On the minimax rate of the Gaussian sequence model under bounded convex constraints

Matey Neykov

Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213

mneykov@stat.cmu.edu

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 L_2 loss is given by $\epsilon^{*2} \wedge \operatorname{diam}(K)^2$ with

$$\epsilon^* = \sup\left\{\epsilon : \frac{\epsilon^2}{\sigma^2} \leq \log M^{\mathrm{loc}}(\epsilon)\right\},$$

where $\log M^{\mathrm{loc}}(\epsilon)$ denotes the local entropy of the set K, and σ^2 is the variance of the noise. We utilize our abstract result to re-derive known minimax rates for some special sets K such as hyperrectangles, ellipses, and more generally quadratically convex orthosymmetric sets. Finally, we extend our results to the unbounded case with known σ^2 to show that the minimax rate in that case is ϵ^{*2}.

1 Introduction

This paper focuses on the Gaussian sequence model $Y_i = \mu_i + \xi_i$ with n observations, where $\xi_i \sim N(0, \sigma^2)$ are independent and identically distirbuted (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:

$$\inf_{\widehat{\nu}} \sup_{\mu \in K} \mathbb{E}\|\widehat{\nu}(Y) - \mu\|^2, \tag{1.1}$$

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 for any bounded convex set K. The existence of such a general scheme should not be a priori obvious, nonetheless we show it does exist. In order to do that we rely on techniques first proposed by LeCam [1973], Birgé [1983]. That being said, while our result may be expected from these works, it is important to note that it cannot be directly derived by using any previously known results. In their work, LeCam [1973], Birgé [1983] metrize the probability space using the squared Hellinger distance, and their loss function between the estimate and the true parameter is also based on the squared Hellinger distance. For two multivariate Gaussians $N(\nu_1, \sigma^2\mathbb{I})$ and $N(\nu_2, \sigma^2\mathbb{I})$ the squared Hellinger distance is given by $1 - \exp\left(-\frac{\|\nu_1-\nu_2\|^2}{8\sigma^2}\right)$ [Pardo, 2018]. This is markedly distinct from the Euclidean norm of the mean difference $\|\nu_1 - \nu_2\|$ which is what we use to metrize the problem, and results in a more natural loss function for the Gaussian sequence model. In particular, the squared Hellinger distance behaves like $\frac{\|\nu_1-\nu_2\|^2}{8\sigma^2}$ when $\|\nu_1-\nu_2\|$ is "small", but is of constant order when $\|\nu_1-\nu_2\|$ is "large". This difference renders it impossible to use directly previous known results. We would also like to be upfront in that we do not propose a fully satisfactory resolution of this problem for any bounded convex set K, as our general algorithm, although very simple to state presents substantial implementational challenges, and is not computationally tractable. We further extend our result to the unbounded case with known variance of the noise.

The constrained Gaussian sequence model setting has numerous applications. For instance, in the special case when the set K is an ellipse, Wei et al. [2020] show two examples — one of constrained ridge regression with fixed design, and one of nonparametric regression with reproducing kernels which can both be viewed through the Gaussian sequence model perspective. In addition, functional regression with shape-constraints, such as isotonic regression or convex regression can often be viewed through the sequence model lens [see, e.g. Bellec et al., 2018, Guntuboyina and Sen, 2018, and references therein]. In the latter literature often times a preferred estimator is the constrained least squares estimator (LSE), which is known to be minimax optimal in some settings. Additional examples of how the Gaussian sequence model encompasses different models are given in Chatterjee [2014], where the author illustrates how both constrained LASSO with fixed design and isotonic regression can be thought of 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 $\sqrt{n}$ (when $\sigma = 1$). Hence the need arises to find other estimators which always enjoy minimaxity.

1.1 Related Literature

There is a tremendous amount of work on the Guassian sequence model. Here we will only scratch the surface. The interested reader can consult with books on the sequence model and nonparametric statistics such as Johnstone [2011], Nemirovski [1998], Tsybakov [2009].

In one of the most classical results, Pinsker [1980] 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 ℓ_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 Casella and Strawderman [1981], Bickel [1981], Ibragimov and Khas' minskii [1985] 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, Donoho et al. [1990] consider almost the exact same problem as we consider here (with ℓ_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. Javanmard and Zhang [2012] derive the minimax rate for symmetric convex polytopes up to logarithmic factors using the truncated series estimator. Javanmard and Zhang [2012] also point out in their introduction, that "it is still largely unkown how to compute the minimax risk for an arbitrary convex body". Zhang [2013] obtains the minimax rate up to a logarithmic factor for ℓ_q balls for $q \leq 1$, by using an estimator which is a mixture of LSE and a linear projection estimator. Chen et al. [2017] extend results of Chatterjee [2014] 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, Ermakov [2020] shows that the linear minimax risk in the sequence model in ℓ_2 can be explicitly quantified for certain convex sets of the form $K = \{x = \{x_i\}_{i=1}^{\infty} : \sup a_k^{-1} \sum_{j=k}^{\infty} x_j^2 \leq P_0\}$ with $a_k > 0$ being a decreasing sequence. Moreover, Ermakov [2020] 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 Yang and Barron [1999] which is also highly relevant (yet does not consider the sequence model per se). Yang and Barron [1999] 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 LeCam [1973], Birgé [1983] in certain problems metrized with the squared Hellinger distance as we alluded to previously. Therefore Yang and Barron [1999] proposed to study the global entropy instead, which is often easier to handle. We must agree, that local entropy (see Definition 2.2) 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, Yang and Barron [1999] 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 3.4 where we illustrate this by considering the example of an ℓ_1 ball.

1.2 Organization

The paper is structured as follows. We present our main results on bounded convex sets K in Section 2. Section 3 is dedicated to some examples. Section 4 argues that the estimator defined in Section 2 is adaptive to the true point, and it also is admissible up to a universal constant. Section 5 extends our main results from the bounded case to the unbounded case with known σ^2. A brief discussion is given in Section 6.

1.3 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 $[m] = \{1, \dots, m\}$. We use $B(\theta, r)$ to denote a closed Euclidean ball centered at the point θ with radius r. We use $\lesssim$ and $\gtrsim$ 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 $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. Throughout the paper we use log to denote the natural logarithm.

2 Main Results

Here we focus on the following problem. We observe n observations $Y_i = \mu_i + \xi_i$, 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.

2.1 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.

Lemma 2.1 (Fano's inequality). *Let $\mu^1, \ldots, \mu^m$ be a collection of ϵ-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 $\xi \sim N(0, \mathbb{I}\sigma^2)$. Then*

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

In the above $I(Y;J)$ is the mutual information between Y and J, and can be upper bounded by $\frac{1}{m}\sum_j D_{KL}(\mathbb{P}_{\mu^j}||\mathbb{P}_\nu) = \frac{1}{m}\sum_j \frac{\|\mu^j - \nu\|^2}{2\sigma^2} \leq \max_j \frac{\|\mu^j - \nu\|^2}{2\sigma^2}$ for any $\nu \in \mathbb{R}^n$. We will now define local packing entropy.

Definition 2.2 (Local Entropy). *Let $\theta \in K$ be a point. Consider the set $B(\theta, \epsilon) \cap K$. Let $M(\epsilon/c, B(\theta, \epsilon) \cap K)$ denote the largest cardinality of an ϵ/c packing set [see Defintion 5.4 Wainwright, 2019, e.g., for a definition of a packing set] in $B(\theta, \epsilon) \cap K$. Let*

$$M^{\text{loc}}(\epsilon) = \sup_{\theta \in K} M(\epsilon/c, B(\theta, \epsilon) \cap K).$$

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

Lemma 2.3. *We have*

$$\inf_{\widehat{\nu}} \sup_{\mu} \mathbb{E}\|\widehat{\nu}(Y) - \mu\|^2 \geq \frac{\epsilon^2}{8c^2},$$

for any ϵ satisfying $\log M^{\text{loc}}(\epsilon) > 4(\epsilon^2/(2\sigma^2) \vee \log 2)$.

Proof. For a given ϵ we can build an ϵ/c-local packing of cardinality $M^{\text{loc}}(\epsilon)$, 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 $\log M^{\text{loc}}(\epsilon) > 2(\epsilon^2/(2\sigma^2) + \log 2)$. From Fano's inequality it immediately follows that the minimax risk is at least $\frac{\epsilon^2}{8c^2}$. The above is implied when $\log M^{\text{loc}}(\epsilon) > 4(\epsilon^2/(2\sigma^2) \vee \log 2)$. ∎

2.2 Upper Bound

In this subsection we focus on the upper bound. Let $d = \operatorname{diam}(K)$. 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 sufficiently large. The reader will notice that our algorithm contains an infinite loop. This means that our estimator can only be achieved in theory. The good news is that if one knows a lower bound on σ, 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 do not provide these details so as to not overburden the presentation.

