The sample complexity of multi-reference alignment*

Amelia Perry†, Jonathan Weed†, Afonso S. Bandeira‡, Philippe Rigollet†, and Amit Singer§

Abstract. The growing role of data-driven approaches to scientific discovery has unveiled a large class of models that involve latent transformations with a rigid algebraic constraint. Three-dimensional molecule reconstruction in Cryo-Electron Microscopy (cryo-EM) is a central problem in this class. Despite decades of algorithmic and software development, there is still little theoretical understanding of the sample complexity of this problem, that is, number of images required for 3-D reconstruction. Here we consider multi-reference alignment (MRA), a simple model that captures fundamental aspects of the statistical and algorithmic challenges arising in cryo-EM and related problems. In MRA, an unknown signal is subject to two types of corruption: a latent cyclic shift and the more traditional additive white noise. The goal is to recover the signal at a certain precision from independent samples. While at high signal-to-noise ratio (SNR), the number of observations needed to recover a generic signal is proportional to $\mathbf{1/SNR}$, we prove that it rises to a surprising $\mathbf{1/SNR^3}$ in the low SNR regime. This precise phenomenon was observed empirically more than twenty years ago for cryo-EM but has remained unexplained to date. Furthermore, our techniques can easily be extended to the heterogeneous MRA model where the samples come from a mixture of signals, as is often the case in applications such as cryo-EM, where molecules may have different conformations. This provides a first step towards a statistical theory for heterogeneous cryo-EM.

Key words. Multi-reference alignment, method of invariants, bispectrum, cryo-EM

AMS subject classifications.

1. Introduction.

Sample complexity is a concept at the cornerstone of statistics and machine learning with far reaching implications for experimental design and data collection strategies, ranging from polling voters for election prediction to training speech recognition systems. Loosely speaking, the sample complexity is the number of measurements needed to estimate model parameters at a prescribed accuracy. Perhaps the most fundamental question associated to sample complexity is its scaling with respect to signal-to-noise ratio (SNR) of the problem at hand. This question is of prime importance especially in modern problems arising in data-driven science that often feature a very low SNR.

In many traditional models, the sample complexity scales as 1/SNR, but this question

*Submitted to the editors September 10, 2018.

Funding: AP is supported in part by NSF grant DMS-1541100, NSF CAREER Award CCF-1453261 and a grant from the MIT NEC Corporation. JW is supported in part by NSF Graduate Research Fellowship DGE-1122374. ASB was partially supported by NSF grants DMS-1317308, DMS-1712730, and DMS-1719545. Part of this work was done while ASB was with the Mathematics Department at MIT. PR is supported in part by NSF grants CAREER DMS-1541099, DMS-1541100, DMS-1712596 and DMS-TRIPODS-1740751; DARPA grant W911NF-16-1-0551, ONR grant N00014-17-1-2147, a grant from the MIT NEC Corporation, grant 2018-182642 from the Chan Zuckerberg Initiative DAF and the MIT Skoltech Seed Fund. AS was partially supported by ARO grant W911NF-17-1-0512, Award Number R01GM090200 from the NIGMS, FA9550-17-1-0291 from AFOSR, Simons Investigator Award and Simons Collaboration on Algorithms and Geometry from Simons Foundation, and the Moore Foundation Data-Driven Discovery Investigator Award.

†Massachusetts Institute of Technology, Cambridge, MA (jweed@mit.edu, rigollet@math.mit.edu).

‡Courant Institute of Mathematical Sciences, New York University, New York, NY (bandeira@cims.nyu.edu).

§Princeton University, Princeton, NJ (amits@math.princeton.edu).

remains elusive in more complex models that feature latent variables in order to account for heterogeneity in the data. In this paper, we examine the sample complexity of such complex models in which the signal undergoes two types of corruption: a latent linear transformation and noise addition. Of particular interest in applications are linear transformations that correspond to a group action. For example, in estimating a two-dimensional image from multiple arbitrarily rotated noisy copies, every measurement corresponds to an unknown element of the group of planar rotations SO(2) that acts linearly on the data. Another example is the reconstruction problem in cryo-EM [16], a fundamental imaging technique that won the 2017 Nobel Prize in Chemistry. In cryo-EM, the goal is to estimate the three-dimensional structure of a molecule from many two-dimensional noisy tomographic projection images taken at unknown viewing angles. Here to every projection image corresponds an unknown element of the 3D rotation group SO(3) and the linear transformation is a composition of a tomographic projection in a fixed direction with the group action of rotating the molecular structure (we ignore possible in-plane translations and other imaging effects). Other estimation problems of similar nature arise in many other scientific and engineering disciplines, such as structure from motion (SfM) in computer vision [3], simultaneous localization and mapping (SLAM) in robotics [28], X-ray free electron lasers (XFEL) in structural biology [12, 17], crystalline simulations [33], and shape matching and image registration and alignment problems arising in geology, medicine, and paleontology, to name a few [14, 15, 30].

Multi-reference alignment (MRA) [6] is one of the simplest models that is able to capture fundamental aspects of this class of problems, rendering it ideal for theoretical study. In this model one observes n independent data points $y_1, \dots, y_n$ given by

$$y_i = R_{\ell_i}\theta + \sigma\xi_i, \tag{1.1}$$

where R_{ℓ_i} is a cyclic shift by an unknown number ℓ_i of coordinates: the jth coordinate of $R_{\ell_i}\theta \in \mathbb{R}^d$ is given by $\big(R_{\ell_i}\theta\big)_j = \theta_{j+\ell_i \pmod d}$. We assume isotropic Gaussian noise $\xi_i \sim \mathcal{N}(0, I_d)$ i.i.d. and independent of $\ell_1, \dots, \ell_n$. We make no assumptions on the shifts $\ell_1, \dots, \ell_n$; however, by applying an independent and uniform random cyclic shift to each observation, we can always reduce the MRA model to the case where $\ell_1, \dots, \ell_n$ are drawn i.i.d. uniformly from $[d]$. We therefore focus on this case for simplicity and generality. The goal is to estimate the unknown vector $\theta \in \mathbb{R}^d$.

The MRA model is illustrated in Figure 1. We refer to $\|\theta\|_2^2/\sigma^2$ as the SNR; without loss of generality we assume in the sequel that $\|\theta\|_2 = 1$, implying $\mathrm{SNR} = 1/\sigma^2$. The latent transformations R_ℓ in MRA correspond to the action of the cyclic group $\mathbb{Z}/d\mathbb{Z}$ on real-valued signals of length d. The simplicity of MRA in the class of problems mentioned earlier stems from the following facts: (i) the group $\mathbb{Z}/d\mathbb{Z}$ is finite (has exactly d elements) and commutative (i.e., $R_\ell R_m = R_m R_\ell$ for all ℓ, m), and (ii) no further linear operation (such as projection as in cryo-EM) is involved.

In this paper, we study the sample complexity of MRA, that is, the number of observations needed to recover a generic signal with a given accuracy as a function of the SNR. Our results reveal a striking difference between the high and low SNR regimes. On the one hand, the picture at a high SNR is fairly standard in signal processing: the sample complexity scales proportionally to 1/SNR. On the other hand, using information theoretic arguments, we show that the presence of the latent cyclic shifts has a profound effect on the sample complexity at low SNR, where the optimal sample complexity becomes proportional to $1/\mathrm{SNR}^3$. Twenty years ago, in a seminal paper by Sigworth [31] that introduced maximum likelihood estimation to the cryo-EM field, an analogous phenomenon was empirically observed (without theoretical explanation) in two-dimensional multi-reference alignment (Figure 2), where the group of transformations are planar rigid motions. Our results shed light on the fundamental reasons behind this behavior of the sample complexity.

More specifically, our results on the sample complexity of MRA highlight the role of the *third moment tensor* (known in signal processing as the *bispectrum*) in the estimation task. From this analysis, we not only show that the $1/\mathrm{SNR}^3$ dependence is unavoidable for any method in the low-SNR regime, but also give a very simple algorithm based on the third moment tensor, which achieves the optimal sample complexity efficiently and provably.

By establishing the correct sample complexity for the MRA model, this work represents the first step towards determining the sample complexity of the reconstruction problem in cryo-EM and other applications involving more complicated group actions. In fact, we complement our results on MRA by showing that a simple extension of our algorithm applies to the heterogeneous case where θ in (1.1) is randomly drawn from a finite family of linearly independent vectors. Using ideas initiated in the present paper, follow-up work [5] has confirmed that similar phenomena arise for molecule reconstruction in cryo-EM, at least in a slightly weaker sense than the one presented in this paper.

2. Overview.

In this section, we give an overview of our contributions and how they fit in the existing literature.

2.1. Existing methods.

The difficulty of the multi-reference alignment problem resides in the fact that both the signal $\theta \in \mathbb{R}^d$ and the shifts $\ell_1, \ldots, \ell_n \in \mathbb{Z}_d$ are unknown. If the shifts were known, one could easily estimate θ by taking the average of $R_{\ell_i}^{-1} y_i, i = 1, \ldots, n$. In fact, this simple observation is the basis of the so-called "synchronization" approach [32, 6, 7]: first estimate the shifts by $\tilde{\ell}_1, \ldots, \tilde{\ell}_n \in \mathbb{Z}_d$ and then estimate θ by averaging the $R_{\tilde{\ell}_i}^{-1} y_i$'s. While the synchronization approach can be employed at high SNR, it is limited by the fact that at low SNR, even alignment of observations to the true signal yields inaccurate shift estimates [4].

