Multireference Alignment is Easier with an Aperiodic Translation Distribution

Emmanuel Abbe[1,2], Tamir Bendory[1], William Leeb[1], João M. Pereira[1], Nir Sharon[1], and Amit Singer[1,3]

[1]The Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA
[2]Electrical Engineering Department, Princeton University, Princeton, NJ, USA
[3]Department of Mathematics, Princeton University, Princeton, NJ, USA

***Abstract*—In the multireference alignment model, a signal is observed by the action of a random circular translation and the addition of Gaussian noise. The goal is to recover the signal's orbit by accessing multiple independent observations. Of particular interest is the sample complexity, i.e., the number of observations/samples needed in terms of the signal-to-noise ratio (the signal energy divided by the noise variance) in order to drive the mean-square error (MSE) to zero. Previous work showed that if the translations are drawn from the uniform distribution, then, in the low SNR regime, the sample complexity of the problem scales as $\omega(1/\operatorname{SNR}^3)$. In this work, using a generalization of the Chapman–Robbins bound for orbits and expansions of the χ^2 divergence at low SNR, we show that in the same regime the sample complexity for any aperiodic translation distribution scales as $\omega(1/\operatorname{SNR}^2)$. This rate is achieved by a simple spectral algorithm. We propose two additional algorithms based on non-convex optimization and expectation-maximization. We also draw a connection between the multireference alignment problem and the spiked covariance model.**

***Index Terms*—multireference alignment, spectral algorithm, method of moments, spiked covariance model, non-convex optimization, expectation-maximization, cryo–EM**

I. Introduction

The problem of multireference alignment (MRA) arises in a variety of engineering and scientific applications, among them structural biology [1], [2], [3], [4], [5], radar [6], [7], robotics [8] and image processing [9], [10], [11]. In these applications, one aims to estimate a signal from its translated or rotated noisy copies. The problem also serves as a simplified model for more general problems like single-particle reconstruction by cryo–electron microscopy (cryo–EM), in which a three-dimensional density is recovered from two-dimensional projections taken at unknown viewing directions [12], [13].

In this paper, we focus on the one-dimensional discrete MRA problem on a circle. In this model, we acquire N measurements from the model

$$Y_j = R_{S_j} x + \sigma G_j, \quad j = 1, \ldots, N, \tag{I.1}$$

where the G_j are i.i.d and drawn from $\mathcal{N}(0, I_L)$, i.e. $G_j \in \mathbb{R}^L$ and its entries are i.i.d standard Gaussian variables. The operator R_s translates a signal $x \in \mathbb{R}^L$ circularly by s elements, namely, $(R_s x)[i] = x[i-s]$, where all indices should be considered as modulo L. The translations S_j are i.i.d. and drawn from some unknown distribution ρ on $\mathbb{Z}_L$. Figure I.1 illustrates the MRA problem in different noise levels.

Previous approaches for estimating x from (I.1) can be broadly classified into two main categories. The first approach is based on estimating the translations S_j, aligning all observations and averaging them to suppress the noise. However, alignment is too erroneous in low signal–to–noise ratio (SNR) [14], defined here as $\operatorname{SNR} := \|x\|^2/\sigma^2$. Note that while the translations S_j are unknown, their estimation is not the primary goal of the problem. The translations are frequently called *latent*, *hidden* or *nuisance* parameters.

An alternative approach aims at estimating the signal x directly. Existing methods bypass the need to estimate the translations by employing expectation-maximization (EM) methods or by using features that are invariant under translation [15]. Section II is devoted to a detailed discussion on existing results and algorithms for MRA. In this paper, we take a different route by trying to estimate both the signal and the distribution of translations ρ simultaneously. When ρ is aperiodic, it turns out this is an easier problem than ignoring the fact that ρ is not uniform and estimating x alone.

In this paper we focus on the regime where both the number of observations and the variance of the noise are diverging. More specifically, our goal is to determine the sample complexity of (I.1), which we define to be the minimal number of measurements, as a function of the SNR, required such that there is a sequence of estimators $\{\hat{X}_N\}$ of x with mean square error (MSE) converging to 0 as N diverges. We define the MSE as

$$\operatorname{MSE} = \frac{1}{\|x\|_2^2} \mathbb{E}\left[\min_{s \in \mathbb{Z}_L} \|R_s \widehat{X} - x\|_2^2\right], \tag{I.2}$$

where the expectation is taken over the estimator $\widehat{X}$, which is a function of the random observations Y_j with distribution determined by (I.1). Allowing for a cyclic shift in (I.2) is intrinsic to the problem: if we apply a shift R_s to x, and its inverse R_{-s} to the right of ρ, we will produce exactly the same samples, thus there is no estimator $\hat{X}$ that is able to distinguish the observations that originate from x and the ones from $R_s x$.

EA was partly supported by the Bell Labs Prize, the NSF CAREER Award CCF–1552131, ARO grant W911NF–16–1–0051, NSF Center for the Science of Information CCF–0939370, and the Google Faculty Research Award. TB, WL, JP, NS, and AS were partially supported by Award Number R01GM090200 from the NIGMS, the Simons Foundation Investigator Award and Simons Collaboration on Algorithms and Geometry, the Moore Foundation Data-Driven Discovery Investigator Award, and AFOSR FA9550-17-1-0291.

In [16], it was proven that when ρ is the uniform distribution, then in the low SNR regime, the sample complexity for estimating signals with non-vanishing Discrete Fourier Transform (DFT) is $\omega(1/\operatorname{SNR}^3)$. In this work, we show that if the translation distribution ρ is aperiodic, meaning there is no $1 \le \ell \le L-1$ where $\rho[k+\ell] = \rho[k]$ for all $0 \le k \le L-1$, the sample complexity for estimating these signals is $\omega(1/\operatorname{SNR}^2)$. This rate is optimal and can be provably achieved by a spectral algorithm based on the first two moments of the data. The main result of this paper is stated as follows:

Main Result (informal): *Consider the model* (I.1) *and suppose that $x \in \mathbb{R}^L$ has a non-vanishing* DFT. *When ρ is aperiodic, the sample complexity of the MRA problem is lower bounded by $\omega(1/\operatorname{SNR}^2)$. This sample complexity is achieved by a spectral algorithm, based on the first two moments of the data. Conversely, the sample complexity for any periodic distribution with periodicity smaller than $L/2$, in particular the uniform distribution, scales like $\omega(1/\operatorname{SNR}^3)$.*

The proposed framework is based on a reliable estimation of the first two moments of the data. Hence, it requires only one pass over the measurements, low storage resources and is computationally efficient. To estimate the signal from the estimated moments, we propose, in addition to the aforementioned spectral algorithm, a non-convex least-squares (LS) algorithm. While the problem is non-convex, it empirically converges to the underlying signal, in the absence of noise, from a random initialization. We also suggest an expectation-maximization (EM) algorithm.

The outline of the paper is as follows. Section II provides a detailed discussion of existing results and algorithms for MRA. In Section III we prove that the sample complexity is lower bounded by $\omega(1/\operatorname{SNR}^2)$. We also show that the sample complexity of any periodic distribution of translations with a period of less than $L/2$ scales as $\omega(1/\operatorname{SNR}^3)$. This is an extension of the results of [16] which considered the uniform distribution case. In Section IV we show that if the distribution is aperiodic, then any signal with non-vanishing DFT can be estimated from its first and second moments, achieving the optimal estimation rate. Section V draws the connections between the MRA model and the well-studied spiked covariance model [17], [18], [19], [20], [21]. Section VI discusses and analyzes alternative algorithmic methods based on LS and EM. Section VII examines the performance of the proposed algorithms by numerical simulations. Section VIII concludes the paper and proposes potential future extensions.

Throughout the paper we use the following notation. We will use capital letter for random variables, and lower case letter for instances of this random variables. An estimator of a signal $z \in \mathbb{R}^L$ is denoted by $\widehat{Z}$. We assume throughout that all signals are defined cyclically; that is, all indices should be considered modulo L. The indices will range from 0 to $L-1$. The DFT of z is defined by $(Fz)[k] = \sum_{i=0}^{L-1} z[i]e^{-2\pi\iota ki/L}$, where $\iota = \sqrt{-1}$. We use C_z for a circulant matrix whose first column is z, namely, $C_z[i,j] = z[i-j]$. A diagonal matrix whose diagonal is z is denoted by D_z. We reserve $\mathbb{E}, *$ and $\odot$ for expectation, convolution and entry-wise product, respectively. The L–simplex is denoted by Δ^L. That is to say, $z \in \Delta^L$ implies that $z[i] \ge 0$ for all i and $\sum_{i=0}^{L-1} z[i] = 1$.

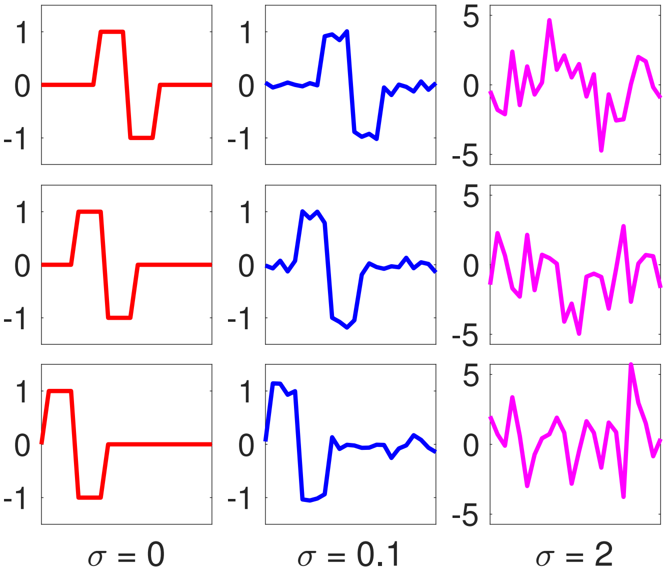


Fig. I.1. The figures illustrates the MRA measurements according to (I.1). The left column presents three measurements with different translations in the absence of noise. In this case, because the solution is defined up to translation, each measurement is a solution. The middle and right columns show measurements with the same translations and low and high noise levels, respectively.

II. Related work

A. Multireference alignment via synchronization

Given the translations s_j, the MRA problem (I.1) is easy. One trivial unbiased estimator of x is given by aligning all measurements and then averaging to suppress the noise, namely,

$$\widehat{X} = \frac{1}{N}\sum_{j=1}^{N} R_{s_j}^{-1} Y_j. \tag{II.1}$$

The variance of this estimator is σ^2/N and therefore the number of measurements N needs to scale like σ^2 to retain a constant estimation error. In other words, the sample complexity grows like $\omega(1/\operatorname{SNR})$. One can replace (II.1) with other estimators, such as James-Stein shrinkage [22], [23], [24], but it will not change the asymptotic sample complexity. In practice, we do not have access to the underlying translations. However, if one can obtain a reliable estimation of the unknown translations $\hat{s}_j$, then one can estimate x by the sample mean as in (II.1) at sample complexity $\omega(1/\operatorname{SNR})$. This motivates the design of synchronization methods that aim to estimate the translations s_j from the data y_j.

A naïve approach for synchronization could be to fix one observation as a template, say Y_1, and estimate the relative translation of each Y_j, with respect to Y_1, by the peak of their cross-correlation:

$$\hat{S}_j = \arg\max_s \sum_{i=0}^{L-1} Y_1[i]Y_j[i+s].$$

This approach may work in the high SNR regimes, but fails as the noise level increases (see for instance Figure I.1 in [15]).

Many alternative synchronization methods were proposed in the literature. For instance, the angular synchronization method aims at aligning all pairwise observations simultaneously [25], [26], [27], [28], [29], [30]. Other methods propose to align through different semidefinite programs (SDPs) [31], [32], [33], [34]. However, alignment is impossible below a critical SNR threshold, no matter how many measurements are acquired. For instance, for the continuous counterpart of (I.1), it has been shown that the Crámer–Rao lower bound is proportional to σ^2 and does not depend on N. This bound holds even if the sought signal is known [14].

B. Multireference alignment in low SNR

This section reviews recent works on MRA in the low SNR regime, in which methods based on alignment fail. The key idea is to estimate the signal directly, without estimating the translations beforehand. As will be emphasized throughout, all these works did not consider the translation distribution ρ, and either assumed or enforced it to be uniform.

In [16], it was shown that if the translations are uniformly distributed, namely, $S \sim \text{Uniform}[0, 1, \ldots, L-1]$, then the number of measurements needs to scale like $\omega(1/\,\mathrm{SNR}^3)$ for the estimator to converge in L^2 to the true signal. A follow-up paper [35] showed that this rate can be achieved by a tensor decomposition algorithm. The analysis of the uniform distribution is of particular interest since, no matter what ρ is, one can always enforce it to be uniform. This can be done simply by reshuffling all measurements by $z_j = R_{S'_j} y_j$, where S'_j are drawn from the uniform distribution. The new set of measurements z_j obeys the MRA model (I.1) with uniform translation distribution. However, as will be shown, this is in general a bad strategy, since the uniform distribution has a sample complexity scaling as $\omega(1/\,\mathrm{SNR}^3)$.

From the algorithmic point–of–view, a recent paper [15] proposes a method that completely overcomes the need to estimate the translations. The core idea is to estimate features of the underlying signal that are invariant under cyclic translation. Particularly, it was proposed to estimate the mean, power spectrum and bispectrum of the signal from the moments of the data. Since these invariant features are polynomials in the signal with degree at most three, they can be estimated at sample complexity growing like $\omega(1/\,\mathrm{SNR}^3)$. Using these invariant features, one can recover the signal as $N \to \infty$ using a variety of algorithms [15]. In [36], it was shown that a similar technique can be used to estimate several signals simultaneously from heterogeneous samples (see also [35, Section 5]). Since the invariant feature technique requires only one pass over the data, it can be performed in a streaming mode, can be parallelized, requires low storage resources of $\mathcal{O}(L^2)$ and has low computational load. The framework proposed in this paper is also based on estimating moments of the data and therefore enjoys the same advantages; however, since we only require second-order moments, we bring the sample complexity down to $\omega(1/\,\mathrm{SNR}^2)$.

Another approach for MRA is to apply an EM algorithm [37]. EM is an iterative algorithm that aims to find the maximum likelihood estimator and is used ubiquitously in many statistical models. For the MRA model (I.1), and under the assumption that the translations are drawn from the uniform distribution, this algorithm takes a simple form and consists of two steps at each iteration [15]. Given a current estimation x_{k-1}, the first step (called the E-step) computes a set of weights which can be understood as the translation distribution of each measurement y_j, if x_{k-1} was the underlying signal. These weights are computed by

$$w_k^{\ell,j} = C_k^j e^{-\frac{1}{2\sigma^2}\|R_\ell x_{k-1} - y_j\|_2^2},$$

where C_k^j is a normalization factor so that $\sum_\ell w_k^{\ell,j} = 1$. Then, the signal estimation is updated by marginalizing over the distributions and averaging (called the M-step):

$$x_k = \frac{1}{N}\sum_{j=1}^{N}\sum_{\ell=0}^{L-1} w_k^{\ell,j} R_\ell^{-1} y_j. \tag{II.2}$$

The EM algorithm enjoys an excellent numerical performance; however, its computational load and storage requirements are heavy since it passes through all the data at each iteration. In Section VI-B, we modify the standard EM algorithm to take the distribution into account.

III. Information theoretic lower bound

In this section, we provide lower bounds for the MSE of an estimator of the signal in terms of the SNR and the number of observations N. In particular, we show that under mild conditions on the signal the MSE is bounded away from zero if $N = O(1/\,\mathrm{SNR}^2)$. As described in Section IV, the MSE of Algorithm 2 converges to 0 if the number of measurements grows like $\omega(1/\,\mathrm{SNR}^2)$. In addition, if the distribution is periodic, the MSE is bounded away from zero if $N = O(1/\,\mathrm{SNR}^3)$. The framework proposed in [15] and described in Section II achieves this sample complexity for any distribution.

