# Optimal rates of estimation for multi-reference alignment

Afonso S. Bandeira*, Philippe Rigollet†, and Jonathan Weed‡

*Courant Institute of Mathematical Sciences, New York University*
*Massachusetts Institute of Technology*
*Massachusetts Institute of Technology*

*Abstract.* This paper describes optimal rates of adaptive estimation of a vector in the multi-reference alignment model, a problem with important applications in fields such as signal processing, image processing, and computer vision, among others. We describe how this model can be viewed as a multivariate Gaussian mixture model under the constraint that the centers belong to the orbit of a group. This enables us to derive matching upper and lower bounds that feature an interesting dependence on the signal-to-noise ratio of the model. Both upper and lower bounds are articulated around a tight local control of Kullback-Leibler divergences that showcases the central role of moment tensors in this problem.

*AMS 2000 subject classifications:* Primary Statistics; secondary Invariant Theory, Signal Processing.
*Key words and phrases:* Multi-reference alignment, Orbit retrieval, Mixtures of Gaussians.

## 1. INTRODUCTION

The multi-reference alignment problem and its variants arise in various scientific and engineering applications such as structural biology [SVN+05, TS12, Sad89], image recognition [Bro92], and signal processing [ZvdHGG03]. A striking feature of this class of problems is that each observation is not only observed in a noisy setting but is also altered by an latent transformation that reflects underlying heterogeneity of the data. The precise nature of this transformation depends on the specific application, but it can often be characterized as the action of the unknown element of a known group.

### 1.1 The multi-reference alignment problem

Consider the multi-reference alignment problem on $\mathbb{R}^d$ under cyclic shifts [BCSZ14]. The goal in this problem is to produce an estimate of a vector $\theta \in \mathbb{R}^d$ (of-

*Part of this work was done while A. S. Bandeira was with the Mathematics Department at MIT and supported by NSF Grant DMS-1317308.
†This work was supported in part by NSF CAREER DMS-1541099, NSF DMS-1541100, DARPA W911NF-16-1-0551, ONR N00014-17-1-2147 and a grant from the MIT NEC Corporation.
‡This work was supported in part by NSF Graduate Research Fellowship DGE-1122374.

ten thought of as a signal) from noisy cyclic shifted observations. More concretely, one observes $n$ independent random vectors $Y_1, \dots, Y_n \in \mathbb{R}^d$ given by $Y_i = R_{\ell_i}\theta + \sigma\xi_i$, where $\ell_i$ is an unknown parameter (shift) in $[d] := \{1, \dots, d\}$; $R_{\ell_i}$ is a latent cyclic shift by $\ell_i$ coordinates: the $j$th coordinate of $R_{\ell_i}\theta \in \mathbb{R}^d$ is given by $\big(R_{\ell_i}\theta\big)_j = \theta_{j+\ell_i (\mathrm{mod}\ d)}$; and $\xi_i \sim \mathcal{N}(0, I_d)$, i.i.d.

Note that $R_{\ell_i}$ is a linear operator from $\mathbb{R}^d$ to $\mathbb{R}^d$ that can be represented by a matrix. For example when $d = 3$, we have

$$R_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad R_2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \text{and } R_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

However, since cyclic shifts in $\mathbb{R}^d$ form a group isomorphic to quotient ring $\mathbb{Z}_d := \mathbb{Z}/d\mathbb{Z}$ of integers modulo $d$, we write $R_{\ell_i} \in \mathbb{Z}_d$ for simplicity.

The action of $\mathbb{Z}_d$ on $\mathbb{R}^d$ also has a simple representation in the Fourier domain, where the group acts on the *phases* of the Fourier coefficients. Denoting by $\hat{\theta}$ the discrete Fourier transform of the vector $\theta$, we obtain that the rotated vector $R_\ell\theta$ satisfies

$$\widehat{R_\ell\theta}_j = e^{2\pi j\ell i/d}\hat{\theta}_j \quad \text{for } -\lfloor d/2\rfloor \le j \le \lceil d/2\rceil. \tag{1.1}$$

In practical applications, the signal $\theta$ arises as the discretization of some underlying continuous signal. Therefore, instead of focusing on the group $\mathbb{Z}_d$, we consider instead the action of the *circle group* $U(1)$ of unit-norm complex numbers, where given $z \in U(1)$ we define the operator $R_z$ on $\mathbb{R}^d$ by its action on the Fourier transform $\hat{\theta}$:

$$\widehat{R_z\theta}_j = z^j\hat{\theta}_j \quad \text{for } -\lfloor d/2\rfloor \le j \le \lfloor d/2\rfloor. \tag{1.2}$$

We define the group of such operators by $\mathcal{F}$ and call $R_z$ a *fractional cyclic shift*. We assume throughout that $d$ is odd, since for real signals not all fractional cyclic shifts are well defined when $j = d/2$. This group is both slightly easier to analyze and better corresponds to the situation in practice. Moreover, a comparison of Equations 1.1 and 1.2 shows when $d$ is large the groups $\mathbb{Z}_d$ and $\mathcal{F}$ are essentially equivalent. For the sake of exposition, in the sequel, we will focus on $\mathcal{F}$ and omit the adjective "fractional" when referring to shifts.

Multi-reference alignment is directly used in structural biology [Dia92, TS12]; radar [ZvdHGG03]; crystalline simulations [SSK13]; and image registration in a number of important contexts, such as in geology, medicine, and paleontology [SSK13, FZB02]. As a results, variants of this problem for groups other than $\mathcal{F}$ have received some attention, but rarely in statistics. We note parallel research efforts that have investigated a Boolean version of this problem [APS17].

Another important related problem is that of molecule reconstruction in Cryo-Electron Microscopy (Cryo-EM). Cryo-EM is an important technique used to determine three-dimensional structures of biological macromolecules. (It was considered the Nature method of the year in 2015 [Edi15, Nog15]). As in the multi-reference alignment problem described above, one of the main difficulties with this imaging technique is that these molecules are imaged at different unknown orientations and each molecule can only be imaged once due to the destructive nature of the process. More precisely, each measurement consists of a tomographic projection of a rotated (by an unknown rotation in $SO(3)$) copy of

the molecule. The task is then to reconstruct the molecule density from many such measurements. This reconstruction problem has also received significant attention, primarily from computational perspectives, but its statistical properties remain largely unexplored.

We present in this paper the first statistical analysis for Euclidean MRA and describe natural classes of signals whose complexity spans the spectrum of statistical rates achievable in this context. Although we focus on $\mathcal{F}$ or its subgroups rather than $SO(3)$, the two frameworks share many features. In particular, our results exhibit non-trivial minimax rates that are inherent to this class of problems.

### 1.2 The Synchronization Approach

The difficulty of the multi-reference alignment problem resides in the fact that both the signal $\theta \in \mathbb{R}^d$ and the shifts $z_1, \dots, z_n \in \mathcal{F}$ are unknown and the latter are therefore latent variables. If the shifts were known, one could easily estimate $\theta$ by taking the average of $R_{z_i}^{-1} Y_i, i = 1, \dots, n$. In fact, this simple observation is the basis of the leading current approach called the "synchronization approach" [BCSZ14, BCS15] to this problem. Specifically, synchronization aims at recovering the latent variables $R_{z_i}$ by solving a problem of the form

$$\min_{z_1', \dots, z_n' \in U(1)} \sum_{1 \le i,j \le n} \left\| R_{z_i'}^{-1} Y_i - R_{z_j'}^{-1} Y_j \right\|^2 . \tag{1.3}$$

Denoting by $\hat{R}_{z_i}$ the solutions of (1.3), one can then estimate $\theta$ by the average of $\hat{R}_{z_i}^{-1} Y_i, i = 1, \dots, n$.

Synchronization problems can be formulated as estimation problems on a graph. More precisely, one can associate each observation $Y_i$ to a graph node, each of which has a hidden label $g_i \in \mathcal{G}$ for some group $\mathcal{G}$ of transformations. (In our case, $\mathcal{G} \cong \mathcal{F}$.) The pairwise data, which we identify with edges of the graph, reveals information about ratios $g_i (g_j)^{-1}$. In the the context of (1.3), this information is simply $\| R_{\ell_i'}^{-1} Y_i - R_{\ell_j'}^{-1} Y_j \|^2$. Despite synchronization problems being computationally hard in general, certain theoretical guarantees have been derived under specific noise models that are unfortunately not realistic for the problems of interest in this paper. For example, it is often assumed that the edge observations are independent instead of the more relevant independence of vertices. Among the most prominent methods are spectral methods [Sin11, BSS11], semidefinite relaxations [BCSZ14, BCS15, ABBS14, BBS16, JMRT16, BBV16], and methods based on Approximate Message Passing [PWBM16] and other modified power methods [Bou16, CC16]. Synchronization also enjoyed many interesting connections with geometry (see [GBM16] for an example).

When the noise is smaller entrywise than the signal $\theta$, synchronization approach yields acceptable results because macroscopic features underlying signal are still visible. However, in applications, the noise level is often significantly larger than signal, and it is on this regime that we focus our attention. An illustration of the difference between these regimes appears in Figure 1.

Unfortunately, for the high noise regimes we are interested here the shifts are impossible to reliably estimate, regardless of the number of samples [ADBS16, WW84]. In fact, it is not difficult to see that even if we had access to $\theta$, one would still not be able to reliably estimate the shifts in a high noise regime.

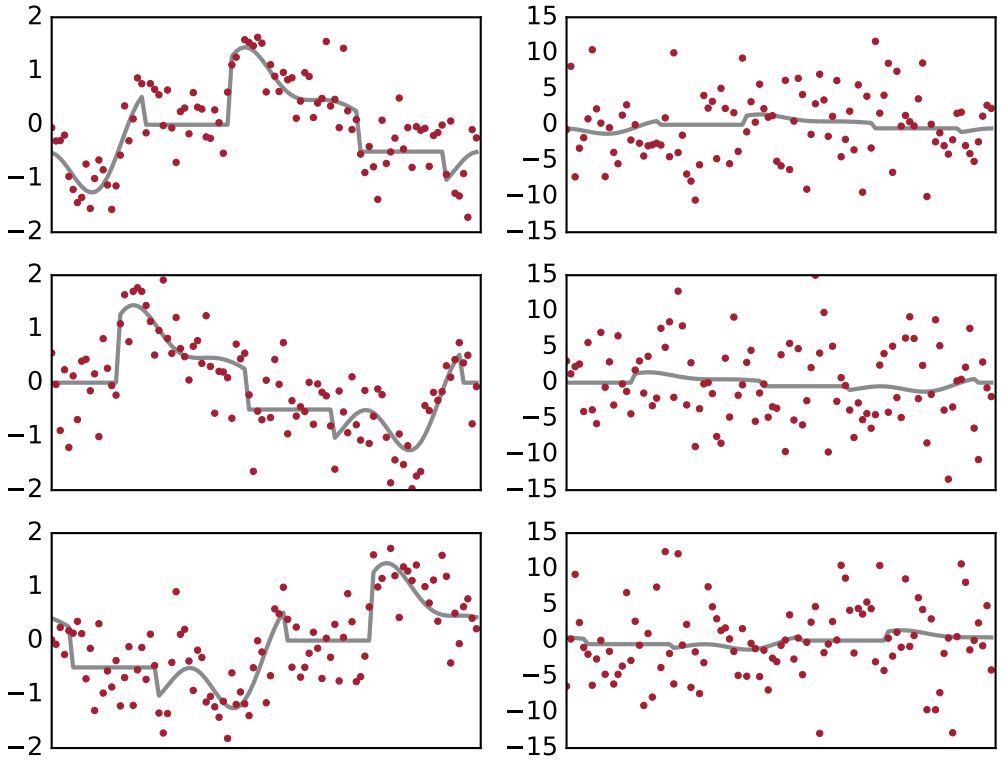


Figure 1: Instances of the multi-reference alignment problem, at low (left column) and high (right column) noise levels. The true underlying signal appears in gray, and the noised version appears in red. When the noise level is low, large features of the signal are still visible despite the noise; in the presence of large noise, however, the signals cannot reliably be synchronized.

### 1.3 Notation

Denote by $W \in \mathcal{C}^{d\times d}$ the discrete Fourier transform matrix, with entries given by

$$W_{jk} = \frac{1}{\sqrt{d}} \exp(2\pi ijk/d)\,, \qquad 1 \le j,k \le d\,.$$

The normalization factor $1/\sqrt{d}$ is chosen so that the resulting matrix $W$ is unitary.

Given $\theta \in \mathbb{R}^d$, let $\hat{\theta} = W\theta$ be its Fourier transform. We will index $\hat{\theta}$ from $-\lfloor d/2 \rfloor$ to $\lfloor d/2 \rfloor$.

The symbol $\|\cdot\|$ denotes the $\ell_2$ norm on $\mathbb{R}^d$. For any positive integer $d$, we write $[d] = \{1, \dots, d\}$.

Given a vector $t$, let $t^{\otimes k}$ denote the order-$k$ tensor formed by taking the $k$-fold tensor product of $t$ with itself. Denote by $\|A\|$ the Hilbert-Schmidt norm of a tensor $A$, defined by $\|A\|^2 = \langle A, A\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the entrywise inner product.

A tensor $A$ is *symmetric* if $A_{i_1\dots i_k} = A_{i_{\pi(1)}\dots i_{\pi(k)}}$ for any permutation $\pi$ of $[k]$. For such tensors, the value $A_{i_1\dots i_k}$ depends only on the multiset $\{i_1, \dots, i_k\}$.

We define a *cyclic shift* of $\theta \in \mathbb{R}^d$ by a unit-norm complex number $z$ by

specifying its action on the Fourier transform of $\theta$:

$$\widehat{R_z(\theta)}_k = z^k \hat{\theta}_k \text{ for } k = -\lfloor d/2 \rfloor, \dots, \lfloor d/2 \rfloor.$$

The group of cyclic shifts is known as the *circle group* and we denote it by $\mathcal{F}$. Note that it is isomorphic to several well known groups such as $SO(2)$ and $\mathbb{R}/\mathbb{Z}$.

Recall that the Kullback-Leibler (KL) divergence between two distributions $P$ and $Q$ such that $P \ll Q$ is given by

$$D(P \,\|\, Q) = \int \log\left(\frac{dP}{dQ}\right) dP\,.$$

It is well known that $D(P \,\|\, Q) \geq 0$, with equality holding iff $P = Q$ almost surely.

### 1.4 Organization of the paper

In Section 2, we propose a new approach to MRA using Gaussian mixture models. Then, in Section 3 we present our main result, Theorem 1, providing minimax rates for the multi-reference alignment problem under cyclic shifts. In Section 4 we draw precise connections between the Kullback-Leibler (KL) divergence of sampled distributions from two different signals and the distance between their invariant moment tensors. This relation does not depend on the precise nature of the cyclic shifts and holds for any compact subgroup of the orthogonal group. In Section 5 we give guarantees for the maximum likelihood estimator (MLE): Theorem 4 gives a general guarantee for the MLE under conditions on the KL divergence that were obtained in Section 4 and in Section 5.2 a modified MLE is developed for the specific problems of shifts. This modified MLE provides the upper bounds in Theorem 1. Section 6 concludes by establishing the lower bound in Theorem 1; the proof involves finding pairs of different signals with several matching invariant moment tensors.

## 2. GAUSSIAN MIXTURE MODEL

We propose an alternative to the synchronization approach discussed above that completely bypasses the estimation of the shifts $\ell_1, \dots, \ell_n$ in favor of estimating $\theta$ directly. To do so, we recast our model as a continuous mixture of Gaussians whose centers are algebraically constrained. Since the Gaussian distribution is invariant under cyclic shift, we can without loss of generality assume that the shifts $R_{\ell_i}$ are independent and uniformly distributed over $\mathcal{F}$. Indeed, to achieve this setup, we transform the observations $Y_i$ into $R_{U_i} Y_i$, where $U_i$ is uniformly distributed over $U(1)$ and independent of all other random variables. Since $U_i + \ell_i$ is also uniformly distributed over $U(1)$, these new observations are drawn from a mixture of Gaussians with uniform mixing weights and centers given by $R_\ell \theta$, $\ell \in U(1)$. In particular, these centers are linked together by a rigid algebraic structure: they are the orbit of $\theta \in \mathbb{R}^d$ under the action of $\mathcal{F}$. We summarize this observation into our basic model.

Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be $n$ independent copies of $Y \in \mathbb{R}^d$ where

$$Y = R\theta + \sigma\xi, \tag{2.1}$$

where $\theta \in \mathbb{R}^d$ is the unknown parameter of interest, $R$ is drawn uniformly (i.e., according to the Haar probability measure) from $\mathcal{F}$ and $\xi \sim \mathcal{N}(0, I_d)$ is independent Gaussian noise. Denote by $\mathrm{P}_\theta$ the distribution of a random variable $Y$ that satisfies (2.1).

Define the maximum likelihood estimator (MLE) $\tilde{\theta}_n$ by

$$\tilde{\theta}_n = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \log \mathbb{E}_R\big[\exp\big(-\frac{1}{2\sigma^2}\|Y_i - R\theta\|^2\big)\big]\,. \tag{2.2}$$

We postpone the discussion of computational efficiency to a companion paper [PWB+17] and focus here on the statistical properties of this estimator. In particular, we show that a small modification of it optimally solves the multi-reference alignment problem in a certain sense.

Although the goal of this problem is to recover the unknown parameter $\theta$, it is not hard to see that we can only hope to identify $\theta$ up to a global cyclic shift. We therefore define a (pseudo-)metric on $\mathbb{R}^d$ that deems two vectors $\theta, \phi \in \mathbb{R}^d$ to be close if they are close up to some element of $\mathcal{F}$. To this end, define

$$\rho(\theta, \phi) = \min_{R \in \mathcal{F}} \|\theta - R\phi\|\,.$$

We assume throughout that the noise variance $\sigma^2$ is known. This assumption is realistic in many applications such as imaging or signal processing. In other circumstances, it may be calibrated using cross-validation, for example. Throughout this paper, we assume that $c \le \|\theta\| \le 1$, where $c > 0$ is a universal constant so that $\sigma$ captures entirely the (inverse) signal-to-noise ratio. We note that, comparatively, $\mathbb{E}\|\sigma\xi_i\|^2 = \sigma^2 d$.

Gaussian mixture models have been extensively studied in the statistical literature since their introduction by Pearson [Pea94] in the nineteenth century (see e.g. [MP00] for an overview). As illustrated by the extant literature, mixture models are quite rich and broadly applicable to a variety of statistical problems ranging from clustering to density estimation. As a result, various statistical perspectives may apply when studying the performance of estimators. In the context of multi-reference alignment, we are naturally interested in estimating $\theta$ or, equivalently, the centers. It has been established that in general, the rate of estimation of the centers may scale like $n^{-C/d}$ for $d$ centers that can be arbitrarily close (see for example [MV10] and more recently [HK15] for an interesting explanation from the point of view of model misspecification). This curse of dimensionality arises from the minimax point of view where centers may be arbitrarily close, with a distance that may depend on the number of observations, thus leading to nonparametric rates of estimation. Instead, when the centers are well separated, the general conditions of [Wal49] are satisfied so that the MLE converges at the parametric rate $1/\sqrt{n}$ . In our case, when the mixture is continuous, Assumption 1, below, plays an analogous role in ensuring convergence at the parametric rate. When such conditions are met, the Fisher information determines how the statistical performance of the MLE scales with the noise level $\sigma$. This question is central to signal processing problems such as multi-reference alignment, where $\sigma$ is quite large, since it determines the order of magnitude of the sample size $n$ required to achieve a certain accuracy.

Exact computation of the Fisher information matrix in this model is out of reach and instead, we focus on the scaling of the quantity $\sqrt{n}\rho(\tilde{\theta}_n, \theta)$ with the signal-to-noise ratio of the problem, where $\tilde{\theta}_n$ is the MLE (2.2).

## 3. MAIN RESULTS

As mentioned above, $\sqrt{n}\rho(\tilde{\theta}, \theta)$ depends asymptotically on the Fisher information of the model, which can be related to the curvature of the Kullback-Leibler divergence around its minimum. Conversely, (lack of) curvature of the Kullback-Leibler divergence around its minimum is what controls minimax lower bounds. To address both problems at once, we follow an idea originally introduced in [LNS99] in the context of functional estimation and further developed by [CL11, WY16]. In the multi-reference alignment model, this approach allows us to relate Kullback-Leibler divergence to moment tensors, which can in turn be controlled using Fourier-theoretic arguments. It will follow from this analysis that the difficulty of estimating a particular signal $\theta$ depends on the support of the Fourier transform $\hat{\theta}$ of $\theta$. In particular, define the *positive support* $\mathrm{psupp}(\hat{\theta})$ of $\hat{\theta}$ by

$$\mathrm{psupp}(\hat{\theta}) = \{j \mid j \in \{1, \dots, d/2\}, \hat{\theta}_j \neq 0\}\,.$$

We make the following assumption, which guarantees that the MLE converges at a parametric rate.

ASSUMPTION 1. *There exists an absolute constant $c > 1$, not depending on $n$, such that $c^{-1} \leq |\hat{\theta}_j| \leq c$ for all $j \in \mathrm{psupp}(\hat{\theta})$. We denote by $\mathcal{T}$ the set of such vectors.*

We emphasize that this is the situation of most interest to practitioners: the existence of very small, but non-zero, coordinates whose values approaches 0 with $n$ should rightly be considered pathological. In a way, Assumption 1 rules out certain artificial difficult situations analogous to classical difficulties arising in estimating mixtures of Gaussians, such as distinguishing the mixture $.5\mathcal{N}(+\varepsilon, 1) + .5\mathcal{N}(-\varepsilon, 1)$ from the single Gaussian $\mathcal{N}(0, 1)$ for very small $\varepsilon$. Moreover, it enables us to restrict our attention to a compact set of parameters and bypass some technical complications.

The following theorem reveals a surprising phenomenon: even under Assumption 1, the multi-reference alignment problem suffers from the curse of dimensionality.

THEOREM 1. *Let $2 \leq s \leq \lfloor d/2 \rfloor$. Let $\mathcal{T}_s$ be the set of vectors $\theta \in \mathcal{T}$ satisfying Assumption 1 and $\mathrm{psupp}(\hat{\theta}) \subset [s]$. Then,*

$$\inf_{T_n} \sup_{\theta \in \mathcal{T}_s} \mathbb{E}_\theta[\rho(T_n, \theta)] \asymp \frac{\sigma^{2s-1}}{\sqrt{n}}(1 + o_n(1))\,, \tag{3.1}$$

*where the infimum is taken over all estimators $T_n$ of $\theta$ and where the symbol $\asymp$ hides constants depending on $d$ but on no other parameter. A modified MLE $\check{\theta}_n$ defined in (5.11) achieves this rate.*

A few remarks are in order. Note first that the curse of dimensionality is inherent to the minimax paradigm. Indeed, our proof of the lower bound describes a class of signals that satisfy Assumption 1 but have a very specific Fourier spectrum. Such signals drive the worst case bound of order $\sigma^{2s-1}/\sqrt{n}$. This limitation is overcome in a companion paper [PWB+17], where we show that even Fourier dense signals $\theta$ may be estimated as the same rate as signals in $\mathcal{T}_2$ as long as they

are generic enough. Second, our proof techniques do not allow us to remove the $\sigma$ dependence of the term $o_n(1)$. In particular, for small values of $n$, this term may actually dominate the upper bound. We conjecture that this issue is an artifact of the proof technique and note that preliminary numerical results in [PWB+17] support this claim.

The rest of this paper is devoted to the proof of the main results in Theorem 1.

## 4. INFORMATION GEOMETRY

In this section, we develop several new tools to obtain precise bounds on the divergence $D(\mathrm{P}_\theta \,\|\, \mathrm{P}_\phi)$ for pairs of signals $\theta$ and $\phi$. On the one hand, it is well known that upper bounds on the divergence translate into minimax lower bounds using LeCam's method [LeC73]. On the other hand, we also show how to transform lower bounds on $D(\mathrm{P}_\theta \,\|\, \mathrm{P}_\phi)$ into *uniform* upper bounds on the performance of the MLE. Note that this analysis departs from the classical *pointwise* rate of convergence for MLE that guarantees a rate of convergence $n^{-1/2}$ *for each fixed choice of parameter* as $n \to \infty$. Our tools strengthen this result considerably. Indeed, we show that for reasonable choices of $\theta$, the MLE achieves a rate of $n^{1/2}$ *uniformly* over all choices of $\theta$. We refer the reader to [HK15] for examples of other Gaussian mixture problems where the pointwise and uniform rates of estimation differ.

For convenience, we abbreviate $D(\mathrm{P}_\theta \,\|\, \mathrm{P}_\phi)$ by $D(\theta \,\|\, \phi)$. The following Lemma collects several useful facts about the function $D(\theta \,\|\, \phi)$. A proof appears in Appendix A.

LEMMA 2. *Let $R$ have uniform distribution over any subgroup $\mathcal{G}$ of the orthogonal group in $d$ dimensions. Fix $\theta, \phi \in \mathbb{R}^d$. The following holds*
*(i) If $\vartheta = \theta - \mathbb{E}R\theta$ and $\varphi = \phi - \mathbb{E}R\phi$, then*

$$D(\theta \,\|\, \phi) = D(\vartheta \,\|\, \varphi) + \frac{1}{2\sigma^2}\|\mathbb{E}[R\theta] - \mathbb{E}[R\phi]\|^2 .$$

*(ii) If $\rho(\theta, \phi) = \varepsilon$ and $\|\theta\|, \|\phi\| \le 1$, then*

$$D(\theta \,\|\, \phi) = \frac{1}{2\sigma^2}\|\mathbb{E}[R\theta] - \mathbb{E}[R\phi]\|^2 + \frac{1}{4\sigma^4}\|\mathbb{E}[(R\theta)^{\otimes 2}] - \mathbb{E}[(R\phi)^{\otimes 2}]\|^2 + \frac{O(\varepsilon^2)}{\sigma^6} ,$$

*where $O(\varepsilon^2)$ hides a constant, which may depend on $d$ but is otherwise independent of $\theta$, $\phi$, and $\sigma$.*

The following Theorem provides a tight bound on the quantity $D(\theta \,\|\, \phi)$ in terms of Hilbert-Schmidt distance between the moment tensors $\mathbb{E}[(R\theta)^{\otimes k}]$ and $\mathbb{E}[(R\phi)^{\otimes k}]$. Specifically, we show that the divergence $D(\theta \,\|\, \phi)$ is of order $\sigma^{-2k}\varepsilon^2$ where $k$ is the smallest natural number such that the moment tensors $\mathbb{E}[(R\theta)^{\otimes k}]$ and $\mathbb{E}[(R\phi)^{\otimes k}]$ differ significantly. This theorem is not specific to cyclic shifts and holds for any compact subgroup of the orthogonal group.

THEOREM 3. *Let $\theta$ be a fixed vector in $\mathbb{R}^d$ such that $\|\theta\| \le 1$. Let $\phi \in \mathbb{R}^d$ be such that $\rho(\theta, \phi) = \varepsilon \le \|\theta\|$. Let $R$ be a random element drawn according to the Haar probability measure on any compact subgroup $\mathcal{G}$ of the orthogonal group*

*in $d$ dimensions. For all $m \geq 1$, let $\Delta_m = \mathbb{E}[(R\theta)^{\otimes m} - (R\phi)^{\otimes m}]$. If there exists $k \geq 1$ such that, as $\varepsilon \to 0$,*

$$\|\Delta_m\| = o(\varepsilon) \text{ for } m = 1, \dots, k-1, \quad \text{and} \quad \|\Delta_k\| = \Omega(\varepsilon)\,,$$

*then $\|\Delta_k\| = \Theta(\varepsilon)$. Moreover, for $\sigma \geq 1$ there exist universal constants $c$ and $\bar{C}$ and constant $\underline{C}_d$ that depends only on $d$, all positive and such that*

$$\frac{c^k}{\sigma^{2k} k!}\|\Delta_k\|^2 - \underline{C}_d \frac{\varepsilon^2}{\sigma^{2k+2}} \leq D(\theta \,\|\, \phi) \leq \frac{2}{\sigma^{2k} k!}\|\Delta_k\|^2 + \bar{C}\frac{\varepsilon^2}{\sigma^{2k+2}}\,.$$

*In particular, there exists positive $\sigma_0, \varepsilon_0$ that depend on $d$ such that for all $\sigma \geq \sigma_0$, and $\theta, \phi$ such that $\|\theta\| \leq 1, \rho(\theta, \phi) \leq \varepsilon_0$, it holds*

$$D(\theta \,\|\, \phi) \asymp \sigma^{-2k} \rho^2(\theta, \phi)\,,$$

*where the symbol $\asymp$ hides constants depending on $d$ but on no other parameters.*

Proof. The divergence $D(\theta \,\|\, \phi)$ and the tensors $\mathbb{E}[(R\phi)^{\otimes m}]$ are unaffected if we replace $\phi$ by $R\phi$ for any $R \in \mathcal{G}$. Hence without loss of generality, we can assume that $\|\theta - \phi\| = \rho(\theta, \phi) = \varepsilon$.

We first establish that for any $m \geq 1$, the quantities $\|\Delta_m\|$ are indeed of order at most $\varepsilon$. Specifically, we prove that

$$\|\mathbb{E}[(R\theta)^{\otimes m} - (R\phi)^{\otimes m}]\|^2 \leq m^2 2^{m-1} \varepsilon^2 \tag{4.1}$$

for all $m \geq 1$ and $\varepsilon \in (0, 1)$. This implies that $\|\Delta_k\| = \Theta(\varepsilon)$.

To prove (4.1), note that by Jensen's inequality,

$$\|\mathbb{E}[(R\theta)^{\otimes m} - (R\phi)^{\otimes m}]\|^2 \leq \mathbb{E}\|(R\theta)^{\otimes m} - (R\phi)^{\otimes m}\|^2 = \|\theta^{\otimes m} - \phi^{\otimes m}\|^2\,.$$

Expanding the norm yields

$$\begin{aligned}\|\theta^{\otimes m} - \phi^{\otimes m}\|^2 &= \|\theta\|^{2m} - 2\langle\theta, \phi\rangle^m + \|\phi\|^{2m} \\ &= \|\theta\|^{2m}(1 - 2(1+\gamma)^m + (1 + 2\gamma + \delta^2)^m)\,,\end{aligned}$$

where $\delta = \varepsilon/\|\theta\|$ and $\gamma = \langle\theta, \phi - \theta\rangle/\|\theta\|^2$ is such that $|\gamma| \leq \delta$ by Cauchy-Schwarz. Using Lemma B.1 in the Supplementary Materials, we conclude that

$$\|\theta\|^{2m} - 2\langle\theta, \phi\rangle^m + \|\phi\|^{2m} \leq \|\theta\|^{2m} m^2 2^{m-1} \delta^2 \leq m^2 2^{m-1} \varepsilon^2\,,$$

as desired.

If $k < 3$, then the claim follows directly from Lemma 2 $(ii)$.

We therefore assume $k \geq 3$. Then $\|\mathbb{E}[R\theta - R\phi]\| = o(\varepsilon)$, so Lemma 2 $(i)$ implies

$$D(\theta \,\|\, \phi) = o(\varepsilon^2) + D(\vartheta \,\|\, \varphi)\,,$$

where $\vartheta = \theta - \mathbb{E}R\theta$ and $\varphi = \phi - \mathbb{E}R\phi$. Hence we can replace $\theta$ and $\phi$ by $\vartheta$ and $\varphi$ without affecting $D(\theta \,\|\, \phi)$ by more than $o(\varepsilon^2)$. We therefore assume without loss of generality that $\mathbb{E}R\theta = \mathbb{E}R\phi = 0$.