Instead, we take a different approach that exploits the connection between MRA and Gaussian mixture models. This connection is based on the fact that in MRA, the data y is distributed according to a uniform mixture of Gaussians whose centers are the rotated vectors $R_1\theta, \ldots, R_d\theta$. To analyze MRA, we therefore rely on techniques from the Gaussian mixture model literature. One insight from this literature, which is crucial to our work, is that there are two separate estimation problems that can be posed for Gaussian mixture models. The first is *clustering*, in which the goal is to assign a label to each datapoint corresponding to the Gaussian from which it was drawn. The second is *parameter estimation*, in which the goal is simply to learn the Gaussians themselves—i.e., to identify the mean vectors and covariance matrices of each Gaussian component, without necessarily assigning a label to each point. Previous theoretical work on MRA has focused on the first task, which forms the basis for the synchronization approach. By contrast, our approach is based on the second task: we seek only to estimate the underlying parameters of the mixture; as they correspond to the underlying signal of interest. This connection also motivates our theoretical approach: we develop an approach based on the *method of moments*, which was introduced in Pearson's

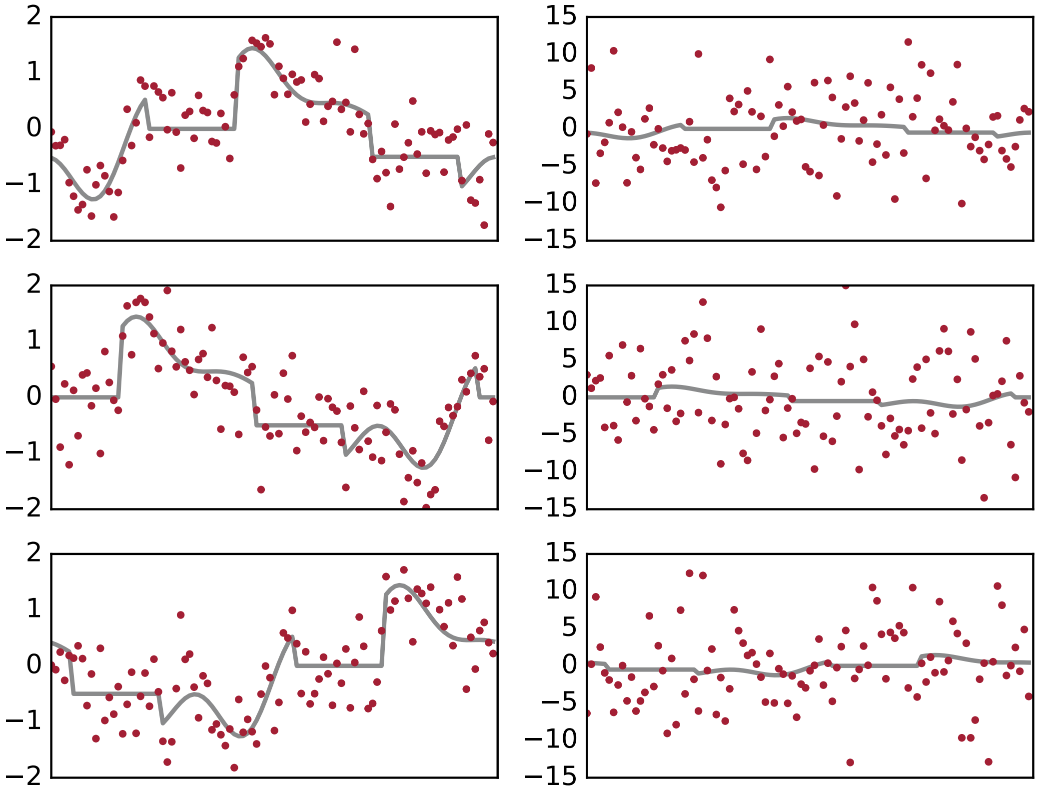

Figure 1. *Instances of the multi-reference alignment problem, at low ($\sigma \approx .5$, left column) and high ($\sigma \approx 3$, right column) noise levels. We plot the values of a vector in $\mathbb{R}^d$ for $d = 100$. Randomly shifted copies of a smoothed version of the underlying signal (θ) appears in gray, and a smoothed version of the noisy observation (y) appears in red. When the noise level is low, salient features of the signal are still visible despite the noise; in the presence of large noise, however, the signals cannot reliably be aligned. We establish the optimal sample complexity of the large noise problem.*

seminal paper on Gaussian mixture models [27] and has recently led to efficient estimators with provably optimal guarantees [26].

2.2. The method of invariants. In this work, we develop a new approach to MRA based on the method of moments. This method focuses on the tensors $T^{(r)}(\theta)$ defined by

$$T^{(r)}(\theta) := \frac{1}{d}\sum_{\ell=1}^{d}(R_\ell \theta)^{\otimes r}\,. \tag{2.1}$$

These tensors are precisely the moments of the uniform distribution over the set of vectors $\{R_1\theta, \dots, R_d\theta\}$. We first establish that the parameter θ can be identified by the moment tensors $\{T^{(r)}(\theta)\}_{r\geq 1}$. We then show that we can estimate the moment tensors accurately enough to recover the original signal.

The method of moments has an alternate interpretation in the context of MRA and similar problems involving group actions. One striking fact is that the moment tensors in

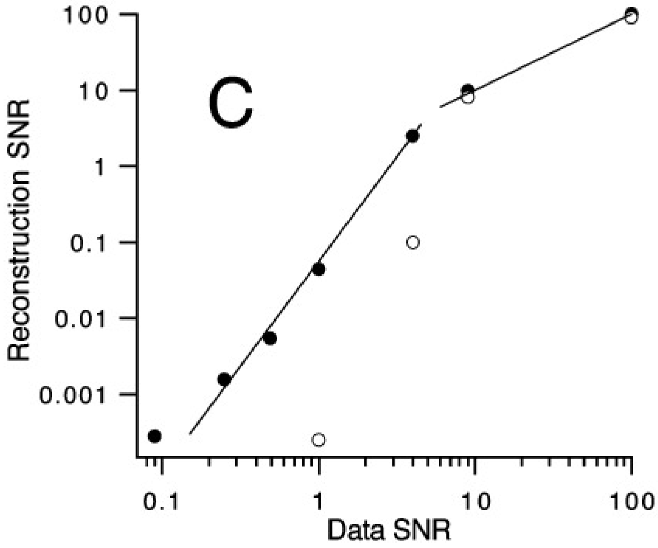

Figure 2. *Figure taken from a paper on cryo-EM [31], illustrating (i) strikingly different behavior for the maximum likelihood estimator in low and high SNR regimes and (ii)* $1/\mathrm{SNR}^3$ *scaling at low SNR. Our theoretical analysis suggests that any estimator, not only the maximum likelihood estimator, is bound to the same limitations. Reprinted from* Journal of Structural Biology*, Vol. 122, F. Sigworth, A maximum-likelihood approach to single-particle image refinement, pp. 328–339, copyright 1998, used with permission.*

MRA capture features of the signal that are *invariant* under cyclic shifts. For example, the first moment tensor reduces to the entrywise mean of the signal (i.e., the vector in $\mathbb{R}^d$ each of whose entries is the average value of θ), which is an example of an invariant feature: it is clearly invariant under cyclic shifts of θ and, as we show below, can easily be estimated consistently in the MRA model. More generally, each entry in the moment tensor $T^{(r)}(\theta)$ is an invariant polynomial in the coordinates of θ, and these invariant polynomials can always be estimated in the MRA model as long as σ is known. We therefore call our approach the *method of invariants*.

In what follows, we focus on the moment tensors $T^{(r)}(\theta)$ for $r \leq 3$. Our core contribution is to show that estimation on the basis of these first three moment tensors yields optimal sample complexity as a function of the SNR for the MRA model. We stress, however, that the focus on moment tensors is not limited to MRA, and that the method of invariants can be used to obtain sample complexity bounds for a wide variety of similar models. In this work, we specialize to MRA since it provides perhaps the simplest nontrivial application of these ideas.

The first three moment tensors in the MRA model correspond to quantities often studied under different names in the signal processing literature. In addition to the first moment tensor—which, as noted above, reduces to the entrywise mean of the signal—the second and third moment tensors are also easy to describe in terms of θ. The second moment tensor $T_2(\theta)$ corresponds to the autocorrelation of the signal θ. Written in the Fourier basis, this autocorrelation corresponds to the power spectrum of θ (the square of the absolute value of the Fourier coefficients of the signal) which is often used as an invariant feature in signal processing. Note that, in general, this quantity does not carry enough information to allow for estimation of θ, since it provides only the magnitudes of the Fourier coefficients, but not their phases.

The crucial object in the case of MRA is the third moment tensor. Written in the Fourier basis, this object is known as the *bispectrum* of the signal, given by

$$\mathcal{B}(k_1, k_2) = \hat{\theta}_{k_1}\hat{\theta}_{k_2}\hat{\theta}_{-k_1-k_2},$$

where $\hat{\theta}$ is the Fourier transform of θ, $k_1, k_2 \in [d]$, and the indices are taken modulo d. It was originally introduced in a statistical context [11, 35] . It is known [21] that the bispectrum uniquely determines the signal θ up to cyclic shift whenever $\hat{\theta}_k \neq 0$ for all $k \in [d]$. We call such signals *generic*. In other words, for generic signals, the moment tensors $T_1(\theta)$, $T_2(\theta)$, and $T^{(3)}(\theta)$ suffice to identify the true signal θ. This fact has been exploited to obtain estimates for alignment problems [29, 18, 10].