Recall that we can estimate the signal only up to cyclic translation. We define the best alignment of $\widehat{X}$ with x by

$$\phi_x(\widehat{X}) = \underset{z\in\{R_s\widehat{X}\}_{s\in\mathbb{Z}_L}}{\operatorname{argmin}} \|z - x\|. \tag{III.1}$$

Accordingly, we write (I.2) as

$$\mathrm{MSE} = \frac{1}{\|x\|^2}\mathbb{E}\left[\|\phi_x(\widehat{X}) - x\|_2^2\right]. \tag{III.2}$$

Since we are interested in estimators that converge to a cyclic shift of x in L^2 as N diverges, we only consider estimators which are asymptotically unbiased, i.e., $\mathbb{E}[\phi_x(\widehat{X})] \to x$ as $N \to \infty$. However the information lower bounds presented in this paper can be adapted to biased estimators (see Theorem III.4). We now present the main results of this section as follows:

Theorem III.1. *Assume that x is not a constant vector. If $\widehat{X}$ is an asymptotically unbiased estimator of x, then*

$$\mathrm{MSE} \geq \frac{1}{8N}\frac{1}{\mathrm{SNR}^2} - O\left(\frac{1}{N\,\mathrm{SNR}^{1.5}}\right). \tag{III.3}$$

Moreover, if ρ is periodic, with a period $\ell < \frac{L}{2}$, then

$$\mathrm{MSE} \geq \frac{1}{54N} \frac{L-2\ell}{2\ell} \frac{1}{\mathrm{SNR}^3} - O\left(\frac{1}{N\,\mathrm{SNR}^{2.5}}\right). \qquad \text{(III.4)}$$

Note that previous work [16] derived the sample complexity for the uniform distribution of translations. Theorem III.1 extends it to any distribution. In addition, we extend [16] by providing the constant that multiplies σ^6, for the uniform distribution case.

In the rest of this section, we develop the main tools required to prove Theorem III.1. Specifically, we start by introducing an auxiliary notation and definitions. Then, in Section III-B we use an adaptation of the Chapman-Robbins lower bound [38], which is a generalization of the Cramér-Rao bound [39], to derive a lower bound on the MSE in terms of the χ^2 divergence, this is Theorem III.4. Then, in Section III-C, we express the χ^2 divergence in terms of the Taylor expansion of the posterior probability density and the moment tensors, obtaining Lemma III.5. Finally in section III-D we combine Theorem III.4 and Lemma III.5 to obtain a general lower bound for MRA, that we particularize for the case when ρ is aperiodic and periodic, respectively. The final details of the proof of Theorem III.1 are given in Appendix D.

A. Notation and definitions

Let $Y^N \in \mathbb{R}^{L\times N}$ be the collection of all measurements as columns in a matrix. Let us denote by $f^N_{x,\rho}$ the probability density of the posterior distribution of Y^N,

$$f^N_{x,\rho}(y^N) = \prod_{j=1}^{N} f_{x,\rho}(y_j), \qquad \text{(III.5)}$$

and the expectation of a function g of the measurements under the measure $f^N_{x,\rho}$ by

$$\mathbb{E}_{x,\rho}\left[g\left(Y^N\right)\right] := \int_{\mathbb{R}^{L\times N}} g\left(y^N\right) f^N_{x,\rho}\left(y^N\right) dy^N.$$

For ease of notation, we write $\mathbb{E}\left[g\left(Y^N\right)\right]$ when the signal and distribution are implicit. The bias-variance trade-off of the MSE is given by

$$\mathrm{MSE} = \frac{\mathrm{tr}(\mathrm{Cov}[\phi_x(\widehat{X})])}{\|x\|^2} + \frac{\|\mathbb{E}[\phi_x(\widehat{X})] - x\|^2}{\|x\|^2}, \qquad \text{(III.6)}$$

with

$$\mathrm{Cov}[\phi_x(\widehat{X})] = \mathbb{E}\left[\phi_x(\widehat{X})\phi_x(\widehat{X})^T\right] - \mathbb{E}[\phi_x(\widehat{X})]\mathbb{E}[\phi_x(\widehat{X})]^T. \qquad \text{(III.7)}$$

We conclude this part with two definitions. First, we define the moment tensors. For a vector $x \in \mathbb{R}^L$, we denote by $x^{\otimes d}$ the L^d dimensional tensor where the entry indexed by $\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}_L^d$ is given by $\prod_{j=1}^d x[k_j]$. The space of d-dimensional tensors forms a vector space, with sum and multiplication defined entry-wise. This vector-space has inner product and norm defined by $\langle A, B\rangle = \sum_{\mathbf{k}\in\mathbb{Z}_L^d} A[\mathbf{k}]B[\mathbf{k}]$ and $\|A\|^2 = \langle A, A\rangle$, respectively.

Definition III.2. The n-th order moment of x over ρ, is the tensor of order n and dimension L^n, defined by

$$M^n_{x,\rho} := \mathbb{E}\left[(R_S x)^{\otimes n}\right],$$

where $S \sim \rho$.

We will explore this notion in more detail in section IV-A, in particular we give explicit formulas for the moments when $n = 1$ (IV.1) and $n = 2$ (IV.4).

Our last definition is of the χ^2 divergence, which gives a measure of how "far" two probability distributions are.

Definition III.3. The χ^2 divergence between two probability densities f_A and f_B, with f_A absolutely continuous with respect to f_B, is defined by

$$\chi^2(f_A||f_B) := \mathbb{E}\left[\left(\frac{f_A(B)}{f_B(B)} - 1\right)^2\right],$$

where $B \sim f_B$.

Due to equation (III.5), the relation between the χ^2 divergence for N and one observations is given by

$$\chi^2(f^N_{\tilde{x},\tilde{\rho}}||f^N_{x,\rho}) = (\chi^2(f_{\tilde{x},\tilde{\rho}}||f_{x,\rho}) + 1)^N - 1. \qquad \text{(III.8)}$$

B. Chapman-Robbins lower bound for an orbit

The classical Chapman-Robbins gives a lower bound on an error metric of the form $\mathbb{E}[\|\widehat{X} - x\|^2]$, i.e., it does not take into consideration a translation-invariant error metric as appears naturally in the MRA problem. Hence, we modify the Chapman-Robbins bound to accommodate error of the form (III.2). We point out that $\mathrm{Cov}[\phi_x(\widehat{X})]$ is related to the MSE by (III.6).

Theorem III.4 (Chapman-Robbins for orbits). *For any $\tilde{x} \in \mathbb{R}^L$ such that $\phi_x(\tilde{x}) \neq x$ and $\tilde{\rho} \in \Delta^L$, we have*

$$\mathrm{Cov}[\phi_x(\widehat{X})] \succeq \frac{zz^T}{\chi^2(f^N_{\tilde{x},\tilde{\rho}}||f^N_{x,\rho})},$$

where $z = \mathbb{E}_{\tilde{x},\tilde{\rho}}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]$.

Proof. See Appendix A. □

C. Fisher information and moment tensors

In this subsection we give a characterization of the χ^2 divergence, which appears in the Chapman-Robbins bound, in terms of the moment tensors.

Instead of considering the posterior probability density of Y^N, we will consider its normalized version $\widetilde{Y}^N = Y^N/\sigma$. We then have

$$\widetilde{Y}_j = \gamma R_{S_j} x + G_j, \qquad \text{(III.9)}$$

where $\gamma = 1/\sigma$, $S_j \sim \rho$ and $G_j \sim \mathcal{N}(0, I)$. While this change of variables does not change the χ^2 divergence, we can now take the Taylor expansion of the probability density around $\gamma = 0$, that is,

$$f_{x,\rho}(y;\gamma) = f_G(y) \sum_{j=0}^{\infty} \alpha^j_{x,\rho}(y) \frac{\gamma^j}{j!}, \qquad \text{(III.10)}$$

where $f_G(y) = f_{x,\rho}(y; 0)$ is the probability density of G_j (since when $\gamma = 0$, $\widetilde{Y}_j = G_j$) and

$$\alpha^j_{x,\rho}(y) := \frac{1}{f_G(y)} \frac{\partial^j f_{x,\rho}}{\partial \gamma^j}(y; 0), \tag{III.11}$$

thus $\alpha^0_{x,\rho}(y) = 1$. We note $f_{x,\rho}(y; \gamma)$ is infinitely differentiable for all $y \in \mathbb{R}^L$, thus $\alpha^j_{x,\rho}(y)$ is always well-defined. We now use (III.10) to give an expression of the χ^2 divergence in terms of the moment tensors.

Lemma III.5. *The divergence $\chi^2(f_{\tilde{x},\tilde{\rho}}||f_{x,\rho})$ can be expressed in terms of the data moments as:*

$$\begin{aligned}\chi^2(f_{\tilde{x},\tilde{\rho}}||f_{x,\rho}) \\ &= \frac{\sigma^{-2d}}{(d!)^2} \mathbb{E}\left[\left(\alpha^d_{\tilde{x},\tilde{\rho}}(G) - \alpha^d_{x,\rho}(G)\right)^2\right] + O(\sigma^{-2d-1}), \qquad \text{(III.12)} \\ &= \frac{\sigma^{-2d}}{d!} \|M^d_{\tilde{x},\tilde{\rho}} - M^d_{x,\rho}\|^2 + O(\sigma^{-2d-1}), \qquad \text{(III.13)}\end{aligned}$$

where $d = \inf\left\{n : \|M^n_{\tilde{x},\tilde{\rho}} - M^n_{x,\rho}\|^2 > 0\right\}$.

Proof. See Appendix B. □

Equation (III.12) is not specific to MRA: one can always obtain this expression as long we are considering the low SNR regime and the observations are independent of the signal in the limit of SNR tending to 0. The particularization to MRA happens in (III.13), due to (III.9) and (III.11).

D. General lower bound for the MRA problem

The following theorem is obtained from the results presented in the previous sections.

Theorem III.6. *Consider the estimation problem given by equation* (I.1). *For any signal $\tilde{x} \in \mathbb{R}^L$ such that $\phi_x(\tilde{x}) \neq x$ and for any $\tilde{\rho} \in \Delta^L$, let $K^n_{\tilde{x},\tilde{\rho}} = \frac{1}{n!}\|M^n_{\tilde{x},\tilde{\rho}} - M^n_{x,\rho}\|^2$, $d_{\tilde{x},\tilde{\rho}} = \inf\left\{n : K^n_{\tilde{x},\tilde{\rho}} > 0\right\}$ and $\bar{d} = \max d_{\tilde{x},\tilde{\rho}}$. Finally let*

$$\lambda^m_N = N/\sigma^{2m}, \quad m \in \mathbb{Z}_+.$$

We have

$$\mathrm{MSE} \geq \sup_{\tilde{x},\tilde{\rho}: d_{\tilde{x},\tilde{\rho}} = \bar{d}} \left\{\frac{\|\phi_x(\tilde{x}) - x\|^2/\|x\|^2}{\exp\left(\lambda^{\bar{d}}_N K^{\bar{d}}_{\tilde{x},\tilde{\rho}}\right) - 1 + O\left(\lambda^{\bar{d}}_N \sigma^{-1}\right)}\right\}, \tag{III.14}$$

thus the MSE is bounded away from zero if $\lambda^{\bar{d}}_N$ is bounded from above, or equivalently $N = O(1/\,\mathrm{SNR}^{\bar{d}})$.

Proof. We first note that $\bar{d} \leq L$, so the maximum is well defined. By Theorem III.4, Lemma III.5, equations (III.7) and (III.8) we obtain

$$\mathrm{MSE} \geq \frac{\|z\|^2/\|x\|^2}{\left(1 + \sigma^{-2d} K^d_{\tilde{x},\tilde{\rho}} + O\left(\sigma^{-2d-1}\right)\right)^N - 1}. \tag{III.15}$$

with $z = \mathbb{E}_{\tilde{x},\tilde{\rho}}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]$. Since $\hat{X}$ is asymptotically unbiased, $\|z\|^2 \to \|\phi_x(\tilde{x}) - x\|^2$ as N diverges. On the other hand we have

$$\begin{aligned}\left(1 + \sigma^{-2d} K^d_{\tilde{x},\tilde{\rho}} + O(\sigma^{-2d-1})\right)^N = \\ \exp\left(\lambda^d_N K^d_{\tilde{x},\tilde{\rho}}\right) + O\left(\lambda^d_N \sigma^{-1}\right)\end{aligned}$$

and (III.14) now follows from taking the supremum over $\tilde{x}$ and $\tilde{\rho}$. □

From this theorem we can obtain (III.3) by providing $\tilde{x}$ and $\tilde{\rho}$ which have $M^1_{\tilde{x},\tilde{\rho}} = M^1_{x,\rho}$, this implies $\bar{d} \geq 2$ and the MSE is bounded away from 0 if $N = O(1/\,\mathrm{SNR}^2)$. Moreover, to obtain (III.4) when ρ is periodic we can provide $\tilde{x}$ and $\tilde{\rho}$ which have $M^d_{\tilde{x},\tilde{\rho}} = M^d_{x,\rho}$ for $d = 1, 2$, similarly to Proposition IV.6, this implies $\bar{d} \geq 3$ and the MSE is bounded away from 0 if $N = O(1/\,\mathrm{SNR}^3)$.

However, when $N = \omega(1/\,\mathrm{SNR}^{\bar{d}})$ the supremum in (III.14) is going to be achieved in the limit $(\tilde{x}, \tilde{\rho}) \to (x, \rho)$. Thus, to prove Theorem III.1, we use intermediate results which explore the limit $(\tilde{x}, \tilde{\rho}) \to (x, \rho)$, and thus provide tighter bounds. However, since considering the limit introduces some technical details, we leave its analysis to Appendix C. The final details of the proof of Theorem III.1 are presented in Appendix D.

IV. Provable algorithm based on the first two moments

In this section, we provide a spectral algorithm to estimate the signal, up to cyclic translation, from the first and second moments of the data, provided that the translation distribution is aperiodic. We prove that this algorithm estimates the signal exactly with high probability in the limit of SNR tending to 0 with a growing number of samples; we will describe the asymptotic model more precisely in Section IV-C. Because the method relies on only second-order information, its sample complexity in this case only grows like $\omega(1/\,\mathrm{SNR}^2)$, compared to sample complexity growing as $\omega(1/\,\mathrm{SNR}^3)$ if the translation distribution is periodic (with period smaller than $L/2$; see Section IV-D). As we proved in Section III, $\omega(1/\,\mathrm{SNR}^2)$ is indeed the sample complexity for aperiodic distributions.

A. Moments of $R_S x$

Before describing the algorithm, we will review a few basic properties of the moments of the random vectors $R_S x$, defined in Definition III.2, and conclude with a theoretical result about the sufficient information they hold.

We will first consider the first moment of the translated signal, $M^1 = \mathbb{E}[R_S x]$, where $S \sim \rho$. This is equal to the convolution of x with ρ; that is,

$$M^1 = x * \rho = C_x \rho = C_\rho x, \tag{IV.1}$$

where C_x is the circulant matrix with x as its first column (and similarly for C_ρ). In this case, the convolution theorem implies

$$F M^1 = F x \odot F \rho, \tag{IV.2}$$

where $\odot$ and F denote entry-wise product and Fourier transform, respectively. We can estimate the first moment from the noisy observations (I.1) by

$$\widehat{M}^1 = \frac{1}{N}\sum_{i=1}^{N} Y_i. \tag{IV.3}$$

Note that if L and σ are fixed, then $\widehat{M}^1$ is a consistent estimator of M^1 as $N \to \infty$.

The second moment of $R_S x$ is defined as

$$M^2 = \mathbb{E}\left[(R_S x)(R_S x)^T\right],$$

where $S \sim \rho$. It can be verified that

$$M^2 = C_x D_\rho C_x^T, \tag{IV.4}$$

where D_ρ is a diagonal matrix of ρ. The unbiased second moment of $R_S x$ is then estimated from the observations Y_j by:

$$\widehat{M}^2 = \frac{1}{N}\sum_{i=1}^{N} Y_i Y_i^T - \sigma^2 I, \tag{IV.5}$$

where I denotes the $L \times L$ identity matrix. As with the first moment, when L and σ are fixed then $\widehat{M}^2$ is a consistent estimator of M^2 as $N \to \infty$.

