Electronic Journal of Statistics
Vol. 14 (2020) 2988–3031
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1739

From Gauss to Kolmogorov: Localized measures of complexity for ellipses

Yuting Wei

Statistics & Data Science Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA, e-mail: ytwei@cmu.edu

Billy Fang[†]

Department of Statistics, University of California, Berkeley, Berkeley, California 94720, USA, e-mail: blfang@berkeley.edu

and

Martin J. Wainwright[*]

Department of Statistics and EECS, University of California, Berkeley, Berkeley, California 94720, USA, e-mail: wainwrig@berkeley.edu

Abstract: The Gaussian width is a fundamental quantity in probability, statistics and geometry, known to underlie the intrinsic difficulty of estimation and hypothesis testing. In this work, we show how the Gaussian width, when localized to any given point of an ellipse, can be controlled by the Kolmogorov width of a set similarly localized. Among other consequences, this connection, when coupled with a previous result due to Chatterjee, leads to a tight characterization of the estimation error of least-squares regression as a function of the true regression vector within the ellipse. This characterization reveals that the rate of error decay varies substantially as a function of location: as a concrete example, in Sobolev ellipses of smoothness α, we exhibit rates that vary from $(\sigma^2)^{\frac{2\alpha}{2\alpha+1}}$, corresponding to the classical global rate, to the faster rate $(\sigma^2)^{\frac{4\alpha}{4\alpha+1}}$. We also show how the local Kolmogorov width can be related to local metric entropy.

AMS 2000 subject classifications: Primary 62F10, 62F30; secondary 62G08.
Keywords and phrases: Complexity measure, ellipse constraint, Kolmogorov width, least squares, adaptive estimation.

Received December 2019.

1. Introduction

The Gaussian width is an important measure of the complexity of a set, and it plays an important role in geometry, statistics and probability theory. Most

[*]Supported partially by a National Science Foundation Graduate Research Fellowship.
[†]Supported in part by Office of Naval Research grant DOD ONR-N00014, and National Science Foundation grant NSF-DMS-1612948.

relevant to this paper is its central role in empirical process theory, where the Gaussian width and its Bernoulli analogue (known as the Rademacher width) can be used to upper bound the error for various types of nonparametric estimators [31, 32, 3, 20, 5, 36]. More recently, these same complexity measures have also been shown to play an important role in high-dimensional testing problems [38, 41].

For a general set, it is non-trivial to provide analytical expressions for its Gaussian or Rademacher widths. There are a variety of techniques for obtaining bounds, including upper bounds via the classical entropy integral of Dudley, as well as lower bounds due to Sudakov-Fernique (see the book [21] for details on these and other results). Talagrand [27] introduced the generic chaining technique that, in principle, leads to sharp lower and upper bounds on Gaussian widths. However, for an arbitrary set, it is generally impossible to evaluate the expressions obtained from the generic chaining; we note this area of research is currently very active (see, e.g., the papers [33, 34]). For applications in statistics, it is of considerable interest to develop techniques that connect and help control various forms of widths.

In this paper, we study a class of Gaussian widths that arise in the context of estimation over (possibly infinite-dimensional) ellipses. As we describe below, many non-parametric problems, among them are regression and density estimation over classes of smooth functions, can be reduced to such ellipse estimation problems. Obtaining sharp rates for such estimation problems requires studying a *localized* notion of Gaussian width, in which the ellipse is intersected with a Euclidean ball around the element θ^* being estimated. The main technical contribution of this paper is to show how this localized Gaussian width can be bounded, from both above and below, using a localized form of the Kolmogorov width [24]. As we show with a number of corollaries, this Kolmogorov width can be calculated in many interesting examples.

Our work makes a connection to the evolving line of work on instance-specific rates in estimation and testing. Within the decision-theoretic framework, the classical approach is to study the (global) minimax risk over a certain problem class. In this framework, methods are compared via their worst-case behavior as measured by performance over the entire problem class. For the ellipse problems considered here, global minimax risks in various norms are well-understood; for instance, see the classic papers [25, 14, 15], as well as the more recent work [17]. When the risk function is near to constant over the set, then the global minimax risk is reflective of the typical behavior. If not, then one is motivated to seek more refined ways of characterizing the hardness of different problems, and the performance of different estimators.

One way of doing so is by studying the notion of an adaptive estimator, meaning one whose performance automatically adapts to some (unknown) property of the underlying function being estimated. For instance, estimators using wavelet bases are known to be adaptive to unknown degree of smoothness [8, 9]. Similarly, in the context of shape-constrained problems, there is a line of work showing that for functions with simpler structure, it is possible to achieve faster rates than the global minimax ones (e.g. [23, 40, 6]). A related line of work,

including some of our own, has studied adaptivity in the context of hypothesis testing (e.g., [30, 2, 37]). The adaptive estimation rates established in this work also share this spirit of being instance-specific.

1.1. Some motivating examples

A primary motivation for our work is to understand the behavior of least-squares estimators over ellipses. Accordingly, let us give a precise definition of the ellipse estimation problem, along with some motivating examples.

Given a fixed integer d and a sequence of non-negative scalars, ordered in the non-decreasing fashion $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d \geq 0$, we can define an elliptical norm on $\mathbb{R}^d$ via

$$\|\theta\|_{\mathcal{E}}^2 := \sum_{j=1}^{d} \frac{\theta_j^2}{\mu_j}.$$

Here for any coefficient $\mu_k = 0$, we interpret the constraint as enforcing that $\theta_k = 0$. For any radius $R > 0$, this semi-norm defines an ellipse of the form

$$\mathcal{E}(R) := \Big\{ \theta \in \mathbb{R}^d \;\; | \;\; \|\theta\|_{\mathcal{E}} \leq R \Big\}. \tag{1}$$

We frequently focus on the case $R = 1$, in which case we adopt the shorthand notation $\mathcal{E}$ for the set $\mathcal{E}(1)$. Whereas equation (1) defines a finite-dimensional ellipse, it should be noted that our theory also applies to infinite-dimensional ellipses for sequences $\{\mu_j\}_{j=1}^{\infty}$ that are summable. Such results can be recovered by studying a truncated version of the ellipse with finite dimension d, and then taking suitable limits. In order to simplify the exposition, we develop our results with finite d, noting how they extend to infinite dimensions after stating our results.

Suppose that for some unknown vector $\theta^* \in \mathcal{E}$, we obtain noisy observations of the form

$$y = \theta^* + \sigma w, \qquad \text{where } w \sim \mathcal{N}(0, \mathbf{I}_d). \tag{2}$$

We assume that the ellipse $\mathcal{E}$ and noise standard deviation σ is known. The goal of ellipse estimation is to specify a mapping $y \mapsto \widehat{\theta}(y)$ such that the associated Euclidean risk

$$\mathbb{E}_y \Big[\|\widehat{\theta}(y) - \theta^*\|_2^2 \Big] = \mathbb{E}_y \Big[\sum_i (\widehat{\theta}_i - \theta_i^*)^2 \Big],$$

is as small as possible.

Let us consider some concrete problems that can be reduced to instances of ellipse estimation.

Example 1 (Linear prediction with correlated designs)**.** Suppose that we have observations from the standard linear model

$$\widetilde{y} = X\beta^* + \nu w,$$

where $\widetilde{y} \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is a (fixed, non-random) design matrix, and $w \sim N(0, \mathbf{I}_n)$ is noise. Suppose moreover that we know a priori that $\|\beta^*\|_2 \leq R$ for some radius $R > 0$. Alternatively, we can think of a condition of this form arising implicitly when using estimators such as ridge regression.

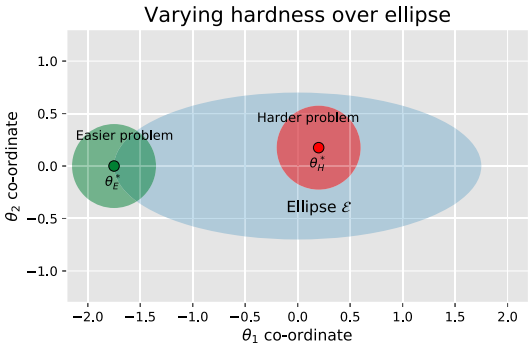
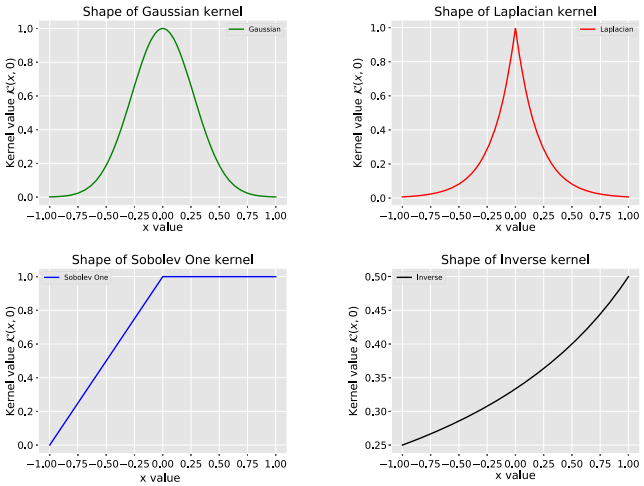


FIG 1. *Illustration of the ellipse estimation problem. The goal to estimate an unknown vector θ^* belonging to an ellipse based on noisy observations. The local geometry of the ellipse controls the difficulty of the problem: due to its proximity to the narrow end of the ellipse, the vector θ_E^* is relatively easy to estimate. By contrast, the vector θ_H^* should be harder, since it lies closest to the center of the ellipse. The theory given in this paper confirms this intuition; see Section 4 for details.*

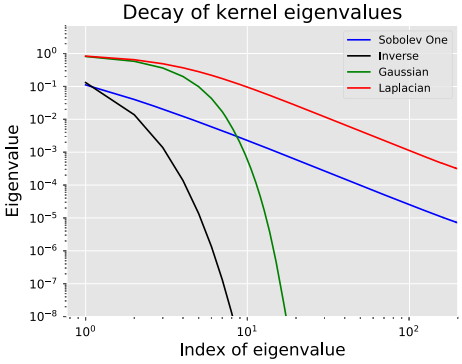
Given an estimate $\widehat{\beta}$, its prediction accuracy can be assessed via the mean-squared error $\mathbb{E}\big[\frac{1}{n}\|X\widehat{\beta} - X\beta^*\|_2^2\big]$, where the expectation is taken over the observation noise. Equivalently, letting $\widehat{\theta} = X\widehat{\beta}/\sqrt{n}$ and $\theta^* = X\beta^*/\sqrt{n}$, our problem is to minimize the mean-squared error $\mathbb{E}\|\widehat{\theta} - \theta^*\|_2^2$. After this transformation, we arrive at the observation model $y = \theta^* + \frac{\nu}{\sqrt{n}}w$, which is a version of our original model (2) with $d = n$ and $\sigma = \frac{\nu}{\sqrt{n}}$. Moreover, the constraint on the ℓ_2-norm of β^* translates into an ellipse constraint on θ^*. In particular, the ellipse is determined by the non-zero eigenvalues of the matrix $\frac{1}{n}XX^\top \in \mathbb{R}^{n \times n}$.

As shown in Figure 1, it is natural to conjecture that the location of θ^* within this ellipse affects the difficulty of estimation. Note that $\mathbb{E}\|y - \theta^*\|_2^2 = \nu^2/n$, so that on average, the observed vector y lies at squared Euclidean distance ν^2/n from the true vector. In certain favorable cases, such as a vector θ_E^* that lies at or close to the boundary of an elongated side of the ellipse, the side-knowledge that $\theta^* \in \mathcal{E}$ is helpful. In other cases, such as a vector θ_H^* that lies closer to the center of the ellipse, the elliptical constraint is less helpful. The theory to be developed in this paper makes this intuition precise. In particular, Section 4 is devoted to a number of consequences of our main results for the problem of estimation in ellipses.

Example 2 (Non-parametric regression using reproducing kernels)**.** We now turn to a class of non-parametric problems that involve a form of ellipse estimation. Suppose that our goal is to predict a response $z \in \mathbb{R}$ based on observing a collection of predictors $x \in \mathcal{X}$. Assuming that pairs (X, Z) are drawn jointly from some unknown distribution $\mathbb{P}$, the optimal prediction in terms of mean-squared error is given by the conditional expectation $f^*(x) := \mathbb{E}[Z \mid X = x]$. Given a collection of samples $\{(x_i, z_i)\}_{i=1}^n$, the goal of non-parametric regression is to produce an estimate $\widehat{f}$ that is as close to f^* as possible.



(a)



(b)

FIG 2. *(a) Illustration of various kernel functions defined on* $[-1, 1] \times [-1, 1]$*. Each plot shows the kernel value* $\mathcal{K}(x, 0)$ *for* $x \in [-1, 1]$*. (b) Illustration of the kernel eigenvalues* $\{\mu_j\}_{j=1}^n$ *for kernel matrices* K *generated from the kernel functions in part (a). Each log-log plot shows the eigenvalue versus the index: note how the Gaussian kernel eigenvalues decay at an exponential rate, whereas those of the Sobolev-One spline kernel decay at a polynomial rate.*

Assuming that the samples are i.i.d., we can rewrite our observations in the form

$$z_i = f^*(x_i) + \gamma v_i, \qquad \text{for } i = 1, \ldots, n, \tag{3}$$

where v_i is an independent sequence of zero-mean noise variables with unit variance. A computationally attractive way of estimating f^* is to perform least-squares regression over a *reproducing kernel Hilbert space*, or RKHS for short [1, 18, 12, 35]. Any such function class is defined by a symmetric, positive definite kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$; standard examples include the Gaussian kernel, Laplace kernel, and the Sobolev (spline) kernels; see Figure 2 for some illustrative examples. Now suppose that f^* belongs to the RKHS induced by the kernel $\mathcal{K}$, say with Hilbert norm $\|f^*\|_{\mathcal{H}} \leq R$. In this case, the representer theorem [18] implies that the observation model (3) is equivalent to

$$z = \sqrt{n} K \alpha^* + \gamma v \qquad \text{for some } \alpha^* \in \mathbb{R}^n,$$

where $K \in \mathbb{R}^{n \times n}$ is the $n \times n$ kernel matrix with entries $K_{ij} = \mathcal{K}(x_i, x_j)/n$ for each $i, j = 1, \ldots, n$, and vector v is a n-dimensional vector formed by v_i. The representer theorem and our choice of scaling ensures that $\|f^*\|_{\mathcal{H}}^2 = (\alpha^*)^\top K \alpha^*$, meaning that α^* belongs to the ellipse of radius R defined by the symmetric and PSD kernel matrix K.

Note that the matrix K can be diagonalized as $K = UDU^\top$, where U is orthonormal, and $D = \operatorname{diag}\{\mu_1, \mu_2, \ldots, \mu_n\}$ is a diagonal matrix of non-negative eigenvalues. Following this transformation, we arrive at an instance of the standard ellipse model

$$y = \theta^* + w \qquad \text{where } w = \gamma U^\top v/\sqrt{n}, \quad y = U^\top z/\sqrt{n},$$

and where $\theta^* = U^\top K \alpha^*$ belongs to the standard ellipse (1) defined by the eigenvalues of K. Note that the noise vector $w = \gamma U^\top v/\sqrt{n}$ has zero-mean entries each with standard deviation $\sigma = \gamma/\sqrt{n}$. The entries of w are not exactly Gaussian (unless the initial noise vector v was jointly Gaussian), but are often well-approximated by Gaussian variables due to central limit behavior for large n.

1.2. Organization and notation

The remainder of this paper is organized as follows. In Section 2, we introduce some background on approximation-theoretic quantities, including the Gaussian width, metric entropy, and the Kolmogorov width. Section 3 is devoted to the statement of our main results, while Section 4 develops a number of their specific consequences for ellipse estimation. In Section 5, we provide the proofs of our main results, with more technical aspects of the arguments provided in the appendices.

Here we summarize some notation that are used throughout this paper. Given any functions $f(\sigma, d)$ and $g(\sigma, d)$, we denote $f(\sigma, d) \lesssim g(\sigma, d)$ to indicate $f(\sigma, d) \leq c g(\sigma, d)$ for some universal constant $c \in (0, \infty)$ that is in-

dependent of any problem parameters, such as σ, d, θ^* etc. Similarly, we define $f(\sigma, d) \gtrsim g(\sigma, d)$. We write $f(\sigma, d) \asymp g(\sigma, d)$ if $f(\sigma, d) \lesssim g(\sigma, d)$ and $f(\sigma, d) \gtrsim g(\sigma, d)$ are both satisfied.

2. Background

Before proceeding to the statements of our main results, we introduce some background on the notion of Gaussian width, Kolmogorov width, as well as setting the estimation problem with ellipse constraint.