Before we proceed we pause to observe a quick fact about the packing sets that are introduced in Algorithm 1. It is simple to see that if one takes the union of all points from the packing sets on all levels, these points form a countable dense subset of K, and hence any point in K is potentially achievable in the limit. Furthermore, as we will see later (see Lemma 5.2) if the point $Y \in K$, Algorithm 1 will always output the point Y. The latter is clearly a desirable property, since when $\sigma = 0$, one needs to pick the observed point to achieve minimaxity, and our estimator is not given knowledge of σ.

Algorithm 1: Upper Bound Algorithm

Input: A point $\nu^* \in K$

$k \leftarrow 1$;

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

while *TRUE* **do**

  Take a $\frac{d}{2^k(C+1)}$ maximal[1]packing set M_k of the set $B\big(\nu^*, \frac{d}{2^{k-1}}\big) \cap K$; `/* The packing sets should be constructed prior to seeing the data */`

  $\nu^* \leftarrow \operatorname{argmin}_{\nu \in M_k} \|Y - \nu\|$; `/* Break ties by taking the point with the least lexicographic ordering */`

  Υ.append(ν^*);

  $k \leftarrow k + 1$;

return ν^* ; `/* Observe that by definition` Υ `forms a Cauchy sequence`[2]`, so` ν^* `can be understood as the limiting point of that sequence. */`

Before we proceed any further we will argue that the so defined estimator $\nu^* = \nu^*(Y)$ is a measurable function of the data. We have

Theorem 2.4. *The function $\nu^* : \mathbb{R}^n \mapsto \mathbb{R}^n$ is measurable (with respect to the Borel σ-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 Υ_j the elements of the array Υ 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

[2]Here 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.

[2]Take any two points Υ_m and $\Upsilon_{m'}$ for $m' > m$. Then $\|\Upsilon_m - \Upsilon_{m'}\| \le \sum_{i=m}^{m'-1} \|\Upsilon_i - \Upsilon_{i+1}\| \le \sum_{i=m}^{m'-1} d/2^{i-1} \le d/2^{m-2}$, so we have a Cauchy sequence.

form an infinite tree of packing sets rooted at the initial point Υ_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 Eucledian 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 tessalation 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 debth j of the tree), shows that Υ_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 (hyperrectangle parallel to the coordinate axes) B. Since

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

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

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

Next

$$\begin{aligned}
&\{y : B_j^L \leq \lim_i \Upsilon_i^j(y) \leq B_j^U\} \\
&\quad = \{y : \inf_{i\geq 1} \sup_{k \geq i} \Upsilon_k^j(y) \leq B_j^U\} \bigcap \{\omega : B_j^L \leq \inf_{i\geq 1} \sup_{k\geq i} \Upsilon_k^j(y)\} \\
&\quad = \bigcap_{l\geq 1} \bigcup_{i \geq 1} \bigcap_{k\geq i} \{y : \Upsilon_k^j(y) \leq B_j^U + l^{-1}\} \bigcap \bigcap_{i\geq 1} \bigcup_{k \geq i} \{y : B_j^L \leq \Upsilon_k^j(y)\}.
\end{aligned}$$

Finally note that the events $\{y : B_j^L \leq \Upsilon_k^j(y)\}$ and $\{y : \Upsilon_k^j(y) \leq B_j^U + l^{-1}\}$ are measurable since as we showed Υ_k are measurable, and the sets $\mathbb{R} \times \dots (-\infty, B_j^U + l^{-1}] \times \mathbb{R}$ and $\mathbb{R} \times \dots [B_j^L, \infty) \times \mathbb{R}$ 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 LeCam [1973], Birgé [1983]. We start with a simple lemma.

Lemma 2.5. *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 $\psi(Y) = \mathbb{1}(\|Y - \nu_1\| \geq \|Y - \nu_2\|)$ satisfies*

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

Proof. Observe that

$$\|Y - \nu_1\|^2 - \|Y - \nu_2\|^2 = 2(\mu + \xi)^{\mathrm{T}}(\nu_2 - \nu_1) + \|\nu_1\|^2 - \|\nu_2\|^2.$$

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

$$\begin{aligned}
&2(\mu + \xi)^{\mathrm{T}}(\nu_2 - \nu_1) + \|\nu_1\|^2 - \|\nu_2\|^2 \\
&\quad = 2{\nu_1}^{\mathrm{T}}(\nu_2 - \nu_1) + 2\xi^{\mathrm{T}}(\nu_2 - \nu_1) + \|\nu_1\|^2 - \|\nu_2\|^2 + 2\eta^{\mathrm{T}}(\nu_2 - \nu_1) \\
&\quad = -\|\nu_1 - \nu_2\|^2 + 2\eta^{\mathrm{T}}(\nu_2 - \nu_1) + 2\xi^{\mathrm{T}}(\nu_2 - \nu_1)
\end{aligned}$$

We have $2\eta^{\mathrm{T}}(\nu_2 - \nu_1) \leq 2\delta\|\nu_1 - \nu_2\| \leq \frac{2}{C}\|\nu_1 - \nu_2\|^2$. Hence the above is a normal with mean at most $(-1 + \frac{2}{C})\|\nu_1 - \nu_2\|^2 < 0$ (assuming $C > 2$) and variance equal to $4\sigma^2\|\nu_1 - \nu_2\|^2$. By a standard bound on the normal distribution cdf [Van Der Vaart and Wellner, 1996, 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

$$\exp\bigg(- \Big(1 - \frac{2}{C}\Big)^2 \frac{\|\nu_1 - \nu_2\|^2}{8\sigma^2}\bigg) \leq \exp\bigg(- (C-2)^2 \frac{\delta^2}{8\sigma^2}\bigg).$$

By symmetry the same argument holds true for the type II error, namely when $\|\mu - \nu_2\| \leq \delta$. □

Suppose now, we are given M points $\nu_1, \ldots, \nu_M \in K' \subset K$ such that $\|\nu_i - \nu_j\| \geq \delta$ and M is maximal[3], i.e., we are given a maximal δ-packing set of K' and it is known that $\mu \in K' \subset K$.

Lemma 2.6. *Under the setting described above, let $i^* = \operatorname{argmin}_i \|Y - \nu_i\|$. We will show that the closest point to Y, ν_{i^*} satisfies*

$$\mathbb{P}(\|\nu_{i^*} - \mu\| \geq (C+1)\delta) \leq M \exp(-(C-2)^2\delta^2/(8\sigma^2)),$$

for any fixed $C > 2$.

Proof. Define the intermediate random variable

$$T_i = \begin{cases} \max_{j\in[M]} \|\nu_i - \nu_j\|, \text{ s.t. } \|Y - \nu_i\| - \|Y - \nu_j\| \geq 0, \|\nu_i - \nu_j\| \geq C\delta \\ 0, \text{if no such } j \text{ exists,} \end{cases}$$

[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.

Without loss of generality assume that $\|\mu - \nu_i\| \leq \delta$ (here note that we have a δ-packing which is also a δ-covering). Next, we have that

$$\begin{aligned}\mathbb{P}(\|\nu_{i^*} - \mu\| \geq \delta + C\delta) &\leq \mathbb{P}(i^* \in \{j : \|\nu_j - \nu_i\| \geq C\delta\}) \\ &\leq P(T_i > 0),\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(T_i > 0) &= \mathbb{P}(\exists j : \|\nu_j - \nu_i\| \geq C\delta \text{ and } \|Y - \nu_i\| - \|Y - \nu_j\| \geq 0) \\ &\leq M \exp(-(C-2)^2\delta^2/(8\sigma^2)),\end{aligned}$$

by Lemma 2.5. This is what we wanted to show.

□

Finally we will need the following simple lemma.

