CHAPTER 5

Minimax Lower Bounds

In the previous chapters, we have proved several upper bounds and the goal of this chapter is to assess their optimality. Specifically, our goal is to answer the following questions:

1. Can our analysis be improved? In other words: do the estimators that we have studied actually satisfy better bounds?

2. Can any estimator improve upon these bounds?

Both questions ask about some form of *optimality*. The first one is about optimality of an estimator, whereas the second one is about optimality of a bound.

The difficulty of these questions varies depending on whether we are looking for a positive or a negative answer. Indeed, a positive answer to these questions simply consists in finding a better proof for the estimator we have studied (question 1.) or simply finding a better estimator, together with a proof that it performs better (question 2.). A negative answer is much more arduous. For example, in question 2., it is a statement about *all estimators*. How can this be done? The answer lies in information theory (see [CT06] for a nice introduction).

In this chapter, we will see how to give a negative answer to question 2. It will imply a negative answer to question 1.

Θ	$\phi(\Theta)$	Estimator	Result
$\mathbb{R}^d$	$\frac{\sigma^2 d}{n}$	$\hat{\theta}^{\text{LS}}$	Theorem 2.2
$\mathcal{B}_1$	$\sigma\sqrt{\frac{\log d}{n}}$	$\hat{\theta}^{\text{LS}}_{\mathcal{B}_1}$	Theorem 2.4
$\mathcal{B}_0(k)$	$\frac{\sigma^2 k}{n}\log(ed/k)$	$\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)}$	Corollaries 2.8-2.9

Table 5.1. Rate $\phi(\Theta)$ obtained for different choices of Θ.

5.1 OPTIMALITY IN A MINIMAX SENSE

Consider the Gaussian Sequence Model (GSM) where we observe $\mathbf{Y} = (Y_1, \dots, Y_d)^\top$, defined by

$$Y_i = \theta^*_i + \varepsilon_i\,, \qquad i = 1, \dots, d\,, \tag{5.1}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)^\top \sim \mathcal{N}_d(0, \frac{\sigma^2}{n} I_d)$, $\theta^* = (\theta^*_1, \dots, \theta^*_d)^\top \in \Theta$ is the parameter of interest and $\Theta \subset \mathbb{R}^d$ is a given set of parameters. We will need a more precise notation for probabilities and expectations throughout this chapter. Denote by $\mathbb{P}_{\theta^*}$ and $\mathbb{E}_{\theta^*}$ the probability measure and corresponding expectation that are associated to the distribution of $\mathbf{Y}$ from the GSM (5.1).

Recall that GSM is a special case of the linear regression model when the design matrix satisfies the ORT condition. In this case, we have proved several performance guarantees (*upper bounds*) for various choices of Θ that can be expressed either in the form

$$\mathbb{E}\big[|\hat{\theta}_n - \theta^*|_2^2\big] \le C\phi(\Theta) \tag{5.2}$$

or the form

$$|\hat{\theta}_n - \theta^*|_2^2 \le C\phi(\Theta)\,, \quad \text{with prob. } 1 - d^{-2} \tag{5.3}$$

For some constant C. The rates $\phi(\Theta)$ for different choices of Θ that we have obtained are gathered in Table 5.1 together with the estimator (and the corresponding result from Chapter 2) that was employed to obtain this rate. Can any of these results be improved? In other words, does there exists another estimator $\tilde{\theta}$ such that $\sup_{\theta^* \in \Theta} \mathbb{E}|\tilde{\theta} - \theta^*|_2^2 \ll \phi(\Theta)$?

A first step in this direction is the Cramér-Rao lower bound [Sha03] that allows us to prove lower bounds in terms of the Fisher information. Nevertheless, this notion of optimality is too stringent and often leads to nonexistence of optimal estimators. Rather, we prefer here the notion of *minimax optimality* that characterizes how fast θ^* can be estimated *uniformly* over Θ.

Definition 5.1. We say that an estimator $\hat{\theta}_n$ is *minimax optimal over* Θ if it satisfies (5.2) and there exists $C' > 0$ such that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\phi^{-1}(\Theta) |\hat{\theta} - \theta|_2^2 \right] \geq C' \tag{5.4}$$

where the infimum is taker over all estimators (i.e., measurable functions of $\mathbf{Y}$). Moreover, $\phi(\Theta)$ is called *minimax rate of estimation over* Θ.

Note that minimax rates of convergence ϕ are defined up to multiplicative constants. We may then choose this constant such that the minimax rate has a simple form such as $\sigma^2 d/n$ as opposed to $7\sigma^2 d/n$ for example.

This definition can be adapted to rates that hold with high probability. As we saw in Chapter 2 (Cf. Table 5.1), the upper bounds in expectation and those with high probability are of the same order of magnitude. It is also the case for lower bounds. Indeed, observe that it follows from the Markov inequality that for any $A > 0$,

$$\mathbb{E}_\theta \left[\phi^{-1}(\Theta) |\hat{\theta} - \theta|_2^2 \right] \geq A \mathbb{P}_\theta \left[\phi^{-1}(\Theta) |\hat{\theta} - \theta|_2^2 > A \right] \tag{5.5}$$

Therefore, (5.6) follows if we prove that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left[|\hat{\theta} - \theta|_2^2 > A\phi(\Theta) \right] \geq C"$$

for some positive constants A and $C"$. The above inequality also implies a lower bound with high probability. We can therefore employ the following alternate definition for minimax optimality.

Definition 5.2. We say that an estimator $\hat{\theta}$ is *minimax optimal over* Θ if it satisfies either (5.2) or (5.3) and there exists $C' > 0$ such that

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left[|\hat{\theta} - \theta|_2^2 > \phi(\Theta) \right] \geq C' \tag{5.6}$$

where the infimum is taker over all estimators (i.e., measurable functions of $\mathbf{Y}$). Moreover, $\phi(\Theta)$ is called *minimax rate of estimation over* Θ.

5.2 REDUCTION TO FINITE HYPOTHESIS TESTING

Minimax lower bounds rely on information theory and follow from a simple principle: if the number of observations is too small, it may be hard to distinguish between two probability distributions that are close to each other. For example, given n i.i.d. observations, it is impossible to reliably decide whether they are drawn from $\mathcal{N}(0, 1)$ or $\mathcal{N}(\frac{1}{n}, 1)$. This simple argument can be made precise using the formalism of *statistical hypothesis testing*. To do so, we reduce our estimation problem to a testing problem. The reduction consists of two steps.

1. **Reduction to a finite number of *hypotheses.*** In this step the goal is to find the largest possible number of hypotheses $\theta_1, \dots, \theta_M \in \Theta$ under the constraint that
$$|\theta_j - \theta_k|_2^2 \geq 4\phi(\Theta)\,. \tag{5.7}$$
This problem boils down to a *packing* of the set Θ.