We conclude this section with the following result, showing conditions which guarantee that there exists only one pair of signal and distribution (up to translation) that exactly agrees with the second moment data. First, recall that a distribution ρ is periodic if and only if there exists a period $1 \le \ell < L$ such that

$$\rho[k] = \rho[k+\ell], \quad k = 0, \dots, L-1.$$

If no period exists we simply call ρ aperiodic distribution.

Theorem IV.1. *Assume that ρ_1 is an aperiodic distribution, and that x_1 is a signal with non-vanishing DFT. Let x_2 and ρ_2 be any other signal and distribution with the same first two moments as x_1 and ρ_1. Then x_2 and ρ_2 are equal to x_1 and ρ_1, respectively, up to a shift. More precisely, there is $s \in \{0, \dots, L-1\}$ with $x_2 = R_s x_1$ and $\rho_2 = R_{-s}\rho_1$.*

The proof is given in Appendix E. Next, we show a constructive method to recover x and ρ from their first two moments M^1 and M^2.

B. Moment inversion when ρ has a unique entry

The key observation driving the algorithm we will describe is that when ρ has at least one distinct entry, and if x has non-zero DFT, then x can be recovered exactly from the first two moments M^1 and M^2.

We first note that the power spectrum of the signal, $P_x[k] := |(Fx)[k]|^2$, is the Fourier transform of the signal's auto-correlation and thus can be derived directly from the second moment. We first recall the factorization $M^2 = C_x D_\rho C_x^T$ from equation (IV.4). The circulant matrix C_x is diagonalized by the Fourier matrix F as follows:

$$C_x = F^{-1} D_{Fx} F,$$

thus we have

$$F M^2 F^{-1} = \frac{1}{L} D_{Fx} C_{F\rho} D_{\overline{Fx}}. \tag{IV.6}$$

The k-th element of the diagonal of (IV.6) is given by

$$\frac{1}{L}(F\rho)[0]\ |(Fx)[k]|^2 = \frac{1}{L} P_x[k],$$

where $(F\rho)[0] = \sum_i \rho[i] = 1$, since ρ is a distribution. Consequently, we can obtain the power spectrum of x from M_2 by $P_x = L\,\mathrm{diag}(F M^2 F^{-1})$.

Now if we conjugate M^2 by the matrix $F^{-1} D_{1/|P_x|^{1/2}} F$, we obtain the matrix $\widetilde{M}^2 = C_{\tilde{x}} D_\rho C_{\tilde{x}}^T$, where $\tilde{x}$ is the vector with the normalized Fourier transform

$$(F\tilde{x})[k] = \frac{(Fx)[k]}{|(Fx)[k]|}. \tag{IV.7}$$

Therefore, the matrix $C_{\tilde{x}}$ is both circulant and real orthonormal, i.e., $C_{\tilde{x}}^{-1} = C_{\tilde{x}}^T$. Consequently, the decomposition $\widetilde{M}^2 = C_{\tilde{x}} D_\rho C_{\tilde{x}}^T$ is an eigendecomposition of $\widetilde{M}^2$, and the eigenvectors are translations of $\tilde{x}$.

If ρ has at least one distinct entry, then the associated eigenvector v will be a translation of $\tilde{x}$, with arbitrary scaling; that is, $v = \alpha \cdot R_s \tilde{x}$ for some number α and shift s. Since the Fourier coefficients are still normalized, we multiply $|P_x|^{1/2}$ and Fv coordinate-wise to get

$$\tilde{v} = \alpha \cdot F^{-1}\left(F(R_s\tilde{x}) \odot |P_x|^{1/2}\right) = \alpha \cdot R_s x.$$

Letting $\mathrm{Sum}(x)$ denote the sum of all elements in x, we have $\alpha = \mathrm{Sum}(\tilde{v})/\mathrm{Sum}(x)$. To uncover α, note that the zeroth Fourier coefficient of $M^1 = x * \rho$ is $(FM^1)[0] = (Fx)[0] \cdot (F\rho)[0]$. But since ρ is a probability vector, $(F\rho)[0] = 1$, and so $\mathrm{Sum}(M^1) = (FM^1)[0] = (Fx)[0] = \mathrm{Sum}(x)$. Consequently, $\alpha = \mathrm{Sum}(\tilde{v})/\mathrm{Sum}(M^1)$, and $R_s x = \tilde{v}/\alpha$.

Note that once we have determined x, we can also determine ρ from $M^1 = x * \rho$ by deconvolution; indeed, since $M^1 = C_x \rho$, we have $\rho = C_x^{-1} M^1$. The algorithm is summarized in Algorithm 1.

Algorithm 1 Exact recovery from the first two moments

```
Input: Moments $M^1$ and $M^2$.
Output: The signal $x$ and distribution $\rho$.
    // Normalize $Fx$
    $P_x \leftarrow L\,\mathrm{diag}(F M^2 F^{-1})$
    $p \leftarrow (P_x)^{-1/2}$
    $Q \leftarrow F^{-1} D_p F$
    $\widetilde{M}^2 \leftarrow Q M^2 Q^*$
    // Extract eigenvector and rescale
    $v \leftarrow \mathrm{UniqEig}(\widetilde{M}^2)$
    $\tilde{v} \leftarrow F^{-1}\left((P_x)^{1/2} \odot Fv\right)$
    $x \leftarrow \left(\mathrm{Sum}(M^1)/\,\mathrm{Sum}(\tilde{v})\right)\tilde{v}$
    $\rho \leftarrow C_x^{-1} M^1$
    return $x$ and $\rho$
```

We have proved the following result:

Proposition IV.2. *Suppose x has non-vanishing* DFT *and ρ has at least one distinct entry and let $M^1 = \mathbb{E}[R_S x]$*

and $M^2 = \mathbb{E}[(R_S x)(R_S x)^T]$ be the first two moments. Then Algorithm 1 returns the signal x and the distribution ρ exactly (up to translation).

C. Estimating x in low SNR

Section IV-B shows that Algorithm 1 recovers x exactly from the exact values of M^1 and M^2, so long as the DFT of x is non-vanishing and ρ has at least one distinct entry. In this section we show that under the same conditions Algorithm 1 is stable under small perturbations of the moments. We also show that if $N = \omega(\sigma^4)$, or equivalently $N = \omega(1/\operatorname{SNR}^2)$, the MSE of the estimate given by Algorithm 1 converges to 0 as N diverges.

We first observe that whenever ρ is aperiodic, we can modify the observations to assume that ρ in fact has all distinct entries. Indeed, we generate a new set of measurements $z_j = R_{S'_j} y_j$, where S'_j are drawn from a new, known distribution θ. In this case, the translations are distributed according to $\rho * \theta$. The following lemma shows that by choosing θ as a random probability distribution on the simplex, we can ensure that all entries of $\rho * \theta$ are distinct with probability 1. Note that if the DFT of θ is non-vanishing (which holds with probability 1 for random θ), then one can recover fully ρ from $\rho * \theta$.

Lemma IV.3. *Let ρ be an aperiodic vector on the simplex and let θ be a random probability density function on the simplex. Then, all entries of $\rho * \theta$ are distinct with probability 1.*

Proof. See Appendix F. □

Using this lemma, we will assume from now on that all entries of ρ are distinct. The following corollary states that Algorithm 1 is stable to perturbations of the moments and power spectrum:

Corollary IV.4. *Suppose x has non-vanishing* DFT *and denote by $\widehat{M}^1$ and $\widehat{M}^2$ the sample moments and power spectrum defined by equations* (IV.3) *and* (IV.5). *Suppose that $\|\widehat{M}^1 - M^1\|_F \le \varepsilon$ and $\|\widehat{M}^2 - M^2\|_F \le \varepsilon$, for sufficiently small $\varepsilon > 0$. Then Algorithm 1, with input data $\widehat{M}^1$ and $\widehat{M}^2$, returns an estimate $\widehat{X}_{\text{Spectral}}$ of x with error at most $C\varepsilon$, where C is a finite and positive constant which depends only on x and ρ.*

Proof. This follows immediately from the variant of the Davis-Kahan theorem found in [40, Theorem 2]. □

The following theorem shows that if N grows like $\omega(\sigma^4)$, the MSE of the estimator converges to 0 as N diverges.

Theorem IV.5. *If $N = \omega(\sigma^4)$, the MSE of $\widehat{X}_{\text{Spectral}}$, defined in Corollary IV.4, converges to 0 as N diverges.*

Proof. See Appendix G. □

Algorithm 2 describes the entire pipeline for estimating x from the noisy measurements (I.1), including randomly shifting the observations, estimating the moments, and using Algorithm 1 to estimate x from the estimated moments.

Algorithm 2 Estimating x and ρ from noisy data

```
Input: $y_j$, $j = 1, \dots, N$ of (I.1) and noise variance $\sigma^2$.
Output: An estimated signal $\hat{x}$ and estimated distribution $\hat{\rho}$.
    // Reshuffling observations (optional)
draw a random distribution $\theta \in \Delta^L$
for each $j = 1, \dots, N$: $y_j \leftarrow R_{S'_j} y_j$ for $S'_j \sim \theta$
    // Moment estimation
$\widehat{M}^1 \leftarrow \frac{1}{N}\sum_{j=1}^N y_j$
$\widehat{M}^2 \leftarrow \frac{1}{N}\sum_{j=1}^N y_j y_j^T - \sigma^2 I$
    // Eigendecomposition and normalization
obtain $\hat{x}$ and $\hat{\rho}'$ from Algorithm 1 with $\widehat{M}^1$ and $\widehat{M}^2$.
return $\hat{x}$ and $\hat{\rho} = C_\theta^{-1}\hat{\rho}'$.
```

D. Non-uniqueness for periodic ρ

We have shown that the first and the second moments suffice to determine the signal if the distribution is aperiodic. In this section, we provide a complementary result, showing that if the distribution is periodic, then having the first two moments is not enough to uniquely determine a signal with non-vanishing DFT. In particular, given a distribution ρ with period ℓ, a signal x_2 has the same first two moments as x_1 if it satisfies:

$$(Fx_2)[k] = \begin{cases} (Fx_1)[k], & k = t\frac{L}{\ell}, \quad t = 0, \dots, \ell - 1, \\ -(Fx_1)[k], & \text{otherwise}. \end{cases} \tag{IV.8}$$

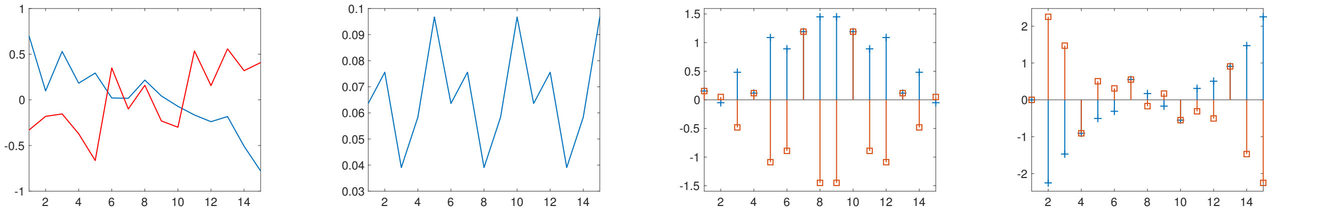
This construction is demonstrated in Figure IV.1.

Proposition IV.6. *Let $\ell < L/2$ be a divisor of $L > 1$. Suppose that ρ is periodic, with period ℓ, and let x_1 be a given signal with non-vanishing* DFT. *Then the signal x_2 defined by* (IV.8) *is not a translation of x_1, and has the same first and second moments as x_1. Therefore, if the distribution is periodic, then any signal with non-vanishing* DFT *is not uniquely determined from its first two moments.*

Proof. See Appendix H. □

In Section III we established this result from an information-theoretic perspective by showing that the sample complexity for periodic distribution grows like $\omega(1/\operatorname{SNR}^3)$, and extending [16] that considered only the uniform distribution. Indeed, the uniform distribution is merely a special case of periodic distributions with minimal period $\ell = 1$. When $\ell > 1$, one can interpret the periodicity as having a uniform distribution over the different cosets of $\mathbb{Z}_L$ with respect to the subgroup generated by a translation in ℓ coordinates. These cosets are exactly the analogue of the sparsity pattern of $F\rho$ attained by jumps of L/ℓ. This also explains why uniformity is the only pathological case for a prime L. Therefore, if one can choose how to sample the signal, a prime number of samples should be considered.

As it turns out, there is one special case where the first two moments are enough to determine x uniquely, up to cyclic translation, even when ρ is periodic. This special case occurs when L is even and ρ is $L/2$-periodic. Note that in this case the information theoretic lower bound presented in section III is also $\omega(1/\operatorname{SNR}^2)$. This result is formulated in the following claim:



(a) The two different real signals x_1 and x_2 of length 15

(b) The 5-periodic distribution

(c) The real parts of Fx_1 and Fx_2

(d) The imaginary parts of Fx_1 and Fx_2

Fig. IV.1. This example demonstrates the constriction of (IV.8) and Proposition IV.6. The figures present two different real signals of length 15 and a 5-periodic distribution. The Fourier transforms of the signals obey (IV.8). The two signals have the same first two moments under the periodic distribution.

Claim IV.7. Suppose that x has non-vanishing DFT, L is even and ρ is $L/2$-periodic. Then, x is uniquely determined from its first two moments, up to global translation.

Proof. See Appendix I. □

V. Connection with the spiked covariance model

In this section, we point out a connection between the spectral algorithm presented in Section IV and the spiked covariance model well-known in statistics [17], [18], [19], [20], [21]. Though somewhat informal, this analysis will provide insight into how the complexity of recovering x depends on the dimension L when the distribution ρ has a fixed support size.

In the spiked model, we observe a matrix

$$\mathbf{Y} = \mathbf{X} + \mathbf{G} \in \mathbb{R}^{L \times N}, \tag{V.1}$$

where $\mathbf{X}$ is a rank r matrix and

$$\mathbf{G} = (G_{ij}), \quad G_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

This model is typically studied in the *high-dimensional regime*, in which L grows proportionally to N; that is, $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$. In this setting, there is a precise understanding of the limiting behavior of the data matrix $\mathbf{Y}$ and the low-rank matrix $\mathbf{X} = [X_1, \ldots, X_N]$.

In [20] (see also [18]), it is shown that when the low-rank matrix $\mathbf{X}$ is random (for instance, its columns may be drawn from a suitable low-rank, mean-zero distribution), then the limiting cosine c of the angles between the top eigenvector of $\mathbf{X}\mathbf{X}^T$ and the top eigenvector of $\mathbf{Y}\mathbf{Y}^T$ is given by the formula:

$$c^2 = \begin{cases} \frac{1-\sigma^4\gamma/\lambda^2}{1+\sigma^2\gamma/\lambda} & \text{if } \lambda > \sigma^2\sqrt{\gamma}, \\ 0 & \text{otherwise,} \end{cases} \tag{V.2}$$

where λ is the top eigenvalue of $\mathbf{X}\mathbf{X}^T/N$.

The key phenomenon is the phase transition at

$$\lambda_{critical} = \sigma^2\sqrt{\gamma}. \tag{V.3}$$

It is only when λ is greater than this critical value that we are guaranteed a non-trivial correlation between the top eigenvector of the observed matrix $\mathbf{Y}\mathbf{Y}^T/N$ and the top eigenvector of $\mathbf{X}\mathbf{X}^T/N$.

We can view the observation model in the one-dimensional MRA model (I.1) as a special instance of the spiked model, by taking the ith column of $\mathbf{X}$ to be $X_i = R_{S_i}x$. As $N \to \infty$, we can write

$$\frac{1}{N}\mathbf{X}\mathbf{X}^T = C_x D_\rho C_x^T. \tag{V.4}$$

Consequently, under the assumption that the DFT of x does not vanish, the rank of $\mathbf{X}$ is the size of the support of ρ. When the support size of ρ is fixed at r, the MRA problem is an instance of the spiked model.