2.1. Gaussian width

Given a bounded subset $\mathcal{S} \subset \mathbb{R}^d$, the *Gaussian width* of $\mathcal{S}$ is defined as

$$\mathscr{G}(\mathcal{S}) := \mathbb{E}[\sup_{u \in \mathcal{S}} \langle u, w \rangle] = \mathbb{E}\Big[\sup_{u \in \mathcal{S}} \sum_{i=1}^{d} w_i u_i\Big], \qquad \text{where } w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1).$$

It measures the size of set $\mathcal{S}$ in a certain sense.

It is also useful to define the classical notions of packing and covering entropy. An *ϵ-cover* of a set $\mathcal{S}$ with respect to the $\|\cdot\|_2$ metric is a discrete set $\{\theta^1, \ldots, \theta^N\} \subset \mathcal{S}$ such that for each $\theta \in \mathcal{S}$, there exists some $i \in \{1, \ldots, N\}$ satisfying $\|\theta - \theta^i\|_2 \leq \epsilon$. The *$\epsilon$-covering number* $N(\epsilon, \mathcal{S})$ is the cardinality of the smallest ϵ-cover, and the logarithm of this number $\log N(\epsilon, \mathcal{S})$ is called the *covering metric entropy* of set $\mathcal{S}$.

Similarly, an *ϵ-packing* of a set $\mathcal{S}$ is a set $\{\theta^1, \ldots, \theta^M\} \subset \mathcal{S}$ satisfying $\|\theta^i - \theta^j\|_2 > \epsilon$ for all $i \neq j$. The size of the largest such packing is called the *ϵ-packing number* of $\mathcal{S}$, which we denote by $M(\epsilon, \mathcal{S})$. It is related to the (covering) metric entropy by the inequalities

$$\log M(2\epsilon, \mathcal{S}) \leq \log N(\epsilon, \mathcal{S}) \leq \log M(\epsilon, \mathcal{S}).$$

For this reason, we use the term metric entropy to refer to either the covering or packing metric entropy, since they differ only in constant terms.

The connection between Gaussian width and metric entropy is well-studied (e.g. [11, 28, 36]). For our future discussion, we collect a few results here as reference. First, Dudley's entropy integral [11] is an upper bound for the Gaussian width—that is,

$$\mathscr{G}(\mathcal{S}) \leq c \int_0^{\operatorname{diam}(\mathcal{S})} \sqrt{\log N(\epsilon, \mathcal{S})}\, d\epsilon,$$

for some universal constant $c > 0$. This upper bound also holds for more general sub-Gaussian processes. Dudley's bound can be much looser than the more refined bounds obtained through Talagrand's generic chaining, which are tight up

to a universal constant [28, Thm. 2.4.1]. For Gaussian processes like ours, Sudakov minoration (e.g., [4, Thm. 13.4]) provides a lower bound on the Gaussian width.

$$\mathscr{G}(\mathcal{S}) \geq \sup_{\epsilon>0} c\,\epsilon\,\sqrt{\log M(\epsilon, \mathcal{S})}. \tag{4}$$

Although we do not directly use this lower bound when proving our main lower bound (Theorem 2) below, we follow its spirit by constructing a large collection of well-separated points.

2.2. Kolmogorov width

In this section, we review the definition of the Kolmogorov width (see, e.g. [24]) and briefly discuss its properties. This geometric quantity plays the central role in our main results.

For a given compact set $\mathcal{S} \subset \mathbb{R}^d$ and integer $k \in [d]$, the *Kolmogorov k-width* of $\mathcal{S}$ is given by

$$\mathscr{W}_k(\mathcal{S}) := \min_{\Pi_k \in \mathcal{P}_k} \max_{\theta \in \mathcal{S}} \|\theta - \Pi_k \theta\|_2, \tag{5}$$

where $\mathcal{P}_k$ denotes the set of all k-dimensional orthogonal linear projections, and $\Pi_k\theta$ denotes the projection of θ to the corresponding k-dimensional linear space. Any projection Π_k achieving the minimum in expression (5) is said to be an *optimal projection* for $\mathscr{W}_k(\mathcal{S})$. Note that the Kolmogorov width $\mathscr{W}_k(\mathcal{S})$ is a non-increasing function of k, meaning that

$$\max_{\theta \in \mathcal{S}} \|\theta\|_2 = \mathscr{W}_0(\mathcal{S}) \geq \mathscr{W}_1(\mathcal{S}) \geq \ldots \geq \mathscr{W}_d(\mathcal{S}) = 0.$$

By definition, the Kolmogorov k-width measures how well the set $\mathcal{S}$ is approximated by the set of k dimensional linear spaces. We also make a note that the Kolmogorov k-width is understood to quantify the performance of the truncated series estimators (e.g. [10, 16]), and play an important role in density estimation and compressed sensing (see, [7, 13]). Recently, it is also shown to determine the local of testing rate in ellipses (see, [37]). We refer the readers to the book by Pinkus [24] for more details on the Kolmogorov width and its properties.

3. Main results

Let us first define the notion of localized Gaussian width formally, and then turn to the statement of our main results.

3.1. Localized Gaussian width

Let $\mathbb{B}(\delta)$ denote the Euclidean ball of radius δ centered at zero, and for a given vector $\theta^* \in \mathcal{E}$, define the shifted ellipse $\mathcal{E}_{\theta^*} := \{\theta - \theta^* \mid \theta \in \mathcal{E}\}$. The *localized*

Gaussian width at θ^* and scale δ is defined as

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) = \mathbb{E}\big[\sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle w, \Delta \rangle\big]. \tag{6}$$

Note that this quantity is simply the ordinary Gaussian width of the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$, and we say that it is localized since the Euclidean ball restricts it to a neighborhood of θ^*. See Figure 3 for an illustration of this set.

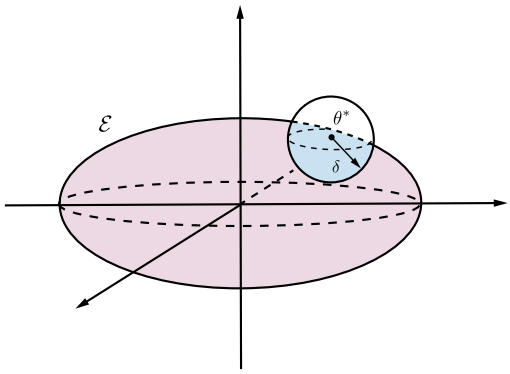


FIG 3. *An illustration of the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$. It is the intersection of the ellipse with Euclidean ball centered at θ^*, and thus varies according to the local geometry of the ellipse.*

We note that localized forms of Gaussian and Rademacher complexity are standard in the literature on empirical processes (e.g., [3, 19]), where it is known that they are needed to obtain sharp rates. In the case of least-squares estimation over convex sets, there is an extremely explicit connection between the localized Gaussian width and the associated estimation error [31, 5, 36]; we describe this relationship in more detail in Section 4 of the current paper as well as in Section D.

Our main results, to be stated in the following subsections, provide conditions under which we can provide a sharp characterization of the localized Gaussian width (6) in terms of the Kolmogorov width.

3.2. Upper bound on the localized Gaussian width

In order to state our first main result, we introduce an approximation-theoretic quantity having to do with the quality of a given k-dimensional projection. For a given integer $k \in \{1, \ldots, d\}$ and any k-dimensional linear projection Π_k, let us define the set

$$\Gamma(\theta^*, \delta, \Pi_k) := \left\{ \gamma \in \mathbb{R}^d \;\middle|\; \gamma > 0, \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \sum_{i=1}^{d} \frac{[\Delta_i - (\Pi_k \Delta)_i]^2}{\gamma_i} \leq 1 \right\}.$$

Here $\gamma > 0$ means that $\gamma_i > 0$ for each coordinate $i = 1, \ldots, n$. It can be verified that the set $\Gamma(\theta^*, \delta, \Pi_k)$ is always non-empty since the constant vector

$\gamma = \mu_1 \delta^2 \mathbf{1}$ always belongs to it. (Here $\mathbf{1}$ denotes the vector of all ones.) To provide some intuition for this definition, the vector $\Delta - \Pi_k(\Delta)$ corresponds to the error incurred by using the subspace associated with Π_k to approximate Δ. The positive vector $\gamma \in \mathbb{R}^d$ allows us to weight the entries of this error vector in computing the Euclidean norm of the weighted error.

We are now ready to state an upper bound on the localized Gaussian width.

Theorem 1. *Given any $\delta > 0$, projection tuple (k, Π_k), and vector $\theta^* \in \mathcal{E}$, we have*

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq \delta \sqrt{k} + \inf_{\gamma \in \Gamma(\theta^*, \delta, \Pi_k)} \sqrt{\sum_{i=1}^{d} \gamma_i}. \tag{7}$$

See Section 5.1 for the proof of this result.

Note that Theorem 1 holds for any dimension and projection pair (k, Π_k). Often the case, we can choose a specific pair for which the set $\Gamma(\theta^*, \delta, \Pi_k)$ is easy to characterize. In particular, given any fixed $\delta > 0$, let us define the *critical dimension*

$$k_*(\theta^*, \delta) := \arg \min_{k=1,\ldots,d} \left\{ \mathscr{W}_k\Big(\mathcal{E}_{\theta^*} \cap \mathbb{B}\big((1-\eta)\delta\big)\Big) \leq \frac{9}{10}\delta \right\}, \tag{8}$$

for some constant $\eta \in (0, 0.1)$. In words, this integer is the minimal dimension for which there exists a k_*-dimensional projection that approximates a neighborhood of the re-centered ellipse to $\frac{9}{10}\delta$-accuracy.[1] Although our notation does not explicitly reflect it, note that $k_*(\theta^*, \delta)$ also depends on the ellipse $\mathcal{E}$.

Given the integer $k_* \equiv k_*(\theta^*, \delta)$, we let $\Pi_{k_*} \in \mathcal{P}_{k_*}$ denote the minimizing projection in the definition (5) of the width, and note that for any vector Δ, the error associated with this projection is given by $\Delta - \Pi_{k_*}(\Delta)$. It can be seen in our later examples, this particular choice $(k_*, \ \Pi_{k_*})$ often yields tight control of the localized Gaussian width. So as to streamline notation, we adopt $\Gamma(\theta^*, \delta)$ as a shorthand for $\Gamma(\theta^*, \delta, \Pi_{k_*})$.

Regularity assumption For many ellipses encountered in practice, the first term in the upper bound (7) dominates the second term involving the set Γ. In order to capture this condition, we say the ellipse $\mathcal{E}$ is *regular at* θ^* if there exists some pair (k, Π_k) such that

$$\inf_{\gamma \in \Gamma(\theta^*, \delta, \Pi_k)} \sum_{i=1}^{d} \gamma_i \ \leq \ c \, \delta^2 \, k \qquad \text{for all } \delta > 0. \tag{9}$$

Here $c < \infty$ is any universal constant. When this condition holds, Theorem 1 implies the existence of another universal constant c' such that

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq c' \, \delta \sqrt{k} \qquad \text{for all } \delta > 0.$$

[1] The constants η and $9/10$ are chosen for the sake of convenience in the proof, but other choices of these quantities (which both must be strictly less than 1) are also possible.

As shown in Section A, the regularity condition (9) is a generalization of a condition previously introduced by Yang et al. [39] in the context of kernel ridge regression, and it holds for many examples encountered in practice.

As a direct consequence of Theorem 1, the following corollary holds.

Corollary 1. *If the regularity assumption (9) is satisfied with dimension and projection pair (k_*, Π_{k_*}), then the localized Gaussian width satisfies*

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq c_u \, \delta \sqrt{k_*} \qquad \text{for all } \delta > 0.$$

Let us illustrate the regularity condition (9) and associated consequences of Theorem 1 with some examples.

Example 3 (Gaussian width of the Euclidean ball)**.** We begin with a simple example: suppose that the ellipse $\mathcal{E}$ is the Euclidean ball in $\mathbb{R}^d$, specified by the aspect ratios $\mu_j = 1$ for all $j = 1, \ldots, d$, and let us use Theorem 1 to upper bound the Gaussian width at $\theta^* = 0$. For $\delta \in (0, \frac{1}{1-\eta})$ and any integer $k < d$, we have $\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) = (1-\eta)\delta$, because any k-dimensional projection must neglect at least one coordinate. Since $1 - \eta > 9/10$, we conclude that $k_*(0, \delta) = d$ for all $\delta \in (0, \frac{1}{1-\eta})$. With this choice of k_*, there is no error in the projection, meaning that $\inf_{\gamma \in \Gamma(\theta^*, \delta)} \sum_{i=1}^d \gamma_i = 0$. Consequently, the regularity condition (9) certainly holds, so that Theorem 1 implies that

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq c' \delta \sqrt{d}.$$

In fact, a direct calculation yields that $\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) = \delta(\sqrt{d} - o(1))$, where $o(1)$ is a quantity tending to zero as d grows (e.g., [36]). Consequently, our bound is asymptotically sharp up to the constant pre-factor in this special case.

We now turn to a second example that arises in non-parametric regression and density estimation under smoothness constraints:

Example 4 (Gaussian width for Sobolev ellipses)**.** Now consider an ellipse $\mathcal{E}$ defined by the aspect ratios $\mu_j = cj^{-2\alpha}$, where $\alpha > 1/2$ is a parameter. Ellipses of this form arise when studying non-parametric estimation problems involving functions that are α-times differentiable with Lebesgue-integrable α-derivative [29]. Let us again use Theorem 1 to upper bound the localized Gaussian width at $\theta^* = 0$. From classical results on Kolmogorov widths of ellipses [24] (see also Wei and Wainwright [37, Appendix C]), we know that $\mathscr{W}_k(\mathcal{E}_0) = \sqrt{\mu_{k+1}}$. Taking into account the intersection with the Euclidean ball, we find that

$$\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) = \min\left\{\sqrt{\mu_{k+1}}, (1-\eta)\delta\right\},$$

valid for any $\delta \in (0, \frac{1}{1-\eta}\sqrt{\mu_1})$ (see also Appendix A for details). Since $1 - \eta > 9/10$, we conclude that

$$k_*(0, \delta) = \arg\min\left\{\sqrt{\mu_{k+1}} \leq \frac{9}{10}\delta\right\} = \left\lceil \left(\frac{10\sqrt{c}}{9\delta}\right)^{1/\alpha} \right\rceil,$$

again valid for all $\delta \in (0, \frac{1}{1-\eta}\sqrt{\mu_1})$. Here the last inequality uses the fact that $\mu_j = cj^{-2\alpha}$.

This argument also shows that the corresponding projection subspace is spanned by the first k_* standard orthogonal vectors $\{e_i\}_{i=1}^{k_*}$. With this projection, any feasible vector $\gamma \in \Gamma(\theta^*, \delta)$ satisfies $\gamma_i \geq \mu_i \mathbf{1}\{i > k_*(0,\delta)\}$, meaning that

$$\inf_{\gamma \in \Gamma(\theta^*,\delta)} \sum_{i=1}^{d} \gamma_i = \sum_{j=k_*+1}^{d} \mu_j = c \sum_{j=k_*+1}^{d} j^{-2\alpha} \leq c \int_{k_*+1}^{\infty} t^{-2\alpha}\, dt = c\delta^{2-1/\alpha}. \quad (10)$$

On the other hand, we also have $\delta^2 k_*(0,\delta) \asymp \delta^{2-1/\alpha}$, so there exists some constant c', such that $\inf_{\gamma\in\Gamma(\theta^*,\delta)} \sum_{i=1}^{d} \gamma_i \leq c'\delta^2 k_*(0,\delta)$ which validates the regularity condition (9). Therefore, Theorem 1 guarantees that

$$\mathscr{G}(\mathcal{E}_0 \cap \mathbb{B}(\delta)) \leq c''\, \delta^{1-(1/2\alpha)}. \quad (11)$$

In fact, the above bound (11) can be shown to be tight up to a constant prefactor. See the discussion following Corollary 2 in the sequel for further details.

3.3. Lower bound on the localized Gaussian width

So far, we have derived an upper bound for the localized Gaussian width. In this section, we use information-theoretic methods to prove an analogous lower bound on the localized Gaussian width. This lower bound involves both the critical dimension $k_*(\theta^*, \delta)$, as previously defined in equation (8), and also a second quantity, one which measures the proximity of θ^* to the boundary of the ellipse. More precisely, for a given $\theta^* \in \mathcal{E}$, define the mapping $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ via

$$\Phi(\delta) = \begin{cases} 1 & \text{if } \delta > \|\theta^*\|_2/(1-\eta) \\ 1 \wedge \min\left\{ r \geq 0 \mid \delta^2 \leq \frac{1}{(1-\eta)^2} \sum_{i=1}^{d} \frac{r^2}{(r+\mu_i)^2} (\theta_i^*)^2 \right\} & \text{otherwise.} \end{cases} \quad (12)$$

As shown by Wei and Wainwright [37, Appendix F], this mapping is well-defined, and has the limiting behavior $\Phi(\delta) \to 0$ as $\delta \to 0^+$; for completeness, we include the verification of these claims in Section G, along with a sketch of the function. Let us denote $\Phi^{-1}(x)$ as the largest positive value of δ such that $\Phi(\delta) \leq x$. Note that by this definition, we have $\Phi^{-1}(1) = \infty$.

