18.S997: High Dimensional Statistics

Lecture Notes

(This version: July 14, 2015)

Philippe Rigollet

Spring 2015

Preface

These lecture notes were written for the course 18.S997: *High Dimensional Statistics* at MIT. They build on a set of notes that was prepared at Princeton University in 2013-14.

Over the past decade, statistics have undergone drastic changes with the development of high-dimensional statistical inference. Indeed, on each individual, more and more features are measured to a point that it usually far exceeds the number of observations. This is the case in biology and specifically genetics where millions of (or combinations of) genes are measured for a single individual. High resolution imaging, finance, online advertising, climate studies . . . the list of intensive data producing fields is too long to be established exhaustively. Clearly not all measured features are relevant for a given task and most of them are simply noise. But which ones? What can be done with so little data and so much noise? Surprisingly, the situation is not that bad and on some simple models we can assess to which extent meaningful statistical methods can be applied. Regression is one such simple model.

Regression analysis can be traced back to 1632 when Galileo Galilei used a procedure to infer a linear relationship from noisy data. It was not until the early 19th century that Gauss and Legendre developed a systematic procedure: the least-squares method. Since then, regression has been studied in so many forms that much insight has been gained and recent advances on high-dimensional statistics would not have been possible without standing on the shoulders of giants. In these notes, we will explore one, obviously subjective giant whose shoulders high-dimensional statistics stand: nonparametric statistics.

The works of Ibragimov and Has'minskii in the seventies followed by many researchers from the Russian school have contributed to developing a large toolkit to understand regression with an infinite number of parameters. Much insight from this work can be gained to understand high-dimensional or sparse regression and it comes as no surprise that Donoho and Johnstone have made the first contributions on this topic in the early nineties.

Acknowledgements. These notes were improved thanks to the careful reading and comments of Mark Cerenzia, Youssef El Moujahid, Georgina Hall, Jan-Christian Hütter, Gautam Kamath, Kevin Lin, Ali Makhdoumi, Yaroslav Mukhin, Ludwig Schmidt, Vira Semenova, Yuyan Wang, Jonathan Weed and Chiyuan Zhang.

These notes were written under the partial support of the National Science Foundation, CAREER award DMS-1053987.

Required background. I assume that the reader has had basic courses in probability and mathematical statistics. Some elementary background in analysis and measure theory is helpful but not required. Some basic notions of linear algebra, especially spectral decomposition of matrices is required for the latter chapters.

Notation

Functions, sets, vectors

$[n]$	Set of integers $[n] = \{1, \dots, n\}$
$\mathcal{S}^{d-1}$	Unit sphere in dimension d
$\mathbb{1}(\,\cdot\,)$	Indicator function
$\lvert x\rvert_q$	ℓ_q norm of x defined by $\lvert x\rvert_q = \left(\sum_i \lvert x_i\rvert^q\right)^{\frac{1}{q}}$ for $q > 0$
$\lvert x\rvert_0$	ℓ_0 norm of x defined to be the number of nonzero coordinates of x
$f^{(k)}$	k-th derivative of f
e_j	j-th vector of the canonical basis
A^c	complement of set A
$\mathrm{conv}(S)$	Convex hull of set S.
$a_n \lesssim b_n$	$a_n \leq C b_n$ for a numerical constant $C > 0$

Matrices

I_p	Identity matrix of $\mathbb{R}^p$
$\mathsf{Tr}(A)$	trace of a square matrix A
$M^\dagger$	Moore-Penrose pseudoinverse of M
$\nabla_x f(x)$	Gradient of f at x
$\nabla_x f(x)\lvert_{x=x_0}$	Gradient of f at x_0

Distributions

$\mathcal{N}(\mu, \sigma^2)$	Univariate Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$
$\mathcal{N}_d(\mu, \Sigma)$	d-variate distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$
$\mathsf{subG}(\sigma^2)$	Univariate sub-Gaussian distributions with variance proxy $\sigma^2 > 0$
$\mathsf{subG}_d(\sigma^2)$	d-variate sub-Gaussian distributions with variance proxy $\sigma^2 > 0$
$\mathsf{subE}(\sigma^2)$	sub-Exponential distributions with variance proxy $\sigma^2 > 0$
$\mathsf{Ber}(p)$	Bernoulli distribution with parameter $p \in [0, 1]$
$\mathsf{Bin}(n, p)$	Binomial distribution with parameters $n \geq 1, p \in [0, 1]$
$\mathsf{Lap}(\lambda)$	Double exponential (or Laplace) distribution with parameter $\lambda > 0$
P_X	Marginal distribution of X

Function spaces

$W(\beta, L)$	Sobolev class of functions
$\Theta(\beta, Q)$	Sobolev ellipsoid of $\ell_2(\mathbb{N})$

Contents

Introduction

This course is mainly about learning a regression function from a collection of observations. In this chapter, after defining this task formally, we give an overview of the course and the questions around regression. We adopt the statistical learning point of view where the task of *prediction* prevails. Nevertheless many interesting questions will remain unanswered when the last page comes: testing, model selection, implementation,...

REGRESSION ANALYSIS AND PREDICTION RISK

Model and definitions

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ where X is called *feature* and lives in a topological space $\mathcal{X}$ and $Y \in \mathcal{Y} \subset \mathbb{R}$ is called *response* or sometimes *label* when $\mathcal{Y}$ is a discrete set, e.g., $\mathcal{Y} = \{0, 1\}$. Often $\mathcal{X} \subset \mathbb{R}^d$, in which case X is called *vector of covariates* or simply *covariate*. Our goal will be to predict Y given X and for our problem to be meaningful, we need Y to depend nontrivially on X. Our task would be done if we had access to the conditional distribution of Y given X. This is the world of the probabilist. The statistician does not have access to this valuable information but rather, has to estimate it, at least partially. The regression function gives a simple summary of this conditional distribution, namely, the conditional expectation.

Formally, the *regression function of Y onto X* is defined by:

$$f(x) = \mathbb{E}[Y|X = x]\,, \quad x \in \mathcal{X}\,.$$

As we will see, it arises naturally in the context of prediction.

Best prediction and prediction risk

Suppose for a moment that you know the conditional distribution of Y given X. Given the realization of $X = x$, your goal is to predict the realization of Y. Intuitively, we would like to find a measurable[1] function $g : X \to \mathcal{Y}$ such that $g(X)$ is close to Y, in other words, such that $|Y - g(X)|$ is small. But $|Y - g(X)|$ is a random variable so it not clear what "small" means in this context. A somewhat arbitrary answer can be given by declaring a random

[1] all topological spaces are equipped with their Borel σ-algebra

variable Z small if $\mathbb{E}[Z^2] = [\mathbb{E}Z]^2 + \mathrm{var}[Z]$ is small. Indeed in this case, the expectation of Z is small and the fluctuations of Z around this value are also small. The function $R(g) = \mathbb{E}[Y - g(X)]^2$ is called the *L_2 risk of g* defined for $\mathbb{E}Y^2 < \infty$.

For any measurable function $g : \mathcal{X} \to \mathbb{R}$, the L_2 risk of g can be decomposed as

$$\begin{aligned}\mathbb{E}[Y - g(X)]^2 &= \mathbb{E}[Y - f(X) + f(X) - g(X)]^2 \\ &= \mathbb{E}[Y - f(X)]^2 + \mathbb{E}[f(X) - g(X)]^2 + 2\mathbb{E}[Y - f(X)][f(X) - g(X)]\end{aligned}$$

The cross-product term satisfies

$$\begin{aligned}\mathbb{E}[Y - f(X)][f(X) - g(X)] &= \mathbb{E}\big[\mathbb{E}\big([Y - f(X)][f(X) - g(X)]\big|X\big)\big] \\ &= \mathbb{E}\big[[\mathbb{E}(Y|X) - f(X)][f(X) - g(X)]\big] \\ &= \mathbb{E}\big[[f(X) - f(X)][f(X) - g(X)]\big] = 0\,.\end{aligned}$$

The above two equations yield

$$\mathbb{E}[Y - g(X)]^2 = \mathbb{E}[Y - f(X)]^2 + \mathbb{E}[f(X) - g(X)]^2 \geq \mathbb{E}[Y - f(X)]^2\,,$$

with equality iff $f(X) = g(X)$ almost surely.

We have proved that the regression function $f(x) = \mathbb{E}[Y|X = x], x \in \mathcal{X}$, enjoys the *best prediction* property, that is

$$\mathbb{E}[Y - f(X)]^2 = \inf_g \mathbb{E}[Y - g(X)]^2\,,$$

where the infimum is taken over all measurable functions $g : \mathcal{X} \to \mathbb{R}$.

Prediction and estimation

As we said before, in a statistical problem, we do not have access to the conditional distribution of Y given X or even to the regression function f of Y onto X. Instead, we observe a *sample* $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ that consists of independent copies of (X, Y). The goal of regression function estimation is to use this data to construct an estimator $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ that has small L_2 risk $R(\hat{f}_n)$.

Let P_X denote the marginal distribution of X and for any $h : \mathcal{X} \to \mathbb{R}$, define

$$\|h\|_2^2 = \int_{\mathcal{X}} h^2 \mathrm{d}P_X\,.$$

Note that $\|h\|_2^2$ is the Hilbert norm associated to the inner product

$$\langle h, h' \rangle_2 = \int_{\mathcal{X}} hh' \mathrm{d}P_X\,.$$

When the reference measure is clear from the context, we will simply write $\|h\|_2 L = \|h\|_{L_2(P_X)}$ and $\langle h, h' \rangle_2 := \langle h, h' \rangle_{L_2(P_X)}$.

It follows from the proof of the best prediction property above that

$$\begin{aligned} R(\hat{f}_n) &= \mathbb{E}[Y - f(X)]^2 + \|\hat{f}_n - f\|_2^2 \\ &= \inf_g \mathbb{E}[Y - g(X)]^2 + \|\hat{f}_n - f\|_2^2 \end{aligned}$$

In particular, the prediction risk will always be at least equal to the positive constant $\mathbb{E}[Y - f(X)]^2$. Since we tend to prefer a measure of accuracy to be able to go to zero (as the sample size increases), it is equivalent to study the *estimation error* $\|\hat{f}_n - f\|_2^2$. Note that if $\hat{f}_n$ is random, then $\|\hat{f}_n - f\|_2^2$ and $R(\hat{f}_n)$ are *random* quantities and we need deterministic summaries to quantify their size. It is customary to use one of the two following options. Let $\{\phi_n\}_n$ be a sequence of positive numbers that tends to zero as n goes to infinity.

1. **Bounds in expectation**. They are of the form:

$$\mathbb{E}\|\hat{f}_n - f\|_2^2 \leq \phi_n \,,$$

   where the expectation is taken with respect to the sample $\mathcal{D}_n$. They indicate the *average behavior* of the estimator over multiple realizations of the sample. Such bounds have been established in nonparametric statistics where typically $\phi_n = O(n^{-\alpha})$ for some $\alpha \in (1/2, 1)$ for example.

   Note that such bounds do not characterize the size of the deviation of the random variable $\|\hat{f}_n - f\|_2^2$ around its expectation. As a result, it may be therefore appropriate to accompany such a bound with the second option below.

2. **Bounds with high-probability**. They are of the form:

$$\mathbb{P}\big[\|\hat{f}_n - f\|_2^2 > \phi_n(\delta)\big] \leq \delta \,, \quad \forall \delta \in (0, 1/3) \,.$$

   Here $1/3$ is arbitrary and can be replaced by another positive constant. Such bounds control the tail of the distribution of $\|\hat{f}_n - f\|_2^2$. They show how large the quantiles of the random variable $\|f - \hat{f}_n\|_2^2$ can be. Such bounds are favored in learning theory, and are sometimes called PAC-bounds (for Probably Approximately Correct).

Often, bounds with high probability follow from a bound in expectation and a concentration inequality that bounds the following probability

$$\mathbb{P}\big[\|\hat{f}_n - f\|_2^2 - \mathbb{E}\|\hat{f}_n - f\|_2^2 > t\big]$$

by a quantity that decays to zero exponentially fast. Concentration of measure is a fascinating but wide topic and we will only briefly touch it. We recommend the reading of [BLM13] to the interested reader. This book presents many aspects of concentration that are particularly well suited to the applications covered in these notes.

Other measures of error

We have chosen the L_2 risk somewhat arbitrarily. Why not the L_p risk defined by $g \mapsto \mathbb{E}|Y - g(X)|^p$ for some $p \geq 1$? The main reason for choosing the L_2 risk is that it greatly simplifies the mathematics of our problem: it is a Hilbert space! In particular, for any estimator $\hat{f}_n$, we have the remarkable identity:

$$R(\hat{f}_n) = \mathbb{E}[Y - f(X)]^2 + \|\hat{f}_n - f\|_2^2 \,.$$

This equality allowed us to consider only the part $\|\hat{f}_n - f\|_2^2$ as a measure of error. While this decomposition may not hold for other risk measures, it may be desirable to explore other distances (or pseudo-distances). This leads to two distinct ways to measure error. Either by bounding a pseudo-distance $d(\hat{f}_n, f)$ (*estimation error*) or by bounding the risk $R(\hat{f}_n)$ for choices other than the L_2 risk. These two measures coincide up to the additive constant $\mathbb{E}[Y - f(X)]^2$ in the case described above. However, we show below that these two quantities may live independent lives. Bounding the estimation error is more customary in statistics whereas, risk bounds are preferred in learning theory.

Here is a list of choices for the pseudo-distance employed in the estimation error.

- **Pointwise error**. Given a point x_0, the pointwise error measures only the error at this point. It uses the pseudo-distance:

$$d_0(\hat{f}_n, f) = |\hat{f}_n(x_0) - f(x_0)| \,.$$

- **Sup-norm error**. Also known as the L_∞-error and defined by

$$d_\infty(\hat{f}_n, f) = \sup_{x \in \mathcal{X}} |\hat{f}_n(x) - f(x)| \,.$$

  It controls the worst possible pointwise error.

- **L_p-error**. It generalizes both the L_2 distance and the sup-norm error by taking for any $p \geq 1$, the pseudo distance

$$d_p(\hat{f}_n, f) = \int_{\mathcal{X}} |\hat{f}_n - f|^p \mathrm{d}P_X \,.$$

  The choice of p is somewhat arbitrary and mostly employed as a mathematical exercise.

Note that these three examples can be split into two families: global (Sup-norm and L_p) and local (pointwise).

For specific problems, other considerations come into play. For example, if $Y \in \{0, 1\}$ is a label, one may be interested in the *classification risk* of a classifier $h : \mathcal{X} \to \{0, 1\}$. It is defined by

$$R(h) = \mathbb{P}(Y \neq h(X)) \,.$$

We will not cover this problem in this course.

Finally, we will devote a large part of these notes to the study of linear models. For such models, $\mathcal{X} = \mathbb{R}^d$ and f is linear (or affine), i.e., $f(x) = x^\top \theta$ for some unknown $\theta \in \mathbb{R}^d$. In this case, it is traditional to measure error directly on the coefficient θ. For example, if $\hat{f}_n(x) = x^\top \hat{\theta}_n$ is a candidate linear estimator, it is customary to measure the distance of $\hat{f}_n$ to f using a (pseudo-)distance between $\hat{\theta}_n$ and θ as long as θ is identifiable.

MODELS AND METHODS

Empirical risk minimization

In our considerations on measuring the performance of an estimator $\hat{f}_n$, we have carefully avoided the question of how to construct $\hat{f}_n$. This is of course one of the most important task of statistics. As we will see, it can be carried out in a fairly mechanical way by following one simple principle: *Empirical Risk Minimization* (ERM[2]). Indeed, an overwhelming proportion of statistical methods consist in replacing an (unknown) expected value ($\mathbb{E}$) by a (known) empirical mean ($\frac{1}{n}\sum_{i=1}^n$). For example, it is well known that a good candidate to estimate the expected value $\mathbb{E}X$ of a random variable X from a sequence of i.i.d copies $X_1, \dots, X_n$ of X, is their empirical average

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \,.$$

In many instances, it corresponds the maximum likelihood estimator of $\mathbb{E}X$. Another example is the sample variance where $\mathbb{E}(X - \mathbb{E}(X))^2$ is estimated by

$$\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \,.$$

It turns out that this principle can be extended even if an optimization follows the substitution. Recall that the L_2 risk is defined by $R(g) = \mathbb{E}[Y - g(X)]^2$. See the expectation? Well, it can be replaced by an average to from the *empirical risk* of g defined by

$$R_n(g) = \frac{1}{n}\sum_{i=1}^n \big(Y_i - g(X_i)\big)^2 \,.$$

We can now proceed to minimizing this risk. However, we have to be careful. Indeed, $R_n(g) \geq 0$ for all g. Therefore any function g such that $Y_i = g(X_i)$ for all $i = 1, \dots, n$ is a minimizer of the empirical risk. Yet, it may not be the best choice (Cf. Figure 1). To overcome this limitation, we need to leverage some prior knowledge on f: either it may belong to a certain class $\mathcal{G}$ of functions (e.g., linear functions) or it is smooth (e.g., the L_2-norm of its second derivative is

[2]ERM may also mean *Empirical Risk Minimizer*

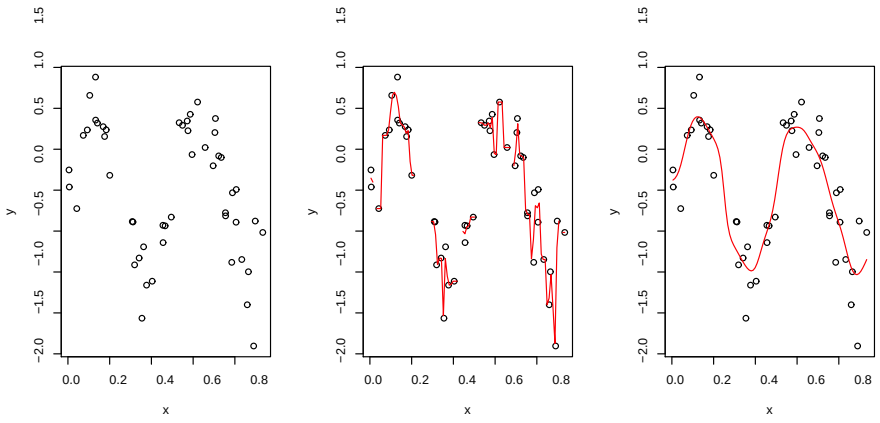


Figure 1. It may not be the best choice idea to have $\hat{f}_n(X_i) = Y_i$ for all $i = 1, \dots, n$.

small). In both cases, this extra knowledge can be incorporated to ERM using either a *constraint*:

$$\min_{g \in \mathcal{G}} R_n(g)$$

or a *penalty*:

$$\min_{g} \Big\{ R_n(g) + \mathrm{pen}(g) \Big\},$$

or both

$$\min_{g \in \mathcal{G}} \Big\{ R_n(g) + \mathrm{pen}(g) \Big\},$$

These schemes belong to the general idea of *regularization*. We will see many variants of regularization throughout the course.

Unlike traditional (low dimensional) statistics, *computation* plays a key role in high-dimensional statistics. Indeed, what is the point of describing an estimator with good prediction properties if it takes years to compute it on large datasets? As a result of this observation, much of the modern estimators, such as the Lasso estimator for sparse linear regression can be computed efficiently using simple tools from convex optimization. We will not describe such algorithms for this problem but will comment on the computability of estimators when relevant.

In particular computational considerations have driven the field of *compressed sensing* that is closely connected to the problem of sparse linear regression studied in these notes. We will only briefly mention some of the results and refer the interested reader to the book [FR13] for a comprehensive treatment.

Linear models

When $\mathcal{X} = \mathbb{R}^d$, an all time favorite constraint $\mathcal{G}$ is the class of linear functions that are of the form $g(x) = x^\top \theta$, that is parametrized by $\theta \in \mathbb{R}^d$. Under this constraint, the estimator obtained by ERM is usually called *least squares estimator* and is defined by $\hat{f}_n(x) = x^\top \hat{\theta}$, where

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \,.$$

Note that $\hat{\theta}$ may not be unique. In the case of a *linear model*, where we assume that the regression function is of the form $f(x) = x^\top \theta^*$ for some unknown $\theta^* \in \mathbb{R}^d$, we will need assumptions to ensure *identifiability* if we want to prove bounds on $d(\hat{\theta}, \theta^*)$ for some specific pseudo-distance $d(\cdot\,, \cdot)$. Nevertheless, in other instances such as regression with fixed design, we can prove bounds on the prediction error that are valid for any $\hat{\theta}$ in the argmin. In the latter case, we will not even require that f satisfies the linear model but our bound will be meaningful only if f can be well approximated by a linear function. In this case, we talk about *misspecified model*, i.e., we try to fit a linear model to data that may not come from a linear model. Since linear models can have good approximation properties especially when the dimension d is large, our hope is that the linear model is never too far from the truth.

In the case of a misspecified model, there is no hope to drive the estimation error $d(\hat{f}_n, f)$ down to zero even with a sample size that tends to infinity. Rather, we will pay a systematic approximation error. When $\mathcal{G}$ is a linear subspace as above, and the pseudo distance is given by the squared L_2 norm $d(\hat{f}_n, f) = \|\hat{f}_n - f\|_2^2$, it follows from the Pythagorean theorem that

$$\|\hat{f}_n - f\|_2^2 = \|\hat{f}_n - \bar{f}\|_2^2 + \|\bar{f} - f\|_2^2 \,,$$

where $\bar{f}$ is the projection of f onto the linear subspace $\mathcal{G}$. The systematic approximation error is entirely contained in the *deterministic* term $\|\bar{f} - f\|_2^2$ and one can proceed to bound $\|\hat{f}_n - \bar{f}\|_2^2$ by a quantity that goes to zero as n goes to infinity. In this case, bounds (e.g., in expectation) on the estimation error take the form

$$\mathbb{E}\|\hat{f}_n - f\|_2^2 \le \|\bar{f} - f\|_2^2 + \phi_n \,.$$

The above inequality is called an *oracle inequality*. Indeed, it says that if ϕ_n is small enough, then $\hat{f}_n$ the estimator mimics the *oracle* $\bar{f}$. It is called "oracle" because it cannot be constructed without the knowledge of the unknown f. It is clearly the best we can do when we restrict our attentions to estimator in the class $\mathcal{G}$. Going back to the gap in knowledge between a probabilist who knows the whole joint distribution of (X, Y) and a statistician who only see the data, the oracle sits somewhere in-between: it can only see the whole distribution through the lens provided by the statistician. In the case, above, the lens is that of linear regression functions. Different oracles are more or less powerful and there is a tradeoff to be achieved. On the one hand, if the oracle is weak,

then it's easy for the statistician to mimic it but it may be very far from the true regression function; on the other hand, if the oracle is strong, then it is harder to mimic but it is much closer to the truth.

Oracle inequalities were originally developed as analytic tools to prove adaptation of some nonparametric estimators. With the development of aggregation [Nem00, Tsy03, Rig06] and high dimensional statistics [CT07, BRT09, RT11], they have become important finite sample results that characterize the interplay between the important parameters of the problem.

In some favorable instances, that is when the X_is enjoy specific properties, it is even possible to estimate the vector θ accurately, as is done in parametric statistics. The techniques employed for this goal will essentially be the same as the ones employed to minimize the prediction risk. The extra assumptions on the X_is will then translate in interesting properties on $\hat{\theta}$ itself, including uniqueness on top of the prediction properties of the function $\hat{f}_n(x) = x^\top \hat{\theta}$.

High dimension and sparsity

These lecture notes are about high dimensional statistics and it is time they enter the picture. By high dimension, we informally mean that the model has more "parameters" than there are observations. The word "parameter" is used here loosely and a more accurate description is perhaps *degrees of freedom*. For example, the linear model $f(x) = x^\top \theta^*$ has one parameter θ^* but effectively d degrees of freedom when $\theta^* \in \mathbb{R}^d$. The notion of degrees of freedom is actually well defined in the statistical literature but the formal definition does not help our informal discussion here.

As we will see in Chapter 2, if the regression function is linear $f(x) = x^\top \theta^*$, $\theta^* \in \mathbb{R}^d$, and under some assumptions on the marginal distribution of X, then the least squares estimator $\hat{f}_n(x) = x^\top \hat{\theta}_n$ satisfies

$$\mathbb{E}\|\hat{f}_n - f\|_2^2 \leq C\frac{d}{n}\,, \tag{1}$$

where $C > 0$ is a constant and in Chapter 5, we will show that this cannot be improved apart perhaps for a smaller multiplicative constant. Clearly such a bound is uninformative if $d \gg n$ and actually, in view of its optimality, we can even conclude that the problem is too difficult statistically. However, the situation is not hopeless if we assume that the problem has actually less degrees of freedom than it seems. In particular, it is now standard to resort to the *sparsity* assumption to overcome this limitation.

A vector $\theta \in \mathbb{R}^d$ is said to be k-sparse for some $k \in \{0, \dots, d\}$ if it has at most k non-zero coordinates. We denote by $|\theta|_0$ the number of nonzero coordinates of θ is also known as sparsity or "ℓ_0-norm" though it is clearly not a norm (see footnote 3). Formally, it is defined as

$$|\theta|_0 = \sum_{j=1}^{d} \mathbb{1}(\theta_j \neq 0)\,.$$

Sparsity is just one of many ways to limit the size of the set of potential θ vectors to consider. One could consider vectors θ that have the following structure for example (see Figure 2):

- Monotonic: $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_d$
- Smooth: $|\theta_i - \theta_j| \leq C|i-j|^\alpha$ for some $\alpha > 0$
- Piecewise constant: $\sum_{j=1}^{d-1} \mathbb{1}(\theta_{j+1} \neq \theta_j) \leq k$
- Structured in another basis: $\theta = \Psi\mu$, for some orthogonal matrix and μ is in one of the structured classes described above.

Sparsity plays a significant role in statistics because, often, structure translate into sparsity in a certain basis. For example a smooth function is sparse in the trigonometric basis and a piecewise constant function has sparse increments. Moreover, as we will real images for example are approximately sparse in certain bases such as wavelet or Fourier bases. This is precisely the feature exploited in compression schemes such as JPEG or JPEG-2000: only a few coefficients in these images are necessary to retain the main features of the image.

We say that θ is *approximately sparse* if $|\theta|_0$ may be as large as d but many coefficients $|\theta_j|$ are small rather than exactly equal to zero. There are several mathematical ways to capture this phenomena, including ℓ_q-"balls" for $q \leq 1$. For $q > 0$, the unit ℓ_q-ball of $\mathbb{R}^d$ is defined as

$$\mathcal{B}_q(R) = \left\{\theta \in \mathbb{R}^d \,:\, |\theta|_q^q = \sum_{j=1}^d |\theta_j|^q \leq 1\right\}$$

where $|\theta|_q$ is often called ℓ_q-norm[3]. As we will see, the smaller q is, the better vectors in the unit ℓ_q ball can be approximated by sparse vectors.

Note that the set of k-sparse vectors of $\mathbb{R}^d$ is a union of $\sum_{j=0}^k \binom{d}{j}$ linear subspaces with dimension at most k and that are spanned by at most k vectors in the canonical basis of $\mathbb{R}^d$. If we knew that θ^* belongs to one of these subspaces, we could simply drop irrelevant coordinates and obtain an oracle inequality such as (1), with d replaced by k. Since we do not know what subspace θ^* lives exactly, we will have to pay an extra term to *find* in which subspace θ^* lives. This it turns out that this term is exactly of the the order of

$$\frac{\log\left(\sum_{j=0}^k \binom{d}{j}\right)}{n} \simeq C\frac{k\log\left(\frac{ed}{k}\right)}{n}$$

Therefore, the price to pay for not knowing which subspace to look at is only a logarithmic factor.

[3]Strictly speaking, $|\theta|_q$ is a norm and the ℓ_q ball is a ball only for $q \geq 1$.

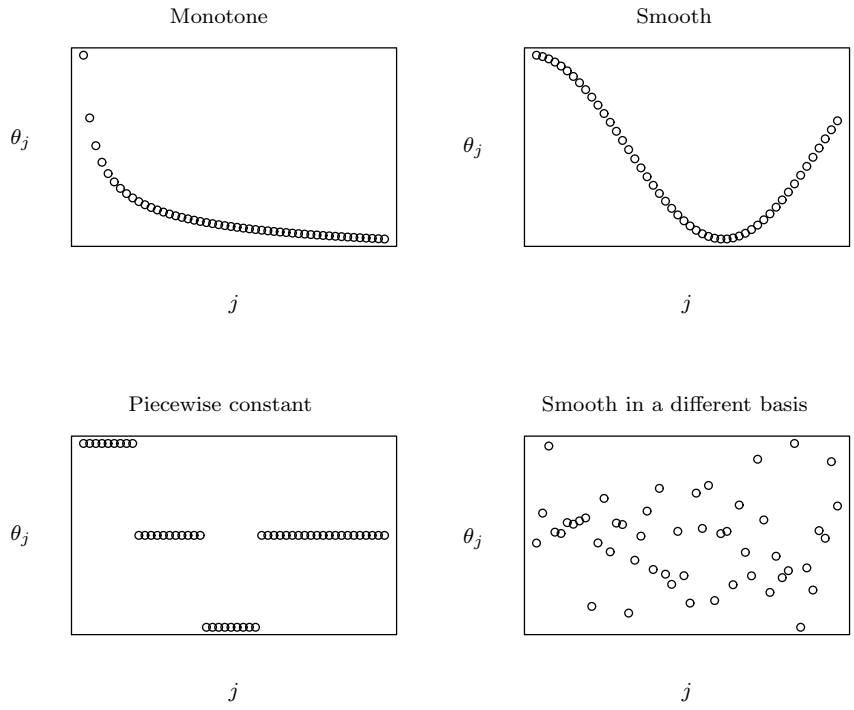


Figure 2. Examples of structures vectors $\theta \in \mathbb{R}^{50}$

Nonparametric regression

Nonparametric does not mean that there is no parameter to estimate (the regression function is a parameter) but rather that the parameter to estimate is infinite dimensional (this is the case of a function). In some instances, this parameter can be identified to an infinite sequence of real numbers, so that we are still in the realm of countable infinity. Indeed, observe that since $L_2(P_X)$ equipped with the inner product $\langle \cdot\,, \cdot \rangle_2$ is a separable Hilbert space, it admits an orthonormal basis $\{\varphi_k\}_{k\in\mathbb{Z}}$ and any function $f \in L_2(P_X)$ can be decomposed as

$$f = \sum_{k\in\mathbb{Z}} \alpha_k \varphi_k \,,$$

where $\alpha_k = \langle f, \varphi_k \rangle_2$.

Therefore estimating a regression function f amounts to estimating the infinite sequence $\{\alpha_k\}_{k\in\mathbb{Z}} \in \ell_2$. You may argue (correctly) that the basis $\{\varphi_k\}_{k\in\mathbb{Z}}$ is also unknown as it depends on the unknown P_X. This is absolutely

correct but we will make the convenient assumption that P_X is (essentially) known whenever this is needed.

Even if infinity is countable, we still have to estimate an infinite number of coefficients using a finite number of observations. It does not require much statistical intuition to realize that this task is impossible in general. What if we know something about the sequence $\{\alpha_k\}_k$? For example, if we know that $\alpha_k = 0$ for $|k| > k_0$, then there are only $2k_0 + 1$ parameters to estimate (in general, one would also have to "estimate" k_0). In practice, we will not exactly see $\alpha_k = 0$ for $|k| > k_0$, but rather that the sequence $\{\alpha_k\}_k$ decays to 0 at a certain polynomial rate. For example $|\alpha_k| \leq C|k|^{-\gamma}$ for some $\gamma > 1/2$ (we need this sequence to be in ℓ_2). It corresponds to a smoothness assumption on the function f. In this case, the sequence $\{\alpha_k\}_k$ can be well approximated by a sequence with only a finite number of non-zero terms.

We can view this problem as a misspecified model. Indeed, for any cut-off k_0, define the oracle

$$\bar{f}_{k_0} = \sum_{|k| \leq k_0} \alpha_k \varphi_k \,.$$

Note that it depends on the unknown α_k and define the estimator

$$\hat{f}_n = \sum_{|k| \leq k_0} \hat{\alpha}_k \varphi_k \,,$$

where $\hat{\alpha}_k$ are some data-driven coefficients (obtained by least-squares for example). Then by the Pythagorean theorem and Parseval's identity, we have

$$\begin{aligned} \|\hat{f}_n - f\|_2^2 &= \|\bar{f} - f\|_2^2 + \|\hat{f}_n - \bar{f}\|_2^2 \\ &= \sum_{|k| > k_0} \alpha_k^2 + \sum_{|k| \leq k_0} (\hat{\alpha}_k - \alpha_k)^2 \end{aligned}$$

We can even work further on this oracle inequality using the fact that $|\alpha_k| \leq C|k|^{-\gamma}$. Indeed, we have[4]

$$\sum_{|k| > k_0} \alpha_k^2 \leq C^2 \sum_{|k| > k_0} k^{-2\gamma} \leq C k_0^{1-2\gamma} \,.$$

The so called *stochastic term* $\mathbb{E} \sum_{|k| \leq k_0} (\hat{\alpha}_k - \alpha_k)^2$ clearly increases with k_0 (more parameters to estimate) whereas the *approximation term* $C k_0^{1-2\gamma}$ decreases with k_0 (less terms discarded). We will see that we can strike a compromise called *bias-variance tradeoff*.