   Then we can use the following trivial observations:
$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\big[|\hat{\theta} - \theta|_2^2 > \phi(\Theta)\big] \geq \inf_{\hat{\theta}} \max_{1\leq j\leq M} \mathbb{P}_{\theta_j}\big[|\hat{\theta} - \theta_j|_2^2 > \phi(\Theta)\big]\,.$$

2. **Reduction to a testing problem.** In this second step, the necessity of the constraint (5.7) becomes apparent.

   For any estimator $\hat{\theta}$, define the minimum distance test $\psi(\hat{\theta})$ that is associated to it by
$$\psi(\hat{\theta}) = \operatorname*{argmin}_{1\leq j\leq M} |\hat{\theta} - \theta_j|_2\,,$$
with ties broken arbitrarily.

   Next observe that if, for some $j = 1, \dots, M$, $\psi(\hat{\theta}) \neq j$, then there exists $k \neq j$ such that $|\hat{\theta} - \theta_k|_2 \leq |\hat{\theta} - \theta_j|_2$. Together with the reverse triangle inequality it yields
$$|\hat{\theta} - \theta_j|_2 \geq |\theta_j - \theta_k|_2 - |\hat{\theta} - \theta_k|_2 \geq |\theta_j - \theta_k|_2 - |\hat{\theta} - \theta_j|_2$$
so that
$$|\hat{\theta} - \theta_j|_2 \geq \frac{1}{2}|\theta_j - \theta_k|_2$$
Together with constraint (5.7), it yields
$$|\hat{\theta} - \theta_j|_2^2 \geq \phi(\Theta)$$
As a result,
$$\begin{aligned}\inf_{\hat{\theta}} \max_{1\leq j\leq M} \mathbb{P}_{\theta_j}\big[|\hat{\theta} - \theta_j|_2^2 > \phi(\Theta)\big] &\geq \inf_{\hat{\theta}} \max_{1\leq j\leq M} \mathbb{P}_{\theta_j}\big[\psi(\hat{\theta}) \neq j\big] \\ &\geq \inf_{\psi} \max_{1\leq j\leq M} \mathbb{P}_{\theta_j}\big[\psi \neq j\big]\end{aligned}$$
where the infimum is taken over all tests based on $\mathbf{Y}$ and that take values in $\{1, \dots, M\}$.

Conclusion: it is sufficient for proving lower bounds to find $\theta_1, \dots, \theta_M \in \Theta$ such that $|\theta_j - \theta_k|_2^2 \geq 4\phi(\Theta)$ and
$$\inf_{\psi} \max_{1\leq j\leq M} \mathbb{P}_{\theta_j}\big[\psi \neq j\big] \geq C'\,.$$

The above quantity is called *minimax probability of error*. In the next sections, we show how it can be bounded from below using arguments from information theory. For the purpose of illustration, we begin with the simple case where $M = 2$ in the next section.

5.3 LOWER BOUNDS BASED ON TWO HYPOTHESES

The Neyman-Pearson Lemma and the total variation distance

Consider two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$ and observations X drawn from either $\mathbb{P}_0$ or $\mathbb{P}_1$. We want to know which distribution X comes from. It corresponds to the following statistical hypothesis problem:

$$\begin{aligned} H_0 &: Z \sim \mathbb{P}_0 \\ H_1 &: Z \sim \mathbb{P}_1 \end{aligned}$$

A test $\psi = \psi(Z) \in \{0, 1\}$ indicates which hypothesis should be true. Any test ψ can make two types of errors. It can commit either an error of type I ($\psi = 1$ whereas $Z \sim \mathbb{P}_0$) or an error of type II ($\psi = 0$ whereas $Z \sim \mathbb{P}_1$). Of course, the test may also be correct. The following fundamental result, called the *Neyman Pearson Lemma* indicates that any test ψ is bound to commit one of these two types of error with positive probability unless $\mathbb{P}_0$ and $\mathbb{P}_1$ have essentially disjoint support.

Let ν be a sigma finite measure satisfying $\mathbb{P}_0 \ll \nu$ and $\mathbb{P}_1 \ll \nu$. For example we can take $\nu = \mathbb{P}_0 + \mathbb{P}_1$. It follows from the Radon-Nikodym theorem [Bil95] that both $\mathbb{P}_0$ and $\mathbb{P}_1$ admit probability densities with respect to ν. We denote them by p_0 and p_1 respectively. For any function f, we write for simplicity

$$\int f = \int f(x)\nu(\mathrm{d}x)$$

Lemma 5.3 (Neyman-Pearson)**.** *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Then for any test ψ, it holds*

$$\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \geq \int \min(p_0, p_1)$$

Moreover, equality holds for the Likelihood Ratio test $\psi^\star = \mathbb{1}(p_1 \geq p_0)$.

Proof. Observe first that

$$\begin{aligned} \mathbb{P}_0(\psi^\star = 1) + \mathbb{P}_1(\psi^\star = 0) &= \int_{\psi^*=1} p_0 + \int_{\psi^*=0} p_1 \\ &= \int_{p_1 \geq p_0} p_0 + \int_{p_1 < p_0} p_1 \\ &= \int_{p_1 \geq p_0} \min(p_0, p_1) + \int_{p_1 < p_0} \min(p_0, p_1) \\ &= \int \min(p_0, p_1)\,. \end{aligned}$$

Next for any test ψ, define its rejection region $R = \{\psi = 1\}$. Let $R^\star = \{p_1 \geq p_0\}$ denote the rejection region of the likelihood ration test $\psi^\star$. It holds

$$\begin{aligned}
\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) &= 1 + \mathbb{P}_0(R) - \mathbb{P}_1(R) \\
&= 1 + \int_R p_0 - p_1 \\
&= 1 + \int_{R\cap R^\star} p_0 - p_1 + \int_{R\cap (R^\star)^c} p_0 - p_1 \\
&= 1 - \int_{R\cap R^\star} |p_0 - p_1| + \int_{R\cap (R^\star)^c} |p_0 - p_1| \\
&= 1 + \int |p_0 - p_1| \big[\mathbb{1}(R \cap (R^\star)^c) - \mathbb{1}(R \cap R^\star)\big]
\end{aligned}$$

The above quantity is clearly minimized for $R = R^\star$. □

The lower bound in the Neyman-Pearson lemma is related to a well known quantity: the total variation distance.

Definition-Proposition 5.4. *The* total variation distance *between two probability measures $\mathbb{P}_0$ and $\mathbb{P}_1$ on a measurable space $(\mathcal{X}, \mathcal{A})$ is defined by*

$$\begin{aligned}
\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \sup_{R\in\mathcal{A}} |\mathbb{P}_0(R) - \mathbb{P}_1(R)| && (i) \\
&= \sup_{R\in\mathcal{A}} \Big| \int_R p_0 - p_1 \Big| && (ii) \\
&= \frac{1}{2} \int |p_0 - p_1| && (iii) \\
&= 1 - \int \min(p_0, p_1) && (iv) \\
&= 1 - \inf_\psi \big[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)\big] && (v)
\end{aligned}$$

where the infimum above is taken over all tests.

