07659, for additional technical details and an exposition on the historical and scientific aspects of diffraction limits, written with a theoretical computer science audience in mind.

### 2 LOWER BOUND PREVIEW

In this section we give a self-contained proof of one of the main technical ingredients in the proof of our main result, Theorem 1.1. It will be convenient to introduce a bit of notation; any outstanding notation we will present Section 3, e.g. our convention for the Fourier transform. Recalling that \( \gamma \pm \sqrt{4/3} \), define

\[
m \pm \frac{2}{(1-\epsilon)\gamma\pi\sigma}
\]

(2)

for any small constant \( \epsilon > 0 \) so that the critical level of separation for which Theorem 1.1 applies is \( \Delta \pm 2/m = \gamma \cdot (1-\epsilon) \cdot \pi \sigma \).\(^1\)

Additionally, let \( k \) be an odd square and define

\[
\nu_{j_1,j_2} = \frac{\Delta}{2} \cdot (j_1, \sqrt{3} \cdot j_2), \quad j_1, j_2 \in \mathcal{J} \pm \frac{\sqrt{k} - 1}{2}, \ldots, \frac{\sqrt{k} - 1}{2}
\]

This construction is illustrated in Figure 3: there, similarly colored points correspond to centers in the same mixture, and our choice of

\(^1\pi \sigma \) is not important to this section but is the natural choice of scaling for Airy disks and will be convenient when we apply the results of this section to prove Theorem 1.1.
\{v_{j_1,j_2}\} ensures that the level of separation between any two points in a particular mixture is \(\Lambda\), which is slightly less than \(\sqrt[3]{4}/3\) times the Abbe limit of \(\sigma\). As such, the following tells us that the answer to Question 1.2 is surprisingly at least \(\sqrt[3]{4}/3\), rather than 1 as the Abbe limit would suggest:

**Lemma 2.1.** There exists a vector \(u \perp (u_{j_1,j_2})_{j_1,j_2} \in \mathbb{R}^k\) for which

\[
\sum_{j_1,j_2 \in J} u_{j_1,j_2} e^{-2\pi i(v_{j_1,j_2}, x)} \leq \exp(-\Omega(\epsilon \sqrt{k}))
\]

for all \(\|x\| \leq 1/\pi \sigma\). Furthermore, \(\text{sgn}(u_{j_1,j_2}) = (-1)^{j_1+j_2}\), and

\[
\sum_{j_1+j_2 \text{ even}} |u_{j_1,j_2}| = \sum_{j_1+j_2 \text{ odd}} |u_{j_1,j_2}| = 1.
\]

We need the following ingredient from the proof of the one-dimensional lower bound in [39].

**Definition 2.1.** The Fejer kernel is given by

\[
K_t(x) = \frac{1}{2\pi} \sum_{j=-t}^{t} (t - |j|) \cos(jx) = \frac{1}{2\pi} \left( \sin t \pi x \right)^2.
\]

We will denote the \(r\)-th power of \(K_t(\cdot)\) by \(K_t^r(\cdot)\).

**Fact 1.** \(K_t\) is even and periodic with period 1. For \(x \in [\pi/2, 0, 1/2]\), \(K_t(x) \leq 1/4\pi^2\).

**Proof.** That \(K_t\) is even and periodic follow from the second definition of \(K_t\) in (4). For the bound on \(K_t\), we can use the elementary bounds \(\sin \pi x \geq 2\pi x\) for \(x \in [0, 1/2]\) and \((\sin \pi x)^2 \leq 1\).

**Proof of Lemma 2.1.** Let \(t = 4/\epsilon\) and \(r = (\sqrt{k} - 1)/2(1 - \epsilon) = \Theta(\epsilon \sqrt{k})\), and assume without loss of generality that \(t\) is even. Consider the function \(H : \mathbb{R}^2 \to \mathbb{R}\) given by

\[
H(x_1, x_2) = K_t(x_1 - 1/2) - K_t(x_2 \sqrt{3} - 1/2).
\]

We know that \(K_t[t] = \frac{1}{\pi} \sum_{j=-t}^{t} (t - |j|) \delta(t - j)\), so it follows that \(K_t^r[t] = \sum_{j_1 \neq j_2} \chi_{j_1} \chi_{j_2} \delta(t - j)\) for nonnegative \(\alpha_j\) which sum to 1. Assuming without loss of generality suppose that \(m\) defined by (2) is an odd integer, we conclude that for \(t = (t_1, t_2) \in \mathbb{C}^2\),

\[
\tilde{H}[t] = \sum_{j_1,j_2 \in J} \hat{h}_{j_1,j_2} e^{-\pi i m(t_1 + t_2/\sqrt{3})} \delta(t_1 - j_1) \cdot \delta(t_2/\sqrt{3} - j_2)
\]

where \(\hat{h}_{j_1,j_2} = \alpha_{j_1} \alpha_{j_2} \geq 0\), where \(\alpha_j = \chi_{j_1} \chi_{j_2} \cdot \mathbb{I}_{\{j = 0\}} + \alpha_{j_1} \cdot \mathbb{I}_{\{j \neq 0\}}\).

We will take

\[
\forall_{j_1,j_2 \in J} \hat{h}_{j_1,j_2} = \alpha_{j_1} \alpha_{j_2} \cdot \mathbb{I}_{\{j \neq 0\}}
\]

Observe that \(\text{sgn}(u_{j_1,j_2}) = (-1)^{j_1+j_2}\) as desired.

By taking the inverse Fourier transform of \(\tilde{H}\), we get that

\[
H(x_1, x_2) = \sum_{j_1,j_2} u_{j_1,j_2} e^{2\pi i(v_{j_1,j_2}, x)}.
\]

Figure 2: The squares correspond to periods of \(K_t^r\), while the ellipses have major and minor axes of length \(\gamma(1-\epsilon)\) and \(2(1-\epsilon)\). The figure is centered around the origin, and the bottom-left ellipse \(K\) is the set of points \(\left(\frac{2m}{m} - \frac{1}{2}, \frac{2m \sqrt{3}}{m} - \frac{1}{2}\right)\) as \((x_1, x_2)\) ranges over the origin-centered \(L_2\) ball of norm \(1/\pi \sigma\). By appropriately translating the four quadrants of this ellipse by distances in \(\mathbb{Z}^2\), we obtain overlapping regions whose union is given by \(R \cup S\), where \(R = [-1/2, 1/2] \times [-1/2, 1/2]\) is given by the central square (green) and \(S\) is the multi-colored set in the middle given by translates of the four connected components of \([-1, 0] \times [-1, 0]\). To complete our proof, it therefore suffices to show that \(H(x) \leq \exp(-\Omega(\epsilon \sqrt{k}))\) for all \(\|x\| \leq 1/\pi \sigma\).

Let \(R \subseteq \mathbb{R}^2\) denote the region \([-1/2, 1/2] \times [-1/2, 1/2]\). But as \(x\) ranges over the \(L_2\) ball of norm \(1/\pi \sigma\), \(\left(\frac{2m}{m} - \frac{1}{2}, \frac{2m \sqrt{3}}{m} - \frac{1}{2}\right)\) ranges over the interior of the ellipse centered at \((-1/2, 1/2)\) with axes of length \(\gamma(1-\epsilon)\) and \(2(1-\epsilon)\). For the subsequent discussion in this paragraph, we refer the reader to Figure 2. By periodicity of \(K_t^r\), the image of this ellipse under \(K_t^r\) is identical to the region \(R \subseteq \mathbb{R}^2\), where \(S\) is defined as follows. Denote the interior of the ellipse by \(B_1\), and denote its translates along the vectors \((0, 1), (1, 0), (1, 1)\) by \(B_2, B_3, B_4\). Define \(S\) to be the set of points in \(R\) that belong to none of \(B_1, B_2, B_3, B_4\).

We claim that \(S\) contains the origin-centered \(L_\infty\) ball of radius \(\epsilon/2\sqrt{2}\). Note that \(S\) is given by translating the four connected components of \([-1, 0] \times [-1, 0]\) \(B_1\), which is nonempty because \(B_1\) consists of points \((x_1, x_2)\) satisfying

\[
\frac{4}{\pi^2(1-\epsilon)^2} (x_1 - 1/2)^2 + \frac{1}{(1-\epsilon)^2} (x_2 - 1/2)^2 \leq 1.
\]

In particular, for \(x_1, x_2 \in [-1, 0]\) satisfying \(|x_1 - 1/2|, |x_2 - 1/2| > (1-\epsilon)/2\), observe that the left-hand quantity in (6) is at least

\[
\left(\frac{4}{\pi^2(1-\epsilon)^2} + \frac{1}{(1-\epsilon)^2}\right) \frac{(1-\epsilon)^2}{4} = 1,
\]
It will be straightforward to extend the above model to take into account error stemming from the fact that the photon detector itself only has finite precision.

**Definition 3.2 (Discretization Error).** Given $\zeta > 0$, we say $x$ is a $\zeta$-granular sample from $\rho$ if it is produced via the following generative process: 1) a point $x'$ is sampled from $\rho$, 2) $x$ is obtained by moving $x'$ an arbitrary distance of at most $\zeta$.

**Fourier Transform.** We will use the following convention in defining the Fourier transform. Given $f \in L^2(\mathbb{R}^d)$,

$$\hat{f}(\omega) \triangleq \int_{\mathbb{R}^d} f(x) \cdot e^{-2\pi i \omega \langle x, x \rangle} \, dx. \quad (7)$$

**Optical Transfer Function.** The following is standard.

**Fact 2.** $A_{\sigma} [\omega] = \frac{\pi}{2} (\arccos(\pi |\sigma| |\omega|) - \pi |\sigma| |\omega|) \sqrt{1 - \pi^2 |\sigma|^2 |\omega|^2}$.