Note that the sample average based estimator for $T^{(3)}$ has a variance of order σ^6/n when σ is large, since it is a cubic polynomial of noisy data. It suggests that in the low SNR regime, any approach relying on the bispectrum requires at least order $1/\mathrm{SNR}^3$ samples. Since this dependency on SNR is very different from the $1/\mathrm{SNR}$ sample complexity of many models, bispectrum approaches seem highly suboptimal.

The main contribution of our work is to show that this number of samples is in fact a *fundamental* requirement of the problem when the shifts are sampled from the uniform distribution, independent of the approach taken (following ideas developed in [8]): all estimators suffer from the same limitations, including the maximum likelihood estimator (see Figure 2). This shows that the latent cyclic transformations fundamentally change the difficulty of the problem. A similar phenomenon has been demonstrated for a Boolean version of MRA [2].

To complement our lower bound, we also propose simple algorithm based on the method of moments capable of provably achieving the optimal $1/\mathrm{SNR}^3$ sample complexity for generic signals. While other algorithms employing the bispectrum exist in the literature [29, 18, 10], ours has the virtue of acting directly to decompose the third moment tensor via a straightforward and principled approach. As we note below, this simple algorithm also extends to the heterogenous setting, for which no algorithms enjoyed theoretical guarantees prior to this work.

2.3. Non-generic signals. The bispectrum-based methods for the multi-reference alignment problem we present work only for generic signals. In fact, non-generic signals can exhibit significantly worse behavior with a sample complexity of order $1/\mathrm{SNR}^d$ rather than $1/\mathrm{SNR}^3$, but this pessimistic scenario does not seem to be representative of signals encountered in practice. In fact, these signals that are hard to estimate form a set of zero Lebesgue measure. See [8] for more details.

2.4. The heterogeneity problem. One of the main challenges in cryo-EM reconstruction is the heterogeneity problem, where one observes noisy projection images of multiple unknown conformations of the same molecule. The MRA model can be extended to accommodate heterogeneity by assuming that in (1.1), the vector $\theta \in \mathbb{R}^d$ is also a latent variable drawn from a finite set of unknown vectors $\mathcal{C} = \{\theta^{(1)}, \dots, \theta^{(K)}\}$. The goal here is to recover the set $\mathcal{C}$ up to a cyclic shift and the proportion of each $\theta^{(j)}$.

Our approach based on the method of invariants coupled with tensor decomposition techniques extends to the heterogeneous setup. It yields the first algorithm capable of provably solving the heterogeneous MRA at arbitrarily low SNR, albeit at a potentially suboptimal sample complexity of $1/\mathrm{SNR}^5$.

2.5. Connections to cryo-EM and XFEL. One of the main motivations to study the multi-reference alignment problem is that it serves as a simpler surrogate for cryo-EM. This paper indicates potentially fruitful directions for future work. Our results offer theoretical support for the use of invariant methods in cryo-EM, a proposal which dates back to Zvi Kam [22]. These methods have also proven effective in XFEL structure determination [36, 13].

Our work serves as a first step towards a complete statistical theory of cryo-EM. In fact, follow-up work to this paper has demonstrated that the method of invariants can be used to characterize the sample complexity of more general models, including the reconstruction problem for cryo-EM [5].

2.6. Notation. We use $[d]$ to represent the set $\{1, \dots, d\}$ and I_d to represent the $d \times d$ identity matrix. The smallest and largest singular values of a matrix are denoted $\sigma_{\min}$ and $\sigma_{\max}$, respectively. The symbol $\mathrm{poly}(\cdot)$ refers to an unspecified polynomial with constant coefficients. C_d is used to refer to a constant that may depend on d but not on other parameters, and it may refer to a different constant in different appearances throughout the text. The expression $f(n) = O(g(n))$ means that there exists a constant C such that $f(n) \le Cg(n)$ for all n, and we write $O_d(g(n))$ when the constant may depend on d. We write $g(n) = \Omega(f(n))$ when $f(n) = O(g(n))$.

3. Fundamental limitations.

In this section, we establish the fundamental limits of MRA and point to shortcomings of existing strategies to achieve optimal sample complexity.

3.1. Lower bounds for sample complexity. Since observations in the MRA model (1.1) are invariant under a global cyclic shift, one may only identify θ up to such a global shift. To account for this fact, it is natural to employ the following shift-invariant distance between vectors $\theta, \tau \in \mathbb{R}^d$:

$$\rho(\theta, \tau) = \min_{\ell \in \mathbb{Z}_d} \|\theta - R_\ell \tau\|_2 .$$

As noted above, by applying an independent and uniform random cyclic shift to each observation, we can always reduce the MRA model to the case where $\ell_1, \dots, \ell_n$ are drawn i.i.d. uniformly from $[d]$. In this case, the distribution of y in (1.1) is a uniform mixture of the d Gaussian distributions $\mathcal{N}(\theta, \sigma^2 I_d), \dots, \mathcal{N}(R_{d-1}\theta, \sigma^2 I_d)$. If y is generated according to this distribution, we call it a "sample from MRA with signal θ." The statistical properties of this Gaussian mixture are analyzed in [8].

If σ is small—that is, if the SNR is sufficiently large—then the signals can be aligned (for example, via the synchronization approach [32]), and therefore θ can be estimated accurately on the basis of n samples from MRA with signal θ as long as $n \ge C/\mathrm{SNR}$ for some constant C. This is the same dependence that would be expected in the absence of shifts. Strikingly, the situation in the high-noise regime (when the SNR is low) is very different: estimation is impossible unless $n \geq C/\mathrm{SNR}^3$ for some constant C.

Theorem 3.1. *Fix $d > 2$, $\varepsilon > 0$ sufficiently small, and $\sigma \geq 1$. There exists a universal constant C such that the following holds with constant probability: for any estimator $\tilde{\theta}$ based on n samples from (1.1) there exists a generic signal $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 = 1$, such that $\rho(\tilde{\theta}, \theta) \geq \varepsilon$ whenever $n \leq C\sigma^6\varepsilon^{-2}$.*

In other words, if we require that our estimator $\tilde{\theta}$ satisfy $\rho(\tilde{\theta}, \theta) < \varepsilon$ with probability close to 1, then we must have $n \geq C\sigma^6\varepsilon^{-2}$. We prove this fact in the supplement using the tight information-theoretic bounds developed in [8], which are based on the method of invariants and, in particular, on the observation given above that the second moment tensor does not carry enough information about θ in general.

3.2. The importance of high frequencies. As noted above, the sample complexity exhibited by the method of invariants, $1/\mathrm{SNR}^3$, is tight for generic signals. For non-generic signals, while the method of invariants still yields optimal results (see [5]), the precise sample complexity depends on specific properties of the support of the Fourier transform of the original signal. This dependence is often counter-intuitive as illustrated by the example below.

Some approaches to the alignment problem implicitly adopt a strategy of first estimating low frequencies of a signal, and then using this initial estimate to estimate higher frequencies (see [9]). In other words, these strategies assume that estimating a low-pass version of a signal is no harder than estimating the original signal.

Surprisingly, this is *not* the case in general, as following example shows. Let us take $d \geq 14$ congruent to 2 (mod 4) and $\theta \in \mathbb{R}^d$ a signal whose Fourier transform $\hat{\theta}$ satisfies $\hat{\theta}_1 = \hat{\theta}_{-1} = 0$ but otherwise has full support. We show in the supplement that we can estimate θ with $O_d\left(1/\mathrm{SNR}^3\right)$ samples. Surprisingly, if we low-pass θ by setting $\hat{\theta}_j = 0$ for all $|j| > 4$, then $\Omega\left(1/\mathrm{SNR}^4\right)$ samples are needed.

To show that θ can be recovered with $O_d(1/\mathrm{SNR}^3)$ samples, it suffices to show that the phases of the Fourier coefficients of θ can be reconstructed uniquely from its bispectrum. Given a complex number z, denote by $\arg(z)$ its phase. By applying a circular shift, we can assume without loss of generality that $\arg(\hat{\theta}_2) \in [0, 4\pi/d)$ and that $\arg(\hat{\theta}^{(3)}) \in [0, \pi)$. It is easy to check that the identity $2\sum_{k=2}^{(d-6)/4} \arg(\mathcal{B}(2, 2k)) + \arg(\mathcal{B}((d-2)/2, (d-2)/2)) = \frac{d}{2}\arg(\hat{\theta}_2)$ holds modulo 2π, and the assumption that $\arg(\hat{\theta}_2) \in [0, 4\pi/d)$ implies that the choice of $\arg(\hat{\theta}_2)$ is unique. This implies that all even-indexed phases can be recovered. We also have the simple identity $\arg(\hat{\theta}_6) + \arg(\mathcal{B}(3, 3)) = 2\arg(\hat{\theta}^{(3)})$ modulo 2π, and the assumption that $\arg(\hat{\theta}^{(3)}) \in [0, \pi)$ implies that the choice is unique. Combined with the knowledge of $\arg(\hat{\theta}_2)$, this implies recoverability of all odd-indexed phases.

To show that the low-pass signals require $n \geq C\sigma^8$ samples, we simply note that the first three moment tensors of the low-pass signals agree. Theorem A.1 therefore implies that the Kullback-Leibler divergence between the relevant distributions is at most $C\sigma^{-8}$. The same argument given in proof of Theorem 3.1 establishes that any test attempting to distinguish between the low-pass signals incurs type-I and type-II error of at least $2/3$ unless $n \geq C\sigma^8$.