Proof. Clearly $(i) = (ii)$ and the Neyman-Pearson Lemma gives $(iv) = (v)$. Moreover, by identifying a test ψ to its rejection region, it is not hard to see that $(i) = (v)$. Therefore it remains only to show that (iii) is equal to any of the other expressions. Hereafter, we show that $(iii) = (iv)$. To that end, observe that

$$\begin{aligned}
\int |p_0 - p_1| &= \int_{p_1\geq p_0} p_1 - p_0 + \int_{p_1<p_0} p_0 - p_1 \\
&= \int_{p_1\geq p_0} p_1 + \int_{p_1<p_0} p_0 - \int \min(p_0, p_1) \\
&= 1 - \int_{p_1<p_0} p_1 + 1 - \int_{p_1\geq p_0} p_0 - \int \min(p_0, p_1) \\
&= 2 - 2\int \min(p_0, p_1)
\end{aligned}$$

□

In view of the Neyman-Pearson lemma, it is clear that if we want to prove large lower bounds, we need to find probability distributions that are close in total variation. Yet, this conflicts with constraint (5.7) and a tradeoff needs to be achieved. To that end, in the Gaussian sequence model, we need to be able to compute the total variation distance between $\mathcal{N}(\theta_0, \frac{\sigma^2}{n} I_d)$ and $\mathcal{N}(\theta_1, \frac{\sigma^2}{n} I_d)$. None of the expression in Definition-Proposition 5.4 gives an easy way to do so. The Kullback-Leibler divergence is much more convenient.

The Kullback-Leibler divergence

Definition 5.5. The Kullback-Leibler divergence between probability measures $\mathbb{P}_1$ and $\mathbb{P}_0$ is given by

$$\mathsf{KL}(\mathbb{P}_1, \mathbb{P}_0) = \begin{cases} \displaystyle\int \log\Big(\frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_0}\Big)\mathrm{d}\mathbb{P}_1\,, & \text{if } \mathbb{P}_1 \ll \mathbb{P}_0 \\ \infty\,, & \text{otherwise}\,. \end{cases}$$

It can be shown [Tsy09] that the integral is always well defined when $\mathbb{P}_1 \ll \mathbb{P}_0$ (though it can be equal to ∞ even in this case). Unlike the total variation distance, the Kullback-Leibler divergence is not a distance. Actually, it is not even symmetric. Nevertheless, it enjoys properties that are very useful for our purposes.

Proposition 5.6. Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures. Then

1. $\mathsf{KL}(\mathbb{P}, \mathbb{Q}) \geq 0$
2. If $\mathbb{P}$ and $\mathbb{Q}$ are product measures, i.e.,

$$\mathbb{P} = \bigotimes_{i=1}^{n} \mathbb{P}_i \quad \text{and} \quad \mathbb{Q} = \bigotimes_{i=1}^{n} \mathbb{Q}_i$$

then

$$\mathsf{KL}(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{n} \mathsf{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,.$$

Proof. If $\mathbb{P}$ is not absolutely continuous then the result is trivial. Next, assume that $\mathbb{P} \ll \mathbb{Q}$ and let $X \sim \mathbb{P}$.

1. Observe that by Jensen's inequality,

$$\mathsf{KL}(\mathbb{P}, \mathbb{Q}) = -\mathbb{E}\log\Big(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(X)\Big) \geq -\log\mathbb{E}\Big(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}(X)\Big) = -\log(1) = 0\,.$$

2. Note that if $X = (X_1, \dots, X_n)$,

$$\begin{aligned}
\mathsf{KL}(\mathbb{P}, \mathbb{Q}) &= \mathbb{E} \log \Big(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X) \Big) \\
&= \sum_{i=1}^{n} \int \log \Big(\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mathbb{Q}_i}(X_i) \Big) \mathrm{d}\mathbb{P}_1(X_1) \cdots \mathrm{d}\mathbb{P}_n(X_n) \\
&= \sum_{i=1}^{n} \int \log \Big(\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mathbb{Q}_i}(X_i) \Big) \mathrm{d}\mathbb{P}_i(X_i) \\
&= \sum_{i=1}^{n} \mathsf{KL}(\mathbb{P}_i, \mathbb{Q}_i)
\end{aligned}$$

□

Point 2. in Proposition 5.6 is particularly useful in statistics where observations typically consist of n independent random variables.

Example 5.7. For any $\theta \in \mathbb{R}^d$, let P_θ denote the distribution of $\mathbf{Y} \sim \mathcal{N}(\theta, \sigma^2 I_d)$. Then

$$\mathsf{KL}(P_\theta, P_{\theta'}) = \sum_{i=1}^{d} \frac{(\theta_i - \theta_i')^2}{2\sigma^2} = \frac{|\theta - \theta'|_2^2}{2\sigma^2} \,.$$

The proof is left as an exercise (see Problem 5.1).

The Kullback-Leibler divergence is easier to manipulate than the total variation distance but only the latter is related to the minimax probability of error. Fortunately, these two quantities can be compared using Pinsker's inequality. We prove here a slightly weaker version of Pinsker's inequality that will be sufficient for our purpose. For a stronger statement, see [Tsy09], Lemma 2.5.

Lemma 5.8 (Pinsker's inequality.)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures such that $\mathbb{P} \ll \mathbb{Q}$. Then*

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\mathsf{KL}(\mathbb{P}, \mathbb{Q})} \,.$$

Proof. Note that

$$\begin{aligned}
\mathsf{KL}(\mathbb{P},\mathbb{Q}) &= \int_{pq>0} p\log\Big(\frac{p}{q}\Big) \\
&= -2\int_{pq>0} p\log\Big(\sqrt{\frac{q}{p}}\Big) \\
&= -2\int_{pq>0} p\log\Big(\Big[\sqrt{\frac{q}{p}}-1\Big]+1\Big) \\
&\geq -2\int_{pq>0} p\Big[\sqrt{\frac{q}{p}}-1\Big] \qquad \text{(by Jensen)} \\
&= 2-2\int\sqrt{pq}
\end{aligned}$$

Next, note that

$$\begin{aligned}
\Big(\int\sqrt{pq}\Big)^2 &= \Big(\int\sqrt{\max(p,q)\min(p,q)}\Big)^2 \\
&\leq \int\max(p,q)\int\min(p,q) \qquad \text{(by Cauchy-Schwarz)} \\
&= \big[2-\int\min(p,q)\big]\int\min(p,q) \\
&= \big(1+\mathsf{TV}(\mathbb{P},\mathbb{Q})\big)\big(1-\mathsf{TV}(\mathbb{P},\mathbb{Q})\big) \\
&= 1-\mathsf{TV}(\mathbb{P},\mathbb{Q})^2
\end{aligned}$$

The two displays yield

$$\mathsf{KL}(\mathbb{P},\mathbb{Q}) \geq 2-2\sqrt{1-\mathsf{TV}(\mathbb{P},\mathbb{Q})^2} \geq \mathsf{TV}(\mathbb{P},\mathbb{Q})^2\,,$$

where we used the fact that $0 \leq \mathsf{TV}(\mathbb{P},\mathbb{Q}) \leq 1$ and $\sqrt{1-x} \leq 1-x/2$ for $x \in [0,1]$. □

Pinsker's inequality yields the following theorem for the GSM.

