On the Minimax Rate of the Gaussian Sequence Model Under Bounded Convex Constraints

Matey Neykov

Abstract—We determine the exact minimax rate of a Gaussian sequence model under bounded convex constraints, purely in terms of the local geometry of the given constraint set $K$. Our main result shows that the minimax risk (up to constant factors) under the squared $\ell_2$ loss is given by $\varepsilon^2 \land \text{diam}(K)^2$ with $\varepsilon^* = \sup \left\{ \varepsilon : (\varepsilon^2 / \sigma^2) \leq \log M_{\text{loc}}(\varepsilon) \right\}$, where $\log M_{\text{loc}}(\varepsilon)$ denotes the local entropy of the set $K$, and $\sigma$ is the variance of the noise. We utilize our abstract result to re-derive known minimax rates for some special sets such as hyperrectangles, ellipses, and more generally quadratically convex orthosymmetric sets. Finally, we extend our results to the unbounded case with known $\sigma^2$ to show that the minimax rate in that case is $\varepsilon^2$.

Index Terms—Estimation, Gaussian distribution, minimax techniques.

I. INTRODUCTION

This paper focuses on the Gaussian sequence model $Y_i = \mu_i + \xi_i$ with $n$ observations (i.e., $i \in \{1, \ldots, n\}$), where $\xi_i \sim N(0, \sigma^2)$ are independent and identically distributed (i.i.d.), and the vector $\mu \in \mathbb{R}^n$ belongs to a known bounded convex set $K$. In particular we would like to determine the minimax rate for this problem. In detail, we would like to quantify (up to proportionality constants) the rate of the following expression, also known as the minimax risk:

$$\inf_{\mathcal{F}} \sup_{\mu \in K} \mathbb{E} \|\hat{\mu}(Y) - \mu\|^2,$$  

where the infimum is taken with respect to all measurable functions (estimators) of the data, and we use the shorthand $\|\cdot\|$ for the Euclidean norm. The minimax risk may appear to be overly pessimistic to some, but everyone will agree that it represents an important measure of the difficulty of the problem. The main contribution of this work is establishing matching (up to constants) upper and lower bounds for the risk (1.1) for any bounded convex set $K$. In particular we would like to single out the upper bound as the main contribution, as the lower bound is a simple consequence of Fano’s inequality. In order to establish the upper bound, we demonstrate that there exists a universal scheme which attains the minimax rate specified by (1.1).

Manuscript received 23 March 2022; revised 28 July 2022; accepted 2 October 2022. Date of publication 10 October 2022; date of current version 20 January 2023. This work was supported in part by Grant NSF DMS-2113684.

The author is with the Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: mneykov@stat.cmu.edu).

Communicated by M. Rodrigues, Associate Editor At Large for Machine Learning and Statistics, Signal Processing and Source Coding.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3213141.

Digital Object Identifier 10.1109/TIT.2022.3213141.
as sequence models under convex constraints. He also shows
that unfortunately the LSE is not minimax optimal in general,
as there exist convex sets where the gap between the minimax
rate and the performance of the LSE can be as large as $\sqrt{n}$
(on the squared risk scale when $\sigma = 1$). This counterexample
naturally leads [5] to ask the question “as to whether there
is a general estimator that is guaranteed to be minimax up
to a universal constant”. Hence the need arises to find other
estimators which always enjoy minimaxity.

A. Related Literature

There is a tremendous amount of work on the Gaussian
sequence model. Here we will only scratch the surface. The
interested reader can consult books on the sequence
model and nonparametric statistics such as [17], [19], and [25].

In one of the most classical results, [22] showed the precise
linear minimax rate when the set $K$ is an ellipse, and in fact he
showed that a linear estimate achieves the minimax rate when
$\sigma \to 0$. Pinsker’s results are valid in a framework more general
than the one we consider in this paper as he looked at ellipses in
the $\ell_2$ space, whereas we consider only subsets of $\mathbb{R}^n$. When
$n = 1$ any bounded convex set is an interval and in that sense
the works of [2], [4], and [15] are very relevant. We will later
see when we consider the example of hyperrectangles that we
are able to recover their result up to constant factors. In a
classic work, [8] consider almost the exact same problem as we
consider here (with $\ell_2$ instead of $\mathbb{R}^n$) and work out a variety
of special cases for $K$ — such as hyperrectangles, ellipses,
and orthosymmetric quadratically convex sets. They show that
a linear projection estimator (also known as the truncated
series estimator) is minimax optimal up to constants in all
of these examples. We will re-derive all of their results (up to
constants) in the Examples section to follow. Reference [16]
derive the minimax rate for symmetric convex polytopes up
to logarithmic factors using the truncated series estimator.
Reference [16] also point out in their introduction, that “it is
still largely unknown how to compute the minimax risk for an
arbitrary convex body”. Reference [31] obtains the minimax
rate up to a logarithmic factor for $\ell_q$ balls for $q \leq 1$, by using
an estimator which is a mixture of LSE and a linear projection
estimator. Reference [6] extend results of [5] to show that the
LSE and other regularized estimators are admissible up
to universal constants in the same setting that we consider.
We will see later on that our estimator, although of different
nature than the aforementioned ones, also has this property due
to the fact that it is minimax up to constant factors. In a recent
paper, [10] shows that the linear minimax risk in the sequence
model in $\ell_2$ can be explicitly quantified for certain convex sets of
the form $K = \{x = \{x_i\}_{i=1}^\infty : \sup_k a_k^{-1} \sum_{j=k}^\infty x_j^2 \leq P_0\}$
with $a_k > 0$ being a decreasing sequence. Moreover, [10]
shows that the asymptotic minimax risk when $a_k = k^{-2\alpha}$
can be precisely quantified as well.

Aside from the aforementioned works which focus on
the Gaussian sequence model, we would like to discuss the
celebrated paper of [30] which is also highly relevant (yet
does not consider the sequence model per se). Reference [30]
based their work on the premise that local entropy is hard
to calculate in general, yet it had been shown that it leads
to optimal rates of convergence by [18] and [3] in certain
problems metrized with the squared Hellinger distance as we
alluded to previously. Therefore [30] proposed to study
the global entropy instead, which is often easier to handle.
We must agree, that local entropy (see Definition II.2) is a
challenging quantity to work with, nevertheless, as our result
shows it is precisely what is needed to calculate in order to
determine the minimax rate for a general convex set $K$. This
is also easy to explain intuitively at this point of the paper even
without going into the mathematical details. Consider, e.g.,
the case where the set $K$ is unbounded, e.g., $K$ is a subspace
(which corresponds to the linear regression setting). The global
entropy of such a set is not even defined (as one cannot pack
an unbounded set), yet its local entropy is well defined and
calculable. We would also further comment that for some sets
$K$ it is sufficient to calculate the global entropy as it is of the
same order as the local entropy. In fact, [30] offer a result (see
Lemma 3 in Section 7 therein), which connects the local and
global entropies. Sometimes, the order of the two quantities
coincides, in which case one may resort to calculating
the global entropy of $K$ instead. See also Subsection III-D where
we illustrate this by considering the example of an $\ell_1$ ball.

B. Organization

The paper is structured as follows. We present our main
results on bounded convex sets $K$ in Section II. Section III is
dedicated to some examples. Section IV argues that the esti-
mator defined in Section II is adaptive to the true point, and it
also is admissible up to a universal constant. Section V extends
our main results from the bounded case to the unbounded case
with known $\sigma^2$. A brief discussion is given in Section VI.

C. Notation

We outline some commonly used notation here. We use $\vee$
and $\wedge$ for max and min of two numbers respectively.
Throughout the paper $\|\cdot\|$ denotes the Euclidean norm. Constants may
change values from line to line. For an integer $m \in \mathbb{N}$ we use the shorthand $[m] = \{1, \ldots, m\}$. We use $B(\theta, r)$ to denote
a closed Euclidean ball centered at the point $\theta$ with radius $r$.
We use $\leq$ to mean $\leq$ and $\geq$ up to absolute constant
factors, and for two sequences $a_n$ and $b_n$ we write $a_n \asymp b_n$
if both $a_n \preceq b_n$ and $a_n \succeq b_n$ hold. Throughout the paper we
use log to denote the natural logarithm.

II. Main Results

Here we focus on the following problem. We observe $n$
observations (i.e., $i \in [n]$) $Y_i = \mu_i + \xi_i$, where $\mu \in K$, for
$K$ being a bounded convex set and $\xi_i \sim N(0, \sigma^2)$ are i.i.d.
random variables. We begin with showing a lower bound.

A. Lower Bound

In this subsection we present our main lower bound. It is a
simple consequence of Fano’s inequality, which we state below
for the convenience of the reader. Throughout this section and
the rest of the paper $c > 0$ is some sufficiently large absolute
constant.
Lemma II.1 (Fano’s Inequality): Let \( \mu^1, \ldots, \mu^m \) be a collection of \( \varepsilon \)-separated points in the parameter space in Euclidean norm. Suppose \( J \) is uniformly distributed over the index set \([m]\), and \( (Y|J=j) = \mu^j + \xi \) for \( \xi \sim N(0, \sigma^2) \). Then

\[
\inf_{\hat{\mu}} \sup_{\mu} \mathbb{E}[\|\hat{\nu}(Y) - \mu\|^2] \geq \frac{\varepsilon^2}{4} \left( 1 - I(Y; J) + \log 2 \right) \frac{\log m}{\log m}.
\]

In the above \( I(Y; J) \) is the mutual information between \( Y \) and \( J \), and can be upper bounded by \( \frac{1}{m} \sum_j D_{KL}(P_{\mu^j} \| P_\nu) \leq \frac{1}{m} \sum_j \|\mu^j - \nu\|^2 \), for any \( \nu \in \mathbb{R}^n \) (see (15.52) [28] e.g.). We will now define local packing entropy.

Definition II.2 (Local Entropy): Let \( \theta \in K \) be a point. Consider the set \( B(\theta, \varepsilon) \cap K \). Let \( M(\varepsilon/c, B(\theta, \varepsilon) \cap K) \) denote the largest cardinality of an \( \varepsilon/c \) packing set [see Definition 5.4 [28], e.g., for a definition of a packing set] in \( B(\theta, \varepsilon) \cap K \). Let

\[
M^{loc}(\varepsilon) = \sup_{\theta \in K} M(\varepsilon/c, B(\theta, \varepsilon) \cap K).
\]

We refer to \( M^{loc}(\varepsilon) \) as local entropy of \( K \). Sometimes we will use \( M^{loc}(\varepsilon) \) if we the set \( K \) is not clear from the context.

Lemma II.3: We have

\[
\inf_{\hat{\nu}} \sup_{\nu} \mathbb{E}[\|\hat{\nu}(Y) - \nu\|^2] \geq \frac{\varepsilon^2}{8c^2},
\]

for any \( \varepsilon \) satisfying \( \log M^{loc}(\varepsilon) > 2(\varepsilon^2/(2\sigma^2) + \log 2) \), where \( c \) is the constant from the Definition II.2 which is fixed to some large enough value.

Proof: For a given \( \varepsilon \) we can build an \( \varepsilon/c \)-local packing of cardinality \( M^{loc}(\varepsilon) \), around some point of \( K \). If such a point does not exist, we can take a sequence of points which achieve this in the limit, which is good enough for our argument to follow. Suppose that \( \log M^{loc}(\varepsilon) > 2(\varepsilon^2/(2\sigma^2) + \log 2) \). From Fano’s inequality it immediately follows that the minimax risk is \( \geq \frac{\varepsilon^2}{8c^2} \). The above is implied when \( \log M^{loc}(\varepsilon) > 4(\varepsilon^2/(2\sigma^2) + \log 2) \). \( \Box \)

B. Upper Bound

In this subsection we focus on the upper bound. Let \( d = \text{diam}(K) \). We propose the estimator described in Algorithm 1, where \( 2(C+1) = c \) is the constant from the definition of local entropy which is assumed to be sufficient large. The reader will notice that our algorithm contains an infinite loop. This means that our estimator can only be achieved in theory. The proof is that one knows a lower bound on \( \sigma \) (including cases when one knows \( \sigma \) exactly), one need not run the procedure ad infinitum. In that case the number of iterations can be determined through a concentration result to follow. We give an updated algorithm with finitely many iterations and additional details of this in Appendix A.

Algorithm 1 Upper Bound Algorithm

Input: A point \( \nu^* \in K \)

1. \( k \leftarrow 1; \)
2. \( Y \leftarrow [\nu^*]; /* This array is needed solely in the proof and is not used by the estimator */ \)
3. while true do
4. \( \text{Take a } \frac{d}{2(C+1)} \text{ maximal packing set } M_k \text{ of the set } B(\nu^*, \frac{d}{2(C+1)}) \cap K; /* The packing sets should be constructed prior to seeing the data */ \)
5. \( \nu^* \leftarrow \arg \min_{\nu \in M_k} \|Y - \nu\|; /* Break ties by taking the point with the least lexicographic ordering */ \)
6. \( Y, \text{append}(\nu^*); \)
7. \( k \leftarrow k + 1; \)

8. return \( \nu^*; /* Observe that by definition } Y \text{ forms a Cauchy sequence, so } \nu^* \text{ can be understood as the limiting point of that sequence. */ \)

1Here the maximality of the packing set is not really important; what is important is that the packing set is a covering. This can be “constructed algorithmically” by greedily taking points one by one and carving balls centered at those points.

2Take any two points \( Y_m \) and \( Y_m' \) for \( m' > m \). Then \( \|Y_m - Y_m'\| \leq \sum_{i=m}^{m'-1} \|Y_i - Y_{i+1}\| \leq \sum_{i=m}^{m'-1} d/2^{i+1} \leq d/2^{m'-2} \) so we have a Cauchy sequence.
Theorem II.4: The function $\nu^* : \mathbb{R}^n \mapsto \mathbb{R}^n$ is measurable (with respect to the Borel $\sigma$-field). As a consequence we have that $\nu^*(Y)$ is a random variable.