The difficulty in recovering the low-pass version arises from the following simple observation: if $\hat{\theta}_j = 0$ for all $j \notin \{\pm 2, \pm 3, \pm 4\}$, then the only nonzero entry of the bispectrum is

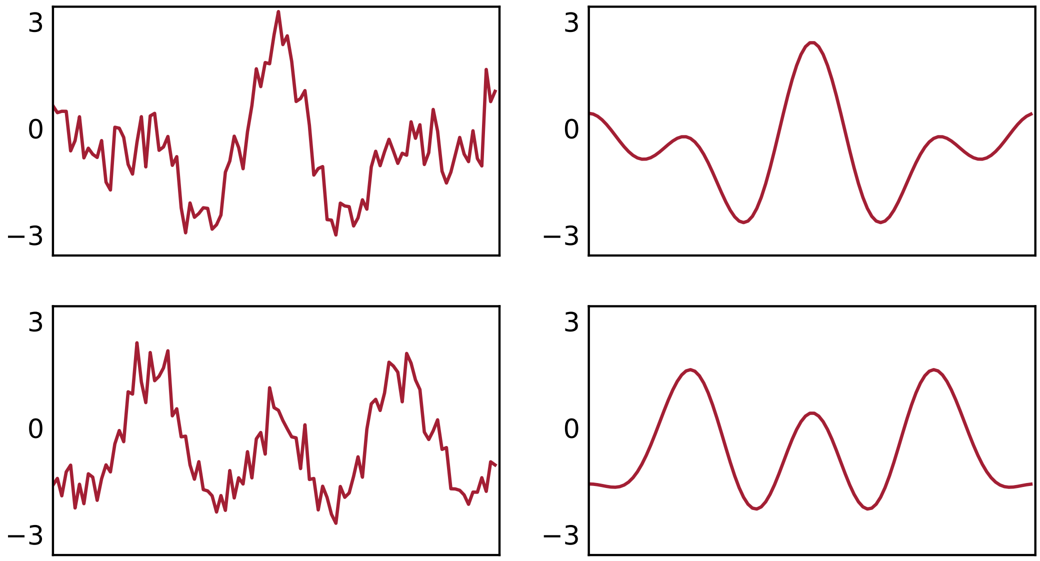

Figure 3. *Two signals whose Fourier transforms have almost full support (left column) and their corresponding low-pass versions (right column). Estimating either of the original signals is possible with $O_d(1/\mathrm{SNR}^3)$ samples. However, the same task for the low-pass versions requires $\Omega(1/\mathrm{SNR}^4)$ samples; in fact, even distinguishing between the two options requires this number of samples. This illustrates the importance of high frequencies in the MRA model.*

$\mathcal{B}(2,2)$. This implies that the bispectrum carries no information about the phase of $\hat{\theta}^{(3)}$ and we show in the supplement that this implies that any two such signals which agree on their second and fourth Fourier coefficient are indistinguishable—in the sense that any procedure to distinguish them fails with constant probability—unless $n \geq C\sigma^8$.

4. Efficient recovery via tensor methods.

Theorem 3.1 implies that the sample complexity of MRA for generic signals is at least $1/\mathrm{SNR}^3$ in the low-SNR regime. In this section, we describe how the method of invariants also yields an efficient algorithm that achieves this optimal sample complexity, that is, it outputs an estimator $\tilde{\theta}$ of θ such that $\rho(\tilde{\theta}, \theta) \leq \varepsilon$ with high probability whenever $n \geq C_d \sigma^6 \varepsilon^{-2}$.

Our approach uses the method of invariants by estimating invariant features in the third moment tensor $T^{(3)}$ defined in (2.1). While other algorithms in the literature have also been based on recovering the signal on the basis of the third moment tensor via iterative methods [29, 18, 10], we propose a simpler procedure which also yields stable recovery guarantees.

First, we estimate $T^{(3)}$ by the following empirical quantity:

$$\tilde{T}_n^{(3)} = \frac{1}{dn} \sum_{i=1}^{n} \sum_{j=1}^{d} ((R_j y_i)^{\otimes 3} - 3\,\mathrm{sym}(R_j y_i \otimes I_d)) \tag{4.1}$$

where

$$\mathrm{sym}(A)_{a_1 \ldots a_k} = \frac{1}{k!} \sum_{\pi \in \mathcal{S}_k} A_{\pi(a_1)\ldots\pi(a_k)}\,. \tag{4.2}$$

Lemma 4.1. *The estimator $\tilde{T}^{(3)}$ is an unbiased estimator of $T^{(3)}$. Moreover, each entry of $\tilde{T}_n^{(3)}$ has a variance of order σ^6/n so long as σ is bounded below by a positive constant.*

Proof. If $\xi_i \sim \mathcal{N}(0, I_d)$, then both $\mathbb{E}[\xi_i]$ and $\mathbb{E}[\xi_i^{\otimes 3}]$ are zero. This implies that

$$\mathbb{E}_{ij}[(R_j y_i)^{\otimes 3}] = \mathbb{E}_{ij}[(R_j\theta + \sigma\xi)^{\otimes 3}] = \mathbb{E}_j(R_j\theta)^{\otimes 3} + 3\,\mathrm{sym}((\mathbb{E}_j R_j\theta) \otimes I_d)\,,$$

so $\tilde{T}_n^{(3)}$ is an unbiased estimator of $T^{(3)}$.

Each entry of y_i is a Gaussian with variance σ^2, so the entries of $\mathrm{sym}(y_i \otimes I_d)$ have variance of order σ^2, and the entries of $y_i^{\otimes 3}$ have variance of order σ^6; the latter dominates for σ bounded away from 0. The claim follows. ∎

Then, we apply a basic decomposition technique, given in the next section, to the tensor $\tilde{T}_n^{(3)}$ to find a vector $\tilde{\theta}$ such that

$$\tilde{T}_n^{(3)} \approx \frac{1}{d}\sum_{\ell=1}^{d}(R_\ell\tilde{\theta})^{\otimes 3}\,.$$

The vector $\tilde{\theta}$ then serves as our estimate of θ.

4.1. Jennrich's Algorithm for Tensor Decomposition.

In this section, we detail a simple decomposition algorithm for the third moment tensor, which in turn provides an efficient algorithm that provably solves MRA for generic signals while achieving optimal sample complexity in terms of SNR. It involves the spectral decomposition of the tensor of empirical third moments. Such decompositions have been long studied and a sophisticated machinery has been developed over the years; see [25, Chapter 3].

The specific algorithm that we use is a standard tensor decomposition algorithm known as *Jennrich's algorithm* (proposed in [20] and credited to Robert Jennrich). The version described below allows the recovery of vectors $u_1, \ldots, u_r$ (up to simple transformations) from a noisy version of the tensor

$$T = \sum_{j=1}^{r} u_j \otimes u_j \otimes v_j \in \mathbb{R}^{m\times m\times p}, \tag{4.3}$$

where $v_1, \ldots, v_r \in \mathbb{R}^p$ are arbitrary nonzero vectors.

Jennrich's Algorithm ([20, 24]).
Input: Tensor $T \approx \sum_{j=1}^r u_j \otimes u_j \otimes v_j \in \mathbb{R}^{m\times m\times p}$.
Output: Matrix $U = [\hat{u}_1, \dots, \hat{u}_r] \in \mathbb{R}^{m\times r}$

- Choose random unit vectors $a, b \in \mathbb{R}^p$, and form matrices $A, B \in \mathbb{R}^{m\times m}$ with entries:
$$A_{ij} = \sum_k T_{ijk} a_k,\ A = \sum_{j=1}^r \langle v_j, a\rangle u_j \otimes u_j,$$
$$B_{ij} = \sum_k T_{ijk} b_k,\ B = \sum_{j=1}^r \langle v_j, b\rangle u_j \otimes u_j$$
- Let W be the matrix whose columns are the first r left singular vectors of A.
- Compute $M = W^\top AW(W^\top BW)^{-1}$.
- Output $U = WP$, where $M = PDP^{-1}$ is the eigendecomposition of M.

Jennrich's algorithm requires only basic matrix operations and can therefore be implemented very efficiently even on large scale problems. It also enjoys the following robustness guarantees. Using the notation of Jennrich's algorithm, it is easy to see that $T^{(3)}$ is indeed a low-rank tensor of the form (4.3), with $m = p = d$, $u_j = v_j = R_{j-1}\theta$ (for $j = 1, \dots, r$) and $U = [\theta, R_1\theta, \dots, R_{d-1}\theta]$. We recall the following recovery guarantee of Jennrich's algorithm when applied to a tensor $\tilde{T}$ that is close to a low rank tensor.

Theorem 4.2 ([19], Theorem 5.2). *Let T be a tensor of the form (4.3) with all u_j linearly independent, and define $\kappa(U) = \sigma_{\max}(U)/\sigma_{\min}(U)$. Moreover, fix $\varepsilon > 0$ and let $\tilde{T}$ satisfy $\|\tilde{T} - T\|_F \le \varepsilon$. Then Jennrich's algorithm applied to $\tilde{T}$ returns unit vectors $\tilde{u}_j, j = 1, \dots, r$ such that there exists a permutation π and scalars β_j satisfying*

$$\max_{j\in[r]} \|\tilde{u}_j - \beta_j u_{\pi(j)}\|_\infty \le \varepsilon\, \mathrm{poly}(m, \kappa) \tag{4.4}$$

with high probability.