Theorem 5.9. *Assume that Θ contains two hypotheses θ_0 and θ_1 such that $|\theta_0-\theta_1|_2^2 = 8\alpha^2\sigma^2/n$ for some $\alpha \in (0,1/2)$. Then*

$$\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{P}_\theta(|\hat\theta-\theta|_2^2 \geq \frac{2\alpha\sigma^2}{n}) \geq \frac{1}{2}-\alpha\,.$$

Proof. Write for simplicity $\mathbb{P}_j = \mathbb{P}_{\theta_j}$, $j=0,1$. Recall that it follows from the

reduction to hypothesis testing that

$$\begin{aligned}
\inf_{\hat\theta}\sup_{\theta\in\Theta}\mathbb{P}_\theta(|\hat\theta-\theta|_2^2\geq\frac{2\alpha\sigma^2}{n}) &\geq \inf_{\psi}\max_{j=0,1}\mathbb{P}_j(\psi\neq j)\\
&\geq \frac{1}{2}\inf_{\psi}\Big(\mathbb{P}_0(\psi=1)+\mathbb{P}_1(\psi=0)\Big)\\
&= \frac{1}{2}\Big[1-\mathsf{TV}(\mathbb{P}_0,\mathbb{P}_1)\Big] && \text{(Prop.-def. 5.4)}\\
&\geq \frac{1}{2}\Big[1-\sqrt{\mathsf{KL}(\mathbb{P}_1,\mathbb{P}_0)}\Big] && \text{(Lemma 5.8)}\\
&= \frac{1}{2}\Big[1-\sqrt{\frac{n|\theta_1-\theta_0|_2^2}{2\sigma^2}}\Big] && \text{(Example 5.7)}\\
&= \frac{1}{2}\Big[1-2\alpha\Big]
\end{aligned}$$

□

Clearly the result of Theorem 5.9 matches the upper bound for $\Theta=\mathbb{R}^d$ only for $d=1$. How about larger d? A quick inspection of our proof shows that our technique, in its present state, cannot yield better results. Indeed, there are only two known candidates for the choice of θ^*. With this knowledge, one can obtain upper bounds that do not depend on d by simply projecting Y onto the linear span of θ_0,θ_1 and then solving the GSM in two dimensions. To obtain larger lower bounds, we need to use more than two hypotheses. In particular, in view of the above discussion, we need a set of hypotheses that spans a linear space of dimension proportional to d. In principle, we should need at least order d hypotheses but we will actually need much more.

5.4 LOWER BOUNDS BASED ON MANY HYPOTHESES

The reduction to hypothesis testing from Section 5.2 allows us to use more than two hypotheses. Specifically, we should find $\theta_1,\dots,\theta_M$ such that

$$\inf_{\psi}\max_{1\leq j\leq M}\mathbb{P}_{\theta_j}\big[\psi\neq j\big]\geq C'\,,$$

for some positive constant C'. Unfortunately, the Neyman-Pearson Lemma no longer exists for more than two hypotheses. Nevertheless, it is possible to relate the minimax probability of error directly to the Kullback-Leibler divergence, without involving the total variation distance. This is possible using a well known result from information theory called *Fano's inequality*. We use it in a form that is tailored to our purposes and that is due to Lucien Birgé [Bir83] and builds upon an original result in [IH81].

Theorem 5.10 (Fano's inequality)**.** *Let $P_1,\dots,P_M, M\geq 2$ be probability distributions such that $P_j\ll P_k$, $\forall\, j,k$. Then*

$$\inf_{\psi}\max_{1\leq j\leq M}P_j\big[\psi(X)\neq j\big]\geq 1-\frac{\frac{1}{M^2}\sum_{j,k=1}^M\mathsf{KL}(P_j,P_k)+\log 2}{\log(M-1)}$$

where the infimum is taken over all tests with values in $\{1, \dots, M\}$.

Proof. Let $Z \in \{1, \dots, M\}$ be a random variable such that $\mathbb{P}(Z = i) = 1/M$ and let $X \sim P_Z$. Note that P_Z is a *mixture distribution* so that for any measure ν such that $P_Z \ll \nu$, we have

$$\frac{\mathrm{d}P_Z}{\mathrm{d}\nu} = \frac{1}{M} \sum_{j=1}^{M} \frac{\mathrm{d}P_j}{\mathrm{d}\nu} \,.$$

For all test ψ, we have

$$\begin{aligned}
&\sum_{j=1}^{M} \mathbb{P}(Z = j|X) \log[\mathbb{P}(Z = j|X)] = \\
&= \mathbb{P}(Z = \psi(X)|X) \log[\mathbb{P}(Z = \psi(X)|X)] + \sum_{j \neq \psi(X)} \mathbb{P}(Z = j|X) \log[\mathbb{P}(Z = j|X)] \\
&= (1 - \mathbb{P}(Z \neq \psi(X)|X)) \log[1 - \mathbb{P}(Z \neq \psi(X)|X)] \\
&\quad + \mathbb{P}(Z \neq \psi(X)|X) \sum_{j \neq \psi(X)} \frac{\mathbb{P}(Z = j|X)}{\mathbb{P}(Z \neq \psi(X)|X)} \log\Big[\frac{\mathbb{P}(Z = j|X)}{\mathbb{P}(Z \neq \psi(X)|X)}\Big] \\
&\quad + \mathbb{P}(Z \neq \psi(X)|X) \log[\mathbb{P}(Z \neq \psi(X)|X)] \\
&= h(\mathbb{P}(Z \neq \psi(X)|X)) + \mathbb{P}(Z \neq \psi(X)|X) \sum_{j \neq \psi(X)} q_j \log(q_j) \,,
\end{aligned}$$

where

$$h(x) = x \log(x) + (1 - x) \log(1 - x)$$

and

$$q_j = \frac{\mathbb{P}(Z = j|X)}{\mathbb{P}(Z \neq \psi(X)|X)}$$

is such that $q_j \geq 0$ and $\sum_{j \neq \psi(X)} q_j = 1$. It implies by Jensen's inequality that

$$\sum_{j \neq \psi(X)} q_j \log(q_j) = - \sum_{j \neq \psi(X)} q_j \log\Big(\frac{1}{q_j}\Big) \geq - \log\Big(\sum_{j \neq \psi(X)} \frac{q_j}{q_j}\Big) = - \log(M - 1) \,.$$