We first show the upper bound. Denote by $\mathsf{g}$ the density of a standard $d$-dimensional Gaussian random variable. For all $\zeta \in \mathbb{R}^d$, let $f_\zeta$ denote the density of $\mathrm{P}_\zeta$. Then

$$f_\zeta(y) = \mathbb{E}_R \frac{1}{\sigma} \mathsf{g}(\sigma^{-1}(y - R\zeta)) = \frac{1}{\sigma} \mathsf{g}(\sigma^{-1} y) e^{-\frac{\|\zeta\|^2}{2\sigma^2}} \mathbb{E}_R e^{\frac{y^\top R\zeta}{\sigma^2}} .$$

Let $\chi^2(\theta, \phi)$ denote the $\chi^2$-divergence between $\mathrm{P}_\theta$ and $\mathrm{P}_\phi$, defined by

$$\chi^2(\theta, \phi) = \int \frac{(f_\theta(y) - f_\phi(y))^2}{f_\theta(y)} \mathrm{d}y \, .$$

Since $\mathbb{E} R\theta = 0$ by assumption, Jensen's inequality implies

$$f_\theta(y) \geq \frac{1}{\sigma} \mathsf{g}(\sigma^{-1} y) e^{-\frac{\|\theta\|^2}{2\sigma^2}} e^{\mathbb{E}_R \frac{y^\top R\theta}{\sigma^2}} = \frac{1}{\sigma} \mathsf{g}(\sigma^{-1} y) e^{-\frac{\|\theta\|^2}{2\sigma^2}} .$$

Hence

$$\frac{(f_\theta(y) - f_\phi(y))^2}{f_\theta(y)} \leq e^{\frac{\|\theta\|^2}{2\sigma^2}} \big( e^{-\frac{\|\theta\|^2}{2\sigma^2}} \mathbb{E}_R e^{\frac{y^\top R\theta}{\sigma^2}} - e^{-\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}_R e^{\frac{y^\top R\phi}{\sigma^2}} \big)^2 \big( \frac{1}{\sigma} \mathsf{g}(\sigma^{-1} y) \big) .$$

Integrating this quantity with respect to $y$ yields a bound on the $\chi^2$ divergence. Let $\xi \sim \mathcal{N}(0, I_d)$ and observe that

$$\begin{aligned} \chi^2(\theta, \phi) &\leq \mathbb{E}_\xi \Big[ e^{\frac{\|\theta\|^2}{2\sigma^2}} \big( e^{-\frac{\|\theta\|^2}{2\sigma^2}} \mathbb{E}_R e^{\frac{\sigma\xi^\top R\theta}{\sigma^2}} - e^{-\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}_R e^{\frac{\sigma\xi^\top R\phi}{\sigma^2}} \big)^2 \Big] \\ &= \mathbb{E}_{\xi, R, R'} \Big[ e^{\frac{\|\theta\|^2}{2\sigma^2}} \big( e^{-\frac{\|\theta\|^2}{\sigma^2}} e^{\frac{\xi^\top (R + R')\theta}{\sigma}} - 2 e^{-\frac{\|\theta\|^2 + \|\phi\|^2}{2\sigma^2}} e^{\frac{\xi^\top (R\theta + R'\phi)}{\sigma}} \\ &\qquad + e^{-\frac{\|\phi\|^2}{\sigma^2}} e^{\frac{\xi^\top (R + R')\phi}{\sigma}} \big) \Big] , \end{aligned}$$

where $R$ and $R'$ are independent elements selected uniformly from $\mathcal{G}$.

Taking expectations with respect to $\xi$ yields

$$\chi^2(\theta, \phi) \leq 2 \mathbb{E}_R \big[ e^{\frac{\theta^\top R\theta}{\sigma^2}} - 2 e^{\frac{\theta^\top R\phi}{\sigma^2}} + e^{\frac{\phi^\top R\phi}{\sigma^2}} \big] ,$$

Where we used that $e^{\|\theta\|^2/(2\sigma^2)} \leq 2$ for $\sigma \geq 1$ and $\|\theta\|^2 \leq 1$.

The random variables $R\theta$ and $R\phi$ have moment generating functions that converge in a neighborhood of the origin, hence

$$\begin{aligned} \chi^2(\theta, \phi) &\leq \sum_{m \geq 0} \frac{2}{\sigma^{2m} m!} \mathbb{E}_R \left[ (\theta^\top R\theta)^m - 2(\theta^\top R\phi)^m + (\phi^\top R\phi)^m \right] \\ &= \sum_{m \geq 0} \frac{2}{\sigma^{2m} m!} \|\Delta_m\|^2 \\ &\leq \frac{2}{\sigma^{2k} k!} \|\Delta_k\|^2 + \varepsilon^2 \sum_{m \geq k+1} \frac{2m^2}{\sigma^{2m} m!} + o(\varepsilon^2) \\ &\leq \frac{2}{\sigma^{2k} k!} \|\Delta_k\|^2 + \bar{C} \frac{\varepsilon^2}{\sigma^{2k+2}} . \end{aligned}$$

The upper bound then follows from the inequality $D(\theta \,\|\, \phi) \leq \chi^2(\theta, \phi)$ [Tsy09].

We now turn to the lower bound. As before, we assume that $k \geq 3$ and define

$$d_1(\theta, \phi) = \int |f_\theta(y) - f_\phi(y)|\, dy\,.$$

Recall that by Pinsker's Inequality, we have $D(\theta \,\|\, \phi) \geq \frac{1}{2} d_1^2(\theta, \phi)$. Moreover,

$$\begin{aligned} d_1(\theta,\phi) &= \frac{1}{(2\pi)^{d/2}\sigma} \int \left| \mathbb{E} e^{-\frac{\|y-R\theta\|^2}{2\sigma^2}} - \mathbb{E} e^{-\frac{\|y-R\phi\|^2}{2\sigma^2}} \right| dy \\ &= \mathbb{E}_\xi \left| \mathbb{E}_R \left[ e^{\frac{2\sigma\xi^\top R\theta - \|\theta\|^2}{2\sigma^2}} \middle| \xi \right] - \mathbb{E}_R \left[ e^{\frac{2\sigma\xi^\top R\phi - \|\phi\|^2}{2\sigma^2}} \middle| \xi \right] \right|, \qquad \xi \sim \mathcal{N}(0, I_d) \\ &= e^{-\frac{\|\phi\|^2}{2\sigma^2}} \mathbb{E}_\xi \left| \mathbb{E}_R \left[ e^{\frac{\xi^\top R\theta}{\sigma}} \middle| \xi \right] e^{\frac{\|\phi\|^2 - \|\theta\|^2}{2\sigma^2}} - \mathbb{E}_R \left[ e^{\frac{\xi^\top R\phi}{\sigma}} \middle| \xi \right] \right|. \end{aligned}$$

We now show that $|\|\theta\|^2 - \|\phi\|^2| = o(\varepsilon)$. Indeed

$$\begin{aligned} \left| \|\theta\|^2 - \|\phi\|^2 \right| &= \left| \mathbb{E} \sum_{i=1}^d \left[ (R\theta)_i^2 - (R\phi)_i^2 \right] \right| \\ &\leq \sum_{i=1}^d \left| \mathbb{E}[(R\theta)_i^2] - \mathbb{E}[(R\phi)_i^2] \right| \\ &\leq \sum_{i,j=1}^d \left| \mathbb{E}[(R\theta)_i (R\theta)_j] - \mathbb{E}[(R\phi)_i (R\phi)_j] \right| \\ &\leq d \|\mathbb{E}(R\theta)^{\otimes 2} - (R\phi)^{\otimes 2}\| = d\|\Delta_2\| = o(\varepsilon)\,, \end{aligned}$$

where we applied Cauchy-Schwarz to get the second inequality and last equality follows from the assumption of the theorem when $k \geq 3$. Since $\|\theta\| \leq 1$ and $\sigma \geq 1$, we have

$$e^{\frac{\|\phi\|^2 - \|\theta\|^2}{2\sigma^2}} = 1 + o(\varepsilon) \quad \text{and} \quad \frac{\|\phi\|^2}{2\sigma^2} \leq 1$$

for $\varepsilon$ small enough. It yields

$$\begin{aligned} d_1(\theta,\phi) &\geq e^{-1} \mathbb{E}_\xi \left| \mathbb{E}_R \left[ e^{\frac{\xi^\top R\theta}{\sigma}} \middle| \xi \right] \big(1 + o(\varepsilon)\big) - \mathbb{E}_R \left[ e^{\frac{\xi^\top R\phi}{\sigma}} \middle| \xi \right] \right| \\ &\geq e^{-1} \mathbb{E}_\xi \left| \mathbb{E}_R \left[ e^{\frac{\xi^\top R\theta}{\sigma}} \middle| \xi \right] - \mathbb{E}_R \left[ e^{\frac{\xi^\top R\phi}{\sigma}} \middle| \xi \right] \right| - e^{\frac{\|\theta\|^2}{2\sigma^2} - 1} o(\varepsilon) \\ &= e^{-1} \mathbb{E}_\xi \Big| \sum_{m \geq 1} \frac{1}{\sigma^m m!} \langle \Delta_m, \xi^{\otimes m} \rangle \Big| - o(\varepsilon) \\ &\geq e^{-1} \mathbb{E}_\xi \Big| \frac{1}{\sigma^k k!} \langle \Delta_k, \xi^{\otimes k} \rangle \Big| - e^{-1} \sum_{m \neq k} \mathbb{E}_\xi \Big| \frac{1}{\sigma^m m!} \langle \Delta_m, \xi^{\otimes m} \rangle \Big| - o(\varepsilon) \\ &\geq \frac{c^k}{\sigma^k \sqrt{k!}} \|\Delta_k\| - e^{-1} \sum_{m \neq k} \frac{\sqrt{(d+m)^m}}{\sigma^m m!} \|\Delta_m\| - o(\varepsilon) \\ &\geq \frac{c^k}{\sigma^k \sqrt{k!}} \|\Delta_k\| - C_d \frac{\varepsilon}{\sigma^{k+1}} \end{aligned}$$

where in the penultimate inequality, we employ a Khinchine-type inequality due to [Bob00], details of which appear as Lemma B.2 in the Supplementary Materials. Together with Pinsker's inequality and the fact that $\|\Delta_k\| = \Theta(\varepsilon)$ where $\varepsilon = \rho(\theta, \phi)$, this completes the proof of the lower bound. □

## 5. MAXIMUM LIKELIHOOD ESTIMATION

Let $Y_1, \dots, Y_n$ be i.i.d observations from the model (2.1) and consider the MLE $\tilde{\theta}_n$ that was defined in (2.2). In this section, we prove our main statistical result, that is a *uniform* upper bound on the rate of convergence of the MLE.

Our proof technique extends beyond the framework of the MRA model and can be broadly applied to derive uniform rates of convergence for the MLE from tight bounds on the KL divergence like Theorem 3. While similar ideas are often employed to obtain *pointwise* rates of convergence, extension to uniform rates requires novel elements. From here on, positive constants may depend on $d$ unless noted otherwise.

We first establish an upper bound for the MLE under a general lower bound for the KL divergence. Then we specialize this result to obtain the minimax upper bounds over $\mathcal{T}_s$ that are presented in Theorem 1.

### 5.1 A general upper bound

THEOREM 4. *Let $R$ be a random element drawn according to the Haar probability measure on any compact subgroup $\mathcal{G}$ of the orthogonal group in $d$ dimensions. Assume that there exist $k \geq 1$ and positive $\sigma_0, \varepsilon_0, c_0$ that depend on $d$ such that the following hold: for all $\sigma \geq \sigma_0$, and $\theta, \phi$ such that $\theta \in \mathcal{T}$,*

$$D(\theta \,\|\, \phi) \geq C\sigma^{-2k}\rho^2(\theta,\phi) \qquad \forall\, \phi \in \mathbb{R}^d \;\; \rho(\theta,\phi) \in [0, \varepsilon_0)\,, \tag{5.1}$$

$$D(\theta \,\|\, \phi) \geq C\sigma^{-\ell} \qquad \forall\, \phi \in \mathbb{R}^d \;\; \rho(\theta,\phi) \in [\varepsilon_0, c_0\sigma)\,, \tag{5.2}$$

$$D(\theta \,\|\, \phi) \geq C\sigma^{-2k}\rho^2(\theta,\phi) \qquad \forall\, \phi \in \mathbb{R}^d \;\; \rho(\theta,\phi) \in [c_0\sigma, \infty)\,. \tag{5.3}$$

*Then the MLE $\tilde{\theta}_n$ satisfies uniformly over such $\theta$,*

$$\mathbb{E}_\theta[\rho(\tilde{\theta}_n, \theta)] \leq C\frac{\sigma^k}{\sqrt{n}} + C_\sigma \frac{\log n}{n}\,, \tag{5.4}$$

*where $C_\sigma \leq C\sigma^{2\ell+10k}$.*

PROOF. The symbols $c$ and $C$ denote constants whose value may change from line to line. In the rest of this proof, we write $\tilde{\theta} = \tilde{\theta}_n$ to denote the MLE.

The main goal of the proof is to combine control of the curvature of the function $D$ with control of the deviations of the log-likelihood function.

Define the event $\mathcal{E} = \{\rho(\tilde{\theta}, \theta) \leq \varepsilon\}$ where $\varepsilon \leq \varepsilon_0$. Since $D$ is invariant under the action of $\mathcal{G}$, we can assume without loss of generality that $\rho(\tilde{\theta}, \theta) = \|\tilde{\theta} - \theta\|$. We first establish that on this event $\|\tilde{\theta} - \theta\|$ can be controlled in terms of the metric induced by the Hessian of $D$ at $\theta$.

Fix $\theta \in \mathbb{R}^d$ and denote by $H$ the Hessian of the function $\phi \mapsto D(\theta \,\|\, \phi)$ evaluated at $\phi = \theta$. For any $u \in \mathbb{R}^d$, define $\|u\|_H = \sqrt{u^\top H u}$.

On $\mathcal{E}$, $D(\theta \,\|\, \tilde{\theta}) \geq C\sigma^{-2k}\rho^2(\theta, \tilde{\theta})$ and $\rho(\theta, \tilde{\theta}) = \|\tilde{\theta} - \theta\|$ so using a third order Taylor expansion, we get

$$\big|D(\theta \,\|\, \tilde{\theta}) - \frac{1}{2}\|\tilde{\theta} - \theta\|_H^2\big| \leq C\frac{\|\tilde{\theta} - \theta\|^3}{\sigma^3} \leq C\varepsilon\sigma^{2k-3}\|\tilde{\theta} - \theta\|_H^2\,.$$

Therefore, there exists a $C$ such that if $\varepsilon < C\sigma^{3-2k}$, we get

$$D(\theta \,\|\, \tilde{\theta}) \geq \frac{1}{4}\|\tilde{\theta} - \theta\|_H^2 \geq C\sigma^{-2k}\|\tilde{\theta} - \theta\|^2\,. \tag{5.5}$$

We now control the geometry of the log-likelihood function near $\theta$. Define

$$D_n(\theta \,\|\, \phi) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta}{f_\phi}(Y_i)\,,$$

where $Y_i$ are i.i.d from $\mathrm{P}_\theta$ and we recall that $f_\zeta$ is the density of $\mathrm{P}_\zeta, \zeta \in \mathbb{R}^d$. Note that $D_n(\theta \,\|\, \theta) = 0$, so $D_n(\theta \,\|\, \tilde{\theta}) \le 0$.

Since $\theta$ is held fixed throughout the proof, we abbreviate $D(\theta \,\|\, \phi)$ and $D_n(\theta \,\|\, \phi)$ as $D(\phi)$ and $D_n(\phi)$, respectively.

