EMPIRICAL PROCESS THEORY AND APPLICATIONS

by

Sara van de Geer

Handout **WS 2006**
ETH Zürich

Contents

Preface

This preface motivates why, from a statistician's point of view, it is interesting to study empirical processes. We indicate that any estimator is some function of the empirical measure. In these lectures, we study convergence of the empirical measure, as sample size increases.

In the simplest case, a data set consists of observations on a single variable, say real-valued observations. Suppose there are n such observations, denoted by $X_1, \ldots, X_n$. For example, X_i could be the reaction time of individual i to a given stimulus, or the number of car accidents on day i, etc. Suppose now that each observation follows the same probability law P. This means that the observations are relevant if one wants to predict the value of a new observation X say (the reaction time of a hypothetical new subject, or the number of car accidents on a future day, etc.). Thus, a common underlying distribution P allows one to generalize the outcomes.

An estimator is any given function $T_n(X_1, \ldots, X_n)$ of the data. Let us review some common estimators.

The empirical distribution. The unknown P can be estimated from the data in the following way. Suppose first that we are interested in the probability that an observation falls in A, where A is a certain set chosen by the researcher. We denote this probability by $P(A)$. Now, from the frequentist point of view, the probability of an event is nothing else than the limit of relative frequencies of occurrences of that event as the number of occasions of possible occurrences n grows without limit. So it is natural to estimate $P(A)$ with the frequency of A, i.e, with

$$P_n(A) = \frac{\text{number of times an observation } X_i \text{ falls in } A}{\text{total number of observations}}$$
$$= \frac{\text{number of } X_i \in A}{n}.$$

We now define the empirical measure P_n as the probability law that assigns to a set A the probability $P_n(A)$. We regard P_n as an estimator of the unknown P.

The empirical distribution function. The distribution function of X is defined as

$$F(x) = P(X \leq x),$$

and the empirical distribution function is

$$\hat{F}_n(x) = \frac{\text{number of } X_i \leq x}{n}.$$

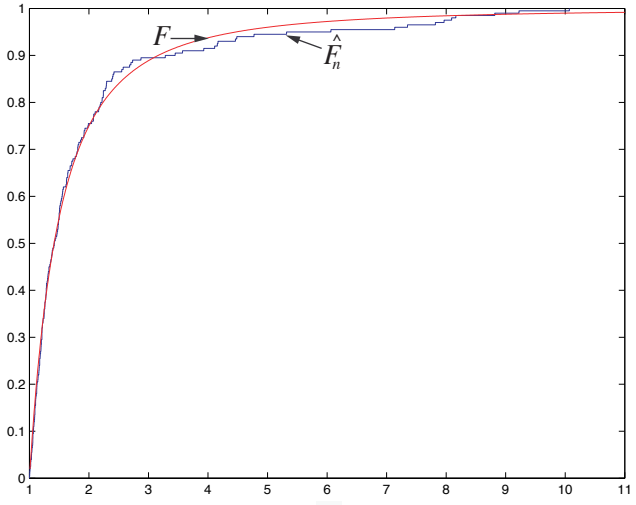

Figure 1

Figure 1 plots the distribution function $F(x) = 1 - 1/x^2$, $x \geq 1$ (smooth curve) and the empirical distribution function $\hat{F}_n$ (stair function) of a sample from F with sample size $n = 200$.

Means and averages. The theoretical mean

$$\mu := E(X)$$

(E stands for Expectation), can be estimated by the sample average

$$\bar{X}_n := \frac{X_1 + \ldots + X_n}{n}.$$

More generally, let g be a real-valued function on $\mathbf{R}$. Then

$$\frac{g(X_1) + \ldots + g(X_n)}{n},$$

is an estimator $Eg(X)$.

Sample median. The median of X is the value m that satisfies $F(m) = 1/2$ (assuming there is a unique solution). Its empirical version is any value $\hat{m}_n$ such that $\hat{F}_n(\hat{m}_n)$ is equal or as close as possible to $1/2$. In the above example $F(x) = 1 - 1/x^2$, so that the theoretical median is $m = \sqrt{2} = 1.4142$. In the ordered sample, the 100^{th} observation is equal to 1.4166 and the 101^{th} observation is equal to 1.4191. A common choice for the sample median is taking the average of these two values. This gives $\hat{m}_n = 1.4179$.

Properties of estimators. Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator of the real-valued parameter θ. Then it is desirable that T_n is in some sense close to θ. A minimum requirement is that the estimator approaches θ as the sample size increases. This is called *consistency*. To be more precise, suppose the sample $X_1, \ldots, X_n$ are the first n of an infinite sequence $X_1, X_2, \ldots$ of independent copies of X. Then T_n is called strongly consistent if, with probability one,

$$T_n \to \theta \text{ as } n \to \infty.$$

Note that consistency of frequencies as estimators of probabilities, or means as estimators of expectations, follows from the (strong) law of large numbers. In general, an estimator T_n can be a complicated function of the data. In that case, it is helpful to know that the convergence of means to their expectations is uniform over a class. The latter is a major topic in empirical process theory.

Parametric models. The distribution P may be partly known beforehand. The unknown parts of P are called *parameters* of the model. For example, if the X_i are yes/no answers to a certain question (the binary case), we know that P allows only two possibilities, say 1 and 0 (yes=1, no=0). There is only one parameter , say the probability of a yes answer $\theta = P(X = 1)$. More generally, in a parametric model, it is assumed that P is known up to a finite number of parameters $\theta = (\theta_1, \cdots, \theta_d)$. We then often write $P = P_\theta$. When there are infinitely many parameters (which is for example the case when P is completely unknown), the model is called nonparametric.

Nonparametric models.
An example of a nonparametric model is where one assumes that the density f of the distribution function F exists, but all one assumes about it is some kind of "smoothness" (e.g. the continuous first derivative of f exists). In that case, one may propose e.g. to use the histogram as estimator of f. This is an example of a nonparametric estimator.

Histograms. Our aim is estimating the density $f(x)$ at a given point x. The density is defined as the derivative of the distribution function F at x:

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \to 0} \frac{P(x, x+h]}{h}.$$

Here, $(x, x+h]$ is the interval with left endpoint x (not included) and right endpoint $x+h$ (included). Unfortunately, replacing P by P_n here does not work, as for h small enough, $P_n(x, x+h]$ will be equal to zero. Therefore, instead of taking the limit as $h \to 0$, we *fix* h at a (small) positive value, called the bandwidth. The estimator of $f(x)$ thus becomes

$$\hat{f}_n(x) = \frac{P_n(x, x+h]}{h} = \frac{\text{number of } X_i \in (x, x+h]}{nh}.$$

A plot of this estimator at points $x \in \{x_0, x_0 + h, x_0 + 2h, \ldots\}$ is called a histogram.

Example . Figure 2 shows the histogram, with bandwidth $h = 0.5$, for the sample of size $n = 200$ from the Pareto distribution with parameter $\theta = 2$. The solid line is the density of this distribution.

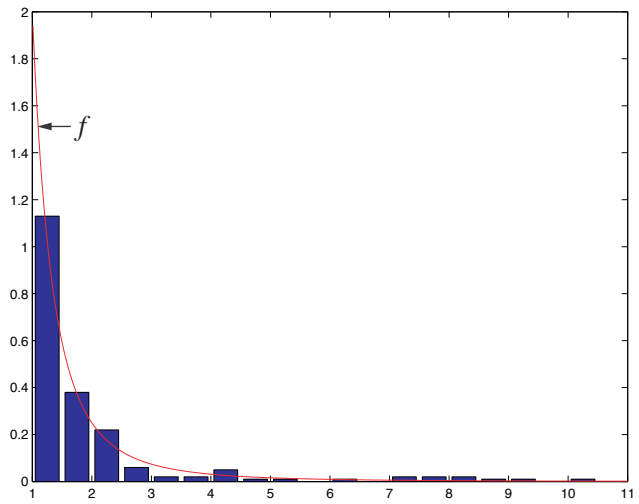

Figure 2

Conclusion. An estimator T_n is some function of the data $X_1, \ldots, X_n$. If it is a symmetric function of the data (which we can in fact assume without loss of generality when the ordering in the data contains no information), we may write $T_n = T(P_n)$, where P_n is the empirical distribution. Roughly speaking, the main purpose in theoretical statistics is studying the difference between $T(P_n)$ and $T(P)$. We therefore are interested in convergence of P_n to P in a broad enough sense. This is what empirical process theory is about.

1. Introduction.

This chapter introduces the notation and (part of the) problem setting.

Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P. The distribution of the sequence $X_1, X_2, \ldots$ (+ perhaps some auxiliary variables) is denoted by $\mathbf{P}$.

Definition. *Let $\{T_n, T\}$ be a collection of real-valued random variables. Then T_n converges* in probability *to T, if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}(|T_n - T| > \epsilon) = 0.$$

Notation: $T_n \to^{\mathbf{P}} T$.

Moreover, T_n converges almost surely (a.s.) *to T if*

$$\mathbf{P}(\lim_{n\to\infty} T_n = T) = 1.$$

Remark. Convergence almost surely implies convergence in probability.

1.1. Law of large numbers for real-valued random variables. Consider the case $\mathcal{X} = \mathbf{R}$. Suppose the mean

$$\mu := EX$$

exists. Define the average

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

Then, by the law of large numbers, as $n \to \infty$,

$$\bar{X}_n \to \mu, \text{ a.s.}$$

Now, let

$$F(t) := P(X \leq t), \ t \in \mathbf{R},$$

be the theoretical distribution function, and

$$F_n(t) := \frac{1}{n}\#\{X_i \leq t, \ 1 \leq i \leq n\}, \ t \in \mathbf{R},$$

be the empirical distribution function. Then by the law of large numbers, as $n \to \infty$,

$$F_n(t) \to F(t), \text{ a.s. for all } t.$$

We will prove (in Chapter 4) the Glivenko-Cantelli Theorem, which says that

$$\sup_t |F_n(t) - F(t)| \to 0, \text{ a.s.}$$

This is a **uniform** law of large numbers.

Application: *Kolmogorov's goodness-of-fit test.* We want to test
$H_0 : \ F = F_0$.
Test statistic:

$$D_n; = \sup_t |F_n(t) - F_0(t)|.$$

Reject H_0 for large values of D_n.

1.2. $\mathbf{R}^d$-valued random variables. Questions:
(i) What is a natural extension of half-intervals in $\mathbf{R}$ to higher dimensions?
(ii) Does Glivenko-Cantelli hold for this extension?

1.3. Definition Glivenko-Cantelli classes of sets. Let for any (measurable[1]) $A \subset \mathcal{X}$,

$$P_n(A) := \frac{1}{n}\#\{X_i \in A,\ 1 \le i \le n\}.$$

We call P_n the empirical measure (based on $X_1, \ldots, X_n$).

Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$.

Definition 1.3.1. The collection $\mathcal{D}$ is called a **Glivenko-Cantelli** (GC) class if

$$\sup_{D \in \mathcal{D}} |P_n(D) - P(D)| \to 0, \text{ a.s.}$$

Example. Let $\mathcal{X} = \mathbf{R}$. The class of half-intervals

$$\mathcal{D} = \{1_{(-\infty,t]} :\ t \in \mathbf{R}\}$$

is GC. But when e.g. P = uniform distribution on $[0,1]$ (i.e., $F(t) = t$, $0 \le t \le 1$), the class

$$\mathcal{B} = \{\text{all (Borel) subsets of } [0,1]\}$$

is **not** GC.

1.4. Convergence of averages to their expectations.

Notation. For a function $g : \mathcal{X} \to \mathbf{R}$, we write

$$P(g) := Eg(X),$$

and

$$P_n(g) := \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Let $\mathcal{G}$ be a collection of real-valued functions on $\mathcal{X}$.

Definition 1.4.1. The class $\mathcal{G}$ is called a **Glivenko-Cantelli** (GC) class if

$$\sup_{g \in \mathcal{G}} |P_n(g) - P(g)| \to 0, \text{ a.s.}$$

We will often use the notation

$$\|P_n - P\|_{\mathcal{G}} := \sup_{g \in \mathcal{G}} |P_n(g) - P(g)|.$$

[1] We will skip measurability issues, and most of the time do not mention explicitly the requirement of measurability of certain sets or functions. This means that everything has to be understood *modulo* measurability.

2. (Exponential) probability inequalities

A statistician is almost never sure about something, but often says that something holds "with large probability". We study probability inequalities for deviations of means from their expectations. These are exponential inequalities, that is, the probability that the deviation is large is exponentially small. (We will in fact see that the inequalities are similar to those obtained if we assume normality.) Exponentially small probabilities are useful indeed when one wants to prove that with large probability a whole collection of events holds simultaneously. It then suffices to show that adding up the small probabilities that one such an event does not hold, still gives something small. We will use this argument in Chapter 4.

2.1. Chebyshev's inequality.

Chebyshev's inequality. *Consider a random variable $X \in \mathbf{R}$ with distribution P, and an increasing function $\phi : \mathbf{R} \to [0, \infty)$. Then for all a with $\phi(a) > 0$, we have*

$$P(X \geq a) \leq \frac{E\phi(X)}{\phi(a)}.$$

Proof.

$$\begin{aligned} E\phi(X) = \int \phi(x)dP(x) &= \int_{X \geq a} \phi(x)dP(x) + \int_{X<a} \phi(x)dP(x) \\ &\geq \int_{X \geq a} \phi(x)dP(x) \geq \int_{X \geq a} \phi(a)dP(x) \\ &= \phi(a) \int_{X \geq a} dP = \phi(a)P(X \geq a). \end{aligned}$$

□

Let X be $\mathcal{N}(0, 1)$-distributed. By Exercise 2.4.1,

$$P(X \geq a) \leq \exp[-a^2/2] \ \forall \ a > 0.$$

Corollary 2.1.1. *Let $X_1, \ldots, X_n$ be independent real-valued random variables, and suppose, for all i, that X_i is $\mathcal{N}(0, \sigma_i^2)$-distributed. Define*

$$b^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Then for all $a > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2b^2}\right].$$

2.2. Bernstein's inequality.

Bernstein's inequality. *Let $X_1, \ldots, X_n$ be independent real-valued random variables with expectation zero. Suppose that for all i,*

$$\mathbf{E}|X_i|^m \leq \frac{m!}{2} K^{m-2} \sigma_i^2, \ m = 2, 3, \ldots.$$

Define

$$b^2 = \sum_{i=1}^{n} \sigma_i^2.$$

We have for any $a > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2(aK + b^2)}\right].$$

Proof. We have for $0 < \lambda < 1/K$,

$$\begin{aligned}\mathbf{E}\exp[\lambda X_i] &= 1 + \sum_{m=2}^{\infty} \frac{1}{m!}\lambda^m \mathbf{E}X_i^m \\ &\leq 1 + \sum_{m=2}^{\infty} \frac{\lambda^2}{2}(\lambda K)^{m-2}\sigma_i^2 \\ &= 1 + \frac{\lambda^2\sigma_i^2}{2(1-\lambda K)} \\ &\leq \exp\left[\frac{\lambda^2\sigma_i^2}{2(1-\lambda K)}\right].\end{aligned}$$

It follows that

$$\begin{aligned}\mathbf{E}\exp\left[\lambda\sum_{i=1}^{n} X_i\right] &= \prod_{i=1}^{n} \mathbf{E}\exp[\lambda X_i] \\ &\leq \exp\left[\frac{\lambda^2 b^2}{2(1-\lambda K)}\right].\end{aligned}$$

Now, apply Chebyshev's inequality to $\sum_{i=1}^n X_i$, and with $\phi(x) = \exp[\lambda x]$, $x \in \mathbf{R}$. We arrive at

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[\frac{\lambda^2 b^2}{2(1-\lambda K)} - \lambda a\right].$$

Take

$$\lambda = \frac{a}{Ka + b^2}$$

to complete the proof.

□

2.3. Hoeffding's inequality.

Hoeffding's inequality. *Let $X_1, \ldots, X_n$ be independent real-valued random variables with expectation zero. Suppose that for all i, and for certain constants $c_i > 0$,*

$$|X_i| \leq c_i.$$

Then for all $a > 0$,

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2\sum_{i=1}^n c_i^2}\right].$$

Proof. Let $\lambda > 0$. By the convexity of the exponential function $\exp[\lambda x]$, we know that for any $0 \leq \alpha \leq 1$,

$$\exp[\alpha\lambda x + (1-\alpha)\lambda y] \leq \alpha\exp[\lambda x] + (1-\alpha)\exp[\lambda y].$$

Define now

$$\alpha_i = \frac{c_i - X_i}{2c_i}.$$

Then

$$X_i = \alpha_i(-c_i) + (1-\alpha_i)c_i,$$

so

$$\exp[\lambda X_i] \leq \alpha_i \exp[-\lambda c_i] + (1-\alpha_i)\exp[\lambda c_i].$$

But then, since $\mathbf{E}\alpha_i = 1/2$, we find

$$\mathbf{E}\exp[\lambda X_i] \leq \frac{1}{2}\exp[-\lambda c_i] + \frac{1}{2}\exp[\lambda c_i].$$

Now, for all x,

$$\exp[-x] + \exp[x] = 2\sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!},$$

whereas

$$\exp[x^2/2] = \sum_{k=0}^{\infty} \frac{x^{2k}}{2^k k!}.$$

Since

$$(2k)! \geq 2^k k!,$$

we thus know that

$$\exp[-x] + \exp[x] \leq 2\exp[x^2/2],$$

and hence

$$\mathbf{E}\exp[\lambda X_i] \leq \exp[\lambda^2 c_i^2/2].$$

Therefore,

$$\mathbf{E}\exp\left[\lambda \sum_{i=1}^{n} X_i\right] \leq \exp\left[\lambda^2 \sum_{i=1}^{n} c_i^2/2\right].$$

It follows now from Chebyshev's inequality that

$$\mathbf{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[\lambda^2 \sum_{i=1}^{n} c_i^2/2 - \lambda a\right].$$

Take $\lambda = a/(\sum_{i=1}^{n} c_i^2)$ to complete the proof.