By the same convexity argument, we get $h(x) \geq - \log 2$. It yields

$$\sum_{j=1}^{M} \mathbb{P}(Z = j|X) \log[\mathbb{P}(Z = j|X)] \geq - \log 2 - \mathbb{P}(Z \neq \psi(X)|X) \log(M - 1) \,. \tag{5.8}$$

Next, observe that since $X \sim P_Z$, the random variable $\mathbb{P}(Z = j|X)$ satisfies

$$\mathbb{P}(Z = j|X) = \frac{1}{M} \frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(X) = \frac{\mathrm{d}P_j(X)}{\sum_{k=1}^{M} \mathrm{d}P_k(X)}$$

It implies

$$\begin{aligned}
&\int \Big\{ \sum_{j=1}^{M} \mathbb{P}(Z=j|X=x) \log[\mathbb{P}(Z=j|X=x)] \Big\} \mathrm{d}P_Z(x) \\
&= \sum_{j=1}^{M} \int \Big\{ \frac{1}{M} \frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(x) \log\Big(\frac{1}{M} \frac{\mathrm{d}P_j}{\mathrm{d}P_Z}(x) \Big) \Big\} \mathrm{d}P_Z(x) \\
&= \frac{1}{M} \sum_{j=1}^{M} \int \log\Big(\frac{\mathrm{d}P_j(x)}{\sum_{k=1}^{M} \mathrm{d}P_k(x)} \Big) \mathrm{d}P_j(x) \\
&\le \frac{1}{M^2} \sum_{j,k=1}^{M} \int \log\Big(\frac{\mathrm{d}P_j(x)}{\mathrm{d}P_k(x)} \Big) \mathrm{d}P_j(x) - \log M \quad \text{(by Jensen)} \\
&= \frac{1}{M^2} \sum_{j,k=1}^{M} \mathsf{KL}(P_j, P_k) - \log M \,,
\end{aligned}$$

Together with (5.8), it yields

$$\frac{1}{M^2} \sum_{j,k=1}^{M} \mathsf{KL}(P_j, P_k) - \log M \ge -\log 2 - \mathbb{P}(Z \neq \psi(X)) \log(M-1)$$

Since

$$\mathbb{P}(Z \neq \psi(X)) = \frac{1}{M} \sum_{j=1}^{M} P_j(\psi(X) \neq j) \le \max_{1 \le j \le M} P_j(\psi(X) \neq j) \,,$$

this implies the desired result. □

Fano's inequality leads to the following useful theorem.

Theorem 5.11. *Assume that Θ contains $M \ge 5$ hypotheses $\theta_1, \dots, \theta_M$ such that for some constant $0 < \alpha < 1/4$, it holds*

(i) $|\theta_j - \theta_k|_2^2 \ge 4\phi$

(ii) $|\theta_j - \theta_k|_2^2 \le \dfrac{2\alpha\sigma^2}{n} \log(M)$

Then

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{P}_\theta\big(|\hat{\theta} - \theta|_2^2 \ge \phi \big) \ge \frac{1}{2} - 2\alpha \,.$$

Proof. in view of (i), it follows from the reduction to hypothesis testing that it is sufficient to prove that

$$\inf_{\psi} \max_{1 \le j \le M} \mathbb{P}_{\theta_j}\big[\psi \neq j \big] \ge \frac{1}{2} - 2\alpha$$

If follows from (ii) and Example 5.7 that

$$\mathsf{KL}(\mathbb{P}_j, \mathbb{P}_k) = \frac{n|\theta_j - \theta_k|_2^2}{2\sigma^2} \le \alpha \log(M)\,.$$

Moreover, since $M \ge 5$,

$$\frac{\frac{1}{M^2}\sum_{j,k=1}^{M} \mathsf{KL}(\mathbb{P}_j, \mathbb{P}_k) + \log 2}{\log(M-1)} \le \frac{\alpha \log(M) + \log 2}{\log(M-1)} \le 2\alpha + \frac{1}{2}\,.$$

The proof then follows from Fano's inequality. □

Theorem 5.11 indicates that we must take $\phi \le \frac{\alpha\sigma^2}{2n}\log(M)$. Therefore, the larger the M, the larger the lower bound can be. However, M cannot be arbitrary larger because of the constraint (i). We are therefore facing a *packing* problem where the goal is to "pack" as many Euclidean balls of radius proportional to $\sigma\sqrt{\log(M)/n}$ in Θ under the constraint that their centers remain close together (constraint (ii)). If $\Theta = \mathbb{R}^d$, this the goal is to pack the Euclidean ball of radius $R = \sigma\sqrt{2\alpha\log(M)/n}$ with Euclidean balls of radius $R\sqrt{2\alpha/\gamma}$. This can be done using a volume argument (see Problem 5.3). However, we will use the more versatile lemma below. It gives a a lower bound on the size of a packing of the discrete hypercube $\{0,1\}^d$ with respect to the *Hamming distance* defined by

$$\rho(\omega, \omega') = \sum_{i=1}^{d} \mathbb{I}(\omega_i \ne \omega'_j)\,, \qquad \forall\, \omega, \omega' \in \{0,1\}^d$$

Lemma 5.12 (Varshamov-Gilbert)**.** *For any $\gamma \in (0, 1/2)$, there exist binary vectors $\omega_1, \dots \omega_M \in \{0,1\}^d$ such that*

(i) $\rho(\omega_j, \omega_k) \ge \big(\frac{1}{2} - \gamma\big)d$ *for all* $j \ne k$,

(ii) $M = \lfloor e^{\gamma^2 d} \rfloor \ge e^{\frac{\gamma^2 d}{2}}$.

Proof. Let $\omega_{j,i}$, $1 \le i \le d, 1 \le j \le M$ to be i.i.d Bernoulli random variables with parameter $1/2$ and observe that

$$d - \rho(\omega_j, \omega_k) = X \sim \mathsf{Bin}(d, 1/2)\,.$$

Therefore it follows from a union bound that

$$\mathbb{P}\big[\exists j \ne k\,, \rho(\omega_j, \omega_k) < \big(\frac{1}{2} - \gamma\big)d\big] \le \frac{M(M-1)}{2}\mathbb{P}\big(X - \frac{d}{2} > \gamma d\big)\,.$$

Hoeffding's inequality then yields

$$\frac{M(M-1)}{2}\mathbb{P}\big(X - \frac{d}{2} > \gamma d\big) \le \exp\Big(-2\gamma^2 d + \log\big(\frac{M(M-1)}{2}\big)\Big) < 1$$

as soon as

$$M(M-1) < 2\exp\left(2\gamma^2 d\right)$$

A sufficient condition for the above inequality to hold is to take $M = \lfloor e^{\gamma^2 d} \rfloor \geq e^{\frac{\gamma^2 d}{2}}$. For this value of M, we have

$$\mathbb{P}\big(\forall j \neq k\,, \rho(\omega_j, \omega_k) \geq \big(\frac{1}{2} - \gamma\big)d\big) > 0$$

and by virtue of the probabilistic method, there exist $\omega_1, \dots \omega_M \in \{0,1\}^d$ that satisfy (i) and (ii) □

5.5 APPLICATION TO THE GAUSSIAN SEQUENCE MODEL

We are now in a position to apply Theorem 5.11 by choosing $\theta_1, \dots, \theta_M$ based on $\omega_1, \dots, \omega_M$ from the Varshamov-Gilbert Lemma.