The main difference here with oracle inequalities is that we make assumptions on the regression function (here in terms of smoothness) in order to

[4]Here we illustrate a convenient notational convention that we will be using throughout these notes: a constant C may be different from line to line. This will not affect the interpretation of our results since we are interested in the order of magnitude of the error bounds. Nevertheless we will, as much as possible, try to make such constants explicit. As an exercise, try to find an expression of the second C as a function of the first one and of γ.

control the approximation error. Therefore oracle inequalities are more general but can be seen on the one hand as less quantitative. On the other hand, if one is willing to accept the fact that approximation error is inevitable then there is no reason to focus on it. This is not the final answer to this rather philosophical question. Indeed, choosing the right k_0 can only be done with a control of the approximation error. Indeed, the best k_0 will depend on γ. We will see that even if the smoothness index γ is unknown, we can select k_0 in a data-driven way that achieves almost the same performance as if γ were known. This phenomenon is called *adaptation* (to γ).

It is important to notice the main difference between the approach taken in nonparametric regression and the one in sparse linear regression. It is not so much about linear vs. nonlinear model as we can always first take nonlinear transformations of the x_j's in linear regression. Instead, sparsity or approximate sparsity is a much weaker notion than the decay of coefficients $\{\alpha_k\}_k$ presented above. In a way, sparsity only imposes that *after ordering* the coefficients present a certain decay, whereas in nonparametric statistics, the order is set ahead of time: we assume that we have found a basis that is ordered in such a way that coefficients decay at a certain rate.

Matrix models

In the previous examples, the response variable is always assumed to be a scalar. What if it is a higher dimensional signal? In Chapter 4, we consider various problems of this form: matrix completion a.k.a. the Netflix problem, structured graph estimation and covariance matrix estimation. All these problems can be described as follows.

Let M, S and N be three matrices, respectively called *observation*, *signal* and *noise*, and that satisfy

$$M = S + N\,.$$

Here N is a random matrix such that $\mathbb{E}[N] = 0$, the all-zero matrix. The goal is to estimate the signal matrix S from the observation of M.

The structure of S can also be chosen in various ways. We will consider the case where S is sparse in the sense that it has many zero coefficients. In a way, this assumption does not leverage much of the matrix structure and essentially treats matrices as vectors arranged in the form of an array. This is not the case of *low rank* structures where one assumes that the matrix S has either low rank or can be well approximated by a low rank matrix. This assumption makes sense in the case where S represents user preferences as in the Netflix example. In this example, the (i, j)th coefficient S_{ij} of S corresponds to the rating (on a scale from 1 to 5) that user i gave to movie j. The low rank assumption simply materializes the idea that there are a few canonical profiles of users and that each user can be represented as a linear combination of these users.

At first glance, this problem seems much more difficult than sparse linear regression. Indeed, one needs to learn not only the sparse coefficients in a given

basis, but also the basis of eigenvectors. Fortunately, it turns out that the latter task is much easier and is dominated by the former in terms of statistical price.

Another important example of matrix estimation is high-dimensional covariance estimation, where the goal is to estimate the covariance matrix of a random vector $X \in \mathbb{R}^d$, or its leading eigenvectors, based on n observations. Such a problem has many applications including principal component analysis, linear discriminant analysis and portfolio optimization. The main difficulty is that n may be much smaller than the number of degrees of freedom in the covariance matrix, which can be of order d^2. To overcome this limitation, assumptions on the rank or the sparsity of the matrix can be leveraged.

Optimality and minimax lower bounds

So far, we have only talked about upper bounds. For a linear model, where $f(x) = x^\top \theta^*$, we will prove in Chapter 2 the following bound for a modified least squares estimator $\hat{f}_n = x^\top \hat{\theta}$

$$\mathbb{E}\|\hat{f}_n - f\|_2^2 \leq C\frac{d}{n}\,.$$

Is this the right dependence in p and n? Would it be possible to obtain as an upper bound: $C(\log d)/n$, C/n or $\sqrt{d}/n^2$, by either improving our proof technique or using another estimator altogether? It turns out that the answer to this question is negative. More precisely, we can prove that for any estimator $\tilde{f}_n$, there exists a function f of the form $f(x) = x^\top \theta^*$ such that

$$\mathbb{E}\|\hat{f}_n - f\|_2^2 > c\frac{d}{n}$$

for some positive constant c. Here we used a different notation for the constant to emphasize the fact that lower bounds guarantee optimality only *up to a constant* factor. Such a lower bound on the risk is called *minimax lower bound* for reasons that will become clearer in chapter 5.

How is this possible? How can we make a statement *for all* estimators? We will see that these statements borrow from the theory of tests where we know that it is impossible to drive both the type I and the type II error to zero simultaneously (with a fixed sample size). Intuitively this phenomenon is related to the following observation: Given n observations $X_1, \dots, X_n$, it is hard to tell if they are distributed according to $\mathcal{N}(\theta, 1)$ or to $\mathcal{N}(\theta', 1)$ for a Euclidean distance $|\theta - \theta'|_2$ is small enough. We will see that it is the case for example if $|\theta - \theta'|_2 \leq C\sqrt{d/n}$, which will yield our lower bound.

CHAPTER 1

Sub-Gaussian Random Variables

1.1 GAUSSIAN TAILS AND MGF

Recall that a random variable $X \in \mathbb{R}$ has Gaussian distribution iff it has a density p with respect to the Lebesgue measure on $\mathbb{R}$ given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big) , \quad x \in \mathbb{R} ,$$

where $\mu = \mathbb{E}(X) \in \mathbb{R}$ and $\sigma^2 = \mathrm{var}(X) > 0$ are the *mean* and *variance* of X. We write $X \sim \mathcal{N}(\mu, \sigma^2)$. Note that $X = \sigma Z + \mu$ for $Z \sim \mathcal{N}(0,1)$ (called standard Gaussian) and where the equality holds in distribution. Clearly, this distribution has unbounded support but it is well known that it has *almost* bounded support in the following sense: $\mathbb{P}(|X - \mu| \leq 3\sigma) \simeq 0.997$. This is due to the fast decay of the tails of p as $|x| \to \infty$ (see Figure 1.1). This decay can be quantified using the following proposition (Mills inequality).

Proposition 1.1. Let X be a Gaussian random variable with mean μ and variance σ^2 then for any $t > 0$, it holds

$$\mathbb{P}(X - \mu > t) \leq \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t} .$$

By symmetry we also have

$$\mathbb{P}(X - \mu < -t) \leq \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t} .$$

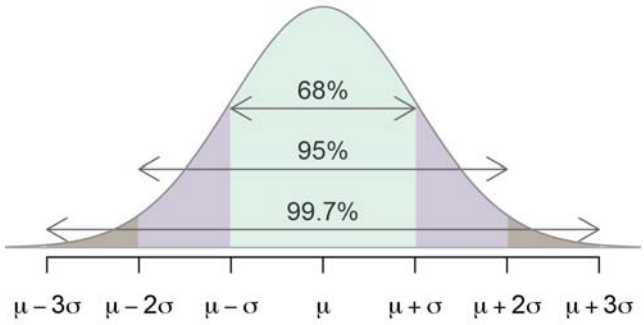


Figure 1.1. Probabilities of falling within 1, 2, and 3 standard deviations close to the mean in a Gaussian distribution. Source http://www.openintro.org/

and

$$\mathbb{P}(|X-\mu|>t) \le \sqrt{\frac{2}{\pi}}\frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t}\,.$$

Proof. Note that it is sufficient to prove the theorem for $\mu = 0$ and $\sigma^2 = 1$ by simple translation and rescaling. We get for $Z \sim \mathcal{N}(0,1)$,

$$\begin{aligned}
\mathbb{P}(Z>t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&\le \frac{1}{\sqrt{2\pi}}\int_t^\infty \frac{x}{t}\exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&= \frac{1}{t\sqrt{2\pi}}\int_t^\infty -\frac{\partial}{\partial x}\exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&= \frac{1}{t\sqrt{2\pi}}\exp(-t^2/2)\,.
\end{aligned}$$

The second inequality follows from symmetry and the last one using the union bound:

$$\mathbb{P}(|Z|>t) = \mathbb{P}(\{Z>t\}\cup\{Z<-t\}) \le \mathbb{P}(Z>t)+\mathbb{P}(Z<-t) = 2\mathbb{P}(Z>t)\,.$$

□

The fact that a Gaussian random variable Z has tails that decay to zero exponentially fast can also be seen in the *moment generating function* (MGF)

$$M\,:\,s\mapsto M(s) = \mathbb{E}[\exp(sZ)]\,.$$

Indeed in the case of a standard Gaussian random variable, we have

$$\begin{aligned} M(s) = \mathbb{E}[\exp(sZ)] &= \frac{1}{\sqrt{2\pi}} \int e^{sz} e^{-\frac{z^2}{2}} \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-s)^2}{2} + \frac{s^2}{2}} \mathrm{d}z \\ &= e^{\frac{s^2}{2}} . \end{aligned}$$

It follows that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[\exp(sX)] = \exp\left(s\mu + \frac{\sigma^2 s^2}{2}\right)$.

1.2 SUB-GAUSSIAN RANDOM VARIABLES AND CHERNOFF BOUNDS

Definition and first properties

Gaussian tails are practical when controlling the tail of an average of independent random variables. Indeed, recall that if $X_1, \ldots, X_n$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \sigma^2/n)$. Using Lemma 1.3 below for example, we get

$$\mathbb{P}(|\bar{X} - \mu| > t) \leq 2\exp\Big(-\frac{nt^2}{2\sigma^2}\Big) .$$

Equating the right-hand side with some confidence level $\delta > 0$, we find that with probability at least[1] $1 - \delta$,

$$\mu \in \left[\bar{X} - \sigma\sqrt{\frac{2\log(2/\delta)}{n}}, \bar{X} + \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right], \tag{1.1}$$

This is almost the confidence interval that you used in introductory statistics. The only difference is that we used an approximation for the Gaussian tail whereas statistical tables or software use a much more accurate computation. Figure 1.2 shows the ration of the width of the confidence interval to that of the confidence interval computer by the software R. It turns out that intervals of the same form can be also derived for non-Gaussian random variables as long as they have sub-Gaussian tails.

Definition 1.2. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall\, s \in \mathbb{R} . \tag{1.2}$$

In this case we write $X \sim \mathsf{subG}(\sigma^2)$. Note that $\mathsf{subG}(\sigma^2)$ denotes a class of distributions rather than a distribution. Therefore, we abuse notation when writing $X \sim \mathsf{subG}(\sigma^2)$.

More generally, we can talk about sub-Gaussian random vectors and matrices. A random vector $X \in \mathbb{R}^d$ is said to be *sub-Gaussian* with variance

[1] We will often commit the statement "at least" for brevity

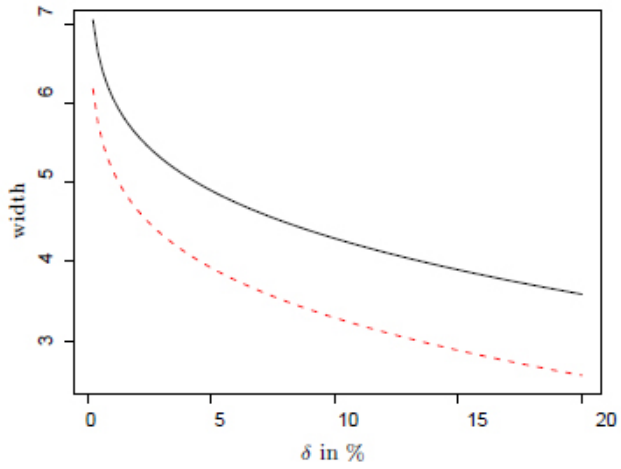


Figure 1.2. Width of confidence intervals from exact computation in R (red dashed) and (1.1) (solid black).

proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^\top X$ is sub-Gaussian with variance proxy σ^2 for any unit vector $u \in \mathcal{S}^{d-1}$. In this case we write $X \sim \mathsf{subG}_d(\sigma^2)$. A random matrix $X \in \mathbb{R}^{d\times T}$ is said to be *sub-Gaussian* with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^\top X v$ is sub-Gaussian with variance proxy σ^2 for any unit vectors $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$. In this case we write $X \sim \mathsf{subG}_{d\times T}(\sigma^2)$.

This property can equivalently be expressed in terms of bounds on the tail of the random variable X.

Lemma 1.3. *Let $X \sim \mathsf{subG}(\sigma^2)$. Then for any $t > 0$, it holds*

$$\mathbb{P}[X > t] \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big), \quad \textit{and} \quad \mathbb{P}[X < -t] \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big). \tag{1.3}$$

Proof. Assume first that $X \sim \mathsf{subG}(\sigma^2)$. We will employ a very useful technique called **Chernoff bound** that allows to to translate a bound on the moment generating function into a tail bound. Using Markov's inequality, we have for any $s > 0$,

$$\mathbb{P}(X > t) \le \mathbb{P}\big(e^{sX} > e^{st}\big) \le \frac{\mathbb{E}\big[e^{sX}\big]}{e^{st}}.$$

Next we use the fact that X is sub-Gaussian to get

$$\mathbb{P}(X > t) \le e^{\frac{\sigma^2 s^2}{2} - st}.$$

The above inequality holds for any $s > 0$ so to make it the tightest possible, we minimize with respect to $s > 0$. Solving $\phi'(s) = 0$, where $\phi(s) = \frac{\sigma^2 s^2}{2} - st$, we find that $\inf_{s>0} \phi(s) = -\frac{t^2}{2\sigma^2}$. This proves the first part of (1.3). The second inequality in this equation follows in the same manner (recall that (1.2) holds for any $s \in \mathbb{R}$).

□

Moments

Recall that the absolute moments of $Z \sim \mathcal{N}(0, \sigma^2)$ are given by

$$\mathbb{E}[|Z|^k] = \frac{1}{\sqrt{\pi}}(2\sigma^2)^{k/2}\Gamma\big(\frac{k+1}{2}\big)$$

where $\Gamma(\cdot)$ denote the Gamma function defined by

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x\,, \quad t > 0\,.$$

The next lemma shows that the tail bounds of Lemma 1.3 are sufficient to show that the absolute moments of $X \sim \mathsf{subG}(\sigma^2)$ can be bounded by those of $Z \sim \mathcal{N}(0, \sigma^2)$ up to multiplicative constants.

Lemma 1.4. *Let X be a random variable such that*

$$\mathbb{P}[|X| > t] \le 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big)\,,$$

then for any positive integer $k \ge 1$,

$$\mathbb{E}[|X|^k] \le (2\sigma^2)^{k/2}k\Gamma(k/2)\,.$$

In particular,

$$\big(\mathbb{E}[|X|^k]\big)^{1/k} \le \sigma e^{1/e}\sqrt{k}\,, \quad k \ge 2\,.$$

and $\mathbb{E}[|X|] \le \sigma\sqrt{2\pi}$.

Proof.

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}(|X| > t^{1/k})\mathrm{d}t \\
&\le 2\int_0^\infty e^{-\frac{t^{2/k}}{2\sigma^2}}\mathrm{d}t \\
&= (2\sigma^2)^{k/2}k\int_0^\infty e^{-u}u^{k/2-1}\mathrm{d}u\,, \qquad u = \frac{t^{2/k}}{2\sigma^2} \\
&= (2\sigma^2)^{k/2}k\Gamma(k/2)
\end{aligned}$$

The second statement follows from $\Gamma(k/2) \leq (k/2)^{k/2}$ and $k^{1/k} \leq e^{1/e}$ for any $k \geq 2$. It yields

$$\big((2\sigma^2)^{k/2} k\Gamma(k/2)\big)^{1/k} \leq k^{1/k}\sqrt{\frac{2\sigma^2 k}{2}} \leq e^{1/e}\sigma\sqrt{k}\,.$$

Moreover, for $k = 1$, we have $\sqrt{2}\Gamma(1/2) = \sqrt{2\pi}$. □

Using moments, we can prove the following reciprocal to Lemma 1.3.

Lemma 1.5. *If (1.3) holds, then for any $s > 0$, it holds*

$$\mathbb{E}[\exp(sX)] \leq e^{4\sigma^2 s^2}\,.$$

As a result, we will sometimes write $X \sim \mathsf{subG}(\sigma^2)$ when it satisfies (1.3).

Proof. We use the Taylor expansion of the exponential function as follows. Observe that by the dominated convergence theorem

$$\begin{aligned}
\mathbb{E}\big[e^{sX}\big] &\leq 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}[|X|^k]}{k!} \\
&\leq 1 + \sum_{k=2}^{\infty} \frac{(2\sigma^2 s^2)^{k/2} k\Gamma(k/2)}{k!} \\
&= 1 + \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^{k} 2k\Gamma(k)}{(2k)!} + \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^{k+1/2}(2k+1)\Gamma(k+1/2)}{(2k+1)!} \\
&\leq 1 + \big(2 + \sqrt{2\sigma^2 s^2}\big) \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^k k!}{(2k)!} \\
&\leq 1 + \big(1 + \sqrt{\frac{\sigma^2 s^2}{2}}\big) \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^k}{k!} && 2(k!)^2 \leq (2k)! \\
&= e^{2\sigma^2 s^2} + \sqrt{\frac{\sigma^2 s^2}{2}}(e^{2\sigma^2 s^2} - 1) \\
&\leq e^{4\sigma^2 s^2}\,.
\end{aligned}$$

□

From the above Lemma, we see that sub-Gaussian random variables can be equivalently defined from their tail bounds and their moment generating functions, up to constants.

Sums of independent sub-Gaussian random variables

Recall that if $X_1, \ldots, X_n$ are i.i.d $\mathcal{N}(0, \sigma^2)$, then for any $a \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(0, |a|_2^2 \sigma^2).$$

If we only care about the tails, this property is preserved for sub-Gaussian random variables.

Theorem 1.6. *Let $X = (X_1, \dots, X_n)$ be a vector of independent sub-Gaussian random variables that have variance proxy σ^2. Then, the random vector X is sub-Gaussian with variance proxy σ^2.*

Proof. Let $u \in \mathcal{S}^{n-1}$ be a unit vector, then

$$\mathbb{E}[e^{su^\top X}] = \prod_{i=1}^n \mathbb{E}[e^{su_i X_i}] \leq \prod_{i=1}^n e^{\frac{\sigma^2 s^2 u_i^2}{2}} = e^{\frac{\sigma^2 s^2 |u|_2^2}{2}} = e^{\frac{\sigma^2 s^2}{2}} .$$

□

Using a Chernoff bound, we immediately get the following corollary

Corollary 1.7. *Let $X_1, \dots, X_n$ be n independent random variables such that $X_i \sim \mathsf{subG}(\sigma^2)$. Then for any $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\Big[\sum_{i=1}^n a_i X_i > t\Big] \leq \exp\Big(-\frac{t^2}{2\sigma^2 |a|_2^2}\Big),$$

and

$$\mathbb{P}\Big[\sum_{i=1}^n a_i X_i < -t\Big] \leq \exp\Big(-\frac{t^2}{2\sigma^2 |a|_2^2}\Big)$$

Of special interest is the case where $a_i = 1/n$ for all i. Then, we get that the average $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, satisfies

$$\mathbb{P}(\bar{X} > t) \leq e^{-\frac{nt^2}{2\sigma^2}} \quad \text{and} \quad \mathbb{P}(\bar{X} < -t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

just like for the Gaussian average.

Hoeffding's inequality

The class of subGaussian random variables is actually quite large. Indeed, Hoeffding's lemma below implies that all randdom variables that are bounded uniformly are actually subGaussian with a variance proxy that depends on the size of their support.

Lemma 1.8 (Hoeffding's lemma (1963))**.** *Let X be a random variable such that $\mathbb{E}(X) = 0$ and $X \in [a, b]$ almost surely. Then, for any $s \in \mathbb{R}$, it holds*

$$\mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}} .$$

In particular, $X \sim \mathsf{subG}(\frac{(b-a)^2}{4})$.

Proof. Define $\psi(s) = \log \mathbb{E}[e^{sX}]$, and observe that and we can readily compute

$$\psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}, \qquad \psi''(s) = \frac{\mathbb{E}[X^2e^{sX}]}{\mathbb{E}[e^{sX}]} - \left[\frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}\right]^2.$$

Thus $\psi''(s)$ can be interpreted as the variance of the random variable X under the probability measure $d\mathbb{Q} = \frac{e^{sX}}{\mathbb{E}[e^{sX}]} d\mathbb{P}$. But since $X \in [a,b]$ almost surely, we have, under any probability,

$$\operatorname{var}(X) = \operatorname{var}\big(X - \frac{a+b}{2}\big) \le \mathbb{E}\Big[\Big(X - \frac{a+b}{2}\Big)^2\Big] \le \frac{(b-a)^2}{4}.$$

The fundamental theorem of calculus yields

$$\psi(s) = \int_0^s \int_0^\mu \psi''(\rho)\, d\rho\, d\mu \le \frac{s^2(b-a)^2}{8}$$

using $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbb{E}X = 0$. □

Using a Chernoff bound, we get the following (extremely useful) result.

Theorem 1.9 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be n independent random variables such that almost surely,*

$$X_i \in [a_i, b_i], \qquad \forall\, i.$$

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then for any $t > 0$,

$$\mathbb{P}(\bar{X} - \mathbb{E}(\bar{X}) > t) \le \exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big),$$

and

$$\mathbb{P}(\bar{X} - \mathbb{E}(\bar{X}) < -t) \le \exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Note that Hoeffding's lemma is for *any* bounded random variables. For example, if one knows that X is a Rademacher random variable. Then

$$\mathbb{E}(e^{sX}) = \frac{e^s + e^{-s}}{2} = \cosh(s) \le e^{\frac{s^2}{2}}$$

Note that 2 is the best possible constant in the above approximation. For such variables $a = -1, b = 1$, $\mathbb{E}(X) = 0$ so Hoeffding's lemma yields

$$\mathbb{E}(e^{sX}) \le e^{\frac{s^2}{2}}.$$

Hoeffding's inequality is very general but there is a price to pay for this generality. Indeed, if the random variables have small variance, we would like to see it reflected in the exponential tail bound (like for the Gaussian case) but the variance does not appear in Hoeffding's inequality. We need a more refined inequality.

1.3 SUB-EXPONENTIAL RANDOM VARIABLES

What can we say when a centered random variable is not sub-Gaussian? A typical example is the double exponential (or Laplace) distribution with parameter 1, denoted by $\mathsf{Lap}(1)$. Let $X \sim \mathsf{Lap}(1)$ and observe that

$$\mathbb{P}(|X| > t) = e^{-t}, \quad t \geq 0\,.$$

In particular, the tails of this distribution do not decay as fast as the Gaussian ones (that decay as $e^{-t^2/2}$). Such tails are said to be *heavier* than Gaussian. This tail behavior is also captured by the moment generating function of X. Indeed, we have

$$\mathbb{E}\big[e^{sX}\big] = \frac{1}{1-s^2} \quad \text{if } |s| < 1\,,$$

and is not defined for $s \geq 1$. It turns out that a rather week condition on the moment generating function is enough to partially reproduce some of the bounds that we have proved for sub-Gaussian random variables. Observe that for $X \sim \mathsf{Lap}(1)$

$$\mathbb{E}\big[e^{sX}\big] \leq e^{2s^2} \quad \text{if } |s| < 1/2\,,$$

In particular, the Laplace distribution has its moment generating distribution that is bounded by that of a Gaussian in a neighborhood of 0 but does not even exist away from zero. It turns out that all distributions that have tails at least as heavy as that of a Laplace distribution satisfy such a property.

Lemma 1.10. *Let X be a centered random variable such that $\mathbb{P}(|X| > t) \leq 2e^{-2t/\lambda}$ for some $\lambda > 0$. Then, for any positive integer $k \geq 1$,*

$$\mathbb{E}[|X|^k] \leq \lambda^k k!\,.$$

Moreover,

$$\big(\mathbb{E}[|X|^k]\big)^{1/k} \leq 2\lambda k\,,$$

and the moment generating function of X satisfies

$$\mathbb{E}\big[e^{sX}\big] \leq e^{2s^2\lambda^2}\,, \qquad \forall |s| \leq \frac{1}{2\lambda}\,.$$

Proof.

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}(|X| > t^{1/k})\mathrm{d}t \\
&\leq \int_0^\infty 2e^{-\frac{2t^{1/k}}{\lambda}}\mathrm{d}t \\
&= 2(\lambda/2)^k k \int_0^\infty e^{-u}u^{k-1}\mathrm{d}u\,, \qquad\qquad u = \frac{2t^{1/k}}{\lambda} \\
&\leq \lambda^k k\Gamma(k) = \lambda^k k!
\end{aligned}$$

The second statement follows from $\Gamma(k) \leq k^k$ and $k^{1/k} \leq e^{1/e} \leq 2$ for any $k \geq 1$. It yields

$$\left(\lambda^k k\Gamma(k)\right)^{1/k} \leq 2\lambda k\,.$$

To control the MGF of X, we use the Taylor expansion of the exponential function as follows. Observe that by the dominated convergence theorem, for any s such that $|s| \leq 1/2\lambda$

$$\begin{aligned}
\mathbb{E}\left[e^{sX}\right] &\leq 1 + \sum_{k=2}^{\infty} \frac{|s|^k \mathbb{E}[|X|^k]}{k!} \\
&\leq 1 + \sum_{k=2}^{\infty} (|s|\lambda)^k \\
&= 1 + s^2\lambda^2 \sum_{k=0}^{\infty} (|s|\lambda)^k \\
&\leq 1 + 2s^2\lambda^2 \qquad\qquad |s| \leq \frac{1}{2\lambda} \\
&\leq e^{2s^2\lambda^2}
\end{aligned}$$

□

This leads to the following definition

Definition 1.11. A random variable X is said to be sub-exponential with parameter λ (denoted $X \sim \mathsf{subE}(\lambda)$) if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2\lambda^2/2}\,, \qquad \forall |s| \leq \frac{1}{\lambda}\,.$$

A simple and useful example of of a sub-exponential random variable is given in the next lemma.

Lemma 1.12. *Let $X \sim \mathsf{subG}(\sigma^2)$ then the random variable $Z = X^2 - \mathbb{E}[X^2]$ is sub-exponential: $Z \sim \mathsf{subE}(16\sigma^2)$.*

Proof. We have, by the dominated convergence theorem,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &= 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}\big[X^2 - \mathbb{E}[X^2]\big]^k}{k!} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 2^{k-1}\big(\mathbb{E}[X^{2k}] + (\mathbb{E}[X^2])^k\big)}{k!} && \text{(Jensen)} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 4^k \mathbb{E}[X^{2k}]}{2(k!)} && \text{(Jensen again)} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 4^k 2(2\sigma^2)^k k!}{2(k!)} && \text{(Lemma 1.4)} \\
&= 1 + (8s\sigma^2)^2 \sum_{k=0}^{\infty} (8s\sigma^2)^k \\
&= 1 + 128 s^2\sigma^4 && \text{for} \quad |s| \le \frac{1}{16\sigma^2} \\
&\le e^{128 s^2\sigma^4} .
\end{aligned}$$

□

Sub-exponential random variables also give rise to exponential deviation inequalities such as Corollary 1.7 (Chernoff bound) or Theorem 1.9 (Hoeffding's inequality) for weighted sums of independent sub-exponential random variables. The significant difference here is that the larger deviations are controlled in by a weaker bound.

Berstein's inequality

Theorem 1.13 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables such that $\mathbb{E}(X_i) = 0$ and $X_i \sim \mathsf{subE}(\lambda)$. Define*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,,$$

Then for any $t > 0$ we have

$$\mathbb{P}(\bar{X} > t) \vee \mathbb{P}(\bar{X} < -t) \le \exp\left[-\frac{n}{2} \left(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda} \right) \right] .$$

Proof. Without loss of generality, assume that $\lambda = 1$ (we can always replace X_i by X_i/λ and t by t/λ. Next, using a Chernoff bound, we get for any $s > 0$

$$\mathbb{P}(\bar{X} > t) \le \prod_{i=1}^{n} \mathbb{E}\big[e^{sX_i}\big] e^{-snt} .$$

Next, if $|s| \leq 1$, then $\mathbb{E}\big[e^{sX_i}\big] \leq e^{s^2/2}$ by definition of sub-exponential distributions. It yields

$$\mathbb{P}(\bar{X} > t) \leq e^{\frac{ns^2}{2} - snt}$$

Choosing $s = 1 \wedge t$ yields

$$\mathbb{P}(\bar{X} > t) \leq e^{-\frac{n}{2}(t^2 \wedge t)}$$

We obtain the same bound for $\mathbb{P}(\bar{X} < -t)$ which concludes the proof. □

Note that usually, Bernstein's inequality refers to a slightly more precise result that is qualitatively the same as the one above: it exhibits a Gaussian tail $e^{-nt^2/(2\lambda^2)}$ and an exponential tail $e^{-nt/(2\lambda)}$. See for example Theorem 2.10 in [BLM13].

1.4 MAXIMAL INEQUALITIES

The exponential inequalities of the previous section are valid for linear combinations of independent random variables, and in particular, for the average $\bar{X}$. In many instances, we will be interested in controlling the *maximum* over the parameters of such linear combinations (this is because of empirical risk minimization). The purpose of this section is to present such results.

Maximum over a finite set

We begin by the simplest case possible: the maximum over a finite set.

Theorem 1.14. *Let $X_1, \ldots, X_N$ be N random variables such that $X_i \sim \mathsf{subG}(\sigma^2)$. Then*

$$\mathbb{E}[\max_{1 \leq i \leq N} X_i] \leq \sigma\sqrt{2\log(N)}\,, \qquad \textit{and} \qquad \mathbb{E}[\max_{1 \leq i \leq N} |X_i|] \leq \sigma\sqrt{2\log(2N)}$$

Moreover, for any $t > 0$,

$$\mathbb{P}\big(\max_{1 \leq i \leq N} X_i > t\big) \leq Ne^{-\frac{t^2}{2\sigma^2}}\,, \qquad \textit{and} \qquad \mathbb{P}\big(\max_{1 \leq i \leq N} |X_i| > t\big) \leq 2Ne^{-\frac{t^2}{2\sigma^2}}$$

Note that the random variables in this theorem need not be independent.

Proof. For any $s > 0$,

$$\begin{aligned}
\mathbb{E}[\max_{1\le i\le N} X_i] &= \frac{1}{s}\mathbb{E}\big[\log e^{s\max_{1\le i\le N} X_i}\big] \\
&\le \frac{1}{s}\log \mathbb{E}\big[e^{s\max_{1\le i\le N} X_i}\big] \qquad \text{(by Jensen)} \\
&= \frac{1}{s}\log \mathbb{E}\big[\max_{1\le i\le N} e^{sX_i}\big] \\
&\le \frac{1}{s}\log \sum_{1\le i\le N} \mathbb{E}\big[e^{sX_i}\big] \\
&\le \frac{1}{s}\log \sum_{1\le i\le N} e^{\frac{\sigma^2 s^2}{2}} \\
&= \frac{\log N}{s} + \frac{\sigma^2 s}{2}
\end{aligned}$$

Taking $s = \sqrt{2(\log N)/\sigma^2}$ yields the first inequality in expectation.

The first inequality in probability is obtained by a simple union bound:

$$\begin{aligned}
\mathbb{P}\big(\max_{1\le i\le N} X_i > t\big) &= \mathbb{P}\Big(\bigcup_{1\le i\le N} \{X_i > t\}\Big) \\
&\le \sum_{1\le i\le N} \mathbb{P}(X_i > t) \\
&\le N e^{-\frac{t^2}{2\sigma^2}}\,,
\end{aligned}$$

where we used Lemma 1.3 in the last inequality.