Using a second order Taylor expansion, we get

$$D(\tilde{\theta}) - D_n(\tilde{\theta}) = -\nabla D_n(\theta)^\top h + \frac{1}{2} h^\top \nabla^2 (D - D_n)(\bar{\theta}) h\,,$$

where $h = \tilde{\theta} - \theta$ and $\bar{\theta}$ lies on a segment between $\tilde{\theta}$ and $\theta$.

For all $\zeta \in \mathcal{T}$, write $H_n(\zeta)$ for the Hessian of $D_n(\phi)$ evaluated at $\phi = \zeta$, and similarly let $H(\zeta)$ be the Hessian of $D(\phi)$ evaluated at $\phi = \zeta$. We obtain

$$\frac{1}{4}\|h\|_H^2 \le D(\tilde{\theta}) - D_n(\tilde{\theta}) \le -\nabla D_n(\theta)^\top h + \frac{1}{2} h^\top (H(\bar{\theta}) - H_n(\bar{\theta})) h\,. \tag{5.6}$$

For the first term, we note that if $H$ were invertible (and hence $\|\cdot\|_H$ a genuine metric), then it is well known (see, e.g. [HUL01]) that

$$\sup_{u:\|u\|_H=1} -\nabla D_n(\theta)^\top u = \|\nabla D_n(\theta)\|_H^* = \|\nabla D_n(\theta)\|_{H^{-1}}\,,$$

where $\|\cdot\|_H^*$, the dual norm to $\|\cdot\|_H$, is such that $\|\cdot\|_H^* = \|\cdot\|_{H^{-1}}$. In general, $H$ is not invertible, but we still have $\|\cdot\|_H^* = \|\cdot\|_{H^\dagger}$, where $H^\dagger$ denotes the Moore-Penrose pseudo-inverse of the matrix $H$.

To control the second, note first that by (5.5), it holds $\|h\| \le C\sigma^k \|h\|_H$. Therefore,

$$h^\top (H(\bar{\theta}) - H_n(\bar{\theta})) h \le C\sigma^k \|h\|_H \|h\| \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}\,,$$

where $\mathcal{B}_\varepsilon := \{\phi \in \mathbb{R}^d \,:\, \rho(\phi, \theta) \le \varepsilon\}$.

Combining the above three bounds and dividing by $\|h\|_H$, we get that on $\mathcal{E}$,

$$\frac{1}{4}\|h\|_H \le \|D_n(\theta)\|_{H^\dagger} + C\sigma^{2k} \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}^2 + \|h\|^2\,,$$

where we applied Young's inequality.

On the other hand, using (5.5), we have $\|h\|_H \ge C\sigma^{-k}\rho(\tilde{\theta}, \theta)$. Therefore, applying the Cauchy-Schwarz inequality and Chebyshev's inequality, we get

$$\begin{aligned}
\mathbb{E}[\rho(\tilde{\theta}, \theta)] &= \mathbb{E}[\rho(\tilde{\theta}, \theta)\mathbb{1}_{\mathcal{E}}] + \mathbb{E}[\rho(\tilde{\theta}, \theta)\mathbb{1}_{\mathcal{E}^c}] \\
&\le C\sigma^k \mathbb{E}[\|h\|_H \mathbb{1}_{\mathcal{E}}] + \big(\mathbb{E}[\rho^2(\tilde{\theta}, \theta)]\big)^{1/2} \big(\mathbb{P}(\mathcal{E}^c)\big)^{1/2} \\
&\le C\sigma^k \mathbb{E}[\|h\|_H \mathbb{1}_{\mathcal{E}}] + \varepsilon^{-1} \mathbb{E}[\rho^2(\tilde{\theta}, \theta)]\,.
\end{aligned}$$

Thus,

$$\mathbb{E}[\rho(\tilde{\theta}, \theta)] \le C\Big(\sigma^k \mathbb{E}\|D_n(\theta)\|_{H^\dagger} + \sigma^{3k}\mathbb{E} \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}^2 + (\sigma^k + \varepsilon^{-1})\mathbb{E}[\rho^2(\tilde{\theta}, \theta)]\Big) \tag{5.7}$$

It suffices to control the right side of the above inequality. The main term is the first one: Jensen's inequality and the second Bartlett identity imply

$$\mathbb{E}\|D_n(\theta)\|_{H^\dagger} \le \sqrt{\operatorname{tr}(H^\dagger \mathbb{E}[D_n(\theta)D_n(\theta)^\top])} = \sqrt{\frac{1}{n}\operatorname{tr}(H^\dagger H)} \le \sqrt{\frac{d}{n}}\,. \tag{5.8}$$

Standard matrix concentration bounds can be applied to show

$$\mathbb{E}\sup_{\phi\in\mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}^2 \le C\frac{\log n}{n\sigma^4}\,. \tag{5.9}$$

A proof of (5.9) appears as Lemma B.4 in the Supplementary Materials.

Likewise, a standard slicing argument, Lemma B.5 in the Supplementary Materials, implies

$$\mathbb{E}[\rho^2(\tilde\theta,\theta)] \le C\frac{\sigma^{(8k-4)\wedge(2\ell+2)}}{n}\,. \tag{5.10}$$

Plugging (5.8), (5.9), and (5.10) into (5.7), we get

$$\mathbb{E}[\rho(\tilde\theta,\theta)] \le C\Big(\frac{\sigma^k}{\sqrt{n}} + \frac{\sigma^{5k-4}\log n}{n} + \frac{\sigma^{(9k-4)\wedge(2\ell+k+2)}}{n\varepsilon}\Big)\,,$$

as desired. □

## 5.2 A modified MLE

The MLE itself may not achieve the optimal rate of convergence because the lower bound (5.1) may not be satisfied with the optimal choice of $k$ when $\phi$ is a specific perturbation of $\theta$. Namely, in the specific case of cyclic shifts, the divergence $D(\phi)$ is not curved enough in directions that perturb a null Fourier coefficient of $\theta$. To overcome this limitation, we split the sample $Y_1,\dots,Y_n$ into two parts: with the first part we estimate the support of $\hat\theta$ under Assumption 1 and with the second part, we compute a maximum likelihood estimator constrained to have the estimated support.

Specifically, assume for simplicity that we have a sample $Y_1,\dots,Y_{2n}$ of size $2n$ and split it into two samples $\mathcal{Y}_1 = \{Y_1,\dots,Y_n\}$ and $\mathcal{Y}_2 = \{Y_{n+1},\dots,Y_{2n}\}$ of equal size.

*5.2.1 Fourier support estimation* We use the first subsample $\mathcal{Y}_1$ to construct a set $\tilde S$ that coincides with $\operatorname{psupp}(\hat\theta)$ with high probability. For any $j = -\lfloor d/2\rfloor,\dots,\lfloor d/2\rfloor$, define,

$$M_j = \frac{1}{n}\sum_{i=1}^n |\widehat{(Y_i)}_j|^2 - \sigma^2\,.$$

Define the set $\tilde S$ by

$$\tilde S = \Big\{ j \in \{-\lfloor d/2\rfloor,\dots,\lfloor d/2\rfloor\} \,:\, M_j \ge C\sigma^2\sqrt{\frac{\log n}{n}}\Big\}$$

The following proposition shows that $\tilde S = \operatorname{psupp}(\hat\theta)$ with high probability.

PROPOSITION 5. *Fix $\theta \in \mathbb{R}^d$. Assume that $n/(\log n) \geq C\sigma^4$. Then*

$$\mathbb{P}[\tilde{S} \neq \mathrm{psupp}(\hat{\theta})] \leq C\frac{\sigma^4}{n}\,.$$

PROOF. This follows from standard concentration arguments. A full proof appears in the Supplementary Materials. □

*5.2.2 Constrained MLE* We use the second sample to construct a constrained MLE. To that end, for any $S \subset \{1, \ldots, \lfloor d/2 \rfloor\}$, define the projection $P_S$ by

$$\widehat{P_S(\phi)}_j = \begin{cases} \hat{\phi}_j & \text{if } j \in S \cup -S \\ \hat{\phi}_0 & \text{if } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The image of $P_S$ is a $(2|S|+1)$-dimensional real vector space. For convenience, write $\phi_S = P_S\phi$ for any vector $\phi \in \mathbb{R}^d$.

Recall that $\mathcal{Y}_2 = \{Y_{n+1}, \ldots, Y_{2n}\}$ denotes the second subsample and define $Y_1^S, \ldots, Y_n^S$ by

$$Y_i^S = P_S Y_{n+i} + \sigma(I - P_S)\xi_i \qquad i = 1, \ldots, n\,.$$

where the $\xi_i \sim \mathcal{N}(0, I_d)$ are i.i.d, independent from $\mathcal{Y}_2$.

Define the modified MLE $\check{\theta}_n$ by

$$\check{\theta}_n = \operatorname*{argmax}_{\phi \in \mathrm{Im}(P_{\tilde{S}})} \frac{1}{n}\sum_{i=1}^{n} \log f_\phi(Y_i^{\tilde{S}})\,. \tag{5.11}$$

PROPOSITION 6. *Fix $2 \leq s \leq \lfloor d/2 \rfloor$, $\theta \in \mathcal{T}_s$ and let $S = \mathrm{psupp}(\hat{\theta})$. Then, there exist positive $\sigma_0, \varepsilon_0, c_0$ that depend on $d$ such that the following hold: for all $\sigma \geq \sigma_0$,*

$$D(\theta \,\|\, \phi) \geq C\sigma^{-4s+2}\rho^2(\theta,\phi) \quad \forall\, \phi \in \mathrm{Im}(P_S) \;\; \rho(\theta,\phi) \in [0, \varepsilon_0)\,, \tag{5.12}$$

$$D(\theta \,\|\, \phi) \geq C\sigma^{-4s+2} \quad \forall\, \phi \in \mathrm{Im}(P_S) \;\; \rho(\theta,\phi) \in [\varepsilon_0, c_0\sigma)\,, \tag{5.13}$$

$$D(\theta \,\|\, \phi) \geq C\sigma^{-2}\rho^2(\theta,\phi) \quad \forall\, \phi \in \mathrm{Im}(P_S) \;\; \rho(\theta,\phi) \in [c_0, \infty)\,. \tag{5.14}$$

A proof appears in Appendix A.

## 5.3 Proof of upper bound in Theorem 1

Define $\mathcal{R} = \{\tilde{S} = \mathrm{psupp}(\hat{\theta})\}$ and observe that

$$\mathbb{E}[\rho(\check{\theta}_n, \theta)] \leq \mathbb{E}[\rho(\check{\theta}_n, \theta)\mathbb{1}_{\mathcal{R}}] + \mathbb{E}[\rho(\check{\theta}_n, \theta)\mathbb{1}_{\mathcal{R}^c}]\,.$$

The first term is controlled by combining Proposition 6 and Theorem 4 to get

$$\mathbb{E}[\rho(\check{\theta}_n, \theta)\mathbb{1}_{\mathcal{R}}] \leq C\frac{\sigma^{2s-1}}{\sqrt{n}} + C_\sigma\frac{\log n}{n}\,,$$

where $C_\sigma \leq C\sigma^{28s-14}$.

To bound the second term, we use the Cauchy-Schwarz inequality and Proposition 5 to get

$$\mathbb{E}[\rho(\check{\theta}_n,\theta)\mathbb{1}_{\mathcal{R}^c}] \le C\frac{\sigma^2}{\sqrt{n}}\sqrt{\mathbb{E}[\rho(\check{\theta}_n,\theta)^2]}$$

We now show that $\mathbb{E}[\rho(\check{\theta}_n,\theta)^2]$ is bounded uniformly over all choices of $\tilde{S}$ by a constant multiple of $\sigma^2$ using a similar slicing argument as the one employed in the proof of Lemma B.5 in the Supplementary Materials.

By the triangle inequality,

$$\rho(\check{\theta}_n,\theta) \le \rho(\check{\theta}_n,\theta_{\tilde{S}}) + \rho(\theta_{\tilde{S}},\theta) \le \rho(\check{\theta}_n,\theta_{\tilde{S}}) + 1\,.$$

In view of (5.14), we have

$$\rho(\check{\theta}_n,\theta_{\tilde{S}})^2 \le (c_0\sigma)^2 + C\sigma^2 D(\theta_{\tilde{S}} \,\|\, \check{\theta}_n) \le (C^\circ\sigma)^2(1 + G_n(\theta_{\tilde{S}} \,\|\, \check{\theta}_n))\,,$$

for some constant $C^\circ$, where

$$G_n(\theta_{\tilde{S}} \,\|\, \check{\theta}_n) = D(\theta_{\tilde{S}} \,\|\, \check{\theta}_n) - D_n(\theta_{\tilde{S}} \,\|\, \check{\theta}_n)\,.$$

For $j \ge 0$, define $S_j = \{\phi \in \mathbb{R}^d \,:\, 2^j\sigma \le \rho(\phi,\theta_{\tilde{S}}) \le 2^{j+1}\sigma\}$ and let $J$ be such that $C^\circ \le 2^J \le 2C^\circ$. Observe that

$$\begin{aligned}
\mathbb{E}[\rho(\check{\theta}_n,\theta_{\tilde{S}})^2] &\le 4(C^\circ\sigma)^2 + \sum_{j\ge J}\mathbb{E}[\rho(\check{\theta}_n,\theta_{\tilde{S}})^2\mathbb{1}(\check{\theta}_n \in S_j)] \\
&\le 4(C^\circ\sigma)^2 + \sigma^2\sum_{j>J} 2^{2j+2}\mathbb{P}[\sup_{\phi\in S_j} G_n(\theta_{\tilde{S}} \,\|\, \phi) > C\frac{2^{2j}}{n}] \\
&\le 4(C^\circ\sigma)^2 + C\sigma^2\sum_{j\ge 0} 2^{2j}\exp(-C2^{2j}) \le C\sigma^2\,,
\end{aligned}$$

where we used (B.7) from the Supplementary Materials in the third inequality. We obtain

$$\mathbb{E}[\rho(\check{\theta}_n,\theta)^2] \le 2\mathbb{E}[\rho(\check{\theta}_n,\theta_{\tilde{S}})^2] + 2 \le C\sigma^2\,.$$

We have established that

$$\mathbb{E}[\rho(\check{\theta}_n,\theta)] \le C\Big(\frac{\sigma^{2s-1}}{\sqrt{n}} + \sigma^{26s-13}\frac{\log n}{n} + \frac{\sigma^3}{n}\Big)\,,$$

which completes the proof of Theorem 1.

## 6. MINIMAX LOWER BOUNDS

Our minimax lower bounds rely ultimately on Le Cam's classical two-point testing method [LeC73]. In particular, the version that we use requires an upper bound on the KL divergence, which can be obtained using Theorem 3 and a moment matching argument.

### 6.1 Moment matching

The lower bound of Theorem 1 follows from Proposition 7.