Proof: First we observe that for each $j$: $\Upsilon_j : \mathbb{R}^n \mapsto \mathbb{R}^n$ are measurable (here we denote by $\Upsilon_j$ the elements of the array $\Upsilon$ which is defined in Algorithm 1). In order to see this, we need to realize that one can (and should) construct the packing sets before one sees the data $Y$. This will form an infinite tree of packing sets rooted at the initial point $\Upsilon_1$. Each packing set splits $\mathbb{R}^n$ into polytopes (some of which may be unbounded) where each point in the packing set is the closest to any point in its corresponding polytope (this is the Voronoi tessellation in Euclidean norm). On the boundaries of these polytopes more than one point can be the closest point — in that case in order to consistently assign a single point always take the point with the least lexicographic order (i.e. it has the smallest 1st coordinate of all points, and the smallest 2nd coordinate of all points with equally small first coordinate and so forth).

Consider the event that $\Upsilon_j(y)$ belongs to a certain packing set, say, $M$ (i.e. the point $y$ is closest to all ancestor nodes of $M$ which essentially means that $y$ belongs to some intersection of polytopes (which is again a polytope call it $Q$)). For a point $m \in M$ we have that $\{y : \Upsilon_j(y) = m\} = \{y \in P \cap \{y : \Upsilon_j(y) \in M\} = \{y \in P \cap \{y \in Q\} = \{y \in P \cap Q\},$ where $P$ is the polytope from the Voronoi tessellation given by $M$, of the point $m$. Since (convex) polytopes are comprised of finitely many linear inequalities they are Borel sets and hence the event $\Upsilon_j(y) = m$ is measurable. Repeating this argument for any point on the same width of the tree on which the point $m$ lies (i.e. on depth $j$ of the tree), shows that $\Upsilon_j$ is a measurable function and $\Upsilon_j(Y)$ is a discrete random variable.

Next, we have $\nu^*(y) = \lim_j \Upsilon_j(y)$, where we know the limit exists since as we mentioned $\Upsilon_j(y)$ form a Cauchy sequence (hence a converging sequence) by definition. It suffices to check whether $\{y : \nu^*(y) \in B\}$ is a Borel set for any closed box $B$ (i.e., $B$ is a hyperrectangle parallel to the coordinate axes). Since

$$\{y : \nu^*(y) \in B\} = \bigcap_{j=1}^n \{y : B^U_j \leq \nu^j(y) \leq B^L_j\},$$

where $\nu^j(y)$ denotes the $j$-th coordinate of $\nu^*$, and $B^L_j$ and $B^U_j$ are the upper and lower bounds of the box $B$ for the $j$-th coordinate, it suffices to show that the sets $\{y : B^L_j \leq \nu^j(y) \leq B^U_j\}$ are measurable. Note that since the sequence is converging

$$\lim_i \Upsilon_i^j(y) = \inf_{i \geq 1} \sup_{k \geq 1} \Upsilon_k^j(y).$$

Next

$$\{y : B^L_j \leq \lim_i \Upsilon_i^j(y) \leq B^U_j\}$$

$$= \{y : \inf_{i \geq 1} \sup_{k \geq 1} \Upsilon_k^j(y) \leq B^U_j\}$$

$$\cap \{y : \Upsilon_k^j(y) \leq B^L_j + l^{-1}\}$$

$$\cap \{y : \Upsilon_k^j(y) \leq B^U_j\}.$$ Finally note that the events $\{y : B^L_j \leq \Upsilon_k^j(y)\}$ and $\{y : \Upsilon_k^j(y) \leq B^U_j + l^{-1}\}$ are measurable since as we showed $\Upsilon_k$ are measurable, and the sets $\mathbb{R} \times \ldots (\mathbb{R}^n, \mathbb{R}^n) \times \mathbb{R}$ are Borel sets in $\mathbb{R}^n$. This completes the proof. □

We will now argue that the estimator from Algorithm 1 attains the minimax rate. The ideas we use are strongly inspired by the works of [18] and [3]. We start with a simple lemma.

Lemma II.5: Suppose we are testing $H_0 : \mu = \nu_1$ vs $H_A : \mu = \nu_2$ for $\|\nu_1 - \nu_2\| \geq C\delta$ for some $C > 2$. Then the test

$$\psi(Y) = 1(\|Y - \nu_1\| \geq \|Y - \nu_2\|)$$

satisfies

$$\sup_{\mu : \|\mu - \nu_1\| \leq \delta} \mathbb{P}_\mu(\psi = 1) \vee \sup_{\mu : \|\mu - \nu_2\| \leq \delta} \mathbb{P}_\mu(\psi = 0) \leq \exp\left(-\frac{(C - 2)^2 \delta^2}{8\sigma^2}\right).$$

Proof: Observe that

$$\|Y - \nu_1\|^2 - \|Y - \nu_2\|^2 = (2\mu + \xi)^\top (\nu_2 - \nu_1) + \|\nu_2\|^2 - \|\nu_2\|^2.$$

Suppose $\|\mu - \nu_1\| \leq \delta$. Then $\mu = \nu_1 + \eta$, $\|\eta\| \leq \delta$ and hence

$$2(\mu + \xi)^\top (\nu_2 - \nu_1) + \|\nu_2\|^2 - \|\nu_2\|^2$$

$$= 2\nu_2^\top \nu_2 - 2\nu_1^\top \nu_1 + 2\xi^\top (\nu_2 - \nu_1)$$

$$+ \|\nu_2\|^2 - \|\nu_2\|^2$$

$$= -2\nu_1^\top \nu_2 + 2\nu_2^\top (\nu_2 - \nu_1) + 2\xi^\top (\nu_2 - \nu_1)$$

We have $2\nu_2^\top (\nu_2 - \nu_1) \leq 2\delta\|\nu_1 - \nu_2\| \leq 2\delta^2\|\nu_1 - \nu_2\|^2$. Hence the above is a normal with mean at most $(-1 + \frac{\delta}{2})\|\nu_1 - \nu_2\|^2 < 0$ (assuming $C > 2$) and variance equal to $4\sigma^2\|\nu_1 - \nu_2\|^2$. By a standard bound on the normal distribution cdf [26] see Section 2.2.1] we have that

$$P(N(m, \tau^2) \geq 0) \leq \exp(-m^2/(2\tau^2)),$$
for $m < 0$, therefore the type I error of the test is bounded by
\[
\exp \left(- \left(1 - \frac{2}{C} \right)^2 \frac{\|v_1 - v_2\|^2}{8\sigma^2} \right) 
\leq \exp \left(- (C - 2)^2 \frac{\delta^2}{8\sigma^2} \right).
\]
By symmetry the same argument holds true for the type II error, namely when $\|\mu - v_2\| \leq \delta$. \hfill \square

**Remark II.6:** It is not too hard to see that this Lemma extends to centered sub-Gaussian noise. In other words if one supposes that $\xi$ satisfies $E\xi = 0$ and $\sup_{v \in S^{n-1}} E \exp(\lambda v^T \xi) \leq \exp(\sigma^2 \lambda^2/2)$ (where $S^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$) for some $\sigma > 0$, the result becomes:
\[
\sup_{\mu, \|\mu - v\| \leq \delta} P_{\mu}(\psi = 1) \leq \sup_{\mu, \|\mu - v\| \leq \delta} P_{\mu}(\psi = 0) 
\leq \exp \left(- (C - 2)^2 \frac{\delta^2}{8\sigma^2} \right).
\]
Since Lemma II.5 is the only place which explicitly uses the Gaussian distribution (in the upper bound analysis), this automatically extends our upper bound results in the bounded $K$ case, for any centered sub-Gaussian noise with the change that $\sigma$ has to be substituted with the variance proxy $\tilde{\sigma}$. 

Suppose now, we are given $M$ points $v_1, \ldots, v_M \in K' \subset K$ such that $\|v_i - v_j\| \geq \delta$ and $M$ is maximal\(^3\), i.e., we are given a maximal $\delta$-packing set $K'$ and it is known that $\mu \in K' \subset K$.

**Lemma II.7:** Under the setting described above, let $i^* \in \arg\min_{v \in K' \setminus K} \|Y - v\|$. We will show that the closest point to $Y$, $\nu_{i^*}$, satisfies
\[
P(\|\nu_{i^*} - \mu\| > (C + 1)\delta) \leq M \exp \left(- \frac{(C - 2)^2 \delta^2}{8\sigma^2} \right),
\]
for any fixed $C > 2$.

**Proof:** Define the intermediate random variable
\[
T_i = \max_{j \in [M]} \|v_j - v_i\|,
\]
s.t. $\|Y - v_i\| > 0$, $\|v_i - v_j\| > 0$.

Without loss of generality assume that $\|\mu - v_i\| \leq \delta$ (here note that we have a $\delta$-packing which is also a $\delta$-covering). Next, we have that
\[
P(\|\nu_{i^*} - \mu\| > \delta + C\delta) \leq \sum_{j} P(i^* \in \{ j : \|v_j - v_i\| > C\delta \}) \leq P(T_i > 0),
\]
where the first inequality follows by the triangle inequality and the second because if $i^* \in \{ j : \|v_j - v_i\| > C\delta \}$ we have $T_i \geq \|v_i - v_{i^*}\| > C\delta$. But
\[
P(T_i > 0) = \sum_{j} \sum_{\|v_j - v_i\| > C\delta} \leq M \exp \left(- \frac{(C - 2)^2 \delta^2}{8\sigma^2} \right),
\]
by Lemma II.5. This is what we wanted to show. \hfill \square

Finally we will need the following simple lemma.

\(\text{Lemma II.8:}\) The function $\varepsilon \mapsto M^{loc}(\varepsilon)$ is monotone non-increasing.

**Remark II.9:** This lemma heavily uses the fact that $K$ is a convex set.

**Proof:** It suffices to show that the function $\varepsilon \mapsto M(\varepsilon/c, B(\theta, \varepsilon) \cap K)$ is non-increasing for any fixed $\theta \in K$.

Upon rescaling one realizes that this is equivalent to packing the set $[\frac{1}{2}(K - \theta)] \cap B(1)$ at a $1/e$ distance, where $B(1) = B(0, 1)$ is the unit ball centered at 0. Now we will show that if $\varepsilon' < \varepsilon$ we have $[\frac{1}{2}(K - \theta)] \cap B(1) \subset [\frac{1}{2}(K - \theta)] \cap B(1)$. Clearly this is implied if we showed that $[\frac{1}{2}(K - \theta)] \subset [\frac{1}{2}(K - \theta)]$. Take a point $x \in [\frac{1}{2}(K - \theta)]$. Hence $x = (k - \theta)/\varepsilon = 0(\varepsilon - \varepsilon')/\varepsilon + \varepsilon'/\varepsilon(k - \theta)/\varepsilon'$ for some $k \in K$. Since $0, (k - \theta)/\varepsilon \subseteq [\frac{1}{2}(K - \theta)]$ and the set $[\frac{1}{2}(K - \theta)]$ is convex, this completes the proof. \hfill \square

Finally we are in a good position to show the main result regarding the estimator of Algorithm 1.

**Theorem II.10:** The estimator from Algorithm 1 returns a vector $\nu^*$ which satisfies the following property
\[
E\|\mu - \nu^*\|^2 \leq C\varepsilon^2,
\]
for some universal constant $C$. Here $\varepsilon^* = \varepsilon_J$, and $J^*$ is the maximal $J \geq 1, J \in \mathbb{N}$, such that $\varepsilon_J := \frac{d(\varepsilon/c/2)^{j+1}}{\varepsilon/c^{j+1}}$ satisfies
\[
\varepsilon_J^2 > 16 \log M^{loc}(\varepsilon_J^{c/(c/2 - 3)})(c/e-c) \sqrt[16]{\log 2}, \quad (II.1)
\]
or $J^* = 1$ if no such $J$ exists. We remind the reader that $c$ is the constant from the definition of local entropy, which is assumed to be sufficiently large.

**Proof:** Combining the results of Lemma II.7 with $c = 2(C + 1)$ where $c$ is the constant from the definition of local packing entropy) and Lemma II.8 we can conclude that for any $2 \leq j \leq J$
\[
P(\|\mu - \nu_j\| > \frac{d}{2^{j-1}}) \leq \frac{d}{2^{j-1}} \nu_j - 1 \leq M^{loc}(\text{d}) \exp \left(- \frac{(C - 2)^2 d^2}{(2d^{j-1})^2(8\sigma^2)} \right).
\]
where $M^{loc}$ is the packing sets from Algorithm 1 corresponding to $\nu_j - 1$. Since the bound does not depend on $\nu_j - 1$ we can drop it from the conditioning. Telescoping this bound (i.e., using that for $k$ events $\{A_k\}_{i \in [k]}$ such that $P(A_i) > 0, i \in [k - 1]$, it always holds that $P(A_k) \leq P(A_k|A_{k-1}^c) + P(A_{k-1}|A_{k-2}^c) + \ldots + P(A_2|A_1^c) + P(A_1)$, which can be proved by induction) we obtain
\[
P(\|\mu - \nu_j\| > \frac{d}{2^{j-1}}) \leq M^{loc}(\text{d}) \sum_{j=1}^{J-1} \exp \left(- \frac{(C - 2)^2 d^2}{(2d^{j-1})^2(8\sigma^2)} \right).
\]

Finally we will need the following simple lemma.