Lemma 2.7. *The function $\epsilon \mapsto M^{\text{loc}}(\epsilon)$ is monotone non-increasing.*

Remark 2.8. *This lemma heavily uses the fact that K is a convex set.*

Proof. It suffices to show that the function $\epsilon \mapsto M(\epsilon/c, B(\theta, \epsilon) \cap K)$ is non-increasing for any fixed $\theta \in K$. Upon rescaling one realizes that this is equivalent to packing the set $[\frac{1}{\epsilon}(K - \theta)] \cap B(1)$ at a $1/c$ distance, where $B(1) = B(0, 1)$ is the unit ball centered at 0. Now we will show that if $\epsilon' < \epsilon$ we have $[\frac{1}{\epsilon}(K - \theta)] \cap B(1) \subset [\frac{1}{\epsilon'}(K - \theta)] \cap B(1)$. Clearly this is implied if we showed that $\frac{1}{\epsilon}(K - \theta) \subset \frac{1}{\epsilon'}(K - \theta)$. Take a point $x \in \frac{1}{\epsilon}(K - \theta)$. Hence $x = (k - \theta)/\epsilon = 0(\epsilon - \epsilon')/\epsilon + \epsilon'/\epsilon(k - \theta)/\epsilon'$ for some $k \in K$. Since $0, (k - \theta)/\epsilon' \in \frac{1}{\epsilon'}(K - \theta)$ and the set $\frac{1}{\epsilon'}(K - \theta)$ is convex, this completes the proof. □

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

Theorem 2.9. *The estimator from Algorithm 1 returns a vector ν^* which satisfies the following property*

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

for some universal constant $\bar{C}$. Here $\epsilon^* = \epsilon_{J^*}$ and J^* is the maximal $J \geq 1$, $J \in \mathbb{N}$, such that $\epsilon_J := \frac{d(c/2-3)}{2^{J-2}c}$ satisfies

$$\frac{\epsilon_J^2}{\sigma^2} > 16 \log M^{\text{loc}}\left(\epsilon_J \frac{c}{(c/2-3)}\right) \vee 16 \log 2, \tag{2.1}$$

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 2.6 (with $c = 2(C+1)$ where c is the constant from the definition of local packing entropy) and Lemma 2.7 we can conclude that

$$\begin{aligned}
\mathbb{P}(\|\mu - \Upsilon_J\| \geq \frac{d}{2^{J-1}}) &\leq \sum_{j=1}^{J-1} |M_j| \exp\left(- \frac{(C-2)^2 d^2}{(2^{2j}(C+1)^2)8\sigma^2}\right) \\
&\leq M^{\mathrm{loc}}\left(\frac{d}{2^{J-2}}\right) \sum_{j=1}^{J-1} \exp\left(- \frac{(C-2)^2 d^2}{(2^{2j}(C+1)^2)8\sigma^2}\right) \\
&\leq M^{\mathrm{loc}}\left(\frac{d}{2^{J-2}}\right) a(1 + a^{4-1} + a^{16-1} + \ldots)\mathbb{1}(J > 1) \\
&\leq M^{\mathrm{loc}}\left(\frac{d}{2^{J-2}}\right) \frac{a}{1-a}\mathbb{1}(J > 1), \qquad (2.2)
\end{aligned}$$

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

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

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

$$\frac{\epsilon_J^2}{\sigma^2} > 16\log M^{\mathrm{loc}}\left(\epsilon_J \frac{2(C+1)}{(C-2)}\right) \vee 16\log 2. \qquad (2.3)$$

By the triangle inequality we have that

$$\|\nu^* - \mu\| \leq \|\nu^* - \Upsilon_J\| + \|\Upsilon_J - \mu\| \leq 3\epsilon_J \frac{C+1}{C-2}, \qquad (2.4)$$

with probability at least $1 - 2\exp(-\epsilon_J^2/(16\sigma^2))$ which holds for all J satisfying (2.3). Here we want to clarify that the last inequality in (2.4) follows from the fact that $\|\nu^* - \Upsilon_J\| \leq d/2^{J-2}$, as seen when we verified that Υ forms a Cauchy sequence. Let J^* be selected as the maximum J such that (2.3) holds, or otherwise if such J does not exist $J^* = 1$. Let $\kappa = 3\frac{C+1}{C-2}$, $\underline{C} = 2$ and $C' = \frac{1}{16}$. We have established that the following bound holds:

$$\mathbb{P}(\|\mu - \nu^*\| \geq \kappa\epsilon_J) \leq \underline{C}\exp(-C'\epsilon_J^2/\sigma^2)\mathbb{1}(J > 1) \leq \underline{C}\exp(-C'\epsilon_J^2/\sigma^2)\mathbb{1}(J^* > 1),$$

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 $\kappa\epsilon_J \geq 6d$ and so

$$\mathbb{P}(\|\mu - \nu^*\| \geq \kappa\epsilon_J) \leq 0 \leq \underline{C}\exp(-C'\epsilon_J^2/\sigma^2)\mathbb{1}(J^* > 1).$$

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

$$\begin{aligned}
\mathbb{P}(\|\mu - \nu^*\| \geq 2\kappa x) &\leq \mathbb{P}(\|\mu - \nu^*\| \geq \kappa\epsilon_{J-1}) \leq \underline{C}\exp(-C'\epsilon_{J-1}^2/\sigma^2)\mathbb{1}(J^* > 1) \\
&\leq \underline{C}\exp(-C'x^2/\sigma^2)\mathbb{1}(J^* > 1),
\end{aligned}$$

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

$$\mathbb{P}(\|\mu - \nu^*\| \geq 2\kappa x) \leq \underline{C}\exp(-C'x^2/\sigma^2)\mathbb{1}(J^* > 1), \tag{2.5}$$

which holds true for $x \geq \epsilon^*$, where $\epsilon^* = \epsilon_{J^*} = \frac{(C-2)d}{(C+1)2^{J^*-1}}$, always (since even if $J^* = 1$ by exception, this bound is still valid).

We have

$$\begin{aligned}
\mathbb{E}\|\mu - \nu^*\|^2 &= \int_0^\infty 2x\mathbb{P}(\|\mu - \nu^*\| \geq x)dx \\
&\leq C'''\epsilon^{*2} + \int_{2\kappa\epsilon^*}^\infty 2x\underline{C}\exp(-C''x^2/\sigma^2)\mathbb{1}(J^* > 1)dx \\
&= C'''\epsilon^{*2} + C''''\sigma^2\exp(-C'''''\epsilon^{*2}/\sigma^2)\mathbb{1}(J^* > 1).
\end{aligned}$$

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

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

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

On the other hand we know that for a constant $C > 1$:

$$4C\epsilon^{*2}/\sigma^2 \geq C\log M^{\text{loc}}(2\epsilon^*) \geq C\log M^{\text{loc}}(2\epsilon^*\sqrt{C}) \geq C\log M^{\text{loc}}\left(2\epsilon^*\sqrt{C}\frac{c}{c/2-3}\right),$$

and so setting $\delta = 2\epsilon^*\sqrt{C}$ we obtain that

$$\delta^2/\sigma^2 \geq C\log M^{\text{loc}}\left(\delta\frac{c}{c/2-3}\right).$$

For $C = 16$ this will satisfy the inequality (2.1) (taking into account that $\epsilon^{*2}/\sigma^2 > 16\log 2$, which implies $\delta^2/\sigma^2 \geq 64\log 2C > 16\log 2$). Since the map $x \mapsto x^2/\sigma^2 - 16\log M^{\text{loc}}\left(x\frac{c}{c/2-3}\right) \vee 16\log 2$ is non-decreasing, we have that $\delta \geq \epsilon_{J^*}/2$. This shows that the rate in this case is ϵ^{*2}.