The remaining two inequalities follow trivially by noting that

$$\max_{1\le i\le N} |X_i| = \max_{1\le i\le 2N} X_i\,,$$

where $X_{N+i} = -X_i$ for $i = 1, \dots, N$. □

Extending these results to a maximum over an infinite set may be impossible. For example, if one is given an infinite sequence of i.i.d $\mathcal{N}(0, \sigma^2)$ random variables $X_1, X_2, \dots,$, then for any $N \ge 1$, we have for any $t > 0$,

$$\mathbb{P}(\max_{1\le i\le N} X_i < t) = [\mathbb{P}(X_1 < t)]^N \to 0\,, \quad N \to \infty\,.$$

On the opposite side of the picture, if all the X_is are equal to the same random variable X, we have for any $t > 0$,

$$\mathbb{P}(\max_{1\le i\le N} X_i < t) = \mathbb{P}(X_1 < t) > 0 \quad \forall\, N \ge 1\,.$$

In the Gaussian case, lower bounds are also available. They illustrate the effect of the correlation between the X_is

Examples from statistics have structure and we encounter many examples where a maximum of random variables over an infinite set is in fact finite. This is due to the fact that the random variable that we are considering are not independent from each other. In the rest of this section, we review some of these examples.

Maximum over a convex polytope

We use the definition of a polytope from [Gru03]: a convex polytope P is a compact set with a finite number of vertices $\mathcal{V}(\mathsf{P})$ called extreme points. It satisfies $\mathsf{P} = \mathrm{conv}(\mathcal{V}(\mathsf{P}))$, where $\mathrm{conv}(\mathcal{V}(\mathsf{P}))$ denotes the convex hull of the vertices of P.

Let $X \in \mathbb{R}^d$ be a random vector and consider the (infinite) family of random variables

$$\mathcal{F} = \{\theta^\top X \,:\, \theta \in \mathsf{P}\}\,,$$

where $\mathsf{P} \subset \mathbb{R}^d$ is a polytope with N vertices. While the family $\mathcal{F}$ is infinite, the maximum over $\mathcal{F}$ can be reduced to the a finite maximum using the following useful lemma.

Lemma 1.15. *Consider a linear form $x \mapsto c^\top x$, $x, c \in \mathbb{R}^d$. Then for any convex polytope $\mathsf{P} \subset \mathbb{R}^d$,*

$$\max_{x\in\mathsf{P}} c^\top x = \max_{x\in\mathcal{V}(\mathsf{P})} c^\top x$$

where $\mathcal{V}(\mathsf{P})$ denotes the set of vertices of P.

Proof. Assume that $\mathcal{V}(\mathsf{P}) = \{v_1, \dots, v_N\}$. For any $x \in \mathsf{P} = \mathrm{conv}(\mathcal{V}(\mathsf{P}))$, there exist nonnegative numbers $\lambda_1, \dots \lambda_N$ that sum up to 1 and such that $x = \lambda_1 v_1 + \cdots + \lambda_N v_N$. Thus

$$c^\top x = c^\top\Big(\sum_{i=1}^N \lambda_i v_i\Big) = \sum_{i=1}^N \lambda_i c^\top v_i \le \sum_{i=1}^N \lambda_i \max_{x\in\mathcal{V}(\mathsf{P})} c^\top x = \max_{x\in\mathcal{V}(\mathsf{P})} c^\top x\,.$$

It yields

$$\max_{x\in\mathsf{P}} c^\top x \le \max_{x\in\mathcal{V}(\mathsf{P})} c^\top x \le \max_{x\in\mathsf{P}} c^\top x$$

so the two quantities are equal. □

It immediately yields the following theorem

Theorem 1.16. *Let P be a polytope with N vertices $v^{(1)}, \dots, v^{(N)} \in \mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector such that, $[v^{(i)}]^\top X, i = 1, \dots, N$ are sub-Gaussian random variables with variance proxy σ^2. Then*

$$\mathbb{E}[\max_{\theta\in\mathsf{P}} \theta^\top X] \le \sigma\sqrt{2\log(N)}\,, \qquad \textit{and} \qquad \mathbb{E}[\max_{\theta\in\mathsf{P}} |\theta^\top X|] \le \sigma\sqrt{2\log(2N)}\,.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\big(\max_{\theta\in\mathsf{P}} \theta^\top X > t\big) \le Ne^{-\frac{t^2}{2\sigma^2}}\,, \qquad \textit{and} \qquad \mathbb{P}\big(\max_{\theta\in\mathsf{P}} |\theta^\top X| > t\big) \le 2Ne^{-\frac{t^2}{2\sigma^2}}$$

Of particular interests are polytopes that have a small number of vertices. A primary example is the ℓ_1 ball of $\mathbb{R}^d$ defined for any radius $R > 0$, by

$$\mathcal{B}_1 = \left\{ x \in \mathbb{R}^d \, : \, \sum_{i=1}^{d} |x_i| \leq 1 \right\}.$$

Indeed, it has exactly $2d$ vertices.

Maximum over the ℓ_2 ball

Recall that the unit ℓ_2 ball of $\mathbb{R}^d$ is defined by the set of vectors u that have Euclidean norm $|u|_2$ at most 1. Formally, it is defined by

$$\mathcal{B}_2 = \left\{ x \in \mathbb{R}^d \, : \, \sum_{i=1}^{d} x_i^2 \leq 1 \right\}.$$

Clearly, this ball is not a polytope and yet, we can control the maximum of random variables indexed by $\mathcal{B}_2$. This is due to the fact that there exists a finite subset of $\mathcal{B}_2$ such that the maximum over this finite set is of the same order as the maximum over the entire ball.

Definition 1.17. Fix $K \subset \mathbb{R}^d$ and $\varepsilon > 0$. A set $\mathcal{N}$ is called an ε-net of K with respect to a distance $d(\cdot, \cdot)$ on $\mathbb{R}^d$, if $\mathcal{N} \subset K$ and for any $z \in K$, there exists $x \in \mathcal{N}$ such that $d(x, z) \leq \varepsilon$.

Therefore, if $\mathcal{N}$ is an ε-net of K with respect to norm $\|\cdot\|$, then every point of K is at distance at most ε from a point in $\mathcal{N}$. Clearly, every compact set admits a finite ε-net. The following lemma gives an upper bound on the size of the smallest ε-net of $\mathcal{B}_2$.

Lemma 1.18. *Fix $\varepsilon \in (0, 1)$. Then the unit Euclidean ball $\mathcal{B}_2$ has an ε-net $\mathcal{N}$ with respect to the Euclidean distance of cardinality $|\mathcal{N}| \leq (3/\varepsilon)^d$*

Proof. Consider the following iterative construction if the ε-net. Choose $x_1 = 0$. For any $i \geq 2$, take any x_i to be any $x \in \mathcal{B}_2$ such that $|x - x_j|_2 > \varepsilon$ for all $j < i$. If no such x exists, stop the procedure. Clearly, this will create an ε-net. We now control its size.

Observe that since $|x - y|_2 > \varepsilon$ for all $x, y \in \mathcal{N}$, the Euclidean balls centered at $x \in \mathcal{N}$ and with radius $\varepsilon/2$ are disjoint. Moreover,

$$\bigcup_{z \in \mathcal{N}} \{ z + \frac{\varepsilon}{2} \mathcal{B}_2 \} \subset (1 + \frac{\varepsilon}{2}) \mathcal{B}_2$$

where $\{z + \varepsilon \mathcal{B}_2\} = \{z + \varepsilon x \, , x \in \mathcal{B}_2\}$. Thus, measuring volumes, we get

$$\mathrm{vol}\left((1 + \frac{\varepsilon}{2}) \mathcal{B}_2 \right) \geq \mathrm{vol}\Big(\bigcup_{z \in \mathcal{N}} \{ z + \frac{\varepsilon}{2} \mathcal{B}_2 \} \Big) = \sum_{z \in \mathcal{N}} \mathrm{vol}\left(\{ z + \frac{\varepsilon}{2} \mathcal{B}_2 \} \right)$$

This is equivalent to

$$(1+\frac{\varepsilon}{2})^d \geq |\mathcal{N}|(\frac{\varepsilon}{2})^d\,.$$

Therefore, we get the following bound

$$|\mathcal{N}| \leq \left(1+\frac{2}{\varepsilon}\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d.$$

□

Theorem 1.19. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. Then*

$$\mathbb{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] = \mathbb{E}[\max_{\theta\in\mathcal{B}_2} |\theta^\top X|] \leq 4\sigma\sqrt{d}\,.$$

Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X = \max_{\theta\in\mathcal{B}_2} |\theta^\top X| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}\,.$$

Proof. Let $\mathcal{N}$ be a $1/2$-net of $\mathcal{B}_2$ with respect to the Euclidean norm that satisfies $|\mathcal{N}| \leq 6^d$. Next, observe that for every $\theta \in \mathcal{B}_2$, there exists $z \in \mathcal{N}$ and x such that $|x|_2 \leq 1/2$ and $\theta = z + x$. Therefore,

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X \leq \max_{z\in\mathcal{N}} z^\top X + \max_{x\in\frac{1}{2}\mathcal{B}_2} x^\top X$$

But

$$\max_{x\in\frac{1}{2}\mathcal{B}_2} x^\top X = \frac{1}{2}\max_{x\in\mathcal{B}_2} x^\top X$$

Therefore, using Theorem 1.14, we get

$$\mathbb{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] \leq 2\mathbb{E}[\max_{z\in\mathcal{N}} z^\top X] \leq 2\sigma\sqrt{2\log(|\mathcal{N}|)} \leq 2\sigma\sqrt{2(\log 6)d} \leq 4\sigma\sqrt{d}\,.$$

The bound with high probability, follows because

$$\mathbb{P}\big(\max_{\theta\in\mathcal{B}_2} \theta^\top X > t\big) \leq \mathbb{P}\big(2\max_{z\in\mathcal{N}} z^\top X > t\big) \leq |\mathcal{N}| e^{-\frac{t^2}{8\sigma^2}} \leq 6^d e^{-\frac{t^2}{8\sigma^2}}\,.$$

To conclude the proof, we find t such that

$$e^{-\frac{t^2}{8\sigma^2}+d\log(6)} \leq \delta \;\Leftrightarrow\; t^2 \geq 8\log(6)\sigma^2 d + 8\sigma^2\log(1/\delta)\,.$$

Therefore, it is sufficient to take $t = \sqrt{8\log(6)}\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}\,.$ □

1.5 PROBLEM SET

Problem 1.1. Let $X_1, \dots, X_n$ be independent random variables such that $\mathbb{E}(X_i) = 0$ and $X_i \sim \mathsf{subE}(\lambda)$. For any vector $a = (a_1, \dots, a_n)^\top \in \mathbb{R}^n$, define the weighted sum

$$S(a) = \sum_{i=1}^{n} a_i X_i \,,$$

Show that for any $t > 0$ we have

$$\mathbb{P}(|S(a)| > t) \leq 2 \exp\left[-C\big(\frac{t^2}{\lambda^2 |a|_2^2} \wedge \frac{t}{\lambda |a|_\infty}\big)\right].$$

for some positive constant C.

Problem 1.2. A random variable X has χ^2_n (chi-squared with n degrees of freedom) if it has the same distribution as $Z_1^2 + \ldots + Z_n^2$, where $Z_1, \dots, Z_n$ are iid $\mathcal{N}(0,1)$.

(a) Let $Z \sim \mathcal{N}(0,1)$. Show that the moment generating function of $Y = Z^2 - 1$ satisfies

$$\phi(s) := E\big[e^{sY}\big] = \begin{cases} \dfrac{e^{-s}}{\sqrt{1-2s}} & \text{if } s < 1/2 \\ \infty & \text{otherwise} \end{cases}$$

(b) Show that for all $0 < s < 1/2$,

$$\phi(s) \leq \exp\Big(\frac{s^2}{1-2s}\Big)\,.$$

(c) Conclude that

$$\mathbb{P}(Y > 2t + 2\sqrt{t}) \leq e^{-t}$$

[Hint: you can use the convexity inequality $\sqrt{1+u} \leq 1 + u/2$].

(d) Show that if $X \sim \chi^2_n$, then, with probability at least $1 - \delta$, it holds

$$X \leq n + 2\sqrt{n \log(1/\delta)} + 2\log(1/\delta)\,.$$

Problem 1.3. Let $X_1, X_2 \dots$ be an infinite sequence of sub-Gaussian random variables with variance proxy $\sigma_i^2 = C(\log i)^{-1/2}$. Show that for C large enough, we get

$$\mathbb{E}\big[\max_{i \geq 2} X_i\big] < \infty\,.$$

Problem 1.4. Let $A = \{A_{i,j}\}_{\substack{1\le i\le n\\ 1\le j\le m}}$ be a random matrix such that its entries are iid sub-Gaussian random variables with variance proxy σ^2.

(a) Show that the matrix A is sub-Gaussian. What is its variance proxy?

(b) Let $\|A\|$ denote the operator norm of A defined by

$$\max_{x\in\mathbb{R}^m} \frac{|Ax|_2}{|x|_2}\,.$$

Show that there exits a constant $C > 0$ such that

$$\mathbb{E}\|A\| \le C(\sqrt{m} + \sqrt{n})\,.$$

Problem 1.5. Recall that for any $q \ge 1$, the ℓ_q norm of a vector $x \in \mathbb{R}^n$ is defined by

$$|x|_q = \Big(\sum_{i=1}^n |x_i|^q\Big)^{\frac{1}{q}}\,.$$

Let $X = (X_1, \dots, X_n)$ be a vector with independent entries such that X_i is sub-Gaussian with variance proxy σ^2 and $\mathbb{E}(X_i) = 0$.

(a) Show that for any $q \ge 2$, and any $x \in \mathbb{R}^d$,

$$|x|_2 \le |x|_q n^{\frac{1}{2}-\frac{1}{q}}\,,$$

and prove that the above inequality cannot be improved

(b) Show that for for any $q > 1$,

$$\mathbb{E}|X|_q \le 4\sigma n^{\frac{1}{q}}\sqrt{q}$$

(c) Recover from this bound that

$$\mathbb{E}\max_{1\le i\le n}|X_i| \le 4e\sigma\sqrt{\log n}\,.$$

Problem 1.6. Let K be a compact subset of the unit sphere of $\mathbb{R}^p$ that admits an ε-net $\mathcal{N}_\varepsilon$ with respect to the Euclidean distance of $\mathbb{R}^p$ that satisfies $|\mathcal{N}_\varepsilon| \le (C/\varepsilon)^d$ for all $\varepsilon \in (0,1)$. Here $C \ge 1$ and $d \le p$ are positive constants. Let $X \sim \mathsf{subG}_p(\sigma^2)$ be a centered random vector.

Show that there exists positive constants c_1 and c_2 to be made explicit such that for any $\delta \in (0,1)$, it holds

$$\max_{\theta\in K} \theta^\top X \le c_1\sigma\sqrt{d\log(2p/d)} + c_2\sigma\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$. Comment on the result in light of Theorem 1.19 .

Problem 1.7. For any $K \subset \mathbb{R}^d$, distance d on $\mathbb{R}^d$ and $\varepsilon > 0$, the ε-covering number $C(\varepsilon)$ of K is the cardinality of the smallest ε-net of K. The ε-packing number $P(\varepsilon)$ of K is the cardinality of the largest set $\mathcal{P} \subset K$ such that $d(z, z') > \varepsilon$ for all $z, z' \in \mathcal{P}$, $z \neq z'$. Show that

$$C(2\varepsilon) \leq P(2\varepsilon) \leq C(\varepsilon)\,.$$

Problem 1.8. Let $X_1, \ldots, X_n$ be n independent and random variables such that $\mathbb{E}[X_i] = \mu$ and $\mathrm{var}(X_i) \leq \sigma^2$. Fix $\delta \in (0, 1)$ and assume without loss of generality that n can be factored into $n = K \cdot G$ where $G = 8\log(1/\delta)$ is a positive integers.

For $g = 1, \ldots, G$, let $\bar{X}_g$ denote the average over the gth group of k variables. Formally

$$\bar{X}_g = \frac{1}{k} \sum_{i=(g-1)k+1}^{gk} X_i\,.$$

1. Show that for any $g = 1, \ldots, G$,

$$\mathbb{P}\big[\bar{X}_g - \mu > \frac{2\sigma}{\sqrt{k}}\big] \leq \frac{1}{4}\,.$$

2. Let $\hat{\mu}$ be defined as the median of $\{\bar{X}_1, \ldots, \bar{X}_G\}$. Show that

$$\mathbb{P}\big[\hat{\mu} - \mu > \frac{2\sigma}{\sqrt{k}}\big] \leq \mathbb{P}\big[\mathcal{B} \geq \frac{G}{2}\big]\,,$$

   where $\mathcal{B} \sim \mathsf{Bin}(G, 1/4)$.

3. Conclude that

$$\mathbb{P}\big[\hat{\mu} - \mu > 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\big] \leq \delta$$

4. Compare this result with 1.7 and Lemma 1.3. Can you conclude that $\hat{\mu} - \mu \sim \mathsf{subG}(\bar{\sigma}^2/n)$ for some $\bar{\sigma}^2$? Conclude.

CHAPTER 2

Linear Regression Model

In this chapter, we consider the following regression model:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n\,, \tag{2.1}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is sub-Gaussian with variance proxy σ^2 and such that $\mathbb{E}[\varepsilon] = 0$. Our goal is to estimate the function f under a linear assumption. Namely, we assume that $x \in \mathbb{R}^d$ and $f(x) = x^\top \theta^*$ for some unknown $\theta^* \in \mathbb{R}^d$.

2.1 FIXED DESIGN LINEAR REGRESSION

Depending on the nature of the *design* points $X_1, \dots, X_n$, we will favor a different measure of risk. In particular, we will focus either on *fixed* or *random* design.

Random design

The case of random design corresponds to the statistical learning setup. Let $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ be $n+1$ i.i.d. random couples. Given $(X_1, Y_1), \dots, (X_n, Y_n)$ the goal is construct a function $\hat{f}_n$ such that $\hat{f}_n(X_{n+1})$ is a good predictor of Y_{n+1}. Note that when $\hat{f}_n$ is constructed, X_{n+1} is still unknown and we have to account for what value it is likely to take.

Consider the following example from [HTF01, Section 3.2]. The response variable Y is the log-volume of a cancerous tumor, and the goal is to predict it based on $X \in \mathbb{R}^6$, a collection of variables that are easier to measure (age of patient, log-weight of prostate, ...). Here the goal is clearly to construct f for *prediction* purposes. Indeed, we want to find an automatic mechanism that

outputs a good prediction of the log-weight of the tumor given certain inputs for a new (unseen) patient.

A natural measure of performance here is the L_2-risk employed in the introduction:

$$R(\hat{f}_n) = \mathbb{E}[Y_{n+1} - \hat{f}_n(X_{n+1})]^2 = \mathbb{E}[Y_{n+1} - f(X_{n+1})]^2 + \|\hat{f}_n - f\|^2_{L^2(P_X)}\,,$$

where P_X denotes the marginal distribution of X_{n+1}. It measures how good the prediction of Y_{n+1} is in average over realizations of X_{n+1}. In particular, it does not put much emphasis on values of X_{n+1} that are not very likely to occur.

Note that if the ε_i are random variables with variance σ^2 then, one simply has $R(\hat{f}_n) = \sigma^2 + \|\hat{f}_n - f\|^2_{L^2(P_X)}$. Therefore, for random design, we will focus on the squared L_2 norm $\|\hat{f}_n - f\|^2_{L^2(P_X)}$ as a measure of accuracy. It measures how close $\hat{f}_n$ is to the unknown f *in average* over realizations of X_{n+1}.

Fixed design

In fixed design, the points (or vectors) $X_1, \dots, X_n$ are *deterministic*. To emphasize this fact, we use lowercase letters $x_1, \dots, x_n$ to denote fixed design. Of course, we can always think of them as realizations of a random variable but the distinction between fixed and random design is deeper and significantly affects our measure of performance. Indeed, recall that for random design, we look at the performance *in average* over realizations of X_{n+1}. Here, there is no such thing as a marginal distribution of X_{n+1}. Rather, since the design points $x_1, \dots, x_n$ are considered deterministic, our goal is estimate f *only* at these points. This problem is sometimes called *denoising* since our goal is to recover $f(x_1), \dots, f(x_n)$ given noisy observations of these values.

In many instances, fixed design can be recognized from their structured form. A typical example is the *regular design* on $[0, 1]$, given by $x_i = i/n, i = 1, \dots, n$. Interpolation between these points is possible under smoothness assumptions.

Note that in fixed design, we observe $\mu^* + \varepsilon$, where $\mu^* = \big(f(x_1), \dots, f(x_n)\big)^\top \in \mathbb{R}^n$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon)^\top \in \mathbb{R}^n$ is sub-Gaussian with variance proxy σ^2. Instead of a functional estimation problem, it is often simpler to view this problem as a vector problem in $\mathbb{R}^n$. This point of view will allow us to leverage the Euclidean geometry of $\mathbb{R}^n$.

In the case of fixed design, we will focus on the *Mean Squared Error* (MSE) as a measure of performance. It is defined by

$$\mathsf{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} \big(\hat{f}_n(x_i) - f(x_i)\big)^2\,.$$

Equivalently, if we view our problem as a vector problem, it is defined by

$$\mathsf{MSE}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \big(\hat{\mu}_i - \mu^*_i\big)^2 = \frac{1}{n}|\hat{\mu} - \mu^*|_2^2\,.$$

Often, the design vectors $x_1, \dots, x_n \in \mathbb{R}^d$ are stored in a $n \times d$ design matrix $\mathbb{X}$, whose jth row is given by $x_j^\top$. With this notation, the linear regression model can be written

$$Y = \mathbb{X}\theta^* + \varepsilon\,, \tag{2.2}$$

where $Y = (Y_1, \dots, Y_n)^\top$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. Moreover,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}|\mathbb{X}(\hat{\theta} - \theta^*)|_2^2 = (\hat{\theta} - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat{\theta} - \theta^*)\,. \tag{2.3}$$

A natural example of fixed design regression is image denoising. Assume that $\mu_i^*, i \in 1, \dots, n$ is the grayscale value of pixel i of an image. We do not get to observe the image μ^* but rather a noisy version of it $Y = \mu^* + \varepsilon$. Given a library of d images $\{x_1, \dots, x_d\}, x_j \in \mathbb{R}^n$, our goal is to recover the original image μ^* using linear combinations of the images $x_1, \dots, x_d$. This can be done fairly accurately (see Figure 2.1).

Figure 2.1. Reconstruction of the digit "6": Original (left), Noisy (middle) and Reconstruction (right). Here $n = 16 \times 16 = 256$ pixels. Source [RT11].

As we will see in Remark 2.3, choosing fixed design properly also ensures that if $\mathsf{MSE}(\hat{f})$ is small for some linear estimator $\hat{f}(x) = x^\top \hat{\theta}$, then $|\hat{\theta} - \theta^*|_2^2$ is also small.

In this chapter we only consider the fixed design case.

2.2 LEAST SQUARES ESTIMATORS

Throughout this section, we consider the regression model (2.2) with fixed design.

Unconstrained least squares estimator

Define the (unconstrained) *least squares estimator* $\hat{\theta}^{\text{LS}}$ to be any vector such that

$$\hat{\theta}^{\text{LS}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} |Y - \mathbb{X}\theta|_2^2\,.$$

Note that we are interested in estimating $\mathbb{X}\theta^*$ and not θ^* itself, so by extension, we also call $\hat{\mu}^{\text{LS}} = \mathbb{X}\hat{\theta}^{\text{LS}}$ least squares estimator. Observe that $\hat{\mu}^{\text{LS}}$ is the projection of Y onto the column span of $\mathbb{X}$.

It is not hard to see that least squares estimators of θ^* and $\mu^* = \mathbb{X}\theta^*$ are maximum likelihood estimators when $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Proposition 2.1. The least squares estimator $\hat{\mu}^{\text{LS}} = \mathbb{X}\hat{\theta}^{\text{LS}} \in \mathbb{R}^n$ satisfies

$$\mathbb{X}^\top \hat{\mu}^{\text{LS}} = \mathbb{X}^\top Y\,.$$

Moreover, $\hat{\theta}^{\text{LS}}$ can be chosen to be

$$\hat{\theta}^{\text{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y\,,$$

where $(\mathbb{X}^\top \mathbb{X})^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbb{X}^\top \mathbb{X}$.

Proof. The function $\theta \mapsto |Y - \mathbb{X}\theta|_2^2$ is convex so any of its minima satisfies

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = 0$$

Where ∇_θ is the gradient operator. Using matrix calculus, we find

$$\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = \nabla_\theta \big\{ |Y|_2^2 + -2Y^\top \mathbb{X}\theta + \theta^\top \mathbb{X}^\top \mathbb{X}\theta \big\} = -2(Y^\top \mathbb{X} - \theta^\top \mathbb{X}^\top \mathbb{X})^\top\,.$$

Therefore, solving $\nabla_\theta |Y - \mathbb{X}\theta|_2^2 = 0$ yields

$$\mathbb{X}^\top \mathbb{X}\theta = \mathbb{X}^\top Y\,.$$

It concludes the proof of the first statement. The second statement follows from the definition of the Moore-Penrose pseudoinverse. □

We are now going to prove our first result on the finite sample performance of the least squares estimator for fixed design.

Theorem 2.2. *Assume that the linear model (2.2) holds where* $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. *Then the least squares estimator* $\hat{\theta}^{\text{LS}}$ *satisfies*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})\big] = \frac{1}{n}\mathbb{E}|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma^2 \frac{r}{n}\,,$$

where $r = \operatorname{rank}(\mathbb{X}^\top \mathbb{X})$. *Moreover, for any* $\delta > 0$, *with probability* $1 - \delta$, *it holds*

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}) \lesssim \sigma^2 \frac{r + \log(1/\delta)}{n}\,.$$

Proof. Note that by definition

$$|Y - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 \le |Y - \mathbb{X}\theta^*|_2^2 = |\varepsilon|_2^2\,. \tag{2.4}$$

Moreover,

$$|Y - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 = |\mathbb{X}\theta^* + \varepsilon - \mathbb{X}\hat{\theta}^{\text{LS}}|_2^2 = |\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 - 2\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) + |\varepsilon|_2^2\,.$$

Therefore, we get

$$|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \leq 2\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) = 2|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2 \frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2} \tag{2.5}$$

Note that it is difficult to control

$$\frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2}$$

as $\hat{\theta}^{\text{LS}}$ depends on ε and the dependence structure of this term may be complicated. To remove this dependency, a traditional technique is "sup-out" $\hat{\theta}^{\text{LS}}$. This is typically where maximal inequalities are needed. Here we have to be a bit careful.

Let $\Phi = [\phi_1, \dots, \phi_r] \in \mathbb{R}^{n \times r}$ be an orthonormal basis of the column span of $\mathbb{X}$. In particular, there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*) = \Phi\nu$. It yields

$$\frac{\varepsilon^\top \mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)}{|\mathbb{X}(\hat{\theta}^{\text{LS}} - \theta^*)|_2} = \frac{\varepsilon^\top \Phi\nu}{|\Phi\nu|_2} = \frac{\varepsilon^\top \Phi\nu}{|\nu|_2} = \tilde{\varepsilon}^\top \frac{\nu}{|\nu|_2} \leq \sup_{u \in \mathcal{B}_2} \tilde{\varepsilon}^\top u\,,$$

where $\mathcal{B}_2$ is the unit ball of $\mathbb{R}^r$ and $\tilde{\varepsilon} = \Phi^\top \varepsilon$. Thus

$$|\mathbb{X}\hat{\theta}^{\text{LS}} - \mathbb{X}\theta^*|_2^2 \leq 4 \sup_{u \in \mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2\,,$$

Next, note that for any $u \in \mathcal{S}^{r-1}$, it holds $|\Phi u|_2^2 = u^\top \Phi^\top \Phi u = u^\top u = 1$ so that for any $s \in \mathbb{R}$, we have

$$\mathbb{E}[e^{s\tilde{\varepsilon}^\top u}] = \mathbb{E}[e^{s\varepsilon^\top \Phi u}] \leq \mathrm{e}^{\frac{s^2\sigma^2}{2}}\,.$$

Therefore, $\tilde{\varepsilon} \sim \mathsf{subG}_r(\sigma^2)$.

To conclude the bound in expectation, observe that Lemma 1.4 yields

$$4\mathbb{E}\big[\sup_{u \in \mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2\big] = 4\sum_{i=1}^r \mathbb{E}[\tilde{\varepsilon}_i^2] \leq 16\sigma^2 r\,.$$

Moreover, with probability $1-\delta$, it follows from the last step in the proof[1] of Theorem 1.19 that

$$\sup_{u \in \mathcal{B}_2} (\tilde{\varepsilon}^\top u)^2 \leq 8\log(6)\sigma^2 r + 8\sigma^2 \log(1/\delta)\,.$$

□

Remark 2.3. If $d \leq n$ and $B := \frac{\mathbb{X}^\top \mathbb{X}}{n}$ has rank d, then we have

$$|\hat{\theta}^{\text{LS}} - \theta^*|_2^2 \leq \frac{\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}})}{\lambda_{\min}(B)}\,,$$

and we can use Theorem 2.2 to bound $|\hat{\theta}^{\text{LS}} - \theta^*|_2^2$ directly.

[1] we could use Theorem 1.19 directly here but at the cost of a factor 2 in the constant.

Constrained least squares estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know *a priori* that $\theta^* \in K$, we may prefer a *constrained least squares* estimator $\hat{\theta}^{\mathrm{LS}}_K$ defined by

$$\hat{\theta}^{\mathrm{LS}}_K \in \operatorname*{argmin}_{\theta \in K} |Y - \mathbb{X}\theta|_2^2 \,.$$

Indeed, the fundamental inequality (2.4) would still hold and the bounds on the MSE may be smaller. Indeed, (2.5) can be replaced by

$$|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 \le 2\varepsilon^\top \mathbb{X}(\hat{\theta}^{\mathrm{LS}}_K - \theta^*) \le 2 \sup_{\theta \in K-K} (\varepsilon^\top \mathbb{X}\theta) \,,$$

where $K - K = \{x - y \,:\, x, y \in K\}$. It is easy to see that if K is symmetric ($x \in K \Leftrightarrow -x \in K$) and convex, then $K - K = 2K$ so that

$$2 \sup_{\theta \in K-K} (\varepsilon^\top \mathbb{X}\theta) = 4 \sup_{v \in \mathbb{X}K} (\varepsilon^\top v)$$

where $\mathbb{X}K = \{\mathbb{X}\theta \,:\, \theta \in K\} \subset \mathbb{R}^n$. This is a measure of the size (width) of $\mathbb{X}K$. If $\varepsilon \sim \mathcal{N}(0, I_d)$, the expected value of the above supremum is actually called *Gaussian width* of $\mathbb{X}K$. Here, ε is not Gaussian but sub-Gaussian and similar properties will hold.

ℓ_1 constrained least squares

Assume here that $K = \mathcal{B}_1$ is the unit ℓ_1 ball of $\mathbb{R}^d$. Recall that it is defined by

$$\mathcal{B}_1 = \Big\{ x \in \mathbb{R}^d \,:\, \sum_{i=1}^d |x_i| \le 1 \Big\} \,,$$

and it has exactly $2d$ vertices $\mathcal{V} = \{e_1, -e_1, \dots, e_d, -e_d\}$, where e_j is the j-th vector of the canonical basis of $\mathbb{R}^d$ and is defined by

$$e_j = (0, \dots, 0, \underbrace{1}_{j\text{th position}}, 0, \dots, 0)^\top \,.$$

It implies that the set $\mathbb{X}K = \{\mathbb{X}\theta, \theta \in K\} \subset \mathbb{R}^n$ is also a polytope with at most $2d$ vertices that are in the set $\mathbb{X}\mathcal{V} = \{\mathbb{X}_1, -\mathbb{X}_1, \dots, \mathbb{X}_d, -\mathbb{X}_d\}$ where $\mathbb{X}_j$ is the j-th column of $\mathbb{X}$. Indeed, $\mathbb{X}K$ is a obtained by rescaling and embedding (resp. projecting) the polytope K when $d \le n$ (resp., $d \ge n$). Note that some columns of $\mathbb{X}$ might not be vertices of $\mathbb{X}K$ so that $\mathbb{X}\mathcal{V}$ might be a strict superset of the set of vertices of $\mathbb{X}K$.