\(^3\)We comment once again, that it is not the maximality that is important; rather it is important for the packing set to also be a covering set.
where for brevity we put
\[ a = \exp\left(-\frac{(C - 2)^2d^2}{(2\sqrt{(J - 1)}(C + 1)^2)(8\sigma^2)}\right), \]
and we are assuming that \( a < 1 \). So if one sets
\[ \varepsilon_J = \frac{(C - 2)^2d^2}{2\sqrt{(C + 1)^2)(8\sigma^2)}}, \]
we have that if \( \varepsilon_J^2/(8\sigma^2) > 2\log M_{\text{loc}}(\varepsilon_J) \) and \( a = \exp(-\varepsilon_J^2/(8\sigma^2)) < 1/2 \), the above probability will be bounded from above by \( 2\exp(-\varepsilon_J^2/(16\sigma^2)) \). Since \( 2\log M_{\text{loc}}(\varepsilon_J) < \frac{2}{\log M_{\text{loc}}(\varepsilon_J)}(\frac{C + 1}{C - 2}) \), this condition is implied when
\[ \frac{\varepsilon_J^2}{\sigma^2} > 16\log M_{\text{loc}}(\varepsilon_J) \cap 16\log 2. \] (II.3)

By the triangle inequality we have that
\[ ||\nu^* - \mu|| \leq ||\nu^* - Y_J|| + ||Y_J - \mu|| \leq 3\varepsilon_J \frac{C + 1}{C - 2}, \] (II.4)
with probability at least \( 1 - 2\exp(-\varepsilon_J^2/(16\sigma^2)) \) which holds for all \( J \) satisfying (II.3). Here we want to clarify that the last inequality in (II.4) follows from the fact that \( ||\nu^* - Y_J|| \leq d/2J^2 \), as seen when we verified that \( Y \) forms a Cauchy sequence. Let \( J^* \) be selected as the maximum \( J \) such that (II.3) holds, or otherwise if such \( J \) does not exist \( J^* = 1 \). Let \( \kappa = \frac{3\varepsilon_J^2}{\sigma^2}, \theta = 2 \) and \( C = \frac{1}{\theta} \). We have established that the following bound holds:
\[ \Pr(||\nu^* - \nu^\dagger|| > \kappa\varepsilon_J) \leq C\exp(-C\varepsilon_J^2/\sigma^2)\mathbb{I}(J > 1) \]
for all \( 1 \leq J \leq J^*, \) where this bound also holds in the case when \( J^* = 1 \) by exception. Observe that we can extend this bound to all \( J \in \mathbb{Z} \) and \( J \leq J^*, \) since for \( J < 1 \) we have \( \kappa\varepsilon > 6 \) and so
\[ \Pr(||\nu^* - \nu^\dagger|| > \kappa\varepsilon_J) \leq C\exp(-C\varepsilon_J^2/\sigma^2)\mathbb{I}(J^* > 1). \]

Now for any \( \varepsilon_{J-1} > x \geq \varepsilon_J \) for \( J \leq J^* \) we have that
\[ \Pr(||\nu - \nu^\dagger|| > 2\kappa x) \leq \Pr(||\nu - \nu^\dagger|| > \kappa\varepsilon_{J-1}) \leq C\exp(-C\varepsilon_{J-1}^2/\sigma^2)\mathbb{I}(J^* > 1) \leq C\exp(-C\varepsilon^2/\sigma^2)\mathbb{I}(J^* > 1), \]
where the last inequality follows due to the fact that the map \( x \mapsto C\exp(-C^2x^2/\sigma^2) \) is monotonically decreasing for positive reals. We will now integrate the tail bound:
\[ \Pr(||\mu - \nu^\dagger|| > 3\kappa x) \leq \Pr(||\mu - \nu^\dagger|| > 2\kappa x) \leq C\exp(-C^2x^2/\sigma^2)\mathbb{I}(J^* > 1). \] (II.5)
which holds true for \( x \geq \varepsilon \) (for \( \varepsilon > 0 \); if \( \varepsilon = 0 \), that means \( \sigma = 0 \) in which case we know the algorithm outputs the correct point), where \( \varepsilon = \varepsilon_{J^*} = \frac{(C - 2)d}{(C + 1)^2(8\sigma^2)}, \) always (since even if \( J^* = 1 \) by exception, this bound is still valid).

We have
\[ \mathbb{E}||\mu - \nu^\dagger||^2 = \int_0^\infty 2x^2\Pr(||\mu - \nu^\dagger|| > x)dx \leq C\varepsilon^2 + \int_0^\infty 2x^2\Pr(-C^2x^2/\sigma^2)\mathbb{I}(J^* > 1)dx = C\varepsilon^2 + 3\varepsilon^2(\varepsilon^2/\sigma^2)\Pr(-C\varepsilon^2/\sigma^2)\mathbb{I}(J^* > 1). \]

Now \( \varepsilon^2/\sigma^2 \) is bigger than a constant \( (16\log 2) \) otherwise \( J^* = 1 \). Hence the above is smaller than \( C\varepsilon^2 \) for some absolute constant \( C \).

We will now formally illustrate that the above estimator achieves the minimax rate. The precise expression of the rate is quantified in the following result:

**Theorem II.11:** Define \( \varepsilon \) as \( \sup\{|\varepsilon_1: \varepsilon^2/\sigma^2 \leq \log M_{\text{loc}}(\varepsilon_1)| \} \), where \( c \) in the definition of local entropy is a sufficiently large absolute constant. Then the minimax rate is given by \( \varepsilon^2/\sigma^2 \) up to absolute constant factors.

**Proof:** First suppose that \( \varepsilon^* \) satisfies \( \varepsilon^2/\sigma^2 > 16\log 2 \). Then for \( \delta^* := \varepsilon^*/4 \) we have \( \log M_{\text{loc}}(\delta^*) \geq \log M_{\text{loc}}(\varepsilon^*) \geq \varepsilon^2/2(\sigma^2) + \varepsilon^2/2(\sigma^2) > 8\varepsilon^2/\sigma^2 + 8\log 2 \) and so this implies the sufficient condition for the lower bound.

On the other hand we know that for a constant \( C > 1 \):
\[ 4C\varepsilon^2/\sigma^2 \geq C\log M_{\text{loc}}(2\varepsilon^*) \geq C\log M_{\text{loc}}(2\varepsilon^2\sqrt{C}) \]
\[ \geq C\log M_{\text{loc}}\left(\frac{\varepsilon^2\sqrt{C}}{c/2 - 3}\right), \]
and so setting \( \delta = 2\varepsilon^*\sqrt{C} - \frac{c}{c/2 - 3} \) we obtain that
\[ \delta^2/\sigma^2 \geq C\log M_{\text{loc}}\left(\frac{\varepsilon^2\sqrt{C}}{c/2 - 3}\right). \]

For \( C = 16 \) this will satisfy the inequality (II.1) (taking into account that \( \varepsilon^2/\sigma^2 > 16\log 2 \), which implies \( \delta^2/\sigma^2 \geq 64\log 2C > 16\log 2 \)). Since the map \( x \mapsto x^2/\sigma^2 \) has diameter \( \sqrt{(64\log 2)\sigma} \) and \( 2\varepsilon^* \leq 16\log 2 \sigma \), we can put points in the diameter of the ball with radius \( 2\varepsilon^* \) such that the packing set has more than \( \exp(64\log 2) \) many points. But that implies that the set \( K \) is entirely inside a ball of radius \( \sqrt{(64\log 2)\sigma} \). In such a case, for the lower bound, we could pick \( \varepsilon \) to be proportional to the diameter of the set (with a small proportionality constant). That will ensure that \( \varepsilon/\sigma \) is upper bounded by some constant (as \( 2\sqrt{(64\log 2)\sigma} \) is bigger than the diameter), and at the same time \( \log M_{\text{loc}}(\varepsilon) \) can be made bigger than a constant (provided that \( c \) in the definition of local packing is large enough) – by taking \( \theta \) (where \( \theta \) is the center of the localized set \( B(\theta, \varepsilon) \cap K \) to be the midpoint of a diameter of the set \( K \) and then placing equispaced points on the diameter. Hence the diameter of the set is a lower bound (up to constant factors) in this case, which is of course always an upper bound too (up to constant factors). So we conclude that either for \( \varepsilon^* \) defined by \( \sup\{\varepsilon: \varepsilon^2/\sigma^2 \leq \log M_{\text{loc}}(\varepsilon)\} \) satisfies \( \varepsilon^2/\sigma^2 > 16\log 2 \) or...
the lower and upper bounds are of the order of the diameter of the set. In summary the rate is given by the $\varepsilon^2 / \sigma^2$. This is true since in the second case, $4\varepsilon^2$ is bigger than the diameter of the set.

In practice it may be challenging to calculate $\varepsilon^*$ precisely, but the following lemma can be useful.

**Lemma II.12:** Suppose that $\varepsilon$ and $\varepsilon'$ are such that $\varepsilon^2 / \sigma^2 > \log M^{loc}(\varepsilon)$ and $\varepsilon^2 / \sigma^2 < \log M^{loc}(\varepsilon')$ and $\varepsilon < \varepsilon'$. Then the rate is given by $\varepsilon^2 / d^2$.

**Proof:** It is clear from the definition of $\varepsilon^*$ that $\varepsilon \geq \varepsilon^*$ while $\varepsilon' \leq \varepsilon^*$. Since $\varepsilon < \varepsilon'$ it follows that $\varepsilon < \varepsilon^*$ which grants the result.

**Remark II.13:** It should be clear that $M^{loc}(\varepsilon)$ can be bounded using Sudakov minoration to yield an upper bound on the minimax rate. We give details in this remark as follows. Suppose that $\varepsilon^2 / \sigma^2 \geq 4 c^{-2} \log M^{loc}(\varepsilon)$. Clearly upon rescaling such an $\varepsilon$ (by $c/2$) we can obtain $\varepsilon' = \varepsilon / c$ (which is of the same order) and is $\geq \varepsilon^*$. The latter follows by the fact that $\frac{\varepsilon^2}{4c^2} \geq \log M^{loc}(\varepsilon) \geq \log M^{loc}(\varepsilon c^2)$ since $c$ is sufficiently large. By Sudakov minoration we have $\log M^{loc}(\varepsilon) \leq \sup_{\theta \in K} \log M^{loc}(\varepsilon) - \frac{\varepsilon^2}{\sigma^2}$, where $w$ denotes the Gaussian width [28, Sec. 5]. It follows that if there exists an $\varepsilon$ such that $\frac{\varepsilon^2}{\sigma^2} \geq \sup_{\theta \in K} \log M^{loc}(\theta \varepsilon) \cap K$ the minimax rate is upper bounded by $\varepsilon^2 / d^2$. An alternative way of seeing that this upper bound on the minimax rate holds, is to use Theorem 2.3. of [1], which shows that the constrained LSE grants this rate. We will also see in our examples, that there exists another universal upper bound on the minimax rate in terms of Kolmogorov complexity. An alternative way of seeing that bound, will be to use the projection estimator $PY$ where $P$ is an orthogonal projection selected in a certain way (cf. Section III-C.1 for more details).

### III. Examples

We now consider several examples, which have been studied previously; nevertheless we find it enlightening to study them from this new perspective. Our examples are also meant to show the reader a couple of methods one can utilize to attain bounds on the local entropy of the constraint set. In addition we will consider an example of convex weak $\ell_p$ balls, and an example of bounded polytopes with $N$ vertices, both of which have not been previously studied to the best of our knowledge. The first example we consider below is concerned with hyperrectangles.

#### A. Hyperrectangles

Let $K = \prod_{i=1}^{n} \left[ -\frac{a_i}{2}, \frac{a_i}{2} \right] \subset \mathbb{R}^n$ be a hyperrectangle. Without loss of generality we will assume that $0 < a_1 \leq a_2 \leq \ldots \leq a_n$. We will show that the following result holds:

**Corollary III.1:** The rate when $K$ is a hyperrectangle as above is given by $(k+2)\sigma^2 \land d^2$ (for $d^2 = \sum_{i=1}^{n} a_i^2$) where $k \in \{0, \ldots, n-1\}$ is such that $(k+1)\sigma^2 \geq \sum_{i=1}^{n-k} a_i^2$ but $(k+2)\sigma^2 > \sum_{i=1}^{n-k-1} a_i^2$, and in the case when $\sum_{i=1}^{n-k} a_i^2 \leq \sigma^2$ the rate is $d^2$.

1) **Upper Bound:** For the upper bound it suffices to consider the case when $\sum_{i=1}^{n} a_i^2 > \sigma^2$ (otherwise the rate is $d^2$ which can trivially be achieved).

Suppose we select $\varepsilon > \varepsilon' \sqrt{k+2} \sigma$, for $\varepsilon'$ being a large constant. We need to make an $\varepsilon/c$ packing of the set $B(\theta, \varepsilon) \cap K$ for any $\theta \in K$. Suppose $M_0$ is the corresponding packing set. Take any two points $x, y \in M_0$. We have

$$\varepsilon/c \leq \|x - y\| \leq \|x_{n-k} - y_{n-k}\| + \|x_{n-k} - y_{n-k}\|$$

$$\leq \sqrt{\sum_{i=1}^{n-k} a_i^2} + \|x_{n-k} - y_{n-k}\|$$

$$\leq \varepsilon/c' \leq \|x_{n-k} - y_{n-k}\| \geq \varepsilon/c''$$

where we denoted by $x_i^n = (x_1, x_{i+1}, \ldots, x_m)^T$. Hence for a large enough $c''$ we will have

$$\|x_{n-k} - y_{n-k}\| \geq \varepsilon/c''$$

2) **Lower Bound:** Next for the lower bound, we will show a lemma first.

**Lemma III.2:** The log cardinality of a maximal packing set of a $k$-dimensional hypercube with side length $\sigma$, to a distance $\sqrt{k}/c$ for some sufficiently large $c$, is at least $\bar{c}k$ for some $\bar{c} > 0$.

**Proof:** For $k = 1$ the assertion is obviously true, so we assume $k \geq 2$. We know that the packing number is at least the ratio between the volumes [28]. The volume of the hypercube is $\sigma^k$. The volume of a sphere of radius $\sqrt{k}/c$ is $\frac{(\sqrt{k}/c)^k \pi^{k/2}}{\Gamma(k/2+1)}$. Taking the ratio we obtain

$$\frac{(k/2+1)}{\sqrt{k}/c} k^{k/2}$$

If $k$ is even, by Stirling’s approximation

$$\Gamma(k/2+1) = (k/2)!$$

$$> \sqrt{2\pi(k/2)^{k/2+1/2}} \exp(-k/2) \exp(1/(6k+1))$$

For $c$ large enough, the log of the ratio can then be lower bounded by $k \log[c/(\sqrt{2}\pi \exp(1/2))] + \frac{1}{2} \log(k/2) + \log(\sqrt{2\pi}) - \frac{1}{6k+1}$. On the other hand, for odd $k$, since $\Gamma$ is increasing (on the interval $[2, \infty)$), we have $\Gamma(k/2+1) \geq \Gamma((k-1)/2+1) > \sqrt{2\pi((k-1)/2)^{(k-1)/2+1/2}} \exp(-(k-1)/2) \exp(1/(6(k-1)+1))$, so that the same conclusion holds.