Next, suppose that ϵ^* defined by $\sup\{\epsilon : \epsilon^2/\sigma^2 \leq \log M^{\text{loc}}(\epsilon)\}$ satisfies $\epsilon^{*2}/\sigma^2 \leq 16\log 2$. For $2\epsilon^*$, we have $64\log 2 > 4\epsilon^{*2}/\sigma^2 \geq \log M^{\text{loc}}(2\epsilon^*)$. If c in the definition of local packing is large enough, we could put points in the diameter of the ball with radius $2\epsilon^*$ 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 $\sqrt{(64\log 2)}\sigma$ (as $\epsilon^{*2} \leq 16\log 2\sigma^2$). In such a case, for the lower bound, we could pick ϵ to be proportional to the diameter of the set (with a small proportionality constant). That will ensure that ϵ/σ is upper bounded by some constant (as $2\sqrt{(64\log 2)}\sigma$ is bigger than the diameter), and at the same time $\log M^{\text{loc}}(\epsilon)$ can be made bigger than a constant (provided that c in the definition of a local packing is large enough) – by taking θ (where θ is the center of the localized set $B(\theta,\epsilon)\cap K$) 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 ϵ^* defined by $\sup\{\epsilon : \epsilon^2/\sigma^2 \leq \log M^{\text{loc}}(\epsilon)\}$ satisfies $\epsilon^{*2}/\sigma^2 > 16\log 2$ or the lower and upper bounds are of the order of the diameter of the set. In summary the rate is given by the $\epsilon^{*2} \wedge d^2$. This is true since in the second case, $4\epsilon^*$ is bigger than the diameter of the set. □

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

Lemma 2.11. *Suppose that ϵ and ϵ' are such that $\epsilon^2/\sigma^2 > \log M^{\text{loc}}(\epsilon)$ and ${\epsilon'}^2/\sigma^2 < \log M^{\text{loc}}(\epsilon')$ and $\epsilon \asymp \epsilon'$. Then the rate is given by $\epsilon^2 \wedge d^2$.*

Proof. It is clear from the definition of ϵ^* that $\epsilon \geq \epsilon^*$ while $\epsilon' \leq \epsilon^*$. Since $\epsilon \asymp \epsilon'$ it follows that $\epsilon \asymp \epsilon^*$ which grants the result. □

Remark 2.12. *It should be clear that $M^{\text{loc}}(\epsilon)$ 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 $\frac{\epsilon^2}{\sigma^2} \geq 4c^{-2}\log M^{\text{loc}}(\epsilon)$. Clearly upon rescaling such an ϵ (by $c/2$) we can obtain $\epsilon' = \frac{\epsilon c}{2}$ (which is of the same order) and is $\geq \epsilon^*$. The latter follows by the fact that $\frac{(\epsilon c)^2}{4} \geq \log M^{\text{loc}}(\epsilon) \geq \log M^{\text{loc}}(\frac{\epsilon c}{2})$ since c is sufficiently large. By Sudakov minoration we have $\log M^{\text{loc}}(\epsilon) \leq \sup_{\theta\in K}\frac{w(B(\theta,\epsilon)\cap K)^2}{\epsilon^2/c^2}$, where w denotes the Gaussian width [Wainwright, 2019, see Section 5]. It follows that if there exists an ϵ such that $\frac{\epsilon^2}{2\sigma} \geq \sup_{\theta\in K} w(B(\theta,\epsilon)\cap K)$ the minimax rate is upper bounded by $\epsilon^2 \wedge d^2$. An alternative way of seeing that this upper bound on the minimax rate holds, is to use Theorem 2.3. of Bellec et al. [2018], 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 3.3.1 for more details).*

3 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. The first example is concerned with hyperrectangles.

3.1 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 rate is given by $(k+2)\sigma^2 \wedge d^2$ (for $d^2 = \sum_{i\in[n]} a_i^2$) where k is an integer such that $(k+1)\sigma^2 \leq \sum_{i=1}^{n-k} a_i^2$ but $(k+2)\sigma^2 > \sum_{i=1}^{n-k-1} a_i^2$, and in the case when $\sum_{i=1}^n a_i^2 \leq \sigma^2$ the rate is d^2.

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

$$\begin{aligned}\epsilon/c \leq \|x - y\| &\leq \|x_1^{n-k-1} - y_1^{n-k-1}\| + \|x_{n-k}^n - y_{n-k}^n\| \\ &\leq \sqrt{\sum_{i=1}^{n-k-1} a_i^2} + \|x_{n-k}^n - y_{n-k}^n\| \\ &\leq \sqrt{k+2}\sigma + \|x_{n-k}^n - y_{n-k}^n\|,\end{aligned}$$

where we denoted by $x_l^m = (x_l, x_{l+1}, \dots, x_m)^\top$. Hence for a large enough c' we will have

$$\|x_{n-k}^n - y_{n-k}^n\| \geq \epsilon/c'',$$

where $c'' = (c'/c - 1)$. This means, that the packing set, also forms a ϵ/c'' packing on the last $k+1$ coordinates. However, this set can at most be a $(k+1)$-sphere with radius ϵ, and so such a packing number will be bounded by $(k+1)\log(1+2c'') \ll (c'\sqrt{k+2})^2$ [Wainwright, 2019] for a large c'.

3.1.2 Lower Bound

Next for the lower bound, we will show a lemma first.

Lemma 3.1. *The log cardinality of a maximal packing set of a k-dimensional hypercube with side length σ, to a distance $\sqrt{k}\sigma/c$ 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 [Wainwright, 2019]. The volume of the hypercube is σ^k. The volume of a sphere of radius $\sqrt{k}\sigma/c$ is $\frac{(\sqrt{k}\sigma/c)^k \pi^{k/2}}{\Gamma(k/2+1)}$. Taking the ratio we obtain

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

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 Γ 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. $\square$

First suppose that $d^2 > \sigma^2$. We will now construct a $\lceil (k+1)/2 \rceil$-dimensional hyperrectangle out of the given points. First, assume that s of the a_i^2 are at least σ^2. If $s \geq k$ then we can build a k-dimensional hyperrectangle of side lengths at least σ. In case $s < k$, we know all of the remaining $n-s$ coordinates are $< \sigma$. Hence by greedily taking coordinates until we reach σ^2 (and note that any such summation will be smaller than $2\sigma^2$) we can construct a hyperrectangle of dimension at least $\lceil (k+1)/2 \rceil$ with sides at least σ (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 $\sqrt{\lceil (k+1)/2 \rceil}\sigma$, this sphere contains a hypercube of side σ, 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 $\exp(\bar{c}\lceil (k+1)/2 \rceil)$ points according to the lemma above. Hence for $\epsilon' = \sqrt{\lceil (k+1)/2 \rceil}\sigma$ we have ${\epsilon'}^2/\sigma^2 \lesssim \log M^{\mathrm{loc}}(\epsilon')$. Thus by rescaling ϵ' we can obtain ${\epsilon'}^2/\sigma^2 < \log M^{\mathrm{loc}}(\epsilon')$. 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 2.10 since $d < \sigma$. This completes the proof.

3.2 Ellipses

Next we consider the example of ellipses. Let $K = \{x : \sum_i \frac{x_i^2}{a_i} \leq 1\}$, where we assume $0 < a_1 \leq \ldots \leq a_n$. Define the Kolmogorov width [Pinkus, 2012] as

$$d_k(K) = \min_{P \in \mathcal{P}_k} \max_{\theta \in K} \|P\theta - \theta\|, \tag{3.1}$$

where $\mathcal{P}_k$ denotes the set of all k-dimensional linear projections. It is known that that $d_k(K) = \sqrt{a_{n-k}}$, where $a_0 = 0$ [see, e.g., Wei et al., 2020, and references therein]. We will show that the minimax rate is $(k+1)\sigma^2 \wedge d^2$, where k is such that $a_{n-k} \leq (k+1)\sigma^2$ but $a_{n-k+1} > k\sigma^2$, $k = 1, \ldots, n$, or d^2 in the case $a_n \leq \sigma^2$.

3.2.1 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 ϵ/c packing set. Take two points x, y in that packing set and let P be the projection achieving the min at (3.1). We have

$$\epsilon/c \leq \|x - y\| \leq \|x - Px - y + Py\| + \|Px - Py\| \leq 2d_k(K) + \|Px - Py\|$$

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

$$\|Px - Py\| \geq \epsilon/c''.$$

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 ϵ to some bigger value, we will obtain $\epsilon^2/\sigma^2 > \log M^{\mathrm{loc}}(\epsilon)$.

3.2.2 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 $a_i \leq \sigma^2$ for all i (which implies that the diameter is also smaller than σ). But that can be handled as in Theorem 2.10 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 Zhang [2013] for instance, see their Lemma 7. For a different example of when the LSE fails refer to Chatterjee [2014].

3.3 Compact Orthosymmetric Quadratically Convex Sets

In this section we consider an example of sets which was first proposed and analyzed in Donoho et al. [1990]. The compact convex set K is called orthosymmetric if for $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in K$ we have $(\pm x_1, \ldots, \pm x_n)^{\mathrm{T}} \in K$ 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 Donoho et al. [1990].

Using the definition of Kolmogorov widths the minimax rate is given by $(k+1)\sigma^2 \wedge d_0(K)^2$ where k is such that $d_k(K)^2 \leq (k+1)\sigma^2$ but $d_{k-1}^2(K) > k\sigma^2$. 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.

3.3.1 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 μ.