Let $\tilde{u}_1$ be the first vector output by Jennrich's algorihtm applied to $\tilde{T}_n^{(3)}$ and let

$$\tilde{\beta}_1 = \tilde{u}_1^\top \mathbf{1}/\tilde{\mu}\,, \quad \tilde{\mu} = \frac{1}{n}\sum_{i=1}^n y_i^\top \mathbf{1}\,, \quad \tilde{\theta} = \tilde{u}_1/\tilde{\beta}_1$$

In the supplement, we show that an algorithm homoJen based on Jennrich's algorithm for tensor decomposition applied to the $\tilde{T}_n^{(3)}$ enjoys the following theoretical guarantees.

Theorem 4.3. *Fix $\sigma > .1$ and $\delta \in (0, 1)$ and assume $.1 \le \|\theta\|_2 \le 10$. Then, for any $\varepsilon > 0$ Jennrich's algorithm applied to $\tilde{T}_n^{(3)}$ outputs $\tilde{\theta}_n$ such that $\rho(\tilde{\theta}_n, \theta) \le \varepsilon$ with probability at least $1 - \delta$ whenever*
$$n \ge \sigma^6 \varepsilon^{-2}\, \mathrm{poly}(d, 1/\min_{j\in[d]} |\hat{\theta}_j|, 1/\delta)$$
in time $O(nd^3 + d^3\, \mathrm{poly}(\log(1/\varepsilon)))$.

Note that the constants .1 and 10 are arbitrary and may be replaced by any other constants. The time complexity is dominated by the time necessary to construct the empirical tensor

$\tilde{T}_n^{(3)}$, which requires only a single pass over the data. Since Jennrich's algorithm relies on basic matrix operations, it requires only $O(d^3 \operatorname{poly}(\log(1/\varepsilon)))$ additional computation time once $\tilde{T}_n^{(3)}$ has been constructed.

In view of the lower bound appearing in Theorem 3.1, the sample complexity of the modified Jennrich algorithm is optimal in terms of the SNR.

Several other bispectrum-based algorithms have appeared in the literature; see [10] for a recent empirical study. These may perform better in practice, but they largely do not come with the theoretical guarantees of the algorithm proposed here, and they do not yield efficient algorithms for the heterogenous case discussed below.

5. Heterogeneity.

In this section, we sketch an extension of the previous results to the heterogenous multi-reference alignment problem. We recall this model here for completeness. In heterogenous MRA, we observe

$$y_i = R_{\ell_i}\theta^{(Z_i)} + \sigma\xi_i\,, \quad i = 1, \ldots, n\,, \tag{5.1}$$

where $Z_1, \ldots, Z_n \in \{1, \ldots, K\}$ are i.i.d. latent variables such that $\Pr(Z_i = k) = \pi_k, k \in [K]$ that are independent of all other variables and $\theta^{(k)} \in \mathbb{R}^d, k \in [K]$ are unknown vectors. The other variables are specified as in the homogeneous model (1.1). The goal here is to recover the set of vectors $\theta^{(k)} \in \mathbb{R}^d, k = 1, \ldots, K$ up to a cyclic shift and the probability mass function $\{\pi_k\}_{k\in[K]}$.

The method of invariants described above can be extended to handle the heterogeneous model (5.1). In this case our method proceeds by estimating the mixtures of signals from an unbiased estimator $\tilde{T}_n^{(5)}$ for the 5-tensor

$$T^{(5)} = \sum_{k=1}^{K}\sum_{\ell=1}^{d} \frac{\pi_k}{d}\big(R_\ell\theta^{(k)}\big)^{\otimes 5}.$$

In our analysis of homogenous MRA, we noted that the moment tensors $T^{(1)} = T^{(1)}(\theta), T^{(2)} = T^{(2)}(\theta), T^{(3)} = T^{(3)}(\theta)$ uniquely determine θ, as long as θ is generic. The method of invariants can also be applied to the heterogenous case to show that the moment tensors $T^{(1)}, \ldots, T^{(5)}$ determine the vectors $\theta^{(1)}, \ldots, \theta^{(K)}$ as long as the vectors satisfy a particular genericity condition. Our proof of this fact is *algorithmic* in the sense that we exhibit an efficient algorithm, which can recover the vectors $\theta^{(1)}, \ldots, \theta^{(K)}$ as long as the collection is suitably generic. The fact that the method of invariants can be extended to the heterogenous case supports the idea that it is a flexible, general approach to models of this kind. This algorithm achieves sample complexity $1/\mathrm{SNR}^5$, whereas the optimal sample complexity for heterogenous MRA is known to be $1/\mathrm{SNR}^3$ in several settings, including when $\theta^{(1)}, \ldots, \theta^{(K)}$ are drawn independently from $\mathcal{N}(0, I/d)$ [5, 37]. Our algorithm therefore does not achieve optimal sample complexity in general; however, it is the first efficient algorithm for the heterogenous problem with provable guarantees.

We now sketch the basic idea of our approach. Using manipulations similar to the ones arising in the proof of Lemma 4.1, it is not hard to show that $T^{(5)}$ can be rewritten as

$$T^{(5)} = \mathbb{E}[y_i^{\otimes 5}] - 10\sigma^2 \operatorname{sym}(T^{(3)} \otimes I_d) - 15\sigma^4 \operatorname{sym}(T_1 \otimes I_d^{\otimes 2})$$

where $T_1 = \mathbb{E}[y_i]$ and $T^{(3)}$ is defined in (2.1). Therefore an unbiased estimator of $T^{(5)}$ is given by

$$\tilde{T}_n^{(5)} = \frac{1}{n}\sum_{i=1}^n y_i^{\otimes 5} - 10\sigma^2 \operatorname{sym}(\tilde{T}_n^{(3)} \otimes I_d) - 15\sigma^4 \operatorname{sym}(\frac{1}{n}\sum_{i=1}^n y_i \otimes I_d^{\otimes 2}),$$

where $\tilde{T}_n^{(3)}$ is given in (4.1). Moreover, each entry of $\tilde{T}_n^{(5)}$ has variance of order σ^{10}/n so long as σ is bounded below by a positive constant.

We propose a method that consists in applying Jennrich's algorithm to an appropriate flattening of $\tilde{T}_n^{(5)}$. We call it heteroJen. It hinges on the following observation: the 5-tensor $T^{(5)}$ can be flattened into a 3-tensor of shape $d^2 \times d^2 \times d$ that admits the following low-rank decomposition:

$$\sum_{k=1}^K \sum_{\ell=1}^d \frac{\pi_k}{d}(R_\ell \theta^{(k)})^{\otimes 2} \otimes (R_\ell \theta^{(k)})^{\otimes 2} \otimes (R_\ell \theta^{(k)}).$$

The algorithm heteroJen then proceeds by plugging $\tilde{T}_n^{(5)}$ into the above flattening operation and then applying Jennrich's algorithm to the resulting 3-tensor of shape $d^2 \times d^2 \times d$. Theorem 4.2 implies that this procedure outputs vectors $\tilde{u}_i$, $1 \le i \le dK$ with the following guarantees: there exist scalars β_i and a bijection $a \times b : [dK] \to [d] \times [K]$ satisfying

$$\|\tilde{u}_i - \beta_i (R_{a(i)}\theta^{(b(i))})^{\otimes 2}\|_\infty \le C_K \frac{\sigma^5}{\sqrt{n}} \operatorname{poly}(d),$$

with high probability.

We compute $\tilde{v}_i$ as the leading eigenvector of the $d \times d$ matrix $\tilde{u}_i$. Letting $\tilde{V}^{(3)}$ be the $d^3 \times dK$ matrix with columns $\tilde{v}_i^{\otimes 3}$, we estimate $\tilde{\alpha} \in \mathbb{R}^{dK}$ as the least-squares solution to $\tilde{V}^{(3)}\tilde{\alpha} = \operatorname{vec}(T^{(3)})$. The vectors $\tilde{w}_i := \tilde{\alpha}^{1/3}\tilde{v}_i$ (with entrywise exponentiation) now comprise dK redundant estimates to the K original signals θ_k; we remove this redundancy by clustering these dK estimates according to the pseudometric $\rho_2(x, y) = \min_{1\le \ell \le d} \|x^{\otimes 2} - (R_\ell y)^{\otimes 2}\|_2$. (We show in the proof of Theorem 5.1, below, that this clustering can be accomplished by a simple thresholding scheme.) Finally, the procedure heteroJen returns one vector from each cluster.

The heteroJen procedure enjoys the following theoretical guarantees that rely on the following condition number.

Let $U = [\operatorname{vec}((R_1\theta^{(1)})^{\otimes 2}), \ldots, \operatorname{vec}((R_d\theta^{(K)})^{\otimes 2})]$, and denote the condition number of U by κ. It can be shown that κ is generically finite. To that end, suppose we have some nonzero linear relation $0 = \sum_{k=1}^K \sum_{\ell=1}^d c_{k,\ell} (R_\ell \theta^{(k)})(R_\ell\theta^{(k)})^\top$. Multiplying by a DFT on the left and its adjoint on the right, and examining the a, b entry, we have $0 = \sum_{k=1}^K (\hat{c_k})_{a-b} \hat{\theta}_a^{(k)} \overline{\hat{\theta}_b^{(k)}}$. Some $(\hat{c_k})_\alpha$ is nonzero, yielding a nontrivial linear relation among the autocorrelation vectors v_k, $1 \le k \le K$, with $v_{k,j} = \hat{\theta}_j^{(k)} \hat{\theta}_{\alpha-j}^{(k)}$. These vectors satisfy the symmetry $v_{k,j} = v_{k,\alpha-j}$, but are generic on this subspace, which has dimension at least $\lceil d/2 \rceil$. Hence generically no such relation exists, and the matrix U has finite condition number.

