Minimax Risk Upper Bounds Based on Shell Analysis of a Quantized Maximum Likelihood Estimator

Noam Gavish and Or Ordentlich
School of Computer Science and Engineering, The Hebrew University of Jerusalem
Jerusalem, Israel
{noam.gavish, or.ordentlich}@mail.huji.ac.il

***Abstract*—This paper develops a unified framework for upper bounding the minimax risk in high-dimensional parameter estimation problems. To this end, we study a quantized maximum likelihood estimator, where the estimator computes the likelihood for all points within a discrete cover, and outputs the candidate with the maximal likelihood. While this concept is straightforward, our analysis is quite delicate. It splits the competing candidates in the cover to small shells, and controls the number of candidates in each shell, as well as the probability that a candidate in the shell outscores a candidate which is close to the true parameter. We demonstrate the utility of our bounds by applying them to different Gaussian problems, and showing that they recover the optimal minimax rate for the Gaussian location model and the spiked Wigner Model. For the multi-reference alignment problem we obtain a novel minimax upper bound, which essentially places no assumptions on the signal of interest.**

I. Introduction

We consider the problem of estimating a parameter $\theta \in \Theta \subset \mathcal{A}^d$ from the measurement $Y^n \sim P^{\theta}_{Y^n}$, under minimax risk with respect to some loss $\ell(\theta, \hat{\theta})$. Many works derived minimax upper bounds using information-theoretic techniques, see e.g. [1]–[3], as well as [4] and references within, for a very partial list of examples. Our goal here, however, is to develop a unified information-theoretic framework for *upper bounding* the minimax risk in various high-dimensional problems where $d \to \infty$.

A common approach for computing such bounds, is analyzing the performance of the maximum likelihood estimator (MLE). While such analysis is classical for models consisting of i.i.d. distributions dictated by a parameter vector of small dimension [5], it often becomes too complex to handle for high-dimensional models of interest [6]. A natural way to overcome the difficulty involved with analyzing the result of a maximization over a (possibly continuous) high-dimensional set of parameters, is to use discretization. Under this paradigm, one fixes some discrete set $\mathcal{C}$ which covers Θ with some predefined resolution τ, computes for each point in the set some score for compatibility with the measurements, and chooses the point with the highest score as the estimator. Examples for works following this approach include that of Le Cam [7] and Birgé [8], and to some extent also that of Yatracos [9] and of Yang and Barron [10]. See also [4, Chapter 32].

When the compatibility score is the likelihood, we refer to the obtained estimator as the *quantized MLE*. The probability $\Pr(\ell(\theta, \theta_f) \geq \delta)$ of forming a "bad" estimate with loss exceeding $\delta > \tau > 0$ is controlled by the *mismatched binary hypothesis testing error probability* of a "bad" point θ_f with $\ell(\theta, \theta_f) \geq \delta$ obtaining a higher likelihood than that of "good" point with $\ell(\theta, \theta_n) \leq \tau < \delta$, times the number of candidates $|\mathcal{C}|$. One shortcoming of this bounding approach, however, is that it fails to exploit the fact that in many problems this mismatched binary hypothesis testing error probability decreases with δ, while the number of competing points $|\{\theta_f \in \mathcal{C} \ : \ \ell(\theta, \theta_f) \approx \delta\}|$ incurring loss δ is increasing in δ, but is typically much smaller than $|\mathcal{C}|$.

To tackle this issue, in this paper we perform a *shell analysis* for the quantized MLE. The contribution of a shell of radius δ' to the error probability is the probability that a point in this distance beats a point with distance τ, times the number of points in the shell. We provide a general upper bound on the estimation error, which depends on the mismatched binary hypothesis testing error probability for the observation model and on the "uniform covering density" of Θ with respect to the loss ℓ. We then show that when $\Theta \subset \mathbb{R}^d$ and $\ell(\theta, \hat{\theta})$ is some (arbitrary) norm, we can form a τ-cover with all shells having size $d \log \frac{\delta'}{\tau} + o(d)$ simultaneously. With this, we are able to successfully apply our bound to various high-dimensional Gaussian problems under quadratic loss, with $\Theta \subset \mathbb{R}^d$ having an essentially unbounded diameter. To illustrate the power and versatility of our bounding approach, we apply it to recover the optimal known minimax rates for the Gaussian location model and the spiked Wigner model. We further apply our bound to the multi-reference alignment (MRA) problem and obtain a novel minimax upper bound, which is essentially assumptions-free. Our bound shows that as long as the number of measurements is $O(d/\log d)$ the minimax risk of MRA is equal to that of the Gaussian location model, up to constants.