Theorem 2.4. *Let $K = \mathcal{B}_1$ be the unit ℓ_1 ball of $\mathbb{R}^d, d \ge 2$ and assume that $\theta^* \in \mathcal{B}_1$. Moreover, assume the conditions of Theorem 2.2 and that the columns of $\mathbb{X}$ are normalized in such a way that $\max_j |\mathbb{X}_j|_2 \le \sqrt{n}$. Then the constrained least squares estimator $\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1}$ satisfies*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1})\big] = \frac{1}{n}\mathbb{E}|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1} - \mathbb{X}\theta^*|_2^2 \lesssim \sigma\sqrt{\frac{\log d}{n}} \,,$$

Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1}) \lesssim \sigma\sqrt{\frac{\log(d/\delta)}{n}}\,.$$

Proof. From the considerations preceding the theorem, we got that

$$|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1} - \mathbb{X}\theta^*|_2^2 \le 4 \sup_{v\in\mathbb{X}K} (\varepsilon^\top v)$$

Observe now that since $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$,then for any column $\mathbb{X}_j$ such that $|\mathbb{X}_j|_2 \le \sqrt{n}$, the random variable $\varepsilon^\top \mathbb{X}_j \sim \mathsf{subG}(n\sigma^2)$. Therefore, applying Theorem 1.16, we get the bound on $\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K)\big]$ and for any $t \ge 0$,

$$\mathbb{P}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K) > t\big] \le \mathbb{P}\big[\sup_{v\in\mathbb{X}K} (\varepsilon^\top v) > nt/4\big] \le 2de^{-\frac{nt^2}{32\sigma^2}}$$

To conclude the proof, we find t such that

$$2de^{-\frac{nt^2}{32\sigma^2}} \le \delta \;\Leftrightarrow\; t^2 \ge 32\sigma^2\frac{\log(2d)}{n} + 32\sigma^2\frac{\log(1/\delta)}{n}\,.$$

□

Note that the proof of Theorem 2.2 also applies to $\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1}$ (exercise!) so that $\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1}$ benefits from the best of both rates.

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_1})\big] \lesssim \min\Big(\frac{r}{n}, \sqrt{\frac{\log d}{n}}\Big)\,.$$

This is called an *elbow effect*. The elbow takes place around $r \simeq \sqrt{n}$ (up to logarithmic terms).

ℓ_0 constrained least squares

We abusively call ℓ_0 norm of a vector $\theta \in \mathbb{R}^d$ it number of non-zero coefficient. It is denoted by

$$|\theta|_0 = \sum_{j=1}^d \mathbb{1}(\theta_j \neq 0)\,.$$

We call a vector θ with "small" ℓ_0 norm a *sparse* vector. More precisely, if $|\theta|_0 \le k$, we say that θ is a k-sparse vector. We also call *support* of θ the set

$$\mathrm{supp}(\theta) = \big\{j \in \{1,\dots,d\} \,:\, \theta_j \neq 0\big\}$$

so that $|\theta|_0 = \mathrm{card}(\mathrm{supp}(\theta)) =: |\,\mathrm{supp}(\theta)|\,.$

Remark 2.5. The ℓ_0 terminology and notation comes from the fact that

$$\lim_{q\to 0^+} \sum_{j=1}^d |\theta_j|^q = |\theta|_0$$

Therefore it is really $\lim_{q\to 0^+} |\theta|_q^q$ but the notation $|\theta|_0^0$ suggests too much that it is always equal to 1.

By extension, denote by $\mathcal{B}_0(k)$ the ℓ_0 ball of $\mathbb{R}^d$, i.e., the set of k-sparse vectors, defined by

$$\mathcal{B}_0(k) = \{\theta \in \mathbb{R}^d \,:\, |\theta|_0 \le k\}\,.$$

In this section, our goal is to control the MSE of $\hat\theta^{\mathrm{LS}}_K$ when $K = \mathcal{B}_0(k)$. Note that computing $\hat\theta^{\mathrm{LS}}_{\mathcal{B}_0(k)}$ essentially requires computing $\binom{d}{k}$ least squares estimators, which is an exponential number in k. In practice this will be hard (or even impossible) but it is interesting to understand the statistical properties of this estimator and to use them as a benchmark.

Theorem 2.6. *Fix a positive integer $k \le d/2$. Let $K = \mathcal{B}_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$ and assume that $\theta^* \in \mathcal{B}_0(k)$. Moreover, assume the conditions of Theorem 2.2. Then, for any $\delta > 0$, with probability $1-\delta$, it holds*

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathrm{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2}{n}\log\binom{d}{2k} + \frac{\sigma^2 k}{n} + \frac{\sigma^2}{n}\log(1/\delta)\,.$$

Proof. We begin as in the proof of Theorem 2.2 to get (2.5):

$$|\mathbb{X}\hat\theta^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 \le 2\varepsilon^\top\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*) = 2|\mathbb{X}\hat\theta^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2\frac{\varepsilon^\top\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*)}{|\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*)|_2}\,.$$

We know that both $\hat\theta^{\mathrm{LS}}_K$ and θ^* are in $\mathcal{B}_0(k)$ so that $\hat\theta^{\mathrm{LS}}_K - \theta^* \in \mathcal{B}_0(2k)$. For any $S \subset \{1,\dots,d\}$, let $\mathbb{X}_S$ denote the $n \times |S|$ submatrix of $\mathbb{X}$ that is obtained from the columns of $\mathbb{X}_j, j \in S$ of $\mathbb{X}$. Denote by $r_S \le |S|$ the rank of $\mathbb{X}_S$ and let $\Phi_S = [\phi_1,\dots,\phi_{r_S}] \in \mathbb{R}^{n\times r_S}$ be an orthonormal basis of the column span of $\mathbb{X}_S$. Moreover, for any $\theta \in \mathbb{R}^d$, define $\theta(S) \in \mathbb{R}^{|S|}$ to be the vector with coordinates $\theta_j, j \in S$. If we denote by $\hat S = \mathrm{supp}(\hat\theta^{\mathrm{LS}}_K - \theta^*)$, we have $|\hat S| \le 2k$ and there exists $\nu \in \mathbb{R}^{r_{\hat S}}$ such that

$$\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*) = \mathbb{X}_{\hat S}(\hat\theta^{\mathrm{LS}}_K(\hat S) - \theta^*(\hat S)) = \Phi_{\hat S}\nu\,.$$

Therefore,

$$\frac{\varepsilon^\top\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*)}{|\mathbb{X}(\hat\theta^{\mathrm{LS}}_K - \theta^*)|_2} = \frac{\varepsilon^\top\Phi_{\hat S}\nu}{|\nu|_2} \le \max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}[\varepsilon^\top\Phi_S]u$$

where $\mathcal{B}_2^{r_S}$ is the unit ball of $\mathbb{R}^{r_S}$. It yields

$$|\mathbb{X}\hat\theta^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 \le 4\max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon_S^\top u)^2\,,$$

$\tilde\varepsilon_S = \Phi_S^\top\varepsilon \sim \mathsf{subG}_{r_S}(\sigma^2)$.

Using a union bound, we get for any $t > 0$,

$$\mathbb{P}\Big(\max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\Big) \le \sum_{|S|=2k}\mathbb{P}\Big(\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\Big)$$

It follows from the proof of Theorem 1.19 that for any $|S| \le 2k$,

$$\mathbb{P}\Big(\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde\varepsilon^\top u)^2 > t\Big) \le 6^{|S|}e^{-\frac{t}{8\sigma^2}} \le 6^{2k}e^{-\frac{t}{8\sigma^2}}\,.$$

Together, the above three displays yield

$$\mathbb{P}(|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 > 4t) \le \binom{d}{2k} 6^{2k} e^{-\frac{t}{8\sigma^2}} . \tag{2.6}$$

To ensure that the right-hand side of the above inequality is bounded by δ, we need

$$t \ge C\sigma^2 \Big\{ \log \binom{d}{2k} + k \log(6) + \log(1/\delta) \Big\} .$$

□

How large is $\log \binom{d}{2k}$? It turns out that it is not much larger than k.

Lemma 2.7. *For any integers $1 \le k \le n$, it holds*

$$\binom{n}{k} \le \Big(\frac{en}{k}\Big)^k$$

Proof. Observe first that if $k = 1$, since $n \ge 1$, it holds,

$$\binom{n}{1} = n \le en = \Big(\frac{en}{1}\Big)^1$$

Next, we proceed by induction and assume that it holds for some $k \le n - 1$.

$$\binom{n}{k} \le \Big(\frac{en}{k}\Big)^k$$

Observe that

$$\binom{n}{k+1} = \binom{n}{k} \frac{n-k}{k+1} \le \Big(\frac{en}{k}\Big)^k \frac{n-k}{k+1} = \frac{e^k n^{k+1}}{(k+1)^{k+1}} \Big(1 + \frac{1}{k}\Big)^k ,$$

where we used the induction hypothesis in the first inequality. To conclude, it suffices to observe that

$$\Big(1 + \frac{1}{k}\Big)^k \le e$$

□

It immediately leads to the following corollary:

Corollary 2.8. *Under the assumptions of Theorem 2.6, for any $\delta > 0$, with probability $1 - \delta$, it holds*

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2 k}{n} \log\Big(\frac{ed}{2k}\Big) + \frac{\sigma^2 k}{n} \log(6) + \frac{\sigma^2}{n} \log(1/\delta) .$$

Note that for any fixed δ, there exits a constant $C_\delta > 0$ such that for any $n \geq 2k$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_0(k)}) \leq C_\delta \frac{\sigma^2 k}{n} \log\left(\frac{ed}{2k}\right).$$

Comparing this result with Theorem 2.2 with $r = k$, we see that the price to pay for not knowing the support of θ^* but only its size, is a logarithmic factor in the dimension d.

This result immediately leads the following bound in expectation.

Corollary 2.9. *Under the assumptions of Theorem 2.6,*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_0(k)})\big] \lesssim \frac{\sigma^2 k}{n} \log\left(\frac{ed}{k}\right).$$

Proof. It follows from (2.6) that for any $H \geq 0$,

$$\begin{aligned}
\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}_{\mathcal{B}_0(k)})\big] &= \int_0^\infty \mathbb{P}(|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 > nu)\mathrm{d}u \\
&\leq H + \int_0^\infty \mathbb{P}(|\mathbb{X}\hat{\theta}^{\mathrm{LS}}_K - \mathbb{X}\theta^*|_2^2 > n(u+H))\mathrm{d}u \\
&\leq H + \sum_{j=1}^{2k} \binom{d}{j} 6^{2k} \int_0^\infty e^{-\frac{n(u+H)}{32\sigma^2}}, \\
&= H + \sum_{j=1}^{2k} \binom{d}{j} 6^{2k} e^{-\frac{nH}{32\sigma^2}} \frac{32\sigma^2}{n} \mathrm{d}u\,.
\end{aligned}$$

Next, take H to be such that

$$\sum_{j=1}^{2k} \binom{d}{j} 6^{2k} e^{-\frac{nH}{32\sigma^2}} = 1\,.$$

In particular, it yields

$$H \lesssim \frac{\sigma^2 k}{n} \log\left(\frac{ed}{k}\right),$$

which completes the proof. □

2.3 THE GAUSSIAN SEQUENCE MODEL

The Gaussian Sequence Model is a toy model that has received a lot of attention, mostly in the eighties. The main reason for its popularity is that it carries already most of the insight of nonparametric estimation. While the model looks very simple it allows to carry deep ideas that extend beyond its framework and in particular to the linear regression model that we are interested in. Unfortunately we will only cover a small part of these ideas and

the interested reader should definitely look at the excellent books by A. Tsybakov [Tsy09, Chapter 3] and I. Johnstone [Joh11]. The model is as follows:

$$Y_i = \theta_i^* + \varepsilon_i\,, \qquad i = 1, \dots, d \tag{2.7}$$

where $\varepsilon_1, \dots, \varepsilon_d$ are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables. Note that often, d is taken equal to ∞ in this sequence model and we will also discuss this case. Its links to nonparametric estimation will become clearer in Chapter 3. The goal here is to estimate the unknown vector θ^*.

The sub-Gaussian Sequence Model

Note first that the model (2.7) is a special case of the linear model with fixed design (2.1) with $n = d$, $f(x) = x^\top \theta^*$, $x_1, \dots, x_n$ form the canonical basis of $\mathbb{R}^n$ and ε has a Gaussian distribution. Therefore, $n = d$ is both the dimension of the parameter θ and the number of observation and it looks like we have chosen to index this problem by d rather than n somewhat arbitrarily. We can bring n back into the picture, by observing that this model encompasses slightly more general choices for the design matrix $\mathbb{X}$ as long as it satisfies the following assumption.

Assumption ORT The design matrix satisfies

$$\frac{\mathbb{X}^\top \mathbb{X}}{n} = I_d\,,$$

where I_d denotes the identity matrix of $\mathbb{R}^d$.

Assumption ORT allows for cases where $d \leq n$ but not $d > n$ (high dimensional case) because of obvious rank constraints. In particular, it means that the d columns of $\mathbb{X}$ are orthogonal in $\mathbb{R}^n$ and all have norm $\sqrt{n}$.

Under this assumption, it follows from the linear regression model (2.2) that

$$\begin{aligned} y := \frac{1}{n}\mathbb{X}^\top Y &= \frac{\mathbb{X}^\top \mathbb{X}}{n}\theta^* + \frac{1}{n}\mathbb{X}^\top \varepsilon \\ &= \theta^* + \xi\,, \end{aligned}$$

where $\xi = (\xi_1, \dots, \xi_d) \sim \mathsf{subG}_d(\sigma^2/n)$. As a result, under the assumption ORT, the linear regression model (2.2) is equivalent to the sub-Gaussian Sequence Model (2.7) up to a transformation of the data Y and a change of variable for the variance. Moreover, for any estimator $\hat{\theta} \in \mathbb{R}^d$, under ORT, it follows from (2.3) that

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat{\theta} - \theta^*) = |\hat{\theta} - \theta^*|_2^2\,.$$

Furthermore, for any $\theta \in \mathbb{R}^d$, the assumption ORT yields,

$$\begin{aligned} |y - \theta|_2^2 &= |\frac{1}{n}\mathbb{X}^\top Y - \theta|_2^2 \\ &= |\theta|_2^2 - \frac{2}{n}\theta^\top \mathbb{X}^\top Y + \frac{1}{n^2} Y^\top \mathbb{X}\mathbb{X}^\top Y \\ &= \frac{1}{n}|\mathbb{X}\theta|_2^2 - \frac{2}{n}(\mathbb{X}\theta)^\top Y + \frac{1}{n}|Y|_2^2 + Q \\ &= \frac{1}{n}|Y - \mathbb{X}\theta|_2^2 + Q\,, \end{aligned} \tag{2.8}$$

where Q is a constant that does not depend on θ and is defined by

$$Q = \frac{1}{n^2} Y^\top \mathbb{X}\mathbb{X}^\top Y - \frac{1}{n}|Y|_2^2$$

This implies in particular that the least squares estimator $\hat{\theta}^{\text{LS}}$ is equal to y.

We introduce a sightly more general model called *sub-Gaussian sequence model*:

$$y = \theta^* + \xi \quad \in \mathbb{R}^d \tag{2.9}$$

where $\xi \sim \mathsf{subG}_d(\sigma^2/n)$.

In this section, we can actually completely "forget" about our original model (2.2). In particular we can define this model independently of Assumption ORT and thus for any values of n and d.

The sub-Gaussian sequence model, like the Gaussian sequence model are called *direct* (observation) problems as opposed to *inverse problems* where the goal is to estimate the parameter θ^* only from noisy observations of its image through an operator. The linear regression model one such inverse problem where the matrix $\mathbb{X}$ plays the role of a linear operator. However, in these notes, we never try to invert the operator. See [Cav11] for an interesting survey on the statistical theory of inverse problems.

Sparsity adaptive thresholding estimators

If we knew a priori that θ was k sparse, we could employ directly Corollary 2.8 to obtain that with probability $1 - \delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)}) \leq C_\delta \frac{\sigma^2 k}{n} \log\left(\frac{ed}{2k}\right).$$

As we will see, the assumption ORT gives us the luxury to not know k and yet *adapt* to its value. Adaptation means that we can construct an estimator that does not require the knowledge of k (the smallest such that $|\theta^*|_0 \leq k$) and yet, perform as well as $\hat{\theta}^{\text{LS}}_{\mathcal{B}_0(k)}$, up to a multiplicative constant.

Let us begin with some heuristic considerations to gain some intuition. Assume the sub-Gaussian sequence model (2.9). If nothing is known about θ^*

it is natural to estimate it using the least squares estimator $\hat{\theta}^{\mathrm{LS}} = y$. In this case,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathrm{LS}}) = |y - \theta^*|_2^2 = |\xi|_2^2 \leq C_\delta \frac{\sigma^2 d}{n}\,,$$

where the last inequality holds with probability at least $1-\delta$. This is actually what we are looking for if $k = Cd$ for some positive constant $C \leq 1$. The problem with this approach is that it does not use the fact that k may be much smaller than d, which happens when θ^* has many zero coordinate.

If $\theta_j^* = 0$, then, $y_j = \xi_j$, which is a sub-Gaussian random variable with variance proxy σ^2/n. In particular, we know from Lemma 1.3 that with probability at least $1-\delta$,

$$|\xi_j| \leq \sigma\sqrt{\frac{2\log(2/\delta)}{n}} = \tau\,. \tag{2.10}$$

The consequences of this inequality are interesting. One the one hand, if we observe $|y_j| \gg \tau$, then it must correspond to $\theta_j^* \neq 0$. On the other hand, if $|y_j| \leq \tau$ is smaller, then, θ_j^* cannot be very large. In particular, by the triangle inequality, $|\theta_j^*| \leq |y_j| + |\xi_j| \leq 2\tau$. Therefore, we loose at most 2τ by choosing $\hat{\theta}_j = 0$. It leads us to consider the following estimator.

Definition 2.10. The **hard thresholding** estimator with threshold $2\tau > 0$ is denoted by $\hat{\theta}^{\mathrm{HRD}}$ and has coordinates

$$\hat{\theta}_j^{\mathrm{HRD}} = \begin{cases} y_j & \text{if } |y_j| > 2\tau\,, \\ 0 & \text{if } |y_j| \leq 2\tau\,, \end{cases}$$

for $j = 1, \ldots, d$. In short, we can write $\hat{\theta}_j^{\mathrm{HRD}} = y_j \mathbb{1}(|y_j| > 2\tau)$.

From our above consideration, we are tempted to choose τ as in (2.10). Yet, this threshold is not large enough. Indeed, we need to choose τ such that $|\xi_j| \leq \tau$ *simultaneously* for all j. This can be done using a maximal inequality. Namely, Theorem 1.14 ensures that with probability at least $1-\delta$,

$$\max_{1\leq j\leq d} |\xi_j| \leq \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$$

It yields the following theorem.

Theorem 2.11. *Consider the linear regression model (2.2) under the assumption* ORT *or, equivalenty, the sub-Gaussian sequence model (2.9). Then the hard thresholding estimator $\hat{\theta}^{\mathrm{HRD}}$ with threshold*

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d/\delta)}{n}}\,, \tag{2.11}$$

enjoys the following two properties on the same event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$:

(i) If $|\theta^*|_0 = k$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{HRD}}) = |\hat{\theta}^{\text{HRD}} - \theta^*|_2^2 \lesssim \sigma^2 \frac{k \log(2d/\delta)}{n} .$$

(ii) if $\min_{j \in \text{supp}(\theta^*)} |\theta_j^*| > 3\tau$, *then*

$$\text{supp}(\hat{\theta}^{\text{HRD}}) = \text{supp}(\theta^*) .$$

Proof. Define the event

$$\mathcal{A} = \Big\{ \max_j |\xi_j| \le \tau \Big\} ,$$

and recall that Theorem 1.14 yields $\mathbb{P}(\mathcal{A}) \ge 1 - \delta$. On the event $\mathcal{A}$, the following holds for any $j = 1, \dots, d$.

First, observe that

$$|y_j| > 2\tau \quad \Rightarrow \quad |\theta_j^*| \ge |y_j| - |\xi_j| > \tau \tag{2.12}$$

and

$$|y_j| \le 2\tau \quad \Rightarrow \quad |\theta_j^*| \le |y_j| + |\xi_j| \le 3\tau \tag{2.13}$$

It yields

$$\begin{aligned} |\hat{\theta}_j^{\text{HRD}} - \theta_j^*| &= |y_j - \theta_j^*| \mathbb{1}(|y_j| > 2\tau) + |\theta_j^*| \mathbb{1}(|y_j| \le 2\tau) \\ &\le \tau \mathbb{1}(|y_j| > 2\tau) + |\theta_j^*| \mathbb{1}(|y_j| \le 2\tau) \\ &\le \tau \mathbb{1}(|\theta_j^*| > \tau) + |\theta_j^*| \mathbb{1}(|\theta_j^*| \le 3\tau) \qquad \text{by (2.12) and (2.13)} \\ &\le 4 \min(|\theta_j^*|, \tau) \end{aligned}$$

It yields

$$|\hat{\theta}^{\text{HRD}} - \theta^*|_2^2 = \sum_{j=1}^d |\hat{\theta}_j^{\text{HRD}} - \theta_j^*|^2 \le 16 \sum_{j=1}^d \min(|\theta_j^*|^2, \tau^2) \le 16 |\theta^*|_0 \tau^2 .$$

This completes the proof of (i).

To prove (ii), note that if $\theta_j^* \neq 0$, then $|\theta_j^*| > 3\tau$ so that

$$|y_j| = |\theta_j^* + \xi_j| > 3\tau - \tau = 2\tau .$$

Therefore, $\hat{\theta}_j^{\text{HRD}} \neq 0$ so that $\text{supp}(\theta^*) \subset \text{supp}(\hat{\theta}^{\text{HRD}})$.

Next, if $\hat{\theta}_j^{\text{HRD}} \neq 0$, then $|\hat{\theta}_j^{\text{HRD}}| = |y_j| > 2\tau$. It yields

$$|\theta_j^*| \ge |y_j| - \tau > \tau$$

Therefore, $|\theta_j^*| \neq 0$ and $\text{supp}(\hat{\theta}^{\text{HRD}}) \subset \text{supp}(\theta^*)$. □

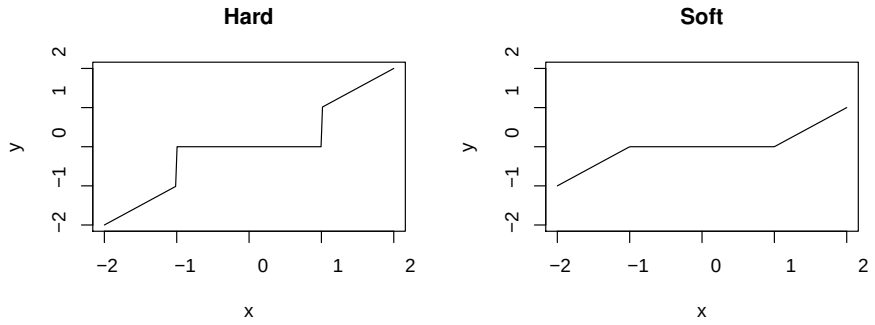


Figure 2.2. Transformation applied to y_j with $2\tau = 1$ to obtain the hard (left) and soft (right) thresholding estimators

Similar results can be obtained for the **soft thresholding** estimator $\hat{\theta}^{\text{SFT}}$ defined by

$$\hat{\theta}_j^{\text{SFT}} = \begin{cases} y_j - 2\tau & \text{if } y_j > 2\tau\,, \\ y_j + 2\tau & \text{if } y_j < -2\tau\,, \\ 0 & \text{if } |y_j| \leq 2\tau\,, \end{cases}$$

In short, we can write

$$\hat{\theta}_j^{\text{SFT}} = \Big(1 - \frac{2\tau}{|y_j|}\Big)_+ y_j$$

2.4 HIGH-DIMENSIONAL LINEAR REGRESSION

The BIC and Lasso estimators

It can be shown (see Problem 2.5) that the hard and soft thresholding estimators are solutions of the following penalized empirical risk minimization problems:

$$\hat{\theta}^{\text{HRD}} = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \Big\{ |y - \theta|_2^2 + 4\tau^2 |\theta|_0 \Big\}$$

$$\hat{\theta}^{\text{SFT}} = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \Big\{ |y - \theta|_2^2 + 4\tau |\theta|_1 \Big\}$$

In view of (2.8), under the assumption ORT, the above variational definitions can be written as

$$\hat{\theta}^{\mathrm{HRD}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + 4\tau^2 |\theta|_0 \right\}$$
$$\hat{\theta}^{\mathrm{SFT}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + 4\tau |\theta|_1 \right\}$$

When the assumption ORT is not satisfied, they no longer correspond to thresholding estimators but can still be defined as above. We change the constant in the threshold parameters for future convenience.

Definition 2.12. Fix $\tau > 0$ and assume the linear regression model (2.2). The BIC[2] estimator of θ^* in is defined by any $\hat{\theta}^{\mathrm{BIC}}$ such that

$$\hat{\theta}^{\mathrm{BIC}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + \tau^2 |\theta|_0 \right\}$$

Moreover the Lasso estimator of θ^* in is defined by any $\hat{\theta}^{\mathcal{L}}$ such that

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + 2\tau |\theta|_1 \right\}$$

Remark 2.13. Numerical considerations. Computing the BIC estimator can be proved to be NP-hard in the worst case. In particular, no computational method is known to be significantly faster than the brute force search among all 2^d sparsity patterns. Indeed, we can rewrite:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + \tau^2 |\theta|_0 \right\} = \min_{0 \le k \le d} \left\{ \min_{\theta \,:\, |\theta|_0 = k} \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + \tau^2 k \right\}$$

To compute $\min_{\theta \,:\, |\theta|_0 = k} \frac{1}{n} |Y - \mathbb{X}\theta|_2^2$, we need to compute $\binom{d}{k}$ least squares estimators on a space of size k. Each costs $O(k^3)$ (matrix inversion). Therefore the total cost of the brute force search is

$$C \sum_{k=0}^{d} \binom{d}{k} k^3 = C d^3 2^d \,.$$

Instead the the Lasso estimator is convex problem and there exists many efficient algorithms to compute it. We will not describe this optimization problem in details but only highlight a few of the best known algorithms:

1. Probably the most popular method among statisticians relies on coordinate gradient descent. It is implemented in the glmnet package in R [FHT10],

[2] Note that it minimizes the Bayes Information Criterion (BIC) employed in the traditional literature of asymptotic statistics if $\tau = \sqrt{\log(d)/n}$. We will use the same value below, up to multiplicative constants (it's the price to pay to get non asymptotic results).

2. An interesting method called LARS [EHJT04] computes the entire *regularization path*, i.e., the solution of the convex problem for all values of τ. It relies on the fact that, as a function of τ, the solution $\hat{\theta}^{\mathcal{L}}$ is a piecewise linear function (with values in $\mathbb{R}^d$). Yet this method proved to be too slow for very large problems and has been replaced by `glmnet` which computes solutions for values of τ on a grid much faster.

3. The optimization community has made interesting contribution to this field by using proximal methods to solve this problem. It exploits the structure of the form: smooth (sum of squares) + simple (ℓ_1 norm). A good entry point to this literature is perhaps the FISTA algorithm [BT09].

4. There has been recently a lot of interest around this objective for very large d and very large n. In this case, even computing $|Y - \mathbb{X}\theta|_2^2$ may be computationally expensive and solutions based on stochastic gradient descent are flourishing.

Note that by Lagrange duality computing $\hat{\theta}^{\mathcal{L}}$ is equivalent to solving an ℓ_1 *constrained* least squares. Nevertheless, the radius of the ℓ_1 constraint is unknown. In general it is hard to relate Lagrange multipliers to the size constraints. The name "Lasso" was given to the constrained version this estimator in the original paper of Robert Tibshirani [Tib96].

Analysis of the BIC estimator

While computationally hard to implement, the BIC estimator gives us a good benchmark for sparse estimation. Its performance is similar to that of $\hat{\theta}^{\text{HRD}}$ but without assumption ORT.

Theorem 2.14. *Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then, the BIC estimator $\hat{\theta}^{\text{BIC}}$ with regularization parameter*

$$\tau^2 = 16\log(6)\frac{\sigma^2}{n} + 32\frac{\sigma^2\log(ed)}{n}\,. \tag{2.14}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\text{BIC}}) = \frac{1}{n}|\mathbb{X}\hat{\theta}^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 \lesssim |\theta^*|_0\sigma^2\frac{\log(ed/\delta)}{n}$$

with probability at least $1-\delta$.

Proof. We begin as usual by noting that

$$\frac{1}{n}|Y - \mathbb{X}\hat{\theta}^{\text{BIC}}|_2^2 + \tau^2|\hat{\theta}^{\text{BIC}}|_0 \le \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + \tau^2|\theta^*|_0\,.$$

It implies

$$|\mathbb{X}\hat{\theta}^{\text{BIC}} - \mathbb{X}\theta^*|_2^2 \le n\tau^2|\theta^*|_0 + 2\varepsilon^\top\mathbb{X}(\hat{\theta}^{\text{BIC}} - \theta^*) - n\tau^2|\hat{\theta}^{\text{BIC}}|_0\,.$$

First, note that

$$2\varepsilon^\top \mathbb{X}(\hat\theta^{\mathrm{BIC}} - \theta^*) = 2\varepsilon^\top \Big(\frac{\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2} \Big) |\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2$$
$$\leq 2\Big[\varepsilon^\top \Big(\frac{\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2} \Big) \Big]^2 + \frac{1}{2} |\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2^2 \,,$$

where we use the inequality $2ab \leq 2a^2 + \frac{1}{2}b^2$. Together with the previous display, it yields

$$|\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2^2 \leq 2n\tau^2 |\theta^*|_0 + 4\big[\varepsilon^\top \mathcal{U}(\hat\theta^{\mathrm{BIC}} - \theta^*)\big]^2 - 2n\tau^2 |\hat\theta^{\mathrm{BIC}}|_0 \tag{2.15}$$

where

$$\mathcal{U}(\hat\theta^{\mathrm{BIC}} - \theta^*) = \frac{\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*}{|\mathbb{X}\hat\theta^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2}$$

Next, we need to "sup out" $\hat\theta^{\mathrm{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbb{R}^d} = \max_{1 \leq k \leq d} \max_{|S| = k} \sup_{\mathrm{supp}(\theta) = S} .$$

Applied to the above inequality, it yields

$$4\big[\varepsilon^\top \mathcal{U}(\hat\theta^{\mathrm{BIC}} - \theta^*)\big]^2 - 2n\tau^2 |\hat\theta^{\mathrm{BIC}}|_0$$
$$\leq \max_{1 \leq k \leq d} \Big\{ \max_{|S| = k} \sup_{\mathrm{supp}(\theta) = S} 4\big[\varepsilon^\top \mathcal{U}(\theta - \theta^*)\big]^2 - 2n\tau^2 k \Big\}$$
$$\leq \max_{1 \leq k \leq d} \Big\{ \max_{|S| = k} \sup_{u \in \mathcal{B}_2^{r_{S,*}}} 4\big[\varepsilon^\top \Phi_{S,*} u\big]^2 - 2n\tau^2 k \Big\} \,,$$

where $\Phi_{S,*} = [\phi_1, \ldots, \phi_{r_{S,*}}]$ is an orthonormal basis of the set $\{\mathbb{X}_j, j \in S \cup \mathrm{supp}(\theta^*)\}$ of columns of $\mathbb{X}$ and $r_{S,*} \leq |S| + |\theta^*|_0$ is the dimension of this column span.

Using union bounds, we get for any $t > 0$,

$$\mathbb{P}\Big(\max_{1 \leq k \leq d} \Big\{ \max_{|S| = k} \sup_{u \in \mathcal{B}_2^{r_{S,*}}} 4\big[\varepsilon^\top \Phi_{S,*} u\big]^2 - 2n\tau^2 k \Big\} \geq t \Big)$$
$$\leq \sum_{k=1}^{d} \sum_{|S| = k} \mathbb{P}\Big(\sup_{u \in \mathcal{B}_2^{r_{S,*}}} \big[\varepsilon^\top \Phi_{S,*} u\big]^2 \geq \frac{t}{4} + \frac{1}{2} n\tau^2 k \Big)$$

Moreover, using the ε-net argument from Theorem 1.19, we get for $|S| = k$,

$$\mathbb{P}\Big(\sup_{u \in \mathcal{B}_2^{r_{S,*}}} \big[\varepsilon^\top \Phi_{S,*} u\big]^2 \geq \frac{t}{4} + \frac{1}{2} n\tau^2 k \Big) \leq 2 \cdot 6^{r_{S,*}} \exp\Big(- \frac{\frac{t}{4} + \frac{1}{2} n\tau^2 k}{8\sigma^2} \Big)$$
$$\leq 2 \exp\Big(- \frac{t}{32\sigma^2} - \frac{n\tau^2 k}{16\sigma^2} + (k + |\theta^*|_0) \log(6) \Big)$$
$$\leq \exp\Big(- \frac{t}{32\sigma^2} - 2k \log(ed) + |\theta^*|_0 \log(12) \Big)$$

where, in the last inequality, we used the definition (2.14) of τ.