Theorem 5.1. *Fix $\sigma > .1$ and $\delta \in (0, 1)$. Assume that $.1 \le \|\theta^{(k)}\|_2 \le 10$ for all $k \in [K]$ and that $K \le \lceil d/2 \rceil$. Then, for any $\varepsilon > 0$, the* heteroJen *applied to $T_n^{(5)}$ outputs $\{\tilde{\theta}_n^{(1)}, \ldots, \tilde{\theta}_n^{(K)}\}$*

such that

$$\sum_k \min_j \rho(\tilde\theta_n^{(j)}, \theta^{(k)}) \le \varepsilon\,,$$

with probability at least $1-\delta$ *whenever*

$$n \ge C_K \sigma^{10} \varepsilon^{-2} \operatorname{poly}(d, \kappa, 1/\delta)$$

in time $O(nd^5 + d^6 \operatorname{poly}(\log(1/\varepsilon)))$.

To the best of our knowledge, heteroJen is the first efficient method for heterogeneous MRA at low SNR. As noted above, follow-up work has shown that similar method-of-moments approaches based on tensor decomposition can efficiently achieve $1/\mathrm{SNR}^3$ sample complexity under the assumption that the components are drawn independently from $\mathcal{N}(0, I/d)$ [37].

6. Concluding remarks.

In this paper, we characterize the sample complexity of MRA, a first step towards a better statistical understanding of cryo-EM. In particular, we show that *any* estimator requires at least $1/\mathrm{SNR}^3$ samples at low SNR.

We also present an algorithm based on the method of invariants that provably solves the MRA problem with optimal sample complexity at low SNR. We further show that the approach can be adapted to heterogenous problems, an extension of particular importance in cryo-EM where different biological molecules or conformations are often imaged together. Our approach is the first to yield theoretical guarantees on any procedure for heterogenous MRA and opens the door to a broader application of the method of invariants.

While this work constitutes a first step towards a statistical theory of 3-D molecule reconstruction in cryo-EM, many questions remain open. Since an earlier version of this manuscript was available, follow-up work has shown that our approach can be extended to molecule reconstruction in cryo-EM [5] and to MRA with nonuniform shifts [1]. These works establish that the method of invariants yields optimal sample complexity in a wide variety of settings.

Appendix A. Proof of lower bounds.

A.1. Proof of Theorem 3.1.

In what follows, let c_d and C be constants (with c_d depending on d) whose value may change from line to line. Let θ be any zero-mean signal such that $\|\theta\|_2 = 1$, and let $\hat\theta$ be its Fourier transform. There exists a coefficient $\hat\theta_j$ with $j \ne 0$ such that $|\hat\theta_j| \ge c_d$. Define τ by setting

$$\hat\tau_k = \begin{cases} e^{i\delta}\hat\theta_j & \text{if } k = j \\ e^{-i\delta}\hat\theta_{-j} & \text{if } k = -j \\ \hat\theta_k & \text{otherwise,} \end{cases}$$

where $\delta = c_d \varepsilon$ for some constant c_d chosen so that $|\hat\tau_j - \hat\theta_j| \ge 2\varepsilon$ as long as ε is sufficiently small. It is clear that $\rho(\theta, \tau) \ge 2\varepsilon$, since when ε is sufficiently small, for any shift R we have $\|\theta - R\tau\|_2 \ge |\hat\theta_j - \widehat{R\tau}_j| \ge |\hat\theta_j - \hat\tau_j| \ge 2\varepsilon$.

We now establish that no procedure can distinguish between MRA with signal θ and MRA with signal τ on the basis of n samples if $n \le C\sigma^6\varepsilon^{-2}$ with probability better than $1/3$. To prove this, we reproduce the following theorem [8, Theorem 9], whose proof we sketch for completeness.

Theorem A.1. *Assume $\sigma \geq 1$. Let θ and τ be two mean-zero signals satisfying $\rho(\theta, \tau) \leq \varepsilon$ and $T^{(r)}(\theta) = T^{(r)}(\tau)$ for $r < k$. If P_θ and P_τ are the Gaussian mixtures corresponding to MRA with signals θ and τ respectively, then the Kullback-Leibler divergence $D(P_\theta \,\|\, P_\tau)$ satisfies*

$$D(P_\theta \,\|\, P_\tau) \leq C_k \sigma^{-2k} \varepsilon^2$$

for some constant C_k depending on k.

Proof. [Sketch] Let $\phi(x)$ be the density of a d-dimensional Gaussian with covariance $\sigma^2 I_d$, and let ϕ_θ and ϕ_τ be the densities of P_θ and P_τ, respectively. Let R be a uniformly distributed random cyclic shift. The convexity of the exponential function implies

$$\phi_\tau(x) = \mathbb{E}\phi(x - R\tau) \geq \phi(x) e^{-\|\tau\|_2^2/2\sigma^2}\,.$$

The χ^2 divergence $\chi^2(P_\theta \,\|\, P_\tau)$ between P_θ and P_τ then satisfies

$$\begin{aligned}\chi^2(P_\theta \,\|\, P_\tau) &:= \int \frac{(\phi_\theta(x) - \phi_\tau(x))^2}{\phi_\tau(x)} \mathrm{d}x \\ &\leq e^{\|\tau\|_2^2/2\sigma^2} \int (e^{-\|\theta\|_2^2/2\sigma^2} \mathbb{E} e^{\frac{x^\top R\theta}{\sigma^2}} - e^{-\|\phi\|_2^2/2\sigma^2} \mathbb{E} e^{\frac{x^\top R\tau}{\sigma^2}})^2 \phi(x) \mathrm{d}x\,.\end{aligned}$$

Expanding the square, collecting terms, and integrating with respect to x yields

$$\chi^2(P_\theta \,\|\, P_\tau) \leq e^{\|\tau\|_2^2/2\sigma^2} \mathbb{E}[e^{(R'\theta)^\top R\theta/\sigma^2} - 2e^{(R'\theta)^\top R\tau/\sigma^2} + e^{(R'\theta)^\top R\theta/\sigma^2}]\,,$$

where R' is an independent copy of R. Expanding this quantity as a Taylor series and applying Fubini's theorem to interchange summation and expectation yields

$$\chi^2(P_\theta \,\|\, P_\tau) \leq e^{\|\tau\|_2^2/2\sigma^2} \sum_{r \geq 1} \frac{\|T^{(r)}(\theta) - T^{(r)}(\tau)\|_{HS}^2}{\sigma^{2r} r!}\,,$$

where $\|\cdot\|_{HS}^2$ represents the Hilbert-Schmidt norm. By [8, Lemma B.12], for ε sufficiently small,

$$\|T^{(r)}(\theta) - T^{(r)}(\tau)\|_{HS}^2 \leq 12 \cdot 2^r \varepsilon^2\,.$$

Combining this with the assumption that $T^{(r)}(\theta) = T^{(r)}(\tau)$ for $r < k$ yields

$$\chi^2(P_\theta \,\|\, P_\tau) \leq e^{\|\tau\|_2^2/2\sigma^2} \varepsilon^2 \sum_{r \geq k} \frac{12 \cdot 2^r}{\sigma^{2r} r!} = C_k \sigma^{-2k} \varepsilon^2\,.$$

Since $D(P_\theta \,\|\, P_\tau) \leq \chi^2(P_\theta \,\|\, P_\tau)$ [34, Lemma 2.7], the claim follows. ■

The two signals θ and τ we have constructed are easily shown to satisfy $T^{(1)}(\theta) = T^{(1)}(\tau)$ and $T^{(2)}(\theta) = T^{(2)}(\tau)$, since their means and power spectra (i.e., the moduli of their Fourier transforms) agree.

By applying Theorem A.1 and Pinsker's inequality, we obtain

$$\mathrm{TV}(P_\theta^{\otimes n}, P_\tau^{\otimes n})^2 \le \frac{1}{2} D(P_\theta^{\otimes n} \,\|\, P_\tau^{\otimes n}) \le \frac{1}{2} C\sigma^{-6}\varepsilon^2 n\,.$$

Therefore if $n \le C^{-1}\sigma^6\varepsilon^{-2}$, [34, Theorem 2.2] implies that for any measurable function $\psi : \mathbb{R}^{d\times n} \to \{\theta, \tau\}$ of the data $y_1, \dots, y_n$, it holds

$$P_\theta^n(\psi(y_1, \dots, y_n) = \tau) + P_\tau^n(\psi(y_1, \dots, y_n) = \theta) \ge 1 - \mathrm{TV}(P_\theta^{\otimes n}, P_\tau^{\otimes n}) \ge 1/2\,.$$

In other words, any hypothesis test ψ must incur type-I and type-II error of at least 1/2. Via Le Cam's two-point testing argument [23], this fact implies that any estimator $\tilde{\theta}$ is bound to incur error ε with constant probability, as claimed. □

Appendix B. Proof of upper bounds.

B.1. Proof of Theorem 4.3. Write $\kappa(\theta)$ for $1/(\min_{j\in[d]} |\hat{\theta}_j|)$. It can be shown that the condition number $\kappa(U)$ of U satisfies $\kappa(U) \le \max_{j,k}\{|\hat{\theta}_j|/|\hat{\theta}_k|\} \le \kappa(\theta)$. Theorem 4.2 implies that Jennrich's algorithm applied to $\tilde{T}_n^{(3)}$ outputs $\tilde{u}_1$ satisfying

$$\|\tilde{u}_1 - \beta_1 R_j \theta\|_\infty \le \frac{\sigma^3 \operatorname{poly}(d)}{\sqrt{n}}, \quad c > 0\,,$$

with high probability for some $j \in [d]$ and some $\beta_j \in \mathbb{R}$.