Lower bounds for estimation

Take $\gamma = 1/4$ and apply the Varshamov-Gilbert Lemma to obtain $\omega_1, \dots, \omega_M$ with $M = \lfloor e^{d/16} \rfloor \geq e^{d/32}$ and such that $\rho(\omega_j, \omega_k) \geq d/4$ for all $j \neq k$. Let $\theta_1, \dots, \theta_M$ be such that

$$\theta_j = \omega_j \frac{\beta\sigma}{\sqrt{n}}\,,$$

for some $\beta > 0$ to be chosen later. We can check the conditions of Theorem 5.11:

(i) $|\theta_j - \theta_k|_2^2 = \dfrac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k) \geq 4\dfrac{\beta^2\sigma^2 d}{16n}$

(ii) $|\theta_j - \theta_k|_2^2 = \dfrac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k) \leq \dfrac{\beta^2\sigma^2 d}{n} \leq \dfrac{32\beta^2\sigma^2}{n}\log(M) = \dfrac{2\alpha\sigma^2}{n}\log(M)\,,$

for $\beta = \frac{\sqrt{\alpha}}{4}$. Applying now Theorem 5.11 yields

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d} \mathbb{P}_\theta\big(|\hat\theta - \theta|_2^2 \geq \frac{\alpha}{256}\frac{\sigma^2 d}{n}\big) \geq \frac{1}{2} - 2\alpha\,.$$

It implies the following corollary.

Corollary 5.13. *The minimax rate of estimation of over $\mathbb{R}^d$ in the Gaussian sequence model is $\phi(\mathbb{R}^d) = \sigma^2 d/n$. Moreover, it is attained by the least squares estimator $\hat\theta^{\text{LS}} = \mathbf{Y}$.*

Note that this rate is minimax over sets Θ that are strictly smaller than $\mathbb{R}^d$ (see Problem 5.4). Indeed, it is minimax over any subset of $\mathbb{R}^d$ that contains $\theta_1, \dots, \theta_M$.

Lower bounds for sparse estimation

It appears from Table 5.1 that when estimating sparse vectors, we have to pay for an extra logarithmic term $\log(ed/k)$ for not knowing the sparsity pattern of the unknown θ^*. In this section, we show that this term is unavoidable as it appears in the minimax optimal rate of estimation of sparse vectors.

Note that the vectors $\theta_1, \dots, \theta_M$ employed in the previous subsection are not guaranteed to be sparse because the vectors $\omega_1, \dots, \omega_M$ obtained from the Varshamov-Gilbert Lemma may themselves not be sparse. To overcome this limitation, we need a sparse version of the Varhsamov-Gilbert lemma.

Lemma 5.14 (Sparse Varshamov-Gilbert)**.** *There exist positive constants C_1 and C_2 such that the following holds for any two integers k and d such that $1 \le k \le d/8$. There exist binary vectors $\omega_1, \dots \omega_M \in \{0,1\}^d$ such that*

(i) $\rho(\omega_i, \omega_j) \ge \dfrac{k}{2}$ *for all* $i \ne j$,

(ii) $\log(M) \ge \dfrac{k}{8} \log(1 + \dfrac{d}{2k})$.

(iii) $|\omega_j|_0 = k$ *for all* j .

Proof. Take $\omega_1, \dots, \omega_M$ independently and uniformly at random from the set

$$C_0(k) = \{\omega \in \{0,1\}^d \,:\, |\omega|_0 = k\}\,,$$

of k-sparse binary random vectors. Note that $C_0(k)$ has cardinality $\binom{d}{k}$. To choose ω_j uniformly from $C_0(k)$, we proceed as follows. Let $U_1, \dots, U_k \in \{1, \dots, d\}$ be k random variables such that U_1 is drawn uniformly at random from $\{1, \dots, d\}$ and for any $i = 2, \dots, k$, conditionally on $U_1, \dots, U_{i-1}$, the random variable U_i is drawn uniformly at random from $\{1, \dots, d\} \setminus \{U_1, \dots, U_{i-1}\}$. Then define

$$\omega = \begin{cases} 1 & \text{if } i \in \{U_1, \dots, U_k\} \\ 0 & \text{otherwise}\,. \end{cases}$$

Clearly, all outcomes in $C_0(k)$ are equally likely under this distribution and therefore, ω is uniformly distributed on $C_0(k)$. Observe that

$$\begin{aligned} \mathbb{P}\big(\exists\, \omega_j \ne \omega_k \,:\, \rho(\omega_j, \omega_k) < k\big) &= \frac{1}{\binom{d}{k}} \sum_{\substack{x \in \{0,1\}^d \\ |x|_0 = k}} \mathbb{P}\big(\exists\, \omega_j \ne x \,:\, \rho(\omega_j, x) < \frac{k}{2}\big) \\ &\le \frac{1}{\binom{d}{k}} \sum_{\substack{x \in \{0,1\}^d \\ |x|_0 = k}} \sum_{j=1}^{M} \mathbb{P}\big(\omega_j \ne x \,:\, \rho(\omega_j, x) < \frac{k}{2}\big) \\ &= M \mathbb{P}\big(\omega \ne x_0 \,:\, \rho(\omega, x_0) < \frac{k}{2}\big) \end{aligned}$$

where ω has the same distribution as ω_1 and x_0 is any k-sparse vector in $\{0,1\}^d$. The last equality holds by symmetry since (i) all the ω_js have the same distribution and (ii) all the outcomes of ω_j are equally likely.

Note that

$$\rho(\omega, x_0) \geq k - \sum_{i=1}^{k} Z_i \,,$$

where $Z_i = \mathbb{1}(U_i \in \mathrm{supp}(x_0))$. Indeed the left hand side is the number of coordinates on which the vectors ω, x_0 disagree and the right hand side is the number of coordinates in $\mathrm{supp}(x_0)$ on which the two vectors disagree. In particular, we have that, $Z_1 \sim \mathsf{Ber}(k/d)$ and for any $i = 2, \ldots, d$, conditionally on $Z_1, \ldots, Z_{i-i}$, we have $Z_i \sim \mathsf{Ber}(Q_i)$, where

$$Q_i = \frac{k - \sum_{l=1}^{i-1} Z_l}{p - (i-1)} \leq \frac{k}{d-k} \leq \frac{2k}{d}$$

since $k \leq d/2$.