It should be noted that while the developed bounds may fall short at achieving the precise constants in

the minimax risk characterization, and tailor-made estimation algorithms often result in sharper bounds, our framework provides a general and relatively simple recipe for computing minimax upper bounds. The main advantage of our bounds is that they are not limited to i.i.d. models $\mathcal{P} = \{P_\theta^{\otimes n} : \theta \in \Theta\}$, and under some assumptions (which hold for many high-dimensional estimation problems of interest) they take a particularly simple form for product models $\mathcal{P} = \{\prod_{i=1}^n P_{g_i(\theta)} : \theta \in \Theta\}$ where $g_1, \ldots, g_n$ are some functions. The three Gaussian examples we consider here can be cast as product models. In the full version of this paper we extend the treatment to general product distributions, but here the details are omitted due to lack of space.

II. Problem Formulation and General Bounds

We begin with several definitions that will be needed for the problem formulation and for stating our results. We follow the notation of [4].

Let $\mathcal{A}$ be some arbitrary alphabet, and $\Theta \subset \mathcal{A}^d$ be some subset in $\mathcal{A}^d$, and consider a class of distributions

$$\mathcal{P}_\Theta \triangleq \left\{P_{Y^n}^\theta \ : \ \theta \in \Theta\right\}. \tag{1}$$

on the alphabet $\mathcal{Y}^n$. Let $\ell : \Theta \times \Theta \to \mathbb{R}_+$ be a loss function. The minimax estimation risk is defined as

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[\ell(\theta, \hat{\theta}(Y^n))\Big],$$

where the notation $\mathbb{E}_\theta$ designates that the expectation is taken with respect to $Y^n \sim P_{Y^n}^\theta$. Our focus is on developing minimax upper bounds on $\Pr(\ell(\theta, \hat{\theta}(Y^n)) > \delta)$. In particular, we will be interested in understanding for which values of δ there exists an estimator $\hat{\theta}$ for which the above probability is small, and this in turn, leads to upper bounds on the minimax risk.

To that end, we study and analyze a proxy for the maximum likelihood estimator (MLE), which we call the *quantized maximum likelihood estimator*. For any $\theta \in \Theta$ and $y^n \in \mathcal{Y}^n$ define the likelihood

$$L_\theta(y^n) \triangleq \frac{dP_{y^n}^\theta}{d\mu}(y^n),$$

where μ is some common dominating measure.

Definition II.1 (quantized MLE)**.** *For a class of distributions defined in* (1)*, a given discrete set $\mathcal{C} \subset \Theta$, and measurements $y^n \in \mathcal{Y}^n$, the quantized MLE outputs*

$$\hat{\theta}(y^n) = \operatorname*{argmax}_{\theta \in \mathcal{C}} L_\theta(y^n) \tag{2}$$

as the estimate.

We say that a discrete set $\mathcal{C}$ forms a τ-cover for Θ (with respect to ℓ) if

$$\forall \theta \in \Theta \ \exists \theta' \in \mathcal{C} \text{ such that } \ell(\theta, \theta') \le \tau.$$

Clearly, if $\mathcal{C}$ is not a τ-cover, the worst-case loss of a quantized MLE with $\mathcal{C}$ is at least τ, as in this case there is $\theta \in \Theta$ that is at least τ-far from any potential output of the quantized MLE. Thus, we would use $\mathcal{C}$ which form a τ-cover for τ smaller than our target risk. In addition, we would like $\mathcal{C}$ to have density as small as possible, such that there will not be many potential candidates far away from the true θ. To that end, we define the population function of the set $\mathcal{C}$ as

$$M_C(\delta) \triangleq \max_{\theta \in \Theta} |\mathcal{C} \cap \mathcal{B}(\theta, \delta, \ell)|$$

where

$$\mathcal{B}(\theta, \delta, \ell) \triangleq \{\theta' \in \Theta \ : \ \ell(\theta, \theta') < \delta\}.$$

We further define the diameter of Θ as

$$\operatorname{diam}(\Theta) = \min\{\delta \in \mathbb{R} \ : \ \forall \theta \in \Theta, \ \mathcal{B}(\theta, \delta, \ell) = \Theta\}$$

We would like to upper bound the probability that the estimation error $\ell(\theta, \hat{\theta}(Y^n))$ of the quantized MLE exceeds $\delta > \tau$. This can only happen if some θ_f with $\ell(\theta, \theta_f) \ge \delta$ received larger likelihood than θ_n with $\ell(\theta, \theta_n) \le \tau$ (such θ_n must exist since $\mathcal{C}$ is a τ-cover). To that end, we need to control the error probability of a mismatched binary hypothesis testing problem where we observe $Y^n \sim P_{Y^n}^\theta$ and need to decide between the $\mathcal{H}_0 = P_{Y^n}^{\theta_n}$ and $\mathcal{H}_1 = P_{Y^n}^{\theta_f}$. We define the function

$$E_{HT}(\tau, \delta) = \min_{\substack{\theta, \theta_n, \theta_f \in \Theta \\ \ell(\theta, \theta_n) \le \tau \\ \ell(\theta, \theta_f) \ge \delta}} -\log P_{Y^n}^\theta \left(\log \frac{dP_{Y^n}^{\theta_f}}{dP_{Y^n}^{\theta_n}}(Y^n) \ge 0\right)$$

The error exponent of mismatched binary hypothesis testing for i.i.d. distributions was studied in [11]. Here, we do not assume the distributions are i.i.d., and furthermore, our quantity of interest is the minimum exponent among all triplets $\theta, \theta_n, \theta_f \in \Theta$ satisfying the loss constraints.