Let us assume that the $|(Fx)[k]| = 1$ for all k. This can be done by estimating the power spectrum first and then normalizing all Fourier coefficients. In this case, C_x is an orthogonal matrix. In other words, $x \perp R_\ell x$ for every $\ell \neq 0$; consequently, the $R_\ell x$ are precisely the top r eigenvectors of $\mathbf{X}\mathbf{X}^T/N$, with corresponding eigenvalues $\|x\|^2\rho[\ell]$. Then, (V.2) tells us exactly how well we expect the spectral algorithm to perform in recovering x; indeed, the theory predicts a non-zero angle between x and the top eigenvector of $\mathbf{Y}\mathbf{Y}^T/N$ whenever:

$$N \geq \frac{L\sigma^4}{\|x\|^4(\max\rho)^2} = \frac{L}{(\max\rho)^2}\frac{1}{\mathrm{SNR}^2}. \tag{V.5}$$

Below this threshold, the output will be essentially random. We see that if the distribution is well-localized, then $\max\rho = \Omega(1)$ (with respect to the growing value of L) and then the sample complexity grows like $\frac{L}{\mathrm{SNR}^2}$. On the other hand, if the distribution is almost uniform, then $\max\rho = \mathcal{O}(1/L)$ as $L \to \infty$, and thus the sample complexity will be proportional to L^3/SNR^2.

To illustrate the relationship between the spiked model and MRA, we ran the following experiment. We generated a signal $x \in \mathbb{R}^{400}$ with i.i.d. normal entries and normalized it so that $\|x\|_2 = 10$. For noise levels σ between 0.1 and 10, we drew N samples of x with noise at level σ, where N is chosen at 100 plus the critical threshold given by (V.5) for $\sigma = \lambda^{1/2}\gamma^{-1/4} = 5.5313$ according to (V.3). For σ large enough, N will not be large enough for the spectral method to produce an estimate better than random. The distribution of translations ρ was taken to be $\rho[i] \propto i^2$, for $i = 1, \ldots, 5$, and zero elsewhere. Each experiment was repeated 200 times. The plots in Figure V.1 display the average values over these 200 runs.

For each draw, we compute the top eigenvalue of the clean data matrix (V.4), denoted by λ, and the associated eigenvector, which is a translated copy of x. We also compute the top eigenvector of the data matrix $\mathbf{Y}\mathbf{Y}^T/N$. The angle between the two eigenvectors is predicted by (V.2). In Figure V.1(a), we plot the predicted cosine against the true cosine. Clearly, we never attain the predicted value of zero in finite samples, but we see a precipitous decline when the noise level σ exceeds its threshold value (the vertical dashed line).

We also measure the relative mean squared error defined by equation (I.2), where $\widehat{X}$ is the top eigenvector multiplied by $\|x\|$. In Figure V.1(b), we plot this error as a function of σ. For reference, we also plot the ordinary error predicted by the spiked model (as derived from the predicted cosine between the vectors), without minimizing over shifts. Of course, minimizing over shifts will decrease the error; however, we still see the same qualitative behavior predicted from the spiked model, namely an increase in error as σ grows, until the critical threshold of σ is reached, after which the error plateaus.

VI. Additional Algorithms

While the spectral algorithm (Algorithm 2) is asymptotically optimal as σ, $N \to \infty$ and for signals with non-vanishing DFT, it may not perform well in small sample size or low DFT values. Therefore, in this section, we present two additional algorithms based on non-convex least-squares minimization and a modification of the EM algorithm presented in Section II that takes the distribution into account. In Appendix K, we also describe and analyze a convex relaxation approach based on semidefinite programming.

A. Non-convex least-squares minimization

The following method aims to find a signal in $\mathbb{R}^L$ and a distribution in Δ^L that fit the observed data as well as possible in the LS sense. We formulate the problem as a smooth, non-convex, optimization problem with the constraint that the distribution lies on a simplex. Given estimators $\widehat{M}^1$ and $\widehat{M}^2$ of the first two moments M^1 and M^2, the problem reads

$$\min_{\tilde{x}\in\mathbb{R}^L,\tilde{\rho}\in\Delta^L} \|\widehat{M}^2 - C_{\tilde{x}} D_{\tilde{\rho}} C_{\tilde{x}}^T\|_{\mathrm{F}}^2 + \lambda\|\widehat{M}^1 - C_{\tilde{x}}\tilde{\rho}\|_2^2, \tag{VI.1}$$

where $\lambda > 0$ is a predefined parameter. It can be verified that, by omitting signal-dependent terms, the variance of the elements of the first moment estimator is proportional to σ^2. It can be also shown that the variance of the elements of the second moment is proportional to $3L\sigma^4$ and $L\sigma^2$ in the low and the high SNR regimes, respectively (again, by omitting signal-dependent terms). Therefore, we set $\lambda = \frac{1}{L(1+3\sigma^2)}$ in our implementation.

B. An expectation-maximization algorithm for estimating x and ρ simultaneously

In Section II, we reviewed the EM algorithm for MRA from [15], which is invariant to the distribution of translations. In this section, we modify the algorithm to take the distribution into account. A similar approach was introduced for the application of cryo-EM in [41].

If we denote $s := \{s_j\}_{1\le j\le N}$, the forward model of the MRA model (I.1) reads:

$$\begin{aligned} f_{x,\rho}(y,s) &= f_{x,\rho}(y|s)\prod_{j=1}^N \rho[s_j] \\ &= \prod_{j=1}^N \rho[s_j]\frac{1}{(2\pi\sigma^2)^{L/2}} e^{-\frac{1}{2\sigma^2}\left\|R_{s_j}x-y_j\right\|^2}. \end{aligned}$$

The log-likelihood function is then given, up to a constant, by

$$\log\mathcal{L}(y,s|x,\rho) = \sum_{j=1}^N \left\{\log\rho[s_j] - \frac{1}{2\sigma^2}\left\|R_{s_j}x-y_j\right\|^2\right\}.$$

The goal of the EM algorithm is to compute the maximum in x,ρ of the marginal likelihood $\mathcal{L}(y|x,\rho)=\sum_s \mathcal{L}(y,s|x,\rho)$. The algorithm proceeds as follows. Start with some initial guesses x_0 and ρ_0 for the signal and distribution. Given x_k and ρ_k, the next guess is given as follows:

$$(x_{k+1},\rho_{k+1}) = \arg\max_{x,\rho} Q(x,\rho|x_k,\rho_k),$$

where

$$Q(x,\rho|x_k,\rho_k) := \mathbb{E}\left[\log\mathcal{L}\left(y,S^k|x,\rho\right)\right]. \tag{VI.2}$$

Here the distribution S^k depends on x_k and ρ_k through

$$w_k^{\ell,j} := \mathbb{P}[S_j^k=\ell] = C_k^j e^{-\frac{1}{2\sigma^2}\|R_\ell x_k - y_j\|^2}\rho_k[\ell],$$

where C_k^j is a normalization term so that $\sum_\ell w_k^{\ell,j}=1$. We can explicitly write (VI.2) (omitting the constant term) as

$$\begin{aligned} &Q(x,\rho|x_k,\rho_k) \\ &\quad= \sum_{j=1}^N \mathbb{E}\left[\log\rho[S_j^k] - \frac{1}{2\sigma^2}\|R_{S_j^k}x-y_j\|^2\right] \\ &\quad= \sum_{j=1}^N\sum_{\ell=0}^{L-1} w_k^{\ell,j}\left\{\log\rho[\ell] - \frac{1}{2\sigma^2}\|R_\ell x - y_j\|^2\right\}, \end{aligned}$$

To maximize Q over x and ρ is simple, since the first term depends only on ρ and the second term depends only on x. Specifically, it is easy to see that the maximum over x is given by a weighted average of the translated observations:

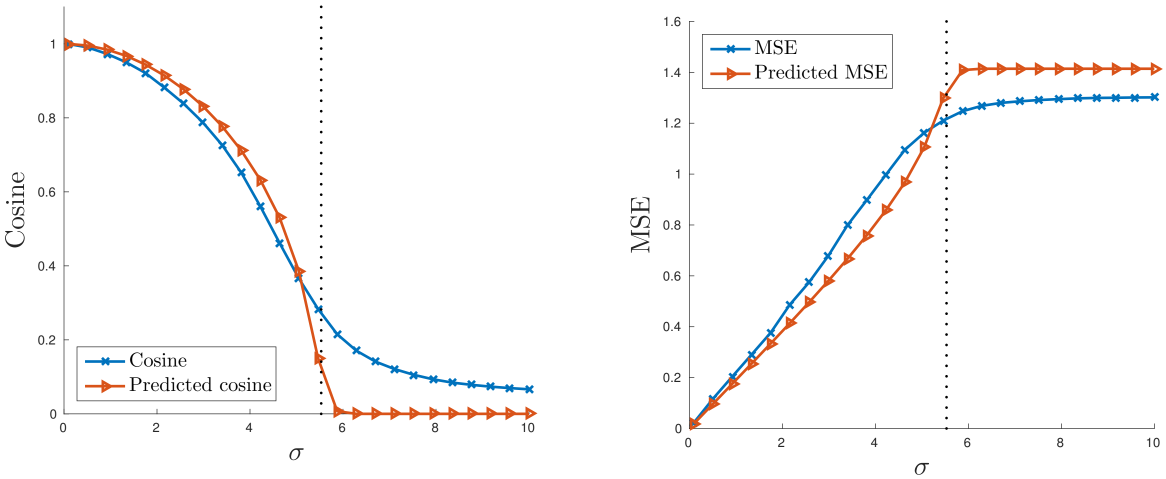
$$x_{k+1} = \frac{1}{N}\sum_{j=1}^N\sum_{\ell=0}^{L-1} w_k^{\ell,j} R_\ell^{-1} y_j. \tag{VI.3}$$

This step is almost identical (up to the values of the weights) to the standard EM update step (II.2).

The maximimizing value of ρ also has a closed formula. First, observe that we can write:

$$\rho_{k+1} = \arg\max_{\rho\in\Delta^L}\sum_{\ell=0}^{L-1} W_k[\ell]\log(\rho[\ell]),$$

where $W_k[\ell] = \sum_{j=1}^N w_k^{\ell,j}$. To maximize a positive weighted combination of logarithms over the simplex, we use the following lemma:



(a) Empirical cosines between the top eigenvectors of the matrices $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ and $\frac{1}{N}\mathbf{Y}\mathbf{Y}^T$ as a function of the noise level compared to asymptotic cosines predicted by spiked model; see (V.2).

(b) Empirical MSE compared to the asymptotic MSE predicted by spiked model. The MSE is defined in (I.2).

Fig. V.1. Experiments related to the connection between the spike model and the MRA problem as discussed in Section V. The dashed line is the predicted threshold value of $\sigma = \lambda^{1/2}\gamma^{-1/4} = 5.5313$.

Lemma VI.1. *If $w[\ell] > 0$ are positive weights, then the maximizer of $\sum_\ell w[\ell]\log(q[\ell])$ over all $q \in \Delta^L$ is*

$$q^*[\ell] = w[\ell] / \sum_{\ell'} w[\ell'].$$

Proof. See Appendix J. □

From this lemma, the maximizing ρ is given by the formula:

$$\rho_{k+1}[\ell] = \frac{W_k[\ell]}{\sum_{\ell'=0}^{L-1} W_k[\ell']}. \tag{VI.4}$$

To conclude, the modified EM updates the signal and the distribution estimations by (VI.3) and (VI.4), respectively. However, compared to the methods which are based on moments estimation like Algorithm 2 or the LS, it passes through the data at each iteration. Therefore, for large sample size, its computational cost may be substantially heavier.

VII. Numerical experiments

In this section, we present numerical results for the algorithms described in Section VI and Algorithm 2. To measure the accuracy of an estimator $\widehat{X}$, we define the recovery relative error as

$$\text{relative error} = \min_{s \in \mathbb{Z}_L} \frac{\|R_s \widehat{X} - x\|_2}{\|x\|_2}. \tag{VII.1}$$

The code of this section, including Matlab implementations and examples, is publicly available online [1].

A. Influence of the number of samples

In the first example, we use a Haar-like signal of length $L = 20$, depicted in Figure VII.1(a). Next, we generate its noisy, translated copies according to the MRA model (I.1), with noise variance of $\sigma = .25$. One example of a data sample corrupted with such noise is illustrated in Figure VII.1(b). We use the EM algorithm of Section VI-B to estimate the signal. This process is repeated three times for different number of samples, $N = 10^3$, $N = 10^5$, and $N = 10^7$. The estimates are presented in Figure VII.1(c)–VII.1(e). As expected, the quality of the estimation improves significantly as N grows.

B. Comparison of EM algorithms

In [15], it is shown that in most cases, an EM method as described in Section II-B, achieves the smallest estimation error compared to the competitor algorithms. The EM algorithm described in that paper is invariant to the distribution ρ. In particular, it treats the data as if it were drawn from the uniform distribution, which requires sample complexity that grows like $\omega(1/\operatorname{SNR}^3)$ rather than $\omega(1/\operatorname{SNR}^2)$. By contrast, the EM algorithm we propose in Section VI-B also estimates the distribution ρ at each iteration. The updated estimation of the distribution is then used to update the signal's estimation.

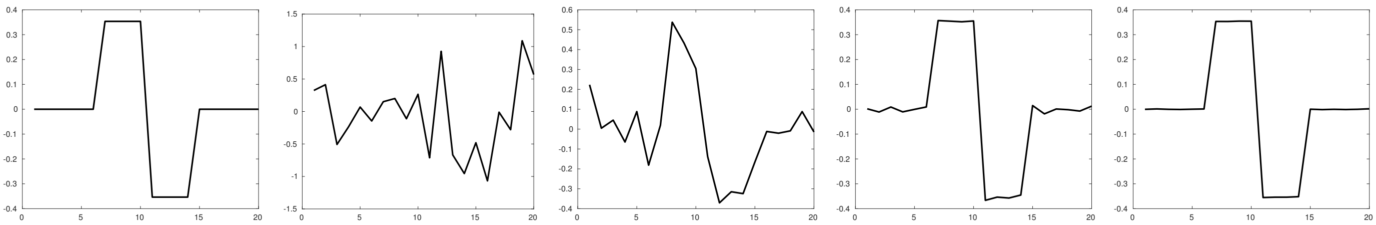
To demonstrate the importance of including the distribution into the model of the estimator, we consider a family of distributions

$$\rho[t] \propto \exp(-t^2/s^2) \tag{VII.2}$$

where the parameter $s > 0$ controls the concentration of ρ, or alternatively its uniformity: the larger s is, the more uniform ρ is. In general, we expect our algorithms to provide better estimations when s is smaller, i.e., when ρ is more concentrated; see Section V.

We compared the standard EM with the EM algorithm described in Section VI-B. The experiments were conducted as follows. We fixed a random signal of length $L = 25$ with i.i.d. normal entries and unit norm, and a series of distributions of the form (VII.2) with the parameter s varying between 3 and 9. Then, for each distribution we generated $N = 2,000$ samples drawn with a fixed level of noise $\sigma = 1$. We repeated the test independently 20 times and averaged the errors. In Figure VII.2, we plot the relative errors of

[1] https://github.com/nirsharon/aperiodicMRA



(a) The clean signal

(b) An example of a data sample (one measurement)

(c) An estimate with $N = 1,000$

(d) An estimate with $N = 100,000$

(e) An estimate with $N = 10,000,000$

Fig. VII.1. An example of the estimation quality of a Haar-like signal with different number of samples (N), using the LS method. In these tests, $\sigma = 1$.

the methods as a function of the uniformity parameter s. As expected, the standard EM is invariant to s. On the other hand, the adapted version of the EM exploits the varying distribution and performs better under more concentrated distributions. As the distribution becomes more uniform, the two methods exhibit similar error rates.