**Proof.** It is enough to consider $\sigma = 1$. Let $G(x) \triangleq J_1(|x|)/|x|$. It is a standard fact that the zeroth-order Hankel transform of the function $r \mapsto J_1(r)$ is the indicator function of the interval $[0, 1]$. Using our convention for the Fourier transform (see (7)), this implies that $\hat{G}(\omega) = 2\pi \cdot 1(|\omega| \in [0, 1/2\pi])$. Because $A_1 = G^2/\pi$, by the convolution theorem we conclude that $A_{1/2} \triangleq \frac{1}{2\pi} \times \hat{G}$ is the convolution of $G$ with itself, which is just $4\pi^2$ times the convolution of the indicator function of the unit disk of radius $1/2\pi$ with itself. By elementary Euclidean geometry one can compute this latter function to be $\omega \mapsto \frac{1}{2\pi} \cdot (\arccos(\pi ||\omega||) - \pi ||\omega|| \sqrt{1 - \pi^2 ||\omega||^2})$, from which the claim follows.

In optics, the two-dimensional Fourier transform of the point-spread function is called the optical transfer function, a term we will occasionally use in the sequel.

Now note that by Fact 2, $\hat{A}_{\sigma}$ is supported only over the disk of radius $\frac{1}{\sqrt{\pi}}$ centered at the origin in the frequency domain. In the spatial domain, this corresponds to a separation of $\pi \sigma$; this is the definition of the Abbe limit. We will need the following elementary estimate for $\hat{A}_{\sigma}$:

**Fact 3.** For all $||\omega||_2 \leq 1$, $\hat{A}_{\sigma}[\omega] \geq (1 - ||\omega||_2)^2$.

**Scaling.** As the algorithms we give will be scale-invariant, we assume that $\sigma = 1/\pi$ in the rest of this work and refer to $A_{1/\pi}$ as $A$.

**Parameter Estimation Accuracy.** The following terminology formalizes what it means for an algorithm to return an accurate estimate for the parameters of a superposition of Airy disks.

**Definition 3.3.** $\left(\hat{\lambda}^i_1, \hat{\mu}^i_1\right)_{i \in [k]}$ is an $(\epsilon_1, \epsilon_2)$-accurate estimate for the parameters of a superposition of $k$ Airy disks $\rho$ with centers $\{\mu^i\}$ and relative intensities $\{\lambda^i\}$ if there exists a permutation $\tau$ for which

$$||\hat{\mu}^i - \tilde{\mu}^\tau(i)||_2 \leq \epsilon_1 \quad \text{and} \quad |\lambda^i - \tilde{\lambda}^\tau(i)| \leq \epsilon_2 \quad \text{for all } i \in [k].$$

**Generalized Eigenvalue Problems.** Given matrices $M, N$, we will denote by $(M, N)$ the generalized eigenvalue problem $Mx = \lambda Nx$. In any solution $(\lambda, x)$ to this, $\lambda$ is called a generalized eigenvalue and $x$ is called a generalized eigenvector.
Matrices, Tensors, and Flattenings. Given a matrix $M \in \mathbb{C}^{a \times b}$, we will denote its $i$-th row vector by $M_i$, its $j$-th column vector by $M^j$, and its $(i, j)$-th entry by $M_{i,j}$.

Given a tensor $T \in \mathbb{C}^{m_1 \times m_2 \times m_3}$ and matrices $M_1 \in \mathbb{C}^{m_1 \times m'_2}$, $M_2 \in \mathbb{C}^{m_2 \times m'_3}$, and $M_3 \in \mathbb{C}^{m_3 \times m'_1}$, define the flattened tensor $T(M_1, M_2, M_3) \in \mathbb{C}^{m_1 \times m_2 \times m_3}$ to have $(i_1, i_2, i_3)$-th entry

$$T_{m_1,m_2,m_3} = (M_1)_{i_1,i_2} (M_2)_{j_2,i_3} (M_3)_{j_3,i_3}.$$

Miscellaneous Notation. Let $\mathbb{S}_d$ denote the Euclidean unit sphere. Given $r > 0$, let $B_d(r)$ denote the Euclidean ball of radius $r$ centered at the origin in $\mathbb{R}^d$.

4 LEARNING SUPERPOSITIONS OF AIRY DISKS

In this section we present the technical details of our algorithmic results. In Sections 4.2 and 4.4, we prove the following formal version of Theorem 1.3.

**Theorem 4.1.** Let $\rho$ be a $\Delta$-separated superposition of $k$ Airy disks with minimum mixing weight $\lambda_{\text{min}}$ and such that $\|\mu_i\| \leq R$ for all $i \in [k]$. For any $\epsilon_1, \epsilon_2 > 0$, there is some $\alpha = \text{poly}(\log(1/\delta), 1/\lambda_{\text{min}}, 1/\epsilon_1, 1/\epsilon_2, R, (\sigma/\Delta)^2)$ for which there exists an algorithm with time and sample complexity $\text{poly}(1/\alpha)$ which, given $\gamma = \text{poly}(\alpha)$-granular sample access to $\rho$, outputs an $(\epsilon_1, \epsilon_2)$-accurate estimate for the parameters of $\rho$ with probability at least $1 - \delta$.

Specifically, in Section 4.2, we show how one can use the matrix pencil method to recover the parameters for $\rho$ given oracle access to the optical transfer function, i.e. the two-dimensional Fourier transform of $\rho$, up to some small additive error. In Section 4.4, we show how to implement this approximate oracle.

In Section 4.3, we also use the oracle of Section 4.4 to prove the following formal version of Theorem 1.2.

**Theorem 4.2.** Let $\rho$ be as in Theorem 4.1, and let

$$\gamma = 2g_0 1/\pi = 1.530\ldots,$$

where $g_0$ is the first positive zero of the Bessel function of the first kind $j_0$. For any $\Delta > \gamma$, $\pi - \gamma$, the following holds.

For any $\epsilon_1, \epsilon_2 > 0$, there is some $\alpha = 1/\text{poly}(k, \mathcal{A}, \pi/\Delta, 1/\lambda_{\text{min}}, 1/\epsilon_1, 1/\epsilon_2, 1/(\Delta - \gamma))$ for which there exists an algorithm with time and sample complexity $\text{poly}(1/\alpha)$ which, given poly($\alpha$)-granular sample access to $\rho$, outputs an $(\epsilon_1, \epsilon_2)$-accurate estimate for the parameters of $\rho$ with probability at least $4/5$.

4.1 Reduction to 2D Superresolution

In this section we reduce the problem of learning superpositions of Airy disks to the problem of learning a convex combination of Dirac deltas given the ability to make noisy, band-limited Fourier measurements.

Formally, suppose we had access to the following oracle:

**Definition 4.1.** An $m$-query, $\eta$-approximate OTF oracle $O$ takes as input a frequency $\omega \in \mathbb{R}^2$ and, given frequencies $\omega_1, \ldots, \omega_m$, outputs numbers $u_1, \ldots, u_m \in \mathbb{R}$ for which $|u_j - \hat{O}([\omega_j])| \leq \eta$ for all $j \in [m]$.

Remark 4.1. As we will see in Section 4.4, $O$ will be constructed by sampling some number of points from $\rho$ and computing empirical averages. The number $m$ and accuracy $\eta$ of queries that $O$ can answer dictates the sample complexity of this procedure. As we will see in the proofs of Lemma 4.6 and Lemma 4.12 below, the $m$ that we need to take will be small, so the reader can ignore $m$ and pretend it is unbounded for most of this section.

Given $\omega \in \mathbb{R}^2$, the Fourier transform of $\rho$ evaluated at frequency $\omega$ is given by

$$\hat{O}([\omega]) = \sum_{j=1}^k \lambda_j A(\omega) e^{-2\pi i (\mu_j, \omega)} = \frac{\gamma}{\mathcal{A}([\omega])} \sum_{j=1}^k \lambda_j e^{-2\pi i (\mu_j, \omega)},$$

(9)

where for $\omega = (r \cos \theta, r \sin \theta)$, we have by Fact 2 that $\mathcal{A}([\omega]) = \frac{2}{\pi} (\arcsin(r) - r \sqrt{1 - r^2})$. In particular, $\mathcal{A}([\omega])$ only depends on $r = \|\omega\|$ (because $\mathcal{A}(\cdot)$ is radially symmetric), so henceforth regard $\mathcal{A}$ as a function merely of $r$.

Define

$F(\omega) = \sum_{j=1}^k \lambda_j e^{-2\pi i (\mu_j, \omega)}$.

This is a trigonometric polynomial to which we have noisy pointwise access using $O$:

**Lemma 4.1.** Let $0 < \rho < 1$. With an $\eta$-approximate OTF oracle $O$, on input $\omega \in B_d^2$ we can produce an estimate of $F(\omega)$ to within $\eta/\mathcal{A}([\omega])$ additive error.