Now, we are in a position to state and prove the most general form of our bound.

Theorem II.2. *Let $\mathcal{C}$ be a τ-cover for Θ, and let $\delta_1 < \delta_2 < \cdots < \delta_k$ be some real numbers, with $\delta_1 = \delta > \tau$ and $\delta_k \ge \operatorname{diam}(\Theta)$ (which implies $M_{\mathcal{C}}(\delta_k) = |\mathcal{C}|$). Assume that for all $i = 1, \ldots, k-1$ we have $\log M_{\mathcal{C}}(\delta_i) \le f(\tau, \delta_i)$ for some function $f : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, and that*

$$\min_{i=1,\ldots,k-1} E_{HT}(\tau, \delta_i) - f(\tau, \delta_{i+1}) - \log(k-1) \ge \gamma, \tag{3}$$

for some $\gamma > 0$. Then, for the quantized MLE, we have

$$P_{Y^n}^\theta \Big[\ell\Big(\theta, \hat{\theta}(Y^n)\Big) > \delta\Big] < \exp(-\gamma)$$

Proof. Since $\mathcal{C}$ forms a τ-cover of Θ, there must exist some θ_n close to θ in the sense that $\ell(\theta, \theta_n) \le \tau$. In order for a point θ_f satisfying $\ell(\theta, \theta_f) \ge \delta > \tau$ to be chosen as the estimator in (2), we must have that

$$L_{\theta_f}(Y^n) \ge L_{\theta_n}(Y^n) \iff \log \frac{dP_{Y^n}^{\theta_f}}{dP_{Y^n}^{\theta_n}}(Y^n) > 0.$$

Thus, the probability that $\ell(\theta, \hat{\theta}(Y^n)) > \delta$ is upper bounded by the probability that some $\theta_f \in \mathcal{C}$ has greater likelihood than θ_n. Let $\mathcal{S} = \mathcal{C} \setminus (\mathcal{C} \cap \mathcal{B}(\theta, \delta, \ell))$ be the set of candidates for the quantized MLE that incur $\ell(\theta, \hat{\theta}(Y^n)) \ge \delta$. We can write $\mathcal{S}$ as a disjoint

partition $\mathcal{S} = \cup_{i=1}^{k-1} \mathcal{S}_i$, of shells, where the ith shell is defined as

$$\mathcal{S}_i = \mathcal{C} \cap (\mathcal{B}(\theta, \delta_{i+1}, \ell) \setminus \mathcal{B}(\theta, \delta_i, \ell)).$$

Note that $|\mathcal{S}_i| \le |\mathcal{C} \cap (\mathcal{B}(\theta, \delta_{i+1}, \ell)| \le M_{\mathcal{C}}(\delta_{i+1})$. We have

$$\begin{aligned}
&P_{Y^n}^{\theta}\left[\ell\Big(\theta, \hat{\theta}(Y^n)\Big) > \delta\right] \\
&\le P_{Y^n}^{\theta}\left(\bigcup_{\theta_f \in \mathcal{S}} \left\{L_{\theta_f}(Y^n) \ge L_{\theta_n}(Y^n)\right\}\right) \\
&= P_{Y^n}^{\theta}\left(\bigcup_{i=1}^{k-1} \bigcup_{\theta_f \in \mathcal{S}_i} \left\{L_{\theta_f}(Y^n) \ge L_{\theta_n}(Y^n)\right\}\right) \\
&\le \sum_{i=1}^{k-1} \sum_{\theta_f \in \mathcal{S}_i} P_{Y^n}^{\theta}\big(L_{\theta_f}(Y^n) \ge L_{\theta_n}(Y^n)\big) \\
&= \sum_{i=1}^{k-1} \sum_{\theta_f \in \mathcal{S}_i} P_{Y^n}^{\theta}\left(\log \frac{dP_{Y^n}^{\theta_f}}{dP_{Y^n}^{\theta_n}}(Y^n) \ge 0\right) \\
&\le \sum_{i=1}^{k-1} \sum_{\theta_f \in \mathcal{S}_i} \exp(-E_{HT}(\tau, \delta_i)) \\
&\le \sum_{i=1}^{k-1} M_{\mathcal{C}}(\delta_{i+1}) \exp(-E_{HT}(\tau, \delta_i)) \\
&= \sum_{i=1}^{k-1} \exp(-(E_{HT}(\tau, \delta_i) - \log M_C(\delta_{i+1}))) \\
&\le \sum_{i=1}^{k-1} \exp(-(E_{HT}(\tau, \delta_i) - f(\tau, \delta_{i+1}))), \qquad (4)
\end{aligned}$$

and the statement immediately follows. □

Applying Theorem II.2 with $k = 2$, $\delta_1 = \delta$, $\delta_k = \mathrm{diam}(\Theta)$, we get the following corollary.