3.3.2 Lower Bound

For the lower bound we may assume

$$\min_{P \in \mathcal{P}_k} \max_{\theta \in K} \|\theta - P\theta\|^2 \geq k\sigma^2.$$

We can only consider projections aligned with the coordinates – there are $\binom{n}{k}$ such projections. Then the optimization is

$$\min_{P \in \mathcal{P}_k} \max_{\theta \in K} \|\theta - P\theta\|^2 \leq \min_{S} \max_{\theta \in K} \sum_i \theta_i^2 - \sum_{i \in S} \theta_i^2,$$

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

$$\min_{P \in \mathcal{P}_k} \max_{\theta} \|\theta - P\theta\|^2 \leq \min_{S} \max_{t \in K^2} \sum_i t_i - \sum_{i \in S} t_i = \min_{w} \max_{t \in K^2} \mathbb{1}^{\mathrm{T}} t - w^{\mathrm{T}} t,$$

whre w ranges in the set $\{e_k : e_k$ has exactly k-entries equal to 1 and the rest are $0\}$. Since the function $-w^{\mathrm{T}} t$ 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

$$\min_{w \in \mathcal{W}_k} \max_{t \in K^2} \mathbb{1}^{\mathrm{T}} t - w^{\mathrm{T}} t = \max_{t \in K^2} \min_{w \in \mathcal{W}_k} \mathbb{1}^{\mathrm{T}} t - w^{\mathrm{T}} t,$$

Now, take t^* maximizing the above, and w^* to be equal to 1 when we have one of the k maximal elements in t^*. We have,

$$\mathbb{1}^\top t^* - {w^*}^\top t^* \geq k\sigma^2.$$

Since the set is orthosymmetric we have the hyperrectangle $\prod_{i\in[n]}[-\sqrt{t_i^*}, \sqrt{t_i^*}] \subset K$. Hence the logic is the same as in the hyperrectangular case — pick the s coefs in t^* which are bigger than σ^2. If $s \geq k$ we are all set. If $s < k$ we know on the remaining they are smaller than σ^2 and they sum up to $k\sigma^2$. Hence we can create a large ($\lceil k/2 \rceil$) hyperrectangle of side lengths at least σ, and the proof can continue as in the hyperrectangle case. The final case to consider is when $d_0(K)^2 \leq \sigma^2$, but that can be handled as in Theorem 2.10.

3.4 ℓ_1 ball

In this section we will replicate a result of Donoho and Johnstone [1994]. Suppose the set $K = \{\theta : \|\theta\|_1 \leq 1\}$. We will use the fact that

$$\log M(\epsilon/c) \geq \log M^{\text{loc}}(\epsilon) \geq \log M(\epsilon/c) - \log M(\epsilon), \tag{3.2}$$

where we denoted with $\log M(\epsilon)$ the log cardinality of the maximal packing set of K at a distance ϵ. The bounds (3.2) follow from Yang and Barron [1999]; actually Yang and Barron [1999] 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 ϵ-packing set of the ℓ_1 ball is given by $\log(\epsilon^2 n)/\epsilon^2$ for $\epsilon \gtrsim 1/\sqrt{n}$, (otherwise it is n if $\epsilon \asymp 1/\sqrt{n}$ and $n \log \frac{1}{\epsilon^2 n}$ when $\epsilon \lesssim 1/\sqrt{n}$ Guedon and Litvak [2000], Schütt [1984]), for c large enough we have that

$$\log M(\epsilon/c) - \log M(\epsilon) \asymp \frac{\log(\epsilon^2 n)}{\epsilon^2} \asymp \log M(\epsilon/c).$$

Hence, for $\epsilon \gtrsim 1/\sqrt{n}$, the equation $\epsilon^2/\sigma^2 \asymp \frac{\log(\epsilon^2 n)}{\epsilon^2}$ determines the minimax rate. Suppose that σ is such that $\log((\sigma^2 \log n)^{1/2} n) \asymp \log n$, and $(\sigma^2 \log n)^{1/4} \gtrsim 1/\sqrt{n}$. Then setting $\epsilon \asymp (\sigma^2 \log n)^{1/4}$ solves the equation up to constant factors. This matches the example after Theorem 3 of Donoho and Johnstone [1994] for $\sigma = 1/\sqrt{n}$.

It is worth pointing out that the orthogonal projection estimator, which works at a minimax rate in all of the aforementioned examples, fails to attain the rate for the ℓ_1 ball [see Zhang, 2013, e.g.].

4 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 4.1. *Let $\theta \in K$ be a point. Consider the set $B(\theta, \epsilon) \cap K$. For $\theta \in K$ Let $M(\theta, \epsilon, c) := M(\epsilon/c, B(\theta, \epsilon) \cap K)$ denote the largest cardinality of an ϵ/c packing set in $B(\theta, \epsilon) \cap K$.*

We first prove the following lemma.

Lemma 4.2. *Suppose ν and μ are two points in K such that $\|\nu - \mu\| < \delta$. Then $M(\nu, \epsilon, c) \leq M(\mu, 2\epsilon, 2c)$ for any $\epsilon > \delta$.*

Proof. It suffices to show that $B(\nu, \epsilon) \cap K \subset B(\mu, 2\epsilon) \cap K$. We will show directly that $B(\nu, \epsilon) \subset B(\mu, 2\epsilon)$. Take any point $x \in B(\nu, \epsilon)$. By the triangle inequality $\|x - \mu\| \leq \|x - \nu\| + \delta \leq 2\epsilon$ since we are assuming $\delta < \epsilon$. This completes the proof. □

Using the above lemma, one can modify the proof of Theorem 2.9 to arrive at the following adaptive version of the result.

Theorem 4.3. *The estimator from Algorithm 1 returns a vector ν^* which satisfies the following property*

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

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

$$\frac{\epsilon_J^2}{\sigma^2} > 16 \log M\left(\mu, 2\epsilon_J \frac{c}{(c/2-3)}, 2c\right) \vee 16 \log 2,$$

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

The main thing that needs to be modified is the local entropy in the bound (2.2). 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 $\widehat{\nu}(Y)$, there exists a point $\theta \in K$ such that

$$\mathbb{E}\|\widehat{\nu}(Y) - \theta\|^2 \geq \bar{c}\epsilon^{*2} \wedge d^2,$$

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

5 Unbounded Sets with known σ^2

In this section we generalize the results of Section 2 to the unbounded case with known σ^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 σ^2 is used is to "split" the sample into two independent samples.

5.1 Lower Bound

Note that for unbounded convex sets, the lower bound remains valid. Namely, as long as, $\log M^{\text{loc}}(\epsilon) > 4\epsilon^2/\sigma^2 \vee 4 \log 2$ the minimax risk is at least $\epsilon^2/8c^2$. 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 [see Lemma 1 Section 2.5 Grünbaum, 2013, e.g.], and therefore, one can position a ball of radius ϵ on that ray so that part of the ray with length 2ϵ is fully in the ball. Then one can put $\exp(4 \log 2)$ balls of radius ϵ/c on that ray centered at equispaced points, which will ensure that $\log M^{\text{loc}}(\epsilon) > 4 \log 2$ for any ϵ.

5.2 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.

Lemma 5.1. *For two convex sets S, S' satisfying $S' \subset S$, we have that $M^{\text{loc}}_{S'}(\epsilon) \le M^{\text{loc}}_{S}(\epsilon)$ for any $\epsilon > 0$.*

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

We first use the knowledge of σ^2 to "split" the sample. To this end let us draw $\eta \sim N(0, \mathbb{I}\sigma^2)$ independently from the observed data Y. Consider the variables $\widetilde{Y}^1 = Y + \eta$ and $\widetilde{Y}^2 = Y - \eta$. These variables are independent. Take any fixed point $\nu \in K$. We consider balls centered at ν with different radiuses $B(\nu, 1) \cap K$,$B(\nu, 2) \cap K$, ..., $B(\nu, 2^m) \cap K$, ... and every time compute the estimator from Algorithm 1 using $\widetilde{Y}^1$ as the "Y value". Denote these estimators with $\{\nu_m\}_{m=1}^\infty$. The intuition for constructing these, is that for large enough m these estimators will have good properties as μ will belong to the set $B(\nu, 2^m) \cap K$. We have the following lemma regarding the sequence of estimators ν_m.