□

2.4. Exercise.

Exercise 1.
Let X be $\mathcal{N}(0,1)$-distributed. Show that for $\lambda > 0$,

$$E\exp[\lambda X] = \exp[\lambda^2/2].$$

Conclude that for all $a > 0$,

$$P(X \geq a) \leq \exp[\lambda^2/2 - \lambda a].$$

Take $\lambda = a$ to find the inequality

$$P(X \geq a) \leq \exp[-a^2/2].$$

3. Symmetrization

Symmetrization is a technique based on the following idea. Suppose you have some estimation method, and want to know how good it performs. Suppose you have a sample of size n, the so-called training set and a second sample, say also of size n, the so-called test set. Then we may use the training set to calculate the estimator, and the test set to check its performance. For example, suppose we want to know how large the maximal deviation is between certain averages and expectations. We cannot calculate this maximal deviation directly, as the expectations are unknown. Instead, we can calculate the maximal deviation between the averages in the two samples. Symmetrization is closely rerlated: it splits the sample of size n randomly in two subsamples.

Let $X \in \mathcal{X}$ be a random variable with distribution P. We consider two independent sets of independent copies of X, $\mathbf{X} := X_1, \ldots, X_n$ and $\mathbf{X}' := X'_1, \ldots, X'_n$.

Let $\mathcal{G}$ be a class of real-valued functions on $\mathcal{X}$. Consider the empirical measures

$$P_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i},\ P'_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}.$$

Here δ_x denotes a point mass at x. Define

$$\|P_n - P\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |P_n(g) - P(g)|,$$

and likewise

$$\|P'_n - P\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |P'_n(g) - P(g)|,$$

and

$$\|P_n - P'_n\|_{\mathcal{G}} := \sup_{g\in\mathcal{G}} |P_n(g) - P'_n(g)|.$$

3.1. Symmetrization with means.

Lemma 3.1.1. *We have*

$$\mathbf{E}\|P_n - P\|_{\mathcal{G}} \leq \mathbf{E}\|P_n - P'_n\|_{\mathcal{G}}.$$

Proof. For a function f on $\mathcal{X}^{2n}$, let $\mathbf{E}_{\mathbf{X}} f(\mathbf{X}, \mathbf{X}')$ denote the conditional expectation of $f(\mathbf{X}, \mathbf{X}')$ given $\mathbf{X}$. Then obviously,

$$\mathbf{E}_{\mathbf{X}} P_n(g) = P_n(g)$$

and

$$\mathbf{E}_{\mathbf{X}} P'_n(g) = P(g).$$

So

$$(P_n - P)(g) = \mathbf{E}_{\mathbf{X}}(P_n - P'_n)(g).$$

Hence

$$\|P_n - P\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |P_n(g) - P(g)| = \sup_{g\in\mathcal{G}} |\mathbf{E}_{\mathbf{X}}(P_n - P'_n)(g)|.$$

Now, use that for any function $f(Z,t)$ depending on a random variable Z and a parameter t, we have

$$\sup_t |Ef(Z,t)| \leq \sup_t E|f(Z,t)| \leq E \sup_t |f(Z,t)|.$$

So

$$\sup_{g\in\mathcal{G}} |\mathbf{E}_{\mathbf{X}}(P_n - P'_n)(g)| \leq \mathbf{E}_{\mathbf{X}}\|P_n - P'_n\|_{\mathcal{G}}.$$

So we now showed that

$$\|P_n - P\|_{\mathcal{G}} \leq \mathbf{E}_{\mathbf{X}}\|P_n - P'_n\|_{\mathcal{G}}.$$

Finally, we use that the expectation of the conditional expectation is the unconditional expectation:

$$\mathbf{E}\mathbf{E}_{\mathbf{X}} f(\mathbf{X}, \mathbf{X}') = \mathbf{E} f(\mathbf{X}, \mathbf{X}).$$

So

$$\mathbf{E}\|P_n - P\|_{\mathcal{G}} \leq \mathbf{E}\mathbf{E}_{\mathbf{X}}\|P_n - P_n'\|_{\mathcal{G}} = \mathbf{E}\|P_n - P_n'\|_{\mathcal{G}}.$$

□

Definition 3.1.2. *A* Rademacher *sequence* $\{\sigma_i\}_{i=1}^n$ *is a sequence of independent random variables* σ_i*, with*

$$\mathbf{P}(\sigma_i = 1) = \mathbf{P}(\sigma_i = -1) = \frac{1}{2} \ \forall \ i.$$

Let $\{\sigma_i\}_{i=1}^n$ be a Rademacher sequence, independent of the two samples $\mathbf{X}$ and $\mathbf{X}'$. We define the symmetrized empirical measure

$$P_n^\sigma(g) := \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i), \ g \in \mathcal{G}.$$

Let

$$\|P_n^\sigma\|_{\mathcal{G}} = \sup_{g\in\mathcal{G}} |P_n^\sigma(g)|.$$

Lemma 3.1.3. *We have*

$$\mathbf{E}\|P_n - P\|_{\mathcal{G}} \leq 2\mathbf{E}\|P_n^\sigma\|_{\mathcal{G}}.$$

Proof. Consider the symmetrized version of the second sample $\mathbf{X}'$:

$$P_n'^{,\sigma}(g) = \frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i').$$

Then $\|P_n - P_n'\|_{\mathcal{G}}$ has the same distribution as $\|P_n^\sigma - P_n'^{,\sigma}\|_{\mathcal{G}}$. So

$$\mathbf{E}\|P_n - P_n'\|_{\mathcal{G}} = \mathbf{E}\|P_n^\sigma - P_n'^{,\sigma}\|_{\mathcal{G}}$$

$$\leq \mathbf{E}\|P_n^\sigma\|_{\mathcal{G}} + \mathbf{E}\|P_n'^{,\sigma}\|_{\mathcal{G}} = 2\mathbf{E}\|P_n^\sigma\|_{\mathcal{G}}.$$

□

3.2. Symmetrization with probabilities.

Lemma 3.2.1. *Let* $\delta > 0$*. Suppose that for all* $g \in \mathcal{G}$*,*

$$\mathbf{P}\left(|P_n(g) - P(g)| > \delta/2\right) \leq \frac{1}{2}.$$

Then

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > \delta\right) \leq 2\mathbf{P}\left(\|P_n - P_n'\|_{\mathcal{G}} > \frac{\delta}{2}\right).$$

Proof. Let $\mathbf{P}_{\mathbf{X}}$ denote the conditional probability given $\mathbf{X}$. If $\|P_n - P\|_{\mathcal{G}} > \delta$, we know that for some random function $g_* = g_*(\mathbf{X})$ depending on $\mathbf{X}$,

$$|P_n(g_*) - P(g_*)| > \delta.$$

Because $\mathbf{X}'$ is independent of $\mathbf{X}$, we also know that

$$\mathbf{P}_{\mathbf{X}}\left(|P_n'(g_*) - P(g_*)| > \delta/2\right) \leq \frac{1}{2}.$$

Thus,

$$\mathbf{P}\left(|P_n(g_*) - P(g_*)| > \delta \text{ and } |P_n'(g_*) - P(g_*)| \leq \frac{\delta}{2}\right)$$

$$= \mathbf{E}\mathbf{P}_{\mathbf{X}}\left(|P_n(g_*) - P(g_*)| > \delta \text{ and } |P_n'(g_*) - P(g_*)| \leq \frac{\delta}{2}\right)$$

$$= \mathbf{E}\mathbf{P}_{\mathbf{X}}\left(|P'_n(g_*) - P)(g_*)| \le \frac{\delta}{2}\right) 1\{|P_n(g_*) - P(g_*)| > \delta\}$$

$$\ge \frac{1}{2}\mathbf{E}1\{(|P_n(g_*) - P(g_*)| > \delta\}$$

$$= \frac{1}{2}\mathbf{P}\left(|P_n(g_*) - P(g_*)| > \delta\right).$$

It follows that

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > \delta\right) \le \mathbf{P}\left(|P_n(g_*) - P(g_*)| > \delta\right)$$

$$\le 2\mathbf{P}\left(|P_n(g^*) - P(g_*)| > \delta \text{ and } |P'_n(g_*) - P(g_*)| \le \frac{\delta}{2}\right)$$

$$\le 2\mathbf{P}\left(|P_n(g_*) - P'_n(g_*)| > \frac{\delta}{2}\right)$$

□

Corollary 3.2.2. *Let $\delta > 0$. Suppose that for all $g \in \mathcal{G}$,*

$$\mathbf{P}\left(|P_n(g) - P(g)| > \delta/2\right) \le \frac{1}{2}.$$

Then

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > \delta\right) \le 4\mathbf{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{\delta}{4}\right).$$

3.3. Some facts about conditional expectations.

Let X and Y be two random variables. We write the conditional expectation of Y given X as

$$\mathbf{E}_X(Y) = \mathbf{E}(Y|X).$$

Then

$$\mathbf{E}(Y) = \mathbf{E}(\mathbf{E}_X(Y).$$

Let f be some function of X and g be some function of (X, Y). We have

$$\mathbf{E}_X(f(X)g(X,Y)) = f(X)\mathbf{E}_X g(X,Y).$$

The conditional probability given X is

$$\mathbf{P}_X((X,Y) \in B) = \mathbf{E}_X 1_B(X,Y).$$

Hence,

$$\mathbf{P}_X(X \in X,\ (X,Y) \in B) = 1_A(X)\mathbf{P}_X((X,Y) \in B).$$

4. Uniform laws of large numbers.

In this chapter, we prove uniform laws of large numbers for the empirical mean of functions g of the individual observations, when g varies over a class $\mathcal{G}$ of functions. First, we study the case where $\mathcal{G}$ is finite. Symmetrization is used in order to be able to apply Hoeffding's inequality. Hoeffding's inequality gives exponential small probabilities for the deviation of averages from their expectations. So considering only a finite number of such averages, the difference between these averages and their expectations will be small for all averages simultaneously, with large probability.

If $\mathcal{G}$ is not finite, we approximate it by a finite set. A δ-approximation is called a δ-covering, and the number of elements of a δ-covering is called the δ-covering number.

We introduce Vapnik Chervonenkis (VC) classes. These are classes with small covering numbers.

Let $X \in \mathcal{X}$ be a random variable with distribution P. Consider a class $\mathcal{G}$ of real-valued functions on $\mathcal{X}$, and consider i.i.d. copies $\{X_1, X_2, \ldots\}$ of X. In this chapter, we address the problem of proving $\|P_n - P\|_{\mathcal{G}} \to^{\mathcal{P}} 0$. If this is the case, we call $\mathcal{G}$ a Glivenko Cantelli (GC) class.

Remark. It can be shown that if $\|P_n - P\|_{\mathcal{G}} \to^{\mathbf{P}} 0$, then also $\|P_n - P\|_{\mathcal{G}} \to 0$ almost surely. This involves e.g. martingale arguments. We will not consider this issue.

4.1. Classes of functions.

Notation. The **sup**-norm of a function g is

$$\|g\|_\infty := \sup_{x \in \mathcal{X}} |g(x)|.$$

Elementary observation. Let $\{A_k\}_{k=1}^N$ be a finite collection of events. Then

$$\mathbf{P}\left(\cup_{k=1}^N A_k\right) \leq \sum_{k=1}^N \mathbf{P}(A_k) \leq N \max_{1 \leq k \leq N} \mathbf{P}(A).$$

Lemma 4.1.1. *Let $\mathcal{G}$ be a finite class of functions, with cardinality $|\mathcal{G}| := N > 1$. Suppose that for some finite constant K,*

$$\max_{g \in \mathcal{G}} \|g\|_\infty \leq K.$$

Then for all

$$\delta \geq 2K\sqrt{\frac{\log N}{n}},$$

we have

$$\mathbf{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \delta\right) \leq 2\exp\left[-\frac{n\delta^2}{4K^2}\right]$$

and

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > 4\delta\right) \leq 8\exp\left[-\frac{n\delta^2}{4K^2}\right].$$

Proof.

• By Hoeffding's inequality, for each $g \in \mathcal{G}$,

$$\mathbf{P}\left(|P_n^\sigma(g)| > \delta\right) \leq 2\exp\left[-\frac{n\delta^2}{2K^2}\right].$$

• Use the elementary observation to conclude that

$$\mathbf{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \delta\right) \leq 2N\exp\left[-\frac{n\delta^2}{2K^2}\right]$$

$$= 2\exp\left[\log N - \frac{n\delta^2}{2K^2}\right] \leq 2\exp\left[-\frac{n\delta^2}{4K^2}\right].$$

• By Chebyshev's inequality, for each $g \in \mathcal{G}$

$$\mathbf{P}\left(|P_n(g) - P(g)| > \delta\right) \leq \frac{\mathrm{var}(g(X))}{n\delta^2} \leq \frac{K^2}{n\delta^2}$$

$$\leq \frac{K^2}{4K^2 \log N} \leq \frac{1}{2}.$$

• Hence, by symmetrization with probabilities

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > 4\delta\right) \leq 4\mathbf{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \delta\right) \leq 8\exp\left[-\frac{n\delta^2}{4K^2}\right].$$

□

Definition 4.1.2. The **envelope** G of a collection of functions $\mathcal{G}$ is defined by

$$G(x) = \sup_{g \in \mathcal{G}} |g(x)|, \ x \in \mathcal{X}.$$

In Exercise 2 of this chapter, the assumption $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq K$ used in Lemma 4.1.1, is weakened to $P(G) < \infty$.

Definition 4.1.3. Let S be some subset of a metric space (Λ, d). For $\delta > 0$, the δ-**covering number** $N(\delta, S, d)$ of S is the minumum number of balls with radius δ, necessary to cover S, i.e. the smallest value of N, such that there exist $s_1, \ldots, s_N$ in Λ with

$$\min_{j=1,\ldots,N} d(s, s_j) \leq \delta, \ \forall \ s \in S.$$

The set $s_1, \ldots, s_N$ is then called a δ-**covering** of S. The logarithm $\log N(\cdot, S, d)$ of the covering number is called the **entropy** of S.

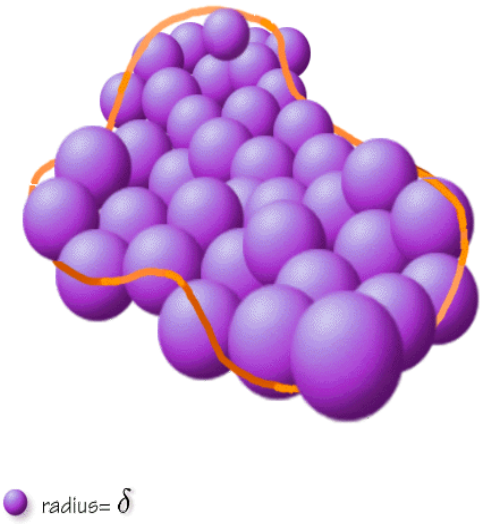

Figure 3

Notation. Let

$$d_{1,n}(g, \tilde{g}) = P_n(|g - \tilde{g}|).$$

Theorem 4.1.4. *Suppose*

$$\|g\|_\infty \leq K, \ \forall \ g \in \mathcal{G}.$$

Assume moreover that

$$\frac{1}{n} \log N(\delta, \mathcal{G}, d_{1,n}) \to^{\mathbf{P}} 0.$$

Then

$$\|P_n - P\|_{\mathcal{G}} \to^{\mathbf{P}} 0.$$

Proof. Let $\delta > 0$. Let $g_1, \ldots, g_N$, with $N = N(\delta, \mathcal{G}, d_{1,n})$, be a δ-covering of $\mathcal{G}$.
• When $P_n(|g - g_j|) \leq \delta$, we have

$$|P_n^\sigma(g)| \leq |P_n^\sigma(g_j)| + \delta.$$

So

$$\|P_n^\sigma\|_\mathcal{G} \leq \max_{j=1,\ldots N} |P_n^\sigma(g_j)| + \delta.$$

• By Hoeffding's inequality and the elementary observation, for

$$\delta \geq 2K\sqrt{\frac{\log N}{n}},$$

we have

$$\mathbf{P_X}\left(\max_{j=1,\ldots,N} |P_n^\sigma(g_j)| > \delta\right) \leq 2\exp\left[-\frac{n\delta^2}{4K^2}\right].$$

• Conclude that for

$$\delta \geq 2K\sqrt{\frac{\log N}{n}},$$

we have

$$\mathbf{P_X}\left(\|P_n^\sigma\|_\mathcal{G} > 2\delta\right) \leq 2\exp\left[-\frac{n\delta^2}{4K^2}\right].$$

• But then

$$\mathbf{P}\left(\|P_n^\sigma\|_\mathcal{G} > 2\delta\right) \leq 2\exp\left[-\frac{n\delta^2}{4K^2}\right] + \mathbf{P}\left(2K\sqrt{\frac{\log N(\delta, \mathcal{G}, d_{1,n})}{n}} > \delta\right).$$

• We thus get as $n \to \infty$,

$$\mathbf{P}\left(\|P_n^\sigma\|_\mathcal{G} > 2\delta\right) \to 0.$$

• No, use the symmetrization with probabilities to conclude

$$\mathbf{P}\left(\|P_n - P\|_\mathcal{G} > 8\delta\right) \to 0.$$

Since δ is arbitrary, this concludes the proof.