Corollary II.3. *Let $N(\Theta, \tau, \ell)$ be the τ-covering number of Θ with respect to ℓ (the size of the smallest τ-net), and assume $E_{HT}(\tau, \delta) > \log|N(\Theta, \tau, \ell)| + \gamma$. Then, there exists a $\mathcal{C}$, such that the quantized MLE attains*

$$P_{Y^n}^{\theta}\left[\ell\Big(\theta, \hat{\theta}(Y^n)\Big) > \delta\right] < \exp(-\gamma).$$

The simplified approach, of using a single shell, is limited to the the bound in Corollary II.3, which becomes quite useless when $\log|N(\Theta, \tau, \ell)|$ is too large. On the other hand, the use of $k-1$ shells in the derivation of Theorem II.2 allows us to exploit the fact that in many models the probability that a candidate θ_f obtains high likelihood decreases with $\ell(\theta, \theta_f)$. This results in Theorem II.2 being useful even for cases where the parameter space Θ has a large covering number. However, as opposed to Corollary II.3 which only requires an upper bound on the covering number, in order to use Theorem II.2 we must find a function $f(\delta, \tau)$ that upper bounds $\log M_{\mathcal{C}}(\delta_i)$ for some τ-cover $\mathcal{C}$, and $i = 1, \ldots, k-1$.

A. Results for high-dimensional Θ

It turns out that whenever $\Theta \subset \mathbb{R}^d$, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some arbitrary norm on $\mathbb{R}^d$, we can obtain simple closed-form expressions.

Lemma II.4. *Let $\Theta \subset \mathbb{R}^d$, and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some arbitrary norm on $\mathbb{R}^d$. Let $k = o(d^2)$ be an integer and let $\delta_1 < \delta_2 < \cdots < \delta_k$ be real numbers with $\delta_1 = \delta > \tau$ and $\delta_k \ge \mathrm{diam}(\Theta)$. There exists a τ-cover $\mathcal{C}$ of Θ satisfying*

$$\log M_{\mathcal{C}}(\delta_i) \le d \log\left(\frac{\delta_i}{\tau}\right) + o(d), \quad \forall i = 1, \ldots, k.$$

Sketch of proof. Any norm $\|\cdot\|$ on $\mathbb{R}^d$ can be expressed as $\|x\| = \inf\{r \ : \ x \in r\mathcal{K}\}$ for some convex symmetric body $\mathcal{K}$ in $\mathbb{R}^d$. Thus, if $\Theta \subset \mathcal{C} + \tau\mathcal{K}$, then $\mathcal{C}$ forms a τ-cover for Θ with respect to $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$. In [12], Erdős and Rogers specify a distribution for an infinite constellation $L \subset \mathbb{R}^d$ with unit density (meaning that $\lim_{r\to\infty} |L \cap r\mathcal{K}|/\mathrm{Vol}(r\mathcal{K}) = 1$ for any convex body $\mathcal{K} \subset \mathbb{R}^d$) with the property that for any convex body $\mathcal{A} \subset \mathbb{R}^d$ with $\mathrm{Vol}(\mathcal{A}) = \Omega(\mathrm{poly}(d))$, we have that with probability $1 - o(d^{-2})$: i)any point in $\mathbb{R}^d$ is covered by $L + \mathcal{A}$; ii)for any $x \in \mathbb{R}^d$ it holds that $|(x + \mathcal{A}) \cap L| < e \cdot \mathrm{Vol}(\mathcal{A})$.[1]

Set $\mathcal{A} = \mathcal{A}_0 = \alpha\tau\mathcal{K}$, and $\mathcal{A}_i = \alpha\delta_i\mathcal{K} = \frac{\delta_i}{\tau}\mathcal{A}$ for $i = 1, \ldots, k-1$, where $\alpha > 0$ is chosen such that $\mathrm{Vol}(\mathcal{A}) = \mathrm{poly}(d)$. Drawing L from the distribution in [12], we get that with high probability $L + \alpha\tau\mathcal{K} = \mathbb{R}^n \supset \Theta$, and that for any $\theta \in \Theta$ it holds that

$$\begin{aligned}
&|L \cap \mathcal{B}(\theta, \alpha\delta_i, \|\cdot\|)| = |(\theta + \alpha\delta_i\mathcal{K}) \cap L| \\
&= |(\theta + \mathcal{A}_i) \cap L| \le e \cdot \mathrm{Vol}(\mathcal{A}_i) = e\left(\frac{\delta_i}{\tau}\right)^d \mathrm{poly}(d).
\end{aligned}$$

Now, scaling by $1/\alpha$ we get that $\tilde{L} = L/\alpha$ forms a τ-cover for $\mathbb{R}^d$ and satisfies $M_{\tilde{L}}(\delta_i) \le e\big(\frac{\delta_i}{\tau}\big)^d \mathrm{poly}(d)$. Taking logarithm establishes the claim. □

With Lemma II.4, we can significantly simplify Theorem II.2 for norm loss on $\mathbb{R}^d$.