Putting everything together, we get

$$
\begin{aligned}
&\mathbb{P}\Big(|\mathbb{X}\hat{\theta}^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2^2 \geq 2n\tau^2|\theta^*|_0 + t\Big) \leq \\
&\quad \sum_{k=1}^{d} \sum_{|S|=k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0 \log(12)\Big) \\
&\quad = \sum_{k=1}^{d} \binom{d}{k} \exp\Big(-\frac{t}{32\sigma^2} - 2k\log(ed) + |\theta^*|_0 \log(12)\Big) \\
&\quad \leq \sum_{k=1}^{d} \exp\Big(-\frac{t}{32\sigma^2} - k\log(ed) + |\theta^*|_0 \log(12)\Big) \qquad \text{by Lemma 2.7} \\
&\quad = \sum_{k=1}^{d} (ed)^{-k} \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0 \log(12)\Big) \\
&\quad \leq \exp\Big(-\frac{t}{32\sigma^2} + |\theta^*|_0 \log(12)\Big)\,.
\end{aligned}
$$

To conclude the proof, choose $t = 32\sigma^2|\theta^*|_0 \log(12) + 32\sigma^2 \log(1/\delta)$ and observe that combined with (2.15), it yields with probability $1-\delta$,

$$
\begin{aligned}
|\mathbb{X}\hat{\theta}^{\mathrm{BIC}} - \mathbb{X}\theta^*|_2^2 &\leq 2n\tau^2|\theta^*|_0 + t \\
&= 64\sigma^2 \log(ed)|\theta^*|_0 + 64\log(12)\sigma^2|\theta^*|_0 + 32\sigma^2\log(1/\delta) \\
&\leq 224|\theta^*|_0\sigma^2\log(ed) + 32\sigma^2\log(1/\delta)\,.
\end{aligned}
$$

□

It follows from Theorem 2.14 that $\hat{\theta}^{\mathrm{BIC}}$ *adapts* to the unknown sparsity of θ^*, just like $\hat{\theta}^{\mathrm{HRD}}$. Moreover, this holds under no assumption on the design matrix $\mathbb{X}$.

Analysis of the Lasso estimator

Slow rate for the Lasso estimator

The properties of the BIC estimator are quite impressive. It shows that under no assumption on $\mathbb{X}$, one can mimic two oracles: (i) the oracle that knows the support of θ^* (and computes least squares on this support), up to a $\log(ed)$ term and (ii) the oracle that knows the sparsity $|\theta^*|_0$ of θ^*, up to a smaller logarithmic term $\log(ed/|\theta^*|_0)$ is replaced by $\log(ed)$. Actually the latter can even be removed by using a modified BIC estimator (see Problem 2.6).

The Lasso estimator is a bit more difficult because, by construction, it should more naturally adapt to the unknown ℓ_1-norm of θ^*. This can be easily shown as in the next theorem, analogous to Theorem 2.4.

Theorem 2.15. *Assume that the linear model (2.2) holds where* $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. *Moreover, assume that the columns of* $\mathbb{X}$ *are normalized in such a way that* $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$. *Then, the Lasso estimator* $\hat{\theta}^{\mathcal{L}}$ *with regularization parameter*

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d)}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,. \tag{2.16}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \leq 4|\theta^*|_1\sigma\sqrt{\frac{2\log(2d)}{n}} + 4|\theta^*|_1\sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1-\delta$. *Moreover, there exists a numerical constant* $C > 0$ *such that*

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}})\big] \leq C|\theta^*|_1\sigma\sqrt{\frac{\log(2d)}{n}}\,.$$

Proof. From the definition of $\hat{\theta}^{\mathcal{L}}$, it holds

$$\frac{1}{n}|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}|_2^2 + 2\tau|\hat{\theta}^{\mathcal{L}}|_1 \leq \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + 2\tau|\theta^*|_1\,.$$

Using Hölder's inequality, it implies

$$\begin{aligned}
|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 &\leq 2\varepsilon^\top\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + 2n\tau\big(|\theta^*|_1 - |\hat{\theta}^{\mathcal{L}}|_1\big)\\
&\leq 2|\mathbb{X}^\top\varepsilon|_\infty|\hat{\theta}^{\mathcal{L}}|_1 - 2n\tau|\hat{\theta}^{\mathcal{L}}|_1 + 2|\mathbb{X}^\top\varepsilon|_\infty|\theta^*|_1 + 2n\tau|\theta^*|_1\\
&= 2(|\mathbb{X}^\top\varepsilon|_\infty - n\tau)|\hat{\theta}^{\mathcal{L}}|_1 + 2(|\mathbb{X}^\top\varepsilon|_\infty + n\tau)|\theta^*|_1
\end{aligned}$$

Observe now that for any $t > 0$,

$$\mathbb{P}(|\mathbb{X}^\top\varepsilon|_\infty \geq t) = \mathbb{P}(\max_{1\leq j\leq d}|\mathbb{X}_j^\top\varepsilon| > t) \leq 2de^{-\frac{t^2}{2n\sigma^2}}$$

Therefore, taking $t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)} = n\tau$, we get that with probability $1-\delta$,

$$|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \leq 4n\tau|\theta^*|_1\,.$$

The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Notice that the regularization parameter (2.16) depends on the confidence level δ. This not the case for the BIC estimator (see (2.14)).

The rate in Theorem 2.15 if of order $\sqrt{(\log d)/n}$ (**slow rate**), which is much slower than the rate of order $(\log d)/n$ (**fast rate**) for the BIC estimator. Hereafter, we show that fast rates can be achieved by the computationally efficient Lasso estimator but at the cost of a much stronger condition on the design matrix $\mathbb{X}$.

Incoherence

Assumption INC(k) We say that the design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if

$$\Big|\frac{\mathbb{X}^\top \mathbb{X}}{n} - I_d\Big|_\infty \leq \frac{1}{14k}$$

where the $|A|_\infty$ denotes the largest element of A in absolute value. Equivalently,

1. For all $j = 1, \dots, d$,
$$\Big|\frac{|\mathbb{X}_j|_2^2}{n} - 1\Big| \leq \frac{1}{14k}\,.$$
2. For all $1 \leq i, j \leq d$, $i \neq j$, we have
$$\big|\mathbb{X}_i^\top \mathbb{X}_j\big| \leq \frac{1}{14k}\,.$$

Note that Assumption ORT arises as the limiting case of INC(k) as $k \to \infty$. However, while Assumption ORT requires $d \leq n$, here we may have $d \gg n$ as illustrated in Proposition 2.16 below. To that end, we simply have to show that there exists a matrix that satisfies INC(k) even for $d > n$. We resort to the *probabilistic method* [AS08]. The idea of this method is that if we can find a probability measure that puts a positive probability of objects that satistify a certain property, then there must exist objects that satisfy said property. In our case, we consider the following probability distribution on random matrices with entries in $\{\pm 1\}$. Let the design matrix $\mathbb{X}$ have entries that are i.i.d Rademacher (± 1) random variables. We are going to show that most realizations of this random matrix satisfy Assumption INC(k) for large enough n.

Proposition 2.16. Let $\mathbb{X} \in \mathbb{R}^{n \times d}$ be a random matrix with entries $X_{ij}, i = 1, \dots, n, j = 1, \dots, d$ that are i.i.d Rademacher (± 1) random variables. Then, $\mathbb{X}$ has incoherence k with probability $1 - \delta$ as soon as

$$n \geq 392k^2 \log(1/\delta) + 784k^2 \log(d)\,.$$

It implies that there exists matrices that satisfy Assumption INC(k) for

$$n \gtrsim k^2 \log(d)\,,$$

for some numerical constant C.

Proof. Let $\varepsilon_{ij} \in \{-1, 1\}$ denote the Rademacher random variable that is on the ith row and jth column of $\mathbb{X}$.

Note first that the jth diagonal entries of $\mathbb{X}^\top \mathbb{X}/n$ is given by

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,j}^2 = 1$$

Moreover, for $j \neq k$, the (j,k)th entry of the $d \times d$ matrix $\frac{\mathbb{X}^\top \mathbb{X}}{n}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{i,j}\varepsilon_{i,k} = \frac{1}{n}\sum_{i=1}^{n} \xi_i^{(j,k)}\,,$$

where for each pair, (j,k), $\xi_i^{(j,k)} = \varepsilon_{i,j}\varepsilon_{i,k}$ so that the random variables $\xi_1^{(j,k)}, \ldots, \xi_n^{(j,k)}$ are iid Rademacher random variables.

Therefore, we get that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\big(\big|\frac{\mathbb{X}^\top\mathbb{X}}{n} - I_d\big|_\infty > t\big) &= \mathbb{P}\Big(\max_{j\neq k}\Big|\frac{1}{n}\sum_{i=1}^{n}\xi_i^{(j,k)}\Big| > t\Big) \\
&\leq \sum_{j\neq k}\mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n}\xi_i^{(j,k)}\Big| > t\Big) && \text{(Union bound)} \\
&\leq \sum_{j\neq k} 2e^{-\frac{nt^2}{2}} && \text{(Hoeffding: Theorem 1.9)} \\
&\leq d^2 e^{-\frac{nt^2}{2}}
\end{aligned}$$

Taking now $t = 1/(14k)$ yields

$$\mathbb{P}\big(\big|\frac{\mathbb{X}^\top\mathbb{X}}{n} - I_d\big|_\infty > \frac{1}{14k}\big) \leq d^2 e^{-\frac{n}{392k^2}} \leq \delta$$

for

$$n \geq 392k^2\log(1/\delta) + 784k^2\log(d)\,.$$

□

For any $\theta \in \mathbb{R}^d$, $S \subset \{1,\ldots,d\}$ define θ_S to be the vector with coordinates

$$\theta_{S,j} = \begin{cases} \theta_j & \text{if } j \in S\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

In particular $|\theta|_1 = |\theta_S|_1 + |\theta_{S^c}|_1$.

The following lemma holds

Lemma 2.17. *Fix a positive integer $k \leq d$ and assume that $\mathbb{X}$ satisfies assumption* ***INC****(k). Then, for any $S \in \{1,\ldots,d\}$ such that $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ that satisfies the* cone condition

$$|\theta_{S^c}|_1 \leq 3|\theta_S|_1\,, \tag{2.17}$$

it holds

$$|\theta_S|_2^2 \leq 2\frac{|\mathbb{X}\theta|_2^2}{n}$$

Proof. We have

$$\frac{|\mathbb{X}\theta|_2^2}{n} = \frac{1}{n}|\mathbb{X}\theta_S + \mathbb{X}\theta_{S^c}|_2^2 \geq \frac{|\mathbb{X}\theta_S|_2^2}{n} + 2\theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_{S^c}$$

If follows now from the incoherence condition that

$$\frac{|\mathbb{X}\theta_S|_2^2}{n} = \theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_S = |\theta_S|_2^2 + \theta_S^\top(\frac{\mathbb{X}^\top\mathbb{X}}{n} - I_d)\theta_S \geq |\theta_S|_2^2 - \frac{|\theta_S|_1^2}{14k}$$

and

$$\left|\theta_S^\top \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta_{S^c}\right| \leq \frac{1}{14k}|\theta_S|_1|\theta_{S^c}|_1 \leq \frac{3}{14k}|\theta_S|_1^2$$

Observe now that it follows from the Cauchy-Schwarz inequality that

$$|\theta_S|_1^2 \leq |S||\theta_S|_2^2$$

Thus for $|S| \leq k$,

$$\frac{|\mathbb{X}\theta|_2^2}{n} \geq \Big(1 - \frac{7|S|}{14k}\Big)|\theta_S|_2^2 \geq \frac{1}{2}|\theta_S|_2^2$$

□

Fast rate for the Lasso

Theorem 2.18. *Fix $n \geq 2$. Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that $|\theta^*|_0 \leq k$ and that $\mathbb{X}$ satisfies assumption $\mathit{INC}(k)$. Then the Lasso estimator $\hat\theta^{\mathcal{L}}$ with regularization parameter defined by*

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}}) = \frac{1}{n}|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}$$

and

$$|\hat\theta^{\mathcal{L}} - \theta^*|_1 \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}\big[\mathsf{MSE}(\mathbb{X}\hat\theta^{\mathcal{L}})\big] \lesssim k\sigma^2\frac{\log(2d)}{n}\,, \quad \textit{and} \quad \mathbb{E}\big[|\hat\theta^{\mathcal{L}} - \theta^*|_1\big] \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}\,.$$

Proof. From the definition of $\hat\theta^{\mathcal{L}}$, it holds

$$\frac{1}{n}|Y - \mathbb{X}\hat\theta^{\mathcal{L}}|_2^2 \leq \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + 2\tau|\theta^*|_1 - 2\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Adding $\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1$ on each side and multiplying by n, we get

$$|\mathbb{X}\hat\theta^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 + n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 \leq 2\varepsilon^\top\mathbb{X}(\hat\theta^{\mathcal{L}} - \theta^*) + n\tau|\hat\theta^{\mathcal{L}} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Applying Hölder's inequality and using the same steps as in the proof of Theorem 2.15, we get that with probability $1-\delta$, we get

$$\begin{aligned}\varepsilon^\top \mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) &\leq |\varepsilon^\top \mathbb{X}|_\infty |\hat{\theta}^{\mathcal{L}} - \theta^*| \\ &\leq \frac{n\tau}{2} |\hat{\theta}^{\mathcal{L}} - \theta^*|_1 \,,\end{aligned}$$

where we used the fact that $|\mathbb{X}_j|_2^2 \leq n + 1/(14k) \leq 2n$. Therefore, taking $S = \operatorname{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 + n\tau|\hat{\theta}^{\mathcal{L}} - \theta^*|_1 &\leq 2n\tau|\hat{\theta}^{\mathcal{L}} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}^{\mathcal{L}}|_1 \\ &= 2n\tau|\hat{\theta}^{\mathcal{L}}_S - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}^{\mathcal{L}}_S|_1 \\ &\leq 4n\tau|\hat{\theta}^{\mathcal{L}}_S - \theta^*|_1 \end{aligned} \tag{2.18}$$

In particular, it implies that

$$|\hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*_{S^c}|_1 \leq 3|\hat{\theta}^{\mathcal{L}}_S - \theta^*_S|_1 \,.$$

so that $\theta = \hat{\theta}^{\mathcal{L}} - \theta^*$ satisfies the cone condition (2.17). Using now the Cauchy-Schwarz inequality and Lemma 2.17 respectively, we get since $|S| \leq k$,

$$|\hat{\theta}^{\mathcal{L}}_S - \theta^*|_1 \leq \sqrt{|S|}|\hat{\theta}^{\mathcal{L}}_S - \theta^*|_2 \leq \sqrt{\frac{2k}{n}}|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2 \,.$$

Combining this result with (2.18), we find

$$|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2^2 \leq 32nk\tau^2 \,.$$

Moreover, it yields

$$\begin{aligned}|\hat{\theta}^{\mathcal{L}} - \theta^*|_1 &\leq 4\sqrt{\frac{2k}{n}}|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*|_2 \\ &\leq 4\sqrt{\frac{2k}{n}}\sqrt{32nk\tau^2} \leq 32k\tau\end{aligned}$$

The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Note that all we required for the proof was not really incoherence but the conclusion of Lemma 2.17:

$$\inf_{|S|\leq k} \inf_{\theta \in \mathcal{C}_S} \frac{|\mathbb{X}\theta|_2^2}{n|\theta_S|_2^2} \geq \kappa \tag{2.19}$$

where $\kappa = 1/2$ and $\mathcal{C}_S$ is the cone defined by

$$\mathcal{C}_S = \left\{ |\theta_{S^c}|_1 \leq 3|\theta_S|_1 \right\} .$$

Condition (2.19) is sometimes called *restricted eigenvalue (RE) condition*. Its name comes from the following observation. Note that all k-sparse vectors θ are in a cone $\mathcal{C}_S$ with $|S| \leq k$ so that the RE condition implies that the smallest eigenvalue of $\mathbb{X}_S$ satisfies $\lambda_{\min}(\mathbb{X}_S) \geq n\kappa$ for all S such that $|S| \leq k$. Clearly, the RE condition is weaker than incoherence and it can actually be shown that a design matrix $\mathbb{X}$ of i.i.d Rademacher random variables satisfies the RE conditions as soon as $n \geq Ck \log(d)$ with positive probability.

2.5 PROBLEM SET

Problem 2.1. Consider the linear regression model with fixed design with $d \le n$. The *ridge* regression estimator is employed when the $\operatorname{rank}(\mathbb{X}^\top \mathbb{X}) < d$ but we are interested in estimating θ^*. It is defined for a given parameter $\tau > 0$ by

$$\hat{\theta}_\tau^{\text{ridge}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + \tau |\theta|_2^2 \right\}.$$

(a) Show that for any τ, $\hat{\theta}_\tau^{\text{ridge}}$ is uniquely defined and give its closed form expression.

(b) Compute the bias of $\hat{\theta}_\tau^{\text{ridge}}$ and show that it is bounded in absolute value by $|\theta^*|_2$.

Problem 2.2. Let $X = (1, Z, \dots, Z^{d-1})^\top \in \mathbb{R}^d$ be a random vector where Z is a random variable. Show that the matrix $\mathbb{E}(XX^\top)$ is positive definite if Z admits a probability density with respect to the Lebesgue measure on $\mathbb{R}$.

Problem 2.3. In the proof of Theorem 2.11, show that $4\min(|\theta_j^*|, \tau)$ can be replaced by $3\min(|\theta_j^*|, \tau)$, i.e., that on the event $\mathcal{A}$, it holds

$$|\hat{\theta}_j^{\text{HRD}} - \theta_j^*| \le 3\min(|\theta_j^*|, \tau).$$

Problem 2.4. For any $q > 0$, a vector $\theta \in \mathbb{R}^d$ is said to be in a weak ℓ_q ball of radius R if the decreasing rearrangement $|\theta_{[1]}| \ge |\theta_{[2]}| \ge \dots$ satisfies

$$|\theta_{[j]}| \le R j^{-1/q}.$$

Moreover, we define the weak ℓ_q norm of θ by

$$|\theta|_{w\ell_q} = \max_{1 \le j \le d} j^{1/q} |\theta_{[j]}|$$

(a) Give examples of $\theta, \theta' \in \mathbb{R}^d$ such that

$$|\theta + \theta'|_{w\ell_1} > |\theta|_{w\ell_1} + |\theta'|_{w\ell_1}$$

What do you conclude?

(b) Show that $|\theta|_{w\ell_q} \le |\theta|_q$.

(c) Show that if $\lim_{d\to\infty} |\theta|_{w\ell_q} < \infty$, then $\lim_{d\to\infty} |\theta|_{q'} < \infty$ for all $q' > q$.

(d) Show that, for any $q \in (0, 2)$ if $\lim_{d\to\infty} |\theta|_{w\ell_q} = C$, there exists a constant $C_q > 0$ that depends on q but not on d and such that under the assumptions of Theorem 2.11, it holds

$$|\hat{\theta}^{\text{HRD}} - \theta^*|_2^2 \le C_q \Big(\frac{\sigma^2 \log 2d}{n} \Big)^{1 - \frac{q}{2}}$$

with probability .99.

Problem 2.5. Show that

$$\hat{\theta}^{\text{HRD}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ |y - \theta|_2^2 + 4\tau^2 |\theta|_0 \right\}$$
$$\hat{\theta}^{\text{SFT}} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ |y - \theta|_2^2 + 4\tau |\theta|_1 \right\}$$

Problem 2.6. Assume that the linear model (2.2) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$ and $\theta^* \neq 0$. Show that the modified BIC estimator $\hat{\theta}$ defined by

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{n} |Y - \mathbb{X}\theta|_2^2 + \lambda |\theta|_0 \log\left(\frac{ed}{|\theta|_0}\right) \right\}$$

satisfies,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) \lesssim |\theta^*|_0 \sigma^2 \frac{\log\left(\frac{ed}{|\theta^*|_0}\right)}{n} .$$

with probability .99, for appropriately chosen λ. What do you conclude?

Problem 2.7. Assume that the linear model (2.2) holds where $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume the conditions of Theorem 2.2 and that the columns of X are normalized in such a way that $\max_j |\mathbb{X}_j|_2 \leq \sqrt{n}$. Then the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter

$$2\tau = 8\sigma \sqrt{\frac{2\log(2d)}{n}} ,$$

satisfies

$$|\hat{\theta}^{\mathcal{L}}|_1 \leq C|\theta^*|_1$$

with probability $1 - (2d)^{-1}$ for some constant C to be specified.

CHAPTER 3

Misspecified Linear Models

Arguably, the strongest assumption that we made in Chapter 2 is that the regression function $f(x)$ is of the form $f(x) = x^\top \theta^*$. What if this assumption is violated? In reality, we do not really believe in the linear model and we hope that good statistical methods should be *robust* to deviations from this model. This is the problem of model misspecified linear models.

Throughout this chapter, we assume the following model:

$$Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \dots, n\,, \tag{3.1}$$

where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is sub-Gaussian with variance proxy σ^2. Here $X_i \in \mathbb{R}^d$. When dealing with fixed design, it will be convenient to consider the vector $g \in \mathbb{R}^n$ defined for any function $g : \mathbb{R}^d \to \mathbb{R}$ by $g = (g(X_1), \dots, g(X_n))^\top$. In this case, we can write for any estimator $\hat{f} \in \mathbb{R}^n$ of f,

$$\mathsf{MSE}(\hat{f}) = \frac{1}{n}|\hat{f} - f|_2^2\,.$$

Even though the model may not be linear, we are interested in studying the statistical properties of various linear estimators introduced in the previous chapters: $\hat{\theta}^{\text{LS}}, \hat{\theta}^{\text{LS}}_K, \hat{\theta}^{\text{LS}}_X, \hat{\theta}^{\text{BIC}}, \hat{\theta}^{\mathcal{L}}$. Clearly, even with an infinite number of observations, we have no chance of finding a consistent estimator of f if we don't know the correct model. Nevertheless, as we will see in this chapter something can still be said about these estimators using *oracle inequalities*.

3.1 ORACLE INEQUALITIES

Oracle inequalities

As mentioned in the introduction, an oracle is a quantity that cannot be constructed without the knowledge of the quantity of interest, here: the regression function. Unlike the regression function itself, an oracle is constrained to take a specific form. For all matter of purposes, an oracle can be viewed as an estimator (in a given family) that can be constructed with an infinite amount of data. This is exactly what we should aim for in misspecified models.

When employing the least squares estimator $\hat{\theta}^{\text{LS}}$, we constrain ourselves to estimating functions that are of the form $x \mapsto x^\top \theta$, even though f itself may not be of this form. Therefore, the oracle $\hat{f}$ is the linear function that is the closest to f.

Rather than trying to approximate f by a linear function $f(x) \approx \theta^\top x$, we make the model a bit more general and consider a dictionary $\mathcal{H} = \{\varphi_1, \dots, \varphi_M\}$ of functions where $\varphi_j : \mathbb{R}^d \to \mathbb{R}$. In the case, we can actually remove the assumption that $X \in \mathbb{R}^d$. Indeed, the goal is now to estimate f using a linear combination of the functions in the dictionary:

$$f \approx \varphi_\theta := \sum_{j=1}^{M} \theta_j \varphi_j \,.$$

Remark 3.1. If $M = d$ and $\varphi_j(X) = X^{(j)}$ returns the jth coordinate of $X \in \mathbb{R}^d$ then the goal is to approximate $f(x)$ by $\theta^\top x$. Nevertheless, the use of a dictionary allows for a much more general framework.

Note that the use of a dictionary does not affect the methods that we have been using so far, namely penalized/constrained least squares. We use the same notation as before and define

1. The least squares estimator:

$$\hat{\theta}^{\text{LS}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 \tag{3.2}$$

2. The least squares estimator constrained to $K \subset \mathbb{R}^M$:

$$\hat{\theta}^{\text{LS}}_K \in \operatorname*{argmin}_{\theta \in K} \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2$$

3. The BIC estimator:

$$\hat{\theta}^{\text{BIC}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 + \tau^2 |\theta|_0 \Big\} \tag{3.3}$$

4. The Lasso estimator:

$$\hat{\theta}^{\mathcal{L}} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^M} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \varphi_\theta(X_i)\big)^2 + 2\tau |\theta|_1 \Big\} \tag{3.4}$$

Definition 3.2. Let $R(\cdot)$ be a risk function and let $\mathcal{H} = \{\varphi_1, \dots, \varphi_M\}$ be a dictionary of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let K be a subset of $\mathbb{R}^M$. The *oracle* on K with respect to R is defined by $\varphi_{\bar{\theta}}$, where $\bar{\theta} \in K$ is such that

$$R(\varphi_{\bar{\theta}}) \leq R(\varphi_\theta)\,, \qquad \forall\, \theta \in K\,.$$

Moreover, $R_K = R(\varphi_{\bar{\theta}})$ is called *oracle risk* on K. An estimator $\hat{f}$ is said to satisfy an oracle inequality (over K) with remainder term ϕ in expectation (resp. with high probability) if there exists a constant $C \geq 1$ such that

$$\mathbb{E} R(\hat{f}) \leq C \inf_{\theta \in K} R(\varphi_\theta) + \phi_{n,M}(K)\,,$$

or

$$\mathbb{P}\big\{ R(\hat{f}) \leq C \inf_{\theta \in K} R(\varphi_\theta) + \phi_{n,M,\delta}(K) \big\} \geq 1 - \delta\,, \qquad \forall\ \delta > 0$$

respectively. If $C = 1$, the oracle inequality is sometimes called *exact*.

Our goal will be to mimic oracles. The finite sample performance of an estimator at this task is captured by an oracle inequality.

Oracle inequality for the least squares estimator

While our ultimate goal is to prove sparse oracle inequalities for the BIC and Lasso estimator in the case of misspecified model, the difficulty of the extension to this case for linear models, is essentially already captured for the least squares estimator. In this simple case, can even obtain an exact oracle inequality.

Theorem 3.3. *Assume the general regression model* (3.1) *with* $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. *Then, the least squares estimator* $\hat{\theta}^{\text{LS}}$ *satisfies for some numerical constant* $C > 0$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{LS}}}) \leq \inf_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 M}{n} \log(1/\delta)$$

with probability at least $1 - \delta$.

Proof. Note that by definition

$$|Y - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 \leq |Y - \varphi_{\bar{\theta}}|_2^2$$

where $\varphi_{\bar{\theta}}$ denotes the orthogonal projection of f onto the linear spam of $\varphi_1, \dots, \varphi_n$. Since $Y = f + \varepsilon$, we get

$$|f - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 \leq |f - \varphi_{\bar{\theta}}|_2^2 + 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}})$$

Moreover, by Pythagoras's theorem, we have

$$|f - \varphi_{\hat{\theta}^{\text{LS}}}|_2^2 - |f - \varphi_{\bar{\theta}}|_2^2 = |\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \,.$$

It yields

$$|\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \leq 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}) \,.$$

Using the same steps as the ones following equation (2.5) for the well specified case, we get

$$|\varphi_{\hat{\theta}^{\text{LS}}} - \varphi_{\bar{\theta}}|_2^2 \lesssim \frac{\sigma^2 M}{n} \log(1/\delta)$$

with probability $1 - \delta$. The result of the lemma follows. □

Sparse oracle inequality for the BIC estimator

The techniques that we have developed for the linear model above also allows to derive oracle inequalities.

Theorem 3.4. *Assume the general regression model (3.1) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Then, the BIC estimator $\hat{\theta}^{\text{BIC}}$ with regularization parameter*

$$\tau^2 = \frac{16\sigma^2}{\alpha n} \log(6eM) \,, \alpha \in (0,1) \tag{3.5}$$

satisfies for some numerical constant $C > 0$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \leq \inf_{\theta \in \mathbb{R}^M} \left\{ \frac{1+\alpha}{1-\alpha} \mathsf{MSE}(\varphi_\theta) + \frac{C\sigma^2}{\alpha(1-\alpha)n} |\theta|_0 \log(eM) \right\} + \frac{C\sigma^2}{\alpha(1-\alpha)n} \log(1/\delta)$$

with probability at least $1 - \delta$.

Proof. Recall the the proof of Theorem 2.14 for the BIC estimator begins as follows:

$$\frac{1}{n}|Y - \varphi_{\hat{\theta}^{\text{BIC}}}|_2^2 + \tau^2 |\hat{\theta}^{\text{BIC}}|_0 \leq \frac{1}{n}|Y - \varphi_\theta|_2^2 + \tau^2 |\theta|_0 \,.$$

This is true for any $\theta \in \mathbb{R}^M$. It implies

$$|f - \varphi_{\hat{\theta}^{\text{BIC}}}|_2^2 + n\tau^2 |\hat{\theta}^{\text{BIC}}|_0 \leq |f - \varphi_\theta|_2^2 + 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta) + n\tau^2 |\theta|_0 \,.$$

Note that if $\hat{\theta}^{\text{BIC}} = \theta$, the result is trivial. Otherwise,

$$\begin{aligned} 2\varepsilon^\top (\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta) &= 2\varepsilon^\top \Big(\frac{\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta}{|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2} \Big) |\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2 \\ &\leq \frac{2}{\alpha} \Big[\varepsilon^\top \Big(\frac{\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta}{|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2} \Big) \Big]^2 + \frac{\alpha}{2} |\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2^2 \,, \end{aligned}$$

where we use Young's inequality $2ab \leq \frac{2}{\alpha}a^2 + \frac{\alpha}{2}b^2$ valif for $a, b \geq 0$, $\alpha > 0$. Next, since

$$\frac{\alpha}{2}|\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta|_2^2 \leq \alpha|\varphi_{\hat{\theta}^{\text{BIC}}} - f|_2^2 + \alpha|\varphi_\theta - f|_2^2\,,$$

we get for $\alpha < 1$,

$$\begin{aligned}(1-\alpha)|\varphi_{\hat{\theta}^{\text{BIC}}} - f|_2^2 &\leq (1+\alpha)|\varphi_\theta - f|_2^2 + n\tau^2|\theta|_0 \\ &\quad + \frac{2}{\alpha}\left[\varepsilon^\top \mathcal{U}(\varphi_{\hat{\theta}^{\text{BIC}}} - \varphi_\theta)\right]^2 - n\tau^2|\hat{\theta}^{\text{BIC}}|_0 \\ &\leq (1+\alpha)|\varphi_\theta - f|_2^2 + 2n\tau^2|\theta|_0 \\ &\quad + \frac{2}{\alpha}\left[\varepsilon^\top \mathcal{U}(\varphi_{\hat{\theta}^{\text{BIC}}-\theta})\right]^2 - n\tau^2|\hat{\theta}^{\text{BIC}} - \theta|_0\end{aligned}$$

We conclude as in the proof of Theorem 2.14. □

A similar oracle can be obtained in expectation (exercise).

The interpretation of this theorem is enlightening. It implies that the BIC estimator will mimic the best tradeoff between the approximation error $\mathsf{MSE}(\varphi_\theta)$ and the complexity of θ as measured by its sparsity. In particular this result, sometimes called *sparse oracle inequality* implies the following oracle inequality. Define the oracle $\bar{\theta}$ to be such that

$$\mathsf{MSE}(\varphi_{\bar{\theta}}) = \min_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta)$$

then, with probability at least $1 - \delta$,

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \leq \frac{1+\alpha}{1-\alpha}\mathsf{MSE}(\varphi_{\bar{\theta}}) + \frac{C\sigma^2}{\alpha(1-\alpha)n}\left[|\bar{\theta}|_0 \log(eM) + \log(1/\delta)\right]\Big\}$$

If the linear model happens to be correct, then, simply, $\mathsf{MSE}(\varphi_{\bar{\theta}}) = 0$.

Sparse oracle inequality for the Lasso

To prove an oracle inequality for the Lasso, we need incoherence on the design. Here the design matrix is given by the $n \times M$ matrix Φ with elements $\Phi_{i,j} = \varphi_j(X_i)$.

Theorem 3.5. *Assume the general regression model (3.1) with $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$. Moreover, assume that there exists an integer k such that the matrix Φ satisfies assumption $\mathsf{INC}(k)$ holds. Then, the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter given by*

$$2\tau = 8\sigma\sqrt{\frac{2\log(2M)}{n}} + 8\sigma\sqrt{\frac{2\log(1/\delta)}{n}} \tag{3.6}$$

satisfies for some numerical constant C,

$$\mathsf{MSE}(\varphi_{\hat\theta^{\mathcal{L}}}) \le \inf_{\substack{\theta\in\mathbb{R}^M\\ |\theta|_0\le k}} \Big\{ \frac{1+\alpha}{1-\alpha}\mathsf{MSE}(\varphi_\theta) + \frac{C\sigma^2}{\alpha(1-\alpha)n}|\theta|_0 \log(eM)\Big\} + \frac{C\sigma^2}{\alpha(1-\alpha)n}\log(1/\delta)$$

with probability at least $1-\delta$.