**Proof.** By dividing by $\mathcal{A}([\omega])$ on both sides of (9), we get that

$$\frac{\hat{O}([\omega])}{\mathcal{A}([\omega])} = \sum_{j=1}^k \lambda_j e^{-2\pi i (\mu_j, \omega)},$$

so given that $O$, on input $\omega$, outputs $u \in \mathbb{R}$ satisfying $|u - \hat{O}([\omega])| \leq \eta$, we have that

$$\left| \frac{u}{\mathcal{A}([\omega])} - F(\omega) \right| \leq \frac{\eta}{\min_{0 < r < \rho} \mathcal{A}(r')} = \frac{\eta}{\mathcal{A}([\omega])},$$

where the last step uses the fact that $\mathcal{A}([\cdot])$ is decreasing on the interval $[0, 1]$.

So concretely, given an $\eta$-approximate OTF oracle, we have reduced the problem of learning superpositions of Airy disks to that of recovering the locations of $\{\mu_i\}$ given the ability to query $F(\omega)$ at arbitrary frequencies $\omega$ for which $\|\omega\|_2 < 1$ to within additive accuracy $\eta/\mathcal{A}([\omega])$.

Lastly, for reasons that will become clear in subsequent sections (see e.g. (17)), it will be convenient to assume that $\mathcal{R} \leq 1/3$. This is without loss of generality, as otherwise, we can scale the data down by a factor of $3\mathcal{R}$ so that they are now i.i.d. samples from the superposition of Airy disks with density $\rho'(x) = \sum_{j=1}^k \lambda_j A_j / \mathcal{A}_j / \mathcal{R} \cdot (x - \mu_j / \mathcal{R})$. Define the rescaled centers $\mu_j' = \mu_j / \mathcal{R}$ and note that by assumption, $\|\mu_j'\|_2 \leq 1/3$ for all $j \in [k]$.

The Fourier transform of $\rho'$ is then given by $\hat{O}([\omega]) = \hat{A}_j / \mathcal{A}_j / \mathcal{R} \cdot \sum_{j=1}^k \lambda_j e^{-2\pi i (\mu_j', \omega)}$, so by the proof of Lemma 4.1 we conclude that with an $\eta$-approximate OTF oracle for $\rho$, for any $0 < r < 1$ on input $\omega \in B_d^2(r \cdot \mathcal{R})$ we can produce an estimate of $\sum_{j=1}^k \lambda_j e^{-2\pi i (\mu_j', \omega)}$ to
within $\eta/\tilde{A}[\ell]$ additive error. Recovering the centers $\{\mu'_j\}$ to within additive error $\epsilon$ then translates to recovering the centers $\{\mu_j\}$ to within additive error $3\Re\epsilon$. For this reason, we will henceforth assume that $\mathcal{R} \leq 1/3$.

4.2 Learning via the Optical Transfer Function

Our basic approach is as follows. To solve the superresolution problem of Section 4.1, we will project in two random, correlated directions $\omega_1, \omega_2 \in \mathbb{R}^2$ and solve the resulting one-dimensional superresolution problems via matrix pencil method (see ModifiedMPM) to recover the projections of $\mu_1, \ldots, \mu_k$ in the directions $\omega_1$ and $\omega_2$, as well as the relative intensities $\lambda_1, \ldots, \lambda_k$. From these projections we can then recover the actual centers for $\rho$ by solving a linear system (PreCONsolidate). Such an approach already achieves constant success probability, and we can amplify this by repeating and running a simple clustering algorithm (see SELECT). The full specification of the algorithm is given as LearnAiryDisks.

4.2.1 Learning in a Random Direction. Fix a unit vector $\nu \in \mathbb{S}^1$. We first show how to leverage Lemma 4.1 and the matrix pencil method to approximate the projection of $\mu_1, \ldots, \mu_k$ along $\nu$.

By the discussion at the end of Section 4.1, we may assume $\|\mu_i\| \leq 1/2$ for all $i \in [k]$, so $\|\mu_j - \mu_i\|_2 \leq 1$ for all $i \neq j$. For $j \in [k]$, let $m_j = \langle \mu_j, \nu \rangle$ and $\omega_j = e^{2\pi i (m_j/4k)}$. In this section we will assume that $m_j \neq 0$ for all $j \in [k]$.

For $\ell \in \mathbb{Z}_{\geq 0}$, let $u_\ell \triangleq \hat{F} \left( \frac{2\pi \nu}{\omega_j} \right) = \sum_{\ell \in [k]} \alpha_j \delta_{\ell j}$. Note that $u_0 = \hat{F}(0) = \sum_j \omega_j \ell_j = 1$. Also note that we do not have access to $\alpha_1, \ldots, \alpha_k$ and would like to recover $m_1, m_2$ given (noisy) access to $\{u_\ell\}_{0 \leq \ell < 2k-1}$.

Consider generalized eigenvalue problem $(VD_j\nu^T, VD_j\nu^T)$ where $V \in \mathbb{C}^{k \times k}$ is given by $V_{ij} = \delta_{j-i}^{-1}$ for $i, j \in [k]$, $D_k \triangleq \text{diag}(\lambda)$, and $D_k \triangleq \text{diag}(\alpha)$. The following facts are standard:

**Observation 4.1.** The eigenvalues of the generalized eigenvalue problem $(VD_j\nu^T, VD_j\nu^T)$ are exactly $\alpha_1, \ldots, \alpha_k$.

**Observation 4.2.** $(VD_j\nu^T)_{ij} = u_{i+j-2}$ and $(VD_j\nu^T)_{ij} = u_{i+j-1}$ for $i, j \in [k]$.

By Lemma 4.1, in reality we only have $\eta'_i$-approximate access to each $u_\ell$, where

$$\eta'_i \leq \frac{\eta}{\tilde{A}([\ell/4k])},$$

so we must instead work with $(VD_j\nu^T + E, VD_j\nu^T + F)$, where the $(i, j)$-th entry of $E$ (resp. $F$) is the noise $\eta'_i u_{i+j-2}$ (resp. $\eta'_i u_{i+j-1}$) in the observation of $u_{i+j-2}$ (resp. $u_{i+j-1}$).

If $V$ is well-conditioned, one can apply standard perturbation bounds to argue that the solutions to this generalized eigenvalue problem are close to those of the original $(VD_j\nu^T, VD_j\nu^T)$. Moreover, given approximations $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k$ to these generalized eigenvalues, we can find approximations $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k$ to $\lambda_1, \ldots, \lambda_k$ by solving the system of equations $v = \tilde{V}\lambda$, where $v = (v_0, \ldots, v_{2k-1})$, $\lambda = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_k)$, and $\tilde{V}$ is given by $V_{ij} = \tilde{\alpha}_{j-i}^{-1}$ for $i, j \in [k]$. The formal specification of the matrix pencil method algorithm ModifiedMPM that we use is given in Algorithm 1.

The following theorem, implicit in the proof of Theorem 2.8 in [39], makes the above reasoning precise. Henceforth, let $\kappa(\Lambda')$ and $\sigma_{\text{min}}(\Lambda')$ respectively denote the condition number and minimum singular value of $V$ when $m_{1, 2}$ have minimum separation $\Delta'$ for all $i \neq j$, and define $\lambda_{\text{min}} = \min_i \lambda_i$, $\lambda_{\text{max}} = \max_i \lambda_i$.

**Theorem 4.3 ([39])** Suppose $m_1, m_2 \in [-1/4, 1/4]$ have separation at least $\Delta'$ and we are given $\eta'_i$-close estimates to $u_\ell$ for $0 \leq \ell \leq 2k-1$. Define $\gamma = \frac{\eta'_{\text{min}}}{\lambda_{\text{min}}} \delta$, $\alpha = \frac{1}{\gamma}$ and $\zeta = \frac{2\lambda_{\text{max}} + \gamma'}{\sigma_{\text{min}}(\Lambda' - 2\gamma)}$.

Then $\|E\| + \|F\| \leq \frac{\sigma_{\text{min}}(\Lambda')^2}{\sigma_{\text{min}}(\Lambda')^2} \lambda_{\text{min}}$ and $\gamma' = \frac{1}{\alpha}$.

ModifiedMPM produces estimates $\hat{\lambda}_i$ for the mixing weights and estimates $\{\tilde{m}_i\}$ for the projected centers such that for some permutation $\tau$:

$$|m_i - \tilde{m}_{\tau(i)}| \leq 8\gamma' \text{ and } |\lambda_i - \hat{\lambda}_i| \leq \zeta'$$

for all $i \in [k]$.

Note that the guarantees of Theorem 4.3 are stated in [39] in terms of wrapparound distance on the interval $[-1/2, 1/2]$, but because $m_{\tau(i)} \in [-1/4, 1/4]$ for all $i \in [k]$, $m_{\tau(i)} \in [-1/4, 1/4]$ have pairwise separation $\Delta'$ both in absolute and wrapparound distance.

In other words, the output of ModifiedMPM converges to the true values for $\{\mu_i, \nu\}_i \in [k]$ and $\{\tilde{m}_i\}_i \in [k]$ at a rate polynomial in the noise rate, condition number of $V$, and relative intensity of the Airy disks, provided $\sigma_{\text{min}}(\Lambda')$ is inverse polynomially large and $\kappa(\Lambda')$ is polynomially small in those parameters.