□

Again, the assumption $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq K$ used in Theorem 4.1.4, can be weakened to $P(G) < \infty$ (G being the envelope of $\mathcal{G}$). See Exercise 3 of this chapter.

4.2. Classes of sets. Let $\mathcal{D}$ be a collection of subsets of $\mathcal{X}$, and let $\{\xi_1, \ldots, \xi_n\}$ be n points in $\mathcal{X}$.

Definition 4.2.1. We write

$$\begin{aligned}\triangle^\mathcal{D}(\xi_1, \ldots, \xi_n) &= \text{card}(\{D \cap \{\xi_1, \ldots, \xi_n\} : \ D \in \mathcal{D}\} \\ &= \text{the number of subsets of } \{\xi_1, \ldots, \xi_n\} \text{ that } \mathcal{D} \text{ can distinguish.}\end{aligned}$$

That is, count the number of sets in $\mathcal{D}$, when two sets D_1 and D_2 are considered as equal if $D_1 \triangle D_2 \cap \{\xi_1, \ldots, \xi_n\} = \emptyset$. Here

$$D_1 \triangle D_2 = (D_1 \cap D_2^c) \cup (D_1^c \cap D_2)$$

is the symmetric difference between D_1 and D_2.

Remark. For our purposes, we will not need to calculate $\triangle^\mathcal{D}(\xi_1, \ldots, \xi_n)$ **exactly**, but only a good enough upper bound.

Example. Let $\mathcal{X} = \mathbf{R}$ and

$$\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}\}.$$

Then for all $\{\xi_1, \ldots, \xi_n\} \subset \mathbf{R}$

$$\triangle^\mathcal{D}(\xi_1, \ldots, \xi_n) \leq n + 1.$$

Example. Let $\mathcal{D}$ be the collection of all finite subsets of $\mathcal{X}$. Then, if the points $\xi_1, \ldots, \xi_n$ are distinct,

$$\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) = 2^n.$$

Theorem 4.2.2. (Vapnik and Chervonenkis (1971)). *We have*

$$\frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathbf{P}} 0,$$

if and only if

$$\sup_{D \in \mathcal{D}} |P_n(D) - P(D)| \to^{\mathbf{P}} 0.$$

Proof of the if-part. This follows from applying Theorem 4.1.4 to $\mathcal{G} = \{1_D : \ D \in \mathcal{D}\}$. Note that a class of indicator functions is uniformly bounded by 1, i.e. we can take $K = 1$ in Theorem 4.1.4. Define now

$$d_{\infty,n}(g, \tilde{g}) = \max_{i=1,\ldots,n} |g(X_i) - \tilde{g}(X_i)|.$$

Then $d_{1,n} \leq d_{\infty,n}$, so also

$$N(\cdot, \mathcal{G}, d_{1,n}) \leq N(\cdot, \mathcal{G}, d_{\infty,n}).$$

But for $0 < \delta < 1$,

$$N(\delta, \{1_D : \ D \in \mathcal{D}\}, d_{\infty,n}) = \Delta^{\mathcal{D}}(X_1, \ldots, X_n).$$

So indeed, if $\frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathbf{P}} 0$, then also $\frac{1}{n} \log N(\delta, \{1_D : \ D \in \mathcal{D}\}, d_{1,n}) \leq \frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \to^{\mathbf{P}} 0$.

□

4.3. Vapnik-Chervonenkis classes.

Definition 4.3.1. Let

$$m^{\mathcal{D}}(n) = \sup\{\triangle^{\mathcal{D}}(\xi_1, \ldots, \xi_n) : \ \xi_1, \ldots, \xi_n \in \mathcal{X}\}.$$

We say that $\mathcal{D}$ is a **Vapnik-Chervonenkis** (VC) class if for certain constants c and V, and for all n,

$$m^{\mathcal{D}}(n) \leq cn^V,$$

i.e., if $m^{\mathcal{D}}(n)$ does not grow faster than a polynomial in n.

Important conclusion: For sets, VC $\Rightarrow$ GC.

Examples.
a) $\mathcal{X} = \mathbf{R}$, $\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}\}$. Since $m^{\mathcal{D}}(n) \leq n + 1$, $\mathcal{D}$ is VC.
b) $\mathcal{X} = \mathbf{R}^d$, $\mathcal{D} = \{1_{(-\infty,t]} : \ t \in \mathbf{R}^d\}$. Since $m^{\mathcal{D}}(n) \leq (n+1)^d$, $\mathcal{D}$ is VC.
c) $\mathcal{X} = \mathbf{R}^d$, $\mathcal{D} = \{\{x : \ \theta^T x > t\}, \begin{pmatrix} \theta \\ t \end{pmatrix} \in \mathbf{R}^{d+1}\}$. Since $m^{\mathcal{D}}(n) \leq 2^d \binom{n}{d}$, $\mathcal{D}$ is VC.

The VC property is closed under measure theoretic operations:

Lemma 4.3.2. *Let $\mathcal{D}$, $\mathcal{D}_1$ and $\mathcal{D}_2$ be VC. Then the following classes are also VC:*
(i) $\mathcal{D}^c = \{D^c : D \in \mathcal{D}\}$,
(ii) $\mathcal{D}_1 \cap \mathcal{D}_2 = \{D_1 \cap D_2 : \ D_1 \in \mathcal{D}_1, \ D_2 \in \mathcal{D}_2\}$,
(iii) $\mathcal{D}_1 \cup \mathcal{D}_2 = \{D_1 \cap D_2 : \ D_1 \in \mathcal{D}_1, \ D_2 \in \mathcal{D}_2\}$.
Proof. Exercise.

□

Examples.
- the class of intersections of two halfspaces,
- all ellipsoids,
- all half-ellipsoids,
- in $\mathbf{R}$, the class $\left\{\{x : \ \theta_1 x + \ldots + \theta_r x^r \leq t\} : \ \begin{pmatrix} \theta \\ t \end{pmatrix} \in \mathbf{R}^{r+1}\right\}$.

There are classes that are GC, but not VC.

Example. Let $\mathcal{X} = [0,1]^2$, and let $\mathcal{D}$ be the collection of all convex subsets of $\mathcal{X}$. Then $\mathcal{D}$ is not VC, but when P is uniform, $\mathcal{D}$ **is** GC.

Definition 4.3.3. The VC dimension of $\mathcal{D}$ is

$$V(\mathcal{D}) = \inf\{n : \ m^{\mathcal{D}}(n) < 2^n\}.$$

The following Lemma is nice to know, but to avoid digressions, we will not provide a proof.

Lemma 4.3.4. *We have that $\mathcal{D}$ is VC if and only if $V(\mathcal{D}) < \infty$. In fact, we have for $V = V(\mathcal{D})$, $m^{\mathcal{D}}(n) \leq \sum_{k=0}^{V} \binom{n}{k}$.* □.

4.4. VC graph classes of functions.

Definition 4.4.1. The **subgraph** of a function $g : \mathcal{X} \to \mathbf{R}$ is

$$\text{subgraph}(g) = \{(x,t) \in \mathcal{X} \times \mathbf{R} : \ g(x) \geq t\}.$$

A collection of functions $\mathcal{G}$ is called a VC class if the subgraphs $\{\text{subgraph}(g) : \ g \in \mathcal{G}\}$ form a VC class.

Example. $\mathcal{G} = \{1_D : D \in \mathcal{D}\}$ is GC if $\mathcal{D}$ is GC.

Examples ($\mathcal{X} = \mathbf{R}^d$).

a) $\mathcal{G} = \{g(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d : \ \theta \in \mathbf{R}^{d+1}\}$,

b) $\mathcal{G} = \{g(x) = |\theta_0 + \theta_1 x_1 + \ldots + \theta_d x_d| : \ \theta \in \mathbf{R}^{d+1}\}$.

c) $d = 1$, $\mathcal{G} = \left\{ g(x) = \begin{cases} a + bx & \text{if } x \leq c \\ d + ex & \text{if } x > c \end{cases}, \ \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} \in \mathbf{R}^5 \right\}$,

d) $d = 1$, $\mathcal{G} = \{g(x) = \mathrm{e}^{\theta x} : \ \theta \in \mathbf{R}\}$.

Definition 4.4.1. Let S be some subset of a metric space (Λ, d). For $\delta > 0$, the δ-**packing number** $D(\delta, S, d)$ of S is the largest value of N, such that there exist $s_1, \ldots, s_N$ in S with

$$d(s_k, s_j) > \delta, \ \forall \ k \neq j.$$

Note. For all $\delta > 0$,

$$N(\delta, S, d) \leq D(\delta, S, d).$$

Theorem 4.4.2. *Let Q be any probability measure on $\mathcal{X}$. Define $d_{1,Q}(g, \tilde{g}) = Q(|g - \tilde{g}|)$. For a VC class $\mathcal{G}$ with VC dimension V, we have for a constant A depending only on V,*

$$N(\delta Q(G), \mathcal{G}, d_{1,Q}) \leq \max(A\delta^{-2V}, \mathrm{e}^{\delta/4}), \ \forall \ \delta > 0$$

.

Proof. Without loss of generality, assume $Q(G) = 1$. Choose $S \in \mathcal{X}$ with distribution $dQ_S = G dQ$. Given $S = s$, choose T uniformly in the interval $[-G(s), G(s)]$. Let $g_1, \ldots, g_N$ be a maximal set in $\mathcal{G}$, such that $Q(|g_j - g_k|) > \delta$ for $j \neq k$. Consider a pair $j \neq k$. Given $S = s$, the probability that T falls in between the two graphs of g_j and g_k is

$$\frac{|g_j(s) - g_k(s)|}{2G(s)}.$$

So the unconditional probability that T falls in between the two graphs of g_j and g_k is

$$\int \frac{|g_j(s) - g_k(s)|}{2G(s)} dQ_S(s) = \frac{Q(|g_j - g_k|)}{2} \geq \frac{\delta}{2}.$$

Now, choose n independent copies $\{(S_i, T_i)\}_{i=1}^n$ of (T, S). The probability that none of these fall in between the graphs of g_j and g_k is then at most

$$(1 - \delta/2)^n.$$

The probability that for some $j \neq k$, none of these fall in between the graphs of g_j and g_k is then at most

$$\binom{N}{2}(1 - \delta/2)^n \leq \frac{1}{2}\exp\left[2\log N - \frac{n\delta}{2}\right] \leq \frac{1}{2} < 1,$$

when we choose n the smallest integer such that

$$n \geq \frac{4\log N}{\delta}.$$

So for such a value of n, with positive probability, for any $j \neq k$, some of the T_i fall in between the graphs of g_j and g_k. Therefore, we must have

$$N \leq cn^V.$$

But then, for $N \geq \exp[\delta/4]$,

$$N \leq c\left(\frac{4\log N}{\delta} + 1\right)^V \leq c\left(\frac{8\log N}{\delta}\right)^V = c\left(\frac{16V\log N^{\frac{1}{2V}}}{\delta}\right)^V$$

$$\leq c\left(\frac{16V}{\delta}\right)^V N^{\frac{1}{2}}.$$

So

$$N \leq c^2\left(\frac{16V}{\delta}\right)^{2V}.$$

□

Corollary 4.4.3. *Suppose $\mathcal{G}$ is VC and that $\int G dP < \infty$. Then by Theorem 4.4.2 and Theorem 4.1.4, we have $\|P_n - P\|_{\mathcal{G}} \to^{\mathbf{P}} 0$.*

4.5. Exercises.

Exercise 1. Let $\mathcal{G}$ be a finite class of functions, with cardinality $|\mathcal{G}| := N > 1$. Suppose that for some finite constant K,

$$\max_{g\in\mathcal{G}} \|g\|_\infty \leq K.$$

Use Bernstein's inequality to show that for

$$\delta^2 \geq \frac{4\log N}{n}\left[\delta K + K^2\right]$$

one has

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > \delta\right) \leq 2\exp\left[-\frac{n\delta^2}{4(\delta K + K^2)}\right].$$

Exercise 2. Let $\mathcal{G}$ be a finite class of functions, with cardinality $|\mathcal{G}| := N > 1$. Suppose that $\mathcal{G}$ has envelope G satisfying

$$P(G) < \infty.$$

Let $0 < \delta < 1$, and take K large enough, so that

$$P(G1\{G > K\}) \leq \delta^2.$$

Show that for

$$\delta \geq 4K\sqrt{\frac{\log N}{n}},$$

$$\mathbf{P}\left(\|P_n - P\|_{\mathcal{G}} > 4\delta\right) \leq 8\exp\left[-\frac{n\delta^2}{16K^2}\right] + \delta.$$

Hint: use

$$|P_n(g) - P(g)| \leq |P_n(g\mathrm{l}\{G \leq K\}) - P((g\mathrm{l}\{G \leq K\})| \\ + P_n(G\mathrm{l}\{G > K\}) + P(G\mathrm{l}\{G > K\}).$$

Exercise 3. Let $\mathcal{G}$ be a class of functions, with envelope G, satisfying $P(G) < \infty$ and $\frac{1}{n}\log N(\delta, \mathcal{G}, d_{1,n}) \to^{\mathbf{P}} 0$. Show that $\|P_n - P\|_{\mathcal{G}} \to^{\mathbf{P}} 0$.

Exercise 4.

Are the following classes of sets (functions) VC? Why (not)?

1) The class of all rectangles in $\mathbf{R}^d$.

2) The class of all monotone functions on $\mathbf{R}$.

3) The class of functions on $[0,1]$ given by

$$\mathcal{G} = \{g(x) = a\mathrm{e}^{bx} + c\mathrm{e}^{dx} : \ (a,b,c,d) \in [0,1]^4\}.$$

4) The class of all sections in $\mathbf{R}^2$ (a section is of the form $\{(x_1,x_2) : \ x_1 = a_1 + r\sin t, x_2 = a_2 + r\cos t, \ \theta_1 \leq t \leq \theta_2\}$, for some $(a_1,a_2) \in \mathbf{R}^2$, some $r > 0$, and some $0 \leq \theta_1 \leq \theta_2 \leq 2\pi$).

5) The class of all star-shaped sets in $\mathbf{R}^2$ (a set D is star-shaped if for some $a \in D$ and all $b \in D$ also all points on the line segment joining a and b are in D).

Exercise 5.

Let $\mathcal{G}$ be the class of all functions g on $[0,1]$ with derivative $\dot{g}$ satisfying $|\dot{g}| \leq 1$. Check that $\mathcal{G}$ is not VC. Show that $\mathcal{G}$ is GC by using partial integration and the Glivenko-Cantelli Theorem for the empirical distribution function.

5. M-estimators

5.1 What is an M-estimator? Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X with values in $\mathcal{X}$ and with distribution P.

Let Θ be a parameter space (a subset of some metric space) and let for $\theta \in \Theta$,

$$\gamma_\theta : \mathcal{X} \to \mathbf{R},$$

be some loss function. We assume $P(|\gamma_\theta|) < \infty$ for all $\theta \in \Theta$. We estimate the unknown parameter

$$\theta_0 := \arg\min_{\theta \in \Theta} P(\gamma_\theta),$$

by the M-estimator

$$\hat{\theta}_n := \arg\min_{\theta \in \Theta} P_n(\gamma_\theta).$$

We assume that θ_0 exists and is unique and that $\hat{\theta}_n$ exists.