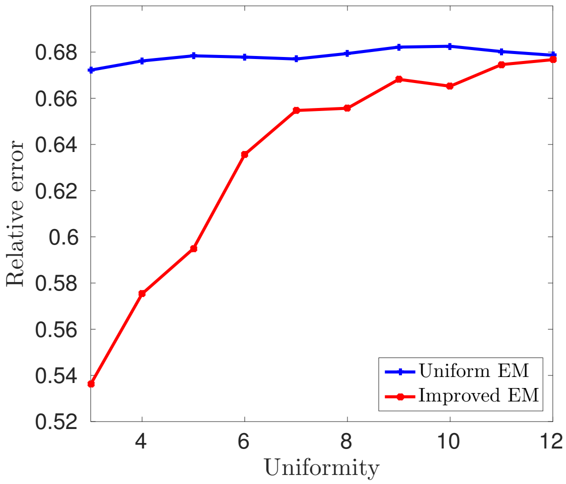


Fig. VII.2. EM comparison: the standard EM described in Section II (uniform EM) versus the EM that includes distribution estimation (improved EM) described in Section VI-B. The algorithms were compared with different distributions of the form VII.2 as a function of the parameter s.

C. Comparison of the different methods

This paper presents three main approaches for solving the MRA: the spectral method described in Algorithm 2, the LS optimization of Section VI-A, and the EM of Section VI-B. In this comparison, we examined the estimation error of these three methods with different noise levels. In detail, we use a random signal of length $L = 25$ with i.i.d. normal entries with unit norm, and a random distribution. We fix the number of samples to be $N = 10,000$. Then, we increase the level of noise σ from 0.001 to 1. In Figure VII.3 we plot the average error. As can be seen, the LS and EM methods are more robust to noise than the spectral method. In addition, the gap between these two methods becomes small as the SNR decreases.

D. Numerical error rates for the EM algorithm

When the distribution ρ is aperiodic, the optimal MSE for recovering x in the low SNR, or large σ, and large N regime is of size $O(\sigma^4/N)$. Since the relative error scales as $\sqrt{\text{MSE}}$, by (VII.1), if the $\log$-error is viewed as a function of $\log(\sigma)$, the slope is expected to be no smaller than 2 when σ is large.

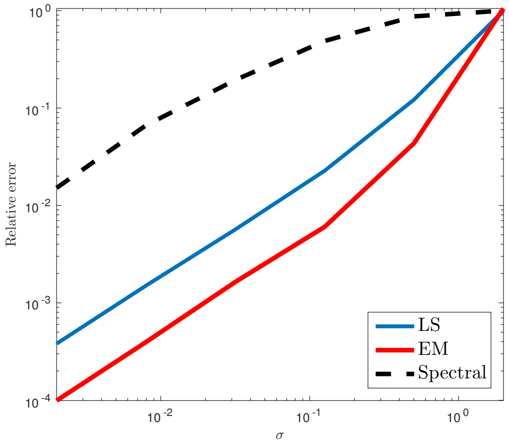


Fig. VII.3. A comparison of three methods: least squares (LS), expectation maximization (EM), and the spectral method, under varying level of noise.

In Figure VII.4 we plot the average log-error of the EM algorithm over 300 trials as a function of $\log(\sigma)$. In each trial, we used the EM algorithm to estimate a randomly generated signal, with translations drawn from a randomly generated probability distribution. When σ is large, the curve is indeed a line with slope close to 2, which is the expected rate. However, when σ is small, the curve is a line with slope close to 1; namely, the error behaves approximately like $O(\sigma/\sqrt{N})$, rather than $O(\sigma^2/\sqrt{N})$. The moderate slope for high SNR suggests that in this regime the recovery problem is easier; for example, we know that alignment is possible in high SNR, as described in Section II-A.

In Figure VII.5 we plot the average log-error (again over 300 experiments) as a function of $\log(\sigma)$, but in this case each experiment used the uniform distribution of translations. In this regime, we know from [16] that the optimal slope is 3, not 2; and indeed, when σ is large the curve has slope close to 3. As in the other plot, when σ is small the curve has slope close to 1. Taken together, these two experiments suggest that the EM algorithm exhibits near-optimal behavior for both periodic and aperiodic distributions.

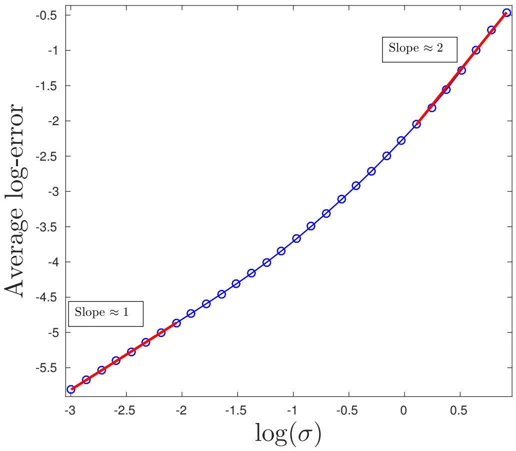


Fig. VII.4. Log-log plot of the error of the EM method versus σ, with random distributions.

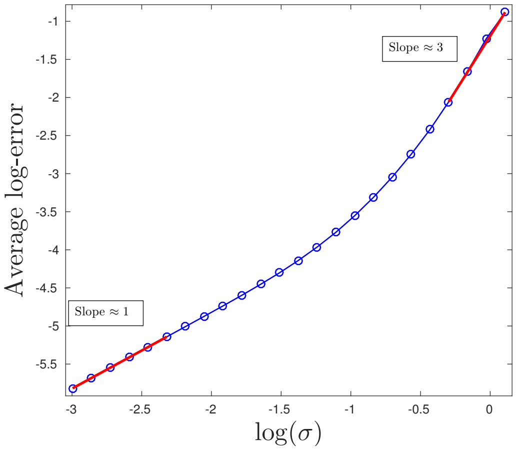


Fig. VII.5. Log-log plot of the error of the EM method versus σ, with uniform distribution

VIII. Discussion

In this paper, we have shown that the sample complexity for MRA with an aperiodic distribution of translations grows like $\omega(1/\operatorname{SNR}^2)$. This sample complexity can be achieved by a simple spectral algorithm. We also examined empirically the LS and EM algorithms. Additionally, we extended previous works by showing that the sample complexity for any periodic distribution scales as $\omega(1/\operatorname{SNR}^3)$.

We drew connections between the MRA problem and the spiked covariance model. This connection implies that the sample complexity is inversely proportional to the square of the maximal value of the distribution. Therefore, the more uniform the distribution is, the higher the sample complexity of the problem.

One of the motivations for considering the MRA model arises from the imaging technique called single particle cryo–electron microscopy (cryo–EM), allowing to visualize molecules at near-atomic resolution [12], [13]. In cryo–EM, noisy two-dimensional tomographic projections of the three-dimensional underlying molecule, taken at unknown viewing direction, are collected. The distribution of viewing directions in cryo–EM is typically non-uniform, as many molecules exhibit some preferred orientation [42].

The MRA model (I.1) can be thought of as a simplified model for the cryo–EM problem, where cyclic translations replace actions of elements of the group $SO(3)$. The tomographic projection does not appear in (I.1). Our technique for MRA, based on the low-order moments of the data, is similar to the framework proposed by Zvi Kam in [43], [44] for cryo–EM. In particular, Kam suggested a method to estimate a molecule directly from the statistics of the projections, rather than estimating the viewing directions. Our work is one step towards understanding the sample complexity of Kam's method in particular, and the cryo–EM problem in general.

Acknowledgments

We would like to thank Afonso Bandeira, Nicolas Boumal, Joseph Kileel, Roy Lederman and Zhizhen Zhao for many insightful discussions.

References

[1] R. Diamond, "On the multiple simultaneous superposition of molecular structures by rigid body transformations," *Protein Science*, vol. 1, no. 10, pp. 1279–1287, 1992.

[2] D. L. Theobald and P. A. Steindel, "Optimal simultaneous superpositioning of multiple structures with missing data," *Bioinformatics*, vol. 28, no. 15, pp. 1972–1979, 2012.

[3] W. Park, C. R. Midgett, D. R. Madden, and G. S. Chirikjian, "A stochastic kinematic model of class averaging in single-particle electron microscopy," *The International journal of robotics research*, vol. 30, no. 6, pp. 730–754, 2011.

[4] W. Park and G. S. Chirikjian, "An assembly automation approach to alignment of noncircular projections in electron microscopy," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 668–679, 2014.

[5] S. H. Scheres, M. Valle, R. Nuñez, C. O. Sorzano, R. Marabini, G. T. Herman, and J.-M. Carazo, "Maximum-likelihood multi-reference refinement for electron microscopy images," *Journal of molecular biology*, vol. 348, no. 1, pp. 139–149, 2005.

[6] J. P. Zwart, R. van der Heiden, S. Gelsema, and F. Groen, "Fast translation invariant classification of HRR range profiles in a zero phase representation," *IEE Proceedings-Radar, Sonar and Navigation*, vol. 150, no. 6, pp. 411–418, 2003.

[7] R. Gil-Pita, M. Rosa-Zurera, P. Jarabo-Amores, and F. López-Ferreras, "Using multilayer perceptrons to align high range resolution radar signals," in *International Conference on Artificial Neural Networks*, pp. 911–916, Springer, 2005.

[8] D. M. Rosen, L. Carlone, A. S. Bandeira, and J. J. Leonard, "A certifiably correct algorithm for synchronization over the special euclidean group," *arXiv preprint arXiv:1611.00128*, 2016.

[9] I. L. Dryden and K. V. Mardia, *Statistical shape analysis*, vol. 4. J. Wiley Chichester, 1998.

[10] H. Foroosh, J. B. Zerubia, and M. Berthod, "Extension of phase correlation to subpixel registration," *IEEE transactions on image processing*, vol. 11, no. 3, pp. 188–200, 2002.

[11] D. Robinson, S. Farsiu, and P. Milanfar, "Optimal registration of aliased images using variable projection with applications to super-resolution," *The Computer Journal*, vol. 52, no. 1, pp. 31–42, 2009.

[12] A. Bartesaghi, A. Merk, S. Banerjee, D. Matthies, X. Wu, J. L. Milne, and S. Subramaniam, "2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor," *Science*, vol. 348, no. 6239, pp. 1147–1151, 2015.

[13] D. Sirohi, Z. Chen, L. Sun, T. Klose, T. C. Pierson, M. G. Rossmann, and R. J. Kuhn, "The 3.8 Å resolution cryo-EM structure of Zika virus," *Science*, vol. 352, no. 6284, pp. 467–470, 2016.

[14] C. Aguerrebere, M. Delbracio, A. Bartesaghi, and G. Sapiro, "Fundamental limits in multi-image alignment," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5707–5722, 2016.

[15] T. Bendory, N. Boumal, C. Ma, Z. Zhao, and A. Singer, "Bispectrum inversion with application to multireference alignment," *IEEE Transactions on Signal Processing*, vol. 66, pp. 1037–1050, Feb 2018.

[16] A. Bandeira, P. Rigollet, and J. Weed, "Optimal rates of estimation for multi-reference alignment," *arXiv preprint arXiv:1702.08546*, 2017.

[17] I. M. Johnstone, “On the distribution of the largest eigenvalue in principal components analysis,” *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[18] D. Paul, “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model,” *Statistica Sinica*, vol. 17, no. 4, pp. 1617–1642, 2007.

[19] M. Gavish and D. L. Donoho, “Optimal shrinkage of singular values,” *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2137–2152, 2017.

[20] F. Benaych-Georges and R. R. Nadakuditi, “The singular values and vectors of low rank perturbations of large rectangular random matrices,” *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, 2012.

[21] E. Dobriban, W. Leeb, and A. Singer, “Optimal prediction in the linearly transformed spiked model,” *arXiv preprint arXiv:1709.03393*, 2017.

[22] W. James and C. Stein, “Estimation with quadratic loss,” in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379, 1961.

[23] B. Efron and C. Morris, “Stein’s estimation rule and its competitors-an empirical Bayes approach,” *Journal of the American Statistical Association*, vol. 68, no. 341, pp. 117–130, 1973.

[24] B. Efron and C. Morris, “Data analysis using Stein’s estimator and its generalizations,” *Journal of the American Statistical Association*, vol. 70, no. 350, pp. 311–319, 1975.

[25] A. Singer, “Angular synchronization by eigenvectors and semidefinite programming,” *Applied and computational harmonic analysis*, vol. 30, no. 1, pp. 20–36, 2011.

[26] N. Boumal, “Nonconvex phase synchronization,” *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2355–2377, 2016.

[27] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, “Message-passing algorithms for synchronization problems over compact groups,” *arXiv preprint arXiv:1610.04583*, 2016.

[28] Y. Chen and E. Candes, “The projected power method: An efficient algorithm for joint alignment from pairwise differences,” *arXiv preprint arXiv:1609.05820*, 2016.

[29] A. S. Bandeira, N. Boumal, and A. Singer, “Tightness of the maximum likelihood semidefinite relaxation for angular synchronization,” *Mathematical Programming*, vol. 163, no. 1, pp. 145–167, 2017.

[30] Y. Zhong and N. Boumal, “Near-optimal bounds for phase synchronization,” *arXiv preprint arXiv:1703.06605*, 2017.

[31] A. S. Bandeira, Y. Chen, and A. Singer, “Non-unique games over compact groups and orientation estimation in cryo-em,” *arXiv preprint arXiv:1505.03840*, 2015.

[32] A. S. Bandeira, M. Charikar, A. Singer, and A. Zhu, “Multireference alignment using semidefinite programming,” in *Proceedings of the 5th conference on Innovations in theoretical computer science*, pp. 459–470, ACM, 2014.

[33] Y. Chen, L. Guibas, and Q. Huang, “Near-optimal joint object matching via convex relaxation,” in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 100–108, 2014.

[34] A. S. Bandeira, N. Boumal, and V. Voroninski, “On the low-rank approach for semidefinite programs arising in synchronization and community detection,” in *Conference on Learning Theory*, pp. 361–382, 2016.

[35] A. Perry, J. Weed, A. Bandeira, P. Rigollet, and A. Singer, “The sample complexity of multi-reference alignment,” *arXiv preprint at arXiv:1707.00943*, 2017.

[36] N. Boumal, T. Bendory, R. R. Lederman, and A. Singer, “Heterogeneous multireference alignment: a single pass approach,” *arXiv preprint arXiv:1710.02590*, 2017.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[38] D. G. Chapman and H. Robbins, “Minimum variance estimation without regularity assumptions,” *Ann. Math. Statist.*, vol. 22, pp. 581–586, 12 1951.

[39] H. Cramér, *Mathematical Methods of Statistics (PMS-9)*, vol. 9. Princeton university press, 2016.

[40] Y. Yu and R. J. Samworth, “A useful variant of the Davis-Kahan theorem for statisticians,” *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015.

[41] N. C. Dvornek, F. J. Sigworth, and H. D. Tagare, “SubspaceEM: A fast maximum-a-posteriori algorithm for cryo-EM single particle reconstruction,” *Journal of structural biology*, vol. 190, no. 2, pp. 200–214, 2015.

[42] M. Radermacher, T. Wagenknecht, A. Verschoor, and J. Frank, “Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli,” *Journal of Microscopy*, vol. 146, no. 2, pp. 113–136, 1987.

[43] Z. Kam, “The reconstruction of structure from electron micrographs of randomly oriented particles,” *Journal of Theoretical Biology*, vol. 82, no. 1, pp. 15–39, 1980.

[44] E. Levin, T. Bendory, N. Boumal, J. Kileel, and A. Singer, “3d ab initio modeling in cryo-em by autocorrelation analysis,” *arXiv preprint arXiv:1710.08076*, 2017.

[45] R. Vershynin, “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, 2010.

[46] M. X. Goemans and D. P. Williamson, “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming,” *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, 1995.

[47] M. Grant, S. Boyd, and Y. Ye, “CVX: Matlab software for disciplined convex programming,” 2008.