To complete the argument, we must establish these bounds on $\sigma_{\text{min}}$ and $\kappa$. Henceforth, let $\Delta' = \min m_{1, 2}/4k$. Theorem 4.3 states that in [39] we can take $\gamma' = \frac{\eta'_{\text{min}}}{\lambda_{\text{min}}} \delta$, $\alpha = 1/\delta$, and $\zeta = 2\lambda_{\text{max}} + \gamma'$. Hence, let

$$\Delta' = \min m_{1, 2}/4k.$$

**Algorithm 1 ModifiedMPM**

1. **Input**: $\omega \in \mathbb{S}^1$, $\eta$-approximate OTP oracle $O$

2. **Output**: Estimates $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_k\}$ for the mixing weights and $\{\tilde{m}_1, \ldots, \tilde{m}_k\}$ for the centers of $\rho$ projected in direction $\omega$

3. Define $\tilde{\nu}_0 = 1$.

4. For $0 \leq \ell \leq 2k - 1$: invoke $O$ on input $\omega \tilde{\nu}_\ell$ to produce $u_\ell \in \mathbb{R}$. Compute $\tilde{v}_\ell \triangleq \frac{u_\ell}{\tilde{A}([\ell/4k])}$.

5. Form the matrices

$$X \triangleq \begin{pmatrix} \tilde{v}_0 & \cdots & \tilde{v}_{k-1} \\ \vdots & \ddots & \vdots \\ \tilde{v}_{k-1} & \cdots & \tilde{v}_{2k-2} \end{pmatrix}, \quad Y \triangleq \begin{pmatrix} \tilde{v}_1 & \cdots & \tilde{v}_k \\ \vdots & \ddots & \vdots \\ \tilde{v}_k & \cdots & \tilde{v}_{2k-1} \end{pmatrix}.$$

6. Solve the generalized eigenvalue problem $(X, Y)$ to produce generalized eigenvalues $\tilde{\lambda}_1, \tilde{\lambda}_2$.

7. For $i = 1, 2$, let $\tilde{m}_i$ be the argument of the projection of $\tilde{\lambda}_i$ onto the complex unit disk.

8. Form the matrix

$$\tilde{V} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \tilde{\alpha}_1 & \tilde{\alpha}_2 & \cdots & \tilde{\alpha}_k \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\alpha}_1^{-1} & \tilde{\alpha}_2^{-1} & \cdots & \tilde{\alpha}_k^{-1} \end{pmatrix}.$$

9. Solve for $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ such that $\tilde{V}\hat{\lambda} = (\tilde{v}_0, \ldots, \tilde{v}_{k-1})$.

10. Output $\{\hat{\lambda}_1\}_i \in [k]$ and $\{\tilde{m}_i\}_i \in [k]$. 


Lemma 4.2. For any $k \geq 2$, we have that
\[ \sigma_{\min}(\Delta')^2 \geq (\Delta/k^2)^{k-1} \ \text{and} \ \kappa(\Delta')^2 \leq k^{2k-1}/\Delta(k-1)^2. \]

Proof. First note that $\sigma_{\max}(\Delta')^2 \leq k^2$. Indeed because the entries of $V$ all have absolute value at most 1, we conclude that for any $v \in \mathbb{S}^{k-1}$ and any row index $j \in [k]$,
\[ \langle V_j, v \rangle^2 \leq \left( \sum_{i=1}^{k} |v_i| \right)^2 \leq k. \]

On the other hand, we also have that
\[ \prod_{i=1}^{k} \sigma_i(V) = \prod_{i<j} |\alpha_i - \alpha_j| \leq e^{2\pi N - 1} \left( \frac{2 - 2 \cos(\lambda')}{\lambda'} \right)^{1/2}, \]
where in the first step we used the standard identity for the determinant of a Vandermonde matrix, and in the second step we used the angular separation of the $\alpha_i$. We can further lower bound by $\Delta^k(k-1)/2$ using the elementary inequality $\cos(\lambda') \leq 1 - \Delta^2/2$, from which the lemma follows. \hfill \Box

This yields the following consequence for ModifiedMPM.

Corollary 4.1. Given $\omega \in S^3$ and access to an $\eta$-approximate OTF oracle $O$, if the projected centers $m_j = (\mu_j, v)$ satisfy $|m_i - m_j| \leq 4k \cdot \Delta'$ for all $i \neq j$ for some $0 < \Delta' \leq 1/16$, then there exists a constant $c_1 > 0$ such that provided that
\[ \eta \leq c_1 \alpha_{\min} \Delta^k \cdot k^{2k-1}/2, \]
then ModifiedMPM produces estimates $(\hat{\lambda}_i)$ for the mixing weights and estimates $(\hat{m}_i)$ for the projected centers such that for some permutation $\tau$: $|m_i - \hat{m}_{\tau(i)}| \leq (k^{2k+1/2} \eta / \alpha_{\min} \Delta^k (k-1)^2)$ and $|\lambda_i - \hat{\lambda}_{\tau(i)}| \leq (k^{2k+1/2} \eta / \alpha_{\min} \Delta^k (k-1)^2)$ for all $i \in [k]$.

Proof. From Lemma 4.2, we have that $\sigma_{\min}(\Delta')^2 \geq (\Delta/k^2)^{k-1}$. Then because $\alpha(\Delta')^2 \leq k^2/\sigma_{\min}(\Delta')^2$, we would like to conclude by Theorem 4.3 that $|m_i - \hat{m}_{\tau(i)}| \leq 8\eta$, where
\[ y = O \left( \frac{\|\eta\|_2 \cdot k^2}{\lambda_{\min} \cdot \sigma_{\min}(\Delta')^2} \right) = O \left( \frac{k^{2k+1/2} \cdot \eta}{\alpha_{\min} \Delta^k (k-1)^2} \right), \]
where in the last step we use the vector $\eta'$ has length $O(k)$ and satisfies $\|\eta\|_\infty \leq O(\eta)$ by (10). To do so, we just need to verify that $\|E| + \|F| < \sigma_{\min}(\Delta')^2 \lambda_{\min}$ and $y < \Delta'/4$. The latter clearly follows from the bound (11) for sufficiently small $c_1$. For the former,
\[ \|E\|_2 \leq \sqrt{k} \cdot \eta_{1/2} \cdots \eta_{2k-1} \leq \eta \sqrt{k} \cdot \prod_{i=1}^{2k-1} \frac{1}{\lambda/4^k} \leq O(\eta \cdot k), \]
where the last step follows by the fact that $\hat{\lambda}[t/4k] \geq \hat{\lambda}[1/2] \geq \Omega(1)$. The same bound holds for $\|F\|_2$. Recalling that $\sigma_{\min}(\Delta')^2 \geq (\Delta^2/k^2)^k$, it is enough for $\eta \leq O(\Delta^2/k^2)^{k-1} \lambda_{\min}/k$, which certainly holds for $\eta$ satisfying (11), for $c_1$ sufficiently small.

Finally, Theorem 4.3 also implies that $|\lambda_i - \hat{\lambda}_i| \leq \zeta$, where
\[ \zeta \leq O \left( \frac{y + k^{1/2} \eta}{\sigma_{\min}(\Delta')^2} \right) \leq O \left( \frac{k^{3k+1/2} \cdot \eta}{\alpha_{\min} \Delta^k (k-1)^2} \right), \]
as claimed. \hfill \Box

4.2.2 Combining Directions. We can run ModifiedMPM to approximately recover $\{\langle \mu_1, \omega_1 \rangle, \omega_1 \}_{i \in [k]}$ and $\{\langle \mu_2, \omega_2 \rangle, \omega_2 \}_{i \in [k]}$ for two randomly chosen directions $\omega_1, \omega_2 \in S^3$. As these directions are random, with high probability we can combine these estimates to obtain an accurate estimate of $\{\langle \mu_j, \omega_1 \rangle, \omega_1 \}_{j \in [k]}$. One subtlety is that the estimates $(\hat{m}_i)$ and $(\hat{m}_j')$ output by ModifiedMPM for the centers projected in directions $\omega_1$ and $\omega_2$ respectively need not be aligned, that is we only know that there exists some permutation $\tau$ for which $\hat{m}_j = \hat{m}_{j'}(\tau)$ for $j \in [k]$.

We first show a “pairing lemma” stating that if $\omega_1$ is chosen randomly and $\omega_2$ is chosen to be close to $\omega_1$, then if one sorts the centers $\mu_1, \ldots, \mu_k$, first in terms of their projections in the $\omega_1$ direction, and then in terms of their projections in the $\omega_2$ direction, the corresponding elements in these two sorted sequences will correspond to the same centers. We require the following elementary fact.

Lemma 4.3. For $\mu \in \mathbb{R}^2$ a unit vector and $\omega \in \mathbb{R}^2$ a random unit vector, $\Pr_{\omega}[|\langle \mu, \omega \rangle| \leq \sin \theta] = 20/\pi$ for all $0 \leq \theta \leq \pi/2$.

Lemma 4.4. Fix an arbitrary $0 < \theta \leq \pi/2$ and let $v = \Delta \sin \theta$. Let $\omega_1, \omega_2 \in \mathbb{S}^2$ be a random unit vector, and let $\omega_1, \omega_2$ be either of the two unit vectors for which $||\omega_1 - \omega_2||_2 = v$. For every $i \in [k]$, define $m_i = \langle \mu_i, \omega_1 \rangle + \langle \mu_i, \omega_2 \rangle$, let $\hat{m}_i, \hat{m}_i' \in \mathbb{R}$ be any numbers for which $\|\hat{m}_i - \hat{m}_i'\| < 2v$.

Then with probability at least $1 - \frac{k(\pi - 2\theta)}{2\pi}$, for every $i \neq j$ the following are equivalent: I) $m_i > m_j$, II) $m_i' > m_j'$, III) $\hat{m}_i > \hat{m}_j$, and IV) $\hat{m}_i' > \hat{m}_j'$.

