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.

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*

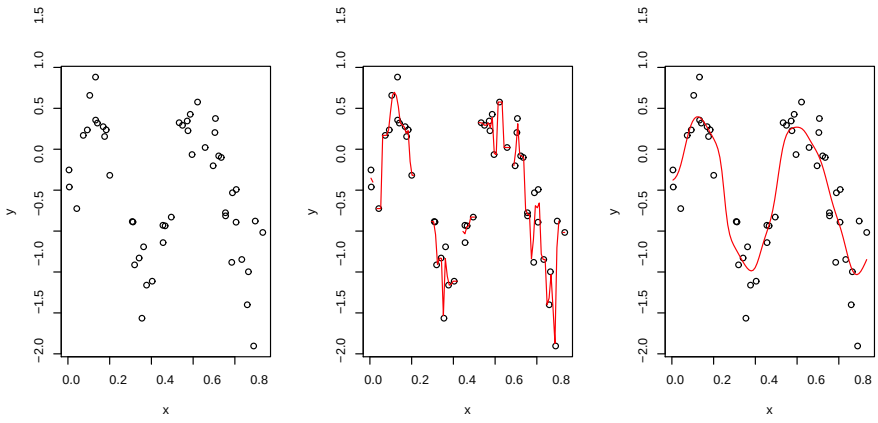

Figure 1. It may not be the best choice idea to have $\hat{f}_n(X_i) = Y_i$ for all $i = 1, \dots, n$.

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 $q \geq 1$.

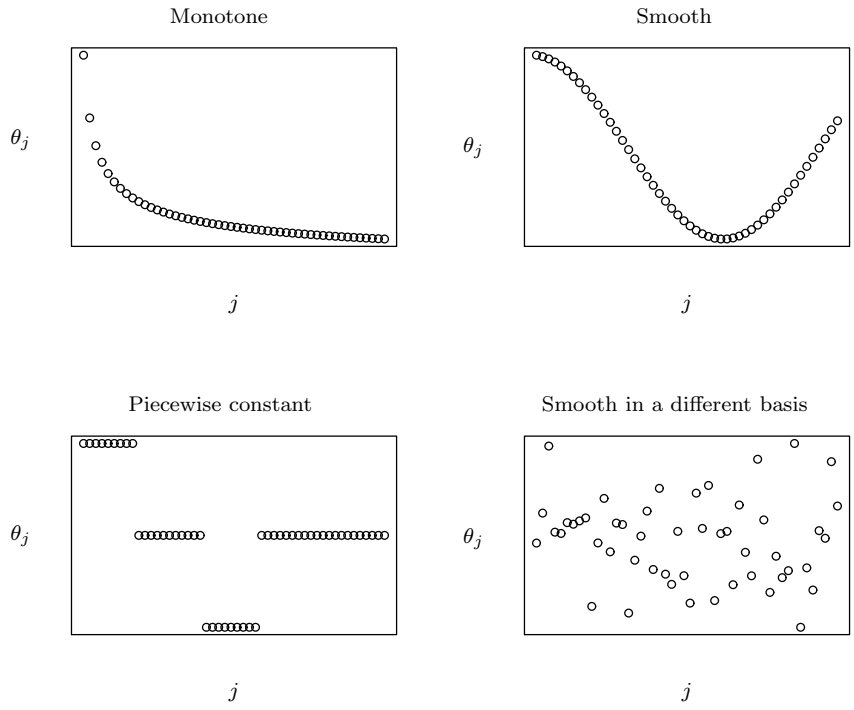

Figure 2. Examples of structures vectors $\theta \in \mathbb{R}^{50}$

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.

18.S997 High-dimensional Statistics
Spring 2015