Proof. From the definition of $\hat\theta^{\mathcal{L}}$, it holds for any $\theta\in\mathbb{R}^M$,

$$\frac{1}{n}|Y-\varphi_{\hat\theta^{\mathcal{L}}}|_2^2 \le \frac{1}{n}|Y-\varphi_\theta|_2^2 + 2\tau|\theta|_1 - 2\tau|\hat\theta^{\mathcal{L}}|_1\,.$$

Adding $\tau|\hat\theta^{\mathcal{L}}-\theta|_1$ on each side and multiplying by n, we get

$$|\varphi_{\hat\theta^{\mathcal{L}}}-f|_2^2 - |\varphi_\theta - f|_2^2 + n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 \le 2\varepsilon^\top(\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta) + n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\,. \tag{3.7}$$

Next, note that $\mathsf{INC}(k)$ for any $k\ge 1$ implies that $|\varphi_j|_2\le 2\sqrt{n}$ for all $j=1,\dots,M$. Applying Hölder's inequality using the same steps as in the proof of Theorem 2.15, we get that with probability $1-\delta$, it holds

$$2\varepsilon^\top(\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta) \le \frac{n\tau}{2}|\hat\theta^{\mathcal{L}}-\theta|_1$$

Therefore, taking $S=\mathrm{supp}(\theta)$ to be the support of θ, we get that the right-hand side of (3.7) is bounded by

$$\begin{aligned} &\le 2n\tau|\hat\theta^{\mathcal{L}}-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}|_1\\ &= 2n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 + 2n\tau|\theta|_1 - 2n\tau|\hat\theta^{\mathcal{L}}_S|_1\\ &\le 4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 \end{aligned} \tag{3.8}$$

with probability $1-\delta$.

It implies that either $\mathsf{MSE}(\varphi_{\hat\theta^{\mathcal{L}}}) \le \mathsf{MSE}(\varphi_\theta)$ or that

$$|\hat\theta^{\mathcal{L}}_{S^c} - \theta_{S^c}|_1 \le 3|\hat\theta^{\mathcal{L}}_S - \theta_S|_1\,.$$

so that $\theta=\hat\theta^{\mathcal{L}}-\theta$ satisfies the cone condition (2.17). Using now the Cauchy-Schwarz inequality and Lemma 2.17 respectively, assume that $|\theta|_0\le k$, we get

$$4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 \le 4n\tau\sqrt{|S|}|\hat\theta^{\mathcal{L}}_S-\theta|_2 \le 4\tau\sqrt{2n|\theta|_0}|\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta|_2\,.$$

Using now the inequality $2ab\le \frac{2}{\alpha}a^2 + \frac{\alpha}{2}b^2$, we get

$$\begin{aligned} 4n\tau|\hat\theta^{\mathcal{L}}_S-\theta|_1 &\le \frac{16\tau^2 n|\theta|_0}{\alpha} + \frac{\alpha}{2}|\varphi_{\hat\theta^{\mathcal{L}}}-\varphi_\theta|_2^2\\ &\le \frac{16\tau^2 n|\theta|_0}{\alpha} + \alpha|\varphi_{\hat\theta^{\mathcal{L}}}-f|_2^2 + \alpha|\varphi_\theta - f|_2^2 \end{aligned}$$

Combining this result with (3.7) and (3.8), we find

$$(1-\alpha)\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathcal{L}}}) \le (1+\alpha)\mathsf{MSE}(\varphi_\theta) + \frac{16\tau^2|\theta|_0}{\alpha}\,.$$

To conclude the proof of the bound with high probability, it only remains to divide by $1-\alpha$ on both sides of the above inequality. The bound in expectation follows using the same argument as in the proof of Corollary 2.9. □

Maurey's argument

From the above section, it seems that the Lasso estimator is strictly better than the BIC estimator as long as incoherence holds. Indeed, if there is no sparse θ such that $\mathsf{MSE}(\varphi_\theta)$ is small, Theorem 3.4 is useless. In reality, no one really believes in the existence of sparse vectors but rater of approximately sparse vectors. Zipf's law would instead favor the existence of vectors θ with absolute coefficients that decay polynomially when ordered from largest to smallest in absolute value. This is the case for example if θ has a small ℓ_1 norm but is not sparse. For such θ, the Lasso estimator still enjoys slow rates as in Theorem 2.15, which can be easily extended to the misspecified case (see Problem 3.2). Fortunately, such vectors can be well approximated by sparse vectors in the following sense: for any vector $\theta \in \mathbb{R}^M$ such that $|\theta|_1 \le 1$, there exists a vector θ' that is sparse and for which $\mathsf{MSE}(\varphi_{\theta'})$ is not much larger than $\mathsf{MSE}(\varphi_\theta)$. The following theorem quantifies exactly the tradeoff between sparsity and MSE. It is often attributed to B. Maurey and was published by Pisier [Pis81]. This is why it is referred to as *Maurey's argument*.

Theorem 3.6. *Let $\{\varphi_1, \dots, \varphi_M\}$ be a dictionary normalized in such a way that*

$$\max_{1\le j\le M} |\varphi_j|_2 \le D\sqrt{n}\,.$$

Then for any integer k such that $1 \le k \le M$ and any positive R, we have

$$\min_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_0\le 2k}} \mathsf{MSE}(\varphi_\theta) \le \min_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_1\le R}} \mathsf{MSE}(\varphi_\theta) + \frac{D^2R^2}{k}\,.$$

Proof. Define

$$\bar{\theta} \in \operatorname*{argmin}_{\substack{\theta\in\mathbb{R}^M \\ |\theta|_1\le R}} |\varphi_\theta - f|_2^2$$

and assume without loss of generality that $|\bar{\theta}_1| \ge |\bar{\theta}_2| \ge \ldots \ge |\bar{\theta}_M|$.

Now decompose $\bar{\theta} = \theta^{(1)} + \theta^{(2)}$, where $\mathrm{supp}(\theta^{(1)}) \subset \{1 \ldots, k\}$ and $\mathrm{supp}(\theta^{(2)}) \subset \{k+1, \dots, M\}$. In particular it holds

$$\varphi_{\bar{\theta}} = \varphi_{\theta^{(1)}} + \varphi_{\theta^{(2)}}\,.$$

Moreover, observe that

$$|\theta^{(2)}|_1 = \sum_{j=k+1}^{M} |\bar{\theta}_j| \le R$$

Let now $U \in \mathbb{R}^n$ be a random vector with values in $\{0, \pm R\varphi_1, \dots, \pm R\varphi_M\}$ defined by

$$\mathbb{P}(U = R\mathrm{sign}(\theta_j^{(2)})\varphi_j) = \frac{|\theta_j^{(2)}|}{R}, \quad j = k+1, \dots, M$$
$$\mathbb{P}(U = 0) = 1 - \frac{|\theta^{(2)}|_1}{R}.$$

Note that $\mathbb{E}[U] = \varphi_{\theta^{(2)}}$ and $|U|_2 \le RD\sqrt{n}$. Let now $U_1, \dots, U_k$ be k independent copies of U define

$$\bar{U} = \frac{1}{k}\sum_{i=1}^{k} U_i .$$

Note that $\bar{U} = \varphi_{\tilde{\theta}}$ for some $\tilde{\theta} \in \mathbb{R}^M$ such that $|\tilde{\theta}|_0 \le k$. Therefore, $|\theta^{(1)} + \tilde{\theta}|_0 \le 2k$ and

$$\begin{aligned}
\mathbb{E}|f - \varphi_{\theta^{(1)}} - \bar{U}|_2^2 &= \mathbb{E}|f - \varphi_{\theta^{(1)}} - \varphi_{\theta^{(2)}} + \varphi_{\theta^{(2)}} - \bar{U}|_2^2 \\
&= \mathbb{E}|f - \varphi_{\theta^{(1)}} - \varphi_{\theta^{(2)}}|_2^2 + |\varphi_{\theta^{(2)}} - \bar{U}|_2^2 \\
&= |f - \varphi_{\tilde{\theta}}|_2^2 + \frac{\mathbb{E}|U - \mathbb{E}[U]|_2^2}{k} \\
&\le |f - \varphi_{\tilde{\theta}}|_2^2 + \frac{(RD\sqrt{n})^2}{k}
\end{aligned}$$

To conclude the proof, note that

$$\mathbb{E}|f - \varphi_{\theta^{(1)}} - \bar{U}|_2^2 = \mathbb{E}|f - \varphi_{\theta^{(1)}+\tilde{\theta}}|_2^2 \ge \min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_0 \le 2k}} |f - \varphi_\theta|_2^2$$

and to divide by n. □

Maurey's argument implies the following corollary.

Corollary 3.7. *Assume that the assumptions of Theorem 3.4 hold and that the dictionary $\{\varphi_1, \dots, \varphi_M\}$ is normalized in such a way that*

$$\max_{1 \le j \le M} |\varphi_j|_2 \le \sqrt{n}\,.$$

Then there exists a constant $C > 0$ such that the BIC estimator satisfies

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathrm{BIC}}}) \le \inf_{\theta \in \mathbb{R}^M} \Big\{ 2\mathsf{MSE}(\varphi_\theta) + C\Big[\frac{\sigma^2 |\theta|_0 \log(eM)}{n} \wedge \sigma|\theta|_1 \sqrt{\frac{\log(eM)}{n}}\Big]\Big\} + C\frac{\sigma^2 \log(1/\delta)}{n}$$

with probability at least $1 - \delta$.

Proof. Choosing $\alpha = 1/3$ in Theorem 3.4 yields

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\text{BIC}}}) \le 2 \inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \right\} + C \frac{\sigma^2 \log(1/\delta)}{n}$$

For any $\theta' \in \mathbb{R}^M$, it follows from Maurey's argument that there exist $\theta \in \mathbb{R}^M$ such that $|\theta|_0 \le 2|\theta'|_0$ and

$$\mathsf{MSE}(\varphi_\theta) \le \mathsf{MSE}(\varphi_{\theta'}) + \frac{2|\theta'|_1^2}{|\theta|_0}$$

It implies that

$$\mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \le \mathsf{MSE}(\varphi_{\theta'}) + \frac{2|\theta'|_1^2}{|\theta|_0} + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n}$$

Taking infimum on both sides, we get

$$\begin{aligned} &\inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C \frac{\sigma^2 |\theta|_0 \log(eM)}{n} \right\} \\ &\qquad \le \inf_{\theta' \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_{\theta'}) + C \min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \right\}. \end{aligned}$$

To control the minimum over k, we need to consider three cases for the quantity

$$\bar{k} = \frac{|\theta'|_1}{\sigma} \sqrt{\frac{\log M}{n}}$$

1. If $1 \le \bar{k} \le M$, then we get

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C\sigma |\theta'|_1 \sqrt{\frac{\log(eM)}{n}}$$

2. If $\bar{k} \le 1$, then

$$|\theta'|_1^2 \le C \frac{\sigma^2 \log(eM)}{n},$$

which yields

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C \frac{\sigma^2 \log(eM)}{n}$$

3. If $\bar{k} \ge M$, then

$$\frac{\sigma^2 M \log(eM)}{n} \le C \frac{|\theta'|_1^2}{M}.$$

Therefore, on the one hand, if $M \ge \frac{|\theta|_1}{\sigma \sqrt{\log(eM)/n}}$, we get

$$\min_k \left(\frac{|\theta'|_1^2}{k} + C \frac{\sigma^2 k \log(eM)}{n} \right) \le C \frac{|\theta'|_1^2}{M} \le C\sigma |\theta'|_1 \sqrt{\frac{\log(eM)}{n}}.$$

On the other hand, if $M \leq \frac{|\theta|_1}{\sigma\sqrt{\log(eM)/n}}$, then for any $\Theta \in \mathbb{R}^M$, we have

$$\frac{\sigma^2|\theta|_0 \log(eM)}{n} \leq \frac{\sigma^2 M \log(eM)}{n} \leq C\sigma|\theta'|_1\sqrt{\frac{\log(eM)}{n}}\,.$$

□

Note that this last result holds for any estimator that satisfies an oracle inequality with respect to the ℓ_0 norm such as the result of Theorem 3.4. In particular, this estimator need not be the BIC estimator. An example is the Exponential Screening estimator of [RT11].

Maurey's argument allows us to enjoy the best of both the ℓ_0 and the ℓ_1 world. The rate adapts to the sparsity of the problem and can be even generalized to ℓ_q-sparsity (see Problem 3.3). However, it is clear from the proof that this argument is limited to squared ℓ_2 norms such as the one appearing in MSE and extension to other risk measures is non trivial. Some work has been done for non Hilbert spaces [Pis81, DDGS97] using more sophisticated arguments.

3.2 NONPARAMETRIC REGRESSION

So far, the oracle inequalities that we have derived do not deal with the approximation error $\mathsf{MSE}(\varphi_\theta)$. We kept it arbitrary and simply hoped that it was small. Note also that in the case of linear models, we simply assumed that the approximation error was zero. As we will see in this section, this error can be quantified under natural smoothness conditions if the dictionary of functions $\mathcal{H} = \{\varphi_1, \ldots, \varphi_M\}$ is chosen appropriately. In what follows, we assume for simplicity that $d = 1$ so that $f : \mathbb{R} \to \mathbb{R}$ and $\varphi_j : \mathbb{R} \to \mathbb{R}$.

Fourier decomposition

Historically, nonparametric estimation was developed before high-dimensional statistics and most results hold for the case where the dictionary $\mathcal{H} = \{\varphi_1, \ldots, \varphi_M\}$ forms an orthonormal system of $L_2([0,1])$:

$$\int_0^1 \varphi_j^2(x)\mathrm{d}x = 1\,, \quad \int_0^1 \varphi_j(x)\varphi_k(x)\mathrm{d}x = 0, \ \forall\, j \neq k\,.$$

We will also deal with the case where $M = \infty$.

When $\mathcal{H}$ is an orthonormal system, the coefficients $\theta_j^* \in \mathbb{R}$ defined by

$$\theta_j^* = \int_0^1 f(x)\varphi_j(x)\mathrm{d}x\,,$$

are called *Fourier coefficients* of f.

Assume now that the regression function f admits the following decomposition

$$f = \sum_{j=1}^{\infty} \theta_j^* \varphi_j \,.$$

There exists many choices for the orthonormal system and we give only two as examples.

Example 3.8. *Trigonometric basis.* This is an orthonormal basis of $L_2([0,1])$. It is defined by

$$\begin{aligned} \varphi_1 &\equiv 1 \\ \varphi_{2k}(x) &= \sqrt{2}\cos(2\pi k x)\,, \\ \varphi_{2k+1}(x) &= \sqrt{2}\sin(2\pi k x)\,, \end{aligned}$$

for $k = 1, 2, \ldots$ and $x \in [0,1]$. The fact that it is indeed an orthonormal system can be easily check using trigonometric identities.

The next example has received a lot of attention in the signal (sound, image, ...) processing community.

Example 3.9. *Wavelets.* Let $\psi : \mathbb{R} \to \mathbb{R}$ be a sufficiently smooth and compactly supported function, called "*mother wavelet*". Define the system of functions

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)\,, \quad j,k \in \mathbb{Z}\,.$$

It can be shown that for a suitable ψ, the dictionary $\{\psi_{j,k}, j,k \in \mathbb{Z}\}$ forms an orthonormal system of $L_2([0,1])$ and sometimes a basis. In the latter case, for any function $g \in L_2([0,1])$, it holds

$$g = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \theta_{jk}\psi_{jk}\,, \quad \theta_{jk} = \int_0^1 g(x)\psi_{jk}(x)\mathrm{d}x\,.$$

The coefficients θ_{jk} are called *wavelet coefficients* of g.

The simplest example is given by the *Haar system* obtained by taking ψ to be the following piecewise constant function (see Figure 3.1). We will not give more details about wavelets here but refer simply point the interested reader to [Mal09].

$$\psi(x) = \begin{cases} 1 & 0 \le x < 1/2 \\ -1 & 1/2 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Sobolev classes and ellipsoids

We begin by describing a class of smooth functions where smoothness is understood in terms of its number of derivatives. Recall that $f^{(k)}$ denotes the k-th derivative of f.

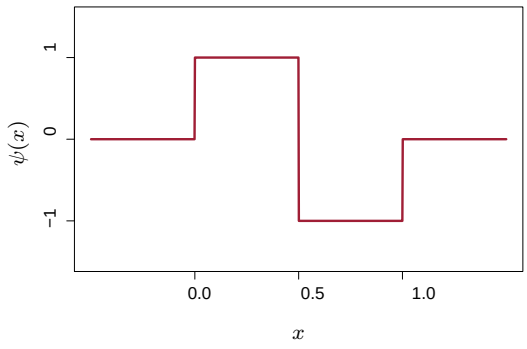


Figure 3.1. The Haar mother wavelet

Definition 3.10. Fix parameters $\beta \in \{1, 2, \dots\}$ and $L > 0$. The Sobolev class of functions $W(\beta, L)$ is defined by

$$W(\beta, L) = \Big\{ f : [0,1] \to \mathbb{R} \ : \ f \in L_2([0,1]), f^{(\beta-1)} \text{ is absolutely continuous and } \int_0^1 [f^{(\beta)}]^2 \le L^2, f^{(j)}(0) = f^{(j)}(1), j = 0, \dots, \beta - 1 \Big\}$$

Any function $f \in W(\beta, L)$ can represented[1] as its Fourier expansion along the trigonometric basis:

$$f(x) = \theta_1^* \varphi_1(x) + \sum_{k=1}^{\infty} \left(\theta_{2k}^* \varphi_{2k}(x) + \theta_{2k+1}^* \varphi_{2k+1}(x) \right), \quad \forall\, x \in [0,1],$$

where $\theta^* = \{\theta_j^*\}_{j \ge 1}$ is in the space of squared summable sequence $\ell_2(\mathbb{N})$ defined by

$$\ell_2(\mathbb{N}) = \Big\{ \theta \ : \ \sum_{j=1}^{\infty} \theta_j^2 < \infty \Big\}.$$

For any $\beta > 0$, define the coefficients

$$a_j = \begin{cases} j^\beta & \text{for } j \text{ even} \\ (j-1)^\beta & \text{for } j \text{ odd} \end{cases} \tag{3.9}$$

Thanks to these coefficients, we can define the Sobolev class of functions in terms of Fourier coefficients.

[1] In the sense that

$$\lim_{k \to \infty} \int_0^1 |f(t) - \sum_{j=1}^{k} \theta_j \varphi_j(t)|^2 \mathrm{d}t = 0$$

Theorem 3.11. *Fix $\beta \geq 1$ and $L > 0$ and let $\{\varphi_j\}_{j\geq 1}$ denote the trigonometric basis of $L_2([0,1])$. Moreover, let $\{a_j\}_{j\geq 1}$ be defined as in (3.9). A function $f \in W(\beta, L)$ can be represented as*

$$f = \sum_{j=1}^{\infty} \theta_j^* \varphi_j \,,$$

where the sequence $\{\theta_j^*\}_{j\geq 1}$ belongs to Sobolev ellipsoid of $\ell_2(\mathbb{N})$ defined by

$$\Theta(\beta, Q) = \Big\{ \theta \in \ell_2(\mathbb{N}) \,:\, \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq Q \Big\}$$

for $Q = L^2/\pi^{2\beta}$.

Proof. Let us first recall the definition of the Fourier coefficients $\{s_k(j)\}_{k\geq 1}$ of the jth derivative $f^{(j)}$ of f for $j = 1, \ldots, \beta$:

$$\begin{aligned}
s_1(j) &= \int_0^1 f^{(j)}(t)\mathrm{d}t = f^{(j-1)}(1) - f^{(j-1)}(0) = 0 \,, \\
s_{2k}(j) &= \sqrt{2} \int_0^1 f^{(j)}(t) \cos(2\pi k t)\mathrm{d}t \,, \\
s_{2k+1}(j) &= \sqrt{2} \int_0^1 f^{(j)}(t) \sin(2\pi k t)\mathrm{d}t \,,
\end{aligned}$$

The Fourier coefficients of f are given by $\theta_k = s_k(0)$.

Using integration by parts, we find that

$$\begin{aligned}
s_{2k}(\beta) &= \sqrt{2} f^{(\beta-1)}(t) \cos(2\pi k t)\Big|_0^1 + (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \sin(2\pi k t)\mathrm{d}t \\
&= \sqrt{2}[f^{(\beta-1)}(1) - f^{(\beta-1)}(0)] + (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \sin(2\pi k t)\mathrm{d}t \\
&= (2\pi k) s_{2k+1}(\beta - 1) \,.
\end{aligned}$$

Moreover,

$$\begin{aligned}
s_{2k+1}(\beta) &= \sqrt{2} f^{(\beta-1)}(t) \sin(2\pi k t)\Big|_0^1 - (2\pi k)\sqrt{2} \int_0^1 f^{(\beta-1)}(t) \cos(2\pi k t)\mathrm{d}t \\
&= -(2\pi k) s_{2k}(\beta - 1) \,.
\end{aligned}$$

In particular, it yields

$$s_{2k}(\beta)^2 + s_{2k+1}(\beta)^2 = (2\pi k)^2 \big[s_{2k}(\beta - 1)^2 + s_{2k+1}(\beta - 1)^2 \big]$$

By induction, we find that for any $k \geq 1$,

$$s_{2k}(\beta)^2 + s_{2k+1}(\beta)^2 = (2\pi k)^{2\beta} \big(\theta_{2k}^2 + \theta_{2k+1}^2 \big)$$

Next, it follows for the definition (3.9) of a_j that

$$\sum_{k=1}^{\infty}(2\pi k)^{2\beta}\big(\theta_{2k}^2+\theta_{2k+1}^2\big) = \pi^{2\beta}\sum_{k=1}^{\infty}a_{2k}^2\theta_{2k}^2 + \pi^{2\beta}\sum_{k=1}^{\infty}a_{2k+1}^2\theta_{2k+1}^2$$
$$= \pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\theta_j^2\,.$$

Together with the Parseval identity, it yields

$$\int_0^1\big(f^{(\beta)}(t)\big)^2\mathrm{d}t = \sum_{k=1}^{\infty}s_{2k}(\beta)^2+s_{2k+1}(\beta)^2 = \pi^{2\beta}\sum_{j=1}^{\infty}a_j^2\theta_j^2\,.$$

To conclude, observe that since $f\in W(\beta,L)$, we have

$$\int_0^1\big(f^{(\beta)}(t)\big)^2\mathrm{d}t \le L^2\,,$$

so that $\theta\in\Theta(\beta,L^2/\pi^{2\beta})$. □

It can actually be shown that the reciprocal is true, that is any function with Fourier coefficients in $\Theta(\beta,Q)$ belongs to if $W(\beta,L)$ but we will not be needing this.

In what follows, we will define smooth functions as functions with Fourier coefficients (with respect to the trigonometric basis) in a Sobolev ellipsoid. By extension, we write $f\in\Theta(\beta,Q)$ in this case and consider any real value for β.

Proposition 3.12. The Sobolev ellipsoids enjoy the following properties

(i) For any $Q>0$,

$$0<\beta'<\beta \;\Rightarrow\; \Theta(\beta,Q)\subset\Theta(\beta',Q)$$

(ii) For any $Q>0$,

$$\beta>\frac{1}{2} \;\Rightarrow\; f \text{ is continuous}$$

The proof is left as an exercise (Problem 3.5)

It turns out that the first functions in the trigonometric basis are orthonormal with respect to the inner product of L_2 but also to the inner predictor associated to fixed design $\langle f,g\rangle := \frac{1}{n}f(X_i)g(X_i)$ when the design is chosen to be *regular*, i.e., $X_i=(i-1)/n$, $i=1,\dots,n$.

Lemma 3.13. *Assume that $\{X_1,\dots,X_n\}$ is the regular design, i.e., $X_i=(i-1)/n$. Then, for any $M\le n-1$, the design matrix $\Phi=\{\varphi_j(X_i)\}_{\substack{1\le i\le n\\ 1\le j\le M}}$ satisfies the* ORT *condition.*

Proof. Note first that for any $j, j' \in \{1, \dots, n-1\}$, $j \neq j'$ the inner product $\varphi_j^\top \varphi_{j'}$ is of the form

$$\varphi_j^\top \varphi_{j'} = 2 \sum_{s=0}^{n-1} u_j(2\pi k_j s/n) v_{j'}(2\pi k_{j'} s/n)$$

where $k_j = \lfloor j/2 \rfloor$ is the integer part of $j/2$ for any $x \in \mathbb{R}$, $u_j(x), v_{j'}(x) \in \{\mathfrak{Re}(e^{\mathrm{i}x}), \mathfrak{Im}(e^{\mathrm{i}x})\}$.

Next, observe that if $k_j \neq k_{j'}$, we have

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{-\frac{\mathrm{i}2\pi k_{j'} s}{n}} = \sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi (k_j - k_{j'}) s}{n}} = 0\,.$$

Moreover, if we define the vectors $a, b, a', b' \in \mathbb{R}^n$ with coordinates such that $e^{\frac{\mathrm{i}2\pi k_j s}{n}} = a_s + \mathrm{i}b_s$ and $e^{\frac{\mathrm{i}2\pi k_{j'} s}{n}} = a'_s + \mathrm{i}b'_s$, we get

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{-\frac{\mathrm{i}2\pi k_{j'} s}{n}} = (a + \mathrm{i}b)^\top (a' - \mathrm{i}b') = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big]$$

and consequently that

$$\frac{1}{2}\varphi_j^\top \varphi_{j'} = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big]$$

with $|a|_2 |b|_2 = |a'|_2 |b'|_2 = 0$, i.e., either $a = 0$ or $b = 0$ and either $a' = 0$ or $b' = 0$. Therefore, in the case where $k_j \neq k_{j'}$, we have

$$a^\top a' = -b^\top b' = 0, \qquad b^\top a' = a^\top b' = 0$$

which implies $\varphi_j^\top \varphi_{j'} = 0$. To conclude the proof, it remains to deal with the case where $k_j = k_{j'}$. This can happen in two cases: $|j' - j| = 1$ or $j' = j$. In the first case, we have that $\{u_j(x), v_{j'}(x)\} = \{\mathfrak{Re}(e^{\mathrm{i}x}), \mathfrak{Im}(e^{\mathrm{i}x})\}$, i.e., one is a $\sin(\cdot)$ and the other is a $\cos(\cdot)$. Therefore,

$$\frac{1}{2}\varphi_j^\top \varphi_{j'} = a^\top a' + b^\top b' + \mathrm{i}\big[b^\top a' - a^\top b'\big] = 0$$

The final case is $j = j'$ for which, on the one hand,

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{\frac{\mathrm{i}2\pi k_j s}{n}} = \sum_{s=0}^{n-1} e^{\frac{\mathrm{i}4\pi k_j s}{n}} = 0$$

and on the other hand

$$\sum_{s=0}^{n-1} e^{\frac{\mathrm{i}2\pi k_j s}{n}} e^{\frac{\mathrm{i}2\pi k_j s}{n}} = |a + \mathrm{i}b|_2^2 = |a|_2^2 - |b|_2^2$$

so that $|a|^2 = |b|^2$. Moreover, by definition,

$$|\varphi_j|_2^2 = \begin{cases} 2|a|_2^2 & \text{if } j \text{ is even} \\ 2|b|_2^2 & \text{if } j \text{ is odd} \end{cases}$$

so that

$$|\varphi_j|_2^2 = 2\frac{|a|_2^2 + |b|_2^2}{2} = \sum_{s=0}^{n-1} \left|e^{\frac{\mathrm{i}2\pi k_j s}{n}}\right|^2 = n$$

Therefore, the design matrix Φ is such that

$$\Phi^\top \Phi = nI_M \,.$$

□

Integrated squared error

As mentioned in the introduction of this chapter, the smoothness assumption allows us to control the approximation error. Before going into the details, let us gain some insight. Note first that if $\theta \in \Theta(\beta, Q)$, then $a_j^2\theta_j^2 \to 0$ as $j \to \infty$ so that $|\theta_j| = o(j^{-\beta})$. Therefore, the θ_js decay polynomially to zero and it makes sense to approximate f by its truncated Fourier series

$$\sum_{j=1}^{M} \theta_j^* \varphi_j =: \varphi_{\theta^*}^M$$

for any fixed M. This truncation leads to a systematic error that vanishes as $M \to \infty$. We are interested in understanding the rate at which this happens.

The Sobolev assumption to control precisely this error as a function of the tunable parameter M and the smoothness β.

Lemma 3.14. *For any integer $M \geq 1$, and $f \in \Theta(\beta, Q)$, $\beta > 1/2$, it holds*

$$\|\varphi_{\theta^*}^M - f\|_{L_2}^2 = \sum_{j>M} |\theta_j^*|^2 \leq QM^{-2\beta} \,. \tag{3.10}$$

and for $M = n - 1$, we have

$$|\varphi_{\theta^*}^{n-1} - f|_2^2 \leq 2n\Big(\sum_{j\geq n} |\theta_j^*|\Big)^2 \lesssim Qn^{2-2\beta} \,. \tag{3.11}$$

Proof. Note that for any $\theta \in \Theta(\beta, Q)$, if $\beta > 1/2$, then

$$\begin{aligned} \sum_{j=2}^{\infty} |\theta_j| &= \sum_{j=2}^{\infty} a_j |\theta_j| \frac{1}{a_j} \\ &\leq \sqrt{\sum_{j=2}^{\infty} a_j^2\theta_j^2 \sum_{j=2}^{\infty} \frac{1}{a_j^2}} \quad \text{by Cauchy-Schwarz} \\ &\leq \sqrt{Q\sum_{j=1}^{\infty} \frac{1}{j^{2\beta}}} < \infty \end{aligned}$$

Since $\{\varphi_j\}_j$ forms an orthonormal system in $L_2([0,1])$, we have

$$\min_{\theta\in\mathbb{R}^M} \|\varphi_\theta - f\|_{L_2}^2 = \|\varphi_{\theta^*} - f\|_{L_2}^2 = \sum_{j>M} |\theta_j^*|^2 \,.$$

When $\theta^* \in \Theta(\beta, Q)$, we have

$$\sum_{j>M} |\theta_j^*|^2 = \sum_{j>M} a_j^2 |\theta_j^*|^2 \frac{1}{a_j^2} \le \frac{1}{a_{M+1}^2} Q \le \frac{Q}{M^{2\beta}} \,.$$

To prove the second part of the lemma, observe that

$$|\varphi_{\theta^*}^{n-1} - f|_2 = \big|\sum_{j\ge n} \theta_j^* \varphi_j\big|_2 \le 2\sqrt{2n} \sum_{j\ge n} |\theta_j^*| \,,$$

where in the last inequality, we used the fact that for the trigonometric basis $|\varphi_j|_2 \le \sqrt{2n}, j \ge 1$ regardless of the choice of the design $X_1, \dots, X_n$. When $\theta^* \in \Theta(\beta, Q)$, we have

$$\sum_{j\ge n} |\theta_j^*| = \sum_{j\ge n} a_j |\theta_j^*| \frac{1}{a_j} \le \sqrt{\sum_{j\ge n} a_j^2 |\theta_j^*|^2} \sqrt{\sum_{j\ge n} \frac{1}{a_j^2}} \lesssim Q n^{\frac{1}{2}-\beta} \,.$$

□

Note the truncated Fourier series φ_{θ^*} is an oracle: this is what we see when we view f through the lens of functions with only low frequency harmonics.

To estimate φ_{θ^*}, consider the estimator $\varphi_{\hat\theta^{\text{LS}}}$ where

$$\hat\theta^{\text{LS}} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^M} \sum_{i=1}^n \big(Y_i - \varphi_\theta(X_i)\big)^2 \,.$$

Which should be such that $\varphi_{\hat\theta^{\text{LS}}}$ is close to φ_{θ^*}. For this estimator, we have proved (Theorem 3.3) an oracle inequality for the MSE that is of the form

$$|\varphi_{\hat\theta^{\text{LS}}}^M - f|_2^2 \le \inf_{\theta\in\mathbb{R}^M} |\varphi_\theta^M - f|_2 + C\sigma^2 M \log(1/\delta) \,, \qquad C > 0 \,.$$

It yields

$$\begin{aligned}
|\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M|_2^2 &\le 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (f - \varphi_{\theta^*}^M) + C\sigma^2 M \log(1/\delta) \\
&= 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j>M} \theta_j^* \varphi_j) + C\sigma^2 M \log(1/\delta) \\
&= 2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j\ge n} \theta_j^* \varphi_j) + C\sigma^2 M \log(1/\delta) \,,
\end{aligned}$$

where we used Lemma 3.13 in the last equality. Together with (3.11) and Young's inequality $2ab \le \alpha a^2 + b^2/\alpha, a, b \ge 0$ for any $\alpha > 0$, we get

$$2(\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M)^\top (\sum_{j\ge n} \theta_j^* \varphi_j) \le \alpha |\varphi_{\hat\theta^{\text{LS}}}^M - \varphi_{\theta^*}^M|_2^2 + \frac{C}{\alpha} Q n^{2-2\beta} \,,$$

for some positive constant C when $\theta^* \in \Theta(\beta, Q)$. As a result,

$$|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}|_2^2 \lesssim \frac{1}{\alpha(1-\alpha)} Q n^{2-2\beta} + \frac{\sigma^2 M}{1-\alpha} \log(1/\delta) \tag{3.12}$$

for any $t \in (0,1)$. Since, Lemma 3.13 implies, $|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}|_2^2 = n\|\varphi^M_{\hat\theta^{\text{LS}}} - \varphi^M_{\theta^*}\|^2_{L_2([0,1])}$, we have proved the following theorem.