Going back to the lower bound let us first suppose that $d^2 > \sigma^2$. We will now construct a $[(k+1)/2]$-dimensional hyperrectangle with side length at least $\sigma$ out of the given points. First, assume that $s$ of the $a_i^2$ are at least $\sigma^2$. If $s \geq k$ then we can build a $k$-dimensional hyperrectangle of side lengths at least $\sigma$. In case $s < k$, we know all of the remaining $n - s$ coordinates are $< \sigma$. Hence by greedily taking coordinates until we reach $\sigma^2$ (and note that any such summation will be smaller than $2\sigma^2$) we can construct a
hyperrectangle of dimension at least \([(k+1)/2]\) with sides at least \(\sigma\) (here we are using the fact that \((k+1)\sigma^2 \leq \sum_{i=1}^{n-k} a_i^2\) by assumption). If we build a sphere centered at the center of this hyperrectangle of radius \(\sqrt{(k+1)/2}\sigma\), this sphere contains a hypercube of side \(\sigma\), which is fully inside the hyperrectangle. When \(c\) from the definition of local packing is sufficiently large, this hypercube can be packed with at least \(\exp\left(\epsilon [(k+1)/2]\right)\) points according to the lemma above. Hence for \(d' = \sqrt{(k+1)/2}\sigma\) we have \(\epsilon^2/\sigma^2 \leq \log M_{\text{loc}}(d')\). Thus by rescaling \(\epsilon'\) we can obtain \(\epsilon^2/\sigma^2 \leq \log M_{\text{loc}}(\epsilon')\). Hence the conclusion.

The last case is to consider \(d^2 < \sigma^2\). This case can be handled by the same logic, as in the proof of Theorem II.11 since \(d < \sigma\). This completes the proof.

B. Ellipses

Next we consider the example of ellipses. Let \(K = \{x : \sum_{i} \frac{x_i^2}{a_i^2} \leq 1\}\), where we assume \(0 < a_1 \leq \ldots \leq a_n\). Define the Kolmogorov width [21] as

\[
d_k(K) = \min_{P \in P_k} \max_{\theta \in K} \|P\theta - \theta\|,
\]

where \(P_k\) denotes the set of all \(k\)-dimensional linear projections. It is known that \(d_k(K) = \sqrt{a_{n-k}}\), where \(a_0 = 0\) [see, e.g., [29] and references therein]. Below we will show the following result:

**Corollary III.3:** The minimax rate for ellipses is \((k+1)\sigma^2 \land d^2\), where \(k \in [n]\) is such that \(a_{n-k} \leq (k+1)\sigma^2\) but \(a_{n-k+1} > k\sigma^2\), or \(d^2\) in the case \(a_n \leq \sigma^2\).

1) **Upper Bound:** The upper bound proof is very similar to the bound for the hyperrectangles. We will only focus on the case \(a_n > \sigma^2\) as otherwise the upper bound is trivial. Suppose \(\epsilon^2 > Ck\sigma^2\). We need an \(\epsilon/c\) packing set. Take two points \(x, y\) in that packing set and let \(P\) be the projection achieving the minimax in (III.1). We have

\[
\epsilon/c \leq \|x - y\| \leq \|Px - P - Py\| + \|P - Py\| \leq 2d_k(K) + \|P - Py\|.
\]

But \(d_k(K) \leq (k+1)\sigma^2\) so when \(C\) is sufficiently large we have

\[
\|P - Py\| \geq \epsilon/c'\).
\]

But this is a \(k\)-dimensional set, which is at most a \(k\)-sphere, which means that the packing set is of cardinality at most \(kC\). Hence by potentially rescaling \(\epsilon\) to some bigger value, we will obtain \(\epsilon^2/\sigma^2 \geq \log M_{\text{loc}}(\epsilon)\).

2) **Lower Bound:** For the lower bound, observe that the ellipse, contains a \(k\)-dimensional ball of radius \(\sqrt{k\sigma^2}\). This can be seen by setting the first \(n - k\) coefficients to 0 and then having the set

\[
\sum_{i \geq n-k+1} \frac{x_i^2}{a_i} \leq 1,
\]

and since \(a_{n-k+1} \geq k\sigma^2\) we have the ball inside. This ball can be packed with at least \(kC\) log-packing. Hence the lower bound upon rescaling \(\epsilon^2 = k\sigma^2\) down a bit.

The only case that we have not handled is if \(a_1 \leq \sigma^2\) for all \(i\) (which implies that the diameter is also smaller than \(\sigma\)). But that can be handled as in Theorem II.11 to yield a rate equal to the diameter of the set.

It is worth pointing out here that the LSE fails to be minimax optimal for certain ellipses. This is shown in [31] for instance, see their Lemma 7. For a different example of when the LSE fails refer to [5].

C. Compact Orthosymmetric Quadratically Convex Sets

In this section we consider an example of sets which was first proposed and analyzed in [8]. The compact convex set \(K\) is called orthosymmetric if for \(x = (x_1, \ldots, x_n) \in K\) we have \((\pm x_1, \ldots, \pm x_n) \in K\) for all possible choices of \(\pm\). The set is called quadratically convex if \(K^2 := \{x^2 : x \in K\}\) is a convex set, where \(x^2\) is \(x\) squared entry-wise. Examples of such sets are hyperrectangles and ellipses. For even more examples refer to [8]. We have

**Corollary III.4:** Using the definition of Kolmogorov widths the minimax rate is given by \((k+1)\sigma^2 \land d_0(K)^2\) where \(k\) is such that \(d_k(K)^2 \leq (k+1)\sigma^2\) but \(d_{k+1}(K) > k\sigma^2\). If \(d_0(K)^2 \leq \sigma^2\) we have that the rate is \(d_0(K)^2\) which is up to constants the diameter of the set.

1) **Upper Bound:** The upper bound is the same as in the ellipse case, and in fact this upper bound is always valid. This reflects the fact that one can always use the optimal projection \(PY\) to estimate \(\mu\).

2) **Lower Bound:** For the lower bound we may assume

\[
\min_{P \in P_{k-1}} \max_{\theta \in K} \|P\theta - \theta\| = \min_{S} \max_{\theta \in K} \sum_{i \in S} \theta_i^2 - \sum_{i \in S} \theta_i^2,
\]

where the minimum over \(S\) is taken with respect to all subsets of \([n]\) with exactly \(k-1\) elements. Since the set is quadratically convex the above can be written as

\[
\min_{P \in P_{k-1}} \max_{\theta} \|P\theta - \theta\|^2 = \min_{S} \max_{\theta \in K} \sum_{i \in K} \theta_i^2 - \sum_{i \in S} \theta_i^2
\]

where \(w\) ranges in the set \(\{e : e \in \mathbb{R}^n\} \text{ has exactly } k-1\text{ entries equal to } 1\text{ and the rest are } 0\} \text{ and } 1 \in \mathbb{R}^n\) denotes the vector comprised of 1’s. Since the function \(-w^T t\) is concave this is the same as the problem where \(w\) ranges in the convex hull of these points (call that set \(\mathcal{W}_k\)). By the minimax theorem (we have that both functions needed in the statement of the minimax theorem are linear hence convex and concave) we have

\[
\min_{w \in \mathcal{W}_k} \max_{t \in K} \|t - w^T t\| = \min_{w \in \mathcal{W}_k} \max_{t \in K} \|t - w^T t\|
\]

Now, take \(t^*\) maximizing the above, and \(w^*\) to be equal to 1 when we have one of the \(k-1\) maximal elements in \(t^*\).
We have,
\[ \mathbb{I}^T t^* - w^* T t^* \geq k\sigma^2. \]

Since the set is orthosymmetric we have the hyperrectangle \( \prod_{i \in [n]} [-\sqrt{\frac{\sigma}{2}}, \sqrt{\frac{\sigma}{2}}] \subseteq K \). Hence the logic is the same as in the hyperrectangular case — pick the \( s \) coefficients in \( t^* \) which are bigger than \( \sigma^2 \). If \( s \geq k \) we are all set. If \( s < k \) we know on the remaining they are smaller than \( \sigma^2 \) and they sum up to \( k\sigma^2 \). Hence we can create a larger \((k/2)\) hyperrectangle of side lengths at least \( \sigma \), and the proof can continue as in the hyperrectangular case. The final case to consider is when \( d_0(K)^2 \leq \sigma^2 \), but that can be handled as in Theorem II.11.

### 4.1 Ball

In this section we will replicate a result of [7]. Suppose the set \( K = \{ \theta : \|\theta\|_1 \leq 1 \} \). We will use the fact that
\[
\log(M(\varepsilon/c)) = \log M^{\text{loc}}(\varepsilon) \geq \log(M(\varepsilon/c)) - \log(M(\varepsilon)),
\]
(III.2)

where we denoted with \( \log M(\varepsilon) \) the log cardinality of the maximal packing set of \( K \) at a distance \( \varepsilon \). The bounds (III.2) follow from [30]; actually [30] only prove the bounds for the special case \( c = 2 \), but their results apply more generally.

Using the fact that the log cardinality of a maximal \( \varepsilon \)-packing set of the \( \ell_1 \) ball is given by \( \log(e^{2n}/\varepsilon^2) \) for \( \varepsilon \geq 1/\sqrt{n} \), (otherwise it is \( n \) if \( c \approx 1/\sqrt{n} \) and \( n \log 1/n \) when \( \varepsilon \approx 1/\sqrt{n} \)), for \( c \) large enough we have that
\[
\log M(\varepsilon/c) - \log M(\varepsilon) \approx \frac{\log(e^{2n}/\varepsilon^2)}{\varepsilon^2} \approx \log M(c/c).
\]

Hence, for \( \varepsilon \geq 1/\sqrt{n} \), the equation \( \varepsilon^2/\sigma^2 \approx \log(e^{2n}/\varepsilon^2) \) determines the minimax rate. Suppose that \( \sigma \) is such that \( \log((\sigma^2 \log n)^{1/2}) \approx \log n \). Then setting \( \varepsilon \approx (\log^2 n)^{1/4} \) solves the equation up to constant factors. This matches the example after Theorem 3 of [7] for \( \sigma = 1/\sqrt{n} \). We conclude that

**Corollary III.5:** The minimax rate for the \( \ell_1 \) ball is \( (\log^2 n)^{1/2} \wedge 4 \) for values of \( \sigma \) such that \( \log((\sigma^2 \log n)^{1/2}) \approx \log n \). It is worth pointing out that the orthogonal projection estimator, which works at a minimax rate in all of the above mentioned examples, fails to attain the rate for the \( \ell_1 \) ball [see [31], e.g.]. On the other hand as we argue below the LSE works optimally for the \( \ell_1 \) ball. For an example of when both LSE and the projection estimator fail refer to Example 8 of [31].

### 4.2 Convex Weak \( \ell_p \) Balls for \( 1 < p < 2 \)

In this section we consider an example inspired by weak \( \ell_p \) balls. Consider the quasi-norm \( \|x\|_{p, \infty} = \max_{i \in [n]} |x_i|^{1/p} x_i^* \) on \( \mathbb{R}^n \) where \( x_i^* \) denotes a decreasing rearrangement of \( |x_1|, \ldots, |x_n| \) where \( 1 < p < 2 \). Unfortunately \( \|x\|_{p, \infty} \) is not a norm (so that its unit ball is not convex), but it admits an equivalent norm as follows. Consider
\[
\|x\|_{p, \infty, *} = \max_{i \in [n]} |x_i|^{1/p} x_i^*,
\]
where \( x_i^* = i^{-1} \sum_{j=1}^i x_j^* \). In this section we derive the minimax rate of the Gaussian sequence model for the convex set \( K = \{ x \in \mathbb{R}^n : \|x\|_{p, \infty, *} \leq 1 \} \). We will refer to \( K \) as the convex weak \( \ell_p \) ball. Using [9, Theorem 2] it is not too hard to see that the log cardinality of a maximal \( \varepsilon \)-packing set of \( K \) in Euclidean norm is given by \( \varepsilon^2 \log(ne^{2\varepsilon^2}) \) for values of \( \varepsilon \approx n^{1/2}/p \). Observe that these bounds actually match the known bounds for \( \ell_p \) balls [see [24], e.g.]. Hence we can apply the same logic as in our \( \ell_1 \) example above, in that we can claim that for large enough \( c \)
\[
\log M(\varepsilon/c) - \log M(\varepsilon) \approx \frac{\log(e^{2n}/\varepsilon^2)}{\varepsilon^2} \approx \log M(\varepsilon/c),
\]
for \( \varepsilon \approx n^{1/2}/p \). Solving the equation \( \frac{\varepsilon^2}{\sigma^2} \approx \log(e^{2n}/\varepsilon^2) \) gives \( \varepsilon \approx \sigma^2 \log(n) \), which satisfies \( \log(n/\varepsilon^2) \approx \log n \). We conclude that

**Corollary III.6:** The minimax rate for the \( \ell_p \) ball above is \( \sigma^2 \log(n) \) \wedge \( \text{diam}(K)^2 \) for values of \( \sigma \) such that \( \log(n/\sigma^2) \approx \log n \). Thus \( \sigma^2 \log(n) \) \wedge \( \text{diam}(K)^2 \) is such that

**Remark III.7:** Finally, let us remark that the same rate is valid for \( \ell_p \) balls for \( 1 < p < 2 \). This was first established in [7] (see their Theorem 3 for \( \sigma = 1/\sqrt{n} \)). However, we would like to point out that the convex weak \( \ell_p \) ball above is a larger set than the \( \ell_p \) ball. This can be seen by the elementary inequality \( \sum_{i=1}^{n} |a_i| \leq \left( \sum_{i=1}^{n} |a_i|^p \right)^{1/p} \) for any real numbers \( \{a_i\}_{k=1}^n \) and \( p > 1 \).