Lemma 5.2. *All estimators ν_m lie in a compact set.*

Proof. For brevity throughout the proof we denote $\widetilde{Y}^1$ with Y. Let $P_{\overline{K}}Y$ denote the projection of Y onto the set $\overline{K}$ which is the closure of K. At some point the radius 2^N will be so big that $P_{\overline{K}}Y$ will be in the set $B(\nu, 2^N) \cap \overline{K}$. From there on, i.e. $m \ge N$, we will argue that the estimators ν_m will be close to the point $P_{\overline{K}}Y$. The first packing set is at distance $\frac{d}{2(C+1)}$ where $d \le 2^{m+1}$ and C is the constant from Algorithm 1 (such that $2(C + 1) = c$). Let $x = \|Y - P_{\overline{K}}Y\|$. For any point $\nu \in K$ we have $\sqrt{x^2 + \|\nu - P_{\overline{K}}Y\|^2} \le \|\nu - Y\| \le x + \|\nu - P_{\overline{K}}Y\|$, 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 $P_{\overline{K}}Y$ satisfies $\|\bar{\nu} - P_{\overline{K}}Y\| \le \frac{d}{2(C+1)}$, and therefore

$$\|\bar{\nu} - Y\| \le x + \|\bar{\nu} - P_{\overline{K}}Y\| \le x + \frac{d}{2(C+1)}.$$

Take $\widehat{\nu}$ to be the closest point to Y. We then have

$$\sqrt{x^2 + \|\widehat{\nu} - P_{\overline{K}}Y\|^2} \le \|\widehat{\nu} - Y\| \le \|\bar{\nu} - Y\| \le x + \frac{d}{2(C+1)}.$$

It follows that

$$\|\widehat{\nu} - P_{\overline{K}}Y\|^2 \le 2x\frac{d}{2(C+1)} + \left(\frac{d}{2(C+1)}\right)^2 \le 3\left(\frac{d}{2(C+1)}\right)^2,$$

assuming that $x \le \frac{d}{2(C+1)}$. Since $C \ge 2$ this implies that $\|\widehat{\nu} - P_{\overline{K}}Y\| \le \frac{d}{2}$, and thus the point $P_{\overline{K}}Y$ will be in the chosen ball for the second step. We can continue this logic until, $x \ge \frac{d}{2^k(C+1)}$. At this point we know that the estimator will be within distance $\frac{d}{2^{k-2}}$ of the central point, which is at distance at most $\frac{d}{2^{k-1}}$ from $P_{\overline{K}}Y$, so that the final estimator will be at distance at most $\frac{3d}{2^{k-1}} \le 6(C+1)x$ from $P_{\overline{K}}Y$. 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. $\square$

Define $\widetilde{C} = \frac{c}{4} - 1$, where c is the local packing constant from Definition 2.2. Once we have established Lemma 5.2, 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.

Algorithm 2: Upper Bound Algorithm (Unbounded Case)

Input: A sequence of estimators $\mathcal{E} := \{\nu_m\}_{m\in\mathbb{N}} \subset K$; d the diameter of $\mathcal{E}$ which is bounded by Lemma 5.2; $\nu^* \in \mathcal{E}$ an arbitrary point.

$k \leftarrow 1$;

$\Upsilon \leftarrow [\nu^*]$;

while *TRUE* **do**

    Take a $\frac{d}{2^{k+1}(\widetilde{C}+1)}$ maximal[4]packing set M_k of the set $B\big(\nu^*, \frac{d}{2^{k-1}}\big) \cap \mathcal{E}$; `/* The packing sets should be constructed in a special way as described in the proof of Theorem 5.3 to ensure measurability */`

    $\nu^* \leftarrow \operatorname{argmin}_{\nu\in M_k} \|\widetilde{Y}^2 - \nu\|$; `/* Break ties by taking the point with smallest index in` $\mathcal{E}$ `*/`

    Υ.append(ν^*);

    $k \leftarrow k + 1$;

return ν^* ; `/* Observe that by definition` Υ `forms a Cauchy sequence, so` ν^* `can be understood as the limiting point of that sequence. */`

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 5.3. *We have that $\nu^* : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^n$ is a measurable function (with respect to the Borel σ-field). As a consequence $\nu^*(Y, \eta)$ is a random variable.*

Proof. We will show that each element in the sequence Υ_j is measurable. Since they form a Cauchy sequence their limit will also be measurable by an argument similar to the one in Theorem 2.4. Throughout the proof, so as to not overburden notation, for the most part we will suppress the dependence of the estimators ν_m on $\widetilde{y}^1 = y + \eta$ and will simply write ν_m. We will also suppress the dependence of Υ_j on y and η.

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 next 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 Υ_1 is measurable. For Υ_1 the big ball on the 1-st step contains all estimators ν_m hence we start from ν_1. We will show that the event $\Upsilon_1 = \nu_j$ is a measurable event, and since as we know from before each ν_j is measurable, and the identity $(y, \eta : \Upsilon_1 \in B) = \cup_j (y, \eta : \Upsilon_1 = \nu_j) \cap (y, \eta : \nu_j(y + \eta) \in B)$ for any hyperrectangle B we will have that Υ_1 is measurable. We will now give a little details about the measurability of the event $(y, \eta : \nu_j(y + \eta) \in B)$. For $(y, \eta : \nu_j(y + \eta) \in B) = (y, \eta : y + \eta \in B')$ for some Borel set B' by the measurability of ν_j. This is a Borel set since the function $(y, \eta) \mapsto y + \eta$ is continuous and hence measurable.

[4] It is not important for the packing set to be maximal as long as it is a covering set. See Theorem 5.3 for a specification of how to construct these sets to ensure measurability.

Let us call the index set of the chosen packing (according to the strategy described above), "the index set". We then have the identity:

$$\{y,\eta : \Upsilon_1 = \nu_j\} = \cup_{S: j\in S, |S|\le M^{\mathrm{loc}}(r)} \Bigg(\{y,\eta : S \text{ is the index set}\} \cap \\ \cap_{i\in S} \{y,\eta : \|\nu_j - \widetilde{y}^2\| \le \|\nu_i - \widetilde{y}^2\|\} \cap_{i\in S, i\le j} \{y,\eta : \|\nu_i - \widetilde{y}^2\| \ne \|\nu_j - \widetilde{y}^2\|\} \Bigg),$$

where we put for brevity $r = d/(4(\widetilde{C}+1))$ and $\widetilde{y}^2 = y-\eta$. 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 measururable since for two measurable functions X and Y the events $X \le Y$ and $X \ne Y$ 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

$$\begin{aligned}\{y,\eta : S \text{ is the index set}\} = {} & \cap_{k=2}^{s_2-1}\{y,\eta : \|\nu_1 - \nu_k\| \le r\} \cap \{y,\eta : \|\nu_1 - \nu_{s_2}\| > r\} \cap \\ & \cap_{k=s_2+1}^{s_3-1} \{y,\eta : \|\nu_1 - \nu_k\| \le r\} \cup \{\omega : \|\nu_{s_2} - \nu_k\| \le r\} \\ & \cap \{y,\eta : \|\nu_1 - \nu_{s_3}\| > r\} \cap \{y,\eta : \|\nu_{s_2} - \nu_{s_3}\| > r\} \cap \\ & \ldots \\ & \cap_{k\ge s_m+1} (\{y,\eta : \|\nu_1 - \nu_k\| \le r\} \cup \{y,\eta : \|\nu_{s_2} - \nu_k\| \le r\} \cup \\ & \ldots \cup \{y,\eta : \|\nu_{s_m} - \nu_k\| \le r\}),\end{aligned}$$

which is clearly measurable (by continuity of $\|\cdot\|$, and the fact that the difference of measurable functions is measurable). This completes the proof that Υ_1 is measurable. We will now argue that Υ_2 is also measurable using the same trick. Observe that the identity:

$$\{y,\eta : \Upsilon_2 = \nu_j\} = \cup_{S: j\in S, |S|\le M^{\mathrm{loc}}(r)} \Bigg(\{y,\eta : S \text{ is the index set}\} \cap \\ \cap_{i\in S} \{y,\eta : \|\nu_j - \widetilde{y}^2\| \le \|\nu_i - \widetilde{y}^2\|\} \cap_{i\in S, i\le j} \{y,\eta : \|\nu_i - \widetilde{y}^2\| \ne \|\nu_j - \widetilde{y}^2\|\} \Bigg),$$

continues to hold for Υ_2 with the only difference that $r = d/(8(\widetilde{C}+1))$. We will now show that the event $\{y,\eta : S \text{ is the index set}\}$ continues to be measurable for Υ_2. We have

$$\begin{aligned}\{y,\eta : S \text{ is the index set}\} = {} & \cap_{k=1}^{s_1-1}\{y,\eta : \|\Upsilon_1 - \nu_k\| > d/2\} \cap \{y,\eta : \|\Upsilon_1 - \nu_{s_1}\| < d/2\} \\ & \cap_{k=s_1+1}^{s_2-1} (\{y,\eta : \|\Upsilon_1 - \nu_k\| > d/2\} \cup \{\omega : \|\nu_{s_1} - \nu_k\| \le r\}) \\ & \cap (\{y,\eta : \|\Upsilon_1 - \nu_{s_2}\| \le d/2\} \cap \{y,\eta : \|\nu_{s_1} - \nu_{s_2}\| > r\}) \cap \\ & \ldots \\ & \cap_{k\ge s_m+1} (\{y,\eta : \|\Upsilon_1 - \nu_k\| > d/2\} \cup \{y,\eta : \|\nu_1 - \nu_k\| \le r\} \\ & \cup \{y,\eta : \|\nu_{s_2} - \nu_k\| \le r\} \cup \ldots \cup \{y,\eta : \|\nu_{s_m} - \nu_k\| \le r\}),\end{aligned}$$

Clearly, all of the above are measurable events, and therefore Υ_2 is measurable. Proving that all subsequent Υ_j are measurable is the same as proving that Υ_2 is measurable which completes the proof.