Examples.
(i) Location estimators. $\mathcal{X} = \mathbf{R}$, $\Theta = \mathbf{R}$, and
(i.a) $\gamma_\theta(x) = (x-\theta)^2$ (estimating the mean),
(i.b) $\gamma_\theta(x) = |x-\theta|$ (estimating the median).
(ii) Maximum likelihood. $\{p_\theta : \theta \in \Theta\}$ family of densities w.r.t. σ-finite dominating measure μ, and

$$\gamma_\theta = -\log p_\theta.$$

If $dP/d\mu = p_{\theta_0}$, $\theta_0 \in \Theta$, then indeed θ_0 is a minimizer of $P(\gamma_\theta)$, $\theta \in \Theta$.
(ii.a) Poisson distribution:

$$p_\theta(x) = \mathrm{e}^\theta \frac{\theta^x}{x!}, \ \theta > 0, \ x \in \{1, 2, \ldots\}.$$

(ii.b) Logistic distribution:

$$p_\theta(x) = \frac{\mathrm{e}^{\theta - x}}{(1 + \mathrm{e}^{\theta - x})^2}, \ \theta \in \mathbf{R}, \ x \in \mathbf{R}.$$

5.2. Consistency. Define for $\theta \in \Theta$,

$$R(\theta) = P(\gamma_\theta),$$

and

$$R_n(\theta) = P_n(\gamma_\theta).$$

We first present an easy proposition with a too stringent condition (•).

Proposition 5.2.1. *Suppose that $\theta \mapsto R(\theta)$ is continuous. Assume moreover that*

$$(\bullet) \qquad \sup_{\theta \in \Theta} |R_n(\theta) - R(\theta)| \to^{\mathbf{P}} 0,$$

i.e., that $\{\gamma_\theta : \theta \in \Theta\}$ is a GC class. Then $\hat{\theta}_n \to^{\mathbf{P}} \theta_0$.

Proof. We have

$$\begin{aligned} 0 &\leq R(\hat{\theta}_n) - R(\theta_0) \\ &= [R(\hat{\theta}_n) - R(\theta_0)] - [R_n(\hat{\theta}_n) - R_n(\theta_0)] + [R_n(\hat{\theta}_n) - R_n(\theta_0)] \\ &\leq [R(\hat{\theta}_n) - R(\theta_0)] - [R_n(\hat{\theta}_n) - R_n(\theta_0)] \to^{\mathbf{P}} 0. \end{aligned}$$

So $R(\hat{\theta}_n) \to^{\mathbf{P}} R(\theta_0)$ and hence $\hat{\theta}_n \to^{\mathbf{P}} \theta_0$.

□

The assumption (•) is hardly ever met, because it is close to requiring compactness of Θ.

Lemma 5.2.2. *Suppose that (Θ, d) is compact and that $\theta \mapsto \gamma_\theta$ is continuous. Moreover, assume that $P(G) < \infty$, where*

$$G = \sup_{\theta \in \Theta} |\gamma_\theta|.$$

Then

$$\sup_{\theta \in \Theta} |R_n(\theta) - R(\theta)| \to^{\mathbf{P}} 0.$$

Proof. Let

$$w(\theta, \rho) = \sup_{\{\tilde{\theta}:\ d(\tilde{\theta}, \theta) < \rho\}} |\gamma_\theta - \gamma_{\tilde{\theta}}|.$$

Then

$$w(\theta, \rho) \to 0,\ \rho \to 0.$$

By dominated convergence

$$P(w(\theta, \rho)) \to 0.$$

Let $\delta > 0$ be arbitrary. Take ρ_θ in such a way that

$$P(w(\theta, \rho_\theta)) \leq \delta.$$

Let $B_\theta = \{\tilde{\theta}:\ d(\tilde{\theta}, \theta) < \rho_\theta\}$ and let $B_{\theta_1}, \ldots, B_{\theta_N}$ be a finite cover of Θ. Then for $\mathcal{G} = \{\gamma_\theta :\ \theta \in \Theta\}$,

$$\mathbf{P}(N(2\delta, \mathcal{G}, d_{1.n}) > N) \to 0.$$

So the result follows from Theorem 4.1.4. □

We give a lemma, which replaces compactness by a convexity assumption.

Lemma 5.2.3. *Suppose that Θ is a convex subset of $\mathbf{R}^r$, and that $\theta \mapsto \gamma_\theta$, $\theta \in \Theta$ is continuous and convex. Suppose $P(G_\epsilon) < \infty$ for some $\epsilon > 0$, where*

$$G_\epsilon = \sup_{\|\theta - \theta_0\| \leq \epsilon} |\gamma_\theta|.$$

Then $\hat{\theta}_n \to^{\mathbf{P}} \theta_0$.

Proof. Because Θ is finite-dimensional, the set $\{\|\theta - \theta_0\| \leq \epsilon\}$ is compact. So by Lemma 5.2.2,

$$\sup_{\|\theta - \theta_0\| \leq \epsilon} |R_n(\theta) - R(\theta)| \to^{\mathbf{P}} 0.$$

Define

$$\alpha = \frac{\epsilon}{\epsilon + \|\hat{\theta}_n - \theta_0\|}$$

and

$$\tilde{\theta}_n = \alpha \hat{\theta}_n + (1 - \alpha)\theta_0.$$

Then

$$\|\tilde{\theta}_n - \theta_0\| \leq \epsilon.$$

Moreover,

$$R_n(\tilde{\theta}_n) \leq \alpha R_n(\hat{\theta}_n) + (1 - \alpha) R_n(\theta_0) \leq R_n(\theta_0).$$

It follows from the arguments used in the proof of Proposition 5.2.1, that $\|\tilde{\theta}_n - \theta_0\| \to^{\mathbf{P}} 0$. But then also

$$\|\hat{\theta}_n - \theta_0\| = \frac{\epsilon \|\tilde{\theta}_n - \theta_0\|}{\epsilon - \|\tilde{\theta}_n - \theta_0\|} \to^{\mathbf{P}} 0.$$

□

5.3. Exercises.

Exercise 1. Let $Y \in \{0,1\}$ be a binary response variable and $Z \in \mathbf{R}$ be a covariable. Assume the logistic regression model

$$P_{\theta_0}(Y = 1|Z = z) = \frac{1}{1 + \exp[\alpha_0 + \beta_0 z]},$$

where $\theta_0 = (\alpha_0, \beta_0) \in \mathbf{R}^2$ is an unknown parameter. Let $\{(Y_i, Z_i)\}_{i=1}^n$ be i.i.d. copies of (Y, Z). Show consistency of the MLE of θ_0.

Exercise 2. Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with density $f_0 = dP/d\mu$ on [0,1]. Here, μ is Lebesgue measure on $[0, 1]$. Suppose it is given that $f_0 \in \mathcal{F}$, with $\mathcal{F}$ the set of all decreasing densities bounded from above by 2 and from below by 1/2. Let $\hat{f}_n$ be the MLE. Can you show consistency of $\hat{f}_n$? For what metric?

6. Uniform central limit theorems

After having studied uniform laws of large numbers, a natural question is: can we also prove uniform central limit theorems? It turns out that precisely defining what a uniform central limit theorem is, is quite involved, and actually beyond our scope. In Sections 6.1-6.4 we will therefore only briefly indicate the results, and not present any proofs. These sections only reveal a glimps of the topic of weak convergence on abstract spaces. The thing to remember from them is the concept asymptotic continuity, *because we will use that concept in our statistical applications. In Section 6.5 we will prove that the empirical process indexed by a VC graph class is asymptotically continuous. This result will be a corollary of another result of interest to (theoretical) statisticians: a result relating the increments of the empirical process to the entropy of* $\mathcal{G}$.

6.1. Real-valued random variables. Let $\mathcal{X} = \mathbf{R}$.

Central limit theorem in R. *Suppose* $EX = \mu$, *and* $\mathrm{var}(X) = \sigma^2$ *exist. Then*

$$\mathbf{P}\left(\sqrt{n}(\frac{\bar{X}_n - \mu}{\sigma}) \leq z\right) \to \Phi(z), \ \textit{for all } z,$$

where Φ *is the standard normal distribution function.* □.

Notation.

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to^{\mathcal{L}} \mathcal{N}(0,1),$$

or

$$\sqrt{n}(\bar{X}_n - \mu) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

6.2. $\mathbf{R}^d$-valued random variables. Let $X_1, X_2, \ldots$ be i.i.d. $\mathbf{R}^d$-valued random variables copies of X, $(X \in \mathcal{X} = \mathbf{R}^d)$, with expectation $\mu = EX$, and covariance matrix $\Sigma = EXX^T - \mu\mu^T$.

Central limit theorem in $\mathbf{R}^d$. *We have*

$$\sqrt{n}(\bar{X}_n - \mu) \to^{\mathcal{L}} \mathcal{N}(0, \Sigma),$$

i.e.

$$\sqrt{n}\left[a^T(\bar{X}_n - \mu)\right] \to^{\mathcal{L}} \mathcal{N}(0, a^T\Sigma a), \ \textit{for all } a \in \mathbf{R}^d.$$

□.

6.3. Donsker's Theorem. Let $\mathcal{X} = \mathbf{R}$. Recall the definition of the distribution function F and the empirical distribution function F_n:

$$F(t) = P(X \leq t), \ t \in \mathbf{R},$$

$$F_n(t) = \frac{1}{n}\#\{X_i \leq t, \ 1 \leq i \leq n\}, \ t \in \mathbf{R}.$$

Define

$$W_n(t) := \sqrt{n}(F_n(t) - F(t)), \ t \in \mathbf{R}.$$

By the central limit theorem in $\mathbf{R}$ (Section 6.1), for all t

$$W_n(t) \to^{\mathcal{L}} \mathcal{N}\left(0, F(t)(1 - F(t))\right).$$

Also, by the central limit theorem in $\mathbf{R}^2$ (Section 6.2), for all $s < t$,

$$\begin{pmatrix} W_n(s) \\ W_n(t) \end{pmatrix} \to^{\mathcal{L}} \mathcal{N}(0, \Sigma(s,t)),$$

where

$$\Sigma(s,t) = \begin{pmatrix} F(s)(1-F(s)) & F(s)(1-F(t)) \\ F(s)(1-F(t)) & F(t)(1-F(t)) \end{pmatrix}.$$

We are now going to consider the **stochastic process** $W_n = \{W_n(t) : \ t \in \mathbf{R}\}$. The process W_n is called the (classical) empirical process.

Definition 6.3.1. Let $\mathcal{K}_0$ be the collection of bounded functions on $[0,1]$ The stochastic process $B(\cdot) \in \mathcal{K}_0$, is called the standard **Brownian bridge** if
- $B(0) = B(1) = 0$,
- for all $r \geq 1$ and all $t_1, \ldots, t_r \in (0,1)$, the vector $\begin{pmatrix} B(t_1) \\ \vdots \\ B(t_r) \end{pmatrix}$ is multivariate normal with mean zero,
- for all $s \leq t$, $\mathrm{cov}(B(s), B(t)) = s(1-t)$.
- the sample paths of B are a.s. continuous.

We now consider the process W_F defined as

$$W_F(t) = B(F(t)) : \ t \in \mathbf{R}.$$

Thus, $W_F = B \circ F$.

Donsker's theorem. *Consider W_n and W_F as elements of the space $\mathcal{K}$ of bounded functions on $\mathcal{R}$. We have*

$$W_n \to^{\mathcal{L}} W_F,$$

that is,

$$\mathbf{E}f(W_n) \to \mathbf{E}f(W_F),$$

for all continuous and bounded functions f. □

Reflection. Suppose F is continuous. Then, since B is almost surely continuous, also $W_F = B \circ F$ is almost surely continuous. So W_n must be approximately continuous as well in some sense. Indeed, we have for any t and any sequence t_n converging to t,

$$|W_n(t_n) - W_n(t)| \to^{\mathbf{P}} 0.$$

This is called **asymptotic continuity**.

6.4. Donsker classes. Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of a random variable X, with values in the space $\mathcal{X}$, and with distribution P. Consider a class $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbf{R}$. The (theoretical) mean of a function g is

$$P(g) := Eg(X),$$

and the (empirical) average (based on the n observations $X_1, \ldots, X_n$) is

$$P_n(g) := \frac{1}{n} \sum_{i=1}^{n} g(X_i).$$

Here P_n is the empirical distribution (based on $X_1, \ldots, X_n$).

Definition 6.4.1. The **empirical process** indexed by $\mathcal{G}$ is

$$\nu_n(g) := \sqrt{n} \left(P_n(g) - P(g) \right), \ g \in \mathcal{G}.$$

Let us recall the central limit theorem for g fixed. Denote the variance of $g(X)$ by

$$\sigma^2(g) := \mathrm{var}(g(X)) = P(g^2) - (P(g))^2.$$

If $\sigma^2(g) < \infty$, we have

$$\nu_n(g) \to^{\mathcal{L}} \mathcal{N}(0, \sigma^2(g)).$$

The central limit theorem also holds for finitely many g simultaneously. Let g_k and g_l be two functions and denote the covariance between $g_k(X)$ and $g_l(X)$ by

$$\sigma(g_k, g_l) := \mathrm{cov}(g_k(X), g_l(X)) = P(g_k g_l) - P(g_k)P(g_l).$$

Then, whenever $\sigma^2(g_k) < \infty$ for $k = 1, \ldots, r$,

$$\begin{pmatrix} \nu_n(g_1) \\ \vdots \\ \nu_n(g_r) \end{pmatrix} \to^{\mathcal{L}} \mathcal{N}(0, \Sigma_{g_1,\ldots,g_r}),$$

where $\Sigma_{g_1,\ldots,g_r}$ is the variance-covariance matrix

$$(*) \qquad \Sigma_{g_1,\ldots,g_r} = \begin{pmatrix} \sigma^2(g_1) & \ldots & \sigma(g_1, g_r) \\ \vdots & \ddots & \vdots \\ \sigma(g_1, g_r) & \ldots & \sigma^2(g_r) \end{pmatrix}.$$

Definition 6.4.2. Let ν be a Gaussian process indexed by $\mathcal{G}$. Assume that for each $r \in \mathbf{N}$ and for each finite collection $\{g_1, \ldots, g_r\} \subset \mathcal{G}$, the r-dimensional vector

$$\begin{pmatrix} \nu(g_1) \\ \vdots \\ \nu(g_r) \end{pmatrix}$$

has a $\mathcal{N}(0, \Sigma_{g_1,\ldots,g_r})$-distribution, with $\Sigma_{g_1,\ldots,g_r}$ defined in (*). We then call ν the **P-Brownian bridge** indexed by $\mathcal{G}$.

Definition 6.4.3. Consider ν_n and ν as bounded functions on $\mathcal{G}$. We call $\mathcal{G}$ a **P-Donsker class** if

$$\nu_n \to^{\mathcal{L}} \nu,$$

that is, if for all continuous and bounded functions f, we have

$$\mathbf{E} f(\nu_n) \to \mathbf{E} f(\nu).$$

Definition 6.4.4. The process ν_n on $\mathcal{G}$ is called **asymptotically continuous** if for all $g_0 \in \mathcal{G}$, and all (possibly random) sequences $\{g_n\} \subset \mathcal{G}$ with $\sigma(g_n - g_0) \to^{\mathbf{P}} 0$, we have

$$|\nu_n(g_n) - \nu_n(g_0)| \to^{\mathbf{P}} 0.$$

We will use the notation

$$\|g\|_{2,P}^2 := P(g^2),$$

i.e., $\|\cdot\|_{2,P}$ is the $L_2(P)$-norm.

Remark. Note that $\sigma(g) \leq \|g\|_{2,P}$.

Definition 6.4.5. The class $\mathcal{G}$ is called **totally bounded** for the metric $d_{2,P}(g, \tilde{g}) := \|g - \tilde{g}\|_{2,P}$ induced by $\|\cdot\|_{2,P}$, if its entropy $\log N(\cdot, \mathcal{G}, d_{2,P})$ is finite.

Theorem 6.4.6. *Suppose that $\mathcal{G}$ is totally bounded. Then $\mathcal{G}$ is a P-Donsker class if and only if ν_n (as process on $\mathcal{G}$) is asymptotically continuous.*

□

6.5. Chaining and the increments of the empirical process.

6.5.1. Chaining. We will consider the increments of the symmetrized empirical process in Section 6.5.2. There, we will work conditionally on $\mathbf{X} = (X_1, \ldots, X_n)$. We now describe the chaining technique in this context

Let

$$\|g\|_{2,n}^2 := P_n(g^2).$$

Let $d_{2,n}(g, \tilde{g}) = \|g - \tilde{g}\|_{2,n}$, i.e., $d_{2,n}$ is the metric induced by $\|\cdot\|_{2,n}$. Suppose that $\|g\|_{2,n} \leq R$ for all $g \in \mathcal{G}$. For notational convenience, we index the functions in $\mathcal{G}$ by a parameter $\theta \in \Theta$: $\mathcal{G} = \{g_\theta : \theta \in \Theta\}$.