Theorem II.5. *Let $\Theta \subset \mathbb{R}^d$ and $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|$ for some norm on $\mathbb{R}^d$. Let $\delta > \tau > 0$, and assume furthermore that $\mathrm{diam}(\Theta) = e^{o(d)}$ and that*

$$\min_{\delta \le \delta' \le \mathrm{diam}(\Theta)} \left[E_{HT}(\tau, \delta') - d \cdot \log\left(\frac{\delta'}{\tau}\right)\right] = \Omega(d) \qquad (5)$$

then

$$\lim_{d\to\infty} P_{Y^n}^{\theta}\left[\|\theta - \hat{\theta}(Y^n)\| > \delta\right] = 0$$

Sketch of proof. Set $k = 1 + \left\lceil d \log \frac{\mathrm{diam}(\Theta)}{\delta} \right\rceil$, and $\delta_i = e^{(i-1)/d}\delta$ for $i = 1, \ldots, k$. With this choice we have

[1] In a more recent work [13], a similar result is shown for the natural distribution on unit co-volume lattices $L \subset \mathbb{R}^d$.

that $\delta_k \geq \mathrm{diam}(\Theta)$. By Lemma II.4 we can find a τ-cover $\mathcal{C}$ such that for $i = 1, \ldots, k$ it holds that

$$\begin{aligned}\log M_{\mathcal{C}}(\delta_{i+1}) &\leq f(\delta_{i+1}, \tau) = d \log \frac{\delta_{i+1}}{\tau} + o(d) \\ &= d \log \frac{\delta_i}{\tau} + d \log \frac{\delta_{i+1}}{\delta_i} + o(d) \\ &= d \log \frac{\delta_i}{\tau} + o(d).\end{aligned}$$

Thus, we have

$$\begin{aligned}&\min_{i=1,\ldots,k-1} E_{HT}(\tau, \delta_i) - f(\tau, \delta_{i+1}) - \log(k-1) \\ &= \min_{i=1,\ldots,k-1} E_{HT}(\tau, \delta_i) - d \log \frac{\delta_i}{\tau} + o(d) \\ &\geq \min_{\delta \leq \delta' \leq \mathrm{diam}(\Theta)} E_{HT}(\tau, \delta') - d \log \frac{\delta'}{\tau} + o(d).\end{aligned}$$

Applying Theorem II.2, our claim follows. □

III. Gaussian Models

In this section we demonstrate how to apply our bounds for several problems. Here, we restrict attention to the setup where $\Theta = \{\theta \in \mathbb{R}^d \ : \ \|\theta\|_2 \leq r\}$ consists of all vectors inside an ℓ_2 ball of radius r in $\mathbb{R}^d$, and the loss is the squared loss $\ell(\theta, \hat{\theta}) = \|\theta - \theta'\|_2^2$. Throughout this section, we assume the dimension d goes to infinity. The number r here is quite arbitrary. We only need it to satisfy $r = \mathrm{diam}(\Theta) = e^{e^{o(d)}}$ in order for Theorem II.5 to hold, which is assumed in our derivations. In order to apply Theorem II.5 for a given class of distribution $\mathcal{P}_\Theta$, it remains to upper bound E_{HT}.

We begin by upper bounding E_{HT} for simple Gaussian models. We will later see that the analysis for the simple case will also be useful for tackling more complicated Gaussian models including the spiked Wigner model and multi-reference alignment.

Let $g : \Theta \to \mathbb{R}^n$ be some function, and assume $P^\theta_{Y^n} = \mathcal{N}(g(\theta), \sigma^2 I_n)$ for all $\theta \in \Theta$, where I_n is the $n \times n$ identity matrix.

The following proposition will be useful.

Proposition III.1. *Let $a, b, c \in \mathbb{R}^n$. Then, for any vector $q \in \mathbb{R}^n$ which satisfies $\|q - c\|_2 \leq \|q - b\|_2$ we have that*

$$\|q - a\|_2 \geq \frac{1}{2}(\|c - a\|_2 - \|b - a\|_2).$$

With this proposition at hand, we can begin bounding E_{HT}. Let $\mathcal{Q}$ denote the Q-function, and define the half-space $\mathcal{H} = \{y^n \ : \ \|y^n - g(\theta_f)\| \leq \|y^n - g(\theta_n)\|\}$. We have

$$\begin{aligned}&P^\theta_{Y^n}\left(\log \frac{dP^{\theta_f}_{Y^n}}{dP^{\theta_n}_{Y^n}}(Y^n) \geq 0\right) \\ &= P^\theta_{Y^n}(\|Y^n - g(\theta_f)\|_2 \leq \|Y^n - g(\theta_n)\|_2) \\ &= P^\theta_{Y^n}(Y^n \in \mathcal{H}) = \mathcal{Q}(\mathrm{dist}(g(\theta), \mathcal{H})/\sigma) \\ &\leq \mathcal{Q}\left(\frac{\|g(\theta_f) - g(\theta)\|_2 - \|g(\theta_n) - g(\theta)\|_2}{2\sigma}\right).\end{aligned}$$