Proof. By Lemma 4.3 and a union bound we have that with probability $1 - \frac{k(\pi - 2\theta)}{2\pi}, |m_i - m_j| > \Delta \sin \theta$ for all $i \neq j$. Fix any $i \neq j$ and suppose that $m_i > m_j$. Then by triangle inequality and Cauchy-Schwarz, we have that
\[ m_i' - m_j' = \langle \mu_1, m_i \rangle + \langle \mu_2, m_i \rangle > \Delta \sin \theta > 4v, \]
where the final inequality follows by the definition of $v$. So I) implies II) and by symmetry we can show II) implies I). We also have that $\hat{m}_i - \hat{m}_j \geq (m_i - m_j) - 4v > 0$, so I) implies III) and by symmetry we can show III) implies IV).

It is enough to show that III) implies I). Suppose $\hat{m}_i > \hat{m}_j$. Then
\[ m_i - m_j = (\hat{m}_i - \hat{m}_j) - 4v < -\frac{1}{2} \Delta \sin \theta > -\Delta \sin \theta, \]
so it must be the case that $m_i - m_j > 0$ given that $|m_i - m_j| > \Delta \sin \theta$.

We now show that we can combine these projected center estimates to approximately recover the two-dimensional centers by solving a linear system. The specification of this algorithm, which we call PreConsolidate, is given in Algorithm 2.

Lemma 4.5. Let $\xi > 0$. Let the parameters $\theta, v$ and the random vectors $\omega_1, \omega_2$ be as in Lemma 4.4. Suppose $\hat{m}_i, \hat{m}_i' \in \mathbb{S}_k$ are collections of numbers for which there exist permutations $\tau_1, \tau_2$ for which
\[ |\langle \omega_1, \mu_i \rangle - \hat{m}_{\tau_1(i)}| \leq \xi \quad \text{and} \quad |\langle \omega_2, \mu_i \rangle - \hat{m}_{\tau_2(i)}'| \leq \xi \]
for all $i \in [k]$. 498
Algorithm 2 PreConsolidate

1. **Input:** Directions \( \omega_1, \omega_2 \in S^1 \) and estimates \( \{ \hat{\lambda}_i \}, \{ \hat{m}_i \} \) and \( \{ \hat{\lambda}'_i \}, \{ \hat{m}'_i \} \) for the parameters of \( \rho \) projected in the directions \( \omega_1 \) and \( \omega_2 \) respectively.
2. **Output:** Estimate \( \{ \hat{\lambda}_i \}, \{ \hat{m}_i \} \) for the parameters of \( \rho \).
3. **for** \( i \in [k] \) **do**
4. Let \( t, t' \in [k] \) be the indices for which \( \hat{m}_{t} \) and \( \hat{m}'_{t} \) are the \( i \)-th largest in \( \{ \hat{m}_j \} \) \( j \in [k] \) and \( \{ \hat{m}'_j \} \) \( j \in [k] \) respectively.
5. Define a formal vector-valued variable \( v^{(i)} \in \mathbb{R}^2 \) and solve the linear system \( \langle \omega_1, v^{(i)} \rangle = \hat{m}_{t} \) and \( \langle \omega_2, v^{(i)} \rangle = \hat{m}'_{t} \).
6. **Output** \( \{ \hat{\lambda}_i \}_{i \in [k]}, \{ \hat{\mu}_i \}_{i \in [k]} \).
7. **Proof.** Condition on the event of Lemma 4.4 occurring, which happens with probability at least \( 1 - \frac{k^{-1}(k-1)\theta}{\pi} \), we have that the output \( \{ \hat{\lambda}_i \}, \{ \hat{\mu}_i \} \) of PreConsolidate satisfies \( \| \hat{\mu}_i - \mu_{\tau(i)} \|_2 \leq \frac{\theta}{\sqrt{k}v^{(i)} \cdot v^{(i)}} \) for some \( \tau \in \mathbb{S}_k \).

Finally, we show how to boost the success probability via the following naive clustering-based algorithm Select (Algorithm 3), whose guarantees we establish below.

We can now give the full specification of our algorithm LearnAiryDisks (see Algorithm 4).

Lemma 4.6. Let \( \rho \) be a \( \Delta \)-separated superposition of \( k \) Airy disks. Then for any \( \varepsilon_1, \varepsilon_2, \delta > 0 \), let

\[
\eta = O \left( \frac{\Delta}{4\kappa} \right) \cdot O \left( \frac{\lambda_{\min}^2}{\lambda_{\max}} \right) \cdot \min \left( \varepsilon_1/M, \varepsilon_2 \right). \tag{12}
\]

Without loss of generality suppose \( \varepsilon_1 < 3\Delta/8 \). Then the output \( \{ \lambda_i, \lambda'_i, \mu_i, \mu'_i \} \) of LearnAiryDisks, given \( \varepsilon_1, \varepsilon_2, \delta \) and access to an \( \eta \)-approximate, \( O(\log(1/\delta)) \)-query OTF oracle \( O \) for \( \rho \), satisfies

\[
\| \hat{\mu}_i - \mu_{\tau(i)} \|_2 \leq \varepsilon_1 \quad \text{and} \quad |\hat{\lambda}_i - \lambda_{\tau(i)}| \leq \varepsilon_2
\]

for some permutation \( \tau \) with probability at least \( 1 - \delta \). Furthermore, the runtime of LearnAiryDisks is dominated by the time it takes to invoke the OTF oracle \( O(\log(1/\delta)) \) times.

Proof. Suppose we are given a valid \( \eta \)-approximate OTF oracle \( O \). By taking \( \theta = \frac{\pi}{3\kappa(k-1)} \) and invoking Lemmas 4.1 and 4.5, we ensure that a single run of PreConsolidate in an iteration of the loop in Step 5 of LearnAiryDisks will yield, with probability at least \( 1 - \frac{1}{\Delta} \), an \( \eta_i', \eta_i'' \)-accurate estimate, where

\[
\eta_i' = \frac{3\pi}{3\kappa(k-1)} \cdot \frac{O \left( k^{3/2} + 1 \right)}{\lambda_{\min}^2} \cdot \frac{\eta}{\lambda_{\max}^2}
\]

and \( \eta_i'' = O \left( \frac{k^{3/2} + 1}{\lambda_{\min}^2} \cdot \frac{\eta}{\lambda_{\max}^2} \right) \).

In this case we say that such an iteration of the loop in LearnAiryDisks "succeeds." We take \( \varepsilon \) such that \( \varepsilon' = \varepsilon_1/3 \) and \( \varepsilon'' = \varepsilon_2 \). The bound in (12) then follows from the elementary inequality \( \sin \theta \geq \theta/2 \) for \( 0 \leq \theta \leq 1 \), together with our choice of \( \theta = \frac{\pi}{3\kappa(k-1)} \).

Each iteration of the loop in Step 5 of LearnAiryDisks individually succeeds with probability at least \( 1 - \frac{1}{\Delta} \). So by a Chernoff bound, by taking \( \varepsilon = O(\log(1/\delta)) \), we conclude that with probability at least \( 1 - \delta \), at least \( 1 - \frac{1}{2\kappa} \) fraction of these iterations will
succeed. So of the $k \cdot T$ elements in $S$, at most $T/2$ correspond to failed iterations.

Now note that all $(t, i)$ for which $t$ corresponds to a successful iteration will be $2\varepsilon_1'$-close to at least $k \cdot T - T/2 > 2T/3$ points. In particular, any such $(t, i)$ will be among the vertices $V$ of $G$ in Algorithm 3. Conversely, for any $(t, i) \in V$, $\tilde{\mu}_t$ is by definition $2\varepsilon_1'$-close to at least $2T/3$ points and there are at most $T/2 < 2T/3$ points which do not correspond to successful iterations. In particular, at least one of the points that $\tilde{\mu}_t$ close to will correspond to a successful iteration, so by the triangle inequality $\|\tilde{\mu}_t - \mu_j\| \leq 3\varepsilon_1'$ for some choice of $j \in [k].$

Observe that $G$ is $k$-partite because every vertex in $V$ is $3\varepsilon_1'$-close to some center of $\rho$, but two vertices which are $3\varepsilon_1'$-close to $\mu_i$ and $\mu_j$ respectively for $i \neq j$ must be distance at least $\Delta - 6\varepsilon_1' > 2\varepsilon_1'$ apart. We conclude that with high probability, SELECT will output $3\varepsilon_1' = \varepsilon_1$-accurate estimates for the centers of $\rho$.

It remains to show that $\lambda_1^*, \lambda_2^*$ are $\varepsilon_2$-accurate estimates for the mixing weights. We know the estimates $\lambda_t^*$ corresponding to successful iterations $t$ and center $\mu_t$ lie in $\{\lambda_t^*\}_{(t,i) \in V(t)}$ for some $t$. Then $\{\lambda_t^*(i)_{(t,i) \in V(t)}$ contains at least $\left(1 - \frac{1}{2k}\right) T > 2T/3$ values that are $\varepsilon_2'$-close to $\lambda_t$, and at most $T/2 < 2T/3$ other values. Call these values "good" and "bad" respectively. Either the median is good, in which case we are done, or the median is bad, in which case because there are strictly more good values than bad values, the median must be upper and lower bounded by good values, in which case we are still done.