Recall that the elliptical norm on $\mathbb{R}^d$ is defined via $\|\theta\|_{\mathcal{E}}^2 := \sum_{j=1}^{d} \frac{\theta_j^2}{\mu_j}$. We are now ready to state our lower bound for the localized Gaussian width.

Theorem 2. *There exist universal constants $c_\ell, c > 0$ such that for all $\theta^* \in \mathcal{E}$,*

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \geq c_\ell\, \delta \sqrt{1 - \|\theta^*\|_{\mathcal{E}}^2}\, \sqrt{k_*(\theta^*, \delta)},$$
$$\text{for all } \delta \in \left(0, c\Phi^{-1}\big((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2\big) \wedge \sqrt{\mu_1}\right).$$

See Section 5.2 for the proof of this theorem.

We remark that the regularity condition (9) is not necessary for this result to hold. Additionally, note that the inequality $\delta < c\Phi^{-1}\big((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2\big)$ is equivalent to $\|\theta^*\|_{\mathcal{E}} < \frac{1}{1+\sqrt{\Phi(\delta/c)}}$. With this assumption, we consider the cases which are slightly bounded away from the boundary of the ellipse. Concretely, if we assume that $\|\theta^*\|_{\mathcal{E}} \leq 1/2$, then $(\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2 \geq 1$ therefore $\Phi^{-1}\big((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2\big) = \infty$.

3.4. Some consequences

One useful consequence of Theorem 1 and Theorem 2 is in providing sufficient conditions for tight control of the localized Gaussian width. If the ellipse $\mathcal{E}$ is regular at θ^*, then the above theorems imply the localized Gaussian width (6) is equivalent to $\delta\sqrt{k_*(\theta^*, \delta)}$ up to a multiplicative constant. Specifically, we have the sandwich relation

$$c_\ell \delta \sqrt{k_*(\theta^*, \delta)} \leq \mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq c_u \delta \sqrt{k_*(\theta^*, \delta)}, \tag{13}$$

for some universal positive constants c_ℓ and c_u and $0 < \delta < c\Phi^{-1}\big((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2\big)$.

Recall our earlier calculation from Example 3, where we showed that the localized Gaussian width scales as $\delta\sqrt{d}$, up to multiplicative constants. The sandwich relation (13) shows that this same scaling holds more generally with d replaced by $k_*(\theta^*, \delta)$. Thus, we can think of $k_*(\theta^*, \delta)$ corresponding to the "effective dimension" of the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$.

It is worthwhile pointing out that our results have a number of corollaries, in particular in terms of how local Gaussian widths and Kolmogorov widths are related to metric entropy. Recall the notion of the metric (packing) entropy as previously defined in Section 2.1. The following corollary provides a sandwich for $k_*(\theta^*, \delta)$ in terms of the metric entropy of the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$.

Corollary 2. *There are universal constants $c_j > 0$ such that for any pair $(\theta^*, \mathcal{E})$ satisfying the regularity condition (9), we have*

$$c_1 \log M\big(\frac{\delta}{2}, \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)\big) \overset{(i)}{\leq} k_*(\theta^*, \delta) \overset{(ii)}{\leq} c_2 \log M\big(c_0\delta, \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)\big), \tag{14}$$

for all $\delta \in (0, 1/e)$.

See Section B for the proof. The lower bound (i) is a relatively straightforward consequence of Sudakov's inequality (4), when combined with our results connecting the Kolmogorov and Gaussian widths. The upper bound (ii) requires a lengthier argument.

Recall that in Example 4, we argued that for the Sobolev ellipse with smoothness $\alpha > 1/2$, the Kolmogorov width at $\theta^* = 0$ is given by $k_*(0, \delta) = c\,(1/\delta)^{(1/\alpha)}$. Combining this calculation with Corollary 2, we find that the metric entropy is $\log M\big(\delta/2, \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)\big) = (1/\delta)^{1/\alpha}$ up to a multiplicative constant. This is a known fact that can be verified by constructing explicit packings of these function classes, but it serves to illustrate the sharpness of our results in this particular context.

4. Consequences for estimation

In the previous section, we established upper and lower bounds on the localized Gaussian width in Theorem 1 and Theorem 2. We now turn to some consequences of these bounds, in particular for the problem of constrained least-squares estimation. Our development involves combining ideas due to Chatterjee [5] on concentration of such least-squares estimators with our analysis of the localized Gaussian width.

Suppose that we are given observations $y \sim \mathcal{N}(\theta^*, \sigma^2 \mathbf{I}_n)$ with $\theta^* \in \mathcal{E}$ according to the earlier model (2), and we consider the constrained least squares estimator (LSE)

$$\widehat{\theta} := \arg\min_{\theta \in \mathcal{E}} \|y - \theta\|_2^2. \tag{15}$$

Let us assume that the ellipse $\mathcal{E}$ is regular at θ^*, so that the localized Gaussian width satisfies the bounds (13) with constants c_ℓ and c_u. Connecting the error $\|\widehat{\theta} - \theta^*\|_2$ to these Gaussian width bounds involves the following two functions

$$\begin{aligned} g^u(\delta) &:= \frac{\delta^2}{2} - \sigma c_\ell \delta \sqrt{k_*(\theta^*, \delta)}, \\ g^\ell(\delta) &:= \frac{\delta^2}{2} - \sigma c_u \delta \sqrt{k_*(\theta^*, \delta)}, \end{aligned} \tag{16}$$

with the critical dimension $k_*(\theta^*, \delta)$ defined in expression (8). We note that similar expressions emerge from the analysis of Chatterjee [5], wherein the above expressions involving the Kolmogorov width are replaced by the localized Gaussian width. Indeed, as we elaborate below, the above relationship (13) provides the link between these two widths.

With these definitions, let us consider the fixed point equation

$$\delta = c_\ell \sigma \sqrt{k_*(\theta^*, \delta)} \qquad \text{for } \delta \leq c\Phi^{-1}\big((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2\big) \wedge \sqrt{\mu_1}. \tag{17}$$

Since $\delta \mapsto k_*(\delta)$ is a non-increasing function of δ (see Wei and Wainwright [37, Appendix E]) while $\delta \mapsto \delta$ is increasing, if this fixed point problem (17) has a solution, then the solution is unique and we denote it as δ_*.

We can now give a precise statement relating the estimation rate of $\widehat{\theta}$ to the solution δ_* of the fixed point equation (17).

Proposition 1 (Least squares on ellipses)**.** *Let $\mathcal{E}$ be regular at θ^*, and let δ_* be the solution to the fixed point problem (17). Suppose furthermore the following conditions hold*

(a) The function g^ℓ is unimodal in δ.
(b) There exists a constant $c_1 \in (0,1)$ such that $c_u^2 k_*(\delta) \leq \frac{1}{4c_1^2} c_\ell^2 k_*(\delta_*)$ for $\delta = c_1 \delta_*$,
(c) There exists a constant $c_2 > 1$ such that $\delta \geq 2\sigma c_u \sqrt{k_*(\delta)}$ for $\delta = c_2 \delta_*$.

Then the error of the least squares estimator (15) *satisfies*

$$c\delta_* \le \|\widehat{\theta} - \theta^*\|_2 \le c'\delta_*, \qquad \text{with prob. } \ge 1 - 3\exp(-c''\delta_*^2/\sigma^2), \tag{18}$$

for some constants that depend only on c_1 *and* c_2.

See Section D for the proof of this result.

Note that this result is stated for the ellipse $\mathcal{E}(R)$ with $R = 1$. For arbitrary R one can easily rescale to obtain similar results; see equation (37) in Section D.1 for more detail. When we say g^ℓ is unimodal, we mean that there is some t such that g^ℓ is nondecreasing for $\delta < t$ and nonincreasing for $\delta > t$.

Equation (18) provides a high probability bound on the least-squares error. If furthermore $\delta_* \gg \sigma$ (which holds true in many cases including the examples shown below), then the probability $1 - 3\exp(-c''\delta_*^2/\sigma^2)$ goes to one, and we are guaranteed that the mean-squared error is sandwiched as

$$c\delta_*^2 \;\le\; \mathbb{E}\|\widehat{\theta} - \theta^*\|_2^2 \;\le\; c'\delta_*^2 \tag{19}$$

for some universal constants (c, c') by integrating the high probability bound; see Section D (in particular the expectation bound (35)) for details.

We claim the conditions of Proposition 1 are relatively mild. Note that the related function $g(t) := \frac{\delta^2}{2} - \sigma\mathcal{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta))$ is strongly convex [5, Thm. 1.1], as mentioned in Section D.1. So it is reasonable to believe that its approximation g^ℓ is unimodal. Moreover, the assumptions (b) and (c) essentially assert that g^ℓ does not change too drastically at two points $c_1\delta_*$ and $c_2\delta_*$ close to the critical radius δ_*. In the next section, we will check these assumptions for different examples.

Note that fixed point problem (17) can be viewed as a kind of a *critical equation* (e.g., [36, Ch. 13] and [39]), whose solution δ_* we call the *critical radius*. The proof of Proposition 1 relies on a result of Chatterjee [5] where the estimation rate is controlled in terms of the optimizer of a particular function that involves the localized Gaussian width. In order to compute this optimizer, one needs "exact" control of the localized Gaussian width. Upper bounds on the localized Gaussian width are tractable, but a *matching* lower bound is usually hard to obtain. The proof of Proposition 1 shows that with two-sided control of the localized Gaussian width (13), the estimation error also satisfies a matching lower bound. To be clear, the result obtained here is only tight up to universal constants, and moreover requires a regularity condition. We believe the latter is an artifact of our proof that is possibly removable, whereas pinning down tight constants, as in the paper [5], seems to require new techniques. In the next section, we illustrate the consequence of this result with some examples.

4.1. Adaptive estimation rates

We now demonstrate the consequences of Proposition 1 via some examples. We begin with the simple problem of estimation for $\theta^* = 0$, where we see a number of standard rates from the ellipse estimation literature. We then consider

some more interesting examples of extremal vectors, and show how the resulting estimation rates differ from the classical ones.

4.1.1. Estimating at $\theta^* = 0$

We begin our exploration by considering ellipse-constrained estimation problem at $\theta^* = 0$. In this section, we focus on two type of ellipses that are specified by aspect ratios μ_j that follow an α-polynomial decay and τ-exponential decay respectively. The first one corresponds to estimating a function in α-smooth Sobolev class—that is, functions that are almost everywhere α-times differentiable, and with the derivative $f^{(\alpha)}$ being Lebesgue integrable. The exponential decay corresponds to functions that are almost infinitely smooth everywhere. Examples of this kind include the reproducing Hilbert spaces with the Gaussian kernel, which satisfies such bound with $\tau = 2$ (real line) or $\tau = 1$ (compact domain) for the Lebesgue measure.

α-polynomial decay Consider an ellipse $\mathcal{E}$ defined by the aspect ratios $\mu_j = cj^{-2\alpha}$ for some $\alpha > 1/2$. In Example 4, inequality (10), it is verified that this ellipse is regular at 0, and that $k_*(\delta) \asymp \delta^{-1/\alpha}$. Thus, solving the fixed point problem (17) yields $\delta_* \asymp \sigma^{\frac{2\alpha}{2\alpha+1}}$, and one can check that the conditions for Proposition 1 are met which in turn gives

$$c\left(\sigma^2\right)^{\frac{2\alpha}{2\alpha+1}} \leq \|\widehat{\theta} - \theta^*\|_2^2 \leq C\left(\sigma^2\right)^{\frac{2\alpha}{2\alpha+1}},$$

with probability $\geq 1 - \exp\left(-c'\sigma^{-\frac{2}{2\alpha+1}}\right)$ for some constants $C > c > 0$ and c'. One may notice that the rate $\left(\sigma^2\right)^{\frac{2\alpha}{2\alpha+1}}$ coincides with the minimax estimation rate for the α-smooth Sobolev function class. We show in Section 4.2 that this is indeed the case more generally.

τ-exponential decay Consider another case where the ellipse $\mathcal{E}$ is defined by the aspect ratios $\mu_j = c_1 \exp(-c_2 j^\tau)$, for some $\tau > 1/2$. Then a slight modification of the computation in Example 4 yields

$$k_*(\delta) = \operatorname{argmin}_k\{\sqrt{\mu_{k+1}}, \frac{9}{10}\delta\} \asymp \log^{\frac{1}{\tau}}\left(\frac{1}{\delta}\right).$$

In order to establish the regularity condition, notice that in this case, the quantity $\inf_{\gamma\in\Gamma(\theta^*,\delta)} \sum_{i=1}^d \gamma_i$ is achieved in limit by $\gamma_i = \mu_i \mathbf{1}\{i > k_*(\delta)\}$ and further more

$$\begin{aligned}\inf_{\gamma\in\Gamma(\theta^*,\delta)} \sum_{i=1}^d \gamma_i &= \sum_{j=k_*+1}^d c_1 e^{-c_2 j^\tau} \\ &\asymp \int_{k_*}^\infty e^{-c_2 t^\tau}\, dt \leq \frac{1}{\tau k_*^{\tau-1}} \int_{k_*^\tau}^\infty e^{-c_2 u}\, du \asymp \frac{\mu_{k_*}}{k_*^{\tau-1}} \leq \delta^2,\end{aligned} \tag{20}$$

which by definition, shows that $\mathcal{E}$ is regular at $\theta^* = 0$.

Solving the fixed point problem (17) yields $\delta_* \asymp \sigma \log^{\frac{1}{2\tau}}\left(\frac{1}{\sigma}\right)$ up to other polylogarithmic factors in σ. One can check that the conditions for Proposition 1 are met, so we have, up to polylogarithmic factors,

$$c\sigma^2 \log^{\frac{1}{\tau}}(\sigma^{-1}) \leq \|\widehat{\theta} - \theta^*\|_2^2 \leq C\sigma^2 \log^{\frac{1}{\tau}}(\sigma^{-1}),$$

with probability $\geq 1 - \exp^{-c' \log^{\frac{1}{\tau}}(1/\sigma)}$ for some constants $C > c > 0$ and c'.

4.1.2. Estimating at extremal vectors

In the previous section, we studied the adaptive estimation rate for $\theta^* = 0$. In this section, we study some non-zero cases of the vector θ^*. For concreteness and simplicity, we restrict our attention to vectors that are non-zero on some coordinate $s \in [d] = \{1, \ldots, d\}$, and zero on all other coordinates. Even for such simple vectors, our analysis reveals some interesting and adaptive scalings.

Given integer $s \in [d]$, consider $\theta^* := (\sqrt{\mu_s} - r)e_s$ for some $r \in [t_\ell^*(s, \mathcal{E}),\ t_u^*(s, \mathcal{E})]$ where $t_\ell^*(s, \mathcal{E}), t_u^*(s, \mathcal{E})$ are small constants that are defined in Wei and Wainwright [37, Corollary 2]. Note that the shrinkage $-r$ away from the boundary is due to the boundary issue in Theorem 2. We believe it is an artifact of our analysis that is possibly removable. We make a note that the extremal vectors considered here are sparse vectors, with most of the entries equal to zero. This sparsity arises because we have considered ellipses that are specified by diagonal quadratic forms; in the more general setting, the "easy" points would not correspond to such sparse vectors.