Thus, recalling that $Q(t) < \exp(-t^2/2)$, we have

$$\begin{aligned}&-\log P^\theta_{Y^n}\left(\log \frac{dP^{\theta_f}_{Y^n}}{dP^{\theta_n}_{Y^n}}(Y^n) \geq 0\right) \\ &\geq \frac{1}{4}\left[\sqrt{\frac{\|g(\theta_f) - g(\theta)\|_2^2}{2\sigma^2}} - \sqrt{\frac{\|g(\theta_n) - g(\theta)\|_2^2}{2\sigma^2}}\right]^2.\end{aligned} \tag{6}$$

Defining the function

$$\psi(\tau, \delta) \triangleq \min_{\substack{\theta, \theta_n, \theta_f \in \Theta \\ \ell(\theta, \theta_n) \leq \tau \\ \ell(\theta, \theta_f) \geq \delta}} \left[\sqrt{\frac{\|g(\theta_f) - g(\theta)\|_2^2}{2\sigma^2}} - \sqrt{\frac{\|g(\theta_n) - g(\theta)\|_2^2}{2\sigma^2}}\right]$$

we obtain that

$$E_{HT}(\tau, \delta) \geq \frac{1}{4}\psi^2(\tau, \delta). \tag{7}$$

Now, we can apply our bound to various Gaussian models (and squared loss). We start with the simplest case of the Gaussian location model, and then move to more complicated models.

A. Gaussian Location model

For the Gaussian location model (GLM) we have that $n = d$ and $g(\theta) = \theta$ is the identity function. Thus, it is immediate to check that

$$\psi(\tau, \delta) = \frac{\sqrt{\delta} - \sqrt{\tau}}{\sqrt{2\sigma^2}}.$$

Substituting this into (7) and using Theorem II.5, we see that if for all $\delta' \geq \delta$ we have

$$\frac{1}{8\sigma^2}(\sqrt{\delta'} - \sqrt{\tau})^2 - d \log \frac{\sqrt{\delta'}}{\sqrt{\tau}} = \Omega(d), \tag{8}$$

then $P^\theta_{Y^n}\left[\left\|\theta - \hat{\theta}(Y^n)\right\|_2^2 > \delta\right] \to_{d \to \infty} 0$. The following proposition, which is straightforward to verify, will be useful.

Proposition III.2. *Let $a, b > 0$ and consider the function*

$$\phi(\tau, \delta) = a(\sqrt{\delta} - \sqrt{\tau})^2 - d \log \frac{\sqrt{\delta}}{\sqrt{\tau}} - d \cdot b.$$

For $\tau = \frac{\log(2)+b}{a} d$ and $\delta^* = \frac{4(\log(2)+b)}{a} d > \tau$, we have that $\phi(\tau, \delta) = \Omega(d)$ for all $\delta > \delta^*$.

Now, applying Proposition III.2 with $a = 1/8\sigma^2$ and $b = 0$, we see that, by (8), for the GLM

$$P^\theta_{Y^n}\left(\|\theta - \hat{\theta}(Y^n))\|_2^2 > 32 \log(2) \sigma^2 d\right) \to 0.$$

Since the minimax risk for the GLM is $\sigma^2 d$ [4], we obtained the optimal behavior up to constants. While the minimax risk for the GLM is trivial to obtain with much simpler method, we now demonstrate that our technique yields the optimal behavior of the minimax risk also in more involved problems.

B. Spiked Wigner Model

In the spiked Wigner model, the parameter space is[2] $\Theta = \{\theta \in \mathbb{R}^d \ : \ 1 \leq ||\theta|| \leq r\}$ and the observation is the $d \times d$ matrix

$$Y = \lambda \theta \theta^T + \frac{1}{\sqrt{d}} W$$

where the entries of W are i.i.d $\mathcal{N}(0, \sigma^2)$. This model can clearly be cast within our framework by taking $n = d^2, \sigma = \frac{1}{\lambda}$, $g(\theta) = \sqrt{d} \cdot \mathrm{vec}[\theta\theta^T]$,where the vec operation simply stacks the columns of a matrix to one long vector. We further require $\lambda > 10$. The goal is to estimate θ. Since this model is invariant to multiplication of θ by -1, the standard quadratic loss is inadequate, and instead, we use the loss $\ell(\theta, \hat{\theta}) = \min\{\|\theta - \hat{\theta}\|_2^2, \|\theta + \hat{\theta}\|_2^2\}$. If $\mathcal{C}$ is a τ-cover under quadratic loss, it is clearly also a τ-cover under the more forgiving loss $\ell(\theta, \hat{\theta})$. We also have that changing the loss from standard quadratic loss to the loss $\ell(\theta, \hat{\theta})$ increases $M_{\mathcal{C}}(\delta)$ by at most a factor 2, which is negligible in our asymptotic analysis. Thus, it only remains to lower bound $\psi(\tau, \delta)$ for $g(\theta) = \sqrt{d} \cdot \mathrm{vec}[\theta\theta^T]$. We have the following proposition (whose proof is omitted due to space limitations).