PROPOSITION 7. *Fix $2 \le s \le \lfloor d/2 \rfloor$ and let $\theta, \phi \in \mathbb{R}^d$ satisfy*

$$\hat{\theta}_m = \hat{\phi}_m = 0 \quad \text{for } m \notin \{\pm(s-1), \pm s\}$$

*and*

$$|\hat{\theta}_m| = |\hat{\phi}_m| \quad \text{for } m \in \{\pm(s-1), \pm s\}.$$

*If $R$ is drawn uniformly from $\mathcal{F}$, then for any $m = 1, \dots, 2s-2$, it holds*

$$\mathbb{E}_R[(R\theta)^{\otimes m}] = \mathbb{E}_R[(R\phi)^{\otimes m}]$$

PROOF. Fix $m \le 2s-2$. Since $\mathbb{E}_R[(R\theta)^{\otimes m}]$ and $\mathbb{E}_R[(R\phi)^{\otimes m}]$ are symmetric tensors, to show that they are equal it suffices to show that

$$\langle \mathbb{E}_R[(R\theta)^{\otimes m}], u^{\otimes m} \rangle = \langle \mathbb{E}_R[(R\phi)^{\otimes m}], u^{\otimes m} \rangle \qquad \forall\, u \in \mathbb{R}^d$$

or equivalently, that

$$\mathbb{E}_R[(u^\top R\theta)^m] = \mathbb{E}_R[(u^\top R\phi)^m] \qquad \forall\, u \in \mathbb{R}^d\,. \tag{6.1}$$

Consider the set $\mathcal{P} = \{\zeta \,:\, |\hat{\zeta}_j| = |\hat{\theta}_j|\,,\ \forall\, j\}$ and note that $\theta, \phi \in \mathcal{P}$. We show that the function $\zeta \mapsto \mathbb{E}_R[(u^\top R\zeta)^m]$ is constant on $\mathcal{P}$, which readily yields (6.1). For a fixed shift $R_z \in \mathcal{F}$, we obtain

$$u^\top R_z \zeta = \langle \hat{u}, \widehat{R_z \zeta} \rangle = \sum_{j=-\lfloor d/2 \rfloor}^{\lfloor d/2 \rfloor} \hat{u}_{-j} \hat{\zeta}_j z^j\,,$$

so

$$(u^\top R_z \zeta)^m = \sum_{j_1, \dots, j_m = -\lfloor d/2 \rfloor}^{\lfloor d/2 \rfloor} z^{j_1 + \cdots + j_m} \prod_{n=1}^{m} \hat{u}_{-j_n} \hat{\zeta}_{j_n}\,.$$

Taking expectations with respect to a uniform choice of $z$ yields

$$\mathbb{E}_R[(u^\top R\zeta)^m] = \sum_{j_1 + \cdots + j_m = 0} \prod_{n=1}^{m} \hat{u}_{-j_n} \hat{\zeta}_{j_n}\,, \tag{6.2}$$

where the sums are over all choices of coordinates $j_1, \dots, j_m \in \{-\lfloor d/2 \rfloor, \dots, \lfloor d/2 \rfloor\}$ whose sum is 0.

The Fourier transform of $\zeta$ is supported only on coordinates $\pm(s-1)$ and $\pm s$, so we may restrict our attention to sums involving only those coordinates. Suppose $j_1 + \cdots + j_m = 0$. Define

$$\alpha = |\{i : j_i = s-1\}| \qquad \beta = |\{i : j_i = -(s-1)\}|$$
$$\gamma = |\{i : j_i = s\}| \qquad \delta = |\{i : j_i = -s\}|$$

By assumption $j_1 + \cdots + j_m = 0$, so the tuple $(\alpha, \beta, \gamma, \delta)$ is a solution to

$$\alpha(s-1) + \beta(-(s-1)) + \gamma(s) + \delta(-s) = 0$$

or, equivalently,

$$(\alpha-\beta)(s-1)+(\gamma-\delta)s=0$$

Since $s-1$ and $s$ are coprime, $(\alpha-\beta)$ and $(\gamma-\delta)$ must be multiples of $s$ and $s-1$, respectively. Since $|\alpha-\beta|+|\gamma-\delta|\le m<2s-1$, in fact $\alpha-\beta=\gamma-\delta=0$.

Therefore the only $m$-tuples $(j_1,\dots,j_m)$ that appear in the sum on the right-hand side of (6.2) are those in which $+(s-1)$ and $-(s-1)$ occur an equal number of times and $+s$ and $-s$ occur an equal number of times. For such $m$-tuples, the product $\prod_{n=1}^m \hat{u}_{-j_n}\hat{\zeta}_{j_n}$ can be reduced to a product of terms of the form $\hat{u}_{-(s-1)}\hat{u}_{s-1}\hat{\zeta}_{s-1}\hat{\zeta}_{-(s-1)}$ and $\hat{u}_{-s}\hat{u}_s\hat{\zeta}_s\hat{\zeta}_{-s}$. Since $u$ and $\zeta$ are real vectors, $\hat{u}_j\hat{u}_{-j}=|u_j|^2$ and $\hat{\zeta}_j\hat{\zeta}_{-j}=|\hat{\zeta}_j|^2$ for all $j=-\lfloor d/2\rfloor,\dots,\lfloor d/2\rfloor$, so

$$\prod_{n=1}^m \hat{u}_{-j_n}\hat{\zeta}_{j_n}=(|\hat{u}_{s-1}|^2|\hat{\zeta}_{s-1}|^2)^a(|\hat{u}_s|^2|\hat{\zeta}_s|^2)^b\,,$$

where $a$ and $b$ are the number of occurrences of the pairs $\pm(s-1)$ and $\pm s$, respectively. This quantity depends only on the moduli $|\hat{\zeta}_s|$ and $|\hat{\zeta}_{s-1}|$, hence it is the same for all $\zeta\in\mathcal{P}$. This completes the proof of (6.1) and therefore the proof of the proposition. □

### 6.2 Proof of lower bound in Theorem 1

Fix $z=e^{i\delta}$ for $\delta=c_1\sigma^{2s-1}/\sqrt{n}$ for some constant $c_1>0$. Let $\tau$ be given by

$$\hat{\tau}_j=\begin{cases}1/2 & \text{if } |j|=s-1 \text{ or } s,\\ 0 & \text{otherwise.}\end{cases}$$

Let $\phi_n$ be given by

$$\widehat{\phi_n}_j=\begin{cases}1/2 & \text{if } |j|=s-1,\\ z/2 & \text{if } j=s,\\ z^*/2 & \text{if } j=-s,\\ 0 & \text{otherwise.}\end{cases}$$

Note that the support of $\widehat{\phi_n}$ and $\hat{\tau}$ lies in $[-s,s]$ and that $\|\phi_n-\tau\|$ and $\rho(\phi_n,\tau)$ are both bounded by $c_2\sigma^{2s-1}/\sqrt{n}$, where $c_2$ can be made arbitrarily small by taking $c_1$ small enough.

Theorem 3 and Proposition 7 imply that

$$D(\mathrm{P}_\tau^n\,\|\,\mathrm{P}_{\phi_n}^n)\le Cn\sigma^{-4s+2}\rho^2(\phi_n,\tau)\le c_3\,,$$

for a positive constant $c_3$ that can be made arbitrarily small by taking $c_1$ small enough. Using standard minimax lower bounds techniques [Tsy09], we get the desired result.

## APPENDIX A: OMITTED PROOFS

Proof of Lemma 2. We first prove the following simple expression:

$$D(\mathrm{P}_\theta\,\|\,\mathrm{P}_\phi)=D(\theta\,\|\,\phi)=\frac{1}{2\sigma^2}(\|\phi\|^2-\|\theta\|^2)+\mathbb{E}\log\frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta}\mid\xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi}\mid\xi]}\,,$$

where $\xi \sim \mathcal{N}(0, I_d)$.

This claim follows directly from the definition of divergence. Denoting by $\mathtt{g}(y)$ the density of a standard Gaussian random variable with respect to the Lebesgue measure on $\mathbb{R}^d$, we can write

$$\begin{aligned}\frac{d\mathrm{P}_\theta}{d\mathrm{P}_\phi}(y) &= \frac{\mathbb{E}[\mathtt{g}((y-R\theta)/\sigma)]}{\mathbb{E}[\mathtt{g}((y-R\phi)/\sigma)]} \\ &= \frac{\mathbb{E}\left[\exp\left(-\frac{1}{2\sigma^2}(\|y\|^2 - 2y^\top R\theta + \|R\theta\|^2)\right)\right)}{\mathbb{E}\left[\exp\left(-\frac{1}{2\sigma^2}(\|y\|^2 - 2y^\top R'\phi + \|R'\phi\|^2)\right)\right)} \\ &= \exp\left(\frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2)\right)\frac{\mathbb{E}\left[\exp\left(\frac{1}{\sigma^2}y^\top R\theta\right]\right)}{\mathbb{E}\left[\exp\left(\frac{1}{\sigma^2}y^\top R'\phi\right]\right)}\,,\end{aligned}$$

since $R$ is orthogonal. Hence

$$\begin{aligned}D(\theta \,\|\, \phi) &= \mathbb{E}_{Y\sim \mathrm{P}_\theta} \log \frac{d\mathrm{P}_\theta}{d\mathrm{P}_\phi}(Y) \\ &= \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \mathbb{E}_{Y\sim\mathrm{P}_\theta}\log\frac{\mathbb{E}[e^{\frac{1}{\sigma^2}Y^\top R\theta} \mid Y]}{\mathbb{E}[e^{\frac{1}{\sigma^2}Y^\top R\phi} \mid Y]}\,.\end{aligned}$$

When $Y \sim \mathrm{P}_\theta$, we can write $Y = R'\theta + \sigma\xi$ for a standard Gaussian vector $\xi$ and rotation $R'$ independent of $R$. Since $R'$ is orthogonal, this has the same distribution as $R'(\theta+\sigma\xi)$. If $R$ and $R'$ are independent and uniform, then $(R')^\top R$ has the same distribution as $R$, so

$$Y^\top R\theta =^d (\theta+\sigma\xi)^\top R\theta\,.$$

Therefore

$$D(\theta \,\|\, \phi) = \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \mathbb{E}_\xi \log\frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi} \mid \xi]}\,. \tag{A.1}$$

We now prove both parts of the Lemma

*(i)* For convenience write $\bar\theta = \mathbb{E}R\theta$ and $\bar\phi = \mathbb{E}R\phi$. These vectors satisfy $R\bar\theta = \bar\theta$ and $R\bar\phi = \bar\phi$ for all $R$. Hence

$$\begin{aligned}(\theta+\sigma\xi)^\top R\theta &= (\bar\theta + \vartheta + \sigma\xi)^\top R(\bar\theta+\vartheta) \\ &= (\vartheta+\sigma\xi)^\top R\vartheta + (\bar\theta+\sigma\xi)^\top\bar\theta\,,\end{aligned}$$

and similarly

$$(\theta+\sigma\xi)^\top R\phi = (\vartheta+\sigma\xi)^\top R\varphi + (\bar\theta+\sigma\xi)^\top\bar\phi\,.$$

Plugging these quantities into the expression for $D(\theta \,\|\, \phi)$ yields

$$\begin{aligned}D(\theta \,\|\, \phi) &= \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \mathbb{E}_\xi\log\frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^\top R\vartheta + (\bar\theta+\sigma\xi)^\top\bar\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^\top R\varphi + (\bar\theta+\sigma\xi)^\top\bar\phi} \mid \xi]} \\ &= \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \frac{1}{\sigma^2}\mathbb{E}_\xi(\bar\theta+\sigma\xi)^\top(\bar\theta-\bar\phi) + \mathbb{E}_\xi\log\frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^\top R\vartheta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\vartheta+\sigma\xi)^\top R\varphi} \mid \xi]} \\ &= \frac{1}{2\sigma^2}(\|\bar\phi\|^2 - \|\bar\theta\|^2) + \frac{1}{\sigma^2}(\|\bar\theta\|^2 - \bar\theta^\top\bar\phi) + D(\vartheta \,\|\, \varphi) \\ &= \frac{1}{2\sigma^2}\|\mathbb{E}[R\theta - R\phi]\|^2 + D(\vartheta \,\|\, \varphi)\,.\end{aligned}$$

*(ii)* The previous claim implies that it suffices to show

$$D(\theta \,\|\, \phi) = \frac{1}{4\sigma^4}\|\mathbb{E}[(R\theta)^{\otimes 2} - (R\phi)^{\otimes 2}]\|^2 + \sigma^{-6}O(\varepsilon^2)$$

for $\theta$ and $\phi$ satisfying $\mathbb{E}R\theta = \mathbb{E}R\phi = 0$.

We accomplish this by expanding the expression for $D(\theta \,\|\, \phi)$ given above as a power series in $\sigma^{-1}$. Recall

$$D(\theta \,\|\, \phi) = \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \mathbb{E}_\xi \log \frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi} \mid \xi]}\,.$$

Given a random variable $X$, write

$$K_X(t) = \log \mathbb{E}\exp(t^\top X)$$

for the cumulant generating function of $X$. Then

$$\log \frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi} \mid \xi]} = K_{R\theta}\left(\frac{1}{\sigma^2}(\theta-\sigma\xi)\right) - K_{R\phi}\left(\frac{1}{\sigma^2}(\theta-\sigma\xi)\right).$$

Denote by $\kappa^m$ and $\lambda^m$ the $m$th cumulant tensors of $R\theta$ and $R\phi$ respectively. Then

$$D(\theta \,\|\, \phi) = \frac{1}{2\sigma^2}(\|\phi\|^2 - \|\theta\|^2) + \mathbb{E}\sum_{m\geq 1}\frac{1}{\sigma^{2m}m!}\langle \kappa^m - \lambda^m, (\theta-\sigma\xi)^{\otimes m}\rangle\,.$$

Since $\kappa^1 = \mathbb{E}R\theta$ and $\lambda^1 = \mathbb{E}R\phi$,

$$\kappa^1 - \lambda^1 = \mathbb{E}R\theta - \mathbb{E}R\phi = 0\,,$$

so the first term in the sum vanishes. Moreover, since $\xi$ is a standard Gaussian, $\mathbb{E}[\xi^{\otimes m}] = 0$ for all odd $m$.

Rearranging to collect powers of $\sigma^{-1}$ yields

$$D(\theta \,\|\, \phi) = \frac{1}{\sigma^2}T_2 + \frac{1}{\sigma^4}T_4 + o(\sigma^{-4})\,,$$

where

$$\begin{aligned}
T_2 &= \frac{1}{2}(\|\phi\|^2 - \|\theta\|^2 + \mathbb{E}\langle \kappa^2 - \lambda^2, \xi^{\otimes 2}\rangle)\\
T_4 &= \frac{1}{2}\langle \kappa^2 - \lambda^2, \theta^{\otimes 2}\rangle + \frac{1}{2}\mathbb{E}\langle \kappa^3 - \lambda^3, \theta\otimes\xi^{\otimes 2}\rangle + \frac{1}{4!}\mathbb{E}\langle \kappa^4 - \lambda^4, \xi^{\otimes 4}\rangle\,.
\end{aligned}$$

Straightforward calculation yields

$$\begin{aligned}
\mathbb{E}\langle \kappa^2 - \lambda^2, \xi^{\otimes 2}\rangle &= \|\theta\|^2 - \|\phi\|^2\\
\langle \kappa^2 - \lambda^2, \theta^{\otimes 2}\rangle &= \mathbb{E}[(\theta^\top R\theta)^2 - (\phi^\top R\theta)^2]\\
\mathbb{E}\langle \kappa^3 - \lambda^3, \theta\otimes\xi^{\otimes 2}\rangle &= 0\\
\mathbb{E}\langle \kappa^4 - \lambda^4, \xi^{\otimes 4}\rangle &= 6\mathbb{E}[(\phi^\top R\phi)^2 - 6(\theta^\top R\theta)^2]\,.
\end{aligned}$$

Combining the above two displays, we conclude

$$\begin{aligned}
T_2 &= 0\\
T_4 &= \frac{1}{4}\mathbb{E}[(\theta^\top R\theta)^2 - 2(\phi^\top R\theta)^2 + (\phi^\top R\phi)] = \frac{1}{4}\|\mathbb{E}[(R\theta)^{\otimes 2} - (R\phi)^{\otimes 2}]\|^2\,,
\end{aligned}$$

and the claim follows. □

PROOF OF PROPOSITION 6. We first suppose $\rho(\theta,\phi) \le \varepsilon_0$.