So as to streamline notation, we adopt $k_*(\delta)$ as a shorthand for $k_*(\theta^*, \delta)$. Wei and Wainwright [37, Appendix D] showed that with $\xi = (1 - \eta)\delta$, we have

$$k_*(\delta) = k_*\Big(\frac{\xi}{1-\eta}\Big) \leq \underbrace{\arg\max_{1 \leq k \leq d} \Big\{\mu_k^2 \geq \frac{1}{64}\xi^2 \mu_s\Big\}}_{=:\, m_u}.$$

This upper bound is proved by considering the projection onto the m_u-dimensional subspace spanned by $\{e_1, \ldots, e_{m_u}\}$. At the same time, we prove in Lemma 6 that

$$k_*(\delta) \geq 0.09 \cdot m_\ell, \qquad \text{where } m_\ell := \arg\max_{1 \leq k \leq d} \Big\{\mu_k^2 \geq \delta^2 \mu_s\Big\}.$$

α-polynomial decay Consider an ellipse $\mathcal{E}$ with $\mu_j = cj^{-2\alpha}$ for some $\alpha > 1/2$. From the above calculation, we can conclude that

$$m_u,\ m_\ell,\ k_* \asymp (\mu_s \delta^2)^{-\frac{1}{4\alpha}},$$

Let us verify the regularity condition (9) with dimension m_u and projection Π_{m_u} to the linear subspace spanned by the vectors $\{e_1, \ldots, e_{m_u}\}$. We make the observation that $\gamma = 4(0, \ldots, 0, \mu_{m_u+1}, \ldots \mu_d)$ is feasible for the set $\Gamma(\theta^*, \delta, \Pi_{m_u})$,

since $\Delta_i - (\Pi_k \Delta)_i$ equals to zero in the first m_u dimensions, equals to Δ_i for $i > m_u$, and further

$$\sum_{i=1}^{d} \frac{\Delta_i^2}{\gamma_i} = \sum_{i=m_u+1}^{d} \frac{(\theta_i - \theta^*)^2}{\gamma_i} \leq \sum_{i=m_u+1}^{d} \frac{2\theta_i^2 + 2\theta^{*2}}{\gamma_i} = \frac{1}{2}\Big(\sum_{i=m_u+1}^{d} \frac{2\theta_i^2}{4\mu_i} + \frac{2\theta^{*2}}{4\mu_i}\Big) \leq 1,$$

where the last step uses the fact that both θ^* and θ belong to $\mathcal{E}$. Therefore one has

$$\inf_{\gamma \in \Gamma(\theta^*, \delta, \Pi_{m_u})} \gamma_i \leq \sum_{i=m_u+1}^{d} \mu_i \asymp \int_{m_u+1}^{\infty} t^{-2\alpha} dt = (m_u+1)^{-2\alpha+1} \leq \delta^2 m_u,$$

where the last step follows from $m_u \asymp (\mu_s \delta^2)^{-\frac{1}{4\alpha}}$. Thereby we establish the regularity condition at θ^*.

As long as $s \lesssim (\sigma^2)^{-2/(4\alpha+1)}$, solving the fixed point problem (17) yields $\delta_* \asymp \sigma^{\frac{4\alpha}{4\alpha+1}}$, and one can check that the conditions for Proposition 1 are met. Thus,

$$c\left(\sigma^2\right)^{\frac{4\alpha}{4\alpha+1}} \leq \|\widehat{\theta} - \theta^*\|_2^2 \leq C\left(\sigma^2\right)^{\frac{4\alpha}{4\alpha+1}},$$

with probability $\geq 1 - \exp\Big(-c'\sigma^{-\frac{2}{4\alpha+1}}\Big)$ for some constants $C > c > 0$ and c'.

τ-exponential decay Now consider ellipse $\mathcal{E}$ with $\mu_j = c_1 \exp(-c_2 j^\tau)$ for some $\tau > 1/2$. From the above calculation, we can conclude that

$$m_u,\ m_\ell,\ k_* \asymp \log^{\frac{1}{\tau}}\left(\frac{1}{\delta}\right).$$

Let us verify the regularity condition (9) with dimension m_u and projection Π_{m_u} to the linear subspace spanned by $\{e_1, \ldots, e_{m_u}\}$. Since the vector $\gamma = 4(0, \ldots, 0, \mu_{m_u+1}, \ldots \mu_d)$ is feasible for the set $\Gamma(\theta^*, \delta, \Pi_{m_u})$, by similar calculation from inequality (20), we can show that the ellipse is regular at θ^*.

Solving the fixed point problem (17) yields $\delta_* \asymp \sigma \log^{\frac{1}{2\tau}}\left(\frac{1}{\sigma}\right)$ up to other polylogarithmic factors in σ. One can check that the conditions for Proposition 1 are met, so we have, up to polylogarithmic factors,

$$c\sigma^2 \log^{\frac{1}{\tau}}(\sigma^{-1}) \leq \|\widehat{\theta} - \theta^*\|_2^2 \leq C\sigma^2 \log^{\frac{1}{\tau}}(\sigma^{-1}),$$

with probability $\geq 1 - \exp\Big(-c' \log^{\frac{1}{\tau}}(\sigma^{-1})\Big)$ for some constants $C > c > 0$ and c'.

Numerical results To illustrate our findings from above, Figure 4 provides a numerical plot of the mean-squared error of the constrained least squared estimator (15) for estimating the vector $\theta^* = 0$ (blue curve) and the vector

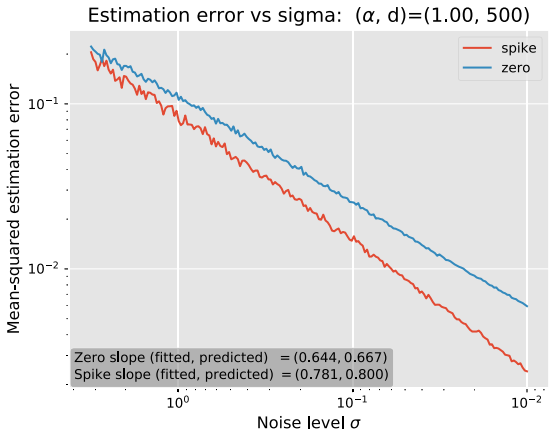


FIG 4. *Log-log plot of* $\mathbb{E}\|\widehat{\theta} - \theta^*\|_2^2$ *against* σ *for the ellipse with polynomial decay* $\mu_j = j^{-2}$ *in* $d = 500$ *dimensions. The blue curve is the case* $\theta^* = 0$*, and the red curve is* $\theta^* = e_1$*.*

$\theta^* = e_1$ (red curve). In each case, the plot shows the error decreases as a function of the inverse noise level $\frac{1}{\sigma^2}$.

The underlying ellipse is defined by the eigenvalues $\mu_j = j^{-2\alpha}$ with $\alpha = 1$. Consequently, the predicted scaling of the mean=squared error is $(\sigma^2)^{\frac{2\alpha}{2\alpha+1}}$ for the zero vector, and $(\sigma^2)^{\frac{4\alpha}{4\alpha+1}}$ for the "spiked" e_1 vector. Based on these predictions, our our theory suggests that on a log-log plot, the mean-squared error should decay at a linear rate with slopes $-2/3$ and $-4/5$ respectively, and indeed the empirical least-squares fit shown in Figure 4 matches this predicted behavior closely.

4.2. Minimax risk bounds

As another consequence of our main results, in this section, we show that the LSE is minimax optimal for ellipse estimation problem that is described above. Here the *minimax risk* over the ellipse $\mathcal{E}$ is defined as

$$\mathfrak{M}(\mathcal{E}) := \inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{E}} \mathbb{E}_\theta^* \|\widehat{\theta} - \theta^*\|_2^2,$$

where the supremum is taken over distributions $\mathcal{N}(\theta^*, \sigma^2 \mathbf{I}_n)$ indexed by $\theta^* \in \mathcal{E}$, and the infimum is taken over all estimators. By this criteria, estimators are compared on their worst-case performance.

In the following, we show that the minimax optimal risk is achieved by the LSE estimator and the risk is characterized through the solution to the fixed point problem (17). Let $\delta_*(0)$ be the solution to the fixed point problem (17) for $\theta^* = 0$.

Corollary 3. *There are universal constants $c, C > 0$, the global minimax risk of estimation over the entire ellipse $\mathcal{E}$ satisfies*

$$\mathfrak{M}(\mathcal{E}) \geq c\sigma^2 k_*(0, \delta_*). \tag{21a}$$

If furthermore the ellipse is regular (9) for all $\theta^* \in \mathcal{E}$, then we also have

$$\mathfrak{M}(\mathcal{E}) \leq C\sigma^2 k_*(0, \frac{1}{2}\delta_*). \tag{21b}$$

We prove this result in Section C.

In contrast to the minimax lower bound of Yang et al. [39], our minimax lower bound (21a) does not require the regularity assumption (9). See Section A for a discussion of how the notion of regularity of Yang et al. [39] is a special case of our notion. The lower bound is proved by showing that the ellipse contains a k_*-dimensional ball, and then applying the standard minimax bound in for estimation in a k_*-dimensional space.

On the other hand, the upper bound (21b) does require the regularity assumption, which allows us to apply Proposition 1. It implies that the risk of the LSE for each problem $\theta^* \in \mathcal{E}$ is upper bounded by $\lesssim \delta_*^2(\theta^*)$. Furthermore, we show that among all θ^*, the largest upper bound $\delta_*^2(\theta^*)$ is the case $\theta^* = 0$, which yields the upper bound in Corollary 3. Thus, the hardest problem for the LSE is estimating $\theta^* = 0$, and its risk there matches the lower bound. In short, the LSE is minimax optimal for ellipses that are regular.

5. Proofs

We now turn to the proofs of our main results, namely Theorem 1 and Theorem 2. The proofs of more technical results are deferred to appendices.

5.1. Proof of Theorem 1

For any dimension and projection pair $(k,\ \Pi_k)$, we can write

$$\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle w, \Delta \rangle \leq \underbrace{\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle w, \Pi_k \Delta \rangle}_{T_1} + \underbrace{\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle w, \Delta - \Pi_k \Delta \rangle}_{T_2}.$$

We now proceed to upper bound the two terms T_1 and T_2.

Bounding T_1 From standard properties of orthogonal projections onto subspaces, we have $\langle w - \Pi_k w, \Pi_k \Delta \rangle = 0$ for any w and Δ. By combining this fact with the Cauchy-Schwarz inequality, the term T_1 is upper bounded as

$$\begin{aligned} T_1 \ = \ \mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle w, \Pi_k \Delta \rangle &= \mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \langle \Pi_k w, \Pi_k \Delta \rangle \\ &\leq \mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \|\Pi_k w\|_2 \|\Pi_k \Delta\|_2. \end{aligned}$$

By the non-expansiveness of projection onto a subspace, we have $\|\Pi_k \Delta\|_2 \leq \|\Delta\|_2 \overset{(i)}{\leq} \delta$, where inequality (i) follows from the inclusion $\Delta \in \mathbb{B}(\delta)$. Thus, we have established that

$$T_1 \leq \delta \mathbb{E}\|\Pi_k w\|_2 \leq \delta\sqrt{k}, \tag{22a}$$

where the last step follows from first applying Jensen's inequality, and then noting that the distribution of $\Pi_k w$ is a k-dimensional standard Gaussian vector.

Bounding T_2 For a given vector $\gamma \in \Gamma(\theta^*, \delta)$, define the diagonal matrix $A := \operatorname{diag}(\sqrt{\gamma_1}, \ldots, \sqrt{\gamma_d})$. Noting that $\langle w, \Delta - \Pi_k \Delta \rangle = \langle Aw, A^{-1}(\Delta - \Pi_k \Delta) \rangle$ and then applying the Cauchy-Schwarz inequality, we find that

$$T_2 \leq \mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)} \|Aw\|_2 \, \|A^{-1}(\Delta - \Pi_k \Delta)\|_2.$$

By the definition of $\Gamma(\theta^*, \delta, \Pi_k)$, we must have $\|A^{-1}(\Delta - \Pi_k \Delta)\|_2 \leq 1$. Thus, we have the upper bound

$$T_2 \leq \mathbb{E}\|Aw\|_2 \leq \sqrt{\sum_{i=1}^{d} \gamma_i}, \tag{22b}$$

where the last step is due to Jensen's inequality. Since our choice of γ was arbitrary, we may add an infimum over $\gamma \in \Gamma(\theta^*, \delta, \Pi_k)$. Combining the two bounds (22a) and (22b) concludes the proof.

5.2. Proof of Theorem 2

As in the preceding proof, we adopt k_* as convenient shorthand for the quantity $k_*(\theta^*, \delta)$. We now divide our analysis into two cases, depending on whether or not $\|\theta^*\|_{\mathcal{E}} \leq 1/2$.

5.2.1. Case I

First, suppose that $\|\theta^*\|_{\mathcal{E}} \leq \frac{1}{2}$, which implies that $\Phi(\delta) \leq (\|\theta^*\|_{\mathcal{E}} - 1)^2 \leq 1$. Under this condition, Lemma 2 from the paper [37] guarantees that

$$\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) \leq \frac{3}{2} \min\left\{(1-\eta)\delta, \sqrt{\mu_{k+1}}\right\}.$$

By definition, the critical dimension $k_* := \arg\min_{k=1,\ldots,d} \{\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) \leq \frac{9}{10}\delta\}$ can be upper bounded as

$$k_* \leq \arg\min_{k=1,\ldots,d} \{\frac{3}{2}\sqrt{\mu_{k+1}} \leq \frac{9}{10}\delta\} =: k'_*, \tag{23}$$

where we have used the fact that $\frac{9}{10} \leq 1 - \eta$, and $\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta))$ is non-decreasing in k.

Let $E_{k'_*}$ denote the k'_*-dimensional subspace of vectors that are zero is their last $d - k'_*$ coordinates. Recalling that $\mathbb{S}(r)$ denotes a Euclidean sphere of radius r, we claim that

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \overset{(i)}{\geq} \mathscr{G}\Big(\mathbb{S}(\frac{3}{10}\delta) \cap E_{k'_*}\Big) \overset{(ii)}{=} \frac{3}{10}\delta\sqrt{k'_*}. \tag{24}$$

Taking this claim as given for the moment, combining it with the bounds $\|\theta^*\|_{\mathcal{E}} \leq 1/2$ and $k_* \leq k'_*$, we find that

$$\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \geq \frac{3}{10}\delta\sqrt{1 - \|\theta^*\|_{\mathcal{E}}^2}\sqrt{k_*(\theta^*, \delta)},$$

which completes the proof of Theorem 2 in this case.

Proof of inequality (24) In this proof, we adopt the convenient shorthand $b = 3/10$. Part (ii) of the inequality can be seen from the spherical example in the discussion of Theorem 1. It only remains to prove part (i). Let us first show that $\mathbb{S}(2b\delta) \cap E_{d-k'_*} \subset \mathcal{E}$. Recalling the definition of k'_* from equation (23), we have

$$\sum_{i=1}^{d} \frac{x_i^2}{\mu_i} = \sum_{i=1}^{k'_*} \frac{x_i^2}{\mu_i} \overset{(iii)}{\leq} \sum_{i=1}^{k'_*} \frac{x_i^2}{\mu_{k'_*}} = \frac{(2b\delta)^2}{\mu_{k'_*}} \overset{(iv)}{\leq} 1,$$

where inequality (iii) follows from the non-increasing order of μ_i and inequality (iv) follows from the definition of k'_*.

In order to establish the inclusion $\mathbb{B}_{k'_*}(b\delta) \subset \mathcal{E}_{\theta^*}$, we make use of the fact that $\|\theta^*\|_{\mathcal{E}} \leq 1/2$. Since $\|2\theta^*\|_{\mathcal{E}} \leq 1$, we have $2\theta^* \in \mathcal{E}$. For any $v \in \mathbb{S}_{k'_*}(b\delta)$, since $\mathbb{B}_{k'_*}(2b\delta) \subset \mathcal{E}$ we have $2v \in \mathcal{E}$. Combining these two facts together and the convexity of set $\mathcal{E}$, we have $v + \theta^* \in \mathcal{E}$. It further implies that $\mathbb{B}_{k'_*}(b\delta) \subset \mathcal{E}_{\theta^*}$ and finishes the proof of inequality (24).

5.2.2. Case II

Otherwise, we may assume that $\|\theta^*\|_{\mathcal{E}} > 1/2$, in which case $\Phi(\delta/c) \leq (\|\theta^*\|_{\mathcal{E}} - 1)^2 < 1$, and hence by definition of the function Φ, we have $\delta < c\|\theta^*\|_2/a$. For the remainder of the proof, we assume that $k_* \geq 160$. The case when $k_* < 160$ is addressed separately at the end of this proof.

The proof of Theorem 2 requires two auxiliary lemmas. The first is a packing lemma, proved in Wei and Wainwright [37, Lem. 4]. Here we state a slightly altered version of this claim that is better suited to our purposes. Let M denote the diagonal matrix with entries $1/\mu_1, \ldots, 1/\mu_d$, and adopt the shorthands $a := 1 - \eta$ and $b := \frac{3}{10}$ based on the definition of the critical dimension (8).