Finally, note that in each iteration of the main loop of LEARN-AiryDisks, $O$ is invoked exactly six times. Furthermore, other than these invocations of $O$, the remaining steps of LEARN-AiryDisks require constant time. So the runtime of LEARN-AiryDisks is indeed dominated by the $O(\log(1/\delta))$ calls to $O$. □

### 4.3 Learning Airy Disks Above the Diffraction Limit

In this subsection we present the proof of Theorem 4.2. Recall that we are assuming that $\sigma = 1/2$ and $\Delta > \sqrt{2}$, where $\Delta$ is defined in (8).

Let $\varepsilon = \frac{1}{2^{3/2}}(\Delta + \sqrt{2})$ and define $R = \frac{\varepsilon}{\sqrt{2}}$ and $r = 1/2 - R$.

We will use the following Algorithm 5 that we call TENSORRESOLVE. While this is only a slight modification of the tensor decomposition algorithm of [30] for high-dimensional supersolution, our analysis is novel and obtains sharper results in low dimensions by using certain extremal functions [10, 23, 26] arising in the study of de Branges spaces (see Theorem 4.4).

Using the notation of TENSORRESOLVE, define the tensor $T \in \mathbb{C}^{m \times m \times k^3}$ given by

$$T_{a,b,i} = \sum_{j=1}^{k} \lambda_j e^{-2\pi i (\mu_j, a + \omega_i^a) + v(1)} \rho_{a,b,i}$$

and note that it admits a low-rank decomposition as

$$T = \sum_{j=1}^{k} \left( V_j \otimes V_j \otimes (W_j D_j) \right),$$

where $D_j$ is the diagonal matrix whose entries consist of the mixing weights $\lambda_j$ and, for every $a, j \in [k]$, $W_{aj} = e^{-2\pi i (\mu_j, a \omega_i)}$ and $V_{aj} = e^{-2\pi i (\mu_j, a \omega_i)}$.

---

#### Algorithm 5 TENSORRESOLVE

1. **Input:** Error parameters $\varepsilon_1, \varepsilon_2$, confidence parameter $\delta$, access to $\eta$-approximate, $\Theta\left(\frac{k^2 \log(1/\delta)}{(1 - 3/2k)}\right)$ query OTF oracle
2. **Output:** With probability at least $1 - \delta$, an $(\varepsilon_1, \varepsilon_2)$-accurate estimate $(\hat{\lambda}_1, (\hat{\mu}_1))$ for the parameters of $\rho$, provided the separation is sufficiently above the diffraction limit (see Lemma 4.12)
3. Define $R = \sqrt{2}/2c$ and $r = 1/2 - R$.
4. Sample $\omega^{(1)}, \ldots, \omega^{(m+2)}$ i.i.d. from the uniform distribution over $B^2(R)$. Also define $\omega^{(m+1)} = (1, 0), \omega^{(m+2)} = (0, 1), \omega^{(m+3)} = (0, 0)$. Define $m' = m + 3$
5. Sample $v$ uniformly from $\mathbb{S}^1$ and define $v(1) = b \cdot v, v(2) = 2b \cdot v,$ and $v(3) = 0$.
6. Define $\xi_{a,b,i} = \omega(a) + \omega(b) + v(1)$ for every $a, b \in [m']$.
7. Query the OTF oracle at $(\xi_{a,b,i})$ to obtain numbers $(u_{a,b,i})$. Construct the tensor $\hat{T} \in \mathbb{C}^{m' \times m' \times k^3}$ given by $\hat{T}_{a,b,i} = u_{a,b,i} / \hat{A}(\xi_{a,b,i})$.
8. For each $j \in [k], i \in [2]$, let $\hat{\mu}_j \in \mathbb{R}^2$ have $i$-th entry equal to the argument of the projection of $\hat{V}_{m+i,j}$ onto the complex disk.
9. Query the OTF oracle at frequencies $\omega(a) \in [m']$ to get numbers $u_{a,b,i} / \hat{A}(\omega(a))$ and form the vector $\hat{v} \in \mathbb{R}^{m''}$ whose $a$-th entry is $u_{a,b,i} / \hat{A}(\omega(a))$ for every $a \in [m']$.
10. Let $\hat{a} = \arg\min_{a} \|\hat{V} - \hat{v}\|^2$.
11. Return $(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ and $(\hat{\mu}_1, \ldots, \hat{\mu}_k)^T$.

$V_{aj} = e^{-2\pi i (\mu_j, a \omega_i)}$. Let $V \in \mathbb{R}^{m' \times k}$ denote the matrix whose $j$-th column is $V_j$.

Note that by our choice of $r, R$ and triangle inequality, we have that $\|\omega(a) + \omega(b) + v(1)\|_2 \leq r + 2R = 1 - c/\sqrt{2} < 1$ for any entry index $a, b, i$. So if $(u_{a,b,i})$ are the numbers obtained from an $\eta$-approximate, $(m + 3)$-query OTF oracle as in Algorithm 5, and $\hat{T}$ is constructed as in Step 6 of TENSORRESOLVE, then by Lemma 4.1 we have that

$$|T_{a,b,i} - \hat{T}_{a,b,i}| \leq \frac{\eta}{A(1 - \frac{c}{\sqrt{2}})} \leq \eta \cdot \left(\frac{r - \sqrt{2}}{2c}\right)^2,$$

where the last step follows by Fact 3.

The following is a consequence of the stability of Jenrich’s algorithm.

#### Lemma 4.7. [e.g. [30], Lemma 3.5] For any $\varepsilon, \delta > 0$, suppose $|T_{a,b,i} - \hat{T}_{a,b,i}| \leq \eta' \leq \eta$ for all $i \in [k]$. Then with probability at least $1 - \delta$ over the randomness of $\omega(i)$, there exists permutation matrix $P$ such that $\|\hat{V} - VP11\|_F \leq \varepsilon$ for all $j \in [k]$. The setting of parameters in [30] is slightly different from ours, so we provide a self-contained proof of Lemma 4.7 in the full version.

We will also need the following basic lemma about the stability of solving for $\hat{a}$ in Step 10 in TENSORRESOLVE.

#### Lemma 4.8. For any $\varepsilon, \varepsilon' > 0$, if $\hat{a} \in \mathbb{R}^k$ satisfies $\hat{V}_{a} = b$ for some $V \in \mathbb{R}^{m' \times k}$ and $b \in \mathbb{R}^m$, and furthermore $\hat{b}$ satisfy $\|\hat{V} - \hat{V}\|_2 \leq \varepsilon$
and \( \| \hat{b} - \tilde{b} \|_2 \leq \epsilon' \), then \( \hat{\lambda} \pm \arg \min \hat{\lambda} \| \hat{V} \hat{\lambda} - \hat{b} \|_2 \leq 2\epsilon \| \hat{V} \|_q \epsilon' \).

Proof. Note that
\[
\| \hat{V} \hat{\lambda} - \hat{b} \|_2 \leq \| (\hat{V} - V) \hat{\lambda} \|_2 + \| \hat{b} - b \|_2 \leq \epsilon \| \hat{\lambda} \|_2 + \epsilon'.
\]
By triangle inequality and definition of \( \hat{\lambda} \), \( \| \hat{V} \hat{\lambda} - \hat{b} \|_2 \leq 2\epsilon \| \hat{\lambda} \|_2 + 2\epsilon' \), so \( \| \hat{\lambda} - \hat{\lambda} \|_2 \leq 2\epsilon \| \hat{\lambda} \|_2 + 2\epsilon' \). The lemma follows as \( \sigma_{\min}(\hat{V}) \geq \sigma_{\min}(V) - \epsilon \).

It remains to show the following condition number bound.

**Lemma 4.9.** For any \( \delta > 0 \), if \( m = \Theta \left( \frac{k^2 \log(\delta / \Delta) \sqrt{\ell}}{\Delta \gamma} \right) \), then \( \kappa(V) \leq O \left( \frac{k \sqrt{k}}{\sqrt{\Delta - \delta}} \right) \) and \( \sigma_{\min}(V) \geq \Omega \left( k^2 \log(\delta / \Delta) \right) \) with probability at least \( 1 - \delta \).

Proof. Let \( V^* \in \mathbb{R}^{m \times k} \) denote the submatrix given by the first \( m \) rows of \( V \). We will need the following basic lemma from [30] relating the condition number of \( V^* \) to that of \( V \):

**Lemma 4.10** ([30], Lemma 3.8). \( \kappa(V) \leq \sqrt{\Delta} \cdot \kappa(V^*) \).

The primary technical component of this section is to upper bound \( \kappa(V^*) \). First, note that given any \( \lambda \in \mathbb{C}^{k-1} \), we have that
\[
\lambda^v V^* \lambda = \sum_{j=1}^{k} \langle \lambda, V_{*,j} \rangle^2 = \sum_{j=1}^{k} \int_{B^{2}(R)} \left| \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} \right|^2 d\psi(\omega),
\]
where \( d\psi(\omega) \) is the uniform measure over \( B^2(R) \). Furthermore, for any \( \omega \in B^2(R) \) and \( i \in [m] \), we have that
\[
0 \leq \langle \lambda, V_{*,i} \rangle^2 \leq \lambda^v \| \lambda \|_2^2 \leq K \cdot \| \lambda \|_2^2.
\]
So by matrix Hoeffding applied to \( m \cdot V^* \cdot V^* \cdot m \) each of which is upper bounded in spectral norm by \( m \cdot k \) based on (13), we conclude that
\[
\Pr \left( \| V^* V^* - \mathbb{E}[V^* V^*] \|_2 > \sqrt{k} \right) \leq k \cdot e^{-\Omega(\epsilon^2)} \forall \epsilon > 0.
\]

**Lemma 4.11** below allows us to bound the quadratic form given by the expectation term evaluated at any \( \lambda \). Taking \( t = O(\sqrt{\log k / \delta}) \) and \( m = \Theta \left( \frac{k^2 \log(k/\delta)}{(\Delta - \delta)} \right) \) in (14) and applying Lemma 4.11, we conclude that with probability at least \( 1 - \delta \), \( \lambda^v V^* \lambda \) is bounded between \( \Omega(\gamma) \cdot \{ (\Delta - \delta) \cdot 1 \} \cdot \| \lambda \|_2^2 \) and \( O(\gamma) \cdot (k + ((\Delta - \delta) \cdot 1)) \cdot \| \lambda \|_2^2 \) from which it follows that with this probability, \( \kappa(V^*) \leq O \left( \frac{k}{(\Delta - \delta)} \right)^{1/2} \), from which the lemma follows by Lemma 4.10.