For any $\phi \in \mathbb{R}^d$, we have $\mathbb{E}[R\phi] = d^{-1/2}\hat{\phi}_0\mathbf{1}$, where $\mathbf{1}$ is the all-ones vector of $\mathbb{R}^d$ and $\hat{\phi}_0$ is known as the DC component of $\phi$. Therefore $\varphi = \phi - \mathbb{E}[R\phi]$ has Fourier transform $\widehat{\varphi} = \hat{\phi} - \hat{\phi}_0 d^{-1/2} W\mathbf{1} = \hat{\phi} - \hat{\phi}_0 e_0$, where $e_0$ is the vector of $\mathbb{R}^d$ indexed over $\{-\lfloor d/2\rfloor, \dots, \lfloor d/2\rfloor\}$ with 1 on the 0 coordinate and 0 on other coordinates. In other words, the Fourier transform of $\varphi$ is the same as that of $\phi$ except that its DC component $\widehat{\varphi}_0$ is set to 0. Similarly, the Fourier transform of $\vartheta_S$ is the same as that of $\theta$ except that its DC component $\widehat{\vartheta}_0$ is set to 0.

By Part *(i)* of Lemma 2,

$$D(\phi) = \frac{1}{2\sigma^2}\|\mathbb{E}[R\theta - R\phi]\|^2 + D(\vartheta \,\|\, \varphi)\,, \tag{A.2}$$

Write $\rho(\theta,\phi) = \varepsilon$ and suppose first that $|\hat{\theta}_0 - \hat{\phi}_0| \ge \frac{1}{2}\varepsilon$. Then (A.2) implies

$$D(\phi) \ge \frac{1}{2\sigma^2}\|\mathbb{E}[R\theta - R\phi]\|^2 = \frac{1}{2\sigma^2}(\hat{\theta}_0 - \hat{\phi}_0)^2 \ge \frac{\varepsilon^2}{8d\sigma^2} \ge C\sigma^{-4s+2}\varepsilon^2\,.$$

Next, if $|\hat{\theta}_0 - \hat{\phi}_0| < \frac{1}{2}\varepsilon$, then

$$\rho(\vartheta_S,\varphi)^2 = \rho(\theta_S,\phi)^2 - |\hat{\theta}_0 - \hat{\phi}_0|^2 \ge 3\varepsilon^2/4\,.$$

Thus, it suffices to show that

$$D(\vartheta \,\|\, \varphi) \ge C\sigma^{-4s+2}\rho(\vartheta,\varphi)^2\,.$$

There are two cases: either $\vartheta_S$ and $\varphi$ have essentially the same power spectrum (i.e., $|\hat{\vartheta}_k| \approx |\hat{\varphi}_k|$ for all $k$) or their power spectra are very different. We will treat these two cases separately.

Recall that for each $j \in S$, by assumption $c^{-1} \le |\hat{\vartheta}_j| \le c$. Consider the polar form $\hat{\varphi}_j/\hat{\vartheta}_j = m_j e^{i\delta_j}$, where $m_j \ge 0$. Since $D(\varphi) = D(R\varphi)$ for all cyclic shifts, we may assume that $\|\vartheta - \varphi\| = \varepsilon$, so that $|1 - m_j| \le C\varepsilon$ for all $j$.

Suppose first that $|1 - m_j| \ge C\varepsilon$ for some $j \in S$. Lemma 2 *(ii)* yields

$$\begin{aligned}
D(\varphi) &= \frac{1}{4\sigma^4}\|\mathbb{E}[(R\vartheta)^{\otimes 2} - (R\varphi)^{\otimes 2}]\|^2 + C\frac{\varepsilon^2}{\sigma^6} \\
&\ge \frac{1}{4\sigma^4}(|\hat{\vartheta}_j|^2 - |\varphi_j|^2)^2 + C\frac{\varepsilon^2}{\sigma^6} \\
&\ge \frac{c^{-2}}{4\sigma^4}|\hat{\vartheta}_j|^2(1 - m_j)^2 + C\frac{\varepsilon^2}{\sigma^6} \\
&\ge \frac{c^{-4}}{4\sigma^4}(1 - m_j)^2 - C\frac{\varepsilon^2}{\sigma^6} \ge C\sigma^{-4}\varepsilon^2\,.
\end{aligned}$$

Hence $D(\varphi) \ge C\sigma^{-4s+2}\varepsilon^2$.

Next, suppose on the contrary that $|1 - m_j| = o(\varepsilon)$ for all $j \in S$. Since $\|\vartheta - \varphi\| = \varepsilon$, we can take the relative phase $\delta_j$ in the polar form to be such that $\delta_j \le C\varepsilon$ for all $j$. By Theorem 3, it is enough to show that there exists an $m \le 2s - 1$ such that

$$\|\mathbb{E}[(R\vartheta)^{\otimes m} - (R\varphi)^{\otimes m}]\|^2 = \Omega(\varepsilon^2)\,.$$

Denote by $p$ the smallest integer in $S$ and observe that

$$\varepsilon^2 = \rho(\vartheta,\varphi)^2 = \min_{z:|z|=1} \sum_{j=-\lfloor d/2\rfloor}^{\lfloor d/2\rfloor} |1 - m_j z^j e^{i\delta_j}|^2 |\hat{\vartheta}_j|^2 \le C \sum_{j\in S} |1 - m_j e^{-j\delta_p/p} e^{i\delta_j}|^2 .$$

Therefore, there exists a coordinate $\ell \in S \setminus \{p\}$ such that

$$|1 - e^{i(p\delta_\ell - \ell\delta_p)/p}|^2 = |1 - m_\ell e^{i(p\delta_\ell - \ell\delta_p)/p}|^2 + o(\varepsilon^2) \ge C\varepsilon^2 . \tag{A.3}$$

Choose $m = \ell + p$. Since $\ell, p \in S \subseteq [s]$ and $\ell \ne p$, the bound $m \le 2s-1$ holds. As in the proof of Proposition 7, we have that

$$\begin{aligned}\|\mathbb{E}[(R\vartheta_S)^{\otimes m} - (R\varphi)^{\otimes m}]\|^2 &= \sum_{j_1+\cdots+j_m=0} \left| \prod_{n=1}^m \hat{\vartheta}_{j_n} - \prod_{n=1}^m \hat{\varphi}_{j_n} \right|^2 \\ &= \sum_{j_1+\cdots+j_m=0} \left| 1 - \prod_{n=1}^m m_{j_n} e^{i\sum_{n=1}^m \delta_{j_n}} \right|^2 \prod_{n=1}^m |\hat{\vartheta}_{j_n}|^2 .\end{aligned}$$

Each term in the above sum is positive. One valid solution to the equation $j_1 + \cdots + j_m = 0$ is $j_1 = \cdots = j_\ell = -p$ and $j_{\ell+1} = \cdots = j_m = \ell$. We obtain

$$\begin{aligned}\|\mathbb{E}[(R\vartheta_S)^{\otimes m} - (R\varphi)^{\otimes m}]\|^2 &\ge C \left| 1 - \prod_{n=1}^m m_{j_n} e^{i(p\delta_\ell - \ell\delta_p)} \right|^2 \\ &= |1 - e^{i(p\delta_\ell - \ell\delta_p)}|^2 + o(\varepsilon^2) .\end{aligned}$$

Observe that for $\delta_\ell$ and $\delta_p$ small enough, it holds

$$|1 - e^{i(p\delta_\ell - \ell\delta_p)}|^2 \ge |1 - e^{i(p\delta_\ell - \ell\delta_p)/p}|^2 \ge C\varepsilon^2$$

where the last inequality follows from (A.3). Combining the above two displays proves (5.12).

Now suppose $\rho(\theta,\phi) \ge \varepsilon_0$. To show that $D(\theta \,\|\, \phi) \ge C\sigma^{-4s+2}$, it suffices to show by the chain rule for divergence that

$$D(P_\theta^N \,\|\, P_\phi^N) \ge c \quad \text{for some } N \le C\sigma^{4s-2} ,$$

where $c$ and $C$ are positive constants. We will show the existence of a test which correctly distinguishes $P_\theta^N$ from $P_\phi^N$ with probability at least $2/3$; the claim will then follow from the Neyman-Pearson lemma and Pinsker's inequality.

We first show the existence of such a test when $\|\phi\|^2 \ge 2$. Let $N = \gamma\sigma^4 \le \gamma\sigma^{4s-2}$ for $s \ge 2$, with $\gamma > 0$ to be chosen later. Let $\{Y_i\}_{i=1}^N$ be samples from either $P_\theta^N$ or $P_\phi^N$, and define a test $\psi : \mathbb{R}^{d\times N} \to \{\theta,\phi\}$ by

$$\psi(Y_1,\dots,Y_N) = \begin{cases} \theta & \text{if } \frac{1}{N}\sum_{i=1}^N \|Y_i\|^2 \le \sigma^2 + 1.5 \\ \phi & \text{otherwise.} \end{cases}$$

An easy computation shows that

$$\mathbb{E}_\theta[\frac{1}{N}\sum_{i=1}^N \|Y_i\|^2] = \sigma^2 + \|\theta\|^2 \le \sigma^2 + 1 \qquad \mathbb{E}_\phi[\frac{1}{N}\sum_{i=1}^N \|Y_i\|^2] = \sigma^2 + \|\phi\|^2$$

$$\mathrm{var}_\theta[\frac{1}{N}\sum_{i=1}^N \|Y_i\|^2] \le \frac{4\sigma^2 + 2\sigma^4}{N} \le \frac{C}{\gamma} \qquad \mathrm{var}_\phi[\frac{1}{N}\sum_{i=1}^N \|Y_i\|^2] \le \frac{C}{\gamma}(\sigma^{-2}\|\phi\|^2 + 1) .$$

Together with Chebyshev's inequality, we get that that for $\|\phi\|^2 \geq 2$,

$$P_\theta(\psi = \phi) + P_\phi(\psi = \theta) \leq \frac{C}{\gamma} \leq 1/3\,,$$

For $\gamma$ large enough, as desired.

Next, suppose that $\|\phi\|^2 \leq 2$. For positive integers $j, k$, denote by $(j,k)$ their greatest common divisor. Given a vector $\zeta \in \mathbb{R}^d$, denote by $\mathcal{P}$ the following set of polynomials in the entries of $\hat{\zeta}$:

$$\begin{aligned} p_0(\zeta) &= \hat{\zeta}_0, \\ p_j(\zeta) &= \|\hat{\zeta}_j\|^2 \quad \text{for } 1 \leq j \leq \lfloor d/2 \rfloor, \\ p_{jk}(\zeta) &= \hat{\zeta}_{-k}^{j/(j,k)} \hat{\zeta}_j^{k/(j,k)} \text{for } 1 \leq j,k \leq \lfloor s \rfloor. \end{aligned}$$

If $\rho(\theta,\phi) > 0$, then by [KI93, Corollary 2], there exists at least one polynomial $p \in \mathcal{P}$ such that

$$p(\theta) \neq p(\phi)\,.$$

It is easy to see that all the polynomials in $\mathcal{P}$ are invariant under the group action; that is,

$$p(\zeta) = p(R\zeta)$$

for any $R \in \mathcal{F}$ and $p \in \mathcal{P}$. This implies that, given a $p \in \mathcal{P}$, the value $p(\zeta)$ can be computed from the moment tensors $\mathbb{E}[(R\zeta + \sigma\xi)^{\otimes k}]$ for $1 \leq k \leq 2s-1$. Indeed, the entries of $\mathbb{E}[(R\zeta + \sigma\xi)^{\otimes k}]$ for $1 \leq k \leq 2s-1$ generate the ring of invariant polynomials in the entries of $\hat{\zeta}$ of degree at most $2s-1$, which includes the set $\mathcal{P}$. Given samples $Y_1, \ldots, Y_n$ from $\mathrm{P}_\zeta$, the tensor $\mathbb{E}[(R\zeta + \sigma\xi)^{\otimes k}]$ can be consistently estimated by computing the empirical moment tensors. We obtain that there exists a constant $M$, depending on $s$ but not on $\sigma$, such that for any $p \in \mathcal{P}$ there exists an unbiased estimator $\tilde{p}(\zeta)$ for $p(\zeta)$ with variance at most $M\sigma^{4s-2}/N$ for any $\zeta \in \{\theta, \phi\}$.

For all $\theta \in \mathbb{R}^d$, define $B_{\theta,\varepsilon_0} = \{\phi : \rho(\theta,\phi) \leq \varepsilon_0, \|\phi\|^2 \leq 2\}$. It is clear that $B_{\theta,\varepsilon_0}$ is compact so that

$$\delta = \inf_{\theta:\|\theta\|^2 \leq 1} \inf_{\phi \in B_{\theta,\varepsilon_0}} \min_{p \in \mathcal{P}: p(\phi) \neq p(\theta)} |p(\phi) - p(\theta)| > 0\,.$$

Note that $\delta$ does not depend on $\theta$ or $\phi$. Set $N = \gamma \delta^{-2} M \sigma^{4s-2}$, where $\gamma$ is to be chosen later. Since $\rho(\theta,\phi) \geq \varepsilon_0$ by assumption, there exists a $p \in \mathcal{P}$ such that $|p(\phi) - p(\theta)| \geq \delta$.

Let $\psi : \mathbb{R}^{d \times N} \to \{\theta, \phi\}$ be the test

$$\psi(Y_1, \ldots, Y_n) = \operatorname*{argmin}_{\zeta \in \{\theta,\phi\}} |\tilde{p}(\zeta) - p(\zeta)|\,.$$

By Chebyshev's inequality, we have

$$P_\theta(\psi = \phi) + P_\phi(\psi = \theta) \leq \frac{C}{\gamma} \leq 1/3\,,$$

for $\gamma$ sufficiently large.

To conclude the proof, observe that by the chain rule, Pinsker's inequality, and the Neyman-Pearson lemma respectively, we get

$$D(\theta \,\|\, \phi) = \frac{1}{N} D(P_\theta^N \,\|\, P_\phi^N) \geq \frac{C}{N} \geq C\sigma^{-4s+2}\,.$$

The desired result (5.13) follows.

Finally, let $\rho(\theta, \phi) \geq c_0\sigma$ where $c_0 = 32\sqrt{2d}$. By (A.1), since $\|\theta\| \leq 1$, we have

$$D(\phi) \geq \frac{1}{2\sigma^2}(\|\phi\|^2 - 1) + \mathbb{E}_\xi \log \frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi} \mid \xi]}\,,$$

where $\xi \sim \mathcal{N}(0, I_d)$. Next, using the Cauchy-Schwarz inequality and Jensen's inequality, we get

$$\mathbb{E}_\xi \log \frac{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\theta} \mid \xi]}{\mathbb{E}[e^{\frac{1}{\sigma^2}(\theta+\sigma\xi)^\top R\phi} \mid \xi]} \geq -\frac{1+\|\phi\|}{\sigma^2}\mathbb{E}\|\theta + \sigma\xi\| \geq -\frac{1+\|\phi\|}{\sigma^2}\sqrt{1+d\sigma^2}$$

Hence, for $\|\phi\| \geq 16\sigma\sqrt{2d}$, we get $D(\phi) \geq \|\phi\|^2/(8\sigma^2)$. Moreover, by the triangle inequality, $\|\phi\|/2 \leq \rho(\phi, \theta) \leq 2\|\phi\|$ for all such $\phi$. We obtainthat for all $\phi \in \mathbb{R}^d$ such that $\rho(\phi, \theta) \geq c_0\sigma$, it holds $D(\phi) \geq \|\phi - \theta\|^2/(32\sigma^2)$, which implies the desired result (5.14).

□

## APPENDIX B: SUPPLEMENTARY MATERIALS

LEMMA B.1. *For any $\delta \in (0, 1), \gamma \in [-\delta, \delta]$, it holds*

$$1 - 2(1+\gamma)^m + (1 + 2\gamma + \delta^2)^m \leq m^2 2^{m-1}\delta^2$$

PROOF. Fix $\delta \in (0, 1)$ and consider the function $\gamma \mapsto f(\gamma, \delta) = 1 - 2(1+\gamma)^m + (1+2\gamma+\delta^2)^m$. Computing the derivative of $f$, it is easy to see that $f$ is decreasing on $(-\delta, -\delta^2)$ and increasing on $(-\delta^2, \delta)$ so that it achieves its maximum either at $-\delta$ or at $\delta$.