Let for $s = 0, 1, 2, \ldots$, $\{g_j^s\}_{j=1}^{N_s}$ be a minimal $2^{-s}R$-covering set of $(\mathcal{G}, d_{2,n})$. So $N_s = N(2^{-s}R, \mathcal{G}, d_{2,n})$, and for each θ, there exists a $g_\theta^s \in \{g_1^s, \ldots, g_{N_s}^s\}$ such that $\|g_\theta - g_\theta^s\|_{2,n} \leq 2^{-s}R$. We use the parameter θ here to indicate which function in the covering set approximates a particular g. We may choose $g_\theta^0 \equiv 0$, since $\|g_\theta\|_{2,n} \leq R$. Then for any S,

$$g_\theta = \sum_{s=1}^{S}(g_\theta^s - g_\theta^{s-1}) + (g_\theta - g_\theta^S).$$

One can think of this as telescoping from g_θ to g_θ^S, i.e. we follow a path taking smaller and smaller steps. As $S \to \infty$, we have $\max_{1\leq i\leq n} |g_\theta(X_i) - g_\theta^S(X_i)| \to 0$. The term $\sum_{s=1}^{\infty}(g_\theta^s - g_\theta^{s-1})$ can be handled by exploiting the fact that as θ varies, each summand involves only finitely many functions.

6.5.2. Increments of the symmetrized process.
We use the notation $a \vee b = \max\{a, b\}$ ($a \wedge b = \min\{a, b\}$).

Lemma 6.5.2.1. *On the set* $\sup_{g\in\mathcal{G}} \|g\|_{2,n} \leq R$, *and*

$$\sqrt{n}\delta \geq \left(14\sum_{s=1}^{\infty} 2^{-s}R\sqrt{\log N(2^{-s}R, \mathcal{G}, d_{2,n})} \vee 70R\log 2\right),$$

we have

$$\mathbf{P_X}\left(\sup_{g\in\mathcal{G}} |P_n^\sigma(g)| \geq \delta\right) \leq 4\exp[-\frac{n\delta^2}{(70R)^2}].$$

Proof. Let $\{g_j^s\}_{j=1}^{N_s}$ be a minimal $2^{-s}R$-covering set of $\mathcal{G}$, $s = 0, 1, \ldots$. So $N_s = N(2^{-s}R, \mathcal{G}, d_{2,n})$. Now, use chaining. Write $g_\theta = \sum_{s=1}^{\infty}(g_\theta^s - g_\theta^{s-1})$. Note that by the triangle inequality,

$$\|g_\theta^s - g_\theta^{s-1}\|_{2,n} \leq \|g_\theta^s - g_\theta\|_{2,n} + \|g_\theta - g_\theta^{s-1}\|_{2,n}$$
$$\leq 2^{-s}R + 2^{-s+1}R = 3(2^{-s}R).$$

Let η_s be positive numbers satisfying $\sum_{s=1}^{\infty}\eta_s \leq 1$. Then

$$\mathbf{P_X}\left(\sup_{\theta\in\Theta} |P_n^\sigma(g_\theta^s - g_\theta^{s-1})| \geq \delta\right)$$
$$\leq \sum_{s=1}^{\infty} \mathbf{P_X}\left(\sup_{\theta\in\Theta} |\frac{1}{n}P_n^\sigma(g_\theta^s - g_\theta^{s-1})| \geq \delta\eta_s\right)$$
$$\leq \sum_{s=1}^{\infty} 2\exp[2\log N_s - \frac{n\delta^2\eta_s^2}{18\times 2^{-2s}R^2}].$$

What is a good choice for η_s? We take

$$\eta_s = \frac{7\times 2^{-s}R\sqrt{\log N_s}}{\sqrt{n}\delta} \vee \frac{2^{-s}\sqrt{s}}{8}.$$

Then indeed, by our condition on $\sqrt{n}\delta$,

$$\sum_{s=1}^{\infty}\eta_s \leq \sum_{s=1}^{\infty}\frac{72^{-s}R\sqrt{\log N_s}}{\sqrt{n}\delta} + \sum_{s=1}^{\infty}\frac{2^{-s}\sqrt{s}}{8} \leq \frac{1}{2} + \frac{1}{2} = 1.$$

Here, we used the bound

$$\sum_{s=1}^{\infty} 2^{-s}\sqrt{s} \leq 1 + \int_1^\infty 2^{-x}\sqrt{x}dx$$
$$\leq 1 + \int_0^\infty 2^{-x}\sqrt{x}dx = 1 + (\pi/\log 2)^{1/2} \leq 4.$$

Observe that

$$\eta_s \geq \frac{7 \times 2^{-s} R\sqrt{\log N_s}}{\sqrt{n}\delta},$$

so that

$$2 \log N_s \leq \frac{2n\delta^2\eta_s^2}{49 \times 2^{-2s} R^2}.$$

Thus,

$$\sum_{s=1}^{\infty} 2\exp[2\log N_s - \frac{n\delta^2\eta_s^2}{18 \times 2^{-2s}R^2}] \leq \sum_{s=1}^{\infty} 2\exp[-\frac{13n\delta^2\eta_s^2}{49 \times 18 \times 2^{-2s}R^2}]$$

$$\leq \sum_{s=1}^{\infty} 2\exp[-\frac{2n\delta^2\eta_s^2}{49 \times 3 \times 2^{-2s}R^2}].$$

Next, invoke that $\eta_s \geq 2^{-s}\sqrt{s}/8$:

$$\sum_{s=1}^{\infty} 2\exp[-\frac{2n\delta^2\eta_s^2}{49 \times 3 \times 2^{-2s}R^2}] \leq \sum_{s=1}^{\infty} 2\exp[-\frac{n\delta^2 s}{49 \times 96R^2}]$$

$$\leq \sum_{s=1}^{\infty} 2\exp[-\frac{n\delta^2 s}{(70R)^2}] = 2(1-\exp[-\frac{n\delta^2}{(70R)^2}])^{-1}\exp[-\frac{n\delta^2}{(70R)^2}]$$

$$\leq 4\exp[-\frac{n\delta^2}{(70R)^2}],$$

where in the last inequality, we used the assumption that

$$\frac{n\delta^2}{(70R)^2} \geq \log 2.$$

Thus, we have shown that

$$\mathbf{P_X}\left(\sup_{g\in\mathcal{G}} |P_n^\sigma(g)| \geq \delta\right) \leq 4\exp[-\frac{n\delta^2}{(70R)^2}].$$

□

Remark. It is easy to see that

$$\sum_{s=1}^{\infty} 2^{-s} R\sqrt{\log N(2^{-s}R, \mathcal{G}, d_{2,n})} \leq 2\int_0^R \sqrt{\log N(u, \mathcal{G}, d_{2,n})}du.$$

6.5.3. Asymptotic equicontinuity of the empirical process.
Fix some $g_0 \in \mathcal{G}$ and let

$$\mathcal{G}(\delta) = \{g \in \mathcal{G} : \|g - g_0\|_{2,P} \leq \delta\}.$$

Lemma 6.5.3.1. *Suppose that $\mathcal{G}$ has envelope G, with $P(G^2) < \infty$, and that*

$$\frac{1}{n}\log N(\delta, \mathcal{G}, d_{2,n}) \to^{\mathbf{P}} 0.$$

Then for each $\delta > 0$ fixed (i.e., not depending on n), and for

$$\frac{a}{4} \geq 28A^{1/2}(\int_0^{2\delta} H^{1/2}(u)du \vee 2\delta),$$

we have

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{g\in\mathcal{G}(\delta)} |\nu_n(g) - \nu_n(g_0)| > a\right) \leq 16\exp[-\frac{a^2}{(140\delta)^2}]$$

$$+\limsup_{n\to\infty} \mathbf{P}\left(\sup_{u>0} \frac{H(u,\mathcal{G},d_{2,n})}{H(u)} > A\right).$$

Proof. The conditions $P(G^2) < \infty$ and imply that

$$\sup_{g\in\mathcal{G}} |\|g-g_0\|_{2,n} - \|g-g_0\|_{2,P}| \to^{\mathbf{P}} 0..$$

So eventually, for each fixed $\delta > 0$, with large probability

$$\sup_{g\in\mathcal{G}(\delta)} \|g-g_0\|_{2,n} \leq 2\delta.$$

The result now follows from Lemma 6.5.2.1. □

Our next step is proving asymptotic equicontinuity of the empirical process. This means that we shall take a small in Lemma 6.5.3.1, which is possible if the entropy integral converges. Assume that

$$\int_0^1 H^{1/2}(u)du < \infty,$$

and define

$$J(\delta) = (\int_0^\delta H^{1/2}(u)du \vee \delta).$$

Roughly speaking, the increment at g_0 of the empirical process $\nu_n(g)$ behaves like $J(\delta)$ for $\|g-g_0\| \leq \delta$. So, since $J(\delta) \to 0$ as $\delta \to 0$, the increments can be made arbitrary small by taking δ small enough.

Theorem 6.5.3.2. *Suppose that $\mathcal{G}$ has envelope G with $P(G^2) < \infty$. Suppose that*

$$\lim_{A\to\infty} \limsup_{n\to\infty} \mathbf{P}\left(\sup_{u>0} \frac{H(u,\mathcal{G},d_{2,n})}{H(u)} > A\right) = 0.$$

Also assume

$$\int_0^1 H^{1/2}(u)du < \infty.$$

Then the empirical process ν_n is asymptotically continuous at g_0, i.e., for all $\eta > 0$, there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{g\in\mathcal{G}(\delta)} |\nu_n(g) - \nu_n(g_0)| > \eta\right) < \eta. \tag{5.9}$$

Proof. Take $A \geq 1$ sufficiently large, such that

$$16\exp[-A] < \frac{\eta}{2},$$

and

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{u>0} \frac{H(u,\mathcal{G},P_n)}{H(u)} > A\right) \leq \frac{\eta}{2}.$$

Next, take δ sufficiently small, such that

$$4 \times 28A^{1/2}J(2\delta) < \eta.$$

Then by Lemma 6.5.3.1,

$$\limsup_{n\to\infty} \mathbf{P}\left(\sup_{g\in\mathcal{G}(\delta)} |\nu_n(g) - \nu_n(g_0)| > \eta\right)$$

$$\leq 16\exp[-\frac{AJ^2(2\delta)}{(2\delta)^2}] + \frac{\eta}{2}$$

$$\leq 16\exp[-A] + \frac{\eta}{2} < \eta,$$

where we used

$$J(2\delta) \geq 2\delta.$$

□

Remark. Because the conditions in Theorem 6.5.3.2 do not depend on g_0, its result holds for each g_0 i.e., we have in fact shown that ν_n is asymptotically continuous.

6.5.4. Application to VC graph classes.

Theorem 6.5.4.1. *Suppose that $\mathcal{G}$ is a VC-graph class with envelope*

$$G = \sup_{g \in \mathcal{G}} |g|$$

satisfying $P(G^2) < \infty$. Then $\{\nu_n(g) : \ g \in \mathcal{G}\}$ is asymptotically continuous, and so $\mathcal{G}$ is P-Donsker.

Proof. Apply Theorem 6.5.3.2 and Theorem 4.4.2.

□

Remark. In particular, suppose that VC -graph class $\mathcal{G}$ with square integrable envelope G is parametrized by θ in some parameter space $\Theta \subset \mathbf{R}^r$, i.e. $\mathcal{G} = \{g_\theta : \ \theta \in \Theta\}$. Let $z_n(\theta) = \nu_n(g_\theta)$. Question: do we have that for a (random) sequence θ_n with $\theta_n \to \theta_0$ (in probability), also

$$|z_n(\theta_n) - z_n(\theta_0)| \to^{\mathbf{P}} 0?$$

Indeed, if $\|g_\theta - g_{\theta_0}\|_{2,P} \to^{\mathbf{P}} 0$ as θ converges to θ_0, the answer is yes.

7. Asymptotic normality of M-estimators.

Consider an M-estimator $\hat{\theta}_n$ of a finite dimensional parameter θ_0. We will give conditions for asymptotic normality of $\hat{\theta}_n$. It turns out that these conditions in fact imply asymptotic linearity. Our first set of conditions include differentiability in θ at each x of the loss function $\gamma_\theta(x)$. The proof of asymptotic normality is then the easiest. In the second set of conditions, only differentiability in quadratic mean of γ_θ is required.

The results of the previous chapter (asymptotic continuity) supply us with an elegant way to handle remainder terms in the proofs.

In this chapter, we assume that θ_0 is an interior point of $\Theta \subset \mathbf{R}^r$. Moreoever, we assume that we already showed that $\hat{\theta}_n$ is consistent.

7.1. Asymptotic linearity.

Definition 7.1.1. The (sequence of) estimator(s) $\hat{\theta}_n$ of θ_0 is called **asymptotically linear** if we may write

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}P_n(l) + o_{\mathbf{P}}(1),$$

where

$$l = \begin{pmatrix} l_1 \\ \vdots \\ l_r \end{pmatrix} : \mathcal{X} \to \mathbf{R}^r,$$

satisfies $P(l) = 0$ and $P(l_k^2) < \infty$, $k = 1, \ldots, r$. The function l is then called the **influence function**. For the case $r = 1$, we call $\sigma^2 := P(l^2)$ the **asymptotic variance**.

Definition 7.1.2. Let $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$ be two asymptotically linear estimators of θ_0, with asymptotic variance σ_1^2 and σ_2^2 respectively. Then

$$e_{1,2} := \frac{\sigma_2^2}{\sigma_1^2}$$

is called the **asymptotic relative efficiency** (of $\hat{\theta}_{n,1}$ as compared to $\hat{\theta}_{n,2}$).

7.2. Conditions a,b and c for asymptotic normality. We start with 3 conditions a,b and c, which are easier to check but more stringent. We later relax them to conditions A,B and C.

Condition a. There exists an $\epsilon > 0$ such that $\theta \mapsto \gamma_\theta$ is differentiable for all $|\theta - \theta_0| < \epsilon$ and all x, with derivative

$$\psi_\theta(x) := \frac{\partial}{\partial \theta}\gamma_\theta(x), \ x \in \mathcal{X}.$$

Condition b. We have as $\theta \to \theta_0$,

$$P(\psi_\theta - \psi_{\theta_0}) = V(\theta - \theta_0) + o(1)|\theta - \theta_0|,$$

where V is a positive definite matrix.

Condition c. There exists an $\epsilon > 0$ such that the class

$$\{\psi_\theta : \ |\theta - \theta_0| < \epsilon\}$$

and is P-Donsker with envelope Ψ satisfying $P(\Psi^2) < \infty$. Moreover,

$$\lim_{\theta \to \theta_0} \|\psi_\theta - \psi_{\theta_0}\|_{2,P} = 0.$$

Lemma 7.2.1. *Suppose conditions a,b and c. Then $\hat{\theta}_n$ is asymptotically linear with influence function*

$$l = -V^{-1}\psi_{\theta_0},$$

so

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{\mathcal{L}} \mathcal{N}(0, V^{-1}JV^{-1}),$$

where

$$J = P(\psi_{\theta_0}\psi_{\theta_0}^T).$$

□

Proof. Recall that θ_0 is an interior point of Θ, and minimizes $P(\gamma_\theta)$, so that $P(\psi_{\theta_0}) = 0$. Because $\hat{\theta}_n$ is consistent, it is eventually a solution of the *score equations*

$$P_n(\psi_{\hat{\theta}_n}) = 0.$$

Rewrite the score equations as

$$0 = P_n(\psi_{\hat{\theta}_n}) = P_n(\psi_{\hat{\theta}_n} - \psi_{\theta_0}) + P_n(\psi_{\theta_0})$$

$$= (P_n - P)(\psi_{\hat{\theta}_n} - \psi_{\theta_0}) + P(\psi_{\hat{\theta}_n}) + P_n(\psi_{\theta_0}).$$

Now, use condition b and the asymptotic equicontinuity of $\{\psi_\theta : |\theta - \theta_0| \leq \epsilon\}$ (see Chapter 6), to obtain

$$0 = o_{\mathbf{P}}(n^{-1/2}) + V(\hat{\theta}_n - \theta_0) + o(|\hat{\theta}_n - \theta_0|) + P_n(\psi_{\theta_0}).$$

This yields

$$(\hat{\theta}_n - \theta_0) = -V^{-1}P_n(\psi_{\theta_0}) + o_{\mathbf{P}}(n^{-1/2}).$$

□

Example: Huber estimator Let $\mathcal{X} = \mathbf{R}$, $\Theta = \mathbf{R}$. The Huber estimator corresponds to the loss function

$$\gamma_\theta(x) = \gamma(x - \theta),$$

with

$$\gamma(x) = x^2 1\{|x| \leq k\} + (2k|x| - k^2)1\{|x| > k\},\ x \in \mathbf{R}.$$

Here, $0 < k < \infty$ is some fixed constant, chosen by the statistician. We will now verify a,b and c.
a)

$$\psi_\theta(x) = \begin{cases} +2k & \text{if } x - \theta \leq k \\ -2(x-\theta) & \text{if } |x - \theta| \leq k \\ -2k & \text{if } x - \theta \geq k \end{cases}.$$

b) We have

$$\frac{d}{d\theta}\int \psi_\theta dP = 2(F(k+\theta) - F(-k+\theta)),$$

where $F(t) = P(X \leq t)$, $t \in \mathbf{R}$ is the distribution function. So

$$V = 2(F(k+\theta_0) - F(-k+\theta_0)).$$

c) Clearly $\psi_\theta : \theta \in \mathbf{R}$ is a VC graph class, with envelope $\Psi \leq 2k$.
So the Huber estimator $\hat{\theta}_n$ has influence function

$$l(x) = \begin{cases} \frac{-k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 \leq -k \\ \frac{x-\theta_0}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } |x - \theta_0| \leq k \\ \frac{k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 \geq k \end{cases}.$$

The asymptotic variance is

$$\sigma^2 = \frac{k^2 F(-k+\theta_0) + \int_{-k+\theta_0}^{k+\theta_0}(x-\theta_0)^2 dF(x) + k^2(1 - F(k+\theta_0))}{(F(k+\theta_0) - F(-k+\theta_0))^2}.$$

7.3. Asymptotics for the median. The median (see Example (i.b) in Chapter 5) can be regarded as the limiting case of a Huber estimator, with $k \downarrow 0$. However, the loss function $\gamma_\theta(x) = |x - \theta|$ is not differentiable, i.e., does not satisfy condition a. For even sample sizes, we do nevertheless have the *score equation* $F_n(\hat{\theta}_n) - \frac{1}{2} = 0$. Let us investigate this closer.