Next we apply a Chernoff bound to get that for any $s > 0$,

$$\mathbb{P}\big(\omega \neq x_0 \,:\, \rho(\omega, x_0) < \frac{k}{2}\big) \leq \mathbb{P}\big(\sum_{i=1}^{k} Z_i > \frac{k}{2}\big) = \mathbb{E}\Big[\exp\big(s \sum_{i=1}^{k} Z_i\big)\Big] e^{-\frac{sk}{2}}$$

The above MGF can be controlled by induction on k as follows:

$$\begin{aligned}
\mathbb{E}\Big[\exp\big(s \sum_{i=1}^{k} Z_i\big)\Big] &= \mathbb{E}\Big[\exp\big(s \sum_{i=1}^{k-1} Z_i\big) \mathbb{E}\exp\big(sZ_k \big| Z_1, \ldots, Z_{k=1}\big)\Big] \\
&= \mathbb{E}\Big[\exp\big(s \sum_{i=1}^{k-1} Z_i\big)(Q_k(e^s - 1) + 1)\Big] \\
&\leq \mathbb{E}\Big[\exp\big(s \sum_{i=1}^{k-1} Z_i\big)\Big]\big(\frac{2k}{d}(e^s - 1) + 1\big) \\
&\;\;\vdots \\
&\leq \big(\frac{2k}{d}(e^s - 1) + 1\big)^k \\
&= 2^k
\end{aligned}$$

For $s = \log(1 + \frac{d}{2k})$. Putting everything together, we get

$$\begin{aligned}
\mathbb{P}\big(\exists\, \omega_j \neq \omega_k \,:\, \rho(\omega_j, \omega_k) < k\big) &\leq \exp\Big(\log M + k\log 2 - \frac{sk}{2}\Big) \\
&= \exp\Big(\log M + k\log 2 - \frac{k}{2}\log(1 + \frac{d}{2k})\Big) \\
&\leq \exp\Big(\log M + k\log 2 - \frac{k}{2}\log(1 + \frac{d}{2k})\Big) \\
&\leq \exp\Big(\log M - \frac{k}{4}\log(1 + \frac{d}{2k})\Big) \qquad (\text{for } d \geq 8k) \\
&< 1\,.
\end{aligned}$$

If we take M such that

$$\log M < \frac{k}{4}\log(1 + \frac{d}{2k})$$

□

Apply the sparse Varshamov-Gilbert lemma to obtain $\omega_1, \dots, \omega_M$ with $\log(M) \geq \frac{k}{8}\log(1 + \frac{d}{2k})$ and such that $\rho(\omega_j, \omega_k) \geq k/2$ for all $j \neq k$. Let $\theta_1, \dots, \theta_M$ be such that

$$\theta_j = \omega_j \frac{\beta\sigma}{\sqrt{n}}\sqrt{\log(1 + \frac{d}{2k})}\,,$$

for some $\beta > 0$ to be chosen later. We can check the conditions of Theorem 5.11:

$$(i)\;\; |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k)\log(1 + \frac{d}{2k}) \geq 4\frac{\beta^2\sigma^2}{8n}k\log(1 + \frac{d}{2k})$$

$$(ii)\;\; |\theta_j - \theta_k|_2^2 = \frac{\beta^2\sigma^2}{n}\rho(\omega_j, \omega_k)\log(1 + \frac{d}{2k}) \leq \frac{2k\beta^2\sigma^2}{n}\log(1 + \frac{d}{2k}) \leq \frac{2\alpha\sigma^2}{n}\log(M)\,,$$

for $\beta = \sqrt{\frac{\alpha}{8}}$. Applying now Theorem 5.11 yields

$$\inf_{\hat\theta}\sup_{\substack{\theta\in\mathbb{R}^d \\ |\theta|_0 \leq k}} \mathbb{P}_\theta\big(|\hat\theta - \theta|_2^2 \geq \frac{\alpha^2\sigma^2}{64n}k\log(1 + \frac{d}{2k})\big) \geq \frac{1}{2} - 2\alpha\,.$$

It implies the following corollary.

Corollary 5.15. *Recall that $\mathcal{B}_0(k) \subset \mathbb{R}^d$ denotes the set of all k-sparse vectors of $\mathbb{R}^d$. The minimax rate of estimation over $\mathcal{B}_0(k)$ in the Gaussian sequence model is $\phi(\mathcal{B}_0(k)) = \frac{\sigma^2 k}{n}\log(ed/k)$. Moreover, it is attained by the constrained least squares estimator $\hat\theta^{\mathrm{LS}}_{\mathcal{B}_0(k)}$.*

Note that the modified BIC estimator of Problem 2.6 is also minimax optimal over $\mathcal{B}_0(k)$ but unlike $\hat\theta^{\mathrm{LS}}_{\mathcal{B}_0(k)}$, it is also adaptive to k. For any $\varepsilon > 0$, the Lasso estimator and the BIC estimator are minimax optimal for sets of parameters such that $k \leq d^{1-\varepsilon}$.

Lower bounds for estimating vectors in ℓ_1 balls

Recall that in Maurey's argument, we approximated a vector θ such that $|\theta|_1 = R$ by a vector θ' such that $|\theta'|_0 = \frac{R}{\sigma}\sqrt{\frac{n}{\log d}}$. We can essentially do the same for the lower bound.

Assume that $d \geq \sqrt{n}$ and let $\beta \in (0,1)$ be a parameter to be chosen later and define k to be the smallest integer such that

$$k \geq \frac{R}{\beta\sigma}\sqrt{\frac{n}{\log(ed/\sqrt{n})}}\,.$$

Let $\omega_1, \ldots, \omega_M$ be obtained from the sparse Varshamov-Gilbert Lemma 5.14 with this choice of k and define

$$\theta_j = \omega_j \frac{R}{k}\,.$$

Observe that $|\theta_j|_1 = R$ for $j = 1, \ldots, M$. We can check the conditions of Theorem 5.11:

$$(i)\ \ |\theta_j - \theta_k|_2^2 = \frac{R^2}{k^2}\rho(\omega_j, \omega_k) \geq \frac{R^2}{2k} \geq 4R\min\big(\frac{R}{8}, \beta^2\sigma\frac{\log(ed/\sqrt{n})}{8n}\big)\,.$$

$$(ii)\ \ |\theta_j - \theta_k|_2^2 \leq \frac{2R^2}{k} \leq 4R\beta\sigma\sqrt{\frac{\log(ed/\sqrt{n})}{n}} \leq \frac{2\alpha\sigma^2}{n}\log(M)\,,$$

for β small enough if $d \geq Ck$ for some constant $C > 0$ chosen large enough. Applying now Theorem 5.11 yields

$$\inf_{\hat\theta}\sup_{\theta\in\mathbb{R}^d} \mathbb{P}_\theta\big(|\hat\theta - \theta|_2^2 \geq R\min\big(\frac{R}{8}, \beta^2\sigma^2\frac{\log(ed/\sqrt{n})}{8n}\big)\big) \geq \frac{1}{2} - 2\alpha\,.$$

It implies the following corollary.