Lemma 1. *For any vector $\theta^* \in \mathcal{E}$ such that $\|\theta^*\|_2 > a\epsilon$, there exists a vector $\theta^\dagger \in \mathcal{E}$, a collection of d-dimensional orthonormal vectors $\{u_i\}_{i=1}^{k_*}$ and an upper triangular matrix of the form*

$$H := \begin{bmatrix} 1 & h_{3,2} & h_{4,2} & \cdots & h_{k_*,2} \\ & 1 & h_{4,3} & \cdots & h_{k_*,3} \\ & & 1 & \cdots & h_{k_*,4} \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix} \in \mathbb{R}^{k_*-1,k_*-1}$$

with ordered singular values $\nu_1 \geq \cdots \geq \nu_{k_*-1} \geq 0$ such that:

(a) The vectors u_1, $M\theta^\dagger$, and $\theta^\dagger - \theta^*$ are all scalar multiples of one another.
(b) We have $\|\theta^\dagger - \theta^*\|_2 = a\delta$.
(c) Letting $H_{\cdot,i}$ denote the ith column of H, for every $i \in [k_* - 1]$, the vector $\theta^\dagger \pm b\delta \underbrace{\begin{bmatrix} u_2 & \cdots u_{k_*} \end{bmatrix}}_{:=U} H_{\cdot,i}$ belongs to the ellipse $\mathcal{E}$.
(d) We have $\|\theta^\dagger\|_{\mathcal{E}} \leq \|\theta^*\|_{\mathcal{E}}$.
(e) For any integers $t_1 \in [k_* - 1]$, $t_2 \in [k_* - 2]$, we have

$$\nu_{t_1} \overset{(i)}{\leq} \frac{a}{3b}\sqrt{\frac{k_*-1}{t_1}}, \quad \text{and} \quad \nu_{t_2+1} \overset{(ii)}{\geq} 1 - \frac{t_2}{k_*-1} - \sqrt{\frac{a^2-9b^2}{9b^2}}.$$

Before proving Theorem 2, let us introduce some notation. Let H, U and $\theta^\dagger$ be as given in the Lemma 1 above and let $X := UH$ have columns $x_1, \ldots, x_{k_*-1}$ Let V be the matrix of right singular vectors of H so that $H^\top H = V\Sigma^2 V^\top$, where Σ^2 is diagonal with the squared singular values $\nu_1^2 \geq \cdots \geq \nu_{k_*-1}^2$ of H in order.

Let $m_1 := \lfloor (k_* - 1)/8 \rfloor$ and $m_2 := \lfloor (k_* - 1)/4 \rfloor$, and define the sparsity level $s := \rho^{\frac{k_*-1}{16}}$ for some constant[2] $\rho \in (0, 1)$. For a given s-sized subset S of $\{m_1, \ldots, m_2\}$, any vector of the form $z^S = (z_1^S, \ldots, z_{k_*-1}^S) \in \{-1, 0, 1\}^{k_*-1}$ with zeros in all positions not indexed by S is called as an *S-valid sign vector*. Any such sign vector can be used to define the perturbed vector

$$\theta^S := \theta^\dagger + b\delta \frac{1}{\sqrt{32s}} UHVz^S \tag{25}$$

The following lemma guarantees the existence of a large collection $\mathcal{T}$ of s-sized subsets of $\{m_1, \ldots, m_2\}$ such that the collection $\{\theta^S : S \in \mathcal{T}\}$ has certain desirable properties.

Lemma 2. *There exists a collection $\mathcal{T}$ of s-sized subsets of $\{m_1, \ldots, m_2\}$ such that:*

(a) The collection $\mathcal{T}$ has cardinality at least $\binom{\lfloor \frac{1}{16}(k_*-1) \rfloor}{s}$.

[2]The arguments that follow do not depend on the specific choice of ρ, and taking $\rho = 1/2$ suffices. However in the proof of Corollary 2, we re-use these arguments for a different value of ρ.

(b) For each $S \in \mathscr{T}$, there is a S-valid sign vector z^S such that the associated perturbation θ^S belongs to the ellipse $\mathcal{E}$, and moreover satisfies the bounds:

$$\delta^2 \overset{(i)}{\leq} \|\theta^S - \theta^*\|_2^2 \overset{(ii)}{\leq} \frac{4}{1 - \|\theta^*\|_{\mathcal{E}}^2}\, \delta^2. \tag{26}$$

See Section E.1 for the proof of this lemma.

Turning back to the proof of Theorem 2, consider those perturbation vectors (25) that are defined via Lemma 2. For each $S \in \mathscr{T}$, we define the vectors

$$\widetilde{\Delta}^S := \theta^S - \theta^*, \quad \text{and} \quad \Delta^S := \frac{\delta}{\|\widetilde{\Delta}^S\|_2} \widetilde{\Delta}^S.$$

Inequality (26) implies that $\frac{\delta}{\|\widetilde{\Delta}^S\|_2} \leq 1$. By the convexity of the set $\mathcal{E}_{\theta^*}$, we have $\Delta^S \in \mathcal{E}_{\theta^*} \cap \mathbb{S}(\delta)$ for each $S \in \mathscr{T}$. By restricting the supremum to a smaller subset, we obtain the lower bound

$$\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{S}(\delta)} \langle w, \Delta \rangle \geq \mathbb{E} \max_{S \in \mathscr{T}} \langle w, \Delta^S \rangle.$$

Re-writing the definition (25) in the form $\theta^S = \theta^\dagger + \frac{b\delta}{\sqrt{32s}} UHVz^S$, it follows that

$$\Delta^S := \frac{\delta}{\|\widetilde{\Delta}^S\|_2} \widetilde{\Delta}^S = \delta \left(\frac{1}{\|\widetilde{\Delta}^S\|_2} (\theta^\dagger - \theta^*) + \frac{b\delta}{\sqrt{32s}\|\widetilde{\Delta}^S\|_2} UHVz^S \right),$$

which further guarantees that

$$\begin{aligned} \mathbb{E} \max_{S \in \mathscr{T}} \langle w, \Delta^S \rangle &\geq \delta \mathbb{E} \max_{S \in \mathscr{T}} \langle w, \frac{b\delta}{\sqrt{32s}\|\widetilde{\Delta}^S\|_2} UHVz^S \rangle + \delta \mathbb{E} \max_{S \in \mathscr{T}} \langle w, \frac{1}{\|\widetilde{\Delta}^S\|_2} (\theta^\dagger - \theta^*) \rangle \\ &= \mathbb{E} \max_{S \in \mathscr{T}} \langle w, \frac{b\delta}{\sqrt{32s}\|\widetilde{\Delta}^S\|_2} UHVz^S \rangle, \end{aligned}$$

where the second equality follows since $\mathbb{E}\langle w, \theta^\dagger - \theta^* \rangle = 0$. The right-hand side is non-negative, since for any fixed choice of $S_0 \in \mathscr{T}$, we have

$$\mathbb{E} \max_{S \in \mathscr{T}} \langle w, \frac{1}{\|\widetilde{\Delta}^S\|_2} UHVz^S \rangle \geq \mathbb{E} \langle w, \frac{1}{\|\widetilde{\Delta}^{S_0}\|_2} UHVz^{S_0} \rangle = 0.$$

Noting that inequality (26)(ii) can be rewritten as $\|\widetilde{\Delta}^S\|_2^2 \leq \frac{4}{1 - \|\theta^*\|_{\mathcal{E}}^2} \delta^2$, we find that

$$\mathbb{E} \max_{S \in \mathscr{T}} \langle w, \frac{1}{\|\widetilde{\Delta}^S\|_2} UHVz^S \rangle \geq \sqrt{\frac{1 - \|\theta^*\|_{\mathcal{E}}^2}{4\delta^2}}\, \mathbb{E} \max_{S \in \mathscr{T}} \langle w, UHVz^S \rangle.$$

Putting together the pieces, we have established that

$$\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{S}(\delta)} \langle w, \Delta \rangle \geq \frac{b\delta}{16} \sqrt{\frac{1 - \|\theta^*\|_{\mathcal{E}}^2}{s}}\, \mathbb{E} \max_{S \in \mathscr{T}} \langle w, UHVz^S \rangle. \tag{27}$$

Our next step is to lower bound the expected maximum on the RHS, and to this end, we state an auxiliary result:

Lemma 3. *Under the conditions of Theorem 2, we have*

$$\mathbb{E}\max_{S\in\mathscr{T}}\langle w, UHVz^S\rangle \geq \frac{1}{4}\mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i\Big]. \tag{28}$$

See Section E.2 for the proof of this lemma.

Let us now control the term on the right-hand side of inequality (28). Let $\mathcal{A}$ be the event that there are least s positive elements among the i.i.d. standard Gaussian random variables $\{w_i\}_{i=m_1}^{m_2}$. By the law of total expectation, we have

$$\mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i\Big] = \underbrace{\mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \;\mid\; \mathcal{A}\Big]}_{T_1}\,\mathbb{P}[\mathcal{A}] + \underbrace{\mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \;\mid\; \mathcal{A}^c\Big]}_{T_2}\,\mathbb{P}[\mathcal{A}^c].$$

Beginning our analysis with T_1, under the event $\mathcal{A}$, there exists some (random) subset $S' \in \mathscr{T}$ of cardinality $|S'| \geq s$ such that $w_i > 0$ for all $i \in S'$. (When there are multiple such sets, we choose one of them uniformly at random.) In terms of this set, we have

$$T_1 = \mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \;\mid\; \mathcal{A}\Big] \geq \mathbb{E}_{w,S'}\Big[\sum_{i\in S'} w_i \;\mid\; \mathcal{A}\Big] = \sum_{S'}\mathbb{E}_w\Big[\sum_{i\in S'} w_i \mid S'\Big]\,\mathbb{P}[S' \mid \mathcal{A}],$$

where $\mathbb{P}[S' \mid \mathcal{A}]$ denotes the conditional probability of the randomly chosen S' given that $\mathcal{A}$ holds. Since we are conditioning on a random set S' on which each w_i is positive, we have

$$\begin{aligned}\mathbb{E}_w\Big[\sum_{i\in S'} w_i \mid S'\Big] &= \mathbb{E}_w\Big[\sum_{i\in S'} w_i \mid w_i > 0\Big] \\ &\geq s\,\mathbb{E}[w_i \mid w_i > 0] \;=\; s\sqrt{2/\pi}.\end{aligned}$$

Since $\sum_{S'}\mathbb{P}[S' \mid \mathcal{A}] = 1$, we have proved that $T_1 \geq s\sqrt{2/\pi}$.

Turning to the term T_2, we begin by observing that for any fixed $S_0 \in \mathscr{T}$, we have

$$\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \geq \sum_{i\in S_0} w_i \geq -\sum_{i\in S_0}|w_i|.$$

Using this observation we can conclude that

$$\mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \;\mid\; \mathcal{A}^c\Big] \geq \mathbb{E}\Big[-\sum_{i\in S_0}|w_i| \;\mid\; \mathcal{A}^c\Big] \overset{(i)}{=} \mathbb{E}\Big[-\sum_{i\in S_0}|w_i|\Big] = -s\sqrt{2/\pi}.$$

where (i) follows from the fact $\mathcal{A}^c$ only depends on the sign of w_i and the distribution of $|w_i|$ is independent of $\mathcal{A}^c$. Combining these two lower bounds, we find that

$$\mathbb{E}\max_{S\in\mathscr{T}}\sum_{i\in S} w_i \geq s\sqrt{2/\pi}(1 - 2\mathbb{P}[\mathcal{A}^c]).$$

We now bound the probability of event $\mathcal{A}^c$. Recall that event $\mathcal{A}$ holds if and only if there are at least s positive elements among the i.i.d. standard Gaussian random variables $\{w_i\}_{i=m_1}^{m_2}$. Since $s := \lfloor (m_2 - m_1)/4 \rfloor$, with probability no larger than $\exp(-(m_2 - m_1)D(\frac{1}{4} \,\|\, \frac{1}{2})) \leq e^{-0.1(m_2 - m_1)}$, there are more than $(m_2 - m_1)/4$ components among $w_{m_1}, \ldots, w_{m_2}$ that are positive, meaning that $\mathbb{P}[\mathcal{A}^c] \leq e^{-0.1(m_2 - m_1)}$. Thus, we have the lower bound

$$\mathbb{E}\Big[\max_{S \in \mathscr{T}} \sum_{i \in S} w_i \geq s\sqrt{2/\pi}(1 - 2e^{-0.1(m_2 - m_1)})\Big] \geq \frac{1}{5}s,$$

where the last step uses the fact that $m_2 - m_1 \geq k_*/16 > 10$.

Combining this last bound with inequalities (27) and (28) yields

$$\begin{aligned}
\mathbb{E} \sup_{\Delta \in \mathcal{E}_{\theta^*} \cap \mathbb{S}(\delta)} \langle w, \Delta \rangle &\geq \frac{b\delta}{16}\sqrt{\frac{1 - \|\theta^*\|_{\mathcal{E}}^2}{s}}\ \mathbb{E} \max_{S \in \mathscr{T}} \langle w, UHVz^S \rangle \\
&\geq \frac{b}{64}\delta\sqrt{\frac{1 - \|\theta^*\|_{\mathcal{E}}^2}{s}}\ \mathbb{E}\Big[\max_{S \in \mathscr{T}} \sum_{i \in S} w_i\Big] \\
&\geq \frac{b}{320}\delta\sqrt{(1 - \|\theta^*\|_{\mathcal{E}}^2)s} \\
&\geq c'\sqrt{1 - \|\theta^*\|_{\mathcal{E}}^2} \cdot \delta\sqrt{k_*},
\end{aligned}$$

where the last step uses the fact that $s = \rho\frac{k_* - 1}{16}$.

In order to finish the proof, we deal with the case of $k_* < 160$ separately. According to part (b) of Lemma 1, if we denote $v_1 := \theta^* - \theta^\dagger$, then $\theta^\dagger \in \mathcal{E}$ and $\|v_1\|_2 = a\delta$. It is also shown in the proof of Wei and Wainwright [37, Lem. 5] that $\theta^* + v_1 \in \mathcal{E}$. Therefore the two points $\pm v_1$ are both contained in $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$ for a sufficiently small δ. As a result, we have $\mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{S}(\delta)) \geq \mathscr{G}(\{\pm v_1\}) = a\delta\sqrt{2/\pi}$, which establishes the lower bound in Theorem 2 with constant $c' = \frac{a}{4\sqrt{5\pi}}$.

6. Discussion

In this paper, we studied the behavior of localized Gaussian widths over ellipses. These localized widths are known to play a fundamental role in controlling the difficulty of associated testing and estimation problems. Despite its fundamental importance, the localized Gaussian width is hard to compute in general. The main contribution of our paper was to show how the localized Gaussian width can be bounded, both from above and below, via the localized Kolmogorov dimension. These Kolmogorov dimensions can be computed in many interesting cases, which leads to an explicit characterization of the estimation error of least-squares regression as a function of the true regression vector within the ellipse. We used this characterization to show how the difficulty of estimating a vector θ^* within the ellipse can vary dramatically as a function of the location of θ^*. Estimating the all-zeros vector ($\theta^* = 0$) is always the hardest sub-problem, and

leads to the global minimax rate. Much faster rates of estimation can be obtained for vectors located near "narrower" portions of the ellipse boundary. While much of the analysis in this paper is specific to ellipses, we do anticipate that the general procedure of moving from Gaussian width to the Kolmogorov width could be useful in studying adaptivity and local geometry in other estimation problems.

References

[1] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337–404. MR0051437

[2] Balakrishnan, S. and L. Wasserman (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics 47*(4), 1893–1927. MR3953439

[3] Bartlett, P. L., O. Bousquet, S. Mendelson, et al. (2005). Local Rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537. MR2166554

[4] Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford, UK: Oxford University Press. MR3185193

[5] Chatterjee, S. (2014). A new perspective on least squares under convex constraint. *The Annals of Statistics 42*(6), 2340–2381. MR3269982

[6] Chatterjee, S., A. Guntuboyina, B. Sen, et al. (2015). On risk bounds in isotonic and other shape restricted regression problems. *The Annals of Statistics 43*(4), 1774–1800. MR3357878

[7] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory 52*(4), 1289–1306. MR2241189

[8] Donoho, D. L. and I. Johnstone (1995, December). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90(432)*, 1200–1224. MR1379464

[9] Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81*(3), 425–455. MR1311089

[10] Donoho, D. L., R. C. Liu, and B. MacGibbon (1990). Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, 1416–1437. MR1062717

[11] Dudley, R. M. (1967). The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. *J. Functional Analysis 1*, 290–330. MR0220340

[12] Gu, C. (2002). *Smoothing spline ANOVA models*. Springer Series in Statistics. New York, NY: Springer. MR1876599

[13] Hasminskii, R., I. Ibragimov, et al. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *The Annals of Statistics 18*(3), 999–1010. MR1062695

[14] Ibragimov, I. A. and R. Z. Khasminskii (1978). Asymptotic properties of some nonparametric estimaters in a Gaussian white noise. In: *Proc 3rd Summer School on Probab, Theory and Math. Stat.*, 31–64. MR0586028

[15] Ibragimov, I. A. and R. Z. Khasminskii (2013). *Statistical Estimation: Asymptotic Theory*, Volume 16. Springer Science & Business Media. MR0620321

[16] Javanmard, A. and L. Zhang (2012). The minimax risk of truncated series estimators for symmetric convex polytopes. In *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 1633–1637. IEEE.