Appendix

A. Proof of Theorem III.4

The proof mimics the one of the classical Chapman and Robbins bound. Recalling equation (III.7) and the definition of positive semidefinite matrices, the statement is equivalent to

$$\mathbb{E}_{x,\rho}\left[\left(w^T(\phi_x(\widehat{X}) - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})])\right)^2\right] \geq \frac{\left[w^T\left(\mathbb{E}_{\tilde{x},\tilde{\rho}}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]\right)\right]^2}{\chi^2(f^N_{\tilde{x},\tilde{\rho}}||f^N_{x,\rho})}, \quad (A.1)$$

for all $w, \tilde{x} \in \mathbb{R}^L$ and $\tilde{\rho} \in \Delta^L$. Define

$$Z = \frac{f_{\tilde{x},\tilde{\rho}}(Y)}{f_{x,\rho}(Y)}.$$

and note that

- $\mathbb{E}_{x,\rho}[g(Y)Z] = \mathbb{E}_{\tilde{x},\tilde{\rho}}[g(Y)]$,
- $\mathbb{E}_{x,\rho}[Z-1] = 0$,
- $\mathbb{E}_{x,\rho}[(Z-1)^2] = \chi^2(f^N_{\tilde{x},\tilde{\rho}}||f^N_{x,\rho})$.

We have

$$\begin{aligned} w^T &\left(\mathbb{E}_{\tilde{x},\tilde{\rho}}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]\right) \\ &= \mathbb{E}_{\tilde{x},\tilde{\rho}}[w^T\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[w^T\phi_x(\widehat{X})] \\ &= \mathbb{E}_{x,\rho}[w^T\phi_x(\widehat{X})(Z-1)] \\ &= \mathbb{E}_{x,\rho}\left[w^T\left(\phi_x(\widehat{X}) - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]\right)(Z-1)\right], \end{aligned}$$

and by Cauchy-Schwarz

$$\begin{aligned} &\left[w^T\left(\mathbb{E}_{\tilde{x},\tilde{\rho}}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]\right)\right]^2 \\ &\quad \leq \mathbb{E}_{x,\rho}[(w^T(\phi_x(\widehat{X}) - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]))^2]\chi^2(f^N_{\tilde{x},\tilde{\rho}}||f^N_{x,\rho}). \end{aligned}$$

B. Proof of Lemma III.5

Equation (III.12) follows from some algebraic manipulations:

$$\begin{aligned}
&\chi^2(f_{\tilde{x},\tilde{\rho}}||f_{x,\rho})\\
&\quad= \int_{\mathbb{R}^L} \left(\frac{f_{\tilde{x},\tilde{\rho}}(y;\gamma)}{f_{x,\rho}(y;\gamma)} - 1\right)^2 f_{x,\rho}(y;\gamma)\,dy\\
&\quad= \int_{\mathbb{R}^L} \frac{\left(\sum_{i=0}^{\infty}(\alpha^i_{\tilde{x},\tilde{\rho}}(y) - \alpha^i_{x,\rho}(y))\frac{\gamma^i}{i!}\right)^2}{\sum_{i=0}^{\infty}\alpha^i_{x,\rho}(y)\frac{\gamma^i}{i!}} f_G(y)\,dy\\
&\quad= \int_{\mathbb{R}^L} \frac{\left(\sum_{i=d}^{\infty}(\alpha^i_{\tilde{x},\tilde{\rho}}(y) - \alpha^i_{x,\rho}(y))\frac{\gamma^i}{i!}\right)^2}{1+\sum_{i=1}^{\infty}\alpha^i_{x,\rho}(y)\frac{\gamma^i}{i!}} f_G(y)\,dy\\
&\quad= \frac{\gamma^{2d}}{(d!)^2}\int_{\mathbb{R}^L}\left(\alpha^d_{\tilde{x},\tilde{\rho}}(y) - \alpha^d_{x,\rho}(y)\right)^2 f_G(y)\,dy + O(\gamma^{2d+1})\\
&\quad= \frac{\gamma^{2d}}{(d!)^2}\mathbb{E}\left[\left(\alpha^d_{\tilde{x},\tilde{\rho}}(G) - \alpha^d_{x,\rho}(G)\right)^2\right] + O(\gamma^{2d+1}),
\end{aligned}$$

where the third equation follows from the definition of d, i.e. $\alpha^n_{\tilde{x},\tilde{\rho}}(z) = \alpha^n_{x,\rho}(z)$ almost surely for all $n < d$. Equation (III.12) now follows from $\gamma = 1/\sigma$.

We now prove (III.13). It is enough to show that

$$\mathbb{E}\left[\alpha^d_{\tilde{x},\tilde{\rho}}(G)\alpha^d_{x,\rho}(G)\right] = d!\left\langle M^d_{\tilde{x},\tilde{\rho}}, M^d_{x,\rho}\right\rangle,$$

Let S and $\tilde{S}$ be two independent random variables such that $S \sim \rho$ and $\tilde{S} \sim \tilde{\rho}$. We have

$$\begin{aligned}
\left\langle M^d_{\tilde{x},\tilde{\rho}}, M^d_{x,\rho}\right\rangle &= \left\langle \mathbb{E}[(R_{\tilde{S}}\tilde{x})^{\otimes d}], \mathbb{E}[(R_S x)^{\otimes d}]\right\rangle\\
&= \mathbb{E}\left[\left\langle (R_{\tilde{S}}\tilde{x})^{\otimes d}, (R_S x)^{\otimes d}\right\rangle\right]\\
&= \mathbb{E}\left[\left\langle R_{\tilde{S}}\tilde{x}, R_S x\right\rangle^d\right]. \qquad \text{(B.1)}
\end{aligned}$$

On the other hand, we can write $f_{x,\rho}$ explicitly by

$$\begin{aligned}
f_{x,\rho}(y;\gamma) &= \frac{1}{\sqrt{2\pi}^L}\sum_{\ell=0}^{L-1}\rho[\ell]\exp\left(-\frac{\|y-\gamma R_\ell x\|^2}{2}\right)\\
&= \mathbb{E}[f_G(y-\gamma R_S x)],
\end{aligned}$$

where $S \sim \rho$, thus by equation (III.11)

$$\begin{aligned}
&\mathbb{E}\left[\alpha^d_{\tilde{x},\tilde{\rho}}(G)\alpha^d_{x,\rho}(G)\right]\\
&\quad= \mathbb{E}\left[\frac{\partial^d}{\partial\tilde{\gamma}^d}\left(\frac{f_{\tilde{x},\tilde{\rho}}(G;\tilde{\gamma})}{f_G(G)}\right|_{\tilde{\gamma}=0}\frac{\partial^d}{\partial\gamma^d}\left(\frac{f_{x,\rho}(G;\gamma)}{f_G(G)}\right|_{\gamma=0}\right]\\
&\quad= \frac{\partial^{2d}}{\partial\tilde{\gamma}^d\partial\gamma^d}\mathbb{E}\left[\frac{f_{\tilde{x},\tilde{\rho}}(G;\tilde{\gamma})}{f_G(G)}\frac{f_{x,\rho}(G;\gamma)}{f_G(G)}\right]_{\tilde{\gamma},\gamma=0}\\
&\quad= \frac{\partial^{2d}}{\partial\tilde{\gamma}^d\partial\gamma^d}\mathbb{E}\left[\frac{f_G(G-\tilde{\gamma}R_{\tilde{S}}\tilde{x})}{f_G(G)}\frac{f_G(G-\gamma R_S x)}{f_G(G)}\right]_{\tilde{\gamma},\gamma=0},
\end{aligned}$$

where S and $\tilde{S}$ are defined as in (B.1). We have

$$\begin{aligned}
&\mathbb{E}\left[\frac{f_G(G-\tilde{\gamma}R_{\tilde{S}}\tilde{x})}{f_G(G)}\frac{f_G(G-\gamma R_S x)}{f_G(G)}\middle|\tilde{S},S\right]\\
&= \frac{1}{\sqrt{2\pi}^L}\int_{\mathbb{R}^L}\exp\left(-\frac{\|z-\tilde{\gamma}R_{\tilde{S}}\tilde{x}\|^2 + \|z-\gamma R_S x\|^2 - \|z\|^2}{2}\right)dz\\
&= \frac{1}{\sqrt{2\pi}^L}\int_{\mathbb{R}^L}\exp\left(-\frac{\|z-\tilde{\gamma}R_{\tilde{S}}\tilde{x}-\gamma R_S x\|^2}{2} + \gamma\tilde{\gamma}\left\langle R_{\tilde{S}}\tilde{x}, R_S x\right\rangle\right)dz\\
&= \exp\left(\gamma\tilde{\gamma}\left\langle R_{\tilde{S}}\tilde{x}, R_S x\right\rangle\right).
\end{aligned}$$

The proof of (III.13) finally follows from equation (B.1) and

$$\begin{aligned}
&\mathbb{E}\left[\alpha^d_{\tilde{x},\tilde{\rho}}(G)\alpha^d_{x,\rho}(G)\right]\\
&\qquad\qquad= \mathbb{E}\left[\frac{\partial^{2d}}{\partial\tilde{\gamma}^d\partial\gamma^d}\exp\left(\gamma\tilde{\gamma}\left\langle R_{\tilde{S}}\tilde{x}, R_S x\right\rangle\right)\right]_{\tilde{\gamma},\gamma=0}\\
&\qquad\qquad= d!\,\mathbb{E}\left[\left\langle R_{\tilde{S}}\tilde{x}, R_S x\right\rangle^d\right].
\end{aligned}$$

C. Analog results for derivatives

This section provides analog results to the ones presented in section III, but involving the limit $(\tilde{x},\tilde{\rho}) \to (x,\rho)$. More specifically, we will take $(\tilde{x},\tilde{\rho}) = (x + hz, \rho + h\theta)$, and study the limit $h \to 0$. For the rest of the section, identify $v = (z,\theta) \in \mathbb{R}^{2L}$. Since $\rho + h\theta$ has to be a probability distribution, we require that $\mathbf{1}^T\theta = 0$ and $\theta[i] \geq 0$ whenever $\rho[i] = 0$.

In comparison with section III, where we used the χ^2 divergence and the moment tensors, in this section we use the Fisher information matrix and directional derivatives of the moment tensors, respectively. We define the Fisher information matrix as the $2L \times 2L$ matrix such that

$$\Gamma^N_{x,\rho} := \mathrm{Cov}[\nabla \log f^N_{x,\rho}].$$

Here $\nabla \log f^N_{x,\rho} \in \mathbb{R}^{2L}$, since there is a component that depends on x and one that depends on ρ. The Fisher information matrix is also the Hessian of the χ^2 divergence, i.e.,

$$\lim_{h\to 0}\frac{\chi^2(f^N_{x+hz,\rho+h\theta}||f^N_{x,\rho})}{h^2} = v^T\Gamma^N_{x,\rho}v. \qquad \text{(C.1)}$$

The Fisher information matrix of N observations is related to the one observation version by

$$\Gamma^N_{x,\rho} = N\Gamma_{x,\rho}. \qquad \text{(C.2)}$$

We define the Jacobian $J_{x,\rho}$ as the $L \times 2L$ matrix such that

$$J_{x,\rho}v = \lim_{h\to 0}\frac{\mathbb{E}_{x+hz,\rho+h\theta}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]}{h}. \qquad \text{(C.3)}$$

We also define the directional derivative of $M^d_{x,\rho}$ along $v = (z,\theta)$ as the d-dimensional tensor

$$\nabla_v M^d_{x,\rho} := \lim_{h\to 0}\frac{M^d_{x+hz,\rho+h\theta} - M^d_{x,\rho}}{h}.$$

This derivative always exists, an explicit formula for $\nabla_v M^d_{x,\rho}$ is given in Lemma D.1. The next corollary is an analog of the Cramér-Rao bound for estimation of an orbit in MRA.

Corollary C.1. *For any $v = (z, \theta) \in \mathbb{R}^{2L}$, such that $\mathbf{1}^T\theta = 0$ and $\theta[i] \geq 0$, whenever $\rho[i] = 0$, we have*

$$\mathrm{Cov}[\phi_x(\widehat{X})] \succeq \frac{J_{x,\rho} v v^T J_{x,\rho}^T}{N v^T \Gamma_{x,\rho} v}.$$

Proof. If θ is under the hypothesis of the theorem, then there exists $h_0 > 0$ such that for all $0 \leq h \leq h_0$, $\rho + h\theta \in \Delta^L$. Letting $(\tilde{x}, \tilde{\rho}) = hv + (x, \rho)$ in Theorem III.4 we obtain for any $w \in \mathbb{R}^L$

$$\begin{aligned} w^T \mathrm{Cov}[\phi_x(\widehat{X})] w & \\ &\geq \lim_{h \to 0} \frac{(w^T(\mathbb{E}_{x+hz,\rho+h\theta}[\phi_x(\widehat{X})] - \mathbb{E}_{x,\rho}[\phi_x(\widehat{X})]))^2}{\chi^2_N(f_{x+hz,\rho+h\theta} || f_{x,\rho})} \\ &= \frac{(w^T J_{x,\rho} v)^2}{N v^T \Gamma_{x,\rho} v}, \end{aligned}$$

by equations (C.1), (C.2) and (C.3), and the corollary follows. □

We now use (III.10) to give an expression of the Fisher information in terms of the directional derivative of the tensor moments.

Lemma C.2. *For any $v = (z, \theta) \in \mathbb{R}^{2L}$,*

$$v^T \Gamma_{x,\rho} v = \frac{\sigma^{-2d}}{(d!)^2} \mathbb{E}\left[\left(v^T \nabla \alpha^d_{x,\rho}(G)\right)^2\right] + O(\sigma^{-2d-1}), \tag{C.4}$$

$$= \frac{\sigma^{-2d}}{d!} \|\nabla_v M^d_{x,\rho}\|^2 + O(\sigma^{-2d-1}), \tag{C.5}$$

where $d = \inf\left\{n : \|\nabla_v M^n_{x,\rho}\|^2 > 0\right\}$.

Proof. In this case we cannot just take the limit $h \to 0$ in (III.12), since the term contained in $O(\sigma^{-2d-1})$ might blow up. Instead we proceed by doing similar algebraic manipulations. Recall that $\nabla f_{x,\rho}(y;\gamma)$ and $\nabla \alpha^i_{x,\rho}(y)$ are in $\mathbb{R}^{2L}$, with $v^T \nabla f_{x,\rho}(y;\gamma)$ being the directional derivative of $f_{x,\rho}(y;\gamma)$ in the direction $v = (z, \theta)$. We have

$$\begin{aligned} v^T \Gamma_{x,\rho} v &= v^T \mathrm{Cov}[\nabla \log f_{x,\rho}(Y;\gamma)] v \\ &= \mathbb{E}_{x,\rho}\left[\left(\frac{v^T \nabla f_{x,\rho}(Y;\gamma)}{f_{x,\rho}(Y;\gamma)}\right)^2\right] \\ &= \int_{\mathbb{R}^L} \frac{\left(\sum\limits_{i=0}^{\infty} v^T \nabla \alpha^i_{x,\rho}(y) \frac{\gamma^i}{i!}\right)^2}{\sum\limits_{i=0}^{\infty} \alpha^i_{x,\rho}(y) \frac{\gamma^i}{i!}} f_G(y)\, dy \end{aligned}$$

where the second line follows from

$$\mathbb{E}_{x,\rho}\left[\frac{\nabla f_{x,\rho}(Y;\gamma)}{f_{x,\rho}(Y;\gamma)}\right] = 0$$

By the definition of d and (C.5), we have $v^T \nabla \alpha^n_{x,\rho}(z) = 0$ almost surely for $n < d$, thus

$$\begin{aligned} v^T \Gamma_{x,\rho} v &= \int_{\mathbb{R}^L} \frac{\left(\sum\limits_{i=d}^{\infty} v^T \nabla \alpha^i_{x,\rho}(y) \frac{\gamma^i}{i!}\right)^2}{1 + \sum\limits_{i=i}^{\infty} \alpha^i_{x,\rho}(y) \frac{\gamma^i}{i!}} f_G(y)\, dy \\ &= \frac{\gamma^{2d}}{(d!)^2} \int_{\mathbb{R}^L} \left(v^T \nabla \alpha^d_{x,\rho}(y)\right)^2 f_G(y)\, dy + O(\gamma^{2d+1}) \\ &= \frac{\gamma^{2d}}{(d!)^2} \mathbb{E}\left[\left(v^T \nabla \alpha^d_{x,\rho}(G)\right)^2\right] + O(\gamma^{2d+1}), \end{aligned}$$

Equation (C.4) now follows since $\gamma = 1/\sigma$.