Theorem 3.15. *Fix $\beta \geq (1+\sqrt{5})/4 \simeq 0.81, Q > 0, \delta > 0$ and assume the general regression model (3.1) with $f \in \Theta(\beta, Q)$ and $\varepsilon \sim \mathsf{subG}_n(\sigma^2), \sigma^2 \leq 1$. Moreover, let $M = \lceil n^{\frac{1}{2\beta+1}} \rceil$ and n be large enough so that $M \leq n-1$. Then the least squares estimator $\hat\theta^{\text{LS}}$ defined in (3.2) with $\{\varphi_j\}_{j=1}^M$ being the trigonometric basis, satisfies with probability $1-\delta$, for n large enough,*

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

where the constant factors may depend on β, Q and σ. Moreover

$$\mathbb{E}\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}}\,.$$

Proof. Choosing $\alpha = 1/2$ for example and absorbing Q in the constants, we get from (3.12) and Lemma 3.13 that for $M \leq n-1$,

$$\|\varphi_{\hat\theta^{\text{LS}}} - \varphi_{\theta^*}\|^2_{L_2([0,1])} \lesssim n^{1-2\beta} + \sigma^2 \frac{M + \log(1/\delta)}{n}\,.$$

Using now Lemma 3.14 and $\sigma^2 \leq 1$, we get

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim M^{-2\beta} + n^{1-2\beta} + \frac{M + \sigma^2 \log(1/\delta)}{n}\,.$$

Taking $M = \lceil n^{\frac{1}{2\beta+1}} \rceil \leq n-1$ for n large enough yields

$$\|\varphi_{\hat\theta^{\text{LS}}} - f\|^2_{L_2([0,1])} \lesssim n^{-\frac{2\beta}{2\beta+1}} + n^{1-2\beta} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

To conclude the proof, simply note that for the prescribed β, we have $n^{1-2\beta} \leq n^{-\frac{2\beta}{2\beta+1}}$. The bound in expectation can be obtained by integrating the tail bound. □

Adaptive estimation

The rate attained by the projection estimator $\varphi_{\hat\theta^{\text{LS}}}$ with $M = \lceil n^{\frac{1}{2\beta+1}} \rceil$ is actually optimal so, in this sense, it is a good estimator. Unfortunately, its implementation requires the knowledge of the smoothness parameter β which is typically unknown, to determine the level M of truncation. The purpose of *adaptive estimation* is precisely to adapt to the unknown β, that is to build an estimator

that does not depend on β and yet, attains a rate of the order of $Cn^{-\frac{2\beta}{2\beta+1}}$ (up to a logarithmic lowdown). To that end, we will use the oracle inequalities for the BIC and Lasso estimator defined in (3.3) and (3.4) respectively. In view of Lemma 3.13, the design matrix Φ actually satisfies the assumption ORT when we work with the trigonometric basis. This has two useful implications:

1. Both estimators are actually thresholding estimators and can therefore be implemented efficiently
2. The condition $\mathsf{INC}(k)$ is automatically satisfied for any $k \geq 1$.

These observations lead to the following corollary.

Corollary 3.16. *Fix $\beta \geq (1+\sqrt{5})/4 \simeq 0.81, Q > 0, \delta > 0$ and n large enough to ensure $n - 1 \geq \lceil n^{\frac{1}{2\beta+1}} \rceil$ assume the general regression model (3.1) with $f \in \Theta(\beta, Q)$ and $\varepsilon \sim \mathsf{subG}_n(\sigma^2), \sigma^2 \leq 1$. Let $\{\varphi_j\}_{j=1}^{n-1}$ be the trigonometric basis. Denote by $\varphi_{\hat{\theta}^{\mathrm{BIC}}}^{n-1}$ (resp. $\varphi_{\hat{\theta}^{\mathcal{L}}}^{n-1}$) the BIC (resp. Lasso) estimator defined in (3.3) (resp. (3.4)) over $\mathbb{R}^{n-1}$ with regularization parameter given by (3.5) (resp. (3.6)). Then $\varphi_{\hat{\theta}}^{n-1}$, where $\hat{\theta} \in \{\hat{\theta}^{\mathrm{BIC}}, \hat{\theta}^{\mathcal{L}}\}$ satisfies with probability $1-\delta$,*

$$\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim n^{-\frac{2\beta}{2\beta+1}} + \sigma^2 \frac{\log(1/\delta)}{n}\,.$$

Moreover,

$$\mathbb{E}\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim \sigma^2 \Big(\frac{\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}\,.$$

where constant factors may depend on β and Q.

Proof. For $\hat{\theta} \in \{\hat{\theta}^{\mathrm{BIC}}, \hat{\theta}^{\mathcal{L}}\}$, adapting the proofs of Theorem 3.4 for the BIC estimator and Theorem 3.5 for the Lasso estimator, for any $\theta \in \mathbb{R}^{n-1}$, with probability $1-\delta$

$$|\varphi_{\hat{\theta}}^{n-1} - f|_2^2 \leq \frac{1+\alpha}{1-\alpha}|\varphi_{\theta}^{n-1} - f|_2^2 + R(|\theta|_0)\,.$$

where

$$R(|\theta|_0) := \frac{C\sigma^2}{\alpha(1-\alpha)}|\theta_0|\log(en) + \frac{C\sigma^2}{\alpha(1-\alpha)}\log(1/\delta)$$

It yields

$$\begin{aligned}|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1}|_2^2 &\leq \frac{2\alpha}{1-\alpha}|\varphi_{\theta}^{n-1} - f|_2^2 + 2(\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1})^\top(\varphi_{\theta}^{n-1} - f) + R(|\theta|_0) \\ &\leq \Big(\frac{2\alpha}{1-\alpha} + \frac{1}{\alpha}\Big)|\varphi_{\theta}^{n-1} - f|_2^2 + \alpha|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta}^{n-1}|_2^2 + R(|\theta|_0)\,,\end{aligned}$$

where we used Young's inequality once again. Choose now $\alpha = 1/2$ and $\theta = \theta_M^*$, where θ_M^* is equal to θ^* on its first M coordinates and 0 otherwise so that $\varphi_{\theta_M^*}^{n-1} = \varphi_{\theta^*}^{M}$. It yields

$$|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta_M^*}^{n-1}|_2^2 \lesssim |\varphi_{\theta_M^*}^{n-1} - f|_2^2 + R(M) \lesssim |\varphi_{\theta_M^*}^{n-1} - \varphi_{\theta^*}^{n-1}|_2^2 + |\varphi_{\theta^*}^{n-1} - f|_2^2 + R(M)$$

Next, it follows from (3.11) that $|\varphi_{\theta^*}^{n-1} - f|_2^2 \lesssim Qn^{2-2\beta}$. Together with Lemma 3.13, it yields

$$\|\varphi_{\hat{\theta}}^{n-1} - \varphi_{\theta_M^*}^{n-1}\|_{L_2([0,1])}^2 \lesssim \|\varphi_{\theta^*}^{n-1} - \varphi_{\theta_M^*}^{n-1}\|_{L_2([0,1])}^2 + Qn^{1-2\beta} + \frac{R(M)}{n}\,.$$

Moreover, using (3.10), we find that

$$\|\varphi_{\hat{\theta}}^{n-1} - f\|_{L_2([0,1])}^2 \lesssim M^{-2\beta} + Qn^{1-2\beta} + \frac{M}{n}\log(en) + \frac{\sigma^2}{n}\log(1/\delta)\,.$$

To conclude the proof, choose $M = \lceil (n/\log n)^{\frac{1}{2\beta+1}} \rceil$ and observe that the choice of β ensures that $n^{1-2\beta} \lesssim M^{-2\beta}$. This yields the high probability bound. The bound in expectation is obtained by integrating the tail. □

While there is sometimes a (logarithmic) price to pay for adaptation, it turns out that the extra logarithmic factor can be removed by a clever use of blocks (see [Tsy09, Chapter 3]). The reason why we get this extra logarithmic factor here is because we use a hammer that's too big. Indeed, BIC and Lasso allow for "holes" in the Fourier decomposition and we use a much weaker version of their potential.

3.3 PROBLEM SET

Problem 3.1. Show that the least-squares estimator $\hat{\theta}^{\mathrm{LS}}$ defined in (3.2) satisfies the following *exact* oracle inequality:

$$\mathbb{E}\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathrm{LS}}}) \le \inf_{\theta \in \mathbb{R}^M} \mathsf{MSE}(\varphi_\theta) + C\sigma^2 \frac{M}{n}$$

for some constant M to be specified.

Problem 3.2. Assume that $\varepsilon \sim \mathsf{subG}_n(\sigma^2)$ and the vectors φ_j are normalized in such a way that $\max_j |\varphi_j|_2 \le \sqrt{n}$. Show that there exists a choice of τ such that the Lasso estimator $\hat{\theta}^{\mathcal{L}}$ with regularization parameter 2τ satisfies the following *exact* oracle inequality:

$$\mathsf{MSE}(\varphi_{\hat{\theta}^{\mathcal{L}}}) \le \inf_{\theta \in \mathbb{R}^M} \left\{ \mathsf{MSE}(\varphi_\theta) + C\sigma|\theta|_1 \sqrt{\frac{\log M}{n}} \right\}$$

with probability at least $1 - M^{-c}$ for some positive constants C, c.

Problem 3.3. Let $\{\varphi_1, \dots, \varphi_M\}$ be a dictionary normalized in such a way that $\max_j |\varphi_j|_2 \le \sqrt{n}$. Show that for any integer k such that $1 \le k \le M$, we have

$$\min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_0 \le 2k}} \mathsf{MSE}(\varphi_\theta) \le \min_{\substack{\theta \in \mathbb{R}^M \\ |\theta|_{w\ell_q} \le 1}} \mathsf{MSE}(\varphi_\theta) + C_q D^2 \frac{\left(k^{\frac{1}{q}} - M^{\frac{1}{q}}\right)^2}{k},$$

where $|\theta|_{w\ell_q}$ denotes the weak ℓ_q norm and $\bar{q}$ is such that $\frac{1}{q} + \frac{1}{\bar{q}} = 1$.

Problem 3.4. Show that the trigonometric basis and the Haar system indeed form an orthonormal system of $L_2([0,1])$.

Problem 3.5. If $f \in \Theta(\beta, Q)$ for $\beta > 1/2$ and $Q > 0$, then f is continuous.

CHAPTER 4

Matrix estimation

Over the past decade or so, matrices have entered the picture of high-dimensional statistics for several reasons. Perhaps the simplest explanation is that they are the most natural extension of vectors. While this is true, and we will see examples where the extension from vectors to matrices is straightforward, matrices have a much richer structure than vectors allowing "interaction" between their rows and columns. In particular, while we have been describing simple vectors in terms of their sparsity, here we can measure the complexity of a matrix by its *rank*. This feature was successfully employed in a variety of applications ranging from *multi-task learning* to *collaborative filtering*. This last application was made popular by the NETFLIX prize in particular.

In this chapter, we study several statistical problems where the parameter of interest θ is a matrix rather than a vector. These problems include: multivariate regression, covariance matrix estimation and principal component analysis. Before getting to these topics, we begin by a quick reminder on matrices and linear algebra.

4.1 BASIC FACTS ABOUT MATRICES

Matrices are much more complicated objects than vectors. In particular, while vectors can be identified with linear operators from $\mathbb{R}^d$ to $\mathbb{R}$, matrices can be identified to linear operators from $\mathbb{R}^d$ to $\mathbb{R}^n$ for $n \geq 1$. This seemingly simple fact gives rise to a profusion of notions and properties as illustrated by Bernstein's book [Ber09] that contains facts about matrices over more than a thousand pages. Fortunately, we will be needing only a small number of such properties, which can be found in the excellent book [GVL96], that has become a standard reference on matrices and numerical linear algebra.

Singular value decomposition

Let $A = \{a_{ij}, 1 \le i \le m, 1 \le j \le n\}$ be a $m \times n$ real matrix of rank $r \le \min(m,n)$. The *Singular Value Decomposition* (SVD) of A is given by

$$A = UDV^\top = \sum_{j=1}^{r} \lambda_j u_j v_j^\top ,$$

where D is a $r \times r$ diagonal matrix with positive diagonal entries $\{\lambda_1, \dots, \lambda_r\}$, U is a matrix with columns $\{u_1, \dots, u_r\} \in \mathbb{R}^m$ that are orthonormal and V is a matrix with columns $\{v_1, \dots, v_r\} \in \mathbb{R}^n$ that are also orthonormal. Moreover, it holds that

$$AA^\top u_j = \lambda_j^2 u_j , \qquad \text{and} \qquad A^\top A v_j = \lambda_j^2 v_j$$

for $j = 1, \dots, r$. The values $\lambda_j > 0$ are called *singular values* of A and are uniquely defined. If rank $r < \min(n,m)$ then the singular values of A are given by $\lambda = (\lambda_1, \dots, \lambda_r, 0, \dots, 0)^\top \in \mathbb{R}^{\min(n,m)}$ where there are $\min(n,m) - r$ zeros. This way, the vector λ of singular values of a $n \times m$ matrix is a vector in $\mathbb{R}^{\min(n,m)}$.

In particular, if A is a $n \times n$ symmetric positive semidefinite (PSD), i.e. $A^\top = A$ and $u^\top A u \ge 0$ for all $u \in \mathbb{R}^n$, then the singular values of A are equal to its eigenvalues.

The largest singular value of A denoted by $\lambda_{\max}(A)$ also satisfies the following variational formulation:

$$\lambda_{\max}(A) = \max_{x \in \mathbb{R}^n} \frac{|Ax|_2}{|x|_2} = \max_{\substack{x \in \mathbb{R}^n \\ y \in \mathbb{R}^m}} \frac{y^\top A x}{|y|_2 |x|_2} = \max_{\substack{x \in \mathcal{S}^{n-1} \\ y \in \mathcal{S}^{m-1}}} y^\top A x .$$

In the case of a $n \times n$ PSD matrix A, we have

$$\lambda_{\max}(A) = \max_{x \in \mathcal{S}^{n-1}} x^\top A x .$$

Norms and inner product

Let $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ be two real matrices. Their size will be implicit in the following notation.

Vector norms

The simplest way to treat a matrix is to deal with it as if it were a vector. In particular, we can extend ℓ_q norms to matrices:

$$|A|_q = \Big(\sum_{ij} |a_{ij}|^q \Big)^{1/q} , \quad q > 0 .$$

The cases where $q \in \{0, \infty\}$ can also be extended matrices:

$$|A|_0 = \sum_{ij} \mathbb{1}(a_{ij} \neq 0) , \qquad |A|_\infty = \max_{ij} |a_{ij}| .$$

The case $q = 2$ plays a particular role for matrices and $|A|_2$ is called the *Frobenius* norm of A and is often denoted by $\|A\|_F$. It is also the Hilbert-Schmidt norm associated to the inner product:

$$\langle A, B\rangle = \mathsf{Tr}(A^\top B) = \mathsf{Tr}(B^\top A)\,.$$

Spectral norms

Let $\lambda = (\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$ be the singular values of a matrix A. We can define spectral norms on A as vector norms on the vector λ. In particular, for any $q \in [1, \infty]$,

$$\|A\|_q = |\lambda|_q\,,$$

is called *Schatten q-norm* of A. Here again, special cases have special names:

- $q = 2$: $\|A\|_2 = \|A\|_F$ is the Frobenius norm defined above.
- $q = 1$: $\|A\|_1 = \|A\|_*$ is called the Nuclear norm (or trace norm) of A.
- $q = \infty$: $\|A\|_\infty = \lambda_{\max}(A) = \|A\|_{\text{op}}$ is called the operator norm (or spectral norm) of A.

We are going to employ these norms to assess the proximity to our matrix of interest. While the interpretation of vector norms is clear by extension from the vector case, the meaning of "$\|A-B\|_{\text{op}}$ is small" is not as transparent. The following subsection provides some inequalities (without proofs) that allow a better reading.

Useful matrix inequalities

Let A and B be two $m \times n$ matrices with singular values $\lambda_1(A) \geq \lambda_2(A) \ldots \geq \lambda_{\min(m,n)}(A)$ and $\lambda_1(B) \geq \ldots \geq \lambda_{\min(m,n)}(B)$ respectively. Then the following inequalities hold:

$$\max_k \big|\lambda_k(A) - \lambda_k(B)\big| \leq \|A - B\|_{\text{op}}\,, \qquad \text{Weyl (1912)}$$

$$\sum_k \big|\lambda_k(A) - \lambda_k(B)\big|^2 \leq \|A - B\|_F^2\,, \qquad \text{Hoffman-Weilandt (1953)}$$

$$\langle A, B\rangle \leq \|A\|_q \|B\|_q\,,\ \frac{1}{p} + \frac{1}{q} = 1, p, q \in [1, \infty]\,, \qquad \text{Hölder}$$

4.2 MULTIVARIATE REGRESSION

In the traditional regression setup, the response variable Y is a scalar. In several applications, the goal is not to predict a variable but rather a vector $Y \in \mathbb{R}^T$, still from a covariate $X \in \mathbb{R}^d$. A standard example arises in genomics data where Y contains T physical measurements of a patient and X contains

the expression levels for d genes. As a result the regression function in this case $f(x) = \mathbb{E}[Y|X = x]$ is a function from $\mathbb{R}^d$ to $\mathbb{R}^T$. Clearly, f can be estimated independently for each coordinate, using the tools that we have developed in the previous chapter. However, we will see that in several interesting scenarios, some structure is shared across coordinates and this information can be leveraged to yield better prediction bounds.

The model

Throughout this section, we consider the following multivariate linear regression model:

$$\mathbb{Y} = \mathbb{X}\Theta^* + E\,, \tag{4.1}$$

where $\mathbb{Y} \in \mathbb{R}^{n\times T}$ is the matrix of observed responses, $\mathbb{X}$ is the $n \times d$ observed design matrix (as before), $\Theta \in \mathbb{R}^{d\times T}$ is the matrix of unknown parameters and $E \sim \mathsf{subG}_{n\times T}(\sigma^2)$ is the noise matrix. In this chapter, we will focus on the prediction task, which consists in estimating $\mathbb{X}\Theta^*$.

As mentioned in the foreword of this chapter, we can view this problem as T (univariate) linear regression problems $Y^{(j)} = \mathbb{X}\theta^{*,(j)} + \varepsilon^{(j)}, j = 1, \dots, T$, where $Y^{(j)}, \theta^{*,(j)}$ and $\varepsilon^{(j)}$ are the jth column of $\mathbb{Y}, \Theta^*$ and E respectively. In particular, an estimator for $\mathbb{X}\Theta^*$ can be obtained by concatenating the estimators for each of the T problems. This approach is the subject of Problem 4.1.

The columns of Θ^* correspond to T different regression tasks. Consider the following example as a motivation. Assume that the Subway headquarters want to evaluate the effect of d variables (promotions, day of the week, TV ads,...) on their sales. To that end, they ask each of their $T = 40,000$ restaurants to report their sales numbers for the past $n = 200$ days. As a result, franchise j returns to headquarters a vector $\mathbb{Y}^{(j)} \in \mathbb{R}^n$. The d variables for each of the n days are already known to headquarters and are stored in a matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$. In this case, it may be reasonable to assume that the same subset of variables has an impact of the sales for each of the franchise, though the magnitude of this impact may differ from franchise to franchise. As a result, one may assume that the matrix Θ^* has each of its T columns that is row sparse and that they *share the same sparsity pattern*, i.e., Θ^* is of the form:

$$\Theta^* = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & \bullet \end{pmatrix},$$

where $\bullet$ indicates a potentially nonzero entry.

It follows from the result of Problem 4.1 that if each task is performed individually, one may find an estimator $\hat{\Theta}$ such that

$$\frac{1}{n}\mathbb{E}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{kT\log(ed)}{n}\,,$$

where k is the number of nonzero coordinates in each column of Θ^*. We remember that the term $\log(ed)$ corresponds to the additional price to pay for not knowing where the nonzero components are. However, in this case, when the number of tasks grows, this should become easier. This fact was proved in [LPTVDG11]. We will see that we can recover a similar phenomenon when the number of tasks becomes large, though larger than in [LPTVDG11]. Indeed, rather than exploiting sparsity, observe that such a matrix Θ^* has rank k. This is the kind of structure that we will be predominantly using in this chapter.

Rather than assuming that the columns of Θ^* share the same sparsity pattern, it may be more appropriate to assume that the matrix Θ^* is low rank or approximately so. As a result, while the matrix may not be sparse at all, the fact that it is low rank still materializes the idea that some structure is shared across different tasks. In this more general setup, it is assumed that the columns of Θ^* live in a lower dimensional space. Going back to the SUBWAY example this amounts to assuming that while there are 40,000 franchises, there are only a few canonical profiles for these franchises and that all franchises are linear combinations of these profiles.

Sub-Gaussian matrix model

Recall that under the assumption ORT for the design matrix, i.e., $\mathbb{X}^\top\mathbb{X} = nI_d$, then the univariate regression model can be reduced to the sub-Gaussian sequence model. Here we investigate the effect of this assumption on the multivariate regression model (4.1).

Observe that under assumption ORT,

$$\frac{1}{n}\mathbb{X}^\top\mathbb{Y} = \Theta^* + \frac{1}{n}\mathbb{X}^\top E\,.$$

Which can be written as an equation in $\mathbb{R}^{d\times T}$ called the *sub-Gaussian matrix model (sGMM)*:

$$y = \Theta^* + F\,, \tag{4.2}$$

where $y = \frac{1}{n}\mathbb{X}^\top\mathbb{Y}$ and $F = \frac{1}{n}\mathbb{X}^\top E \sim \mathsf{subG}_{d\times T}(\sigma^2/n)$.

Indeed, for any $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$, it holds

$$u^\top F v = \frac{1}{n}(\mathbb{X}u)^\top E v = \frac{1}{\sqrt{n}} w^\top E v \sim \mathsf{subG}(\sigma^2/n)\,,$$

where $w = \mathbb{X}u/\sqrt{n}$ has unit norm: $|w|_2^2 = u^\top \frac{\mathbb{X}^\top\mathbb{X}}{n} u = |u|_2^2 = 1$.

Akin to the sub-Gaussian sequence model, we have a *direct* observation model where we observe the parameter of interest with additive noise. This

enables us to use thresholding methods for estimating Θ^* when $|\Theta^*|_0$ is small. However, this also follows from Problem 4.1. The reduction to the vector case in the sGMM is just as straightforward. The interesting analysis begins when Θ^* is low-rank, which is equivalent to sparsity in its unknown eigenbasis.

Consider the SVD of Θ^*:

$$\Theta^* = \sum_j \lambda_j u_j v_j^\top .$$

and recall that $\|\Theta^*\|_0 = |\lambda|_0$. Therefore, if we knew u_j and v_j, we could simply estimate the λ_js by hard thresholding. It turns out that estimating these eigenvectors by the eigenvectors of y is sufficient.

Consider the SVD of the observed matrix y:

$$y = \sum_j \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top .$$

Definition 4.1. The **singular value thresholding** estimator with threshold $2\tau \geq 0$ is defined by

$$\hat{\Theta}^{\text{SVT}} = \sum_j \hat{\lambda}_j \mathbb{1}(|\hat{\lambda}_j| > 2\tau)\hat{u}_j \hat{v}_j^\top .$$

Recall that the threshold for the hard thresholding estimator was chosen to be the level of the noise with high probability. The singular value thresholding estimator obeys the same rule, except that the norm in which the magnitude of the noise is measured is adapted to the matrix case. Specifically, the following lemma will allow us to control the operator norm of the matrix F.

Lemma 4.2. *Let A be a $d \times T$ random matrix such that $A \sim \mathsf{subG}_{d\times T}(\sigma^2)$. Then*

$$\|A\|_{\text{op}} \leq 4\sigma\sqrt{\log(12)(d \vee T)} + 2\sigma\sqrt{2\log(1/\delta)}$$

with probability $1 - \delta$.

Proof. This proof follows the same steps as Problem 1.4. Let $\mathcal{N}_1$ be a 1/4-net for $\mathcal{S}^{d-1}$ and $\mathcal{N}_2$ be a 1/4-net for $\mathcal{S}^{T-1}$. It follows from Lemma 1.18 that we can always choose $|\mathcal{N}_1| \leq 12^d$ and $|\mathcal{N}_2| \leq 12^T$. Moreover, for any $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$, it holds

$$\begin{aligned}
u^\top A v &\leq \max_{x\in\mathcal{N}_1} x^\top A v + \frac{1}{4}\max_{u\in\mathcal{S}^{d-1}} u^\top A v \\
&\leq \max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y + \frac{1}{4}\max_{x\in\mathcal{N}_1}\max_{v\in\mathcal{S}^{T-1}} x^\top A v + \frac{1}{4}\max_{u\in\mathcal{S}^{d-1}} u^\top A v \\
&\leq \max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y + \frac{1}{2}\max_{u\in\mathcal{S}^{d-1}}\max_{v\in\mathcal{S}^{T-1}} u^\top A v
\end{aligned}$$

It yields

$$\|A\|_{\text{op}} \leq 2\max_{x\in\mathcal{N}_1}\max_{y\in\mathcal{N}_2} x^\top A y$$

So that for any $t \geq 0$, by a union bound,

$$\mathbb{P}\big(\|A\|_{\mathrm{op}} > t\big) \leq \sum_{\substack{x \in \mathcal{N}_1 \\ y \in \mathcal{N}_2}} \mathbb{P}\big(x^\top A y > t/2\big)$$

Next, since $A \sim \mathsf{subG}_{d\times T}(\sigma^2)$, it holds that $x^\top A y \sim \mathsf{subG}(\sigma^2)$ for any $x \in \mathcal{N}_1, y \in \mathcal{N}_2$.Together with the above display, it yields

$$\mathbb{P}\big(\|A\|_{\mathrm{op}} > t\big) \leq 12^{d+T} \exp\Big(-\frac{t^2}{8\sigma^2}\Big) \leq \delta$$

for

$$t \geq 4\sigma\sqrt{\log(12)(d \vee T)} + 2\sigma\sqrt{2\log(1/\delta)}\,.$$

□

The following theorem holds.

Theorem 4.3. *Consider the multivariate linear regression model (4.1) under the assumption* ***ORT*** *or, equivalently, the sub-Gaussian matrix model (4.2). Then, the singular value thresholding estimator $\hat{\Theta}^{\mathrm{SVT}}$ with threshold*

$$2\tau = 8\sigma\sqrt{\frac{\log(12)(d \vee T)}{n}} + 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\,, \tag{4.3}$$

satisfies

$$\frac{1}{n}\|\mathbb{X}\hat{\Theta}^{\mathrm{SVT}} - \mathbb{X}\Theta^*\|_F^2 = \|\hat{\Theta}^{\mathrm{SVT}} - \Theta^*\|_F^2 \leq 144\,\mathrm{rank}(\Theta^*)\tau^2 \lesssim \frac{\sigma^2\,\mathrm{rank}(\Theta^*)}{n}\Big(d \vee T + \log(1/\delta)\Big)\,.$$

with probability $1-\delta$.

Proof. Assume without loss of generality that the singular values of Θ^* and y are arranged in a non increasing order: $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$. Define the set $S = \{j : |\hat{\lambda}_j| > 2\tau\}$.

Observe first that it follows from Lemma 4.2 that $\|F\|_{\mathrm{op}} \leq \tau$ for τ chosen as in (4.3) on an event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$. The rest of the proof is on $\mathcal{A}$.

Note that it follows from Weyl's inequality that $|\hat{\lambda}_j - \lambda_j| \leq \|F\|_{\mathrm{op}} \leq \tau$. It implies that $S \subset \{j : |\lambda_j| > \tau\}$ and $S^c \subset \{j : |\lambda_j| \leq 3\tau\}$.

Next define the oracle $\bar{\Theta} = \sum_{j\in S} \lambda_j u_j v_j^\top$ and note that

$$\|\hat{\Theta}^{\mathrm{SVT}} - \Theta^*\|_F^2 \leq 2\|\hat{\Theta}^{\mathrm{SVT}} - \bar{\Theta}\|_F^2 + 2\|\bar{\Theta} - \Theta^*\|_F^2 \tag{4.4}$$

Using Cauchy-Schwarz, we control the first term as follows

$$\|\hat{\Theta}^{\mathrm{SVT}} - \bar{\Theta}\|_F^2 \leq \mathrm{rank}(\hat{\Theta}^{\mathrm{SVT}} - \bar{\Theta})\|\hat{\Theta}^{\mathrm{SVT}} - \bar{\Theta}\|_{\mathrm{op}}^2 \leq 2|S|\|\hat{\Theta}^{\mathrm{SVT}} - \bar{\Theta}\|_{\mathrm{op}}^2$$

Moreover,

$$\begin{aligned}\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\|_{\text{op}} &\leq \|\hat{\Theta}^{\text{SVT}} - y\|_{\text{op}} + \|y - \Theta^*\|_{\text{op}} + \|\Theta^* - \bar{\Theta}\|_{\text{op}} \\ &\leq \max_{j \in S^c} |\hat{\lambda}_j| + \tau + \max_{j \in S^c} |\lambda_j| \leq 6\tau\,.\end{aligned}$$

Therefore,

$$\|\hat{\Theta}^{\text{SVT}} - \bar{\Theta}\|_F^2 \leq 72|S|\tau^2 = 72\sum_{j \in S}\tau^2\,.$$

The second term in (4.4) can be written as

$$\|\bar{\Theta} - \Theta^*\|_F^2 = \sum_{j \in S^c} |\lambda_j|^2\,.$$

Plugging the above two displays in (4.4), we get

$$\|\hat{\Theta}^{\text{SVT}} - \Theta^*\|_F^2 \leq 144\sum_{j \in S}\tau^2 + \sum_{j \in S^c}|\lambda_j|^2$$

Since on S, $\tau^2 = \min(\tau^2, |\lambda_j|^2)$ and on S^c, $|\lambda_j|^2 \leq 3\min(\tau^2, |\lambda_j|^2)$, it yields,

$$\begin{aligned}\|\hat{\Theta}^{\text{SVT}} - \Theta^*\|_F^2 &\leq 432\sum_j \min(\tau^2, |\lambda_j|^2) \\ &\leq 432\sum_{j=1}^{\text{rank}(\Theta^*)}\tau^2 \\ &= 432\,\text{rank}(\Theta^*)\tau^2\,.\end{aligned}$$

□

In the next subsection, we extend our analysis to the case where $\mathbb{X}$ does not necessarily satisfy the assumption ORT.

Penalization by rank

The estimator from this section is the counterpart of the BIC estimator in the spectral domain. However, we will see that unlike BIC, it can be computed efficiently.

Let $\hat{\Theta}^{\text{RK}}$ be any solution to the following minimization problem:

$$\min_{\Theta \in \mathbb{R}^{d \times T}} \left\{\frac{1}{n}\|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + 2\tau^2\,\text{rank}(\Theta)\right\}.$$

This estimator is called *estimator by rank penalization with regularization parameter* τ^2. It enjoys the following property.

Theorem 4.4. *Consider the multivariate linear regression model (4.1). Then, the estimator by rank penalization $\hat{\Theta}^{\mathrm{RK}}$ with regularization parameter τ^2, where τ is defined in (4.3) satisfies*

$$\frac{1}{n}\|\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 8\,\mathrm{rank}(\Theta^*)\tau^2 \lesssim \frac{\sigma^2\,\mathrm{rank}(\Theta^*)}{n}\Big(d \vee T + \log(1/\delta)\Big)\,.$$

with probability $1-\delta$.