Let $X \in \mathbf{R}$ have distribution F, and let F_n be the empirical distribution. The population median θ_0 is a solution of the equation

$$F(\theta_0) = 0.$$

We assume this solution exists and also that F has positive density f in a neighborhood of θ_0. Consider now for simplicity even sample sizes n and let the sample median $\hat{\theta}_n$ be any solution of

$$F_n(\hat{\theta}_n) = 0.$$

Then we get

$$\begin{aligned} 0 &= F_n(\hat{\theta}_n) - F(\theta_0) \\ &= \Big[F_n(\hat{\theta}_n) - F(\hat{\theta}_n)\Big] + \Big[F(\hat{\theta}_n) - F(\theta_0)\Big] \\ &= \frac{1}{\sqrt{n}} W_n(\hat{\theta}_n) + \Big[F(\hat{\theta}_n) - F(\theta_0)\Big], \end{aligned}$$

where $W_n = \sqrt{n}(F_n - F)$ is the empirical process. Since F is continuous at θ_0, and $\hat{\theta}_n \to \theta_0$, we have by the asymptotic continuity of the empirical process (Section 6.3), that $W_n(\hat{\theta}_n) = W_n(\theta_0) + o_{\mathbf{P}}(1)$. We thus arrive at

$$\begin{aligned} 0 &= W_n(\theta_0) + \sqrt{n}\Big[F(\hat{\theta}_n) - F(\theta_0)\Big] + o_{\mathbf{P}}(1) \\ &= W_n(\theta_0) + \sqrt{n}[f(\theta_0) + o(1)][\hat{\theta}_n - \theta_0]. \end{aligned}$$

In other words,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{W_n(\theta_0)}{f(\theta_0)} + o_{\mathbf{P}}(1).$$

So the influence function is

$$l(x) = \begin{cases} -\frac{1}{2f(\theta_0)} & \text{if } x \leq \theta_0 \\ +\frac{1}{2f(\theta_0)} & \text{if } x > \theta_0 \end{cases},$$

and the asymptotic variance is

$$\sigma^2 = \frac{1}{4f(\theta_0)^2}.$$

We can now compare median and mean. It is easily seen that the asymptotic relative efficiency of the mean as compared to the median is

$$\mathrm{e}_{1,2} = \frac{1}{4\sigma_0^2 f(\theta_0)^2},$$

where $\sigma_0^2 = \mathrm{var}(X)$. So $\mathrm{e}_{1,2} = \pi/2$ for the normal distribution, and $\mathrm{e}_{1,2} = 1/2$ for the double exponential (Laplace) distribution. The density of the double exponential distribution is

$$f(x) = \frac{1}{\sqrt{2\sigma_0^2}} \exp\left[-\frac{\sqrt{2}|x - \theta_0|}{\sigma_0}\right], \ x \in \mathbf{R}.$$

7.4. Conditions A,B and C for asymptotic normaility. We are now going to relax the condition of differentiability of γ_θ.

Condition A. (Differentiability in quadratic mean.) There exists a function $\psi_0 : \mathbf{X} \to \mathbf{R}^r$, with $P(\psi_{0,k}^2) < \infty$, $k = 1, \ldots, r$, such that

$$\lim_{\theta \to \theta_0} \frac{\big\|\gamma_\theta - \gamma_{\theta_0} - (\theta - \theta_0)^T \psi_0\big\|_{2,P}}{|\theta - \theta_0|} = 0.$$

Condition B. We have as $\theta \to \theta_0$,

$$P(\gamma_\theta) - P(\gamma_{\theta_0}) = \frac{1}{2}(\theta - \theta_0)^T V (\theta - \theta_0) + o(1)|\theta - \theta_0|^2,$$

with V a positive definite matrix.

Condition C. Define for $\theta \neq \theta_0$,

$$g_\theta = \frac{\gamma_\theta - \gamma_{\theta_0}}{|\theta - \theta_0|}.$$

Suppose that for some $\epsilon > 0$, the class $\{g_\theta : \ 0 < |\theta - \theta_0| < \epsilon\}$ is a P-Donsker class with envelope G satisfying $P(G^2) < \infty$.

Lemma 7.4.1. *Suppose conditions A,B and C are met. Then $\hat{\theta}_n$ has influence function*

$$l = -V^{-1}\psi_0,$$

and so

$$\sqrt{(}\hat{\theta}_n - \theta_0) \to^{\mathcal{L}} \mathcal{N}(0, V^{-1}JV^{-1}),$$

where $J = \int \psi_0 \psi_0^T dP$.

Proof. Since $\{g_\theta : \ |\theta - \theta_0| \leq \epsilon\}$ is a P-Donsker class, and $\hat{\theta}_n$ is consistent, we may write

$$\begin{aligned} 0 \geq P_n(\gamma_{\hat{\theta}_n} - \gamma_{\theta_0}) &= (P_n - P)(\gamma_{\hat{\theta}_n} - \gamma_{\theta_0}) + P(\gamma_{\hat{\theta}_n} - \gamma_{\theta_0}) \\ &= (P_n - P)(g_\theta)|\theta - \theta_0| + P(\gamma_{\hat{\theta}_n} - \gamma_{\theta_0}) \\ &= (P_n - P)(\hat{\theta}_n - \theta_0)^T\psi_0 + o_{\mathbf{P}}(n^{-1/2}) + P(\gamma_{\hat{\theta}_n} - \gamma_{\theta_0}) \\ = (P_n - P)(\hat{\theta}_n - \theta_0)^T\psi_0 + o_{\mathbf{P}}(n^{-1/2})|\hat{\theta}_n - \theta_0| &+ \frac{1}{2}(\hat{\theta}_n - \theta_0)^T V(\hat{\theta}_n - \theta_0) + o(|\hat{\theta}_n - \theta_0|^2). \end{aligned}$$

This implies $|\hat{\theta}_n - \theta_0| = O_{\mathbf{P}}(n^{-1/2})$. But then

$$|V^{1/2}(\hat{\theta}_n - \theta_0) + V^{-1/2}(P_n - P)(\psi_0) + o_{\mathbf{P}}(n^{-1/2})|^2 \leq o_{\mathbf{P}}(\frac{1}{n}).$$

Therefore,

$$\hat{\theta}_n - \theta_0 = -V^{-1}(P_n - P)(\psi_0) + o_{\mathbf{P}}(n^{-1/2}).$$

Because $P(\psi_0) = 0$, the result follows, and the asymptotic covariance matrix is $V^{-1}JV^{-1}$. □

7.5.Exercises.

Exercise 1. Suppose X has the logistic distribution with location parameter θ (see Example (ii.b) of Chapter 5). Show that the maximum likelihood estimator has asymptotic variance equal to 3, and the median has asymptotic variance equal to 4. Hence, the asymptotic relative efficiency of the maximum likelihood estimator as compared to the median is 4/3.

Exercise 2. Let (X_i, Y_i), $i = 1, \ldots, n, \ldots$ be i.i.d. copies of (X, Y), where $X \in \mathbf{R}^d$ and $Y \in \mathbf{R}$. Suppose that the conditional distribution of Y given $X = x$ has median $m(x) = \alpha_0 + \alpha_1 x_1 + \ldots \alpha_d x_d$, with

$$\alpha = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_d \end{pmatrix} \in \mathbf{R}^{d+1}.$$

Assume moreover that given $X = x$, the random variable $Y - m(x)$ has a density f not depending on x, with f positive in a neighborhood of zero. Suppose moreover that

$$\Sigma = E \begin{pmatrix} 1 & X \\ X & XX^T \end{pmatrix}$$

exists. Let

$$\hat{\alpha}_n = \arg \min_{a \in \mathbf{R}^{d+1}} \frac{1}{n} \sum_{i=1}^n |Y_i - a_0 - a_1 X_{i,1} - \ldots - a_d X_{i,d}|,$$

be the least absolute deviations (LAD) estimator. Show that

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \to^{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4f^2(0)}\Sigma^{-1}\right),$$

by verifying conditions A,B and C.

8. Rates of convergence for least squares estimators

Probability inequalities for the least squares estimator are obtained, under conditions on the entropy of the class of regression functions. In the examples, we study smooth regression functions, functions of bounded variation, concave functions, analytic functions, and image restoration. Results for he entropies of various classes of functions is taken from the literature on approximation theory.

Let $Y_1, \ldots, Y_n$ be real-valued observations, satisfying

$$Y_i = g_0(z_i) + W_i, \ i = 1, \ldots, n,$$

with $z_1, \ldots, z_n$ (fixed) covariates in a space $\mathcal{Z}$, $W_1, \ldots, W_n$ independent errors with expectation zero, and with the unknown regression function g_0 in a given class $\mathcal{G}$ of regression functions. The least squares estimator is

$$\hat{g}_n := \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} (Y_i - g(z_i))^2.$$

Throughout, we assume that a minimizer $\hat{g}_n \in \mathcal{G}$ of the sum of squares exists, but it need not be unique.

The following notation will be used. The empirical measure of the covariates is

$$Q_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}.$$

For g a function on $\mathcal{Z}$, we denote its squared $L_2(Q_n)$-norm by

$$\|g\|_n^2 := \|g\|_{2,Q_n}^2 := \frac{1}{n} \sum_{i=1}^{n} g^2(z_i).$$

The empirical inner product between error and regression function is written as

$$(w, g)_n = \frac{1}{n} \sum_{i=1}^{n} W_i g(z_i).$$

Finally, we let

$$\mathcal{G}(R) := \{g \in \mathcal{G} : \ \|g - g_0\|_n \leq R\}$$

denote a ball around g_0 with radius R, intersected with $\mathcal{G}$.

The main idea to arrive at rates of convergence for $\hat{g}_n$ is to invoke the basic inequality

$$\|\hat{g}_n - g_0\|_n^2 \leq 2(w, \hat{g}_n - g_0)_n.$$

The modulus of continuity of the process $\{(w, g - g_0)_n : \ g \in \mathcal{G}(R)\}$ can be derived from the entropy of $\mathcal{G}(R)$, endowed with the metric

$$d_n(g, \tilde{g}) := \|g - \tilde{g}\|_n.$$

8.1. Gaussian errors.

When the errors are Gaussian, it is not hard to extend the maximal inequality of Lemma 6.5.2.1. We therefore, and to simplify the exposition, will assume in this chapter that

$$W_1, \ldots, W_n \text{ are i.i, d, } \ \mathcal{N}(0, 1)-\text{distributed.}$$

Then, as in Lemma 6.5.2.1, for

$$\sqrt{n}\delta \geq 28 \int_0^R \sqrt{\log N(u, \mathcal{G}(R), d_n)} du \vee 70 R \log 2,$$

we have

$$\mathbf{P}\left(\sup_{g \in \mathcal{G}(R)} (w, g - g_0)_n \geq \delta \right) \leq 4 \exp\left[-\frac{n\delta^2}{(70R)^2} \right]. \qquad (*)$$

8.2. Rates of convergence.

Define

$$J(\delta, \mathcal{G}(\delta), d_n) = \int_0^\delta \sqrt{\log N(u, \mathcal{G}(\delta), d_n)} du \vee \delta.$$

Theorem 8.2.1. *Take $\Psi(\delta) \geq J(\delta, \mathcal{G}(\delta), d_n)$ in such a way that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ. Then for a constant c, and for*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n)$$

we have for all $\delta \geq \delta_n$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > \delta) \leq c \exp[-\frac{n\delta^2}{c^2}].$$

Proof. We have

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > \delta) \leq$$

$$\sum_{s=0}^{\infty} \mathbf{P}\left(\sup_{g \in \mathcal{G}(2^{s+1}\delta)} (w, g - g_0)_n \geq 2^{2s-1}\delta^2 \right) := \sum_{s=0}^{\infty} \mathbf{P}_s.$$

Now, if

$$\sqrt{n}\delta_n^2 \geq c_1 \Psi(\delta_n),$$

then also for all $2^{s+1}\delta > \delta_n$,

$$\sqrt{n} 2^{2s+2}\delta^2 \geq c_1 \Psi(2^{s+1}\delta).$$

So, for an appropriate choice of c_1, we may apply (*) to each $\mathbf{P}_s$. This gives, for some c_2, c_3,

$$\sum_{s=0}^{\infty} \mathbf{P}_s \leq \sum_{s=0}^{\infty} c_2 \exp[-\frac{n 2^{4s-2}\delta^4}{c_2 2^{2s+2}\delta^2}] \leq c_3 \exp[-\frac{n\delta^2}{c_3^2}].$$

Take $c = \max\{c_1, c_2, c_3\}$. □

8.3. Examples.

Example 8.3.1. Linear regression. Let

$$\mathcal{G} = \{g(z) = \theta_1 \psi_1(z) + \ldots + \theta_r \psi_r(z) : \ \theta \in \mathbf{R}^r\}.$$

One may verify

$$\log N(u, \mathcal{G}(\delta), d_n) \leq r \log(\frac{\delta + 4u}{u}), \text{ for all } 0 < u < \delta, \ \delta > 0.$$

So

$$\int_0^\delta \sqrt{\log N(u, \mathcal{G}(\delta), d_n)} du \leq r^{1/2} \int_0^\delta \log^{1/2}(\frac{\delta + 4u}{\delta}) du$$

$$= r^{1/2}\delta \int_0^1 \log^{1/2}(1 + 4v) dv := A_0 r^{1/2}\delta.$$

So Theorem 8.2.1 can be applied with

$$\delta_n \geq c A_0 \sqrt{\frac{r}{n}}.$$

It yields that for some constant c (not the same at each appearance) and for all $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > T\sqrt{\frac{r}{n}}) \leq c \exp[-\frac{T^2 r}{c^2}].$$

(Note that we made extensive use here from the fact that it suffices to calculate the *local* entropy of $\mathcal{G}$.)