[17] Juditsky, A., A. Nemirovski, et al. (2018). Near-optimality of linear recovery in gaussian observation scheme under $\|\cdot\|_2^2$-loss. *The Annals of Statistics 46*(4), 1603–1629. MR3819111

[18] Kimeldorf, G. and G. Wahba (1971). Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl. 33*, 82–95. MR0290013

[19] Koltchinski, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics 34*(6), 2593–2656. MR2329442

[20] Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory 47*(5), 1902–1914. MR1842526

[21] Ledoux, M. and M. Talagrand (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag. MR1102015

[22] Massart, P. (2003). *Concentration Inequalties and Model Selection*. Ecole d'Eté de Probabilités, Saint-Flour. New York: Springer. MR2319879

[23] Meyer, M. and M. Woodroofe (2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*, 1083–1104. MR1810920

[24] Pinkus, A. (2012). *N-Widths in Approximation Theory*. New York: Springer-Verlag. MR0774404

[25] Pinsker, M. S. (1980). Optimal filtering of square-integrable signals in Gaussian noise. *Problemy Peredachi Informatsii 16*(2), 52–68. MR0624591

[26] Pisier, G. (1986). Probabilistic methods in the geometry of Banach spaces. In *Probability and analysis*, pp. 167–241. Springer. MR0864714

[27] Talagrand, M. (2000). *The Generic Chaining*. New York, NY: Springer-Verlag. MR2133757

[28] Talagrand, M. (2014). *Upper and lower bounds for stochastic processes: modern methods and classical problems*, Volume 60. Springer Science & Business Media. MR3184689

[29] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer. MR2724359

[30] Valiant, G. and P. Valiant (2014). An automatic inequality prover and instance optimal identity testing. In *IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 51–60. IEEE. MR3344854

[31] van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

[32] van der Vaart, A. W. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. New York, NY: Springer-Verlag. MR1385671

[33] van Handel, R. (2018a). Chaining, interpolation, and convexity. *J. Eur. Math. Soc. (JEMS) 20*(10), 2413–2435. MR3852183

[34] van Handel, R. (2018b). Chaining, interpolation and convexity II: The con-

traction principle. *Ann. Probab. 46*(3), 1764–1805. MR3785599
[35] Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PN: SIAM. MR1045442
[36] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press. MR3967104
[37] Wei, Y. and M. J. Wainwright (2020). The local geometry of testing in ellipses: Tight control via localized kolmogorov widths. *IEEE Transactions on Information Theory*.
[38] Wei, Y., M. J. Wainwright, A. Guntuboyina, et al. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *The Annals of Statistics 47*(2), 994–1024. MR3909958
[39] Yang, Y., M. Pilanci, and M. J. Wainwright (2017). Randomized sketches for kernels: Fast and optimal non-parametric regression. *The Annals of Statistics, 2017 45*(3), 991–1023. MR3662446
[40] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *The Annals of Statistics 30*(2), 528–555. MR1902898
[41] Celentano, M., A. Montanari, and Y. Wei (2020). The Lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint* arXiv:2007.13716.

Apendix

This appendix is organized as follows. We first provide more explanation to our ellipse regularity by relating with the kernel regularity concept defined in Yang et al. [39] in Section A. It is then followed by the proof of Corollary 2 in Section B and the proof of Corollary 3 in Section C. We provide the proof of Proposition 1 in Section D and the details needed to establish Theorem 2 in Section E. A number of auxiliary results that are used for proving our main results are collected in Section F. Finally, for completeness, the well-definedness of the function Φ is provided in Section G.

Appendix A: Properties of kernel regularity

In this section, we relate our definition of regularity (9) to a concept introduced in previous work by Yang et al. [39]. In the context of kernel ridge regression, they defined the quantity

$$\widetilde{k}_* \equiv \widetilde{k}_*(\delta) := \arg\min_k \{\mu_{k+1} \leq \delta^2\} \tag{29}$$

with the convention $\widetilde{k}_* = d$ if the minimization is over an empty set. They said that an ellipse is *regular* if

$$\sum_{j=\widetilde{k}_*+1}^{d} \mu_j \leq c\widetilde{k}_*\delta^2, \qquad \text{for all } \delta > 0, \tag{30}$$

where $c > 0$ is some universal constant that does not depend on d. They used this property to prove a minimax lower bound on the prediction error for kernel ridge regression.

Let us now show that our regularity assumption (9) is a generalization of the condition (30), in that it reduces to it in the special case $\theta^* = 0$. In order to establish this claim, we begin by observing that for any $k \in \{1, \ldots, d-1\}$, we have $\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) = \min\{\mu_{k+1}^{1/2}, (1-\eta)\delta\}$ because the minimization in the definition (5) is achieved by the projection onto the subspace $\operatorname{span}\{e_1, \ldots, e_k\}$, and the maximization is achieved by $\theta = \min\{\mu_{k+1}^{1/2}, (1-\eta)\delta\} e_{k+1}$. On the other hand, for $k = d$ we have $\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) = 0$. Putting these two together gives

$$k_*(0, \delta) = \min\left\{k \mid \mu_{k+1} \leq \frac{81}{100}\delta^2\right\},$$

(with the convention $k_* = d$ if the minimum is over an empty set). Thus, we have recovered definition (29) up to a constant factor in δ.

Since the optimal projection Π_{k_*} is the projection onto the linear subspace $\operatorname{span}\{e_1, \ldots, e_{k_*}\}$, we can consider a sequence of positive vectors approaching $\gamma := (\mu_i \mathbf{1}\{i > k\})_{i=1}^d$ to obtain

$$\inf_{\gamma \in \Gamma(\theta^*, \delta)} \sum_{i=1}^d \gamma_i \leq \sum_{i=k_*+1}^d \mu_i.$$

Consequently, our regularity condition (9) holds as long as $\sum_{i=k_*+1}^d \mu_i \leq c k_* \delta^2$. Thus, it matches the notion of regularity (30) considered in Yang et al. [39].

Appendix B: Proof of Corollary 2

Throughout this proof, we use c, c', c'' etc. to denote universal constants that do not depend on any problem parameters such as δ, μ_i and θ^* and their values can vary from line to line.

The proof of inequality (i) in equation (14) is straightforward. By combining the Sudakov minoration (4) with our upper bound (13) on the localized Gaussian width, we find that

$$c' \delta \sqrt{\log M(\delta/2, \mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta))} \leq \mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)) \leq c_u \delta \sqrt{k_*(\theta^*, \delta)}.$$

Thus, we have proved inequality (i) in equation (14).

We now turn to the proof the second inequality (ii). It is convenient to divide our analysis into two cases depending on whether or not $\|\theta^*\|_{\mathcal{E}} \leq \frac{1}{2}$.

Case 1 $\|\theta^*\|_{\mathcal{E}} \leq \frac{1}{2}$. As shown earlier in equation (24) from the proof of Theorem 2, the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$ contains the k'_*-dimensional sphere $\mathbb{S}(\frac{3}{10}\delta) \cap E_{k'_*}$. Thus, by a standard volume argument [26, 36], it must have log packing number bounded from below by $c k'_* \log \frac{1}{\delta}$. This quantity is lower bounded by k_* up to some universal constant, which establishes inequality (ii) in this case.

Case 2 $\|\theta^*\|_{\mathcal{E}} > \frac{1}{2}$. We follow the notation from Section 5. In the proof of Theorem 2 (in particular, see equation (25) and Lemma 2), we constructed a set of vectors θ^S that after rescaling, all lie in our set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$. Each such vector θ^S is formed by taking a certain point $\theta^\dagger$ near θ^*, and adding certain combinations of orthogonal vectors u_i. We argue here that there is a subset of these scaled vectors of size $\gtrsim k_*$ that are pairwise separated from each other by a distance $\gtrsim \delta$.

We are only interested in proving bounds up to constant factors, meaning that we may assume without loss of generality that $k_* \geq 32 \times 10^4$; otherwise the result (14) holds immediately with a sufficiently large choice of c'.

Recall the earlier definition $s := \rho^{\frac{k_*-1}{16}}$ for a fixed constant $\rho \in (0,1)$; for this argument, we take $\rho = 10^{-4}$. By Lemma 4.10 in Massart [22], we can find a subset of s-sparse vectors contained in the binary hypercube $\{0,1\}^{\frac{1}{16}(k_*-1)}$ with log cardinality at least

$$s \log \frac{\frac{1}{16}(k_*-1)}{s} \gtrsim k_*,$$

and such that any pair of distinct elements differs in at least $(2-2\rho)s$ entries. Transferring this result to the context of Lemma 2, we are guaranteed a collection of vectors of log cardinality $\gtrsim k_*$ such that

$$\|z^S - z^{S'}\|_2^2 > (2-2\rho)s$$

for $z^S \neq z^{S'}$ in our packing.

Recalling that $V^\top H^\top H V = \Sigma^2$ and the definition (25) of θ^S, we then have

$$\begin{aligned}\|\theta^S - \theta^{S'}\|_2^2 &= \frac{b^2\delta^2}{32s}\|UHV(z^S - z^{S'})\|_2^2 \\ &= \frac{b^2\delta^2}{32s}(z^S - z^{S'})^\top \Sigma^2 (z^S - z^{S'}).\end{aligned}$$

Since z^S and $z^{S'}$ are zero in their first $m_1 - 1$ components, we can use inequality (ii) from Lemma 1 to bound the relevant diagonal entries of Σ. Doing so yields

$$\|\theta^S - \theta^{S'}\|_2^2 \geq \frac{b^2\delta^2}{32s}\left(1 - \frac{1}{8} - \sqrt{\frac{a^2-9b^2}{9b^2}}\right)^2 \|z^S - z^{S'}\|_2^2.$$

Thus, we have obtained a collection of vectors θ^S, indexed by subsets S, such that $\|\theta^S - \theta^{S'}\|_2 \gtrsim \delta$ for $S \neq S'$.

Finally, we need to show that after shrinking these θ^S toward θ^* and recentering, we obtain a packing of $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$. For each S recall the definitions $\widetilde{\Delta}^S := \theta^S - \theta^*$ and $\Delta^S := \frac{\delta}{\|\widetilde{\Delta}^S\|_2}\widetilde{\Delta}^S$. From discussion below Lemma 2, we have already showed that each vector Δ^S lies in $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$; it only remains to verify that distinct pairs are well-separated.

First, direct computation yields

$$\|\Delta^S - \Delta^{S'}\|_2^2 = 2\delta^2\Big(1 - \frac{\langle \widetilde{\Delta}^S, \widetilde{\Delta}^{S'}\rangle}{\|\widetilde{\Delta}^S\|_2\|\widetilde{\Delta}^{S'}\|_2}\Big). \tag{31}$$

In order to show that the right-hand side is lower bounded by a constant multiple of δ^2, it suffices to upper bound the inner product term. Using the fact that $\theta^\dagger - \theta^*$ has norm $a\delta$ and is orthogonal to the columns of U (see Lemma 1), we have

$$\begin{aligned}\langle \widetilde{\Delta}^S, \widetilde{\Delta}^{S'}\rangle &= \langle \theta^\dagger - \theta^* + \frac{b\delta}{\sqrt{32s}}UHVz^S, \theta^\dagger - \theta^* + \frac{b\delta}{\sqrt{32s}}UHVz^{S'}\rangle \\ &= a^2\delta^2 + \frac{b^2\delta^2}{32s}z^S\Sigma^2 z^{S'}.\end{aligned}$$

If $z^S \neq z^{S'}$ are from our packing, then by construction they differ on at least $(2-2\rho)s$ components, so they must agree on at most ρs components. Applying the inequality (i) from Lemma 1 to bound the relevant entries of Σ^2, we can continue from above to obtain

$$\begin{aligned}\langle \widetilde{\Delta}^S, \widetilde{\Delta}^{S'}\rangle &\leq a^2\delta^2 + \frac{b^2\delta^2}{32}\cdot 8(\frac{a}{3b})^2\cdot\rho \\ &\leq a^2\Big(1 + \frac{\rho}{36}\Big)\delta^2 < \delta^2.\end{aligned}$$

The last inequality follows from our earlier choice of $a := 1 - 10^{-5}$ and $\rho := 10^{-4}$. Dividing both sides by $\|\widetilde{\Delta}^S\|_2\|\widetilde{\Delta}^{S'}\|_2 \geq \delta^2$ (where this inequality follows from Lemma 2), we can continue from our earlier step (31) to obtain

$$\|\Delta^S - \Delta^{S'}\|_2^2 \geq c\delta^2.$$

Putting together the pieces, we have exhibited the claimed packing of $\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)$ of log cardinality $\gtrsim k_*$ and packing radius $\gtrsim \delta$.

Appendix C: Proof of Corollary 3

We divide our proof into two parts, corresponding to the upper and lower bounds respectively.

Upper bound Let us start with the proof of the upper bound. Under the regularity assumption, we may apply Proposition 1 to bound the mean-squared error $\mathbb{E}_{\theta^*}\|\widehat{\theta} - \theta^*\|_2^2$ of the LSE; in particular, it is upper bounded by $\delta_*^2(\theta^*)$ up to an universal constant. (Recall that $\delta_*(\theta^*)$ is the solution to the fixed point equation (17).)

In order to arrive at the desired minimax upper bound, we need to show that the function $\theta^* \mapsto \delta_*(\theta^*)$ is maximized at $\theta^* = 0$. Since k_* is a non-increasing function of δ (see the paper Wei and Wainwright [37, Appendix E]), a larger

$k_*(\theta^*)$ corresponds to a larger value of $\delta_*(\theta^*)$. These two quantities are related via the equation

$$\delta_*(\theta^*) = c_\ell \sigma \sqrt{k_*(\theta^*, \delta_*)}.$$

The following lemma bounds the supremum of k_*.

Lemma 4. *The critical dimensions at any θ^* can be controlled as*

$$k_*(\theta^*, \delta) \leq k_*(0, \frac{1}{2}\delta) + 1 \qquad \text{for all } \delta \in \Big(0,\ \Phi^{-1}((\|\theta^*\|_{\mathcal{E}}^{-1} - 1)^2) \wedge \sqrt{\mu_1}\Big).$$

The proof of this lemma is given in Section F.2. Note that it implies the claimed upper bound upper bound (21b).

Lower bound By definition, the minimax risk decreases when the supremum is taken over a smaller subset. In order to establish the lower bound, we restrict the supremum to a ball around zero. Recall our calculations from Example 4, where we showed that the Kolmogorov width of a local ball around $\theta^* = 0$ is given by

$$\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)) = \min\big\{\sqrt{\mu_{k+1}}, (1-\eta)\delta\big\}.$$

The corresponding $k_*(0, \delta)$ is given by

$$k_*(0, \delta) := \arg\min_{k=1,\ldots,d} \Big\{\mathscr{W}_k\Big(\mathcal{E} \cap \mathbb{B}\big((1-\eta)\delta\big)\Big) \leq \frac{9}{10}\delta\Big\}.$$

By inspection, we have the upper bound $k_*(0, \delta) = \arg\min_{k=1,\ldots,d}\big\{\sqrt{\mu_{k+1}} \leq \frac{9}{10}\delta\Big\}$. We also have the lower bound $\sqrt{\mu_{k_*(0,\delta)}} \geq \frac{9}{10}\delta$ for every $\delta \leq \sqrt{\mu_1}$. Note that the ellipse $\mathcal{E}$ always contains a k-dimensional ball centered at zero with radius $\sqrt{\mu_k}$. Combined with the bounds just stated, for every $\delta \in (0, \sqrt{\mu_1}]$, the ellipse also contains a ball of radius $\frac{9}{10}\delta$ centered at zero of dimension $k_*(0, \delta)$.

Now we are ready to control the minimax risk. First notice that

$$\mathfrak{M}(\mathcal{E}) := \inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{E}} \mathbb{E}_{\theta^*}\|\widehat{\theta} - \theta^*\|_2^2 \geq \inf_{\widehat{\theta}} \sup_{\theta^* \in \mathbb{B}(\frac{9}{10}\delta) \cap E_{k_*(0,\delta)}} \mathbb{E}_\theta\|\widehat{\theta} - \theta^*\|_2^2, \tag{32}$$

where recall that E_m denotes the space which contains d-dimensional vectors with their last $d - m$ coordinates all equal to zero.

By standard results (e.g., see the book [36]), estimating a m-dimensional vector in a r radius ball has minimax risk lower bounded as

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathbb{B}(r) \cap E_m} \mathbb{E}_\theta\|\widehat{\theta} - \theta^*\|_2^2 \gtrsim \min\{r^2, m\sigma^2\}.$$

Substituting this lower bound into inequality (32), we find that

$$\mathfrak{M}(\mathcal{E}) \gtrsim \min\{\big(\frac{9}{10}\delta\big)^2, k_*(0, \delta)\sigma^2\}, \tag{33}$$

for each $\delta \leq \sqrt{\mu_1}$. From the definition (17), we have $\delta_*(0) = c_\ell \sigma \sqrt{k_*(0, \delta_*)}$. Taking $\delta = \delta_*(0)$ in inequality (33) yields the claimed lower bound (21a).

Appendix D: Proof of Proposition 1

This section is devoted to the proof of Proposition 1.