$\square$

Next we prove a modification of Lemma 2.6. The setting is as follows. We are given M points $\nu_1, \dots, \nu_M \in K$ such that $\min \|\nu_i - \mu\| \le \rho$.

Lemma 5.4. *Let $i^* = \operatorname{argmin}_i \|\widetilde{Y}^2 - \nu_i\|$. We will show that the closest point to $\widetilde{Y}^2$, ν_{i^*} satisfies*

$$\mathbb{P}(\|\nu_{i^*} - \mu\| \ge (C+1)\rho) \le M \exp(-(C-2)^2\rho^2/(16\sigma^2)),$$

for any fixed $C > 2$.

Proof. Define the intermediate random variable

$$T_i = \begin{cases} \max_{j \in [M]} \|\nu_i - \nu_j\|, \text{ s.t. } \|\widetilde{Y}^2 - \nu_i\| - \|\widetilde{Y}^2 - \nu_j\| \ge 0, \|\nu_i - \nu_j\| \ge C\rho \\ 0, \text{if no such } j \text{ exists,} \end{cases}$$

Without loss of generality assume that $\|\mu - \nu_i\| \le \rho$. Next, we have that

$$\begin{aligned} \mathbb{P}(\|\nu_{i^*} - \mu\| \ge \rho + C\rho) &\le \mathbb{P}(i^* \in \{j : \|\nu_j - \nu_i\| \ge C\rho\}) \\ &\le P(T_i > 0), \end{aligned}$$

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

$$\begin{aligned} \mathbb{P}(T_i > 0) = \mathbb{P}(\exists j : \|\nu_j - \nu_i\| \ge C\rho \text{ and } \|\widetilde{Y}^2 - \nu_i\| - \|\widetilde{Y}^2 - \nu_j\| \ge 0) \\ \le M \exp(-(C-2)^2\rho^2/(16\sigma^2)), \end{aligned}$$

by Lemma 2.5 (here we used the fact that $\xi_i - \eta_i \sim N(0, 2\sigma^2)$). This is what we wanted to show. ☐

Theorem 5.5. *The estimator from Algorithm 2 returns a vector ν^* which satisfies the following property*

$$\mathbb{E}\|\mu - \nu^*\|^2 \le \bar{C}\epsilon^{*2},$$

for some universal constant $\bar{C}$, where ϵ^* is the smallest solution to

$$\frac{\epsilon^2}{\sigma^2} > 32 \log M^{\text{loc}}\left(\epsilon \frac{c}{(c/2-3)}\right) \vee 32 \log 2. \tag{5.1}$$

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

Remark 5.6. *For c large enough inequality (5.1) is equivalent to simply*

$$\frac{\epsilon^2}{\sigma^2} > 32 \log M^{\text{loc}}\left(\epsilon \frac{c}{2(c/4-3)}\right),$$

since one can always take the center of the ball lying on an infinite ray (which exists [see Lemma 1 Section 2.5 Grünbaum, 2013, e.g.]), and then there will exist at least $\exp(\log 2)$ equispaced points on that ray.

Remark 5.7. *Note that the expected value in (5.1) is taken with respect to both ξ and η. It is clear by Jensen's inequality, that the estimator $\mathbb{E}_\eta \nu^*(Y,\eta)$ satisfies*

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

Note that since $\mathbb{E}_\eta \nu^*(Y,\eta) = \mathbb{E}[\nu^*(Y,\eta)|Y]$ it is a measurable function of the data Y, and therefore achieves the minimax rate as shown in Proposition 5.8.

Proof. Let $\rho = \inf_j \|\mu - \nu_j\|$, and let $\bar{\nu}$ be a limiting point of ν_j such that $\rho = \|\mu - \bar{\nu}\|$. Note that ρ is fixed given $\widetilde{Y}^1$. 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 2.9 are fulfilled and by (2.5) therefore

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

which holds true for $x \geq \epsilon^*$, where $\epsilon^* = \epsilon_{J^*} = \frac{(C-2)\operatorname{diam}(B(\nu,2^N)\cap K)}{(C+1)2^{J^*-1}}$, and where J^* is the maximum J selected so that $\frac{\epsilon_J^2}{2\sigma^2} > 16 \log M^{\text{loc}}_{B(\nu,2^N)\cap K}\left(\epsilon_J \frac{2(C+1)}{(C-2)}\right) \vee 16\log 2$ of $J^* = 1$ if such J does not exist. Here we have $2\sigma^2$ in the denominator since $\xi_i + \eta_i \sim N(0, 2\sigma^2)$.

For any J such that $\frac{d}{2^{J+1}(\widetilde{C}+1)} \geq \rho$ by Lemma 5.4 we have the following bound (recall that $c = 4(\widetilde{C}+1)$ where c is the constant from the definition of local packing entropy):

$$\begin{aligned}
&\mathbb{P}\left(\|\bar{\nu} - \Upsilon_J\| \geq \frac{d}{2^{J-1}} \middle| \|\bar{\nu} - \Upsilon_{J-1}\| \leq \frac{d}{2^{J-2}}, \widetilde{Y}^1\right) \\
&\quad \leq \mathbb{P}\left(\|\bar{\nu} - \Upsilon_J\| \geq \rho + (\widetilde{C}+1)(\frac{d}{2^J(\widetilde{C}+1)} + \rho) \middle| \|\bar{\nu} - \Upsilon_{J-1}\| \leq \frac{d}{2^{J-2}}, \widetilde{Y}^1\right) \\
&\quad \leq \mathbb{P}\left(\|\mu - \Upsilon_J\| \geq (\widetilde{C}+1)(\frac{d}{2^J(\widetilde{C}+1)} + \rho) \middle| \|\bar{\nu} - \Upsilon_{J-1}\| \leq \frac{d}{2^{J-2}}, \widetilde{Y}^1\right) \\
&\quad \leq |M_{J-1}| \exp(-(\widetilde{C}-2)^2(d/(2^J(\widetilde{C}+1)) + \rho)^2/(16\sigma^2))).
\end{aligned}$$

Telescoping this bound by the union bound gives us that

$$\begin{aligned}
\mathbb{P}(\|\mu - \Upsilon_J\| \geq \rho + \frac{d}{2^{J-1}}|\widetilde{Y}^1) &\leq \sum_{j=2}^J |M_{j-1}| \exp(-(\widetilde{C}-2)^2(d/(2^j(\widetilde{C}+1)) + \rho)^2/(16\sigma^2))) \\
&\leq M^{\text{loc}}\left(\frac{d}{2^{J-2}}\right) \sum_{j=2}^J \exp(-(\widetilde{C}-2)^2(d/(2^j(\widetilde{C}+1)) + \rho)^2/(16\sigma^2))) \\
&\leq M^{\text{loc}}\left(\frac{d}{2^{J-2}}\right) \sum_{j=2}^J \exp(-(\widetilde{C}-2)^2(d/(2^j(\widetilde{C}+1))^2/(16\sigma^2))) \\
&\leq M^{\text{loc}}\left(\frac{d}{2^{J-2}}\right) a(1 + a^{4-1} + a^{16-1} + \ldots)\mathbb{1}(J > 1) \\
&\leq M^{\text{loc}}\left(\frac{d}{2^{J-2}}\right) \frac{a}{1-a}\mathbb{1}(J > 1)
\end{aligned}$$

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

So if one sets $\epsilon_J = \frac{(\widetilde{C}-2)d}{2^J(\widetilde{C}+1)}$, we have that if $\epsilon_J^2/(16\sigma^2) > 2\log M^{\mathrm{loc}}\Big(\epsilon_J \frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big)$ 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

$$2\log M^{\mathrm{loc}}\Big(\epsilon_J \frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big) \le 2\Big(\log 2 \vee \log M^{\mathrm{loc}}\Big(\epsilon_J \frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big)\Big),$$

this condition is implied when $\frac{\epsilon_J^2}{\sigma^2} > 32\log M^{\mathrm{loc}}\Big(\epsilon_J \frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big) \vee 32\log 2$.