Proof. We begin as usual by noting that

$$\|\mathbb{Y} - \mathbb{X}\hat{\Theta}^{\mathrm{RK}}\|_F^2 + 2n\tau^2\,\mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) \le \|\mathbb{Y} - \mathbb{X}\Theta^*\|_F^2 + 2n\tau^2\,\mathrm{rank}(\Theta^*)\,,$$

which is equivalent to

$$\|\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 2\langle E, \mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\rangle - 2n\tau^2\,\mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) + 2n\tau^2\,\mathrm{rank}(\Theta^*)\,.$$

Next, by Young's inequality, we have

$$2\langle E, \mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\rangle = 2\langle E, U\rangle^2 + \frac{1}{2}\|\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\|_F^2\,,$$

where

$$U = \frac{\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*}{\|\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\|_F}\,.$$

Write

$$\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^* = \Phi N\,,$$

where Φ is a $n \times r, r \le d$ matrix whose columns form orthonormal basis of the column span of $\mathbb{X}$. The matrix Φ can come from the SVD of $\mathbb{X}$ for example: $\mathbb{X} = \Phi\Lambda\Psi^\top$. It yields

$$U = \frac{\Phi N}{\|N\|_F}$$

and

$$\|\mathbb{X}\hat{\Theta}^{\mathrm{RK}} - \mathbb{X}\Theta^*\|_F^2 \le 4\langle \Phi^\top E, N/\|N\|_F\rangle^2 - 4n\tau^2\,\mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) + 4n\tau^2\,\mathrm{rank}(\Theta^*)\,. \tag{4.5}$$

Note that $\mathrm{rank}(N) \le \mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) + \mathrm{rank}(\Theta^*)$. Therefore, by Hölder's inequality, we get

$$\begin{aligned}
\langle E, U\rangle^2 &= \langle \Phi^\top E, N/\|N\|_F\rangle^2 \\
&\le \|\Phi^\top E\|_{\mathrm{op}}^2 \frac{\|N\|_1^2}{\|N\|_F^2} \\
&\le \mathrm{rank}(N)\|\Phi^\top E\|_{\mathrm{op}}^2 \\
&\le \|\Phi^\top E\|_{\mathrm{op}}^2\big[\mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) + \mathrm{rank}(\Theta^*)\big]\,.
\end{aligned}$$

Next, note that Lemma 4.2 yields $\|\Phi^\top E\|_{\mathrm{op}}^2 \le n\tau^2$ so that

$$\langle E, U\rangle^2 \le n\tau^2\big[\mathrm{rank}(\hat{\Theta}^{\mathrm{RK}}) + \mathrm{rank}(\Theta^*)\big]\,.$$

Together with (4.5), this completes the proof. □

It follows from Theorem 4.4 that the estimator by rank penalization enjoys the same properties as the singular value thresholding estimator even when $\mathbb{X}$ does not satisfies the ORT condition. This is reminiscent of the BIC estimator which enjoys the same properties as the hard thresholding estimator. However this analogy does not extend to computational questions. Indeed, while the rank penalty, just like the sparsity penalty, is not convex, it turns out that $\mathbb{X}\hat{\Theta}^{\text{RK}}$ can be computed efficiently.

Note first that

$$\min_{\Theta\in\mathbb{R}^{d\times T}} \frac{1}{n}\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 + 2\tau^2\,\mathrm{rank}(\Theta) = \min_k \left\{ \frac{1}{n} \min_{\substack{\Theta\in\mathbb{R}^{d\times T}\\ \mathrm{rank}(\Theta)\le k}} \|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 + 2\tau^2 k \right\}.$$

Therefore, it remains to show that

$$\min_{\substack{\Theta\in\mathbb{R}^{d\times T}\\ \mathrm{rank}(\Theta)\le k}} \|\mathbb{Y}-\mathbb{X}\Theta\|_F^2$$

can be solved efficiently. To that end, let $\bar{\mathbb{Y}} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^\dagger\mathbb{X}^\top\mathbb{Y}$ denote the orthogonal projection of $\mathbb{Y}$ onto the image space of $\mathbb{X}$: this is a linear operator from $\mathbb{R}^{d\times T}$ into $\mathbb{R}^{n\times T}$. By the Pythagorean theorem, we get for any $\Theta\in\mathbb{R}^{d\times T}$,

$$\|\mathbb{Y}-\mathbb{X}\Theta\|_F^2 = \|\mathbb{Y}-\bar{\mathbb{Y}}\|_F^2 + \|\bar{\mathbb{Y}}-\mathbb{X}\Theta\|_F^2\,.$$

Next consider the SVD of $\bar{\mathbb{Y}}$:

$$\bar{\mathbb{Y}} = \sum_j \lambda_j u_j v_j^\top$$

where $\lambda_1\ge\lambda_2\ge\ldots$. The claim is that if we define $\tilde{\mathbb{Y}}$ by

$$\tilde{\mathbb{Y}} = \sum_{j=1}^k \lambda_j u_j v_j^\top$$

which is clearly of rank at most k, then it satisfies

$$\|\bar{\mathbb{Y}}-\tilde{\mathbb{Y}}\|_F^2 = \min_{Z:\mathrm{rank}(Z)\le k} \|\bar{\mathbb{Y}}-Z\|_F^2\,.$$

Indeed,

$$\|\bar{\mathbb{Y}}-\tilde{\mathbb{Y}}\|_F^2 = \sum_{j>k}\lambda_j^2\,,$$

and for any matrix Z such that $\mathrm{rank}(Z)\le k$ with SVD

$$Z = \sum_{j=1}^k \mu_j x_j y_j^\top,$$

where $\mu_1\ge\mu_2\ge\ldots$, we have by Hoffman-Weilandt

$$\|Z-\bar{Y}\|_F^2 \ge \sum_{j\ge1}|\lambda_j-\mu_j|^2 \ge \sum_{j>k}\lambda_j^2\,.$$

Therefore, any minimizer of $\mathbb{X}\Theta \mapsto \|\mathbb{Y} - \mathbb{X}\Theta\|_F^2$ over matrices of rank at most k can be obtained by truncating the SVD of $\mathbb{Y}$ at order k.

Once $\mathbb{X}\hat{\Theta}^{\mathrm{RK}}$ has been found, one may obtain a corresponding $\hat{\Theta}^{\mathrm{RK}}$ by least squares but this is not necessary for our results.

Remark 4.5. While the rank penalized estimator can be computed efficiently, it is worth pointing out that a convex relaxation for the rank penalty can also be used. The estimator by nuclear norm penalization $\hat{\Theta}$ is defined to be any solution to the minimization problem

$$\min_{\Theta \in \mathbb{R}^{d \times T}} \Big\{ \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + \tau \|\Theta\|_1 \Big\}$$

Clearly this criterion is convex and it can actually be implemented efficiently using semi-definite programming. It has been popularized by matrix completion problems. Let $\mathbb{X}$ have the following SVD:

$$\mathbb{X} = \sum_{j=1}^{r} \lambda_j u_j v_j^\top ,$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$. It can be shown that for some appropriate choice of τ, it holds

$$\frac{1}{n} \|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\lambda_1}{\lambda_r} \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n} d \vee T$$

with probability .99. However, the proof of this result is far more involved than a simple adaption of the proof for the Lasso estimator to the matrix case (the readers are invited to see that for themselves). For one thing, there is no assumption on the design matrix (such as INC for example). This result can be found in [KLT11].

4.3 COVARIANCE MATRIX ESTIMATION

Empirical covariance matrix

Let $X_1, \ldots, X_n$ be n i.i.d. copies of a random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}\big[XX^\top\big] = \Sigma$ for some unknown matrix $\Sigma \succ 0$ called *covariance matrix*. This matrix contains information about the moments of order 2 of the random vector X. A natural candidate to estimate Σ is the *empirical covariance matrix* $\hat{\Sigma}$ defined by

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top .$$

Using the tools of Chapter 1, we can prove the following result.

Theorem 4.6. *Let $X_1, \ldots, X_n$ be n i.i.d. sub-Gaussian random vectors such that $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Then*

$$\|\hat{\Sigma} - \Sigma\|_{\mathrm{op}} \lesssim \|\Sigma\|_{\mathrm{op}} \Big(\sqrt{\frac{d + \log(1/\delta)}{n}} \vee \frac{d + \log(1/\delta)}{n} \Big) ,$$

with probability $1-\delta$.

Proof. Observe first that without loss of generality we can assume that $\Sigma = I_d$. Indeed, note that since $\mathbb{E}\big[XX^\top\big] = \Sigma \succ 0$, then $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Moreover, $Y = \Sigma^{-1/2}X \sim \mathsf{subG}_d(1)$ and $\mathbb{E}[YY^\top] = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_d$. Therefore,

$$\begin{aligned}\frac{\|\hat{\Sigma}-\Sigma\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} &= \frac{\|\frac{1}{n}\sum_{i=1}^n X_iX_i^\top - \Sigma\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \\ &\le \frac{\|\Sigma^{1/2}\|_{\mathrm{op}}\|\frac{1}{n}\sum_{i=1}^n Y_iY_i^\top - I_d\|_{\mathrm{op}}\|\Sigma^{1/2}\|_{\mathrm{op}}}{\|\Sigma\|_{\mathrm{op}}} \\ &= \|\frac{1}{n}\sum_{i=1}^n Y_iY_i^\top - I_d\|_{\mathrm{op}}\,.\end{aligned}$$

Let $\mathcal{N}$ be a $1/4$-net for $\mathcal{S}^{d-1}$ such that $|\mathcal{N}| \le 12^d$. It follows from the proof of Lemma 4.2 that

$$\|\hat{\Sigma} - I_d\|_{\mathrm{op}} \le 2\max_{x,y\in\mathcal{N}} x^\top(\hat{\Sigma}-I_d)y$$

So that for any $t \ge 0$, by a union bound,

$$\mathbb{P}\big(\|\hat{\Sigma}-I_d\|_{\mathrm{op}} > t\big) \le \sum_{x,y\in\mathcal{N}} \mathbb{P}\big(x^\top(\hat{\Sigma}-I_d)y > t/2\big)\,. \tag{4.6}$$

It holds,

$$x^\top(\hat{\Sigma}-I_d)y = \frac{1}{n}\sum_{i=1}^n \big\{(X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big]\big\}\,.$$

Using polarization, we also have

$$(X_i^\top x)(X_i^\top y) = \frac{Z_+^2 - Z_-^2}{4}\,,$$

here $Z_+ = X_i^\top(x+y)$ and $Z_- = X_i^\top(x-y)$. It yields

$$\begin{aligned}&\mathbb{E}\big[\exp\big(s\big((X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big]\big)\big)\big] \\ &\quad= \mathbb{E}\big[\exp\big(\frac{s}{4}\big(Z_+^2 - \mathbb{E}[Z_+^2]\big) - \frac{s}{4}\big(Z_-^2 - \mathbb{E}[Z_-^2]\big)\big)\big] \\ &\quad\le \Big(\mathbb{E}\big[\exp\big(\frac{s}{2}\big(Z_+^2 - \mathbb{E}[Z_+^2]\big)\big)\big]\mathbb{E}\big[\exp\big(-\frac{s}{2}\big(Z_-^2 - \mathbb{E}[Z_-^2]\big)\big)\big]\Big)^{1/2}\,,\end{aligned}$$

where in the last inequality, we used Cauchy-Schwarz. Next, since $X \sim \mathsf{subG}_d(1)$, we have $Z_+, Z_- \sim \mathsf{subG}(2)$, and it follows from Lemma 1.12 that

$$Z_+^2 - \mathbb{E}[Z_+^2] \sim \mathsf{subE}(32)\,, \qquad \text{and} \qquad Z_-^2 - \mathbb{E}[Z_-^2] \sim \mathsf{subE}(32)$$

Therefore for any $s \le 1/16$, we have for any $Z \in \{Z_+, Z_-\}$, we have

$$\mathbb{E}\big[\exp\big(\frac{s}{2}\big(Z^2 - \mathbb{E}[Z^2]\big)\big)\big] \le e^{128s^2}\,,$$

It yields that

$$(X_i^\top x)(X_i^\top y) - \mathbb{E}\big[(X_i^\top x)(X_i^\top y)\big] \sim \mathsf{subE}(16)\,.$$

Applying now Bernstein's inequality (Theorem 1.13), we get

$$\mathbb{P}\big(x^\top(\hat\Sigma - I_d)y > t/2\big) \le \exp\Big[-\frac{n}{2}\big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\big)\Big]\,.$$

Together with (4.6), this yields

$$\mathbb{P}\big(\|\hat\Sigma - I_d\|_{\mathrm{op}} > t\big) \le 144^d \exp\Big[-\frac{n}{2}\big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\big)\Big]\,. \tag{4.7}$$

In particular, the right hand side of the above inequality is at most $\delta \in (0,1)$ if

$$\frac{t}{32} \ge \Big(\frac{2d}{n}\log(144) + \frac{2}{n}\log(1/\delta)\Big) \vee \Big(\frac{2d}{n}\log(144) + \frac{2}{n}\log(1/\delta)\Big)^{1/2}$$

This concludes our proof. □

Theorem 4.6 indicates that for fixed d, the empirical covariance matrix is a consistent estimator of Σ (in any norm as they are all equivalent in finite dimension). However, the bound that we got is not satisfactory in high-dimensions when $d \gg n$. To overcome this limitation, we can introduce sparsity as we have done in the case of regression. The most obvious way to do so is to assume that few of the entries of Σ are non zero and it turns out that in this case thresholding is optimal. There is a long line of work on this subject (see for example [CZZ10] and [CZ12]).

Once we have a good estimator of Σ, what can we do with it? The key insight is that Σ contains information about the projection of the vector X onto *any* direction $u \in \mathcal{S}^{d-1}$. Indeed, we have that $\mathrm{var}(X^\top u) = u^\top \Sigma u$, which can be readily estimated by $\widehat{\mathrm{Var}}(X^\top u) = u^\top \hat\Sigma u$. Observe that it follows from Theorem 4.6

$$\begin{aligned}\big|\widehat{\mathrm{Var}}(X^\top u) - \mathrm{Var}(X^\top u)\big| &= \big|u^\top(\hat\Sigma - \Sigma)u\big| \\ &\le \|\hat\Sigma - \Sigma\|_{\mathrm{op}} \\ &\lesssim \|\Sigma\|_{\mathrm{op}}\Big(\sqrt{\frac{d+\log(1/\delta)}{n}} \vee \frac{d+\log(1/\delta)}{n}\Big)\end{aligned}$$

with probability $1-\delta$.

The above fact is useful in the Markowitz theory of portfolio section for example [Mar52], where a portfolio of assets is a vector $u \in \mathbb{R}^d$ such that $|u|_1 = 1$ and the risk of a portfolio is given by the variance $\mathrm{Var}(X^\top u)$. The goal is then to maximize reward subject to risk constraints. In most instances, the empirical covariance matrix is plugged into the formula in place of Σ.

4.4 PRINCIPAL COMPONENT ANALYSIS

Spiked covariance model

Estimating the variance in all directions is also useful for *dimension reduction*. In *Principal Component Analysis (PCA)*, the goal is to find one (or more) directions onto which the data $X_1, \dots, X_n$ can be projected without loosing much of its properties. There are several goals for doing this but perhaps the most prominent ones are data visualization (in few dimensions, one can plot and visualize the cloud of n points) and clustering (clustering is a hard computational problem and it is therefore preferable to carry it out in lower dimensions). An example of the output of a principal component analysis is given in Figure 4.1. In this figure, the data has been projected onto two orthogonal directions PC1 and PC2, that were estimated to have the most variance (among all such orthogonal pairs). The idea is that when projected onto such directions, points will remain far apart and a clustering pattern will still emerge. This is the case in Figure 4.1 where the original data is given by $d = 500,000$ gene expression levels measured on $n \simeq 1,387$ people. Depicted are the projections of these $1,387$ points in two dimension. This image has become quite popular as it shows that gene expression levels can recover the structure induced by geographic clustering. How is it possible to "compress" half a million dimensions into only two? The answer is that the data is intrinsically low dimensional. In this case, a plausible assumption is that all the $1,387$ points live close to a two-dimensional linear subspace. To see how this assumption (in one dimension instead of two for simplicity) translates into the structure of the covariance matrix Σ, assume that $X_1, \dots, X_n$ are Gaussian random variables generated as follows. Fix a direction $v \in \mathcal{S}^{d-1}$ and let $Y_1, \dots, Y_n \sim \mathcal{N}_d(0, I_d)$ so that $v^\top Y_i$ are i.i.d. $\mathcal{N}(0,1)$. In particular, the vectors $(v^\top Y_1)v, \dots, (v^\top Y_n)v$ live in the one-dimensional space spanned by v. If one would observe such data the problem would be easy as only two observations would suffice to recover v. Instead, we observe $X_1, \dots, X_n \in \mathbb{R}^d$ where $X_i = (v^\top Y_i)v + Z_i$, and $Z_i \sim \mathcal{N}_d(0, \sigma^2 I_d)$ are i.i.d. and independent of the Y_is, that is we add a isotropic noise to every point. If the σ is small enough, we can hope to recover the direction v (See Figure 4.2). The covariance matrix of X_i generated as such is given by

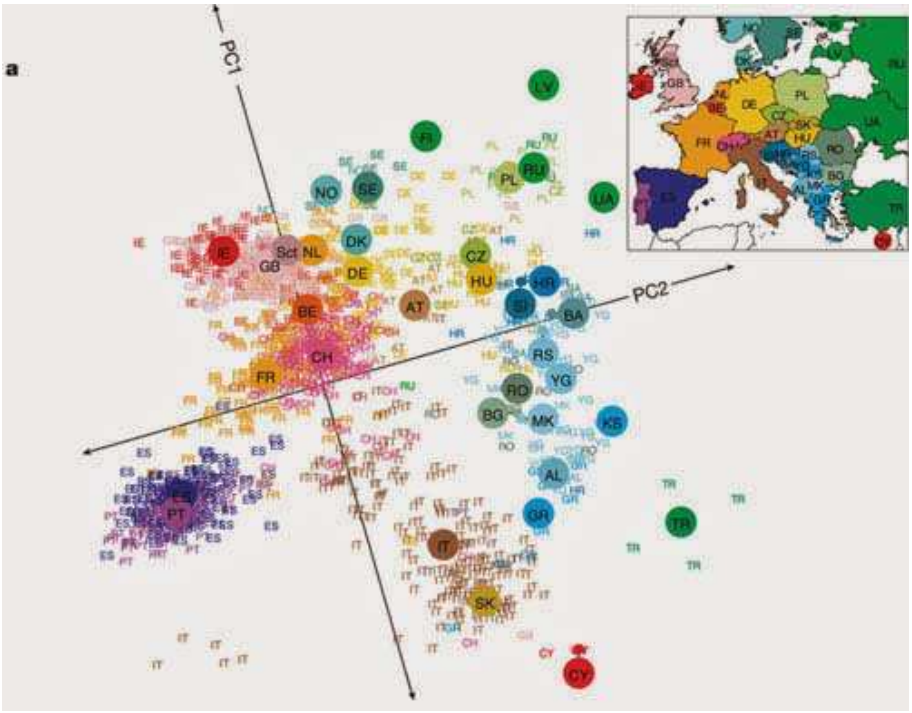
$$\Sigma = \mathbb{E}\big[XX^\top\big] = \mathbb{E}\big[((v^\top Y)v + Z)((v^\top Y)v + Z)^\top\big] = vv^\top + \sigma^2 I_d\,.$$

This model is often called the *spiked covariance model*. By a simple rescaling, it is equivalent to the following definition.

Definition 4.7. A covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is said to satisfy the spiked covariance model if it is of the form

$$\Sigma = \theta vv^\top + I_d\,,$$

where $\theta > 0$ and $v \in \mathcal{S}^{d-1}$. The vector v is called the *spike*.



Courtesy of Macmillan Publishers Ltd. Used with permission.

Figure 4.1. Projection onto two dimensions of $1,387$ points from gene expression data. Source: Gene expression blog.

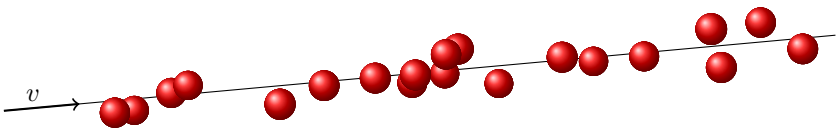


Figure 4.2. Points are close to a one dimensional space space by v.

This model can be extended to more than one spike but this extension is beyond the scope of these notes.

Clearly, under the spiked covariance model, v is the eigenvector of the matrix Σ that is associated to its largest eigenvalue $1 + \theta$. We will refer to this vector simply as *largest eigenvector*. To estimate it, a natural candidate is the largest eigenvector $\hat{v}$ of $\tilde{\Sigma}$, where $\tilde{\Sigma}$ is any estimator of Σ. There is a caveat: by symmetry, if u is an eigenvector, of a symmetric matrix, then $-u$ is also an eigenvector associated to the same eigenvalue. Therefore, we may only estimate v up to a sign flip. To overcome this limitation, it is often useful to describe proximity between two vectors u and v in terms of the principal angle

between their linear span. Let us recall that for two unit vectors the principal angle between their linear spans is denoted by $\angle(u, v)$ and defined as

$$\angle(u, v) = \arccos(|u^\top v|)\,.$$

The following result form perturbation theory is known as the Davis-Kahan $\sin(\theta)$ theorem as it bounds the sin of the principal angle between eigenspaces. This theorem exists in much more general versions that extend beyond one-dimensional eigenspaces.

Theorem 4.8 (Davis-Kahan $\sin(\theta)$ theorem)**.** *Let Σ satisfy the spiked covariance model and let $\tilde{\Sigma}$ be any PSD estimator of Σ. Let $\tilde{v}$ denote the largest eigenvector of $\tilde{\Sigma}$. Then we have*

$$\min_{\varepsilon\in\{\pm 1\}} |\varepsilon\tilde{v} - v|_2^2 \le 2\sin^2\big(\angle(\tilde{v}, v)\big) \le \frac{8}{\theta^2}\|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}}^2\,.$$

Proof. Note that for any $u \in \mathcal{S}^{d-1}$, it holds under the spiked covariance model that

$$u^\top \Sigma u = 1 + \theta(v^\top u)^2 = 1 + \theta\cos^2(\angle(u, v))\,.$$

Therefore,

$$v^\top \Sigma v - \tilde{v}^\top \Sigma \tilde{v} = \theta[1 - \cos^2(\angle(\tilde{v}, v))] = \theta\sin^2(\angle(\tilde{v}, v))\,.$$

Next, observe that

$$\begin{aligned}
v^\top \Sigma v - \tilde{v}^\top \Sigma \tilde{v} &= v^\top \tilde{\Sigma} v - \tilde{v}^\top \Sigma \tilde{v} - v^\top\big(\tilde{\Sigma} - \Sigma\big)v \\
&\le \tilde{v}^\top \tilde{\Sigma} \tilde{v} - \tilde{v}^\top \Sigma \tilde{v} - v^\top\big(\tilde{\Sigma} - \Sigma\big)v \\
&= \langle \hat{\Sigma} - \Sigma, \tilde{v}\tilde{v}^\top - vv^\top\rangle \qquad (4.8)\\
&\le \|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}}\|\tilde{v}\tilde{v}^\top - vv^\top\|_1 \qquad \text{(Hölder)}\\
&\le \sqrt{2}\|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}}\|\tilde{v}\tilde{v}^\top - vv^\top\|_F \qquad \text{(Cauchy-Schwarz)}\,.
\end{aligned}$$

where in the first inequality, we used the fact that $\tilde{v}$ is the largest eigenvector of $\tilde{\Sigma}$ and in the last one, we used the fact that the matrix $\tilde{v}\tilde{v}^\top - vv^\top$ has rank at most 2.

Next, we have that

$$\|\tilde{v}\tilde{v}^\top - vv^\top\|_F^2 = 2(1 - (v^\top\tilde{v})^2) = 2\sin^2(\angle(\tilde{v}, v))\,.$$

Therefore, we have proved that

$$\theta\sin^2(\angle(\tilde{v}, v)) \le 2\|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}}\sin(\angle(\tilde{v}, v))\,,$$

so that

$$\sin(\angle(\tilde{v}, v)) \le \frac{2}{\theta}\|\tilde{\Sigma} - \Sigma\|_{\mathrm{op}}\,.$$

To conclude the proof, it remains to check that

$$\min_{\varepsilon\in\{\pm 1\}} |\varepsilon\tilde{v} - v|_2^2 = 2 - 2|\tilde{v}^\top v| \le 2 - 2(\tilde{v}^\top v)^2 = 2\sin^2(\angle(\tilde{v}, v))\,.$$

□

Combined with Theorem 4.6, we immediately get the following corollary.

Corollary 4.9. *Let $X_1, \dots, X_n$ be n i.i.d. copies of a sub-Gaussian random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Assume further that $\Sigma = \theta vv^\top + I_d$ satisfies the spiked covariance model. Then, the largest eigenvector $\hat{v}$ of the empirical covariance matrix $\hat{\Sigma}$ satisfies,*

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat{v} - v|_2 \lesssim \frac{1+\theta}{\theta}\Big(\sqrt{\frac{d+\log(1/\delta)}{n}} \vee \frac{d+\log(1/\delta)}{n}\Big)$$

with probability $1-\delta$.

This result justifies the use of the empirical covariance matrix $\hat{\Sigma}$ as a replacement for the true covariance matrix Σ when performing PCA in low dimensions, that is when $d \ll n$. In the high-dimensional case, where $d \gg n$, the above result is uninformative. As before, we resort to sparsity to overcome this limitation.

Sparse PCA

In the example of Figure 4.1, it may be desirable to interpret the meaning of the two directions denoted by PC1 and PC2. We know that they are linear combinations of the original 500,000 gene expression levels. A natural question to ask is whether only a subset of these genes could suffice to obtain similar results. Such a discovery could have potential interesting scientific applications as it would point to a few genes responsible for disparities between European populations.

In the case of the spiked covariance model this amounts to have v to be sparse. Beyond interpretability as we just discussed, sparsity should also lead to statistical stability as in the case of sparse linear regression for example. To enforce sparsity, we will assume that v in the spiked covariance model is k-sparse: $|v|_0 = k$. Therefore, a natural candidate to estimate v is given by $\hat{v}$ defined by

$$\hat{v}^\top \hat{\Sigma} \hat{v} = \max_{\substack{u \in \mathcal{S}^{d-1} \\ |u|_0 = k}} u^\top \hat{\Sigma} u\,.$$

It is easy to check that $\lambda^k_{\max}(\hat{\Sigma}) = \hat{v}^\top \hat{\Sigma} \hat{v}$ is the largest of all leading eigenvalues among all $k \times k$ sub-matrices of $\hat{\Sigma}$ so that the maximum is indeed attained, though there my be several maximizers. We call $\lambda^k_{\max}(\hat{\Sigma})$ the k-sparse leading eigenvalue of $\hat{\Sigma}$ and $\hat{v}$ a k-sparse leading eigenvector.

Theorem 4.10. *Let $X_1, \dots, X_n$ be n i.i.d. copies of a sub-Gaussian random vector $X \in \mathbb{R}^d$ such that $\mathbb{E}[XX^\top] = \Sigma$ and $X \sim \mathsf{subG}_d(\|\Sigma\|_{\mathrm{op}})$. Assume further that $\Sigma = \theta vv^\top + I_d$ satisfies the spiked covariance model for v such that $|v|_0 = k \le d/2$. Then, the k-sparse largest eigenvector $\hat{v}$ of the empirical covariance matrix satisfies,*

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat{v} - v|_2 \lesssim \frac{1+\theta}{\theta}\Big(\sqrt{\frac{k\log(ed/k)+\log(1/\delta)}{n}} \vee \frac{k\log(ed/k)+\log(1/\delta)}{n}\Big)\,.$$

with probability $1-\delta$.

Proof. We begin by obtaining an intermediate result of the Davis-Kahan $\sin(\theta)$ theorem (Theorem 4.8). Note that we get from (4.8) that

$$v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} \le \langle \hat{\Sigma} - \Sigma, \hat{v}\hat{v}^\top - vv^\top \rangle$$

Since both $\hat{v}$ and v are k sparse, there exists a (random) set $S \subset \{1,\dots,d\}$ such that $|S| \le 2k$ and $\{\hat{v}\hat{v}^\top - vv^\top\}_{ij} = 0$ if $(i,j) \notin S^2$. It yields

$$\langle \hat{\Sigma} - \Sigma, \hat{v}\hat{v}^\top - vv^\top \rangle = \langle \hat{\Sigma}(S) - \Sigma(S), \hat{v}(S)\hat{v}(S)^\top - v(S)v(S)^\top \rangle$$

Where for any $d \times d$ matrix M, we defined the matrix $M(S)$ to be the $|S| \times |S|$ sub-matrix of M with rows and columns indexed by S and for any vector $x \in \mathbb{R}^d$, $x(S) \in \mathbb{R}^{|S|}$ denotes the sub-vector of x with coordinates indexed by S. It yields, by Hölder's inequality that

$$v^\top \Sigma v - \hat{v}^\top \Sigma \hat{v} \le \|\hat{\Sigma}(S) - \Sigma(S)\|_{\mathrm{op}} \|\hat{v}(S)\hat{v}(S)^\top - v(S)v(S)^\top\|_1 .$$

Following the same steps as in the proof of Theorem 4.8, we get now that

$$\min_{\varepsilon \in \{\pm 1\}} |\varepsilon \hat{v} - v|_2^2 \le 2\sin^2\big(\angle(\hat{v}, v)\big) \le \frac{8}{\theta^2} \sup_{S\,:\,|S|=2k} \|\hat{\Sigma}(S) - \Sigma(S)\|_{\mathrm{op}} .$$

To conclude the proof, it remains to control $\sup_{S\,:\,|S|=2k} \|\hat{\Sigma}(S) - \Sigma(S)\|_{\mathrm{op}}$. To that end, observe that

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{S\,:\,|S|=2k} \|\hat{\Sigma}(S) - \Sigma(S)\|_{\mathrm{op}} > t\|\Sigma\|_{\mathrm{op}} \Big] \\
&\qquad \le \sum_{S\,:\,|S|=2k} \mathbb{P}\Big[\sup_{S\,:\,|S|=2k} \|\hat{\Sigma}(S) - \Sigma(S)\|_{\mathrm{op}} > t\|\Sigma(S)\|_{\mathrm{op}} \Big] \\
&\qquad \le \binom{d}{2k} 144^{2k} \exp\Big[-\frac{n}{2}\Big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\Big)\Big] .
\end{aligned}$$

where we used (4.7) in the second inequality. Using Lemma 2.7, we get that the right-hand side above is further bounded by

$$\exp\Big[-\frac{n}{2}\Big(\Big(\frac{t}{32}\Big)^2 \wedge \frac{t}{32}\Big) + 2k\log(144) + k\log\Big(\frac{ed}{2k}\Big)\Big]$$

Choosing now t such that

$$t \ge C\sqrt{\frac{k\log(ed/k) + \log(1/\delta)}{n}} \vee \frac{k\log(ed/k) + \log(1/\delta)}{n} ,$$

for large enough C ensures that the desired bound holds with probability at least $1-\delta$. ∎

4.5 PROBLEM SET

Problem 4.1. Using the results of Chapter 2, show that the following holds for the multivariate regression model (4.1).

1. There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ such that
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{rT}{n}$$
with probability .99, where r denotes the rank of $\mathbb{X}$.

2. There exists an estimator $\hat{\Theta} \in \mathbb{R}^{d\times T}$ such that
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \sigma^2 \frac{|\Theta^*|_0 \log(ed)}{n}\,.$$
with probability .99.

Problem 4.2. Consider the multivariate regression model (4.1) where $\mathbb{Y}$ has SVD:
$$\mathbb{Y} = \sum_j \hat{\lambda}_j \hat{u}_j \hat{v}_j^\top\,.$$
Let M be defined by
$$\hat{M} = \sum_j \hat{\lambda}_j \mathbb{I}(|\hat{\lambda}_j| > 2\tau)\hat{u}_j\hat{v}_j^\top\,, \tau > 0\,.$$

1. Show that there exists a choice of τ such that
$$\frac{1}{n}\|\hat{M} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$
with probability .99.

2. Show that there exists a matrix $n \times n$ matrix P such that $P\hat{M} = \mathbb{X}\hat{\Theta}$ for some estimator $\hat{\Theta}$ and
$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$
with probability .99.

3. Comment on the above results in light of the results obtain in Section 4.2.

Problem 4.3. Consider the multivariate regression model (4.1) and define $\hat{\Theta}$ be the any solution to the minimization problem
$$\min_{\Theta\in\mathbb{R}^{d\times T}} \left\{\frac{1}{n}\|\mathbb{Y} - \mathbb{X}\Theta\|_F^2 + \tau\|\mathbb{X}\Theta\|_1\right\}$$

1. Show that there exists a choice of τ such that

$$\frac{1}{n}\|\mathbb{X}\hat{\Theta} - \mathbb{X}\Theta^*\|_F^2 \lesssim \frac{\sigma^2 \operatorname{rank}(\Theta^*)}{n}(d \vee T)$$

with probability .99.

[Hint:Consider the matrix

$$\sum_j \frac{\hat{\lambda}_j + \lambda_j^*}{2}\hat{u}_j\hat{v}_j^\top$$

where $\lambda_1^* \geq \lambda_2^* \geq \ldots$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots$ are the singular values of $\mathbb{X}\Theta^*$ and $\mathbb{Y}$ respectively and the SVD of $\mathbb{Y}$ is given by

$$\mathbb{Y} = \sum_j \hat{\lambda}_j\hat{u}_j\hat{v}_j^\top$$

2. Find a closed form for $\mathbb{X}\hat{\Theta}$.

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