### F. Bounds for a Bounded Convex Polytope With N Vertices

In this subsection we derive an upper bound on the minimax rate in the case when the set \( K \subset \mathbb{R}^p \) is a bounded convex polytope with \( N \) vertices. Without loss of generality suppose \( K \) is a polytope of diameter smaller than 1, and it has exactly \( N \) vertices.

#### 1) Upper Bound

By Maurey’s empirical method, one can establish that \( \log M(\varepsilon) \leq (C + 4C/\varepsilon^2 N)^1/2 \) for some absolute constant \( C \) (see Corollary 0.0.4 and Exercise 0.0.6 of [27] and use the fact that the cardinality of a packing set of radius \( 2\varepsilon \) is smaller than the cardinality of a covering set of radius \( \varepsilon \), see (III.4) below). By (III.2) we have
\[
\log M^{\text{loc}}(\varepsilon) \leq \log M(\varepsilon/c) \leq \left[ 4C/\varepsilon^2 \right]^1/2 \log(C + 4C/\varepsilon^2 N/\varepsilon^2). \]

Thus an upper bound on the minimax rate is given by \( \varepsilon^2 \wedge \text{diam}(K) \) where \( \varepsilon := \min \left\{ \varepsilon : \frac{\varepsilon^2}{\sigma^2} \leq \left[ 4C/\varepsilon^2 \right]^1/2 \log(C + 4C/\varepsilon^2 N/\varepsilon^2) \right\} \). As illustrated in Section III-D this rate is in fact achieved for the \( \ell_1 \) ball at least for a regime of \( \sigma \) values. It is worth pointing out that since the upper bound based on Maurey’s argument is nearly the same as that given by Sudakov minoration [27, Corollary 7.4.4], it follows that the LSE will achieve (nearly) the same upper bound on the rate.

#### 2) Lower Bound

In addition, we can show a matching lower bound for some convex polytopes as follows. Suppose there are \( R \geq N \) points \( v_i \in K \) for \( i \in [R] \) satisfying the
following condition

$$\left\| \sum_{i \in [R]} v_i \theta_i \right\| \geq \kappa_c \|\theta\| - f(n, p, R), \quad (\text{III.3})$$

for any $\theta$ in the $2 \times \ell_1$ ball of $\mathbb{R}^R$ for some small non-negative function $f(n, p, R)$, and for some positive constant $\kappa_c > 0$. By a sparse Varshamov-Gilbert lemma [11, Lemma 10.12] one can find $L \geq \exp(c_1 \log R/4k)$ vectors $\{w_i\}_{i \in [L]}$ in the set $\{w \in \{0, 1\}^R : \rho_H(w) = k\}$ where $\rho_H$ is the Hamming distance, such that $\rho_H(w_i, w_j) \geq c_k$. Now set $x_i = \sum_{j \in [R]} v_j w_{ij}$, and observe that $\|x_i \geq \|x_j \|- f(n, p, R) \geq \kappa_c$. It follows that for $k = \frac{c_2}{1 + f(n, p, R)}$, the set $\{x_i\}_{i \in [L]}$ is a $(\varepsilon)$ packing set. Thus log $M'(\varepsilon) \geq c_1 \left(1 + f(n, p, R)\right) \log R/\varepsilon^2$. Now one can use (III.2) coupled with the upper bound on log $M'(\varepsilon)$ via Maurey’s argument above, to claim that for a sufficiently large $c$ the log $M'(\varepsilon) \geq \frac{c_2}{\varepsilon^2} \log \frac{R}{\varepsilon^2}$ for $\varepsilon \leq f(n, p, R)$. It follows that if the solution $\varepsilon^*$ to the equation $\frac{c_2}{\varepsilon^2} \geq f(n, p, N)$ is $\varepsilon^2 \wedge \text{diam}(K)$ is a lower bound on the rate. Further since $R \geq N$ then the lower and upper bounds would match provided that $\varepsilon^* \geq f(n, p, R)$.

One instance when a similar scenario can appear in practice is when $K = X\beta$ for $\beta \in \ell_1^p(1)$, where we denote the unit $\ell_1$ ball in $\mathbb{R}^p$ with $\ell_1^p(1)$. Assuming that $\max_{j \in [p]} \|X_j\| \leq 1$, it follows that $K$ is a symmetric polytope with at most $N \leq 2p$ vertices. In this case one can see that the calculations above recover the bounds given in [23, Theorems 3 and 4] for the $\ell_1$ ball in the case when $\sigma = \frac{1}{\sqrt{p}}$. Here the quantity $f(n, p, N)$ can be taken as $\sqrt{\log p/N}$. One example of a matrix $X$ that satisfies condition (III.3) with high probability is if the rows of $X$ consist of i.i.d. $N(0, I_p)/\sqrt{CN}$ for a sufficiently large $C^2$ variables. Then with high probability it can be shown the columns of $X$ are bounded in $\ell_2$ norm see [23, Appendix I], and also by [23, Proposition 1] (III.3) is satisfied by $R = p \geq N$ points.

### G. Cartesian Product of Sets

In this section we consider the example when $K = K_1 \times K_2$ is a Cartesian product of two closed bounded convex sets. Intuitively it should be clear that if one has a minimax rate optimal estimator on $K_1$ and a minimax rate optimal estimator on $K_2$ by running them separately one will obtain at most twice the maximum of the two rates. On the other hand, for the lower bound it is clear that either of the two minimax rates are lower bounds on the minimax rate over $K$. Below we make this intuition precise by using local packing entropy calculations.

1) **Upper Bound:** We begin by reminding the reader that

$$M(2\delta, S) \leq N(\delta, S) \leq M(\delta, S), \quad (\text{III.4})$$

where $M$ and $N$ denote the maximal packing and minimum covering numbers of the (totally) bounded set $S \subset \mathbb{R}^n$ in Euclidean norm, and the $\delta$ (or $2\delta$) indicates at what distance we are packing or covering (see e.g. [28, Lemma 5.5]).

Consider now a fixed point $(x^0, y^0) \in K$ such that $x^0 \in K_1$ and $y^0 \in K_2$ are arbitrary points. Let $N_1$ be a minimal covering of the set $B(x^0, \varepsilon) \cap K_1$ and $N_2$ be a minimal covering of the set $B(y^0, \varepsilon) \cap K_2$ at a distance $\varepsilon/4$. Put $\tilde{N} = N_1 \times N_2$. Consider $N' = \Pi_B((x^0, y^0), \varepsilon) \cap K$ which is the projection of $\tilde{N}$ onto the closed convex set $B((x^0, y^0), \varepsilon) \cap K$. We will show that $N'$ is a covering of $B((x^0, y^0), \varepsilon) \cap K$.

First let us verify that for a point $(x, y) \in N'$ we have $\|\langle x, y \rangle - \langle x^0, y^0 \rangle\| \leq \varepsilon$. This is so simply by the fact that we projected $\tilde{N}$ onto the set $B((x^0, y^0), \varepsilon) \cap K$. Now for an arbitrary point $(\bar{x}, \bar{y}) \in B((x^0, y^0), \varepsilon) \cap K$ let us find $(x', y') \in N'$ such that $\|\langle \bar{x}, \bar{y} \rangle - \langle x', y' \rangle\|$ is small. Let $\bar{x}$ be the point closest to $\bar{x}$ from $N_1$ and similarly let $\bar{y}$ be the point closest to $\bar{y}$ from $N_2$. Define $(x', y') = \Pi_B((x^0, y^0), \varepsilon) \cap K(\bar{x}, \bar{y}) \in N'$. We have

$$\|\langle \bar{x}, \bar{y} \rangle - \langle x', y' \rangle\| \leq \|\bar{x} - x'\| + \|\bar{y} - y'\| \leq \frac{\varepsilon}{2\varepsilon},$$

where in the above the first inequality follows by the fact that $\|\bar{x}, \bar{y}\| = B((x^0, y^0), \varepsilon) \cap K$ and the projection does not increase the distance between the point and any point in the set $B((x^0, y^0), \varepsilon) \cap K$, and the last inequality is true because $\|\bar{x} - x^0\| \leq \varepsilon$ and similarly $\|\bar{y} - y^0\| \leq \varepsilon$ and the definitions of $N_1$ and $N_2$. Now using (III.4), we conclude that log $M'(\varepsilon) \geq 2(\log M(K_1(\varepsilon/4C)) \cap \log M(K_2(\varepsilon/4C)))$, where we denoted with $M'(\varepsilon) \varepsilon$ the local packing entropy of $K_1$ of radius $\varepsilon$ at a distance $\varepsilon/4C$ (instead of $\varepsilon/c$) and similarly for the term $M'(\varepsilon, c)$.

### 2) Lower Bound

In this section we establish an lower bound on the rate. Let $(x^0, y^0) \in K$ be a point where $x^0 \in K_1$ and $y^0 \in K_2$ are arbitrary points. Consider two maximal packing sets $M_1$ and $M_2$ of $B(x^0, \varepsilon/2) \cap K_1$ and $B(y^0, \varepsilon/2) \cap K_2$ at a distance $\sqrt{2}\varepsilon/c$. Let $M$ be a maximal packing set of $B((x^0, y^0), \varepsilon) \cap K$ at a distance $\varepsilon/c$. We claim that

$$\log |M| \geq \log |M_1| + \log |M_2|. \quad (\text{III.5})$$

This is so since the set $M' = M_1 \times M_2$ forms a packing set of $B((x^0, y^0), \varepsilon) \cap K$. To see this we first verify that for all $(x, y) \in M'$ we have $\|\langle x, y \rangle - \langle x^0, y^0 \rangle\| \leq \varepsilon$. This is true since $\|\langle x, y \rangle - \langle x^0, y^0 \rangle\| \leq \|x-x^0\| + \|y-y^0\|$, and the requirements for the points in $M_1$ and $M_2$. Next for any two distinct points in $(x, y), (x', y') \in M'$ (i.e., $x \neq x'$ and/or $y 
eq y'$) we have $\|\langle x, y \rangle - \langle x', y' \rangle\| \geq \|x-x'\|\|y-y'\| \geq \varepsilon/c$. This finishes the proof. Next, (III.5) implies that

$$\log M(K_1(\varepsilon/2\sqrt{2}\varepsilon/c)) \cap \log M(K_2(\varepsilon/2\sqrt{2}\varepsilon/c)) \geq \log M(K_1(\varepsilon/2\sqrt{2}\varepsilon/c)) \cap \log M(K_2(\varepsilon/2\sqrt{2}\varepsilon/c),$$

where as in the upper bound we denoted with $M(K_1(\varepsilon/2\sqrt{2}\varepsilon/c))$ the local packing entropy of $K_1$ of radius $\varepsilon/2$ (instead of $\varepsilon$) at a distance $\varepsilon/2\varepsilon/c$ and similarly for the term $M(K_2(\varepsilon/2\sqrt{2}\varepsilon/c)$, and in the last inequality we used Lemma II.8.
Combining the results from the previous two subsections, and the fact our results are robust to changes in $c$, i.e., to selecting $c$ to be slightly bigger or smaller sufficiently large constant we conclude that:

**Corollary III.8**: The minimax rate up to constant factors is given by $\varepsilon^* = \sup \{ \varepsilon : \varepsilon^2 / \sigma^2 \leq \log M_{K_0}^{\text{loc}}(\varepsilon) \lor \log M_{K_2}^{\text{loc}}(\varepsilon) \}$. (III.6)

**Remark III.9**: Let us remark that the corollary above can give rise to many examples where the minimax rate can be quantified with more interpretable quantities than the local entropies, for instance when $K_1$ and $K_2$ are an ellipse and a hyperrectangle. Of course this bound also extends to the case when $K = \prod_{j=1}^K K_j$ as long as the number of sets $k$ remains fixed, i.e., it does not scale with $n$ (or $\sigma$). Finally we remark that the same logic shows that if one has a set $K$ which is a direct sum $K = K_1 \oplus K_2$, where $K_1 \perp K_2$ are orthogonal bounded and closed convex sets the minimax rate on the sum would be given by $\varepsilon^2 \lor \text{diam}(K)^2$ where $\varepsilon^*$ is determined via equation (III.6). This is so since for any two points $z = x + y$, $z' = x' + y' \in K$ where $x, x' \in K_1$ and $y, y' \in K_2$ we have

$$
\left( \|x - x'\| + \|y - y'\| \right) ^ 2 \geq \|x + y - (x' + y')\|^2 \\
= \|x - x'\|^2 + \|y - y'\|^2 \\
\geq \left( \|x - x'\| + \|y - y'\| \right) ^ 2 / 2,
$$

so that the same proof as above will apply.

**IV. Adaptivity and Admissibility up to a Universal Constant**

In this section we argue that the estimator constructed in Algorithm 1 is adaptive to the true point. It will be beneficial to define local entropy in a slightly different manner than before.

**Definition IV.1**: Let $\theta \in K$ be a point. Consider the set $B(\theta, \varepsilon) \cap K$. For $\theta \in K$ let $M(\nu, \varepsilon, c) := M(\nu/c, B(\theta, \varepsilon) \cap K)$ denote the largest cardinality of an $\varepsilon/c$ packing set in $B(\theta, \varepsilon) \cap K$.

**Remark IV.2**: We would like to underscore the fact that Definition IV.1 does not take a supremum over all points in the set $K$. This small but key difference is what enables us to formalize the adaptive result below.

We first prove the following lemma.

**Lemma IV.3**: Suppose $\nu$ and $\mu$ are two points in $K$ such that $\|\nu - \mu\| < \delta$. Then $M(\nu, \varepsilon, c) \leq M(\mu, 2\varepsilon, 2c)$ for any $\varepsilon > \delta$.