Proposition III.3. *Let $c_1 < 100$ be some universal constant, $\delta^* = \frac{c_1}{\lambda^2}$, and $\tau = \frac{2(\sqrt{2}-1)}{9}\frac{\delta^*}{4}$. Then for any $\delta > \delta^*$ we have that*

$$\psi(\tau, \delta) \geq \lambda\sqrt{d}\left(\sqrt{(\sqrt{2}-1)}\sqrt{\delta} - \frac{3}{\sqrt{2}}\sqrt{\tau}\right).$$

Thus, using Theorem II.5, and Proposition III.2, after some algebraic manipulations we see that there is a $\mathcal{C}$ for the quantized MLE for which $P^{\theta}_{Y^n}\left(\|\theta - \hat{\theta}(Y^n)\|_2^2) > \delta^*\right)$ vanishes with d, with

$$\delta^* = \frac{c_1}{\lambda^2}$$

It is known, see e.g. [4, Theorem 33.18] that if $\lambda < 1$, then no estimator can achieve non trivial (< 1) error for all $\theta \in \Theta$. Thus, our bounds recover the optimal scaling of the signal-to-noise-ratio (SNR) threshold. Furthermore, given the SNR is above the threshold, we obtain a loss with the same scaling as those achieved by spectral methods [14], [15] (although the constants are better for the spectral methods).

C. Multi-Reference Alignment

In the multi-reference alignment (MRA) problem [16] we have $\theta \in \{\theta \in \Theta = \mathbb{R}^d \ : \ \|\theta\|_2 \leq r\}$ and we observe n measurements of the form

$$Y_j = R_{k_j}\theta + Z_j, \quad j = 1, \ldots, m,$$

where $R_k\theta$ is a cyclic shift of θ by k indices, $k_j \overset{i.i.d.}{\sim} \mathrm{Unif}([d])$ (here $[d] = \{0, \ldots, d-1\}$), and $Z_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$ are statistically independent of θ. The challenge in estimating θ up to a cyclic shift: $\ell(\theta, \theta') = \min_{i=0,\ldots,k-1} \|\theta - R_i\theta'\|_2^2$ is that the cyclic shifts $R_{k_1}, \ldots, R_{k_m}$ are unknown. Were they known, the problem would become equivalent to the GLM. Consequently, the quadratic minimax risk for the MRA must satisfy $R^* \geq \sigma^2 d/m$.

In order to obtain a lower bound for this problem using our technique, we cast the problem as a Gaussian problem in md dimensions, with the parameter space $\tilde{\Theta} = \mathbb{R}^d \times [d]^m$ such that for $\tilde{\theta} = (\theta, k_1, \cdots, k_m) \in \mathbb{R}^d \times [d]^n$ we have that $g(\tilde{\theta}) = \mathrm{vec}[R_{k_1}\theta, \ldots, R_{k_m}\theta] \in \mathbb{R}^{md}$, where vec$[\cdot]$ is as defined in III-B. Thus, setting $n = md$ we have that $P^{\tilde{\theta}}_{Y^n} = \mathcal{N}(g(\tilde{\theta}), \sigma^2 I_n)$, which is the simple Gaussian model we have already studied. For the loss $\tilde{\ell}(\tilde{\theta}, \hat{\tilde{\theta}}) = \frac{1}{m}\|g(\tilde{\theta}) - g(\hat{\tilde{\theta}})\|_2^2$, we clearly have (by (6)) that

$$\psi(\tau, \delta) = \frac{\sqrt{\delta} - \sqrt{\tau}}{\sqrt{2\sigma^2/m}}.$$

Furthermore, note that if $\mathcal{C}$ is a τ-cover for $\mathbb{R}^d$ under quadratic loss, then

$$\tilde{\mathcal{C}} = \mathcal{C} \times [d]^m$$

is a τ-cover for $\tilde{\Theta}$ under the loss $\tilde{\ell}$. Furthermore, it can be shown that $\log M_{\tilde{\mathcal{C}}}(\delta) \leq \log M_{\mathcal{C}}(\delta) + m \log d$, where $M_{\tilde{\mathcal{C}}}(\delta)$ is defined with respect to the loss $\tilde{\ell}$ and $M_{\mathcal{C}}(\delta)$ with respect to the ℓ_2 loss on $\mathbb{R}^d$. Thus, using Theorem II.5, we see that if for all $\delta' \geq \delta$ we have

$$\frac{m}{8\sigma^2}(\sqrt{\delta'} - \sqrt{\tau})^2 - d \log \frac{\sqrt{\delta'}}{\sqrt{\tau}} - d \cdot m \frac{\log d}{d} = \Omega(d),$$