We now prove (C.5) We let $(\tilde{x}, \tilde{\rho}) = (x, \rho) + hv$ in (III.13) and take the limit $h \to 0$ to get

$$\begin{aligned} \mathbb{E}\left[\left(v^T \nabla \alpha^d_{x,\rho}(G)\right)^2\right] & \\ &= \lim_{h \to 0} \frac{\mathbb{E}\left[\left(\alpha^d_{x+hz,\rho+h\theta}(G) - \alpha^d_{x,\rho}(G)\right)^2\right]}{h^2} \\ &= d! \lim_{h \to 0} \frac{\|M^d_{x+hz,\rho+h\theta} - M^d_{x,\rho}\|^2}{h^2} \\ &= d! \|\nabla_v M^d_{x,\rho}\|^2. \end{aligned}$$

□

Finally, from Corollary C.1 and Lemma C.2, we obtain a result analog to Theorem III.6.

Corollary C.3. *For any $v = (z, \theta) \in \mathbb{R}^{2L}$, such that $\mathbf{1}^T\theta = 0$ and $\theta[i] \geq 0$ whenever $\rho[i] = 0$, let $Q^n_v = \frac{1}{n!}\|\nabla_v M^d_{x,\rho}\|^2$, $q_v = \inf\{n : Q^n_v > 0\}$ and $\bar{q} = \max q_v$. Then*

$$\mathrm{MSE} \geq \sup_{v : q_v = \bar{q}} \left\{ \frac{\|z\|^2}{\lambda^{\bar{q}}_N Q^{\bar{q}}_v + O\left(\lambda^{\bar{q}}_N \sigma^{-1}\right)} \right\}. \tag{C.6}$$

D. Proof of Theorem III.1

Before proving Theorem III.1, we need the following lemma.

Lemma D.1. *The entries with index $\mathbf{k} = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d_L$ of $M^d_{x,\rho}$ and $\nabla_v M^d_{x,\rho}$ can be explicitly written as*

$$M^d_{x,\rho}[\mathbf{k}] := \sum_{\ell=0}^{L} \rho[\ell] \prod_{i=1}^{d} x[k_i - \ell], \tag{D.1}$$

and

$$(\nabla_v M^d_{x,\rho})[\mathbf{k}] = \sum_{\ell=0}^{L-1} \left(\rho[\ell] \sum_{i=1}^{d} \frac{z[k_i - \ell]}{x[k_i - \ell]} + \theta[\ell]\right) \prod_{i=1}^{d} x[k_i - \ell], \tag{D.2}$$

where we use the convention $x[k_i - \ell]/x[k_i - \ell] = 1$ when $x[k_i - \ell] = 0$. Moreover, denote the d-dimensional Fourier Transform by F_d. For any $\mathbf{a} = (a_1, a_2, \ldots, a_d) \in \mathbb{Z}^d_L$ we have

$$F_d M^d_{x,\rho}[\mathbf{a}] = F\rho\left[\sum_{j=1}^{d} a_j\right] \prod_{j=1}^{d} Fx[a_j], \tag{D.3}$$

and

$$F_d(\nabla_v M^d_{x,\rho})[\mathbf{a}] = \tag{D.4}$$

$$\left(\sum_{j=1}^{d} \frac{Fz[a_j]}{Fx[a_j]} + \frac{F\theta\left[\sum_{j=1}^{d} a_j\right]}{F\rho\left[\sum_{j=1}^{d} a_j\right]} \right) F_d M^d_{x,\rho}[\mathbf{a}], \tag{D.5}$$

again using the convention $Fx[a_j]/Fx[a_j] = 1$ *when* $Fx[a_j] = 0$. *The denote the* d*-dimensional Fourier Transform preserves the* L_2 *norm of the tensors, i.e.*

$$\|M^d_{\tilde{x},\tilde{\rho}} - M^d_{x,\rho}\|^2 = \frac{1}{L^d}\|F_d M^d_{\tilde{x},\tilde{\rho}} - F_d M^d_{x,\rho}\|^2, \tag{D.6}$$

and

$$\|\nabla_v M^d_{x,\rho}\|^2 = \frac{1}{L^d}\|F_d \nabla_v M^d_{x,\rho}\|^2. \tag{D.7}$$

Also,

Proof. We first prove equation (D.1). By equation (III.2), we have

$$\begin{aligned} M^d_{x,\rho}[\mathbf{k}] &= \mathbb{E}\left[\prod_{i=1}^{d} (R_S x)[k_i]\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{d} x[k_i - S]\right] \\ &= \sum_{\ell=0}^{L} \rho[\ell] \prod_{i=1}^{d} x[k_i - \ell]. \end{aligned}$$

Equation (D.2) follows from the formula of the derivative of the product:

$$\begin{aligned} &\left(\rho[\ell] \prod_{i=1}^{d} x[k_i - \ell]\right)' \\ &\qquad = \left(\rho'[\ell] + \rho[\ell] \sum_{i=1}^{d} \frac{x'[k_i - \ell]}{x[k_i - \ell]}\right) \prod_{i=1}^{d} x[k_i - \ell]. \end{aligned}$$

We finally prove (D.3); the proof of (D.4) is analogous.

$$\begin{aligned} F_d M^d_{x,\rho}[\mathbf{a}] &= \sum_{\mathbf{k} \in \mathbb{Z}_L^d} M^d_{x,\rho}[\mathbf{k}] \exp\left(-\frac{2\pi\iota}{L}\langle \mathbf{k}, \mathbf{a}\rangle\right) \\ &= \sum_{\mathbf{k} \in \mathbb{Z}_L^d} \sum_{\ell=0}^{L-1} \rho[\ell] \prod_{j=1}^{d} x[k_j - \ell] \exp\left(-\frac{2\pi\iota}{L} k_j a_j\right) \\ &= \sum_{\mathbf{k} \in \mathbb{Z}_L^d} \sum_{\ell=0}^{L-1} \rho[\ell] \prod_{j=1}^{d} x[k_j] \exp\left(-\frac{2\pi\iota}{L} a_j (k_j + \ell)\right) \\ &= \sum_{\mathbf{k} \in \mathbb{Z}_L^d} \sum_{\ell=0}^{L-1} \rho[\ell] \prod_{j=1}^{d} x[k_j] \exp\left(-\frac{2\pi\iota}{L} (k_j a_j + \ell a_j)\right) \\ &= \sum_{\ell=0}^{L-1} \rho[\ell] \exp\left(-\frac{2\pi\iota}{L}\left(\ell \sum_{j=1}^{d} a_j\right)\right) \prod_{j=1}^{d} Fx[a_j] \\ &= F\rho\left[\sum_{j=1}^{d} a_j\right] \prod_{j=1}^{d} Fx[a_j]. \end{aligned}$$

□

We are now ready to prove Theorem III.1, starting by (III.3). Since $\widehat{X}$ is asymptotically unbiased, $\mathbb{E}_{x,\rho}[\phi_x(\widehat{X})] \to x$ and $J_{x,\rho} \to [I_L \; 0_{L\times L}]$ as $N \to \infty$. By (III.6) and Corollary C.1 we have

$$\begin{aligned} &\lim_{N\to\infty} N \cdot \mathrm{MSE} \\ &\qquad \geq \lim_{N\to\infty} \frac{N \operatorname{tr}(\operatorname{Cov}[\phi_x(\widehat{X})])}{\|x\|^2} && \text{(D.8)} \\ &\qquad \geq \lim_{N\to\infty} \frac{\sigma^{2d}}{d!} \frac{\|J_{x,\rho} v\|^2}{\|x\|^2} \frac{1}{\|\nabla_v M^d_{x,\rho}\|^2} - O\left(\sigma^{2d-1}\right) \\ &\qquad = \frac{\sigma^{2d}}{d!} \frac{\|z\|^2}{\|x\|^2} \frac{1}{\|\nabla_v M^d_{x,\rho}\|^2} - O\left(\sigma^{2d-1}\right). && \text{(D.9)} \end{aligned}$$

We will choose $z = x - \frac{\mathbf{1}^T x}{L}\mathbf{1}$, and $\theta = \frac{1}{L}\mathbf{1} - \rho$. This choice of θ is under the theorem assumptions, since $\mathbf{1}^T\theta = 0$ and $\theta[i] = \frac{1}{L} \geq 0$ whenever $\rho[i] = 0$. By the linearity of the Fourier transform, this definition is equivalent to $Fz = Fx - Fx[0]\delta_0$ and $F\theta = F\rho[0]\delta_0 - F\rho = \delta_0 - F\rho$. Since the d-dimensional Fourier Transform is unitary, we can write using Lemma D.1

$$\|\nabla_v M^d_{x,\rho}\|^2 = \frac{1}{L^d} \sum_{\mathbf{a} \in \mathbb{Z}_L^d} |F_d \nabla_v M^d_{x,\rho}[\mathbf{a}]|^2. \tag{D.10}$$

For $d = 1, 2$ we have

$$F_1 \nabla_v M^1_{x,\rho}[a] = F\rho[a]Fz[a] + F\theta[a]Fx[a],$$

and

$$\begin{aligned} F_2 \nabla_v M^2_{x,\rho}[a_1, a_2] \; =& F\rho\left[a_1 + a_2\right] Fz[a_1]Fx[a_2] \\ &+ F\rho\left[a_1 + a_2\right] Fx[a_1]Fz[a_2] \\ &+ F\theta\left[a_1 + a_2\right] Fx[a_1]Fx[a_2]. \end{aligned} \tag{D.11}$$

Now by our choice of z and θ we have $F\rho[a]Fz[a] = -F\theta[a]Fx[a]$ for all $a \in \mathbb{Z}_L$, so $\|\nabla_v M^1_{x,\rho}\| = 0$. On the other hand, by some algebra manipulation of (D.11) we obtain

$$\begin{aligned} &\|\nabla_v M^2_{x,\rho}\|^2 \\ &\qquad = \frac{1}{L^2}\left(3\|Fz\|_4^4 + \sum_{\mathbf{a} \in \mathbb{Z}_L^2} |F\rho[a_1 + a_2]Fz[a_1]Fz[a_2]|^2\right) \\ &\qquad \leq \frac{4}{L^2}\|Fz\|^4 \\ &\qquad \leq 4\|z\|^2\|x\|^2, \end{aligned}$$

where we used $|F\rho[a_1 + a_2]| \leq 1$ and $\|Fz\|_4 \leq \|Fz\| \leq \|Fx\|$, and (III.3) follows.

We now proceed to prove (III.4). Suppose that ρ is periodic with period $\ell < \frac{L}{2}$, and let $b = \frac{L}{\ell}$, so that $b > 2$. Then $F\rho[k] = 0$ if b does not divide k. For a positive integer $i \leq \lceil \frac{b-2}{2} \rceil$, define $z_i \in \mathbb{R}^L$ such that

$$Fz_i[k] = \begin{cases} Fx[k]\iota & \text{if } b | k - i, \\ -Fx[k]\iota & \text{if } b | k + i, \\ 0 & \text{otherwise}, \end{cases}$$

where $b|k$ means that b divides k. Assume $z_i \neq 0$, let $\theta_i = 0_L$ and $v_i = (z_i, \theta_i)$. Since $\widehat{X}$ is asymptotically unbiased and

$\{z_i\}_{1\leq i\leq\lceil\frac{b-2}{2}\rceil}$ is a set of orthogonal vectors, we have by (D.8) and Corollary C.1:

$$\begin{aligned}\lim_{N\to\infty} N\cdot\text{MSE}\\ &\geq \lim_{N\to\infty}\frac{N\,\text{tr}(\text{Cov}[\phi_x(\widehat{X})])}{\|x\|^2}\\ &\geq \lim_{N\to\infty}\frac{1}{\|x\|^2}\sum_{i=1}^{\lceil\frac{b-2}{2}\rceil}\frac{Nz_i^T\,\text{Cov}[\phi_x(\widehat{X})]z_i}{\|z_i\|^2}\\ &\geq \frac{1}{\|x\|^2}\sum_{i=1}^{\lceil\frac{b-2}{2}\rceil}\frac{\sigma^{2d_i}}{d_i!}\frac{\|z_i\|^2}{\|\nabla_{v_i}M_{x,\rho}^{d_i}\|^2}-O\left(\sigma^{2d_i-1}\right),\end{aligned}$$

where $d_i = \inf\left\{n:\|\nabla_{v_i}M_{x,\rho}^{n}\|^2>0\right\}$. Recalling equation (D.10) and since $\theta_i=0$, we have now for $d=1,2,3$,

$$F_1\nabla_{v_i}M_{x,\rho}^1[a]=F\rho[a]Fz_i[a], \tag{D.12}$$

$$\begin{aligned}F_2\nabla_{v_i}M_{x,\rho}^2[a_1,a_2]=F\rho\,[a_1+a_2]\,(Fz_i[a_1]Fx[a_2]\\ +Fx[a_1]Fz_i[a_2]),\end{aligned} \tag{D.13}$$

and

$$\begin{aligned}F_3\nabla_{v_i}M_{x,\rho}^3[a_1,a_2,a_3]=&F\rho\,[a_1+a_2+a_3]\\ &(Fz_i[a_1]Fx[a_2]Fx[a_3]\\ &+Fx[a_1]Fz_i[a_2]Fx[a_3]\\ &+Fx[a_1]Fx[a_2]Fz_i[a_3]).\end{aligned} \tag{D.14}$$

Since $F\rho[a]\neq 0\Rightarrow b|a\Rightarrow Fz_i[a]=0$, $(D.12)=0$ $\forall a\in\mathbb{Z}_L$. Also $F\rho[a_1+a_2]\neq 0$ implies $b|a_1+a_2$. Let $\tilde{a}_j=\text{mod}\,(a_j,b)$ for $j=1$ and 2. Since $b|a_1+a_2$, $\tilde{a}_1+\tilde{a}_2=b$, so assume with out loss of generality that $\tilde{a}_1\leq\frac{b}{2}$. If $\tilde{a}_1\neq i$, then $Fz[a_1]=Fz[a_2]=0$. On the other hand, if $\tilde{a}_1=i$, then

$$\begin{aligned}&Fz[a_1]Fx[a_2]+Fx[a_1]Fz[a_2]\\ &\qquad=\iota Fx[a_1]Fx[a_2]-\iota Fx[a_1]Fx[a_2]\\ &\qquad=0,\end{aligned}$$

so (D.13)$=0$ $\forall\mathbf{a}\in\mathbb{Z}_L^2$. Finally since $|F\rho[\cdot]|\leq 1$ we have

$$\begin{aligned}\|\nabla_{v_i}M_{x,\rho}^3\|^2&\leq\frac{9}{L^3}\sum_{\mathbf{a}\in\mathbb{Z}_L^3}|Fz_i[a_1]Fx[a_2]Fx[a_3]|^2\\ &=9\|z_i\|^2\|x\|^4,\end{aligned}$$

and the result follows. Finally, if $z_i=0$, we can alternatively choose

$$F\tilde{z}_i[k]=\begin{cases}\iota & \text{if } b|k-i,\\ -\iota & \text{if } b|k+i,\\ 0 & \text{otherwise.}\end{cases}$$

We still have (D.12) $=0$ $\forall a\in\mathbb{Z}_L$ and (D.13) $=0$ for all $\mathbf{a}\in\mathbb{Z}_L^2$ except if $\tilde{a}_1=i$. But $z_i=0$ implies $Fx[a]=0$ if $\text{mod}\,(a,b)=\pm i$, so (D.13) $=0$ also if $\tilde{a}_1=i$.