**Proof**: It suffices to show that $B(\nu, \varepsilon) \cap K \subset B(\mu, 2\varepsilon) \cap K$. We will show directly that $B(\nu, \varepsilon) \subset B(\mu, 2\varepsilon)$. Take any point $x \in B(\nu, \varepsilon)$.

By the triangle inequality $\|x - \mu\| \leq \|x - \nu\| + \delta \leq 2\varepsilon$ since we are assuming $\delta < \varepsilon$. This completes the proof.

Using the above lemma, one can modify the proof of Theorem II.10 to arrive at the following adaptive version of the result.

**Theorem IV.4**: The estimator from Algorithm 1 returns a vector $\nu^*$ which satisfies the following property

$$
\mathbb{E} \|\mu - \nu^*\|^2 \leq \tilde{C}\varepsilon^2,
$$

for some universal constant $\tilde{C}$, where $\varepsilon^* = \varepsilon J^*$ and $J^*$ is the maximal $J \geq 1$ such that $\varepsilon J := \frac{d(c/2 - 3)}{2x - c^2}$ satisfies

$$
\frac{\varepsilon^2}{\sigma^2} > 16 \log M \left( \varepsilon J, 2\varepsilon J, \frac{c}{(c/2 - 3) - 2c} \right) \lor 16 \log 2,
$$

of $J^* = 1$ if no such $J$ exists.

The main thing that needs to be modified is the local entropy in the bound (II.2). We omit the details.

The final remark of this section is to observe that due to the minimaxity of the estimator in Algorithm 1, we have that it is admissible up to a universal constant. This is a trivial observation. For any estimator $\tilde{\nu}(Y)$, there exists a point $\theta \in K$ such that

$$
\mathbb{E}[\|\tilde{\nu}(Y) - \theta\|^2] \geq \tilde{c}\varepsilon^2 \lor d^2,
$$

where $\tilde{c}$ is a universal constant. On the other hand we know that $\mathbb{E}[\|\nu^*(Y) - \theta\|^2] \leq C\varepsilon^2 \lor d^2$ where $C$ is another universal constant. Hence the conclusion.

**V. Unbounded Sets With Known $\sigma^2$**

In this section we generalize the results of Section II to the unbounded case with known $\sigma^2$. A new algorithm is needed which runs multiple bounded algorithms and “aggregates” them in a way similar to how we constructed the bounded case algorithm. The only place where knowledge of $\sigma^2$ is used is to “split” the sample into two independent samples.

**A. Lower Bound**

Note that for unbounded convex sets, the lower bound remains valid. Namely, as long as $\log M^{\text{loc}}(\varepsilon) > 4\varepsilon^2 / \sigma^2 \lor 4 \log 2$ the minimax risk is at least $\varepsilon^2 / 8c^2$. Observe also, that for a sufficiently large $c$ the term $4 \log 2$ does not have effect on the lower bound. This is so since any unbounded convex set in $\mathbb{R}^n$ contains a ray [12, see Lemma 1 Section 2.5 e.g.], and therefore, one can position a ball of radius $\varepsilon$ on that ray so that part of the ray with length $2\varepsilon$ is fully in the ball. Then one can put $\exp(4 \log 2)$ balls of radius $\varepsilon/c$ on that ray centered at equispaced points, which will ensure that $\log M^{\text{loc}}(\varepsilon) > 4 \log 2$ for any $\varepsilon$.

**B. Upper Bound**

In this section we describe an algorithm for unbounded convex sets, and show it achieves the minimax rate. We start with a simple lemma. For simplicity we will assume that the given set $K$ is closed, but we remark how to fix our argument for sets that are not necessarily closed in Remark V.10.

**Lemma V.1**: For two convex sets $S, S'$ satisfying $S' \subset S$, we have that $M^{\text{loc}}_{S_0}(\varepsilon) \leq M^{\text{loc}}_{S_0}(\varepsilon)$ for any $\varepsilon > 0$.

**Proof**: Since for any $\theta \in S'$ we have $B(\theta, \varepsilon) \cap S' \subset B(\theta, \varepsilon) \cap S$ the proof is complete.

We first use the knowledge of $\sigma^2$ to “split” the sample. To this end let us draw $\eta \sim N(0, \sigma^2)$ independently from the observed data $Y$. Consider the variables $\tilde{Y}^1 = Y + \eta$ and $\tilde{Y}^2 = Y - \eta$. These variables are independent. Take any fixed point $\nu \in K$. We consider balls centered at $\nu$ with different radii $B(\nu, 1) \cap K, B(\nu, 2) \cap K, \ldots, B(\nu, 2^m) \cap K, \ldots$ and every
time compute the estimator from Algorithm 1 using \( \bar{Y} \) as the “\( Y \) value”. Denote these estimators with \( \{\nu_m\}_{m=1}^{\infty} \). Note that since \( K \) is closed all of these estimators are proper (i.e. they output values in \( K \)). The intuition for constructing these, is that for large enough \( m \) these estimators will have good properties as \( \mu \) will belong to the set \( B(\nu, 2^m) \cap K \). We have the following lemma regarding the sequence of estimators \( \nu_m \).

**Lemma V.2:** All estimators \( \nu_m \) lie in a compact set.

**Remark V.3:** We would like to remark that this compact set depends on \( \bar{Y} \) and the true point \( \mu \). This is not an issue for our analysis since the two samples \( \bar{Y} \) and \( \bar{Y}^2 \) are independent by construction, hence we may consider the first sample as “frozen”.

**Proof:** For brevity throughout the proof we denote \( \bar{Y} \) with \( Y \). Let \( P_K Y \) denote the projection of \( Y \) onto the set \( K \) (this is a well defined operator since \( K \) is assumed to be closed). At some point the radius \( 2^N \) will be so big that \( P_K Y \) will be in the set \( B(\nu, 2^N) \cap K \). From there on, i.e. \( m \geq N \), we will argue that the estimators \( \nu_m \) will be close to the point \( P_K Y \). The first packing set is at distance \( d \left( \frac{2^N - 1}{2^N} \right) \) where \( d \leq 2^m + 1 \) and \( C \) is the constant from Algorithm 1 (such that \( 2(N + 1) = C \)). Let \( x = \|Y - P_K Y\| \). For any point \( \nu \in K \) we have \( \sqrt{x^2 + \|\nu - P_K Y\|^2} \leq \|\nu - Y\| \leq x + \|\nu - P_K Y\| \), where the first inequality follows by the cosine theorem, and the second one from the triangle inequality. On the other hand the closest point \( \bar{\nu} \) from the packing set to \( P_K Y \) satisfies \( \|\bar{\nu} - P_K Y\| \leq \frac{d}{2(N + 1)} \), and therefore

\[
\|\bar{\nu} - Y\| \leq x + \|\bar{\nu} - P_K Y\| \leq x + \frac{d}{2(N + 1)}.
\]

Take \( \bar{\nu} \) to be the closest point to \( Y \). We then have

\[
\sqrt{x^2 + \|\bar{\nu} - P_K Y\|^2} \leq \|\bar{\nu} - Y\| \leq \|\bar{\nu} - Y\| \leq x + \frac{d}{2(N + 1)}.
\]

It follows that

\[
\|\bar{\nu} - P_K Y\|^2 \leq \frac{d}{2(N + 1)} \left( \frac{d}{2(N + 1)} \right)^2 \leq 3 \left( \frac{d}{2(N + 1)} \right)^2,
\]

assuming that \( x \leq \frac{d}{2(N + 1)} \). Since \( C \geq 2 \) this implies that \( \|\bar{\nu} - P_K Y\| \leq \frac{d}{4} \), and thus the point \( P_K Y \) will be in the chosen ball for the second step. We can continue this logic until \( x \geq \frac{d}{2(N + 1)} \). At this point we know that the estimator will be within distance \( \frac{d}{2(N + 1)} \) of the central point, which is at distance at most \( \frac{d}{2(N + 1)} \) from \( P_K Y \), so that the final estimator will be at distance at most \( \frac{d}{2(N + 1)} \) from \( P_K Y \). This completes the proof that all estimators will be on a compact set since the initial ones fall into a ball of radius \( 2^N \) and are also in a compact set.

**Remark V.4:** The lemma above extends to the case where \( K \) is not closed. The only thing that needs to be modified in the proof is that \( P_K Y \) should be interpreted as \( \bar{P}_K Y \) where as usual \( \bar{K} \) is the closure of \( K \).

Define \( C = \frac{\bar{c}}{\bar{c}} - 1 \), where \( \bar{c} \) is the local packing constant from Definition II.2. Once we have established Lemma V.2, we can proceed to propose Algorithm 2. As we mentioned previously, this algorithm runs multiple bounded algorithms and “aggregates” them in a way similar to how Algorithm 1 works.

Before we proceed with the proof of why Algorithm 2 works, we will show that the estimator produced by it is measurable. We have

**Theorem V.5:** We have that \( \nu^* : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a measurable function (with respect to the Borel \( \sigma \)-field). As a consequence \( \nu^*(Y, \eta) \) is a random variable.

**Proof:** We will show that each element in the sequence \( \nu_j \) is measurable. Since they form a Cauchy sequence their limit will also be measurable by an argument similar to the one in Theorem II.4. Throughout the proof, so as to not overburden notation, for the most part we will suppress the dependence of the estimators \( \nu_m \) on \( \bar{y} = y + \eta \) and will simply write \( \nu_m \).

We will also suppress the dependence of \( \nu_j \) on \( y \) and \( \eta \).

We will select a packing set greedily starting with the minimum index that belongs to the ball on the \( k \)-th step, then carving a ball out centered at that minimum index, and exit considering the minimum index that is in the bigger ball but is out of the carved out ball and so on. We will first show that \( \nu_j \) is measurable. For \( \nu_1 \) the big ball on the 1-st step contains all estimators \( \nu_m \) hence we start from \( \nu_1 \). We will show that the event \( \nu_1 = \nu_j \) is a measurable event, and since as we know from before each \( \nu_j \) is measurable, and the identity \( (\omega : \nu_j \in B) = \bigcup_j (y, \eta) : \nu_j = \nu_j \) \( \cap (y, \eta) = \nu_j (y + \eta) \in B \) for any hyperrectangle \( B \) we will have that \( \nu_j \) is measurable. We will now give a little details about the measurability of the

---

*It is not important for the packing set to be maximal as long as it is a covering set. See Theorem V.5 for a specification of how to construct these sets to ensure measurability.
event \((y, \eta : \nu_j(y + \eta) \in B)\). For \((y, \eta : \nu_j(y + \eta) \in B) = (\omega : y + \eta \in B')\) for some Borel set \(B'\) by the measurability of \(\nu_j\). This is a Borel set since the function \((y, \eta) \mapsto y + \eta\) is continuous and hence measurable.

Let us call the index set of the chosen packing (according to the strategy described above), “the index set”. Let for brevity here and throughout the proof \(\omega := (y, \eta)\). We then have the identity:

\[
\begin{align*}
\{ \omega : \Upsilon_1 = \nu_j \} &= \\
\cup_{S,j \in S,|S| \leq M^\omega(r)} \left( \{ \omega : S \text{ is the index set} \} \cap \\
\cap_{s \in S} \{ \omega : \| \nu_j - \tilde{y} \| \leq \| \nu_i - \tilde{y} \| \} \\
\cap_{s \in S, i \leq j} \{ \omega : \| \nu_i - \tilde{y} \| \neq \| \nu_j - \tilde{y} \| \} \right),
\end{align*}
\]

where we put for brevity \(r = d/(4(\tilde{C} + 1))\) and \(\tilde{y}^2 = y - \eta\). Let \(S = (s_1, s_2, \ldots, s_m)\) (note that \(s_1 = 1\) always has to belong in \(S)\). The above events in the latter two intersections are measurable since for two measurable functions \(X\) and \(Y\) the events \(X \leq Y\) and \(X \neq Y\) are measurable, the function \(\| \cdot \|\) is continuous hence measurable, the sum (difference) of two measurable functions is measurable, and the maps \(\nu_j(y + \eta)\) and \(y - \eta\) are measurable (as we argued earlier and by continuity). Now, the event that \(S\) is the index set is

\[
\begin{align*}
\{ \omega : \Upsilon_1 = \nu_j \} &= \\
\cap_{k \geq s_2}^{s_1 - 1} \{ \omega : \| \nu_1 - \nu_k \| \leq r \} \cap \{ \omega : \| \nu_1 - \nu_{s_2} \| > r \} \cap \\
\cap_{k = s_1 + 1}^{s_2 - 1} \{ \omega : \| \nu_1 - \nu_k \| \leq r \} \cup \{ \omega : \| \nu_{s_2} - \nu_k \| \leq r \} \\
\cap \{ \omega : \| \nu_{s_2} - \nu_k \| > r \} \cap \{ \omega : \| \nu_{s_2} - \nu_{s_1} \| > r \} \cap \\
\cdots \\
\cap_{k \geq s_2}^{s_2-1} \{ \omega : \| \nu_1 - \nu_k \| \leq r \} \cap \\
\{ \omega : \| \nu_{s_2} - \nu_k \| \leq r \} \cup \cdots \cup \{ \omega : \| \nu_m - \nu_k \| \leq r \},
\end{align*}
\]

which is clearly measurable (by continuity of \(\| \cdot \|\), and the fact that the difference of measurable functions is measurable). This completes the proof that \(\Upsilon_1\) is measurable. We will now argue that \(\Upsilon_2\) is also measurable using the same trick. Observe that the identity:

\[
\begin{align*}
\{ \omega : \Upsilon_2 = \nu_j \} &= \\
\cup_{S,j \in S,|S| \leq M^\omega(r)} \left( \{ \omega : S \text{ is the index set} \} \cap \\
\cap_{s \in S} \{ \omega : \| \nu_j - \tilde{y} \| \leq \| \nu_i - \tilde{y} \| \} \\
\cap_{s \in S, i \leq j} \{ \omega : \| \nu_i - \tilde{y} \| \neq \| \nu_j - \tilde{y} \| \} \right),
\end{align*}
\]

continues to hold for \(\Upsilon_2\) with the only difference that \(r = d/(8(\tilde{C} + 1))\). We will now show that the event \(\{ \omega : S \text{ is the index set} \} \) continues to be measurable for \(\Upsilon_2\). We have

\[
\begin{align*}
\{ \omega : S \text{ is the index set} \} &= \\
\cap_{k \geq s_2}^{s_1 - 1} \{ \omega : \| \Upsilon_1 - \nu_k \| > \frac{r}{2} \} \cap \{ \omega : \| \Upsilon_1 - \nu_{s_2} \| \leq \frac{r}{2} \} \\
\cap_{k = s_1 + 1}^{s_2 - 1} \{ \omega : \| \Upsilon_1 - \nu_k \| > \frac{r}{2} \} \\
\cup \{ \omega : \| \nu_{s_2} - \nu_k \| \leq r \} \cup \{ \omega : \| \Upsilon_1 - \nu_{s_2} \| \leq \frac{r}{2} \}
\end{align*}
\]

Clearly, all of the above are measurable events, and therefore \(\Upsilon_2\) is measurable. Proving that all subsequent \(\Upsilon_j\) are measurable is the same as proving that \(\Upsilon_2\) is measurable which completes the proof.