It remains to show Lemma 4.11 below, the key technical ingredient of this section. We will require the following special case of a result of [23], which essentially follows from results of [10, 26]. This can be thought of as the high-dimensional generalization of the well-known Beurling-Selberg minorant (see, e.g., [47] for a discussion of the one-dimensional case).

**Theorem 4.4** ([23], Theorem 1). For any \( d \in \mathbb{N} \) and \( \frac{\log_1 - 1}{2} < \frac{1}{r} \leq \frac{1}{2} \), there exists a function \( M \in L^1(\mathbb{R}^d) \) whose Fourier transform is supported in \( B^2(d') \), and which satisfies \( M(x) \leq \mathbb{I} \cdot x \in B^2(d') \) for all \( x \in \mathbb{R}^d \) and \( M[0] = \left( \frac{2 \pi \gamma}{d'^2} \right) \cdot \frac{C(d, r)}{\gamma} \cdot \frac{d'(d - 1)}{d'} \), where \( d' - 1 \) denotes the surface area of \( \mathbb{S}^{d-1} \) and \( C(d, r) = \frac{\pi d}{d'} \cdot \frac{1}{\Gamma \left( \frac{d + 2}{d'} \right)^2} \cdot \left( \frac{d'}{d} \right) \).

**Lemma 4.11.** \( \int_{B^2(R)} \left| \sum_{j=1}^{k} \lambda_j e^{-2\pi i \langle \mu_j, \omega \rangle} \right|^2 d\psi(\omega) \) is bounded between \( \Omega((\Delta - \delta) \cdot 1) \cdot \| \lambda \|_2^2 \) and \( k \| \lambda \|_2^2 \) where \( d\psi(\omega) \) denotes the uniform probability measure over \( B^2(R) \).

Proof. The upper bound follows by (13). We now show the lower bound. By Theorem 4.4 applied to \( d = 2 \), for any \( \frac{\pi}{2} < r < \frac{1}{2} \), there is a function \( M \) which minimizes the indicator function of \( B^2(1) \) and has Fourier transform supported in \( B^2(r) \). Take \( r = \{ \Delta R \times \frac{\pi^{d'}/d'}{2} \} \) which satisfies \( \pi / 2 < r < \frac{1}{2} \). This implies that the function \( M(\omega) = \frac{1}{\pi} \cdot \frac{1}{\hat{M}(\omega)} \cdot M(\omega) \) minimizes the density \( \psi(\omega) \), has Fourier transform supported in \( B^2(r) \) \( \subseteq B^2(\Delta) \), and satisfies
\[
\tilde{M}[0] = \frac{1}{\pi R^2} \left( \frac{2r^2}{R^2} \right) \cdot \left( \frac{C(2, r)}{1 + C(2, r)^2} \right) \geq \frac{1}{\pi R^2} \geq 4 - 2\pi \cdot \frac{1}{\pi R^2} \geq \Omega((\Delta - \delta) \cdot 1).
\]

Putting everything together, we have the following guarantee:

**Lemma 4.12.** Let \( \rho \) be a \( \Delta \)-separated superposition of \( k \) Airy disks. For any \( \epsilon_1, \epsilon_2, \delta > 0 \), let
\[
m = \Theta \left( \frac{k^2 \log(k/\delta)}{(\Delta - \delta)} \right), \eta = O \left( \frac{4 \Delta \delta^2 \min(\Delta - \delta, k) \cdot (15)}{\tau \cdot \sqrt{k^2 \log(k/\delta)}} \right) \cdot \left( \frac{\log(1/\epsilon)}{\epsilon} \right)^{1/2}.
\]

WLOG suppose \( \epsilon_1 < 1/6 \). Then the output \( (\lambda_1', \lambda_2', \mu_1', \mu_2') \) of TENSOR-RESOLVE, given \( \epsilon_1, \epsilon_2, \delta \) and access to an \( \eta \)-approximate, \( m \)-query OTF oracle \( O \) for \( \rho \), satisfies
\[
\| \mu_1 - \mu_1' \|_2 \leq \epsilon_1 \quad \text{and} \quad |\lambda_1 - \lambda_1'| \leq \epsilon_2
\]
for some permutation \( \tau \) with probability at least \( 1 - \delta \). Furthermore, the runtime of LEARNARYDISKS is polynomial in \( k \), the number of OTF oracle queries, and the time it takes to make those queries.

**Proof.** By Lemma 4.7, if we take \( m = \Theta \left( \frac{k^2 \log(k/\delta)}{(\Delta - \delta)} \right) \) and \( \eta = O \left( \frac{4 \Delta \delta^2 \min(\Delta - \delta, k) \cdot (15)}{\tau \cdot \sqrt{k^2 \log(k/\delta)}} \right) \cdot \left( \frac{\log(1/\epsilon)}{\epsilon} \right)^{1/2} \), then the output \( \hat{V} \) of JENNIRICH(T) satisfies
\[
\| \hat{V} - \hat{V}' \|_F \leq \epsilon_1 \text{ for some permutation matrix } P \text{.}
\]

In fact, one can improve the upper bound by using a suitable majorant for the indicator of the ball.
the elementary inequality \(|e^{-2\pi ix} - 1| \geq 2|x|| for any \( |x| \leq 2/3\) and the fact that
\[
\left( \hat{\mu}_j - \mu_j, \omega(f) \right) \leq \left\| \hat{\mu}_j - \mu_j \right\| \|\omega(f)\|_2 \leq 2R \leq 2/3,
\]
we conclude that \(|\hat{\mu}_j - \mu_j, \omega(f)\| \leq \varepsilon_1/2\) for all \( j \in [k], \ell \in [m] \). In particular, this holds for all \( \ell = m+1 \) and \( \ell = m+2 \), so \(|\hat{\mu}_j - \mu_j| \leq \varepsilon_1/2\). By dividing \( \varepsilon_1 \) by \( \sqrt{2} \) and absorbing constants, we get that the estimates \( \hat{\mu}_j \) for the centers are \( \varepsilon_1/2\)-close to the true centers.

To show that the mixing weights are \( \varepsilon_2\)-close to the true mixing weights, we can apply Lemma 4.8 to conclude that
\[
\| \lambda - \hat{\lambda} \|_2 \leq O\left( \frac{\varepsilon_1 + \varepsilon_2}{k^2 \log(1/\delta) - \varepsilon_1} \right) = O\left( \frac{\varepsilon_1}{k^2 \log(k/\delta)} \right),
\]
so, possibly by modifying \( \varepsilon_1 \) to be \( \varepsilon_2/k^2 \log(1/\delta) \), we get that the estimates for the mixing weights are \( \varepsilon_2\)-close to the ground truth.  

Note that we can also amplify the success probability of TensorResolve by running Select from Section 4.2.

4.4 Approximating the OTF

**Algorithm 6 DFT**

1. **Input**: Error tolerance \( \eta > 0 \), sample access to \( \rho \), confidence parameter \( \beta > 0 \), frequencies \( \omega_1, \ldots, \omega_m \)
2. **Output**: With probability at least \( 1 - \beta \), numbers \( u_1, \ldots, u_m \) such that for each \( j \in [m], |u_j - \bar{\rho}(\omega_j)| \leq \eta \)
3. \( N \leftarrow O(\log(m/\beta)/\eta^2) \)
4. Draw samples \( x_1, \ldots, x_N \) from \( \mathbb{R}^d \)
5. For each \( j \in [m] \), compute \( u_j \leftarrow 1/N \sum_{i=1}^{N} \cos(2\pi \cdot \langle \omega_j, x_i \rangle) \)
6. Output \( u_1, \ldots, u_m \).

**Lemma 4.13.** For \( 0 < \beta < 1, \eta > 0 \), and \( \omega_1, \ldots, \omega_m \in \mathbb{R}^d \), \( DFT([\omega_i]_{i \in [m]}) \) draws \( N = O(\log(m/\beta)/\eta^2) \) samples and in time \( T = O(N \cdot m) \) outputs \( u_1, \ldots, u_m \) for which \( |u_j - \bar{\rho}(\omega_j)| \leq \eta \).

**Proof.** By a union bound, it suffices to show that for any \( j \) \in \([m], |u_j - \bar{\rho}(\omega_j)| \leq \eta \) with probability at least \( 1 - \beta/m \). Note that
\[
E[u_j] = E_{x \sim \rho}[\cos(2\pi \cdot \langle \omega_j, x \rangle)] = E[\Re \hat{\rho}(\omega_j)] = \bar{\rho}(\omega_j),
\]
where the last step follows by the fact that \( \bar{\rho} \) is real-valued (by circular symmetry). Furthermore, the summands in \( \sum_{i=1}^{N} \cos(2\pi \cdot \langle \omega_j, x_i \rangle) \) are \([-1, 1]-\)valued, so the lemma follows by Chernoff.