As we are only concerned with polynomial dependence and not detailed bounds, we write $A \approx B$ if we can bound $|A - B| \le \alpha \operatorname{poly}(d, \kappa(\theta), \delta^{-1}) + \sigma/\sqrt{n}\operatorname{poly}(d, \kappa(\theta), \delta^{-1})$ so long as[1] $\alpha \le 1$ and $\sigma/\sqrt{n} \le 1$; we apply this also to vectors in 2-norm or (equivalently) most other common norms.

Theorem 4.2 guarantees us that $\tilde{u} \approx \beta_i R_i \theta$. Taking norms, we have $1 \approx |\beta_i|\|\theta\|_2$; as $\|\theta\|_2$ is bounded above and below by constants, we have that $|\beta_i| \approx 1/\|\theta\|_2$ is also of constant order. Note also that by Chebyshev we have that $|\tilde{\mu} - \mu| \le \sigma\sqrt{\delta}/\sqrt{2nd}$ with probability $1 - \delta/2$, so that $\mu \approx \tilde{\mu}$. From $\tilde{u} \approx \beta_i R_i \theta$ we also derive that

$$\langle \tilde{u}, \mathbf{1}\rangle/d \approx \beta_i \mu \approx \beta_i \tilde{\mu}. \tag{B.1}$$

We know that $\|\theta\|_2 \le \sigma_{\max}(U)$ and that $\sigma_{\min}(U) \le \|U \cdot \frac{1}{\sqrt{d}}\mathbf{1}\|_2 = d|\mu|$, so that $|\tilde{\mu}| \approx |\mu| \ge d\kappa(\theta)/\|\theta\|_2$; we are thus justified in dividing (B.1) by $\tilde{\mu}$ to obtain $\tilde{\beta} \approx \beta_i$, and $\beta_i/\tilde{\beta} \approx 1$. We now bound the total estimation error as follows:

$$\begin{aligned}
\|\tilde{\beta}^{-1}\tilde{u} - R_i\theta\|_2 &\le \|\tilde{\beta}^{-1}\tilde{u} - \beta_i^{-1}\tilde{u}\|_2 + \|\beta_i^{-1}\tilde{u} - R_i\theta\|_2 \\
&\le |\tilde{\beta}^{-1} - \beta_i^{-1}| + |\beta_i|^{-1}\alpha \\
&= |\beta_i|^{-1}\left(\alpha + \left|\frac{\beta}{\tilde{\beta}} - 1\right|\right) \approx 0.
\end{aligned}$$

[1]The precise bound of 1 here is arbitrary.

Thus in order to bound this estimation error to within ε, it suffices to require bounds of the form $\sigma/\sqrt{n} \le \varepsilon/\operatorname{poly}(d,\kappa(\theta),\delta^{-1})$ and $\alpha \le \varepsilon/\operatorname{poly}(d,\kappa(\theta),\delta^{-1})$. By Theorem 4.2, we achieve this bound on α from Jennrich's algorithm so long as $\|\tilde{T}_n^{(3)} - T^{(3)}\|_F \le \varepsilon/\operatorname{poly}(d,\kappa(\theta),\delta^{-1})$. This estimation error is achieved with probability $1-\delta/2$ so long as $n \ge \sigma^6 \operatorname{poly}(d,\kappa(\theta),\delta^{-1})$, which also subsumes the explicit bound on $\sigma/\sqrt{n}$. By a union bound over the two probabilistic steps in this argument, the desired accuracy guarantee holds with probability $1-\delta$. □

B.2. Proof of Theorem 5.1. As in the proof of Theorem 4.3, we write $A \approx B$ if we can bound

$$|A-B| \le \alpha \operatorname{poly}(d,\kappa,\delta^{-1}) + \sigma^3/\sqrt{n}\operatorname{poly}(d,\kappa,\delta^{-1})$$

given that[2] $\alpha \le 1$ and $\sigma^3/\sqrt{n} \le 1$; we apply this also to vectors in 2-norm or (equivalently) most other common norms.

From Theorem 4.2, we are guaranteed that $\tilde{u}_i \approx \beta_i (R_{a(i)}\theta^{(b(i))})^{\otimes 2}$; taking norms, we have $1 \approx |\beta_i| \|\theta^{(b(i))}\|_2$, so that $|\beta_i| \approx \|\theta^{(b(i))}\|^{-1}$ is of constant order. Now by the Davis–Kahan theorem, if $v_{\max}(M)$ denotes either choice of unit-length eigenvector of M corresponding to the eigenvalue of largest magnitude, we have

$$\tilde{v}_i := v_{\max}(\tilde{u}_i) \approx \varepsilon_i R_{a(i)}\theta^{(b(i))}/\|\theta^{(b(i))}\|_2,$$

for some sign $\varepsilon_i = \pm 1$. Then we have $\tilde{V}^{(3)} \approx V^{(3)}$, where as above, $\tilde{V}^{(3)}$ is the $d^3 \times dK$ matrix whose columns are $\tilde{v}_i^{\otimes 3}$, and $V^{(3)}$ has columns $\varepsilon_i (R_{a(i)}\theta^{(b(i))})^{\otimes 3}/\|\theta^{(b(i))}\|_2^3$. Estimating $T^{(3)}$ by $\tilde{T}_n^{(3)}$ according to Lemma 4.1, we have $\tilde{T}_n^{(3)} \approx T^{(3)}$ by Chebyshev, with probability $1-\delta/2$. Note then that $V^{(3)}\alpha = \operatorname{vec}(T^{(3)})$, where $\alpha_i = \varepsilon_i \|\theta^{(b(i))}\|_2^3/dK$. By the perturbation theory of linear systems, we are now guaranteed that, letting $\tilde{\alpha}$ be the least squares solution to $\tilde{V}^{(3)}\tilde{\alpha} = \operatorname{vec}(\tilde{T}_n^{(3)})$, we have $\tilde{\alpha} \approx \alpha$, so long as the system is well-conditioned, which we defer to the following lemma:

Lemma B.1. *If $\kappa(V^{(3)})$ denotes the condition number of $V^{(3)}$, then $\kappa(V^{(3)}) \le \kappa \operatorname{poly}(d)$.*

As α_i is of constant order, it follows that $\tilde{w}_i := \tilde{\alpha}_i^{1/3}\tilde{v}_i \approx R_{a(i)}\theta^{(b(i))}$, so that $\rho(\tilde{\alpha}_i^{1/3}\tilde{v}_i, \theta^{(b(i))}) \approx 0$. We are thus guaranteed dK good estimates to the original K signals. We next discuss how to remove this redundancy by clustering.

Define the pseudometric on $\mathbb{R}^d$ defined by $\rho_2(x,y) = \min_{1\le \ell \le d} \|x^{\otimes 2} - (R_\ell y)^{\otimes 2}\|_2$. Note that

$$\rho_2(w_i, w_{i'}) \approx \rho_2(R_{a(i)}\theta^{(b(i))}, R_{a(i')}\theta^{(b(i'))}) = \rho_2(\theta^{(b(i))}, \theta^{(b(i'))}).$$

If $b(i) = b(i')$, so that the two estimates w_i and $w_{i'}$ should represent the same signal, we thus have $\rho(w_i, w_{i'}) \approx 0$. If $b(i) \ne b(i')$, we have

$$\rho_2(w_i, w_{i'}) \approx \rho_2(\theta^{(b(i))}, \theta^{(b(i'))}) = \min_\ell \|U(e_{b(i),0} - e_{b(i'),\ell})\|_2 \ge \sqrt{2}\,\sigma_{\min}(U) = 1/\operatorname{poly}(d,\kappa),$$

where $e_{b,\ell} \in \mathbb{R}^{dK}$ is the standard basis vector corresponding to signal b and rotation ℓ. It follows that, provided α and σ^6/n are inverse-polynomially small in d, κ, δ^{-1}, we exactly

[2]The precise bound of 1 here is arbitrary.

recover the clusters of estimates $\tilde{w}_i$ corresponding to the same signal θ_k, simply by comparing on the metric ρ_2 and thresholding. Drawing one estimate $\tilde{w}_i$ from each cluster, we obtain one estimate of each signal.

To conclude, in order to bound this estimation error to within ε, it suffices to require bounds of the form $\sigma^3/\sqrt{n} \leq \varepsilon/\operatorname{poly}(d,\kappa,\delta^{-1})$ and $\alpha \leq \varepsilon/\operatorname{poly}(d,\kappa,\delta^{-1})$. By Theorem 4.2, we achieve this bound on α from Jennrich's algorithm so long as

$$\|\tilde{T}_n^{(5)} - T^{(5)}\|_F \leq \varepsilon/\operatorname{poly}(d,\kappa,\delta^{-1}).$$

This estimation error is achieved with probability $1-\delta/2$ so long as $n \geq \sigma^{10}\operatorname{poly}(d,\kappa,\delta^{-1})$, which also subsumes the explicit bound on $\sigma^3/\sqrt{n}$. By a union bound over the two probabilistic steps in this argument, the desired accuracy guarantee holds with probability $1-\delta$. $\square$

B.3. Proof of Lemma B.1. We apply the following transformations which do not alter the condition number: we transform the rows by the third tensor power of a DFT, we permute the columns to sort by signal and rotation, and we negate columns according to the signs ε_i. It thus suffices to control the condition number of the $d^3 \times dK$ matrix $V^{(3)\prime}$ whose columns are $(\hat{R}_j\hat{\theta}_k)^{\otimes 3}/\|\theta_k\|_2^3$, where $\hat{R}_i = \operatorname{diag}(\{\omega^{ij}\}_j)$ is the Fourier representation of a rotation action ($\omega = e^{2i\pi/d}$), and $\hat{\theta}$ is the Fourier transform of θ. Meanwhile, let V_2' be the $d^2 \times dK$ matrix with columns $(\hat{R}_j\hat{\theta}_k)^{\otimes 2}$, the Fourier transform of U, so that $\kappa(V_2') = \kappa$.