Example 8.3.2. Smooth functions. Let

$$\mathcal{G} = \{g : [0,1] \to \mathbf{R}, \int (g^{(m)}(z))^2 dz \leq M^2\}.$$

Let $\psi_k(z) = z^{k-1}$, $k = 1, \ldots, m$, $\psi(z) = (\psi_1(z), \ldots, \psi_m(z))^T$ and $\Sigma_n = \int \psi\psi^T dQ_n$. Denote the smallest eigenvalue of Σ_n by λ_n, and assume that

$$\lambda_n \geq \lambda > 0, \text{ for all } n \geq n_0.$$

One can show (Kolmogorov and Tihomirov (1959)) that

$$\log N(\delta, \mathcal{G}(\delta), d_n) \leq A\delta^{-\frac{1}{m}}, \text{ for small } \delta > 0,$$

where the constant A depends on λ. Hence, we find from Theorem 8.2.1 that for $T \geq c$, ,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-\frac{m}{2m+1}}) \leq c \exp[-\frac{T^2 n^{\frac{1}{2m+1}}}{c^2}].$$

Example 8.3.3. Functions of bounded variation in R. Let

$$\mathcal{G} = \{g : \mathbf{R} \to \mathbf{R}, \int |g^{'}(z)| dz \leq M\}.$$

Without loss of generality, we may assume that $z_1 \leq \ldots \leq z_n$. The derivative should be understood in the generalized sense:

$$\int |g^{'}(z)| dz := \sum_{i=2}^{n} |g(z_i) - g(z_{i-1})|.$$

Define for $g \in \mathcal{G}$,

$$\alpha := \int g dQ_n.$$

Then it is easy to see that,

$$\max_{i=1,\ldots,n} |g(z_i)| \leq \alpha + M.$$

One can now show (Birman and Solomjak (1967)) that

$$\log N(\delta, \mathcal{G}(\delta), d_n) \leq A\delta^{-1}, \text{ for small } \delta > 0,$$

and therefore, for all $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-1/3}) \leq c \exp[-\frac{T^2 n^{1/3}}{c^2}].$$

Example 8.3.4. Functions of bounded variation in $\mathbf{R}^2$. Suppose that $z_i = (u_k, v_l)$, $i = kl$, $k = 1, \ldots, n_1$, $l = 1, \ldots, n_2$, $n = n_1 n_2$, with $u_1 \leq \ldots \leq u_{n_1}$, $v_1 \leq \ldots \leq v_{n_2}$. Consider the class

$$\mathcal{G} = \{g : \mathbf{R}^2 \to \mathbf{R}, \ I(g) \leq M\}$$

where

$$I(g) := I_0(g) + I_1(g_{1\cdot}) + I_2(g_{\cdot 2}),$$

$$I_0(g) := \sum_{k=2}^{n_2} \sum_{l=2}^{n_2} |g(u_k, v_l) - g(u_{k-1}, v_l) - g(u_k, v_{l-1}) + g(u_{k-1}, v_{l-1})|,$$

$$g_{1\cdot}(u) := \frac{1}{n_2} \sum_{l=1}^{n_2} g(u, v_l),$$

$$g_{\cdot 2}(v) := \frac{1}{n_1}\sum_{k=1}^{n_1} g(u_k, v),$$

$$I_1(g_{1\cdot}) := \sum_{k=2}^{n_1} |g_{1\cdot}(u_k) - g_{1\cdot}(u_{k-1})|,$$

and

$$I_2(g_{\cdot 2}) := \sum_{l=2}^{n_2} |g_{\cdot 2}(v_l) - g_{\cdot 2}(v_{l-1})|.$$

Thus, each $g \in \mathcal{G}$ as well as its marginals have total variation bounded by M. We apply the result of Ball and Pajor (1990) on convex hulls. Let

$$\Lambda := \{\text{all distribution functions } F \text{ on } \mathbf{R}^2\},$$

and

$$\mathcal{K} := \{1_{(-\infty, z]} : \ z \in \mathbf{R}^2\}.$$

Clearly, $\Lambda = \overline{\text{conv}}(\mathcal{K})$, and

$$N(\delta, \mathcal{K}, d_n) \leq c\frac{1}{\delta^4}, \text{ for all } \delta > 0.$$

Then from Pall and Pajor (1990),

$$\log N(\delta, \Lambda, d_n) \leq A\delta^{-\frac{4}{3}}, \text{ for all } \delta > 0.$$

The same bound holds therefore for any uniformly bounded subset of $\mathcal{G}$. Now, any function $g \in \mathcal{G}$ can be expressed as

$$g(u, v) = \tilde{g}(u, v) + \tilde{g}_{1\cdot}(u) + \tilde{g}_{\cdot 2}(v) + \alpha,$$

where

$$\alpha := \frac{1}{n_1 n_2}\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} g(u_k, v_k),$$

and where

$$\sum_{k=1}^{n_1}\sum_{l=1}^{n_2} \tilde{g}(u_k, v_l) = 0,$$

$$\sum_{k=1}^{n_1} \tilde{g}_{1\cdot}(u_k) = 0,$$

as well as

$$\sum_{l=1}^{n_2} \tilde{g}_{\cdot 2}(v_l) = 0.$$

It is easy to see that

$$|\tilde{g}(u_k, v_l)| \leq I_0(\tilde{g}) = I_0(g), \ k = 1, \ldots, n_1, \ l = 1, \ldots, n_2,$$

$$|\tilde{g}_{1\cdot}(u_k)| \leq I_1(\tilde{g}_{1\cdot}) = I_1(g_{1\cdot}), \ k = 1, \ldots, n_1,$$

and

$$|g_{\cdot 2}(v_l)| \leq I_2(\tilde{g}_{\cdot 2}) = I_2(g_{\cdot 2}), \ l = 1, \ldots, n_2.$$

Whence

$$\{\tilde{g} + \tilde{g}_{1\cdot} + \tilde{g}_{\cdot 2} : \ g \in \mathcal{G}\}$$

is a uniformly bounded class, for which the entropy bound $A_1\delta^{-4/3}$ holds, with A_1 depending on M. It follows that

$$\log N(u, \mathcal{G}(\delta), d_n) \leq A_1 u^{-\frac{4}{3}} + A_2 \log(\frac{\delta}{u}), \ 0 < u \leq \delta.$$

From Theorem 8.2.1, we find for all $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-\frac{3}{10}}) \leq c \exp[-\frac{T^2 n^{\frac{2}{5}}}{c^2}].$$

The result can be extended to functions of bounded variation in $\mathbf{R}^r$. Then one finds the rate $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-\frac{1+r}{2+4r}})$.

Example 8.3.5. Concave functions. Let

$$\mathcal{G} = \{g : [0,1] \to \mathbf{R},\ 0 \leq g^{'} \leq M,\ g^{'} \text{ decreasing}\}.$$

Then $\mathcal{G}$ is a subset of

$$\{g : [0,1] \to \mathbf{R},\ \int_0^1 |g^{''}(z)|dz \leq 2M\}.$$

Birman and Solomjak (1967) prove that for all $m \in \{2, 3, \ldots\}$,

$$\log N(\delta, \{g : [0,1] \to [0,1] :\ \int_0^1 |g^{(m)}(z)|dz \leq 1\}, d_\infty) \leq A\delta^{-\frac{1}{m}},\ \text{for all } \delta > 0.$$

Again, our class $\mathcal{G}$ is not uniformly bounded, but we can write for $g \in \mathcal{G}$,

$$g = g_1 + g_2,$$

with $g_1(z) := \theta_1 + \theta_2 z$ and $|g_2|_\infty \leq 2M$. Assume now that $\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})^2$ stays away from 0. Then, we obtain for $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-\frac{2}{5}}) \leq c \exp[-\frac{T^2 n^{\frac{1}{5}}}{c^2}].$$

Example 8.3.6. Analytic functions. Let

$$\mathcal{G} = \{g : [0,1] \to \mathbf{R} :\ g^{(k)} \text{ exists for all } k \geq 0,\ |g^{(k)}|_\infty \leq M \text{ for all } k \geq m\}.$$

.

Lemma 8.3.1 *We have*

$$\log N(u, \mathcal{G}(\delta), d_n) \leq (\frac{\log(\frac{3M}{u})}{\log 2} + 1) \vee m) \log(\frac{3\delta + 6u}{u}),\ 0 < u < \delta.$$

Proof. Take

$$d = (\lfloor \frac{\log(\frac{M}{u})}{\log 2} \rfloor + 1) \vee m,$$

where $\lfloor x \rfloor$ is the integer part of $x \geq 0$. For each $g \in \mathcal{G}$, we can find a polynomial f of degree $d-1$ such that

$$|g(z) - f(z)| \leq M|z - \frac{1}{2}|^d \leq M(\frac{1}{2})^d \leq u.$$

Now, let $\mathcal{F}$ be the collection of all polynomials of degree $d-1$, and let $f_0 \in \mathcal{F}$ be the approximating polynomial of g_0, with $|g_0 - f_0|_\infty \leq u$.

If $\|g - g_0\|_n \leq \delta$, we find $\|f - f_0\|_n \leq \delta + 2u$. We know that

$$\log N(u, \mathcal{F}(\delta + 2u), d_n) \leq d \log(\frac{\delta + 6u}{u}),\ u > 0,\ \delta > 0.$$

If $\|f - \tilde{f}\|_n \leq u$ and $|g - f|_\infty \leq u$ as well as $|\tilde{g} - \tilde{f}|_\infty \leq u$, we obtain $\|g - \tilde{g}\|_n \leq 3u$. So,

$$H(3u, \mathcal{G}_n(\delta), d_n) \leq (\frac{\log(\frac{M}{u})}{\log 2} + 1) \vee m) \log(\frac{\delta + 6u}{u}),\ u > 0,\ \delta > 0.$$

□

From Theorem 8.2.1, it follows that for $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-1/2}\log^{1/2} n) \leq c\exp[-\frac{T^2\log n}{c^2}].$$

Example 8.3.7. Image restoration.

Case (i). Let $\mathcal{Z} \subset \mathbf{R}^2$ be some subset of the plane. Each site $z \in \mathcal{Z}$ has a certain gray-level $g_0(z)$, which is expressed as a number between 0 and 1, i.e., $g_0(z) \in [0,1]$. We have noisy data on a set of $n = n_1 n_2$ pixels $\{z_{kl} : \ k = 1,\ldots,n_1, \ l = 1,\ldots,n_2\} \subset \mathcal{Z}$:

$$Y_{kl} = g_0(z_{kl}) + W_{kl},$$

where the measurement errors $\{W_{kl} : \ k = 1,\ldots,n_1, \ l = 1,\ldots,n_2\}$ are independent $\mathcal{N}(0,1)$ random variables. Now, each patch of a certain gray-level is a mixture of certain amounts of black and white. Let

$$\mathcal{G} = \overline{\text{conv}}(\mathcal{K}),$$

where

$$\mathcal{K} := \{1_{\mathcal{D}} : \ D \in \mathcal{D}\}.$$

Assume that

$$N(\delta, \mathcal{K}, d_n) \leq c\delta^{-w}, \text{ for all } \delta > 0.$$

Then from Ball and Pajor(1990),

$$\log N(\delta, \mathcal{G}, d_n) \leq A\delta^{-\frac{2w}{2+w}}, \text{ for all } \delta > 0.$$

It follows from Theorem 8.2.1 that for $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-\frac{2+w}{4+4w}}) \leq c\exp[-\frac{T^2 n^{\frac{w}{2+2w}}}{c^2}].$$

Case (ii). Consider a black-and-white image observed with noise. Let $\mathcal{Z} = [0,1]^2$ be the unit square, and

$$g_0(z) = \begin{cases} 1, & \text{if } z \text{ is black,} \\ 0, & \text{if } z \text{ is white .} \end{cases}$$

The black part of the image is

$$D_0 := \{z \in [0,1]^2 : \ g_0(z) = 1\}.$$

We observe

$$Y_{kl} = g(z_{kl}) + W_{kl},$$

with $z_{kl} = (u_k, v_l)$, $u_k = k/m$, $v_l = l/m$, $k,l \in \{1,\ldots,m\}$. The total number of pixels is thus $n = m^2$.

Suppose that

$$D_0 \in \mathcal{D} = \{\text{all convex subsets of } [0,1]^2\},$$

and write

$$\mathcal{G} := \{1_D : \ D \in \mathcal{D}\}.$$

Dudley (1984) shows that for all $\delta > 0$ sufficiently small

$$\log N(\delta, \mathcal{G}, d_n) \leq A\delta^{-\frac{1}{2}},$$

so that for $T \geq c$,

$$\mathbf{P}(\|\hat{g}_n - g_0\|_n > Tn^{-\frac{2}{5}}) \leq c\exp[-\frac{T^2 n^{\frac{1}{5}}}{c^2}].$$

Let $\hat{D}_n$ be the estimate of the black area, so that $\hat{g}_n = 1_{\hat{D}_n}$. For two sets D_1 and D_2, denote the symmetric difference by

$$D_1 \Delta D_2 := (D_1 \cap D_2^c) \cup (D_1^c \cap D_2).$$

Since $Q_n(D) = \|l_D\|_n^2$, we find

$$Q_n(\hat{D}_n \Delta D_0) = O_{\mathbf{P}}(n^{-\frac{4}{5}}).$$

Remark. In higher dimensions, say $\mathcal{Z} = [0,1]^r$, $r \geq 2$, the class $\mathcal{G}$ of indicators of convex sets has entropy

$$\log N(\delta, \mathcal{G}, d_n) \leq A\delta^{-\frac{r-1}{2}}, \ \delta \downarrow 0,$$

provided that the pixels are on a regular grid (see Dudley (1984)). So the rate is then

$$Q_n(\hat{D}_n \Delta D_0) = \begin{cases} O_{\mathbf{P}}(n^{-\frac{4}{r+3}}) & , \text{if } r \in \{2,3,4\}, \\ O_{\mathbf{P}}(n^{-\frac{1}{2}} \log n), & \text{if } r = 5, \\ O_{\mathbf{P}}(n^{-\frac{2}{r-1}}), & \text{if } r \geq 6 \ . \end{cases}$$

For $r \geq 5$, the least squares estimator converges with suboptimal rate.

8.4. Exercises.

8.1. Let $Y_1, \ldots, Y_n$ be independent, uniformly sub-Gaussian random variables, with $EY_i = \alpha_0$ for $i = 1, \ldots, \lfloor n\gamma_0 \rfloor$, and $EY_i = \beta_0$ for $i = \lfloor n\gamma_0 \rfloor + 1, \ldots, n$, where α_0, β_0 and the change point γ_0 are completely unknown. Write $g_0(i) = g(i; \alpha_0, \beta_0, \gamma_0) = \alpha_0 1\{1 \leq i \leq \lfloor n\gamma_0 \rfloor\} + \beta_0 1\{\lfloor n\gamma_0 \rfloor + 1 \leq i \leq n\}$. We call the parameter $(\alpha_0, \beta_0, \gamma_0)$ identifiable if $\alpha_0 \neq \beta_0$ and $\gamma_0 \in (0,1)$. Let $\hat{g}_n = g(\cdot; \hat{\alpha}_n, \hat{\beta}_n, \hat{\gamma}_n)$ be the least squares estimator. Show that if $\alpha_0, \beta_0, \gamma_0$ is identifiable, then $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-1/2})$, and $|\hat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}(n^{-1/2})$, $|\hat{\beta}_n - \beta_0| = O_{\mathbf{P}}(n^{-1/2})$, and $|\hat{\gamma}_n - \gamma_0| = O_{\mathbf{P}}(n^{-1})$. If $(\alpha_0, \beta_0, \gamma_0)$ is not identifiable, show that $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-1/2}(\log \log n)^{1/2})$.

8.2. Let $z_i = i/n$, $i = 1, \ldots, n$, and let $\mathcal{G}$ consist of the functions

$$g(z) = \begin{cases} \alpha_1 + \alpha_2 z, & \text{if } z \leq \gamma \\ \beta_1 + \beta_2 z, & \text{if } z > \gamma \end{cases} .$$

Suppose g_0 is continuous, but does have a kink at γ_0: $\alpha_{1,0} = \alpha_{2,0} = 0$, $\beta_{1,0} = -\frac{1}{2}$, $\beta_{2,0} = 1$, and $\gamma_0 = \frac{1}{2}$. Show that $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-1/2})$, and that $|\hat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}(n^{-1/2})$, $|\hat{\beta}_n - \beta_0| = O_{\mathbf{P}}(n^{-1/2})$ and $|\hat{\gamma}_n - \gamma_0| = O_{\mathbf{P}}(n^{-1/3})$.

8.3. If $\mathcal{G}$ is a uniformly bounded class of increasing functions, show that it follows from Theorem 8.2.1 that $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-1/3}(\log n)^{1/3})$. (Actually, by a more tight bound on the entropy one has the rate $O_{\mathbf{P}}(n^{-1/3})$, see Example 8.3.3.).

9. Penalized least squares

We revisit the regression problem of the previous chapter (but use a slightly different notation). One has observations $\{(x_i, Y_i)\}_{i=1}^n$, with $x_1, \ldots, x_n$ fixed co-variables, and $Y_1, \ldots, Y_n$ response variables, satisfying the regression

$$Y_i = f_0(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and centered noise variables, and f_0 is an unknown function on $\mathcal{X}$. The errors are assumed to be $\mathcal{N}(0, \sigma^2)$-distributed.

Let $\bar{\mathcal{F}}$ be a collection of regression functions. The penalized least squares estimator is

$$\hat{f}_n = \arg\min_{f \in \bar{\mathcal{F}}} \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - f(x_i)|^2 + \text{pen}(f) \right\}.$$

Here $\text{pen}(f)$ is a penalty on the complexity of the function f. Let Q_n be the empirical distribution of $x_1, \ldots, x_n$ and $\|\cdot\|_n$ be the $L_2(Q_n)$-norm. Define

$$f_* = \arg\min_{f \in \bar{\mathcal{F}}} \left\{ \|f - f_0\|_n^2 + \text{pen}(f) \right\}.$$

Our aim is to show that

$$(*) \qquad \mathbf{E}\|\hat{f}_n - f_0\|_n^2 \leq \text{const.} \left\{ \|f_* - f_0\|_n^2 + \text{pen}(f_*) \right\}.$$

When this aim is indeed reached, we loosely say that $\hat{f}_n$ satisfies an oracle inequality. In fact, what (*) says it that $\hat{f}_n$ behaves as the noiseless version f_*. That means so to speak that we "overruled" the variance of the noise.

In Section 9.1, we recall the definitions of estimation and approximation error. Section 9.2 calculates the estimation error when one employs least squares estimation, without penalty, over a finite model class. The estimation error turns out to behave as the log-cardinality of the model class. Section 9.3 shows that when considering a collection of nested finite models, a penalty $\text{pen}(f)$ proportional to the log-cardinality of the smallest class containing f will indeed mimic the oracle over this collection of models. In Section 9.4, we consider general penalties. It turns out that the (local) *entropy* of the model classes plays a crucial rule. The local entropy a finite-dimensional space is proportional to its dimension. For a finite class, the entropy is (bounded by) its log-cardinality.