D.1. Reduction to bounding localized Gaussian width

Chatterjee [5] provided one way of obtaining upper and lower bounds on the error $\|\widehat{\theta} - \theta^*\|_2$ of the least squares estimator for a general convex set, under the Gaussian sequence model (2). Define the function

$$g(t) := \frac{\delta^2}{2} - \sigma \mathscr{G}(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\delta)), \tag{34}$$

which can be shown to be strongly convex on $(0, \infty)$ with a unique minimizer $\delta_0 > 0$. Then:

Theorem 3 ([5, Thm. 1.1, Cor. 1.2])**.** *The least squares estimator $\widehat{\theta}$ satisfies*

$$\big|\|\widehat{\theta} - \theta^*\|_2 - \delta_0\big| \leq t\sqrt{\delta_0}, \quad \text{w.p. } \geq 1 - 3\exp\left(-\frac{t^4}{32\sigma^2(1 + t/\sqrt{\delta_0})^2}\right),$$

for any $t > 0$. Furthermore, there is a universal constant $C > 0$ such that

$$\big|\mathbb{E}\|\widehat{\theta} - \theta^*\|_2^2 - \delta_0^2\big| \leq C\delta_0^{3/2}\sigma^{1/2}, \qquad \text{if } \delta_0 \geq \sigma. \tag{35}$$

In particular, if we take $t = c\sqrt{\delta_0}$, it is guaranteed that

$$\big|\|\widehat{\theta} - \theta^*\|_2 - \delta_0\big| \leq c\delta_0, \quad \text{w.p. } \geq 1 - 3\exp(-c'\delta_0^2/\sigma^2). \tag{36}$$

The following simple lemma shows how sandwiching g between two functions allows us to obtain upper and lower bounds for its minimizer δ_0.

Lemma 5. *Suppose that there are functions g^ℓ, g^u such that $g^\ell(\delta) \leq g(\delta) \leq g^u(\delta)$ for all $\delta \in [0, \infty)$. Then for any $r \geq \inf_{\delta \geq 0} g^u(\delta)$, we have*

$$\delta_0 \in \{\delta \geq 0 : g^\ell(\delta) \leq r\}.$$

In particular, if g^ℓ is unimodal, then this sub-level set is an interval.

The proof of this lemma is simple. For a given $r \geq \inf_{\delta \geq 0} g^u(\delta)$, we have

$$g^\ell(\delta_0) \overset{(i)}{\leq} g(\delta_0) \overset{(ii)}{=} \inf_{\delta \geq 0} g(\delta) \overset{(iii)}{\leq} \inf_{\delta \geq 0} g^u(\delta) \ \leq \ r$$

where inequalities (i) and (iii) follow from the assumed sandwich relation, and equality (ii) follows from the fact that δ_0 is the minimizer of g.

Lemma 5 and the bound (36) together show that bounds on the localized Gaussian width that appears in the definition (34) of g can be used to obtain high probability upper and lower bounds on the error of the LSE.

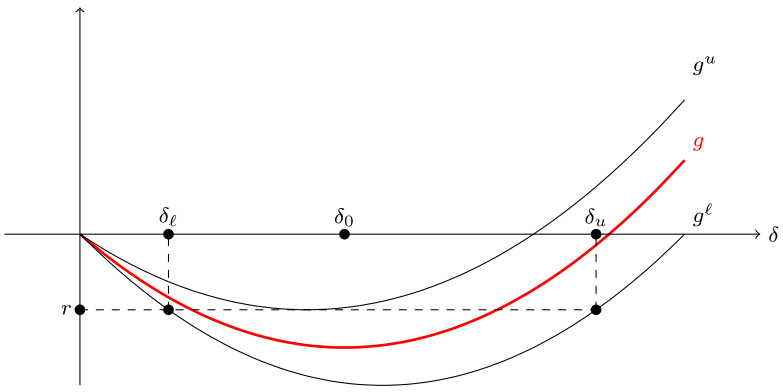


FIG 5. *Visualization of Lemma 5 when* $r = \inf_{\delta \geq 0} g^u(\delta)$, *and* g^ℓ *is convex.*

We remark that the case for estimation over $\mathcal{E}(R)$ for $R > 0$ reduces to the case $R = 1$ by rescaling. Let $\mathcal{E}(R)_{\theta^*} := \{\theta - \theta^* : \theta \in \mathcal{E}(R)\}$ denote the re-centered ellipse. Note that g can be rewritten as

$$g(\delta) := \frac{\delta^2}{2} - \sigma \mathscr{G}(\mathcal{E}(R)_{\theta^*} \cap \mathbb{B}(\delta)) = R^2 \underbrace{\left[\frac{\widetilde{\delta}^2}{2} - \widetilde{\sigma}\mathscr{G}(\mathcal{E}_{\widetilde{\theta}^*} \cap \mathbb{B}(\widetilde{\delta}))\right]}_{\widetilde{g}(\widetilde{\delta})} \tag{37}$$

after the changes of variables $\widetilde{\delta} := \delta/R$, $\widetilde{\theta}^* := \theta^*/R$, and $\widetilde{\sigma} := \sigma/R$. Then one can focus on bounding $\widetilde{g}$ and ultimately re-scale by R any bounds obtained for the minimizer of $\widetilde{g}$ in order to obtain bounds for the original minimizer δ_0.

D.2. Main portion of the proof

Recall the two functions defined in equation (16). Under our assumptions, the bounds (13) hold, so that the critical function g from equation (34) is sandwiched as $g^\ell(\delta) \leq g(\delta) \leq g^u(\delta)$ for all δ. Now Lemma 5 is applicable for ellipse $\mathcal{E}$, constant $r = -\frac{\delta_*^2}{2} = g^u(\delta_*)$ and function pair (g^u, g^ℓ), so we know $\delta_0 \in \{\delta \geq 0 \mid g^\ell(\delta) \leq r\}$.

Since function g^ℓ is convex in δ, there are two solutions δ' and δ'' to the equation

$$g^\ell(\delta) = -\frac{\delta_*^2}{2}, \tag{38}$$

and Lemma 5 guarantees that

$$\delta' \leq \delta_0 \leq \delta''. \tag{39}$$

Moreover, we show below that $c_1\delta_* \leq \delta_0 \leq c_2\delta_*$. Taking this inequality to be true for the moment, combining it with equation (36) yields

$$(1-c)c_1\delta_* \leq \|\widehat{\theta} - \theta^*\|_2 \leq (1+c)c_2\delta_*, \qquad \text{with prob. } \geq 1 - 3\exp(-c'\delta_*^2/\sigma^2),$$

which concludes the proof. Note that we arrive at the expectation bounds (19) by simply applying the earlier result (35).

It remains to show that $c_1\delta_* \leq \delta'$ and $\delta'' \leq c_2\delta_*$. After some manipulation using the fixed point equation (17), equation (38) can be rewritten as

$$\left(\delta - \sigma\sqrt{c_u^2 k_*(\delta)}\right)^2 = \sigma^2(c_u^2 k_*(\delta) - c_\ell^2 k_*(\delta_*)).$$

Note that the solutions δ to the equality (38) must satisfy $c_u^2 k_*(\delta) \geq c_\ell^2 k_*(\delta_*)$, as required for the right-hand side to be non-negative. In addition, they must satisfy one of the following two equations:

$$\delta = \sigma\sqrt{c_u^2 k_*(\delta)} + \sigma\sqrt{c_u^2 k_*(\delta) - c_\ell^2 k_*(\delta_*)} := h_+(\delta), \tag{40a}$$

$$\delta = \sigma\sqrt{c_u^2 k_*(\delta)} - \sigma\sqrt{c_u^2 k_*(\delta) - c_\ell^2 k_*(\delta_*)} := h_-(\delta). \tag{40b}$$

Note that any solution δ'' to the first equation (40a) is larger than any solution δ' to the second equation (40b). Indeed, we have $\delta' = h_-(\delta') < h_+(\delta')$, so the non-increasing nature of h_+ guarantees that the solution δ'' to the equation $\delta = h_+(\delta)$ must be larger than δ'.

- We first consider the solution δ'' to the first equation (40a). It is easy to check that

  $$\sigma c_u\sqrt{k_*(\delta)} \leq h_+(\delta) \leq 2\sigma c_u\sqrt{k_*(\delta)}.$$

  Recall $k_*(\delta)$ is non-increasing in δ. We know δ'' is smaller than the solution to $\delta = 2\sigma c_u\sqrt{k_*(\delta)}$, which in turn is smaller than $c_2\delta_*$ (by assumption (c) of Proposition 1). We thus have $\delta_* \leq \delta'' \leq c_2\delta_*$.
- Next we consider the solution δ' to the second equation (40b). We claim that $\delta' \geq c_1\delta_*$. In order to show this, we prove that $h_-(\delta)$ satisfies

  $$h_-(c_1\delta_*) \overset{(i)}{\geq} c_1\delta_*, \text{ for some } c_1 \in (0,1) \qquad \text{and} \quad h_-(\delta_*) \overset{(ii)}{\leq} \delta_*. \tag{41}$$

  Take the above inequalities as given for now, we can combine them with the fact that $h_-(\delta)$ is a non-decreasing function of δ to conclude that the fixed point solution δ' of (40a) satisfies $c_1\delta_* \leq \delta' \leq \delta_*$.

Putting these two pieces together with inequality (39), we conclude the proof of Proposition 1. It remains to prove the inequalities (41).

Proof of part (i) Applying the simple inequality $c_u^2 k_*(c_1\delta_*) - c_\ell^2 k_*(\delta_*) \leq c_u^2 k_*(c_1\delta_*)$ yields

$$h_-(c_1\delta_*) = \frac{\sigma c_\ell^2 k_*(\delta_*)}{\sqrt{c_u^2 k_*(c_1\delta_*)} + \sqrt{c_u^2 k_*(c_1\delta_*) - c_\ell^2 k_*(\delta_*)}} \geq \frac{\sigma c_\ell^2 k_*(\delta_*)}{2\sqrt{c_u^2 k_*(c_1\delta_*)}}$$
$$\geq c_1 \sigma \sqrt{c_\ell^2 k_*(\delta_*)},$$

where the last inequality follows by the fact that $c_u^2 k_*(c_1\delta) \leq \frac{1}{4c_1^2} c_\ell^2 k_*(\delta_*)$ (cf. Assumption (b) in Proposition 1). The fixed point equation (17) further implies that

$$h_-(c_1\delta_*) \geq c_1 \sigma \sqrt{c_\ell^2 k_*(\delta_*)} = c_1 \delta_*,$$

which proves our claim (i).

Proof of part (ii) Using the fact that $c_u^2 k_*(\delta_*) \geq c_\ell^2 k_*(\delta_*)$, we find that

$$h_-(\delta_*) = \frac{\sigma c_\ell^2 k_*(\delta_*)}{\sqrt{c_u^2 k_*(\delta_*)} + \sqrt{c_u^2 k_*(\delta_*) - c_\ell^2 k_*(\delta_*)}} \leq \frac{\sigma c_\ell^2 k_*(\delta_*)}{\sqrt{c_u^2 k_*(\delta_*)}} \leq \sigma \sqrt{c_\ell^2 k_*(\delta_*)} = \delta_*,$$

where the last equality follows from the fact that δ_* is a solution of the fixed point equation. This completes the proof of claim (ii).

Appendix E: Auxiliary proofs for Theorem 2

In this section, we collect the proofs of various auxiliary results that underlie Theorem 2.

E.1. Proof of Lemma 2

The set class $\mathscr{T}$ to be demonstrated consists of all s-sized subsets of a particular subset $T \subset \{m_1, \ldots, m_2\}$; the subset T is constructed to have cardinality at least $\lfloor \frac{k_*-1}{16} \rfloor$, so that the set class $\mathscr{T}$ has at least $\binom{\lfloor \frac{1}{16}(k_*-1) \rfloor}{s}$ elements.

Consider the $k_* - 1$ diagonal elements of the matrix $V^\top X^\top B X V$. The sum of these diagonal elements is $\operatorname{tr}\left(V^\top X^\top B X V\right)$. Furthermore, the pigeonhole principle ensures that the smallest $\frac{15}{16}(k_* - 1)$ of the diagonal elements are each at most

$$\frac{16}{k_* - 1} \operatorname{tr}\left(V^\top X^\top B X V\right) = \frac{16}{k_* - 1} \operatorname{tr}\left(X^\top B X\right) \leq 16 \max_{i \leq k_* - 1} \|x_i\|_{\mathcal{E}}^2. \tag{42}$$

Let T be the indices of those $\frac{15}{16}(k_* - 1)$ diagonal elements that are also in $\{m_1, \ldots, m_2\}$. By construction, we have $|T| \geq m_2 - m_1 - \frac{1}{16}(k_* - 1) = \frac{k_*-1}{16}$, as desired.

Given the set class $\mathscr{T}$ defined by the subset T, we now show that inequality (i) in equation (26) holds. Note that Lemma 1 implies that any sign vector z^S supported on S satisfies

$$\|\theta^S - \theta^*\|_2^2 = \|\theta^\dagger - \theta^*\|_2^2 + \frac{b^2\delta^2}{32s}\|UHVz^S\|_2^2.$$

Here the decomposition uses the fact that $\theta^\dagger - \theta^*$ is parallel to u_1, as guaranteed by part (a) of Lemma 1; this property ensures that $\theta^\dagger - \theta^*$ is orthogonal to $u_2, \ldots, u_{k_*}$. Since the columns of U are orthogonal unit vectors, we have $\|UHVz^S\|_2^2 = \|HVz^S\|_2^2$. Then recalling that $V^\top H^\top HV = \Sigma^2$ is a diagonal matrix containing the squared singular values of H, we may use inequality (ii) in Lemma 1 to obtain

$$\begin{aligned}\|\theta^S - \theta^*\|_2^2 &= a^2\delta^2 + \frac{b^2\delta^2}{32s}\|HVz^S\|_2^2 \\ &\geq \Big[a^2 + \frac{b^2}{32}\Big(1 - \frac{m_2}{k_* - 1} - \sqrt{\frac{a^2 - 9b^2}{9b^2}}\Big)^2\Big]\delta^2 \\ &\geq \Big(a^2 + \frac{b^2}{2^9}\Big)\delta^2\end{aligned}$$

where the last step follows from inequality (48). Here let us take η small enough, for instance 10^{-5} such that the right hand side above is greater than δ^2. (We have made these choices of constants for the sake of convenience in the proof, but note that other choices of these quantities are also possible.)

Now, we prove that $\theta^S \in \mathcal{E}$ and inequality (ii) in equation (26) holds, in particular by using a probabilistic argument. Recall that $B := \operatorname{diag}(\mu_1^{-1}, \ldots, \mu_d^{-1})$ so that $\|x\|_{\mathcal{E}}^2 = x^\top Bx$. For a given subset S, we specify a random choice of z^S, in which for each $j \in S$, the value $z_j^S \in \{-1, +1\}$ is an independent Rademacher variable. Using this random choice of z^S, we then let θ^S be defined as in equation (25), so that it is now a random vector.

Now part (a) of Lemma 1 guarantees that the vector $B\theta^\dagger$ is orthogonal to $u_2, \ldots, u_{k_*}$. As a consequence, we have $\|\theta^S\|_{\mathcal{E}}^2 = (\theta^\dagger)^\top B\theta^\dagger + \frac{b^2\delta^2}{32s}\|XVz^S\|_{\mathcal{E}}^2$.

Let us focus on the expectation of the second term in the equation above. By the linearity and cyclic invariance properties of trace, we have

$$\begin{aligned}\mathbb{E}\|XVz^S\|_{\mathcal{E}}^2 &= \mathbb{E}[(z^S)^\top V^\top X^\top BXVz^S] \\ &= \operatorname{tr}\big(V^\top X^\top BXV\mathbb{E}[z^S(z^S)^\top]\big) \\ &= \operatorname{tr}\big(V^\top X^\top BXVI_S\big),\end{aligned}$$

where $I_S = \mathbb{E}[z^S(z^S)^\top]$ is the diagonal matrix whose ith diagonal entry is 1 if $i \in S$ and zero otherwise. The last expression is the sum of s diagonal entries of $V^\top X^\top BXV$ indexed by elements of T, so that our earlier bound (42) implies that

$$\mathbb{E}\|XVz^S\|_{\mathcal{E}}^2 \leq 16\, s \max_{i \leq k_* - 1}\|x_i\|_{\mathcal{E}}^2.$$

Letting i^* denote the maximizer of the right-hand side, then combining the previous few displays yields

$$\mathbb{E}\|\theta^S\|_{\mathcal{E}}^2 = (\theta^\dagger)^\top B\theta^\dagger + \frac{b^2\delta^2}{32s}\mathbb{E}\|XVz^S\|_{\mathcal{E}}^2 \;\le\; (\theta^\dagger)^\top B\theta^\dagger + \frac{b^2\delta^2}{2}\|x_{i^*}\|_{\mathcal{E}}^2.$$

Again using the fact that $B\theta^\dagger$ and x_{i^*} are orthogonal, we have $\|\theta^\dagger + b\delta x_{i^*}\|_{\mathcal{E}}^2 = \|\theta^\dagger\|_{\mathcal{E}}^2 + b^2\delta^2\|x_{i^*}\|_{\mathcal{E}}^2$, and thus

$$\mathbb{E}\|\theta^S\|_{\mathcal{E}}^2 \le \frac{1}{2}\|\theta^\dagger\|_{\mathcal{E}}^2 + \frac{1}{2}\|\theta^\dagger + b\delta x_{i^*}\|_{\mathcal{E}}^2 \le \frac{\|\theta^*\|_{\mathcal{E}}^2 + 1}{2}, \tag{43}$$

where the last step is due to the fact that $\|\theta^\dagger\|_{\mathcal{E}} \le \|\theta^*\|_{\mathcal{E}}$ by construction, as well as $\|\theta^\dagger + b\delta x_{i^*}\|_{\mathcal{E}}^2 \le 1$, by claim (c) in Lemma 1.