Corollary 5.16. *Recall that $\mathcal{B}_1(R) \subset \mathbb{R}^d$ denotes the set vectors $\theta \in \mathbb{R}^d$ such that $|\theta|_1 \leq R$. Then there exist a constant $C > 0$ such that if $d \geq n^{1/2+\varepsilon}$, $\varepsilon > 0$, the minimax rate of estimation over $\mathcal{B}_1(R)$ in the Gaussian sequence model is*

$$\phi(\mathcal{B}_0(k)) = \min(R^2, R\sigma\frac{\log d}{n})\,.$$

Moreover, it is attained by the constrained least squares estimator $\hat\theta^{\mathrm{LS}}_{\mathcal{B}_1(R)}$ if $R \geq \sigma\frac{\log d}{n}$ and by the trivial estimator $\hat\theta = 0$ otherwise.

Proof. To complete the proof of the statement, we need to study risk of the trivial estimator equal to zero for small R. Note that if $|\theta^*|_1 \leq R$, we have

$$|0 - \theta^*|_2^2 = |\theta^*|_2^2 \leq |\theta^*|_1^2 = R^2\,.$$

□

Remark 5.17. Note that the inequality $|\theta^*|_2^2 \leq |\theta^*|_1^2$ appears to be quite loose. Nevertheless, it is tight up to a multiplicative constant for the vectors of the form $\theta_j = \omega_j \frac{R}{k}$ that are employed in the lower bound. Indeed, if $R \leq \sigma \frac{\log d}{n}$, we have $k \leq 2/\beta$

$$|\theta_j|_2^2 = \frac{R^2}{k} \geq \frac{\beta}{2} |\theta_j|_1^2 .$$

PROBLEM SET

Problem 5.1. (a) Prove the statement of Example 5.7.

(b) Let P_θ denote the distribution of $X \sim \mathsf{Ber}(\theta)$. Show that

$$\mathsf{KL}(P_\theta, P'_\theta) \geq C(\theta - \theta')^2 .$$

Problem 5.2. Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Prove that for any test ψ, it holds

$$\max_{j=0,1} \mathbb{P}_j(\psi \neq j) \geq \frac{1}{4} e^{-\mathsf{KL}(\mathbb{P}_0, \mathbb{P}_1)} .$$

Problem 5.3. For any $R > 0$, $\theta \in \mathbb{R}^d$, denote by $\mathcal{B}_2(\theta, R)$ the (Euclidean) ball of radius R and centered at θ. For any $\varepsilon > 0$ let $N = N(\varepsilon)$ be the largest integer such that there exist $\theta_1, \dots, \theta_N \in \mathcal{B}_2(0, 1)$ for which

$$|\theta_i - \theta_j| \geq 2\varepsilon$$

for all $i \neq j$. We call the set $\{\theta_1, \dots, \theta_N\}$ an ε-packing of $\mathcal{B}_2(0, 1)$.

(a) Show that there exists a constant $C > 0$ such that $N \leq C/\varepsilon^d$.

(b) Show that for any $x \in \mathcal{B}_2(0, 1)$, there exists $i = 1, \dots, N$ such that $|x - \theta_i|_2 \leq 2\varepsilon$.

(c) Use (b) to conclude that there exists a constant $C' > 0$ such that $N \geq C'/\varepsilon^d$.

Problem 5.4. Show that the rate $\phi = \sigma^2 d/n$ is the minimax rate of estimation over:

(a) The Euclidean Ball of $\mathbb{R}^d$ with radius $\sigma^2 d/n$.

(b) The unit ℓ_∞ ball of $\mathbb{R}^d$ defined by

$$\mathcal{B}_\infty(1) = \{\theta \in \mathbb{R}^d \,:\, \max_j |\theta_j| \leq 1\}$$

as long as $\sigma^2 \leq n$.

(c) The set of nonnegative vectors of $\mathbb{R}^d$.

(d) The discrete hypercube $\frac{\sigma}{16\sqrt{n}}\{0, 1\}^d$.

Problem 5.5. Fix $\beta \geq 5/3, Q > 0$ and prove that the minimax rate of estimation over $\Theta(\beta, Q)$ with the $\|\cdot\|_{L_2([0,1])}$-norm is given by $n^{-\frac{2\beta}{2\beta+1}}$.
`[Hint:Consider functions of the form`

$$f_j = \frac{C}{\sqrt{n}} \sum_{i=1}^{N} \omega_{ji} \varphi_i$$

`where` C `is a constant,` $\omega_j \in \{0, 1\}^N$ `for some appropriately chosen` N `and` $\{\varphi_j\}_{j \geq 1}$ `is the trigonometric basis.]`

Bibliography

[AS08] Noga Alon and Joel H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, Inc., Hoboken, NJ, third edition, 2008. With an appendix on the life and work of Paul Erdős.

[Ber09] Dennis S. Bernstein. *Matrix mathematics*. Princeton University Press, Princeton, NJ, second edition, 2009. Theory, facts, and formulas.

[Bil95] Patrick Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, third edition, 1995. A Wiley-Interscience Publication.

[Bir83] Lucien Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete*, 65(2):181–237, 1983.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[Cav11] Laurent Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, volume 203 of *Lect. Notes Stat. Proc.*, pages 3–96. Springer, Heidelberg, 2011.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[CT07] Emmanuel Candes and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[CZ12] T. Tony Cai and Harrison H. Zhou. Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica*, 22(4):1319–1349, 2012.

[CZZ10] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144, 2010.

[DDGS97] M.J. Donahue, C. Darken, L. Gurvits, and E. Sontag. Rates of convex approximation in non-hilbert spaces. *Constructive Approximation*, 13(2):187–220, 1997.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.

[FHT10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[Gru03] Branko Grunbaum. *Convex polytopes*, volume 221 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 2003. Prepared and with a preface by Volker Kaibel, Victor Klee and Günter M. Ziegler.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[IH81] I. A. Ibragimov and R. Z. Hasminskiĭ. *Statistical estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, New York, 1981. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[Joh11] Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Unpublished Manuscript., December 2011.

[KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329, 2011.

[LPTVDG11] K. Lounici, M. Pontil, A.B. Tsybakov, and S. Van De Geer. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

[Mal09] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier/Academic Press, Amsterdam, third edition, 2009. The sparse way, With contributions from Gabriel Peyré.

[Mar52] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.

[Nem00] Arkadi Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

[Pis81] G. Pisier. Remarques sur un résultat non publié de B. Maurey. In *Seminar on Functional Analysis, 1980–1981*, pages Exp. No. V, 13. École Polytech., Palaiseau, 1981.

[Rig06] Philippe Rigollet. Adaptive density estimation using the blockwise Stein method. *Bernoulli*, 12(2):351–370, 2006.

[RT11] Philippe Rigollet and Alexandre Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[Sha03] Jun Shao. *Mathematical statistics*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2003.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[Tsy03] Alexandre B. Tsybakov. Optimal rates of aggregation. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 303–313. Springer, 2003.

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

18.S997 High-dimensional Statistics
Spring 2015