Below constants can change values from line to line. By the triangle inequality we have that $\|\nu^* - \mu\| \le \|\nu^* - \Upsilon_J\| + \|\Upsilon_J - \mu\| \le \rho + 6\epsilon_J \frac{\widetilde{C}+1}{\widetilde{C}-2} \le 7\epsilon_J \frac{\widetilde{C}+1}{\widetilde{C}-2}$ with probability at least $1 - 2\exp(-\epsilon_J^2/(32\sigma^2))$. Let J^{**} be selected as the maximum J such that $\frac{\epsilon_J^2}{\sigma^2} > 32\log M^{\mathrm{loc}}\Big(\epsilon_J \frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big) \vee 32\log 2$ otherwise if such J does not exist $J^{**} = 1$. We have shown that for all $J \le J^{**}$ we have

$$\begin{aligned}\mathbb{P}(\|\mu - \nu^*\| \ge \frac{7}{2}\frac{d}{2^{J-1}}) &\le \underline{\underline{C}}\exp(-C'(d/2^{J-1})^2/\sigma^2)\mathbb{1}(J^{**} > 1)\\ &+ \mathbb{1}\Big(\frac{d}{2^{J+1}(\widetilde{C}+1)} \le 2\kappa\epsilon^*\Big) + C''\exp(-C'''(d/2^{J-1})^2/\sigma^2)\mathbb{1}(J^* > 1),\end{aligned}$$

where the last two summands, come from controlling the probability of the event $\frac{d}{2^{J+1}(\widetilde{C}+1)} < \rho$. Hence for any $x \ge \epsilon^{**}$ we have

$$\begin{aligned}\mathbb{P}(\|\mu - \nu^*\| \ge 7x) &\le \underline{\underline{C}}\exp(-C'x^2/\sigma^2)\mathbb{1}(J^{**} > 1)\\ &+ \mathbb{1}\Big(\frac{x}{4(\widetilde{C}+1)} \le 2\kappa\epsilon^*\Big) + C''\exp(-C'''x^2/\sigma^2)\mathbb{1}(J^* > 1),\end{aligned}$$

where $\epsilon^{**} = \epsilon_{J^{**}}$.

Integrating the tail bound as before we have

$$\begin{aligned}\mathbb{E}\|\mu - \nu^*\|^2 &\le C'''\epsilon^{**2} + C''''\sigma^2\exp(-C''\epsilon^{**2}/\sigma^2)\mathbb{1}(J^{**} > 1)\\ &+ C'''''\epsilon^{*2} + C''''''\sigma^2\exp(-\underline{\underline{C}}\epsilon^{*2}/\sigma^2)\mathbb{1}(J^* > 1).\end{aligned}$$

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

$$\frac{\epsilon^2}{\sigma^2} > 32\log M^{\mathrm{loc}}\Big(\epsilon\frac{2(C+1)}{(C-2)}\Big) \vee 32\log 2,$$

since $M^{\mathrm{loc}}(x) \ge M^{\mathrm{loc}}_{B(\nu,2^N)\cap K}(x)$ for any x. In addition, since $M^{\mathrm{loc}}\Big(\epsilon\frac{2(C+1)}{(C-2)}\Big) \ge M^{\mathrm{loc}}\Big(\epsilon\frac{4(\widetilde{C}+1)}{(\widetilde{C}-2)}\Big)$ (which follows since we have $\epsilon\frac{2\widetilde{C}+1}{\widetilde{C}-2} > \epsilon\frac{C+1}{C-2}$ and $c = 4(\widetilde{C}+1) = 2(C+1)$) we conclude that $2\epsilon^{***} \ge \epsilon^{**}$. This completes the proof. □

Proposition 5.8. *Define ϵ^* as $\sup\{\epsilon : \epsilon^2/\sigma^2 \le \log M^{\mathrm{loc}}(\epsilon)\}$, where c in the definition of local entropy is a sufficiently large absolute constant. Then the minimax rate is given by ϵ^{*2} up to absolute constant factors.*

Proof. For $\delta^* := \epsilon^*/4$ we have $\log M^{\mathrm{loc}}(\delta^*) \geq \log M^{\mathrm{loc}}(\epsilon^*) \geq \epsilon^{*2}/\sigma^2 = 16\delta^{*2}/\sigma^2$ 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 5.2).

On the other hand we know that for a constant $C > 1$:

$$4C\epsilon^{*2}/\sigma^2 \geq C \log M^{\mathrm{loc}}(2\epsilon^*) \geq C \log M^{\mathrm{loc}}(2\epsilon^*\sqrt{C}) \geq C \log M^{\mathrm{loc}}\left(2\epsilon^*\sqrt{C}\frac{c}{c/2-3}\right),$$

and so setting $\delta = 2\epsilon^*\sqrt{C}$ we obtain that

$$\delta^2/\sigma^2 \geq C \log M^{\mathrm{loc}}\left(\delta\frac{c}{c/2-3}\right).$$

Plugging in $C = 32$ grants the requirement of Remark 5.6, which completes the proof. □

6 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 σ^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. Finally an exciting question that remains is whether knowledge of σ^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 σ^2 seems to be required.

7 Acknowledgements

The author is grateful to Siva Balakrishnan for helpful discussions and for pointing him to the relevant papers by Li Zhang, to Ramon van Handel for helpful discussions on the entropy numbers of ellipses, and to Larry Wasserman for encouragements.

References

P. C. Bellec et al. Sharp oracle inequalities for least squares estimators in shape restricted regression. *The Annals of Statistics*, 46(2):745–780, 2018.

P. J. Bickel. Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *The Annals of Statistics*, 9(6):1301–1309, 1981.

L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65(2):181–237, 1983.

G. Casella and W. E. Strawderman. Estimating a bounded normal mean. *The Annals of Statistics*, 9(4):870–878, 1981.

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

X. Chen, A. Guntuboyina, and Y. Zhang. A note on the approximate admissibility of regularized estimators in the gaussian sequence model. *Electronic Journal of Statistics*, 11(2):4746–4768, 2017.

D. L. Donoho and I. M. Johnstone. Minimax risk overl p-balls forl p-error. *Probability Theory and Related Fields*, 99(2):277–303, 1994.

D. L. Donoho, R. C. Liu, and B. MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, pages 1416–1437, 1990.

M. Ermakov. Minimax nonparametric estimation on maxisets. *Journal of Mathematical Sciences*, 245(1), 2020.

B. Grünbaum. *Convex polytopes*, volume 221. Springer Science & Business Media, 2013.

O. Guedon and A. Litvak. Euclidean projections of a p-convex body. In *Geometric aspects of functional analysis*, pages 95–108. Springer, 2000.

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *Statistical Science*, 33(4): 568–594, 2018.

I. A. Ibragimov and R. Z. Khas' minskii. On nonparametric estimation of the value of a linear functional in gaussian white noise. *Theory of Probability & Its Applications*, 29(1):18–32, 1985.

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

I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. *Unpublished manuscript*, 2011.

L. LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53, 1973.

A. Nemirovski. Lectures on probability theory and statistics. part ii: topics in non-parametric statistics. *Probability Summer School, Saint Flour, Springer-Verlag, Berlin*, 1998.

L. Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2018.

A. Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

M. Pinsker. Optimal filtration of square-integrable signals in gaussian noise. *Prob. Info. Transmission*, 16(2):120–133, 1980.

C. Schütt. Entropy numbers of diagonal operators between symmetric banach spaces. *Journal of approximation theory*, 40(2):121–128, 1984.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

A. W. Van Der Vaart and J. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Wei, B. Fang, and M. J. Wainwright. From gauss to kolmogorov: Localized measures of complexity for ellipses. *Electronic Journal of Statistics*, 14(2):2988–3031, 2020.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

L. Zhang. Nearly optimal minimax estimator for high-dimensional sparse linear regression. *The Annals of Statistics*, 41(4):2149–2175, 2013.