Consider the function $\delta \mapsto g(\delta) = f(\delta, \delta) = 1 - 2(1+\delta)^m + (1+\delta)^{2m}$ and observe that $g(-\delta) = f(-\delta, \delta)$. Since $g(0) = g'(0) = 0$, and

$$g''(\delta) = 2m(1+\delta)^{m-2}[(2m-1)(1+\delta)^m - (m-1)] \leq m^2 2^m \quad \forall\, \delta \in [-1, 1]\,,$$

it follows from a second order Taylor expansion that

$$g(\delta) \leq m^2 2^{m-1}\delta^2$$

□

LEMMA B.2. *Let $\xi \sim \mathcal{N}(0, I_d)$ be a standard $d$-dimensional Gaussian. Then for any symmetric order-$m$ tensor and $A \in \mathbb{R}^{d^m}$ there exists constant $c = c(A)$, independent of $m$ and $d$, such that*

$$c^m\sqrt{m!}\|A\| \leq \mathbb{E}|\langle A, \xi^{\otimes m}\rangle| \leq \sqrt{(d+m)^m}\|A\|\,.$$

PROOF. Assume without loss of generality that $\|A\| = 1$. The upper bound follows readily from the Cauchy-Schwarz inequality and the fact that $\|\xi\|^2 \sim \chi^2_d$,

$$\mathbb{E}_\xi |\langle A, \xi^{\otimes m}\rangle| \le \mathbb{E}_\xi \|\xi^{\otimes m}\| = \mathbb{E}_\xi \|\xi\|^m \le (d+m)^{m/2}\,.$$

To prove the lower bound, we first show that

$$\mathbb{E}_\xi [\langle A, \xi^{\otimes m}\rangle^2] \ge m!\,. \tag{B.1}$$

We do so by expressing the left side in terms of multivariate Hermite polynomials.

Recall that the Hermite polynomials $\{h_k(x)\}_{k\ge 0}$ can be normalized to satisfy the following two properties:

1. The function $h_k(x)$ is a polynomial with leading term $x^k/\sqrt{k!}$,
2. The functions $\{h_k\}_{k\ge 0}$ form an orthonormal basis of of $L_2(\mu)$, where $\mu$ denotes the standard Gaussian measure on $\mathbb{R}$.

Given a multi-index $\alpha \in \mathbb{N}^d$, define the multivariate Hermite polynomial $h_\alpha$ by

$$h_\alpha(x_1, \dots, x_d) = \prod_{i=1}^d h_{\alpha_i}(x_i)\,.$$

The multivariate Hermite polynomials form an orthonormal basis for the space $\mathbb{R}[x_1, \dots, x_d]$ of $d$-variate polynomial functions with respect to the inner product over $L_2(\mu^{\otimes d})$.

Multivariate Hermite polynomials satisfy the following useful property: Given two multi-indices $\alpha, \beta \in \mathbb{N}^d$ such that $|\alpha| = |\beta|$, we now show that

$$\langle x^\beta, h_\alpha\rangle = \sqrt{\alpha!}\delta_{\alpha\beta}\,, \tag{B.2}$$

where $\delta$ denotes the Kronecker symbol. Indeed, on the one hand if $\alpha \ne \beta$, then since $|\alpha| = |\beta|$ there exists an index $i$ such that $\alpha_i > \beta_i$. By the definition of the univariate Hermite polynomials, $x^{\beta_i} \in \operatorname{span}(h_1(x), \dots, h_{\beta_i}(x))$, hence orthogonality of the polynomials $h_{\alpha_i}$ and $h_j$ for $1 \le j \le \beta_i$ implies

$$\langle x^\beta, h_\alpha\rangle = \prod_{i=1}^d \langle x^{\beta_i}, h_{\alpha_i}\rangle = 0\,.$$

On the other hand, if $\alpha = \beta$, then

$$\langle x^\alpha, h_\alpha\rangle = \prod_{i=1}^d \langle x^{\alpha_i}, h_{\alpha_i}\rangle = \prod_{i=1}^d \sqrt{\alpha_i!} = \sqrt{\alpha!}\,.$$

The order-$m$ tensor $A$ can be identified with a multilinear—thus polynomial—map from $\mathbb{R}^d$ to $\mathbb{R}$:

$$A(x_1, \dots, x_d) = \sum_{i_1, \dots, i_m} A_{i_1 \dots i_m} x_{i_1} \dots x_{i_m} = \sum_{\alpha \in \mathbb{N}^d : |\alpha| = m} \frac{m!}{\alpha!} A_\alpha x^\alpha\,,$$

where in the second equality, we used the fact that $A$ is symmetric. Together with (B.2), it yields that for any $|\alpha| = m$,

$$\langle A, h_\alpha\rangle = \frac{m!}{\sqrt{\alpha!}} A_\alpha\,.$$

Moreover, since $\|A\| = 1$, we also have

$$\sum_{\alpha \in \mathbb{N}^d : |\alpha| = m} \frac{m!}{\alpha!} A_\alpha^2 = 1\,,$$

so that

$$\sum_{\alpha \in \mathbb{N}^d : |\alpha| = m} \langle A, h_\alpha \rangle^2 = m!\,.$$

Therefore, using Plancherel's formula, we get that for any $m \geq 0$,

$$\mathbb{E}_\xi[\langle A, \xi^{\otimes m} \rangle^2] = \langle A, A \rangle = \sum_{\alpha \in \mathbb{N}^d} \langle A, h_\alpha \rangle^2 \geq \sum_{\alpha \in \mathbb{N}^d : |\alpha| = m} \langle A, h_\alpha \rangle^2 = m!\,,$$

as claimed.

To prove the claimed lower bound on $\mathbb{E}_\xi |\langle A, \xi^{\otimes m} \rangle|$, we employ the following Khinchine-type inequality:

THEOREM B.3 ([Bob00, Theorem 2]). *Let $f = f(x_1, \ldots, x_m)$ be a degree $m$ polynomial and let $X \sim \mathcal{N}(0, I_m)$ be a standard Gaussian vector in $\mathbb{R}^m$. Then there exists a universal constant $c$ such that*

$$\sqrt{\mathbb{E}[f(X)^2]} \leq c^m \mathbb{E}[|f(X_1, \ldots, X_m)|]\,.$$

Applying Theorem B.3 to (B.1) yields the claim. □

LEMMA B.4. *Let $H(\zeta)$ and $H_n(\zeta)$ be the Hessians of $D(\phi)$ and $D_n(\phi)$, respectively, evaluated at $\phi = \zeta$. If $\mathcal{B}_\varepsilon := \{\phi \in \mathbb{R}^d : \rho(\phi, \theta) \leq \varepsilon\}$, then*

$$\mathbb{E} \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{op}^2 \leq C \frac{\log n}{n \sigma^4}\,.$$

PROOF. The matrix $H_n(\phi)$ can be written as a sum of independent random matrices:

$$H_n(\phi) = \frac{1}{n} \sum_{i=1}^n H_i(\phi)\,, \qquad H_i(\phi) = \nabla_\phi^2 \log \frac{p_\theta}{p_\phi}(Y_i)\,.$$

Using symmetrization, we get

$$\mathbb{E} \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}^2 \leq \frac{4}{n^2} \mathbb{E} \sup_{\phi \in \mathcal{B}_\varepsilon} \Big\| \sum_{i=1}^n \varepsilon_i H_i(\phi) \Big\|_{\mathrm{op}}^2\,, \tag{B.3}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d Rademacher random variables that are independent of the observations $Y_1, \ldots, Y_1$ and $\mathbb{E} H_n(\phi) = H(\phi)$.

Using a third order derivative calculation, we find that $\phi \mapsto H_i(\phi)$ satisfies the following Lipschitz property. For any $u, \phi, \eta \in \mathbb{R}^d$, such that $\|u\| = 1$ and $\phi, \eta \in \mathcal{B}_\varepsilon$, we have

$$|u^\top H_i(\phi) u - u^\top H_i(\eta) u| \leq C \frac{1 + |\xi_i|^3}{\sigma^3} \|\phi - \eta\|$$

where $\xi_i$ is the noise in (2.1).

Fix $\gamma \in (0, \varepsilon)$ and le $\mathcal{Z}$ be a $\gamma$-net of $\mathcal{B}_\varepsilon$. In other words, we require that

$$\max_{\phi \in \mathcal{B}_\varepsilon} \min_{\eta \in \mathcal{Z}} \|\eta - \phi\| \le \gamma \,.$$

We can always choose $\mathcal{Z}$ to have cardinality $|\mathcal{Z}| \le (C\varepsilon/\gamma)^d$ for some universal constant $C > 0$. Then, by Young's inequality, we get

$$\sup_{\phi \in \mathcal{B}_\varepsilon} \Big\| \sum_{i=1}^n \varepsilon_i H_i(\phi) \Big\|_{\mathrm{op}}^2 \le C \frac{\gamma^2}{\sigma^6} \Big( \sum_{i=1}^n |\xi_i|^3 \Big)^2 + C \max_{\phi \in \mathcal{Z}} \Big\| \sum_{i=1}^n \varepsilon_i H_i(\phi) \Big\|_{\mathrm{op}}^2 \tag{B.4}$$

The expectation of the first term is controlled using the fact that

$$\mathbb{E} \Big( \sum_{i=1}^n |\xi_i|^3 \Big)^2 = \sum_{i,j=1}^n \mathbb{E}[|\xi_i \xi_j|^3] \le C n^2 \,. \tag{B.5}$$

For second term, we employ a standard matrix concentration bound [Tro15, Theorem 4.6.1] to get that

$$\mathbb{E}\big[ \exp\big( t \| \sum_{i=1}^n \varepsilon_i H_i(\phi) \|_{\mathrm{op}}^2 \big) \big| Y_1, \dots, Y_n \big] \le d \exp\big( \frac{t^2}{2} \| \sum_{i=1}^n H_i(\phi)^2 \|_{\mathrm{op}} \big) \,.$$

Using standard arguments (see, e.g., [BLM13]), this implies

$$\begin{aligned} \mathbb{E}\big[ \max_{\phi \in \mathcal{Z}} \Big\| \sum_{i=1}^n \varepsilon_i H_i(\phi) \Big\|_{\mathrm{op}}^2 \big] &\le C(\log |\mathcal{Z}|) \mathbb{E}\big[ \max_{\phi \in \mathcal{Z}} \Big\| \sum_{i=1}^n H_i(\phi)^2 \Big\|_{\mathrm{op}} \big] \\ &\le C(\log |\mathcal{Z}|) n \mathbb{E}\big[ \max_{\phi \in \mathcal{Z}} \|H_1(\phi)^2\|_{\mathrm{op}} \big] \end{aligned}$$

As before, computing a second order derivative, we can show that

$$\|H_1(\phi)^2\|_{\mathrm{op}} \le \|H_1(\phi)\|_{\mathrm{op}}^2 \le C(1 + |\xi|^4)/\sigma^4 \,.$$

The above two displays yield

$$\mathbb{E}\big[ \max_{\phi \in \mathcal{Z}} \Big\| \sum_{i=1}^n \varepsilon_i H_i(\phi) \Big\|_{\mathrm{op}}^2 \big] \le C \frac{\log(\varepsilon/\gamma)}{\sigma^4} n$$

Combining the last display with (B.3), (B.4), and (B.5), we get

$$\mathbb{E} \sup_{\phi \in \mathcal{B}_\varepsilon} \|H(\phi) - H_n(\phi)\|_{\mathrm{op}}^2 \le C \Big( \frac{\gamma^2}{\sigma^6} + \frac{\log(\varepsilon/\gamma)}{n\sigma^4} \Big) \le C \frac{\log n}{n\sigma^4} \,, \tag{B.6}$$

for $\gamma = n^{-1/2}$. □

LEMMA B.5. *Assume the conditions of Theorem 4 hold. Then the MLE $\tilde{\theta}_n$ satisfies*

$$\mathbb{E}[\rho(\tilde{\theta}, \theta)^2] \le C \frac{\sigma^{(8k-4) \wedge (2\ell+2)}}{n} \,.$$

PROOF. As in the proof of Theorem 4, since $\theta$ is fixed, we simply write $D(\phi) = D(\theta \,\|\, \phi)$ and define

$$D_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta}{f_\phi}(Y_i)\,,$$

where $Y_i$ are i.i.d from model (2.1) and we recall that $f_\zeta$ is the density of $\mathrm{P}_\zeta, \zeta \in \mathbb{R}^d$.

We first establish using Lemma B.7 that the process $\{G_n(\phi)\}_{\phi\in\mathbb{R}^d}$ defined by $G_n(\phi) = D(\phi) - D_n(\phi)$ is a subgaussian process with respect to the Euclidean distance with variance proxy $c\sigma^2/n$ for some constant $c > 0$, i.e., that for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\lambda(G_n(\phi) - G_n(\zeta)))] \le \exp\big(c\frac{\lambda^2\sigma^2}{n}\|\phi - \zeta\|^2\big)\,.$$

We then apply the following standard tail bound.

PROPOSITION B.6 ([Ver17, Theorem 8.5.4]). *If $\{X_\phi\}_\phi$ is a (standard) subgaussian process on $\mathbb{R}^d$ with respect to the Euclidean metric and $B_\delta(\theta)$ is a ball of radius $\delta$ around $\theta$, then*

$$\mathbb{P}[\sup_{\phi\in B_\delta(\theta)} (X_\phi - X_\theta) \ge C\delta + x] \le Ce^{-Cx^2/\delta^2}\,.$$

The rescaled process $\sigma\sqrt{n}G_n$ is standard subgaussian process with respect to the Euclidean metric, so applying Proposition B.6 and noting that $D_n(\theta) = 0$ yields

$$\mathbb{P}\left[\sup_{\phi\in B_\delta(\theta)} G_n(\phi) \ge C\frac{\delta}{\sigma\sqrt{n}} + x\right] \le C\exp\big(-C\frac{n\sigma^2x^2}{\delta^2}\big)\,. \tag{B.7}$$

For convenience, write $v_n = \sqrt{n}(\tilde\theta - \theta)$, where $\tilde\theta$ is a MLE satisfying $\|\tilde\theta - \theta\| = \rho(\tilde\theta, \theta)$. We wish to show that $\mathbb{E}[\|v_n\|^2]$ is bounded by a constant that depends on $\sigma$ but not on $n$.

Define $E$ to be the compact subset of $\mathbb{R}^d$ defined by

$$E = \big\{\phi \in \mathbb{R}^d \,:\, \varepsilon_0 \le \rho(\theta, \phi) \le c_0\sigma\big\}\,,$$

where $c_0$ and $\varepsilon_0$ are defined in Theorem 4. In particular, for any $\phi \notin E$, it holds $D(\phi) \ge C\sigma^{-2k}\rho(\theta,\phi)^2$.