Let $v \in \mathbb{R}^{dK}$; then we have

$$\begin{aligned}
\|V^{(3)\prime}v\|_2^2 &= \sum_{\ell=1}^{d} \left\| V_2'\operatorname{diag}\left(\{\omega^{j\ell}(\hat{\theta}_k)_\ell \|\theta_k\|_2^{-3}\}_{jk}\right) v\right\|_2^2 \\
&\geq \sum_{\ell=1}^{d} \sigma_{\min}(U)^2 \left\|\operatorname{diag}\left(\{\omega^{j\ell}(\hat{\theta}_k)_\ell \|\theta_k\|_2^{-3}\}_{jk}\right) v\right\|_2^2 \\
&= \sum_{\ell} \sigma_{\min}(U)^2 \sum_{jk} |(\tilde{\theta}_k)_\ell|^2 \|\theta_k\|_2^{-6} |v|_{jk}^2 \\
&= \sigma_{\min}(U)^2 \sum_{k} \|\theta_k\|_2^{-4} \left(\sum_{j} |v|_{jk}^2\right) \\
&\geq \sigma_{\min}(U)^2 10^{-4} \|v\|_2^2,
\end{aligned}$$

so that $\sigma_{\min}(V^{(3)\prime}) \geq \sigma_{\min}(U)10^{-2}$. Observing the norms of columns, it is clear that $\sigma_{\max}(U)$ and $\sigma_{\max}(V^{(3)\prime})$ are bounded above by $\operatorname{poly}(d)$, so we conclude that $\kappa(V^{(3)}) = \kappa(V^{(3)\prime}) \leq \kappa\operatorname{poly}(d)$, as desired. $\square$

Acknowledgments. The authors thank Alex Wein for many insightful discussions on the topic of this paper.

REFERENCES

[1] E. Abbe, T. Bendory, W. Leeb, J. Pereira, N. Sharon, and A. Singer, *Multireference alignment is easier with an aperiodic translation distribution*, arXiv preprint arXiv:1710.02793, (2017).

[2] E. Abbe, J. Pereira, and A. Singer, *Sample complexity of the boolean multireference alignment problem*, in 2017 IEEE International Symposium on Information Theory (ISIT), July 2017.

[3] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, *Building rome in a day*, in Twelfth IEEE International Conference on Computer Vision (ICCV 2009), Kyoto, Japan, September 2009, IEEE.

[4] C. Aguerrebere, M. Delbracio, A. Bartesaghi, and G. Sapiro, *Fundamental limits in multi-image alignment*, IEEE Trans. Signal Process., 64 (2016), pp. 5707–5722, https://doi.org/10.1109/TSP.2016.2600517, http://dx.doi.org/10.1109/TSP.2016.2600517.

[5] A. S. Bandeira, B. Blum-Smith, J. Kileel, A. Perry, J. Weed, and A. S. Wein, *Estimation under group actions: recovering orbits from invariants*, arXiv preprint arXiv:1712.10163, (2018).

[6] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu, *Multireference alignment using semidefinite programming*, in ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science, ACM, New York, 2014, pp. 459–470.

[7] A. S. Bandeira, Y. Chen, and A. Singer, *Non-unique games over compact groups and orientation estimation in cryo-EM*, Available online at arXiv:1505.03840 [cs.CV], (2015).

[8] A. S. Bandeira, P. Rigollet, and J. Weed, *Optimal rates of estimation for multi-reference alignment*, Available online at arXiv:1702.08546 [math.ST], (2017).

[9] A. Barnett, L. Greengard, A. Pataki, and M. Spivak, *Rapid solution of the cryo-EM reconstruction problem by frequency marching*, SIAM J. Imaging Sci., 10 (2017), pp. 1170–1195, https://doi.org/10.1137/16M1097171, https://doi.org/10.1137/16M1097171.

[10] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, *Bispectrum inversion with application to multireference alignment*, Available online at arXiv:1705.00641 [cs.IT], (2017).

[11] D. R. Brillinger, *Some history of the study of higher-order moments and spectra*, Statist. Sinica, 1 (1991), pp. 465–476.

[12] H. N. Chapman, A. Barty, M. J. Bogan, S. Boutet, M. Frank, S. P. Hau-Riege, S. Marchesini, B. W. Woods, S. Bajt, W. H. Benner, et al., *Femtosecond diffractive imaging with a soft-x-ray free-electron laser*, Nature Physics, 2 (2006), p. 839.

[13] J. J. Donatelli, J. A. Sethian, and P. H. Zwart, *Reconstruction from limited single-particle diffraction data via simultaneous determination of state, orientation, intensity, and phase*, Proceedings of the National Academy of Sciences, 114 (2017), pp. 7222–7227.

[14] I. L. Dryden and K. V. Mardia, *Statistical shape analysis*, Wiley series in probability and statistics, Wiley, Chichester, 1998.

[15] H. Foroosh, J. Zerubia, and M. Berthod, *Extension of phase correlation to subpixel registration*, IEEE Trans. Image Processing, 11 (2002), pp. 188–200, https://doi.org/10.1109/83.988953, https://doi.org/10.1109/83.988953.

[16] J. Frank, *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*, Oxford University Press, 2006.

[17] K. Gaffney and H. Chapman, *Imaging atomic structure and dynamics with ultrafast x-ray scattering*, Science, 316 (2007), pp. 1444–1448.

[18] G. B. Giannakis, *Signal reconstruction from multiple correlations: Frequency-and time-domain approaches*, JOSA A, 6 (1989), pp. 682–697.

[19] N. Goyal, S. Vempala, and Y. Xiao, *Fourier PCA and robust tensor decomposition*, in Proceedings of the 46th Annual ACM Symposium on Theory of Computing, ACM, 2014, pp. 584–593.

[20] R. Harshman, *Foundations of the PARAFAC procedure: Model and conditions for an explanatory multimodal factor analysis*, tech. report, Tech. Rep. UCLA Working Papers in Phonetics 16, University of California, Los Angeles, Los Angeles, CA, December. 13, 27, 1970.

[21] R. Kakarala, *Completeness of bispectrum on compact groups*, arXiv preprint arXiv:0902.0196, (2009).

[22] Z. Kam, *The reconstruction of structure from electron micrographs of randomly oriented particles*, Journal of Theoretical Biology, 82 (1980), pp. 15–39.

[23] L. Le Cam, *Convergence of estimates under dimensionality restrictions*, The Annals of Statistics, (1973),

pp. 38–53.

[24] S. Leurgans, R. Ross, and R. Abel, *A decomposition for three-way arrays*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 1064–1083.

[25] A. Moitra, *Algorithmic aspects of machine learning*, Lecture notes (MIT), (2014).

[26] A. Moitra and G. Valiant, *Settling the polynomial learnability of mixtures of gaussians*, in 51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA, IEEE Computer Society, 2010, pp. 93–102, https://doi.org/10.1109/FOCS.2010.15, https://doi.org/10.1109/FOCS.2010.15.

[27] K. Pearson, *Contributions to the mathematical theory of evolution*, Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 185 (1894), pp. 71–110.

[28] D. Rosen, L. Carlone, A. Bandeira, and J. Leonard, *A certifiably correct algorithm for synchronization over the special Euclidean group*, in Intl. Workshop on the Algorithmic Foundations of Robotics (WAFR), San Francisco, CA, Dec. 2016.

[29] B. M. Sadler and G. B. Giannakis, *Shift- and rotation-invariant object reconstruction using the bispectrum*, Oct. Soc. Am. A, 9 (1992), pp. 57–69.

[30] J. Schnitzbauer, Y. Wang, S. Zhao, M. Bakalar, T. Nuwal, B. Chen, and B. Huang, *Correlation analysis framework for localization-based superresolution microscopy*, Proceedings of the National Academy of Sciences, 115 (2018), pp. 3219–3224.

[31] F. Sigworth, *A maximum-likelihood approach to single-particle image refinement*, Journal of structural biology, 122 (1998), pp. 328–339.

[32] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 20 – 36.

[33] B. Sonday, A. Singer, and I. G. Kevrekidis, *Noisy dynamic simulations in the presence of symmetry: Data alignment and model reduction*, Computers & Mathematics with Applications, 65 (2013), pp. 1535 – 1557.

[34] A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer Series in Statistics, Springer, New York, 2009, https://doi.org/10.1007/b13794, http://dx.doi.org/10.1007/b13794. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[35] J. W. Tukey, *The spectral representation and transformation properties of the higher moments of stationary time series*, in The Collected Works of John W. Tukey, D. R. Brillinger, ed., vol. 1, Wadsworth, 1984, ch. 4, pp. 165–184.

[36] B. von Ardenne, M. Mechelke, and H. Grubmüller, *Structure determination from single molecule x-ray scattering with three photons per image*, Nature communications, 9 (2018), p. 2375.

[37] A. Wein, *Statistical Estimation in the Presence of Group Actions*, PhD thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2018.