Similarly, part (a) of Lemma 1 implies the vector $\theta^\dagger - \theta^*$ is orthogonal to all of the vectors $u_2, \ldots, u_{k_*}$, whence

$$\|\theta^S - \theta^*\|_2^2 = \|\theta^\dagger - \theta^*\|_2^2 + \frac{b^2\delta^2}{2s}\|UHVz^S\|_2^2.$$

By properties of the trace along with the fact that $I_S = \mathbb{E}[z^S(z^S)^\top]$ is the diagonal matrix with ith diagonal entry equal to 1 if $i \in S$ and zero otherwise, we then have

$$\mathbb{E}\|\theta^S - \theta^*\|_2^2 = \|\theta^\dagger - \theta^*\|_2^2 + \frac{b^2\delta^2}{32s}V^\top H^\top HVI_S.$$

By noting $V^\top H^\top HV = \Sigma^2$ is diagonal with the squared singular values of H and applying the bound (i) from Lemma 1, we have

$$\mathbb{E}\|\theta^S - \theta^*\|_2^2 \le \Big(a^2 + \frac{b^2}{32}\cdot\frac{a^2}{9b^2}\cdot\frac{k_*-1}{m_1}\Big)\delta^2 \le \Big(1 + \frac{1}{36}\Big)a^2\delta^2 < 2\delta^2. \tag{44}$$

By a union bound and Markov's inequality, the two inequalities (43) and (44) imply

$$\begin{aligned}\mathbb{P}\Big(\|\theta^S\|_{\mathcal{E}}^2 > 1 \text{ or } \|\theta^S - \theta^*\|_2^2 > \frac{4}{1-\|\theta^*\|_{\mathcal{E}}^2}\delta^2\Big) &\le \mathbb{E}\|\theta^S\|_{\mathcal{E}}^2 + \frac{1-\|\theta^*\|_{\mathcal{E}}^2}{4\delta^2}\mathbb{E}\|\theta^S - \theta^*\|_2^2 \\ &< \frac{\|\theta^*\|_{\mathcal{E}}^2 + 1}{2} + \frac{1-\|\theta^*\|_{\mathcal{E}}^2}{2} = 1.\end{aligned}$$

We conclude that there exists some sign vector z^S satisfying both inequalities (i) and (iii).

E.2. Proof of Lemma 3

We prove Lemma 3 via two successive applications of the Sudakov-Fernique comparison inequality. In order to keep our presentation self-contained, let us

restate a version of this result here (e.g., see Theorem 3.15 in Ledoux and Talagrand [21]). For a given a pair of centered Gaussian vectors $\{X_j,\ j = 1, \ldots, N\}$ and $\{Y_j,\ j = 1, \ldots, N\}$, suppose that

$$\mathrm{var}(X_i - X_j) \leq \mathrm{var}(Y_i - Y_j) \qquad \text{for all } (i, j) \in [N] \times [N].$$

The Sudakov-Fernique comparison then asserts that $\mathbb{E}[\max_{j\in[N]} X_j] \leq \mathbb{E}[\max_{j\in[N]} Y_j]$.

Using this result, we now prove our claim. For each $S \in \mathscr{T}$, define the zero-mean Gaussian random variable $g^S := \langle w, UHVz^S\rangle$. First, define a diagonal matrix $D := \mathrm{diag}(0, \ldots, 0, \underbrace{1, \ldots, 1}_{m_1:m_2}, 0, \ldots, 0)$, and the zero-mean Gaussian random variables $\widetilde{g}^S := \frac{3}{4}\langle w, Dz^S\rangle$. We claim that the Sudakov-Fernique comparison implies that

$$\mathbb{E}\max_{S\in\mathscr{T}}\langle w, UHVz^S\rangle \geq \frac{1}{4}\mathbb{E}\max_{S\in\mathscr{T}}\langle w, Dz^S\rangle. \tag{45}$$

See below for the details of this claim. Second, we introduce the vector $\widehat{z}^S$ with components $\widehat{z}_i^S := |z_i^S|$, and define a third Gaussian process using the variables $\widehat{g}^S := \langle w, D(\widehat{z}^S - \widehat{z}^{S'})\rangle$. We also claim that

$$\mathbb{E}\max_{S\in\mathscr{T}}\langle w, Dz^S\rangle \geq \mathbb{E}\Big[\max_{S\in\mathscr{T}}\sum_{i\in S} w_i\Big]. \tag{46}$$

These two claims in conjunction imply the claim of Lemma 3. Let us now prove inequalities (45) and (46).

Proof of inequality (45) We claim that the processes $\{g^S, S \in \mathscr{T}\}$ and $\{\widetilde{g}^S, S \in \mathscr{T}\}$ satisfy the Sudakov-Fernique conditions. In order to prove this claim, we need to verify that for all subsets $S, S' \in \mathscr{T}$, we have relation $\mathrm{var}(g^S - g^{S'}) \geq \mathrm{var}(\widetilde{g}^S - \widetilde{g}^{S'})$. On one hand, we have

$$\begin{aligned}\mathrm{var}(g^S - g^{S'}) &= \mathbb{E}\langle w, UHV(z^S - z^{S'})\rangle^2 = \|UHV(z^S - z^{S'})\|_2^2 \\ &= \|HV(z^S - z^{S'})\|_2^2,\end{aligned}$$

where the last step uses the orthonormality of U. On the other hand, we have the equality $\mathrm{var}(\widetilde{g}^S - \widetilde{g}^{S'}) = \|D(z^S - z^{S'})\|_2^2$. Consequently, it suffices to show that there exists an orthogonal matrix V such that

$$(HV)^\top HV \succeq \frac{1}{16}D^2. \tag{47}$$

In order to see this fact, part (e) of Lemma 1 implies that the m_2 largest eigenvalues of $H^\top H = V\Sigma^2 V^\top$ is lower bounded by $1 - \frac{m_2}{k_*-1} - \sqrt{\frac{a^2-9b^2}{9b^2}}$. With the choice of the constants (a, b) specified above (see paragraph below Lemma 1),

it is guaranteed that $\frac{a^2-9b^2}{9b^2} \leq \frac{1}{4}$. This observation and the definition $m_2 := \lfloor (k_* - 1)/4 \rfloor$ together imply that

$$1 - \frac{m_2}{k_* - 1} - \sqrt{\frac{a^2 - 9b^2}{9b^2}} \geq 1 - \frac{1}{4} - \frac{1}{2} = \frac{1}{4}, \tag{48}$$

which implies the claim (47), and further completes the proof of the lower bound (45).

Proof of inequality (46) The vector $\widehat{z}^S$ defined above is an indicator vector for the support of z^S. Defining a third Gaussian process using the variables $\widehat{g}^S := \langle w, D(\widehat{z}^S - \widehat{z}^{S'}) \rangle$, we have

$$\mathrm{var}(\widetilde{g}^S - \widetilde{g}^{S'}) = \|D(z^S - z^{S'})\|_2^2 \ \geq \ \|D(\widehat{z}^S - \widehat{z}^{S'})\|_2^2 \ = \ \mathrm{var}(\widehat{g}^S - \widehat{g}^{S'}).$$

A second application of the Sudakov-Fernique inequality then yields

$$\mathbb{E} \max_{S \in \mathscr{T}} \langle w, Dz^S \rangle \geq \mathbb{E} \max_{S \in \mathscr{T}} \langle w, D\widehat{z}^S \rangle = \mathbb{E}\Big[\max_{S \in \mathscr{T}} \sum_{i \in S} w_i \Big],$$

where in the last step we recall the fact that S is supported on the set $\{m_1, \ldots, m_2\}$.

Appendix F: Proof of auxiliary lemmas

In this section, we collect the proofs of various auxiliary lemmas.

F.1. Proof of Lemma 6

Let us first state the lemma used in Section 4.1.2.

Lemma 6. *For an extremal vector of the form $\theta^* = \sqrt{\mu_s} e_s - r e_s$, the critical dimension (8) is lower bounded as*

$$k_*(\delta) \geq 0.09 \cdot \arg \max_{1 \leq k \leq d} \Big\{ \mu_k^2 \geq \delta^2 \mu_s \Big\}.$$

The rest of this section is devoted to the proof of this lemma.

Defining the integer $m := \max\{2, \arg\max_{1 \leq k \leq d} \Big\{ \mu_k^2 \geq \delta^2 \mu_s \Big\}\}$, Wei and Wainwright [37] show that we can inscribe an $(m-1)$-dimensional ℓ_∞ ball with radius $\delta/\sqrt{m-1}$ into the ellipse $\mathcal{E}$; in particular, see [37, Appendix D]. We claim that the Kolmogorov k-widths of the s-dimensional ℓ_∞ ball of radius $\frac{1}{\sqrt{s}}$ are lower bounded as

$$\mathscr{W}_k(\mathbb{B}_\infty(1/\sqrt{s})) \geq 1 - \frac{k}{s}. \tag{49}$$

Taking this claim as given for the moment, we use it to complete the proof of Lemma 6. Using the lower bound (49), we have

$$\mathscr{W}_k\Big(\mathcal{E}_{\theta^*} \cap \mathbb{B}(\xi)\Big) \geq \mathscr{W}_k\Big(\mathbb{B}_\infty(\frac{\xi}{\sqrt{m-1}})\Big) \geq \left(1 - \frac{k}{m-1}\right)\xi.$$

With $\xi := (1-\eta)\delta$ and $k = (1 - \frac{0.9}{1-\eta})(m-1)$ the above becomes $\mathscr{W}_k\Big(\mathcal{E}_{\theta^*} \cap \mathbb{B}((1-\eta)\delta)\Big) \geq 0.9\delta$, so by the definition of the critical dimension (8), we have

$$k_*(\delta) \geq (1 - \frac{0.9}{1-\eta})(m-1) \geq 0.09 \cdot \arg\max_{1 \leq k \leq d} \Big\{\mu_k^2 \geq \delta^2 \mu_s\Big\},$$

as claimed.

The only remaining detail is to prove inequality (49).

Proof of inequality (49) Define the set $\mathcal{V} := \{v \in \mathbb{R}^s \mid v_i = \pm\frac{1}{\sqrt{s}}\}$ with cardinality $M = 2^s$. We claim that for any k-dimensional subspace $W \subseteq \mathbb{R}^s$, there exists some $v \in \mathcal{V}$ such that

$$\|v - \Pi_W(v)\|_2^2 \geq 1 - \frac{k}{s}. \tag{50}$$

Then by definition of Kolmogorov width, the inequality (49) holds. In order to prove the lower bound (50), we take an orthonormal basis $z_1, \ldots, z_k$ of W and extend it to an orthonormal basis $z_1, \ldots, z_m$ for $\mathbb{R}^s$. We then have

$$\sum_{v \in \mathcal{V}} \|v - \Pi_W(v)\|_2^2 = \sum_{v \in \mathcal{V}} \sum_{j=k+1}^{s} \langle v, z_j\rangle^2 = \sum_{j=k+1}^{s} \sum_{v \in \mathcal{V}} \langle v, z_j\rangle^2 = (s-k) \cdot \frac{M}{s},$$

where we have used the fact that

$$\sum_{v \in \mathcal{V}} \langle v, z_j\rangle^2 = M \cdot \frac{1}{s}\|z_j\|_2^2 = \frac{M}{s}.$$

Therefore, there must exist some $v \in \mathcal{V}$ such that $\|v - \Pi_W(v)\|_2^2 \geq 1 - \frac{k}{s}$, which establishes the inequality (49).

F.2. Proof of Lemma 4

Recalling our calculations from Example 4, we found that

$$k_*(0, \frac{1}{2}\delta) = \operatorname{argmin}_k\{\sqrt{\mu_{k+1}} \leq \frac{9}{10} \cdot \frac{1}{2}\delta\}.$$

Consequently, in order to prove Lemma 4, it suffices to show that

$$k_*(\theta^*, \delta) \leq \arg\min_k\{2\sqrt{\mu_k} \leq \frac{9}{10}\delta\} \qquad \text{for all } \delta \leq \sqrt{\mu_1}.$$

By definition of the critical dimension (8), it is sufficient to show that the Kolmogorov width is upper bounded as

$$\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}(a\delta)) \leq \min\{a\delta,\ 2\sqrt{\mu_k}\}, \tag{51}$$

where $a := 1 - \eta$.

We claim that the set $\mathcal{E}_{\theta^*} \cap \mathbb{B}(a\delta)$ is contained within the set $2\mathcal{E} \cap \mathbb{B}(a\delta)$. Indeed, note that any $v \in \mathcal{E}_{\theta^*} \cap \mathbb{B}(a\delta)$ has Euclidean norm bounded as $\|v\|_2 \leq a\delta$ and Hilbert norm bounded as $\|v + \theta^*\|_{\mathcal{E}} \leq 1$. The Cauchy-Schwarz further guarantees that

$$\|v\|_{\mathcal{E}}^2 = \|v + \theta^* - \theta^*\|_{\mathcal{E}}^2 \leq 2\|v + \theta^*\|_{\mathcal{E}}^2 + 2\|\theta^*\|_{\mathcal{E}}^2 \leq 4,$$

where the last step follows from the fact that both θ^* and $v + \theta^*$ lie in ellipse $\mathcal{E}$. We have thus established the claimed set inclusion.

From this set inclusion, we have

$$\mathscr{W}_k(\mathcal{E}_{\theta^*} \cap \mathbb{B}(a\delta)) \leq \mathscr{W}_k(2\mathcal{E} \cap \mathbb{B}(a\delta)) = \min\{a\delta,\ 2\sqrt{\mu_k}\},$$

which establishes the claim (51). Putting pieces together completes the proof of Lemma 4.

Appendix G: Well-definedness of the function Φ

In this section, we verify that the function Φ from equation (12) is well-defined. We again use the shorthand $a := 1 - \eta$. In order to provide intuition, Figure 6 provides an illustration of Φ.

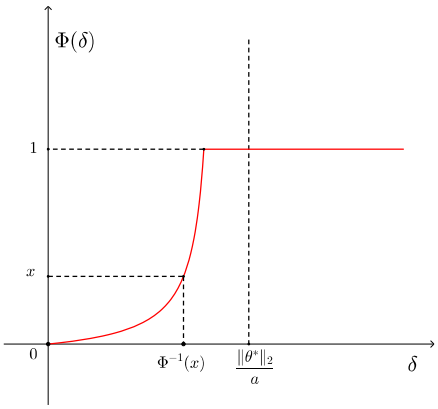


FIG 6. *Illustration of the function* Φ.

We begin with the case when $\delta < \|\theta^*\|_2/a$. For simplicity of notation, let

$$r(\delta) := \min\left\{r \geq 0 \ \mid a^2\delta^2 \leq \sum_{i=1}^{d} \frac{r^2}{(r + \mu_i)^2}(\theta_i^*)^2\right\}.$$

Note that for each for each $\mu_i \geq 0$, the function $f(r) := \sum_{i=1}^{d} \frac{r^2}{(r+\mu_i)^2}(\theta_i^*)^2$ is non-decreasing in r. It is also easy to check that

$$\lim_{r \to 0^+} f(r) = 0, \qquad \text{and } \lim_{r \to \infty} f(r) = \|\theta^*\|_2^2.$$

Then the quantity $r(\delta)$ is uniquely defined and positive whenever $\delta < \|\theta^*\|_2/a$. Note that as $\delta \to \frac{\|\theta^*\|_2}{a}$, $a^2\delta^2 \to \|\theta^*\|_2^2$ therefore $r(\delta) \to \infty$.

It is worth noticing that given any θ^* where $\|\theta^*\|_2$ does not depend on δ, r goes to zero when $\delta \to 0$, namely $\lim_{\delta \to 0^+} \Phi(\delta) = 0$.