We now show that for general granularity \( \zeta > 0 \), the output of DFT still achieves error \( \eta + O(\zeta) \).

**Corollary 4.2.** For \( 0 < \beta < 1, \eta, \zeta > 0 \), and \( \omega_1, \ldots, \omega_m \in \mathbb{R}^d \), if \( DFT([\omega_i]_{i \in [m]}) \) draws \( N = O(\log(m/\beta)/\eta^2) \) samples of granularity \( \zeta \), then in time \( T = O(N \cdot m) \) it outputs numbers \( u_1, \ldots, u_m \) for which \( |u_j - \bar{\rho}(\omega_j)| \leq \eta + O(\zeta \cdot \left\| \omega_j \right\|_2) \).

**Proof.** Note that \( \cos(\cdot) \) is \( \alpha \)-Lipschitz for some \( \alpha < 3/4 \). This implies that for any \( \omega \in \mathbb{R}^d \), the function \( x \mapsto \cos(2\pi \langle x, \omega \rangle) \) is at most \( O(|\omega|) \)-Lipschitz with respect to \( \ell_2 \).

Take any collection of \( 0 \)-granular samples \( x_1', \ldots, x_N' \) for which the averages \( u_1', \ldots, u_m' \) computed by DFT would be \( \eta \)-accurate. If DFT were instead passed \( \zeta \)-granular samples \( x_1, \ldots, x_N \) for which \( \left\| x_i' - x_i \right\|_2 \leq \zeta \) for each \( i \in [N] \), then by triangle inequality, the averages \( u_1, \ldots, u_m \) computed by DFT with these samples would satisfy \( |u_j - u_j'| \leq \eta + O(\zeta \cdot \left\| \omega_j \right\|_2) \) for each \( j \in [m] \), as claimed.

We can finally complete the proof of Theorem 4.1.

**Proof of Theorem 4.1.** By Lemma 4.6, it suffices to produce an \( \eta \)-approximate, \( m \)-query OTO oracle for \( \eta \) defined in (12) and \( m = O(\log 1/\delta) \). By Corollary 4.2, this can be done using
\[
\log(1/\delta)/\eta^2 = O\left( \log(1/\delta) \cdot \poly(1/\min(1, \varepsilon_1, 1/\varepsilon_2, (6k/\Delta)^k) \right)
\]
samples of granularity \( \eta/2 \) with probability at least \( 1 - \delta \). Theorem 4.1 then follows by a union bound over the failure probabilities of LearnAiryDisks and DFT, and replacing \( 2\delta \) with \( \delta \) and absorbing constant factors. Finally, note that the dependence on \( R \) follows by the discussion at the end of Section 4.1.

**Proof of Theorem 4.2.** By Lemma 4.12, it suffices to produce an \( \eta \)-approximate, \( m \)-query OTO oracle for \( \eta \) defined in (16) and \( m = \Theta\left( \frac{k^2}{(\Delta - \| \omega \|)^2} \right) \). By Corollary 4.2, this can be done with probability \( 9/10 \) using
\[
\log(10m)/\eta^2 = O\left( \poly(k, 1/\Delta, 1/\min(1, \varepsilon_1, 1/\varepsilon_2, k, (\Delta - \| \omega \|) \wedge 1) \right)
\]
samples of granularity \( \eta/2 \) with probability at least \( 1 - \delta \). Theorem 4.1 then follows by a union bound over the failure probabilities of TensorResolveCorrect and DFT.

5 CONCLUSION AND OPEN PROBLEMS

We hope that our work will be a stepping-stone towards developing a rigorous theory of resolution limits in more sophisticated optical systems. The setting that we study, diffraction through a perfectly circular aperture under incoherent illumination, is arguably the most basic model one can study in Fourier optics. One can ask whether the techniques developed in this work can be pushed to answer questions about the following more challenging setting:

**Coherent illumination.** In the presence of light emanating from a single point source, the (complex-valued) amplitude of the electric field at a point \( P \) on the observation plane is proportional to \( e^{i\phi \cdot f_j(z)/|z|^2} \), where \( e^{i\phi} \) is some phase factor, \( z \) is the angular displacement of the point \( P \) from the optical axis, and \( \sigma \) is the spread parameter which depends on the wavelength of the light and the radius of the aperture. This means that the actual probability distribution over where on the observation plane a photon gets detected is proportional to the squared modulus of this, i.e. \( J_1(\phi/|z|^2) \). In this work, we assumed that in the presence of many point sources, the light emanating from the various point sources is incoherent. In other words, there is no interference introduced by the extra phase factors, and mathematically this translates to a probability distribution given by a nonnegative linear combination of the probability densities coming from the individual sources of light, and this is what gives rise to the mixture model we studied.

The **coherent setting** is quite different. Suppose that for point source \( j \), the extra phase factor in the electric field at any given point is \( e^{i\phi_j} \) for some \( \langle \omega_j \rangle \). Under coherent illumination from multiple point sources, it is the electric field which is a linear combination, namely of the electric fields associated to each individual point source. This gives rise to the following natural probabilistic model:
We thank Elchanan Mossel and Tim Roughgarden for helpful feedback on earlier versions of this work.

REFERENCES

[1] Ernst Abbe. 1873. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. Archiv für mikroskopische Anatomie 9, 1 (1873), 413–418.

[2] D. Achlioptas and F. McSherry. 2005. On Spectral Learning of Mixtures of Distributions. In Proceedings of the Eighteenth Annual Conference on Learning Theory (COLT). 458–469.

[3] George Biddell Airy. 1835. On the diffraction of an object-glass with circular aperture. Transactions of the Cambridge Philosophical Society 5 (1835), 283.

[4] S. Arora and R. Kannan. 2001. Learning mixtures of arbitrary Gaussians. In Proceedings of the 33rd Symposium on Theory of Computing, 247–257.

[5] Mikhail Belkin and Kanishk Sinha. 2015. Polynomial learning of distribution densities at specified by phases of well-separated gaussians. In Proceedings of the thiry-fourth annual ACM symposium on Theory of Computing, 755–766.

[6] Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. 2012. Nearly optimal sparse fourier transform. In Proceedings of the forty-fourth annual ACM symposium on Theory of computing, 563–578.

[7] Jeffrey J. Holt, Jeffrey D. Vaaler, et al. 1996. The Beurling-Selberg extremal functions for a ball in Euclidean space. Duke Mathematical Journal 83, 1 (1996), 203–248.

[8] Samuel B. Hopkins and Jerry Li. 2018. Mixture models, robustness, and sum of squares proofs. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, 1021–1034.

[9] William V. Houston. 1927. A compound interferometer for fine structure work. Physical Review 29, 3 (1927), 478.

[10] Mikhail Belkin and Kaushik Sinha. 2015. Polynomial learning of distribution densities at specified by phases of well-separated gaussians. In Proceedings of the thiry-fourth annual ACM symposium on Theory of Computing, 755–766.

[11] Sanjoy Dasgupta. 1999. Learning mixtures of Gaussians. In Proceedings of the thirty-fourth annual ACM symposium on Theory of Computing, 1047–1058.

[12] Animesh Vakarelov. 2016. Sparse fourier transform in any constant dimension via sum-of-squares. arXiv preprint arXiv:1607.03619 (2016).

[13] S. Vempala and G. Wang. 2002. A Spectral Algorithm for learning mixtures of spherical Gaussians. In Proceedings of the 43rd Annual Symposium on Foundations of Computer Science. 1047–1058.

[14] Jeffrey J. Holt, Jeffrey D. Vaaler, et al. 1996. The Beurling-Selberg extremal functions for a ball in Euclidean space. Duke Mathematical Journal 83, 1 (1996), 203–248.

[15] Piotr Indyk, Michael Kapralov, and Eric Price. 2014. (Nearly) sample-optimal sparse fourier transform. In Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. SIAM, 480–499.

[16] Yiyan Jin, Daoguo Liu, and Zhen Song. 2020. A robust multi-dimensional sparse Fourier transform in the continuous setting. arXiv preprint arXiv:2005.06156 (2020).

[17] Pravesh K. Khot and David Steurer. 2017. Outlier-robust moment-estimation via sum-of-squares. arXiv preprint arXiv:1711.11581 (2017).

[18] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. 2010. Efficiently learning mixtures of two Gaussians. In Proceedings of the forty-second ACM symposium on Theory of Computing, 553–562.

[19] Yoav Freund. 1997. An efficient algorithm for learning linear threshold functions. Information and Computation 129, 2 (1996), 87–95.

[20] Ruoyu An and Daniel M. Kane. 2016. Near-optimal sparse fourier transform. In Proceedings of the forty-seventh annual ACM symposium on Theory of computing, 563–578.

[21] Jeffrey J. Holt, Jeffrey D. Vaaler, et al. 1996. The Beurling-Selberg extremal functions for a ball in Euclidean space. Duke Mathematical Journal 83, 1 (1996), 203–248.

[22] Jeffrey J. Holt, Jeffrey D. Vaaler, et al. 1996. The Beurling-Selberg extremal functions for a ball in Euclidean space. Duke Mathematical Journal 83, 1 (1996), 203–248.

[23] S. Vempala and G. Wang. 2002. A Spectral Algorithm for learning mixtures of spherical Gaussians. In Proceedings of the 43rd Annual Symposium on Foundations of Computer Science. 1047–1058.