Next we prove a modification of Lemma II.7. The setting is as follows. We are given \(M\) points \(\nu_1, \ldots, \nu_M \in K\) such that \(\min \{ \| \nu_i - \mu \| \} \leq \rho\).

**Lemma V.6:** Let \(i^* = \arg\min_i \| \tilde{y}^2 - \nu_i \|\). We will show that the closest point to \(\tilde{y}^2\), \(\nu_{i^*}\), satisfies

\[
\Pr(\| \nu_{i^*} - \mu \| > (C + 1)\rho) \leq M \exp(-(C - 2)^2\rho^2/(16\sigma^2)),
\]

for any fixed \(C > 2\).

**Proof:** Define the intermediate random variable

\[
T_i = \begin{cases} 
\max_{j \in [M]} \| \nu_j - \nu_{i^*} \|, \\
\text{s.t. } \| \tilde{y}^2 - \nu_i \| - \| \tilde{y}^2 - \nu_{i^*} \| \geq 0, \\
\quad \text{and } \| \nu_i - \nu_{i^*} \| > C\rho \\
0, \text{if no such } j \text{ exists},
\end{cases}
\]

Without loss of generality assume that \(\| \mu - \nu_i \| \leq \rho\). Next, we have that

\[
\Pr(\| \nu_{i^*} - \mu \| > \rho + C\rho) \leq \Pr(i^* \in \{ j : \| \nu_j - \nu_i \| > C\rho \}) \leq \Pr(T_i > 0),
\]

where the first inequality follows by the triangle inequality and the second because if \(i^* \in \{ j : \| \nu_j - \nu_i \| > C\rho \}\) we have \(T_i \geq \| \nu_i - \nu_{i^*} \| > C\rho\). But

\[
\Pr(T_i > 0) = \Pr(\exists j : \| \nu_j - \nu_i \| > C\rho, \\
\quad \text{and } \| \tilde{y}^2 - \nu_i \| - \| \tilde{y}^2 - \nu_{i^*} \| \geq 0) \leq M \exp(-(C - 2)^2\rho^2/(16\sigma^2)),
\]

by Lemma II.5 (here we used the fact that \(\xi_i - \eta_i \sim N(0,2\sigma^2)\)). This is what we wanted to show.

**Theorem V.7:** The estimator from Algorithm 2 returns a vector \(\nu^*\) which satisfies the following property

\[
E\| \mu - \nu^* \|^2 \leq C\varepsilon^2,
\]

for some universal constant \(C\), where \(\varepsilon^*\) is the smallest solution to

\[
\frac{\varepsilon^2}{\sigma^2} > 32 \log M^{\loc} \left( \frac{\varepsilon}{c/2 - 3} \right) + 32 \log 2.
\]

We remind the reader that \(c\) is the constant from the definition of local entropy, which is assumed to be sufficiently large.

**Remark V.8:** For \(c\) large enough inequality (V.1) is equivalent to simply

\[
\frac{\varepsilon^2}{\sigma^2} > 32 \log M^{\loc} \left( \frac{\varepsilon}{c/2 - 3} \right),
\]

since one can always take the center of the ball lying on an infinite ray (which exists [12, see Lemma 1 Section 2.5 e.g.],
and then there will exist at least $\exp(\log 2)$ equispaced points on that ray.

**Remark V.9:** Note that the expected value in (V.1) is taken with respect to both $\xi$ and $\eta$. It is clear by Jensen's inequality, that the estimator $E_n^{\nu^*}(Y, \eta)$ satisfies

$$E_n\|\mu - E_n^{\nu^*}(Y, \eta)\|^2 \leq E\|\mu - \nu^*\|^2 \leq C\epsilon^2.$$  

Note that since $E_n^{\nu^*}(Y, \eta) = E[\nu^*(Y, \eta) | Y]$ it is a measurable function of the data $Y$, and therefore achieves the minimax rate as described in Proposition V.11.

**Proof:** Let $\rho = \inf_{\nu} \|\mu - \nu\|$, and let $\bar{\nu}$ be a limiting point of $\nu_j$ such that $\|\mu - \bar{\nu}\|$. Note that $\rho$ is fixed given $Y^{-1}$. We know that for the $N$-th estimator where $N$ is such that $2^N \geq \|\mu - \nu\|$ we have that the conditions of Theorem II.10 are fulfilled and by (II.5) therefore

$$P(\rho > 2\kappa x) \leq P(\|\mu - \nu_N\| > 2\kappa x) \leq C\exp(-C' x^2/\sigma^2)1(J^* > 1),$$

where for brevity we put $a = \exp\left(-\frac{(C-2)^2\sigma^2}{(2^{2J-2}C+1)^2(16\sigma^2)}\right)$, and we are assuming that $\sigma < 1$.

So if one sets $\epsilon_j = \frac{(C-2)^2\sigma^2}{2^{2J-2}C+1}$, we have that if $\epsilon_j^2/(16\sigma^2) > 2 \log M^{loc}(\epsilon_j \frac{4\hat{C}+1}{(C-2)})$ and $\exp(-\epsilon_j^2/(16\sigma^2)) < 1/2$, the above probability will be bounded from above by $2\exp(-\epsilon_j^2/(32\sigma^2))$. Since

$$2 \log M^{loc}(\epsilon_j \frac{4\hat{C}+1}{(C-2)}) \leq 2 \left(\log 2 \vee \log M^{loc}(\epsilon_j \frac{4\hat{C}+1}{(C-2)})\right),$$

this condition is implied when $\epsilon_j^2 > 32 \log M^{loc}(\epsilon_j \frac{4\hat{C}+1}{(C-2)}) \vee \log 2.$

Below constants can change values from line to line. By the triangle inequality we have that $||\nu^* - \mu|| \leq ||\nu^* - \bar{\nu}_j|| + ||\bar{\nu}_j - \mu|| \leq \rho + \epsilon_j \frac{\hat{C}+1}{C-2} \leq \epsilon_j \frac{\hat{C}+1}{C-2}$ with probability at least $1 - 2\exp(-\epsilon_j^2/(32\sigma^2))$. Let $J^{**}$ be selected as the maximum $J$ such that $\frac{\epsilon_j^2}{2^J} > 32 \log M^{loc}(\epsilon_j \frac{4\hat{C}+1}{(C-2)}) \vee \log 2$ otherwise if such $J$ does not exist $J^{**} = 1$. We have shown that for all $J \leq J^{**}$ we have

$$P(||\mu - \nu^*|| > \frac{7 \cdot d}{2^{2J-1}})$$

where the last two summands, come from controlling the probability of the event $\frac{d}{2^{2J-1}(C+1)} < \rho$. Hence for any $x \geq \epsilon^{**} > 0$ (since if $\epsilon^{**} = 0$ then necessarily $\sigma = 0$ in which case the algorithm will return the point $Y^{-1} = \bar{Y}^2 = \mu$) we have

$$P(||\mu - \nu^*|| > 8 x) \leq P(||\mu - \nu^*|| > 7x) \leq C\exp(-C'' x^2/\sigma^2)1(J^{**} > 1)$$

where $\epsilon^{**} = \epsilon_{J^{**}}$. Integrating the tail bound as before we have

$$E||\mu - \nu^*||^2$$

where for brevity we put $a = \exp\left(-\frac{(C-2)^2\sigma^2}{(2^{2J-2}C+1)^2(16\sigma^2)}\right)$, and we are assuming that $\sigma < 1$.

Now $\epsilon^{**2}/\sigma^2$ is bigger than a constant (32 log 2) otherwise $J^{**} = 1$, and similarly for $\epsilon^*$ and $J^*$. Hence the above is smaller than $C\max(\epsilon^{**2}, \epsilon^{**2})$ for some absolute constant $C$. Finally observe that $\epsilon^*$ is smaller than $2\epsilon^{**}$ which is defined as the infimum $\epsilon$ such that

$$\frac{\epsilon^2}{\sigma^2} > 32 \log M^{loc}(\epsilon_2 \frac{(C+1)}{(C-2)}) \vee 32 \log 2,$$
since $M^\text{loc}(x) \geq M^\text{loc}_{B_{\nu,2^N}}(x)$ for any $x$. In addition, since
$$M^\text{loc}(\frac{2(C+1)}{C-2}) \geq M^\text{loc}(\frac{4(C+1)}{(C-2)})$$
(which follows since we have $\frac{2(C+1)}{C} > \frac{C+1}{C-2}$ and $c = 4(C+1) = 2(C+1)$) we conclude that $2^* \geq \epsilon^*$. This completes the proof.

Remark VI.10: In this remark we explain how to fix the above proof for the case when the set $K$ is not necessarily closed. The issue lies in that in this case the estimators $\nu_m$ may not belong to the set $K$, and therefore we might not have a bound on the entropies localized at these points. The fix is simple. Since each estimator $\nu_m \in K$ (where $K$ is the closure of $K$), we can consider a sequence of points $\{\nu_m\}_{i \in \mathbb{N}}$ which has $\nu_m$ as its limiting point and each point $\nu_m \in K$. For instance select each $\nu_m = \alpha_i \nu + (1 - \alpha_i) \nu_m$ for some appropriately chosen $\alpha_i$ which converges to 0 (e.g. $\alpha_i = 1/i$). Note that this preserves measurability, and the selected $\nu_m$ still belong to a compact set, yet are now points in the set $K$

(i.e. this sequence is usually used to prove that the rational numbers are countable). Note that inequality (V.2) continues (i.e. this sequence is usually used to prove that the rational numbers are countable). Note that inequality (V.2) continues. Hence all arguments of the proof will remain valid.

Proposition VI.11: Define $\epsilon^*$ as $\sup \{ \epsilon : \epsilon^2/\sigma^2 \leq \log M^\text{loc}(\epsilon) \}$, where $\epsilon$ in the definition of local entropy is a sufficiently large absolute constant. Then the minimax rate is given by $\epsilon^2$ up to absolute constant factors.

Proof: For $\epsilon^* := \epsilon^*/4$ we have $\log M^\text{loc}(\epsilon^*) \geq \log M^\text{loc}(\epsilon^*/4) \geq \epsilon^2/\sigma^2 = 16\epsilon^2/\sigma^2$ and so this implies the sufficient condition for the lower bound (note that here we don’t have a constant $4 \log 2$ per the comment in Section V-B).

On the other hand we know that for a constant $C > 1$:
$$4C\epsilon^2/\sigma^2 \geq C \log M^\text{loc}(2\epsilon^*) \geq C \log M^\text{loc}(2\epsilon^*\sqrt{C}) \geq C \log M^\text{loc}(2\epsilon^*\sqrt{C}) \geq C \log M^\text{loc}(\frac{2\epsilon^*\sqrt{C}}{c/2 - 3});$$
and so setting $\delta = 2\epsilon^*\sqrt{C}$ we obtain that
$$\delta^2/\sigma^2 \geq C \log M^\text{loc}(\frac{\delta}{c/2 - 3}).$$
Plugging in $C = 32$ grants the requirement of Remark V.8, which completes the proof.

VI. DISCUSSION

In this paper we studied the minimax rate of the Gaussian sequence model under convex constraints. We proposed a method which is minimax optimal up to constant factors for any bounded convex set $K$, and an extension of the method which is minimax optimal for unbounded sets provided that $\sigma^2$ is known. Unfortunately, our algorithm is not computationally tractable. A natural open question is whether there exist computationally feasible general schemes which achieve the minimax rate for any set $K$. In addition, it is clear that

the algorithm we proposed in this paper has something in common with the constrained LSE, as at each step it is looking for points which are closest to the observed point $Y$. It will be interesting if this connection is studied more closely — in particular if there exist sufficient conditions for $K$ under which the two estimators are sufficiently close. Furthermore, throughout the paper we assumed that the model is well-specified, i.e., that $\mu \in K$. In future work we would like to see whether the techniques proposed here can capture the misspecified case. Another interesting open question is whether one can borrow ideas from this analysis to study the minimax risk under different loss functions, such as $\ell_p$ norms e.g. The biggest roadblock in terms of the upper bound that we currently see is extending Lemma II.5 to this more general setting. Finally an exciting question that remains is whether knowledge of $\sigma^2$ is necessary for the unbounded sets case. Our conjecture is that this is not the case, but at the moment we can only guarantee minimaxity by aggregating bounded estimators for which the knowledge of $\sigma^2$ seems to be required.

APPENDIX A

FINITE STEP ALGORITHM IN THE PRESENCE OF A LOWER BOUND OF $\sigma$

The notation in this section is identical to the one used in Section II-B.