then $P^{\tilde{\theta}}_{Y^n}\left[\tilde{\ell}(\tilde{\theta}, \hat{\tilde{\theta}}(Y^n) > \delta\right] \to_{d\to\infty} 0$. Using Proposition III.2 we deduce that there is a $\mathcal{C}$ for the quantized MLE for which $P^{\tilde{\theta}}_{Y^n}\left[\tilde{\ell}(\tilde{\theta}, \hat{\tilde{\theta}}(Y^n) > \delta^*\right] \to_{d\to\infty} 0$, where

$$\delta^* = \frac{32\sigma^2 d}{m}(\log 2 + m\frac{\log d}{d}). \tag{9}$$

In the MRA problem, the shifts $R_{k_1}, \ldots, R_{k_m}$ are nuisance parameters, and the goal is only to estimate θ up to a cyclic shift under the loss $\ell(\theta, \hat{\theta}) = \min_{i\in[d]} \|\theta - R_i\hat{\theta}\|_2^2$. Clearly, if we find an estimator $\hat{\tilde{\theta}} = (\hat{\theta}, \hat{k}_1, \cdots, \hat{k}_m)$ for $\tilde{\theta} = (\theta, k_1, \cdots, k_m)$ with $\tilde{\ell}(\tilde{\theta}, \hat{\tilde{\theta}}) < \delta$, then also $\min_{i\in[d]} \|\theta - R_i\hat{\theta}\|_2^2 < \delta$. Thus, δ^* from (9) upper bounds the minimax risk for MRA in high dimensions. The main insight from (9) is that if the number of measurements is not too large, namely, $m = O(\frac{d}{\log d})$, the minimax risk of MRA scales exactly like that of the GLM. On the other hand, our bound becomes quite useless for $m = \omega(\frac{d}{\log d})$.

While the risk for MRA in high dimensions was previously studied for a given prior distribution [17] or in a minimax setting under the assumption of "generic" signals [18], to the best of our knowledge, this is the first minimax upper bound for this problem which only assumes $\|\theta\|_2 = e^{e^{o(d)}}$.

IV. Acknowledgements

This work was supported by ISF under Grant 1641/21.

[2]It is more common to assume that $||\theta||$ is constant, or drawn from a prior for which $||\theta||$ concentrates. The parameter space Θ defined above is more general.

REFERENCES

[1] L. LeCam, "Convergence of estimates under dimensionality restrictions," The Annals of Statistics, pp. 38–53, 1973.

[2] I. A. Ibragimov and R. Z. Has' Minskii, Statistical estimation: asymptotic theory. Springer Science & Business Media, 2013, vol. 16.

[3] M. S. Pinsker, "Optimal filtering of square-integrable signals in gaussian noise," Problemy Peredachi Informatsii, vol. 16, no. 2, pp. 52–68, 1980.

[4] Y. Polyanskiy and Y. Wu, Information Theory ,From Coding to Learning. Cambridge university press, 2023, to appear. Draft available at https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf.

[5] H. L. Van Trees, Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory. John Wiley & Sons, 2004.

[6] M. J. Wainwright, High-dimensional statistics: A non-asymptotic viewpoint. Cambridge university press, 2019, vol. 48.

[7] L. Le Cam, Asymptotic methods in statistical decision theory. Springer Science & Business Media, 2012.

[8] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, vol. 65, pp. 181–237, 1983.

[9] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and kolmogorov's entropy," The Annals of Statistics, vol. 13, no. 2, pp. 768–774, 1985.

[10] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," Annals of Statistics, pp. 1564–1599, 1999.

[11] P. Boroumand and A. Guillén i Fàbregas, "Mismatched binary hypothesis testing: Error exponent sensitivity," IEEE Transactions on Information Theory, vol. 68, no. 10, pp. 6738–6761, 2022.

[12] P. Erdos and C. Rogers, "Covering space with convex bodies," Acta Arithmetica, vol. 7, no. 3, pp. 281–285, 1962.

[13] O. Ordentlich, O. Regev, and B. Weiss, "Bounds on the density of smooth lattice coverings," to be submitted, 2023.

[14] D. Féral and S. Péché, "The largest eigenvalue of rank one deformation of large wigner matrices," Communications in mathematical physics, vol. 272, pp. 185–228, 2007.

[15] A. Perry, A. S. Wein, A. S. Bandeira, and A. Moitra, "Optimality and sub-optimality of pca i: Spiked random matrix models," The Annals of Statistics, vol. 46, no. 5, pp. 2416–2451, 2018.

[16] A. Perry, J. Weed, A. S. Bandeira, P. Rigollet, and A. Singer, "The sample complexity of multireference alignment," SIAM Journal on Mathematics of Data Science, vol. 1, no. 3, pp. 497–517, 2019.

[17] E. Romanov, T. Bendory, and O. Ordentlich, "Multi-reference alignment in high dimensions: sample complexity and phase transition," SIAM Journal on Mathematics of Data Science, vol. 3, no. 2, pp. 494–523, 2021.

[18] Z. Dou, Z. Fan, and H. Zhou, "Rates of estimation for high-dimensional multi-reference alignment," arXiv preprint arXiv:2205.01847, 2022.