We employ the so-called *slicing* (a.k.a *peeling*) method. Define the sequence $\{\alpha_j\}_{j\ge0}$ where $\alpha_0 = 0$ and $\alpha_j = C\sigma^{j(2k-1)}$ for $j \ge 1$ for some large enough constant $C > 0$. For any $j \ge 0$, define $S_j = \{\phi \in \mathbb{R}^d \,:\, \alpha_j \le \sqrt{n}\rho(\phi,\theta) \le \alpha_{j+1}\} \setminus E$ and observe that

$$\begin{aligned}\mathbb{E}[\|v_n\|^2] &= \mathbb{E}[\|v_n\|^2 \mid \tilde\theta \in E]\mathbb{P}[\tilde\theta \in E] + \sum_{j\ge0}\mathbb{E}[\|v_n\|^2 \mid \tilde\theta \in S_j]\mathbb{P}[\tilde\theta \in S_j] \\ &\le c_0^2\sigma^2\mathbb{P}[\tilde\theta \in E] + C\sigma^{8k-4} + \sum_{j\ge2}\alpha_{j+1}^2\mathbb{P}[\tilde\theta \in S_j]\,. \end{aligned} \tag{B.8}$$

We now show that if $\tilde\theta \in S_j, j \ge 2$, then $G_n(\tilde\theta) = D(\tilde\theta) - D_n(\tilde\theta)$ is large. To that end, observe that on the one hand, by definition of the MLE, we have

$D_n(\tilde\theta) \le D_n(\theta) = 0$. On the other hand, $D(\tilde\theta) \ge C\sigma^{-2k}\rho(\tilde\theta,\theta)^2$. Hence, if $\tilde\theta \in S_j$, then $G_n(\tilde\theta) \ge C\sigma^{-2k}\rho(\tilde\theta,\theta)^2 \ge C\sigma^{-2k}\alpha_j^2/n$. It yields

$$\mathbb{P}[\tilde\theta \in S_j] \le \mathbb{P}\big[\sup_{\phi\in S_j} G_n(\phi) \ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\big] \le \mathbb{P}\big[\sup_{\phi\in B_{\frac{\alpha_{j+1}}{\sqrt{n}}}(\theta)} G_n(\phi) \ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\big]\,.$$

Recall $\alpha_j = C\sigma^{j(2k-1)}$ so that $\sigma^{-2k}\alpha_j^2 \ge C\alpha_{j+1}/\sigma, j \ge 2$ and apply (B.7) with $\delta = \alpha_{j+1}/\sqrt{n}$ and $x = C\sigma^{-2k}\alpha_j^2/n$ to get

$$\begin{aligned}\mathbb{P}\big[\sup_{\phi\in B_{\frac{\alpha_{j+1}}{\sqrt{n}}}(\theta)} G_n(\phi) \ge C\sigma^{-2k}\frac{\alpha_j^2}{n}\big] &\le C\exp\big(-C\frac{\alpha_j^4}{\alpha_{j+1}^2\sigma^{4k-2}}\big)\\ &\le C\exp\big(-C\sigma^{2j(2k-1)}\big)\,.\end{aligned}$$

It yields

$$\sum_{j\ge 2}\alpha_{j+1}^2\mathbb{P}[\tilde\theta\in S_j] \le C\sum_{j\ge 2}\sigma^{4j(2k-1)}\exp\big(-C\sigma^{2j(2k-1)}\big) \le C\,. \tag{B.9}$$

When $\tilde\theta \in E$, we use (5.2) to conclude that if $n \ge \sigma^{2\ell}$ then

$$\mathbb{P}[\tilde\theta\in E] \le \mathbb{P}\big[\sup_{\phi\in E} G_n(\phi) \ge C\sigma^{-\ell}\big] \le C\exp\big(-Cn\sigma^{-2\ell}\big)\,,$$

where in the last inequality, we used (B.7). Together with (B.8) and (B.9), it implies the desired result. □

LEMMA B.7. *The process $\{G_n(\phi)\}_{\phi\in\mathbb{R}^d}$ defined by $G_n(\phi) = D(\phi) - D_n(\phi)$ is a subgaussian process with respect to the $\ell_2$ distance on $\mathbb{R}^d$ with variance proxy $c/(n\sigma^2)$ for some constant $c > 0$, i.e., that for any $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}[\exp(\lambda(G_n(\phi) - G_n(\zeta)))] \le \exp\big(\lambda^2\frac{c}{n\sigma^2}\|\phi-\zeta\|^2\big)\,.$$

PROOF. By definition of $G_n$ and the densities $f_\zeta$ and $f_\phi$, we have

$$\begin{aligned}G_n(\phi) - G_n(\zeta) &= D(\phi) - D(\zeta) - D_n(\phi) + D_n(\zeta)\\ &= \mathbb{E}[\log f_\zeta(Y) - \log f_\phi(Y)] - \frac{1}{n}\sum_{i=1}^n[\log f_\zeta(Y_i) - \log f_\phi(Y_i)]\\ &= \mathbb{E}\Big[\log\frac{\mathbb{E}_R\exp(-\frac{\|Y-R\zeta\|^2}{2\sigma^2})}{\mathbb{E}_R\exp(-\frac{\|Y-R\phi\|^2}{2\sigma^2})}\Big] - \frac{1}{n}\sum_{i=1}^n\log\frac{\mathbb{E}_R\exp(-\frac{\|Y_i-R\zeta\|^2}{2\sigma^2})}{\mathbb{E}_R\exp(-\frac{\|Y_i-R\phi\|^2}{2\sigma^2})}\\ &= \mathbb{E}\Big[\log\frac{\mathbb{E}_R\exp(Y^\top R\zeta/\sigma^2)}{\mathbb{E}_R\exp(Y^\top R\phi/\sigma^2)}\Big] - \frac{1}{n}\sum_{i=1}^n\log\frac{\mathbb{E}_R\exp(Y_i^\top R\zeta/\sigma^2)}{\mathbb{E}_R\exp(Y_i^\top R\phi/\sigma^2)}\\ &= \mathbb{E}[\Delta(Y)] - \frac{1}{n}\sum_{i=1}^n\Delta(Y_i)\,,\end{aligned}$$

where

$$\Delta(Y) = \log\frac{\mathbb{E}_R\exp(Y^\top R\zeta/\sigma^2)}{\mathbb{E}_R\exp(Y^\top R\phi/\sigma^2)}\,.$$

Next, using a standard symmetrization argument, we get

$$\mathbb{E}[\exp(\lambda(G_n(\phi) - G_n(\zeta)))] \le \prod_{i=1}^{n} \mathbb{E}[\exp(\frac{2\lambda}{n}\varepsilon_i \Delta(Y_i))]\,, \tag{B.10}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d Rademacher random variables that are independent of the sample $Y_1, \dots, Y_n$. Next, observe that using the Cauchy-Schwarz inequality, we get that the function $\zeta \mapsto \log \mathbb{E}_R \exp(Y_i^\top R\zeta/\sigma^2)$ is $\|Y_i\|/\sigma^2$-Lipschitz with respect to the Euclidean distance. Hence $|\Delta(Y_i)| \le \|Y_i\|\|\phi - \zeta\|/\sigma^2$. Moreover, it follows from (2.1) that conditionally on $R_{\ell_i}$, the random variable $\|Y_i\|^2/\sigma^2 \sim \chi^2_d(1/\sigma^2)$ follows a noncentral $\chi^2$ distribution with $d$ degrees of freedom and noncentrality parameter $\|\theta\|^2/\sigma^2 = 1/\sigma^2 \le 1$. In particular $\|Y_i\|^2/\sigma^2$ has a finite moment generating function in the neighborhood of the origin, and we obtain for some constant $c$ small enough that

$$\mathbb{E}\big[\exp\big(c\frac{\sigma^2(\varepsilon_i\Delta(Y_i))^2}{\|\phi-\zeta\|^2}\big)\big] \le \mathbb{E}\big[\exp\big(c\frac{\|Y_i\|^2}{\sigma^2}\big)\big] \le e\,.$$

It implies (see, e.g., [Ver17] proposition 2.5.2) that

$$\mathbb{E}\big[\exp(\frac{2\lambda}{n}\varepsilon_i\Delta(Y_i))\big] \le \exp(\lambda^2 \frac{c}{n^2\sigma^2}\|\phi-\zeta\|^2)$$

for all $\lambda \in \mathbb{R}$. Together with (B.10), this yields the desired result. □

LEMMA B.8. *Fix $\theta \in \mathbb{R}^d$. Assume that $n/(\log n) \ge C\sigma^4$. For any $j = -\lfloor d/2\rfloor, \dots, \lfloor d/2\rfloor$, define,*

$$M_j = \frac{1}{n}\sum_{i=1}^{n} |\widehat{(Y_i)}_j|^2 - \sigma^2\,.$$

*Define the set $\tilde{S}$ by*

$$\tilde{S} = \Big\{ j \in \{-\lfloor d/2\rfloor, \dots, \lfloor d/2\rfloor\} \,:\, M_j \ge C\sigma^2\sqrt{\frac{\log n}{n}}\Big\}$$

*Then*

$$\mathbb{P}[\tilde{S} \ne \operatorname{psupp}(\hat{\theta})] \le C\frac{\sigma^4}{n}\,.$$

PROOF. It is straightforward to check that $\mathbb{E}[M_j] = |\hat{\theta}_k|^2$. To calculate the variance, we note that

$$|\widehat{(Y_i)}_k|^2 - |\hat{\theta}_k|^2 - \sigma^2 = 2\sigma\Re(\widehat{(R_i\theta)}_k\widehat{(\xi_i)}_k) + \sigma^2(|\widehat{(\xi_i)}_k|^2 - 1)\,.$$

Hence $\operatorname{var}[|\widehat{(Y_i)}_k|^2 - \sigma^2] \le C\sigma^4$, and $\operatorname{var}[M_j] \le C\sigma^4/n$.

Moreover, the random variable $|\widehat{(Y_i)}_k|^2 - \sigma^2$ is a $\sigma^2$-subexponential random variable. Indeed,

$$|\widehat{(Y_i)}_k|^2 - |\hat{\theta}_k|^2 - \sigma^2 \le 2\sigma|\widehat{(\xi_i)}_k| + \sigma^2(|\widehat{(\xi_i)}_k|^2 - 1) \le 2\sigma^2(|\widehat{(\xi_i)}_k|^2 - 1) + \sigma^2 + 1\,.$$

The variable $\widehat{(\xi_i)}_0$ is a standard Gaussian, and for $k \neq 0$ the variable $\widehat{(\xi_i)}_k$ is a standard complex Gaussian, hence $|\widehat{(\xi_i)}_k|^2$ is a rescaled $\chi^2$ random variable with at most 2 degrees of freedom. Hence $|\widehat{(\xi_i)}_k|^2$ is sub-exponential with constant variance proxy, so that $|\widehat{(Y_i)}_k|^2 - \sigma^2$ is $\sigma^2$-subexponential.

If $j \in \mathrm{psupp}(\hat{\theta})$, then by Chebyshev's inequality

$$\mathbb{P}[j \notin \tilde{S}] = \mathbb{P}\big[M_j \leq C\sigma^2\sqrt{\frac{\log n}{n}}\big] \leq C\frac{\sigma^4}{n\big(|\hat{\theta}_j|^2 - C\sigma^2\sqrt{\frac{\log n}{n}}\big)^2} \leq C\frac{\sigma^4}{n}\,.$$

On the other hand, if $j \notin \mathrm{psupp}(\hat{\theta})$, then $\mathbb{E}[M_j] = 0$ and by Bernstein's inequality

$$\mathbb{P}[j \in \tilde{S}] = \mathbb{P}[M_j \geq C\sigma^2\sqrt{\frac{\log n}{n}}] \leq \frac{C}{n}\,.$$

The proof follows using a union bound. □

## REFERENCES


[ABBS14] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *Network Science and Engineering, IEEE Transactions on*, 1(1):10–22, Jan 2014.

[ADBS16] Cecilia Aguerrebere, Mauricio Delbracio, Alberto Bartesaghi, and Guillermo Sapiro. Fundamental limits in multi-image alignment. *Available online at arXiv:1602.01541 [cs.CV]*, 2016.

[APS17] Emmanuel Abbe, Joao Pereira, and Amit Singer. Sample complexity of the boolean multireference alignment problem. *Available online at arXiv:1701.07540 [cs.IT]*, 2017.

[BBS16] Afonso S. Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, pages 1–23, 2016.

[BBV16] A. S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *COLT*, 2016.

[BCS15] A. S. Bandeira, Y. Chen, and A. Singer. Non-unique games over compact groups and orientation estimation in cryo-em. *Available online at arXiv:1505.03840 [cs.CV]*, 2015.

[BCSZ14] Afonso S. Bandeira, Moses Charikar, Amit Singer, and Andy Zhu. Multireference alignment using semidefinite programming. In *ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science*, pages 459–470. ACM, New York, 2014.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, with a foreword by Michel Ledoux.

[Bob00] S. G. Bobkov. Remarks on the growth of $L^p$-norms of polynomials. In *Geometric aspects of functional analysis*, volume 1745 of *Lecture Notes in Math.*, pages 27–35. Springer, Berlin, 2000.

[Bou16] Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal of Optimization, to appear*, 2016.

[Bro92] Lisa Gottesfeld Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376, 1992.

[BSS11] A. S. Bandeira, A. Singer, and D. Spielman. The $so(3)$ cheeger inequality. *Unpublished Draft*, 2011.

[CC16] Yuxin Chen and Emmanuel Candes. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Available online at arXiv:1609.05820 [cs.IT]*, 2016.

[CL11] T. Tony Cai and Mark G. Low. Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.*, 39(2):1012–1041, 04 2011.

[Dia92] R. Diamond. On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Science*, 1(10):1279–1287, October 1992.

[Edi15] Editorial. Method of the year 2015. *Nature Methods*, 13:1, 2015.

[FZB02] H. Foroosh, J. B. Zerubia, and M. Berthod. Extension of phase correlation to subpixel registration. *IEEE Transactions on Image Processing*, 11(3):188–200, 2002.

[GBM16] Tingran Gao, Jacek Brodzki, and Sayan Mukherjee. The geometry of synchronization problems and learning group actions. *Available online at arXiv:1610.09051 [math.ST]*, 2016.

[HK15] Philippe Heinrich and Jonas Kahn. Optimal rates for finite mixture estimation. *arXiv:1507.04313*, 2015.

[HUL01] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

[JMRT16] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 2016.

[KI93] R. Kakarala and G. J. Iverson. Uniqueness of results for multiple correlations of periodic functions. *Oct. Soc. Am. A*, 10:1517–1528, 1993.

[LeC73] L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

[LNS99] O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the l r norm of a regression function. *Probability Theory and Related Fields*, 113(2):221–253, 1999.

[MP00] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[MV10] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. *Arxiv:1004.4223v1*, 2010.

[Nog15] Eva Nogales. The development of cryo-em into a mainstream structural biology technique. *Nature Methods*, 13:24–27, 2015.

[Pea94] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 185:71–110, 1894.

[PWB+17] A. Perry, J. Weed, A. S. Bandeira, P. Rigollet, and A. Singer. The sample complexity of multi-reference alignment. Manuscript, 2017.

[PWBM16] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra. Message-passing algorithms for synchronization problems over compact groups. arXiv:1610.04583, 2016.

[Sad89] B. M. Sadler. Shift and rotation invariant object reconstruction using the bispectrum. In *Workshop on Higher-Order Spectral Analysis*, pages 106–111, Jun 1989.

[Sin11] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.

[SSK13] B. Sonday, A. Singer, and I. G. Kevrekidis. Noisy dynamic simulations in the presence of symmetry: Data alignment and model reduction. *Computers & Mathematics with Applications*, 65(10):1535 – 1557, 2013.

[SVN+05] Sjors H.W. Scheres, Mikel Valle, Rafael Nuñez, Carlos O.S. Sorzano, Roberto Marabini, Gabor T. Herman, and Jose-Maria Carazo. Maximum-likelihood multi-reference refinement for electron microscopy images. *Journal of Molecular Biology*, 348(1):139 – 149, 2005.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[TS12] D. L. Theobald and P. A. Steindel. Optimal simultaneous superpositioning of multiple structures with missing data. *Bioinformatics*, 28(15):1972–1979, 2012.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[Ver17] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press (to appear), 2017.

[Wal49] Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics*, 20:595–601, 1949.

[WW84] E. Weinstein and A. J. Weiss. Fundamental limitations in passive timedelay estimationpart ii: Wide-band systems. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(5):1064–1078, 1984.

[WY16] Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, June 2016.

[ZvdHGG03] 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. *Radar, Sonar and Navigation, IEE Proceedings*, 150(6):411–418, 2003.

Afonso S. Bandeira
Department of Mathematics
Courant Institute of Mathematical Sciences
Center for Data Science
New York Univeristy,
New York, NY 10012, USA
(bandeira@cims.nyu.edu)

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(rigollet@math.mit.edu)

Jonathan Weed
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Avenue,
Cambridge, MA 02139-4307, USA
(jweed@mit.edu)