E. Proof of Theorem IV.1

We show that if the first two moments of two pairs, signal and distribution, are equal then the pairs are identical up to a translation. Specifically, suppose that x_1 and ρ_1 have the same first two moments as x_2 and ρ_2. Equality of the first moments means that $x_1*\rho_1=x_2*\rho_2$, and therefore:

$$(Fx_1)[k]\cdot(F\rho_1)[k]=(Fx_2)[k]\cdot(F\rho_2)[k].$$

Since Fx_1 is non-vanishing, we define the ratio

$$r[k]=\frac{(Fx_2)[k]}{(Fx_1)[k]}.$$

Then,

$$(F\rho_1)[k]=(F\rho_2)[k]\cdot r[k]. \tag{E.1}$$

Furthermore, from the equality of second moments $C_{x_1}D_{\rho_1}C_{x_1}^T = C_{x_2}D_{\rho_2}C_{x_2}^T$, or equivalently (after taking Fourier transforms) $D_{Fx_1}C_{F\rho_1}D_{Fx_1}^* = D_{Fx_2}C_{F\rho_2}D_{Fx_2}^*$. Consequently, for k, $p=0,\ldots,L-1$:

$$\begin{aligned}&(Fx_1)[k]\cdot(F\rho_1)[k-p]\cdot(Fx_1)[p]^*\\ &\qquad=(Fx_2)[k]\cdot(F\rho_2)[k-p]\cdot(Fx_2)[p]^*,\end{aligned}$$

or equivalently,

$$(F\rho_1)[k-p]=(F\rho_2)[k-p]\cdot r[k]\cdot r[p]^*. \tag{E.2}$$

Because ρ_1 and ρ_2 are probability distributions, $(F\rho_1)[0]=(F\rho_2)[0]=1$. Therefore, taking $k=p$ in (E.2) implies $\left|r[k]\right|=1$. By (E.1), $r[0]=1$, and $F\rho_1$ and $F\rho_2$ have the same support.

We will denote by $\text{GCD}\,(a_1,\ldots,a_\ell)$ the greatest common divisor of the positive numbers $a_1,\ldots,a_\ell$.

Lemma E.1. *If a distribution ρ is aperiodic then*

$$\text{GCD}\left(\{k\mid 1\leq k\leq L,\ (F\rho)[k]\neq 0\}\right)=1.$$

Proof of Lemma E.1. A necessary and sufficient condition for a distribution ρ to have period ℓ is that $(F\rho)[m]\neq 0$ only for m of the form $k(L/\ell)$, $k=0,1,\ldots,\ell-1$. Therefore, the aperiodicity of a distribution ρ means that the shared greatest common divisor of all the indices of nonzero entries in $F\rho$ (which includes L, since $\rho[0]=\rho[L]=1$) is 1. In fact, if the GCD were equal to some $d>1$, then the distribution would be periodic with a period of L/d as all nonzero entries would be of the form kd, $k\in\{0,1,\ldots L/d\}$. □

Let $m_1,\ldots,m_\ell$ be the indices of the support of $F\rho_1$ (and $F\rho_2$). Because the greatest common divisor GCD is associative – that is, $\text{GCD}(a,b,c)=\text{GCD}(\text{GCD}(a,b),c)$ – by Lemma E.1 there exist integers $a_1,\ldots,a_\ell$ such that

$$\sum_{j=1}^{n}a_jm_j=1\mod L. \tag{E.3}$$

Taking $k-p=m_j$ in (E.2), we obtain:

$$r[p+m_j]=\tilde{\omega}_j\cdot r[p] \tag{E.4}$$

where

$$\tilde{\omega}_j=\frac{(F\rho_1)[m_j]}{(F\rho_2)[m_j]}.$$

From (E.3), repeated application of (E.4) yields:

$$r[p+1] = \tilde{\omega}_1^{a_1}\cdots\tilde{\omega}_\ell^{a_\ell}\cdot r[p] = \omega\cdot r[p], \tag{E.5}$$

where $\omega = \tilde{\omega}_1^{a_1}\cdots\tilde{\omega}_\ell^{a_\ell}$. Repeatedly applying (E.5), we obtain $r[m] = \omega^m r[0] = \omega^m$, or equivalently:

$$(Fx_2)[m] = \omega^m\cdot(Fx_1)[m]. \tag{E.6}$$

Furthermore, when $m = L$, we see:

$$1 = r[0] = r[L] = \omega^L\cdot r[0] = \omega^L,$$

i.e., ω is an L^{th} root of unity. Equation (E.6) then implies x_2 is a translation of x_1. Finally, (E.1) then shows that $(F\rho_1)[m] = \omega^m(F\rho_2)[m]$, so that ρ_1 is also a translation of ρ_2. This completes the proof.

F. *Proof of Lemma IV.3*

For any $0\le i\le L-1$, we can write

$$(\rho*\theta)[i] = e_i^T C_\rho\theta,$$

with e_i the unit vector with one in its ith entry. Consequently, equality of two distinct entries i and j implies

$$(e_i - e_j)^T C_\rho\theta = 0. \tag{F.1}$$

However, for a random choice of θ, if (F.1) holds with non-zero probability, then

$$(e_i - e_j)^T C_\rho = 0,$$

or,

$$C_\rho^T e_i = C_\rho^T e_j.$$

The latter implies that ρ shifted by i equals ρ shifted by j, i.e., $\rho[k-i] = \rho[k-j]$, or

$$\rho[k] = \rho[k+i-j], \quad \forall k.$$

Therefore, ρ is periodic.

G. *Proof of Theorem IV.5*

Since the residuals $\widehat{M}^1 - M^1$, $\widehat{M}^2 - M$ and $\widehat{P}_x - P_x$ are subexponential, we can apply the Bernstein-type inequality for subexponential random variables found in [45], together with Corollary IV.4, to obtain

$$\begin{aligned}\mathbb{P}&\left[\min_{s\in\mathbb{Z}_L}\|R_s\widehat{X}_{\text{Spectral}} - x\|^2\ge t\right]\\ &\le C_1\exp\left(-\frac{N}{\sigma^4}\min\left\{\frac{t}{C_2},\frac{\sqrt{t}}{C_3}\right\}\right),\end{aligned} \tag{G.1}$$

where C_1, C_2 and C_3 are finite, positive constants that depend on x and ρ. We have

$$\begin{aligned}\text{MSE}\cdot\|x\|^2 &= \mathbb{E}\left[\min_{s\in\mathbb{Z}_L}\|R_s\widehat{X}_{\text{Spectral}} - x\|^2\right]\\ &= \int_0^\infty\mathbb{P}\left[\min_{s\in\mathbb{Z}_L}\|R_s\widehat{X}_{\text{Spectral}} - x\|^2\ge t\right]dt\\ &\le C_1\int_0^\infty\exp\left(-\frac{N}{\sigma^4}\min\left\{\frac{t}{C_2},\frac{\sqrt{t}}{C_3}\right\}\right)dt\\ &= C_4\frac{\sigma^4}{N}\left[C_5 + \left(C_5 + 2\frac{\sigma^4}{N}\right)\exp\left(-C_5\frac{N}{\sigma^4}\right)\right],\end{aligned} \tag{G.2}$$

with $C_4 = C_1C_3^2$ and $C_5 = C_2/C_3^2$, thus if $N = \omega(\sigma^4)$, (G.2) converges to 0 as n diverges, and $\widehat{X}_{\text{Spectral}}$ converges to the true signal in L^2, up to a cyclic shift.

H. *Proof of Proposition IV.6*

It is clear that, as $L>1$, $x_1\neq x_2$. In addition, since x_1 is real, the construction ensures that x_2 is real as well.

The ℓ periodicity of ρ means a sparsity pattern for $F\rho$. Particularly, $F\rho$ is zero everywhere besides

$$(F\rho)\left[kL/\ell\right]\neq 0 \iff kL/\ell \text{ is integer}, \tag{H.1}$$

for $k = 0,\ldots,\ell-1$. It is easy to verify that

$$(Fx_1)[k](F\rho)[k] = (Fx_2)[k](F\rho)[k], \quad k = 0,\ldots,L-1.$$

Therefore, x_1 and x_2 share the same first moment.

For the second moments, we will show the equality

$$C_{x_1}D_\rho C_{x_1}^T = C_{x_2}D_\rho C_{x_2}^T.$$

Applying the Fourier matrix, due to the realness of ρ, the latter is equivalent to

$$D_{Fx_1}C_{F\rho}D_{Fx_1} = D_{Fx_2}C_{F\rho}D_{Fx_2},$$

Similar to (E.2) and by the sparsity pattern of (H.1), this equality should hold only if

$$(Fx_1)[i]\,(Fx_1)\left[i+tL/\ell\right]^* = (Fx_2)[i]\,(Fx_2)\left[i+tL/\ell\right]^*,$$

for all $t = 0,\ldots,\ell$ and $i = 0,\ldots,L-1$. By the construction (IV.8), this equation holds true.

I. *Proof of Claim IV.7*

Throughout the proof, we assume that each period has no repeated values. This property is guaranteed by reshuffling the measurements with random $\theta\in\Delta_L$; see Lemma IV.3. Additionally, we assume without loss of generality that $|Fx|[k] = 1$ for all k.

Observe that both x and $R_{L/2}x$ are eigenvectors of $\widehat{M}^2 = C_xD_\rho C_x^T$ (we assume exact knowledge of the moments) with the same eigenvalue. Also, x and $R_{L/2}x$ are orthogonal as columns in the orthogonal matrix C_x. Then, if u is an eigenvector, we can write for some scalars $\alpha,\beta\in\mathbb{R}^L$:

$$u = \alpha x + \beta R_{L/2}x,$$

and therefore,

$$R_{L/2}u = \alpha R_{L/2}x + \beta x,$$

as $R_{L/2} = R_{L/2}^{-1}$. Then, one can verify that the inner product of u and $R_{L/2}u$ is $2\alpha\beta\|x\|^2$. Since the signals are orthogonal, their inner product is zero. This means that α or β must be zero. This in turn implies that u was either x or $R_{L/2}x$ in the first place. Therefore, x is the unique eigenvector of $\widehat{M}^2$ that is orthogonal to its translation by $L/2$. This completes the proof.

J. Proof of Lemma VI.1

It is easy to check that the condition $q[\ell] > 0$ is automatically enforced whenever $w[\ell] > 0$ (otherwise the objective is $-\infty$). So the simplex constraint is equivalent to $\sum_{\ell=0}^{L-1} q[\ell] = 1$. The Lagrangian for this problem is the function:

$$(q, \nu) = \sum_{\ell=0}^{L-1} w[\ell]\log(q[\ell]) + \nu\left(1 - \sum_{\ell=0}^{L-1} q[\ell]\right),$$

and the KKT conditions imply $q^*[\ell] = \frac{w[\ell]}{\nu^*}$. Since q is on the simplex, we conclude that $\nu^* = \sum_{\ell'=0}^{L-1} w[\ell']$.

K. Convex relaxation with semidefinite program

In this section, we propose an additional algorithm for non-uniform MRA based on a semidefinite program (SDP) relaxation.

Since the power spectrum of the signal can be estimated from the data at sample complexity scaling as $\omega(1/\operatorname{SNR}^2)$ according to (IV.6), we assume in this section, without loss of generality, that $|Fx|[k] = 1$ for all k. Note, that as in Algorithm 2, the normalization is done on the second moment matrix, not the individual observations, in order to retain the noise statistics.

The SDP relaxation is based on considering the second moment matrix in the Fourier domain, namely,

$$M_*^2 = F\left(M^2\right)F^{-1} = D_{Fx}C_{F\rho}^T D_{Fx}^*. \tag{K.1}$$

The last expression can be also written as

$$M_*^2 = C_{F\rho}^T \odot (FxFx^*),$$

or

$$M_*^2 \odot \overline{X} = C_{F\grave{\rho}}, \tag{K.2}$$

where $X = (Fx)(Fx)^*$. and $\grave{\rho} := F^{-1}(\overline{F\rho})$.

The formulation of (K.2) suggests to pose the recovery problem as,

$$\begin{aligned} \min_{\tilde{\rho},\tilde{X}} \quad & \left\|\widehat{M}_*^2 \odot \overline{\tilde{X}} - C_{F\tilde{\rho}}\right\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \operatorname{diag}(\tilde{X}) = 1, \quad \operatorname{rank}(\tilde{X}) = 1, \\ & \tilde{X}[1,0] = 1, \quad \tilde{X} \succeq 0, \quad \tilde{\rho}[0] = 1, \\ & \tilde{\rho}[k] = \overline{\tilde{\rho}[-k]}, \forall k. \end{aligned} \tag{K.3}$$

The constraint $\tilde{X}[1,0] = 1$ follows the assumption that $(Fx)[0] = (Fx)[1] = 1$. While we can easily estimate $(Fx)[0]$ and therefore fix it, the assumption of fixed $(Fx)[1] = 1$ is more delicate. Recall that the solution for the MRA problem is always up to cyclic translation. In the Fourier domain, it means that the first entry of the Fourier transform of the signal is determined up to an arbitrary modulation by $e^{2\pi i\ell/L}$ for some $\ell \in \mathbb{Z}$. If $L \to \infty$, this allows us to fix this coefficient arbitrarily.

Similarly to the well-known SDP relaxation of the Max-Cut problem [46], the non-convex problem (K.3) can be relaxed to a convex program by omitting the rank constraint as follows,

$$\begin{aligned} \min_{\tilde{\rho},\tilde{X}} \quad & \left\|\widehat{M}_*^2 \odot \overline{\tilde{X}} - C_{F\tilde{\rho}}\right\|_{\mathrm{F}}^2 \\ \text{subject to} \quad & \operatorname{diag}(\tilde{X}) = 1, \quad \tilde{X}[1,0] = 1, \\ & \tilde{X} \succeq 0, \quad \tilde{\rho}[0] = 1, \quad \tilde{\rho}[k] = \overline{\tilde{\rho}[-k]}, \forall k. \end{aligned} \tag{K.4}$$

This relaxation is convex and can be solved in polynomial time using off–the–shelf software, such as CVX [47].

The SDP relaxation (K.4) recovers the Fourier phases of the signal and the distribution exactly for $N \to \infty$ and fixed noise level, since in this regime we can estimate the first two moments arbitrarily well.

Theorem K.1. *Assume that* $|Fx|[k] = 1$ *for all* k *and that* $F\rho$ *is non-vanishing. In addition, assume that* $(Fx)[0] = (Fx)[1] = 1$. *Then, if* $N \to \infty$ *and* σ *is fixed, the solution of* (K.4) *is given by* $\tilde{X} = (Fx)(Fx)^*$ *and* $\tilde{\rho} = F\grave{\rho}$.

Proof. Since σ is fixed and $N \to \infty$, one can estimate M_*^2 as in (K.1) exactly. Then, since (K.4) admits at least one solution (the underlying signal and distribution), the objective is zero at the solution and we get the relation:

$$C_{\tilde{\rho}} = M_*^2 \odot \overline{\tilde{X}} = C_{F\grave{\rho}} \odot (FxFx^*) \odot \overline{\tilde{X}}, \tag{K.5}$$

where we use $\grave{\rho} := F^{-1}(\overline{F\tilde{\rho}})$. Let $u = \tilde{\rho}/F\grave{\rho}$. Since $\tilde{X} \succeq 0$ we conclude that $C_u \succeq 0$ and hence $Fu \geq 0$ (the Fourier transform of u is non-negative). By the constraints of (K.4), we also have $u[0] = 1$. By examining the $(1,0)$th entry of (K.5), we also conclude that

$$(Fx)[1]\overline{(Fx)[0]}(F\grave{\rho})[1]\overline{\tilde{X}[1,0]} = \tilde{\rho}[1] \Rightarrow u[1] = \overline{\tilde{X}[1,0]} = 1,$$

where the last equality holds because of the constraints of (K.4).

Until now, we have shown that the vector u satisfies $u[0] = u[1] = 1$, it is conjugate-symmetric and its Fourier transform is non-negative. Therefore, by Lemma IV.2 of [15], we conclude that $u[n] = 1$ for all n, or $\tilde{\rho} = F\grave{\rho}$. Next, we substitute $\tilde{\rho} = F\grave{\rho}$ in (K.5) and get

$$1 = (FxFx^*) \odot \overline{\tilde{X}},$$

where the equality holds entry-wise. Since all entries of $\hat{x}$ are normalized, we conclude that $\tilde{X} = (Fx)(Fx)^*$. This concludes the proof. □