Algorithm 3 Upper Bound Algorithm With Finite Steps Given a Lower Bound on $\sigma$ 

Input: A point $\nu^* \in K$, $\mathcal{J}$ specified in Theorem A.1 

1. $k \leftarrow 1$; 

2. $\Upsilon \leftarrow [\nu^*]$; /* This array is needed solely in the proof and is not used by the estimator */ 

3. for $k \leq \mathcal{J}$ do 

4. Take a $d \leftarrow \frac{d}{2\epsilon^*c}$ maximal packing set $\mathcal{M}_k$ of the set $B(\nu^*, \frac{c}{2\epsilon^*c}) \cap K$; /* The packing sets should be constructed prior to seeing the data */ 

5. $\nu^* \leftarrow \arg\min_{\nu \in \mathcal{M}_k} \| \nu - Y \|$; /* Break ties by taking the point with the least lexicographic ordering */ 

6. $\Upsilon$.append($\nu^*$); 

7. $k \leftarrow k + 1$; 

8. return $\nu^*$

Theorem A.1: Suppose $\underline{\sigma}$ is a known lower bound on $\sigma$. Let $\mathcal{J}$ be defined as the maximum integer $J$ such that
$$\frac{\epsilon^2}{\underline{\sigma}^2} \geq 16 \log M^\text{loc}(\frac{\epsilon}{c/2 - 3}) \vee 16 \log 2, \quad (A.1)$$
where $\epsilon \geq \frac{d(c/2 - 3)}{2\epsilon^*c}$, and let $\mathcal{J} = 1$ if no such integer exists. Then estimator from Algorithm 3 returns a vector $\nu^*$ which satisfies the following property
$$\mathbb{E} \| \mu - \nu^* \|^2 \leq \bar{C}\epsilon^2,$$
for some universal constant $\bar{C}$. Here $\varepsilon^*$ is the same as the one defined in equation (II.1) in Theorem II.10.

Proof: Combining the results of Lemma II.7 (with $c = 2(C + 1)$ where $c$ is the constant from the definition of local packing entropy) and Lemma II.8 we can conclude that

$$\mathbb{P}(\|\mu - Y\| > \frac{d}{2J+1})$$

$$\leq M_{\log}(\frac{d}{2J+1}) \sum_{j=1}^{c-2} \exp\left(-\frac{(C-2)^2 d^2}{(2^{2j}(C+1)^2)(8\sigma^2)}\right)$$

$$\leq M_{\log}(\frac{d}{2J+1}) a(1 + a^{4-1} + a^{16-1} + \ldots) \mathbb{I}(J > 1)$$

$$\leq M_{\log}(\frac{d}{2J+1}) \frac{a}{1-a} \mathbb{I}(J > 1),$$

where $M_j$ are the packing sets from Algorithm 1, and for brevity we put

$$a = \exp\left(-\frac{(C-2)^2 d^2}{(2^{2j}(C+1)^2)(8\sigma^2)}\right),$$

and we are assuming that $a < 1$. So if one sets $\varepsilon_J = \frac{2\log M_{\log}(\frac{d}{2J+1}) a}{2^{2j}(C+1)}$, we have that if $\varepsilon_J^*/(8\sigma^2) > 2\log M_{\log}(\frac{d}{2J+1}) a = \exp(-\varepsilon_J^2/(8\sigma^2)) < 1/2$, the above probability will be bounded from above by $2 \exp(-\varepsilon_J^2/(16\sigma^2))$. Since $2\log M_{\log}(\frac{d}{2J+1}) a < 2 \left(\log 2 \vee \log M_{\log}(\frac{d}{2J+1}) a\right)$ this condition is implied when

$$\frac{\varepsilon_J^2}{\sigma^2} > 16 \log M_{\log}(\frac{d}{2J+1}) a \vee 16 \log 2. \quad (A.2)$$

By the triangle inequality we have that

$$\|\nu^* - \mu\| = \|Y_J - \mu\| \leq \|Y_J - Y\| + \|Y - \mu\|$$

$$\leq 3\varepsilon_J^2 \frac{C+1}{C^2 - 2}, \quad (A.3)$$

with probability at least $1 - 2 \exp(-\varepsilon_J^2/(16\sigma^2))$ which holds for all $J$ satisfying (A.2) which include $J$. Here we want to clarify that the last inequality in (A.3) follows from the fact that $\|Y_J - Y\| < d/2l^2$ when $J > 1$, as seen when we verified that $Y$ forms a Cauchy sequence. Let $J^*$ be selected as the maximum $J$ such that (A.2) holds, or otherwise if such $J$ does not exist $J^* = 1$. Observe that the so defined $J^* \leq \mathcal{J}$, since $\sigma \leq \sigma$ (which also holds in the case when $\mathcal{J} = 1$, because this implies $J^* = 1$). Let $\kappa = \frac{3\varepsilon_J^2}{\sigma^2}$, $\bar{C} = 2$ and $C^* = \frac{1}{16}$. We have established that the following bound holds:

$$\mathbb{P}(\|\mu - \nu^*\| > \kappa \varepsilon_J) \leq \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_J^2}{\sigma^2}) \mathbb{I}(J^* > 1),$$

for all $1 \leq J \leq J^*$, where this bound also holds in the case when $J^* = 1$ by exception. Observe that we can extend this bound to all $J \in \mathbb{Z}$ and $J \leq J^*$, since for $J < 1$ we have $\kappa \varepsilon_J \geq 6$ and so

$$\mathbb{P}(\|\mu - \nu^*\| > \kappa \varepsilon_J)$$

$$\leq 0 \leq \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_J^2}{\sigma^2}) \mathbb{I}(J^* > 1).$$

Now for any $\varepsilon_{J-1} > x \geq \varepsilon_J$ for $J \leq J^*$ we have that

$$\mathbb{P}(\|\mu - \nu^*\| > 2\kappa x) \leq \mathbb{P}(\|\mu - \nu^*\| \geq \kappa \varepsilon_{J-1})$$

$$\leq \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_{J-1}^2}{\sigma^2}) \mathbb{I}(J^* > 1)$$

$$\leq \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_J^2}{\sigma^2}) \mathbb{I}(J^* > 1),$$

where the last inequality follows due to the fact that the map $x \mapsto \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_J^2}{\sigma^2})$ is monotonically decreasing for positive reals. We will now integrate the tail bound:

$$\mathbb{P}(\|\mu - \nu^*\| \geq 3\kappa x) \leq \mathbb{P}(\|\mu - \nu^*\| > 2\kappa x)$$

$$\leq \mathcal{C}_{\exp}(\frac{-C^* \varepsilon_J^2}{\sigma^2}) \mathbb{I}(J^* > 1),$$

which holds true for $x \geq \varepsilon^*$ (for $\varepsilon^* > 0$ if $\varepsilon^* = 0$ we know $\sigma = 0$ and therefore $\sigma = 0$ so we need to run the algorithm ad infinity (or simply output $Y$ in that case)), where $\varepsilon^* = \varepsilon_J = \frac{3\varepsilon_J^2}{\sigma^2} \frac{C+1}{C^2 - 2}$, always (since even if $J^* = 1$ by exception, this bound is still valid).

We have

$$\mathbb{E}(\|\mu - \nu^*\|^2)$$

$$= \int_0^{\infty} 2 x^2 \mathbb{P}(\|\mu - \nu^*\| \geq x) dx$$

$$\leq C'' \varepsilon^* + \int_{3\kappa x^*}^{\infty} 2 x^2 \mathcal{C}_{\exp}(\frac{-C'' \varepsilon^2}{\sigma^2}) \mathbb{I}(J^* > 1) dx$$

$$= C'' \varepsilon^* + C'' \varepsilon^2 \mathcal{C}_{\exp}(\frac{-C'' \varepsilon^2}{\sigma^2}) \mathbb{I}(J^* > 1).$$

Now $\varepsilon^2/\sigma^2$ is bigger than a constant (16 log 2) otherwise $J^* = 1$. Hence the above is smaller than $\bar{C}\varepsilon^2$ for some absolute constant $\bar{C}$.

Acknowledgment

The author is grateful to Siva Balakrishnan for helpful discussions and for pointing him to the relevant articles by Li Zhang, to Ramon van Handel for enlightening discussions on entropy numbers, and to Larry Wasserman for encouragements. He also thank Shamindra Shrotiya who helped with plotting Figure 1. Furthermore, he would like to thank an AE and three anonymous referees for their insightful suggestions which greatly improved the presentation of this article.

References

[1] P. C. Bellec, “Sharp Oracle inequalities for least squares estimators in shape restricted regression,” Ann. Statist., vol. 46, no. 2, pp. 745–780, Apr. 2018.

[2] P. J. Bickel, “Minimax estimation of the mean of a normal distribution when the parameter space is restricted,” Ann. Statist., vol. 9, no. 6, pp. 1301–1309, Nov. 1981.

[3] L. Birgé, “Approximation dans les espaces métriques et théorie de l’estimation,” Zeitshrift für Wahrscheinlichkeitstheorie Verwandte Gebiete, vol. 65, no. 2, pp. 181–237, 1983.

[4] G. Casella and W. E. Strawderman, “Estimating a bounded normal mean,” Ann. Statist., vol. 9, no. 4, pp. 870–878, 1981.

[5] S. Chatterjee, “A new perspective on least squares under convex constraint,” Ann. Statist., vol. 42, no. 6, pp. 2340–2381, 2014.

[6] X. Chen, A. Guntuboyina, and Y. Zhang, “A note on the approximate admissibility of regularized estimators in the Gaussian sequence model,” Electron. J. Statist., vol. 11, no. 2, pp. 4746–4768, Jan. 2017.

[7] D. L. Donoho and I. M. Johnstone, “Minimax risk over $l_p$-balls for $l_p$-estimation,” Probab. Theory Rel. Fields, vol. 99, no. 2, pp. 277–303, 1994.

[8] D. L. Donoho, R. C. Liu, and B. MacGibbon, “Minimax risk over hyperrectangles, and implications,” Ann. Statist., vol. 18, no. 3, pp. 1416–1437, Sep. 1990.
[9] D. E. Edmunds, “Entropy numbers of embeddings of Sobolev spaces in Zygmund spaces,” *Studia Mathematica*, vol. 128, no. 1, pp. 71–102, 1998.

[10] M. Ermakov, “Minimax nonparametric estimation on maxisets,” *J. Math. Sci.*, vol. 244, no. 5, pp. 779–788, Feb. 2020.

[11] S. Foucart and H. Rauhut, “An invitation to compressive sensing,” in *A Mathematical Introduction to Compressive Sensing*. Cham, Switzerland: Springer, 2013, pp. 1–39.

[12] B. Grünbaum, *Convex Polytopes*, vol. 221. Cham, Switzerland: Springer, 2013.

[13] O. Guedon and A. Litvak, “Euclidean projections of a p-convex body,” in *Geometric Aspects of Functional Analysis*. Cham, Switzerland: Springer, 2000, pp. 95–108.

[14] A. Guntuboyina and B. Sen, “Nonparametric shape-restricted regression,” *Stat. Sci.*, vol. 33, no. 4, pp. 568–594, Nov. 2018.

[15] I. A. Ibragimov and R. Z. Khas’minskii, “On nonparametric estimation of the value of a linear functional in Gaussian white noise,” *Theory Probab. Appl.*, vol. 29, no. 1, pp. 18–32, Jan. 1985.

[16] A. Javanmard and L. Zhang, “The minimax risk of truncated series estimators for symmetric convex polytopes,” in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 1633–1637.

[17] I. Johnstone, “Gaussian estimation: Sequence and wavelet models,” Stanford Univ., Palo Alto, CA, USA, Tech. Rep., 2019. [Online]. Available: https://imjohnstone.su.domains/GE_09_16_19.pdf

[18] L. LeCam, “Convergence of estimates under dimensionality restrictions,” *Ann. Statist.*, vol. 1, no. 1, pp. 38–53, Jan. 1973.

[19] A. Nemirovski, “Lectures on probability theory and statistics. Part II: Topics in non-parametric statistics,” in *Probability Summer School*. Berlin, Germany: Springer-Verlag, 1998.

[20] L. Pardo, *Statistical Inference Based on Divergence Measures*. Boca Raton, FL, USA: CRC Press, 2018.

[21] A. Pinkus, *N-Widths in Approximation Theory*, vol. 7. Cham, Switzerland: Springer, 2012.

[22] M. Pinsker, “Optimal filtration of square-integrable signals in Gaussian noise,” *Prob. Info. Transm.*., vol. 16, no. 2, pp. 120–133, 1980.

[23] G. Raskutti, M. J. Wainwright, and B. Yu, “Minimax rates of estimation for high-dimensional linear regression over \( \ell_q \)-balls,” *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[24] C. Schütt, “Entropy numbers of diagonal operators between symmetric Banach spaces,” *J. Approx. Theory*, vol. 40, no. 2, pp. 121–128, 1984.

[25] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Cham, Switzerland: Springer, 2009.

[26] A. W. Van Der Vaart and J. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*. Cham, Switzerland: Springer, 1996.

[27] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[28] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[29] Y. Wei, B. Fang, and M. J. Wainwright, “From Gauss to Kolmogorov: Localized measures of complexity for ellipses,” *Electron. J. Statist.*, vol. 14, no. 2, pp. 2988–3031, Jan. 2020.

[30] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, Oct. 1999.

[31] L. Zhang, “Nearly optimal minimax estimator for high-dimensional sparse linear regression,” *Ann. Statist.*, vol. 41, no. 4, pp. 2149–2175, Aug. 2013.

Matey Neykov was born in Sofia, Bulgaria. He received the B.S. degree in applied mathematics from Sofia University, Bulgaria, in 2009, and the Ph.D. degree in biostatistics from Harvard University, in 2015. He was a Post-Doctoral Researcher at Operations Research and Financial Engineering Department, Princeton University for two years. He was an Assistant Professor of statistics and data science at Carnegie Mellon University, where he has been an Associate Professor, since 2022. His research interests include theoretical statistics, machine learning, and statistical applications to biomedical sciences.