Whether or not (*) holds true depends on the choice of the penalty. In Section 9.4, we show that when the penalty is taken "too small" there will appear an additional term showing that not all variance was "killed". Section 9.5 presents an example.

Throughout this chapter, we assume the noise level $\sigma > 0$ to be known. In that case, by a rescaling argument, one can assume without loss of generality that $\sigma = 1$. In general, one needs a good estimate of an upper bound for σ, because the penalties considered in this chapter depend on the noise level. When one replaces the unknown noise level σ by an estimated upper bound, the penalty in fact becomes data dependent.

9.1. Estimation and approximation error. Let $\mathcal{F}$ be a model class. Consider the least squares estimator without penalty

$$\hat{f}_n(\cdot, \mathcal{F}) = \arg\min_{f \in \mathcal{F}} \ \frac{1}{n} \sum_{i=1}^n |Y_i - f(x_i)|^2.$$

The excess risk $\|\hat{f}_n(\cdot, \mathcal{F}) - f_0\|_n^2$ of this estimator is the sum of estimation error and approximation error.

Now, if we have a collection of models $\{\mathcal{F}\}$, a penalty is usually some measure of the complexity of the model class $\mathcal{F}$. With some abuse of notation, write this penalty as $\text{pen}(\mathcal{F})$. The corresponding penalty on the functions f is then

$$\text{pen}(f) = \min_{\mathcal{F}: \ f \in \mathcal{F}} \text{pen}(\mathcal{F}).$$

We may then write

$$\hat{f}_n = \arg\min_{\mathcal{F} \in \{\mathcal{F}\}} \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{f}_n(x_i, \mathcal{F})|^2 + \text{pen}(\mathcal{F}) \right\},$$

where $\hat{f}_n(\cdot, \mathcal{F})$ is the least squares estimator over $\mathcal{F}$. Similarly,

$$f_* = \arg \min_{\mathcal{F} \in \{\mathcal{F}\}} \{\|f_*(\cdot, \mathcal{F}) - f_0\|_n^2 + \text{pen}(\mathcal{F})\},$$

where $f_*(\cdot, \mathcal{F})$ is the best approximation of f_0 in the model $\mathcal{F}$.

As we will see, taking $\text{pen}(\mathcal{F})$ proportional to (an estimate) of the estimation error of $\hat{f}_n(\cdot, \mathcal{F})$ will (up to constants and possibly $(\log n)$-factors) balance estimation error and approximation error.

In this chapter, the empirical process takes the form

$$\nu_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(x_i),$$

with the function f ranging over (some subclass of) $\bar{\mathcal{F}}$. Probability inequalities for the empirical process are derived using Exercise 2.4.1. The latter is for normally distributed random variables. It is exactly at this place where our assumption of normally distributed noise comes in. Relaxing the normality assumption is straightforward, provided a proper probability inequality, an inequality of *sub-Gaussian* type, goes through. In fact, at the cost of additional, essentially technical, assumptions, an inequality of *exponential type* on the errors is sufficient as well (see van de Geer (2000)).

9.2. Finite models. Let $\mathcal{F}$ be a finite collection of functions, with cardinality $|\mathcal{F}| \geq 2$. Consider the least squares estimator over $\mathcal{F}$

$$\hat{f}_n = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |Y_i - f(x_i)|^2.$$

In this section, $\mathcal{F}$ is fixed, and we do not explicitly express the dependency of $\hat{f}_n$ on $\mathcal{F}$. Define

$$\|f_* - f_0\|_n = \min_{f \in \mathcal{F}} \|f - f_0\|_n.$$

The dependence of f_* on $\mathcal{F}$ is also not expressed in the notation of this section. Alternatively stated, we take here

$$\text{pen}(f) = \begin{cases} 0 & \forall f \in \mathcal{F} \\ \infty & \forall f \in \bar{\mathcal{F}} \backslash \mathcal{F} \end{cases} .$$

The result of Lemma 9.2.1 below implies that the estimation error is proportional to $\log |\mathcal{F}|/n$, i.e., it is logarithmic in the number of elements in the parameter space. We present the result in terms of a probability inequality. An inequality for e.g., the average excess risk follows from this (see Exercise 9.1).

Lemma 9.2.1. *We have for all $t > 0$ and $0 < \delta < 1$,*

$$\mathbf{P}\left(\|\hat{f}_n - f_0\|_n^2 \geq (\frac{1+\delta}{1-\delta}) \left\{ \|f_* - f_0\|_n^2 + \frac{4 \log |\mathcal{F}|}{n\delta} + \frac{4t^2}{\delta} \right\} \right) \leq \exp[-nt^2].$$

Proof. We have the basic inequality

$$\|\hat{f}_n - f_0\|_n^2 \leq \frac{2}{n} \sum_{i=1}^n \epsilon_i (\hat{f}_n(x_i) - f_*(x_i)) + \|f_* - f_0\|_n^2.$$

By Exercise 2.1.1, for all $t > 0$,

$$\mathbf{P}\left(\max_{f \in \mathcal{F},\ \|f - f_*\|_n > 0} \frac{\frac{1}{n} \sum_{i=1}^n \epsilon_i (f(x_i) - f_*(x_i))}{\|f - f_*\|_n} > \sqrt{2 \log |\mathcal{F}|/n + 2t^2} \right)$$

$$\leq |\mathcal{F}| \exp[-(\log |\mathcal{F}| + nt^2)] = \exp[-nt^2].$$

If $\frac{1}{n} \sum_{i=1}^n \epsilon_i (\hat{f}_n(x_i) - f_*(x_i)) \leq (2 \log |\mathcal{F}|/n + 2t^2)^{1/2} \|\hat{f}_n - f_*\|_n$, we have, using $2\sqrt{ab} \leq a + b$ for all non-negative a and b,

$$\|\hat{f}_n - f_0\|_n^2 \leq 2(2 \log |\mathcal{F}|/n + 2t^2)^{1/2} \|\hat{f}_n - f_*\|_n + \|f_* - f_0\|_n^2$$

$$\leq \delta \|\hat{f}_n - f_0\|_n^2 + 4\log|\mathcal{F}|/(n\delta) + 4t^2/\delta + (1+\delta)\|f_* - f_0\|_n^2.$$

□

9.3. Nested, finite models. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ be a collection of nested, finite models, and let $\bar{\mathcal{F}} = \cup_{m=1}^{\infty} \mathcal{F}_m$. We assume $\log|\mathcal{F}_1| > 1$.

As indicated in Section 9.1, it is a good strategy to take the penalty proportional to the estimation error. In the present context, this works as follows. Define

$$\mathcal{F}(f) = \mathcal{F}_{m(f)},\ m(f) = \arg\min\{m:\ f \in \mathcal{F}_m\},$$

and for some $0 < \delta < 1$,

$$\text{pen}(f) = \frac{16\log|\mathcal{F}(f)|}{n\delta}.$$

In coding theory, this penalty is quite familiar: when encoding a message using an encoder from $\mathcal{F}_m$, one needs to send, in addition to the encoded message, $\log_2|\mathcal{F}_m|$ bits to tell the receiver which encoder was used.

Let

$$\hat{f}_n = \arg\min_{f\in\bar{\mathcal{F}}} \left\{ \frac{1}{n}\sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \text{pen}(f) \right\},$$

and

$$f_* = \arg\min_{f\in\bar{\mathcal{F}}} \left\{ \|f - f_0\|_n^2 + \text{pen}(f) \right\}.$$

Lemma 9.3.2. *We have, for all $t > 0$ and $0 < \delta < 1$,*

$$\mathbf{P}\left(\|\hat{f}_n - f_0\|_n^2 > (\frac{1+\delta}{1-\delta}) \left\{ \|f_* - f_0\|_n^2 + \text{pen}(f_*) + 4t^2/\delta \right\} \right) \leq \exp[-nt^2].$$

Proof. Write down the basic inequality

$$\|\hat{f}_n - f_0\|_n^2 + \text{pen}(\hat{f}_n) \leq \frac{2}{n}\sum_{i=1}^{n} \epsilon_i(\hat{f}_n(x_i) - f_*(x_i)) + \|f_* - f_0\|_n^2 + \text{pen}(f_*).$$

Define $\bar{\mathcal{F}}_j = \{f:\ 2^j < |\log\mathcal{F}(f)| \leq 2^{j+1}\}$, $j = 0, 1, \ldots$. We have for all $t > 0$, using Lemma 3.8,

$$\mathbf{P}\left(\exists\ f \in \bar{\mathcal{F}}:\ \frac{1}{n}\sum_{i=1}^{n} \epsilon_i(f(x_i) - f_*(x_i)) > (8\log|\mathcal{F}(f)|/n + 2t^2)^{1/2}\|f - f_*\|_n \right)$$

$$\leq \sum_{j=0}^{\infty} \mathbf{P}\left(\exists\ f \in \bar{\mathcal{F}}_j,\ \frac{1}{n}\sum_{i=1}^{n} \epsilon_i(f(x_i) - f_*(x_i)) > (2^{j+3}/n + 2t^2)^{1/2}\|f - f_*\|_n \right)$$

$$\leq \sum_{j=0}^{\infty} \exp[2^{j+1} - (2^{j+2} + nt^2)] = \sum_{j=0}^{\infty} \exp[-(2^{j+1} + nt^2)]$$

$$\leq \sum_{j=0}^{\infty} \exp[-(j + 1 + nt^2)] \leq \int_0^{\infty} \exp[-(x + nt^2)] = \exp[-nt^2].$$

But if $\sum_{i=1}^{n} \epsilon_i(\hat{f}_n(x_i) - f_*(x_i))/n \leq (8\log|\mathcal{F}(\hat{f}_n)|/n + 2t^2)^{1/2}\|\hat{f}_n - f_*\|_n$, the basic inequality gives

$$\|\hat{f}_n - f_0\|_n^2 \leq 2(8\log|\mathcal{F}(\hat{f}_n)|/n + 2t^2)^{1/2}\|\hat{f}_n - f_*\|_n + \|f_* - f_0\|_n^2 + \text{pen}(f_*) - \text{pen}(\hat{f}_n)$$

$$\leq \delta\|\hat{f}_n - f_0\|_n^2 + 16\log|\mathcal{F}(\hat{f})|/(n\delta) - \text{pen}(\hat{f}) + 4t^2/\delta + (1+\delta)\|f_* - f_0\|_n^2 + \text{pen}(f_*)$$

$$= \delta\|\hat{f}_n - f_0\|_n^2 + 4t^2/\delta + (1+\delta)\|f_* - f_0\|_n^2 + \text{pen}(f_*),$$

by the definition of $\text{pen}(f)$.

□

9.4. General penalties. In the general case with possibly infinite model classes $\mathcal{F}$, we may replace the log-cardinality of a class by its entropy.

Definition. *Let $u > 0$ be arbitrary and let $N(u, \mathcal{F}, d_n)$ be the minimum number of balls with radius u necessary to cover $\mathcal{F}$. Then $\{H(u, \mathcal{F}, d_n) := \log N(u, \mathcal{F}, d_n) : u > 0\}$ is called the* entropy *of $\mathcal{F}$ (for the metric d_n induced by the norm $\|\cdot\|_n$).*

Recall the definition of the estimator

$$\hat{f}_n = \arg\min_{f \in \bar{\mathcal{F}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(x_i)|^2 + \text{pen}(f) \right\},$$

and of the noiseless version

$$f_* = \arg\min_{f \in \bar{\mathcal{F}}} \left\{ \|f - f_0\|_n^2 + \text{pen}(f) \right\}.$$

We moreover define

$$\mathcal{F}(t) = \{f \in \bar{\mathcal{F}} : \ \|f - f_*\|_n^2 + \text{pen}(f) \leq t^2\}, \ t > 0.$$

Consider the entropy $H(\cdot, \mathcal{F}(t), d_n)$ of $\mathcal{F}(t)$. Suppose it is finite for each t, and in fact that the square root of the entropy is integrable, i.e. that for some continuous upper bound $\bar{H}(\cdot, \mathcal{F}(t), d_n)$ of $H(\cdot, \mathcal{F}(t), d_n)$. one has

$$\Psi(t) = \int_0^t \sqrt{\bar{H}(u, \mathcal{F}(t), d_n)} du < \infty, \ \forall t > 0. \qquad (**)$$

This means that near $u = 0$, the entropy $H(u, \mathcal{F}(t), d_n)$ is not allowed to grow faster than $1/u^2$. Assumption (**) is related to asymptotic continuity of the empirical process $\{\nu_n(f) : \ f \in \mathcal{F}(t)\}$. If (**) does not hold, one can still prove inequalities for the excess risk. To avoid digressions we will skip that issue here.

Lemma 9.4.1. *Suppose that $\Psi(t)/t^2$ does not increase as t increases. There exists constants c and c' such that for*

$$(\bullet) \qquad \sqrt{n} t_n^2 \geq c \left(\Psi(t_n) \vee t_n \right),$$

we have

$$\mathbf{E} \left\{ \|\hat{f}_n - f_0\|_n^2 + \text{pen}(\hat{f}_n) \right\} \leq 2 \left\{ \|f_* - f_0\|_n^2 + \text{pen}(f_*) + t_n^2 \right\} + \frac{c'}{n}.$$

Lemma 9.4.1 is from van de Geer (2001). Comparing it to e.g. Lemma 9.3.1, one sees that there is no arbitrary $0 < \delta < 1$ involved in the statement of Lemma 9.4.1. In fact, van de Geer (2001) has fixed δ at $\delta = 1/3$ for simplicity.

When $\Psi(t)/t^2 \leq \sqrt{n}/C$ for all t, and some constant C, condition $(\bullet)$ is fulfilled if $t_n \geq c n^{-1/2}$, and, in addition, $C \geq c$. In that case one indeed has overruled the variance. We stress here, that the constant C depends on the penalty, i.e. the penalty has to be chosen carefully.

9.5. Application to the "classical" penalty. Suppose $\mathcal{X} = [0, 1]$. Let $\bar{\mathcal{F}}$ be the class of functions on $[0, 1]$ which have derivatives of all orders. The s-th derivative of a function $f \in \bar{\mathcal{F}}$ on $[0, 1]$ is denoted by $f^{(s)}$. Define for a given $1 \leq p < \infty$, and given smoothness $s \in \{1, 2, \ldots\}$,

$$I^p(f) = \int_0^1 |f^{(s)}(x)|^p dx, \ f \in \bar{\mathcal{F}}.$$

We consider two cases. In Subsection 9.5.1, we fix a smoothing parameter $\lambda > 0$ and take the penalty $\text{pen}(f) = \lambda^2 I^p(f)$. After some calculations, we then show that in general the variance has not been "overruled", i.e., we do not arrive at an estimator that behaves as a noiseless version, because there still is an additional term. However, this additional term can now be "killed" by including it in the penalty. It all boils down in Subsection 9.5.2 to a data dependent choice for λ, or alternatively viewed, a penalty of the form $\text{pen}(f) = \tilde{\lambda}^2 I^{\frac{2}{2s+1}}(f)$, with $\tilde{\lambda} > 0$ depending on s and n. This penalty allows one to adapt to small values for $I(f_0)$.

9.5.1. Fixed smoothing parameter. For a function $f \in \bar{\mathcal{F}}$, we define the penalty

$$\text{pen}(f) = \lambda^2 I^p(f),$$

with a given $\lambda > 0$.

Lemma 9.5.1. *The entropy integral Ψ can be bounded by*

$$\Psi(t) \leq c_0 \left(t^{\frac{2ps+2-p}{2ps}} \lambda^{-\frac{1}{ps}} + t\sqrt{\log(\frac{1}{\lambda} \vee 1)} \right) \ t > 0.$$

Here, c_0 is a constant depending on s and p.

Proof. This follows from the fact that

$$H(u, \{f \in \bar{\mathcal{F}} : \ I(f) \leq 1, \ |f| \leq 1\}, d_\infty) \leq Au^{-1/s}, \ u > 0$$

where the constant A depends on s and p (see Birman and Solomjak (1967)). For $f \in \mathcal{F}(t)$, we have

$$I(f) \leq \left(\frac{t}{\lambda}\right)^{\frac{2}{p}},$$

and

$$\|f - f_*\|_n \leq t.$$

We therefore may write $f \in \mathcal{F}(t)$ as $f_1 + f_2$, with $|f_1| \leq I(f_1) = I(f)$ and $\|f_2 - f_*\|_n \leq t + I(f)$. It is now not difficult to show that for some constant C_1

$$H(u, \mathcal{F}(t), \|\cdot\|_n) \leq C_1 \left((\frac{t}{\lambda})^{\frac{2}{ps}} u^{-\frac{1}{s}} + \log(\frac{t}{(\lambda \wedge 1)u}) \right), \ 0 < u < t.$$

□

Corollary 9.5.2. *By applying Lemma 9.4.1, we find that for some constant c_1,*

$$\mathbf{E}\{\|\hat{f}_n - f_0\|_n^2 + \lambda^2 I^p(\hat{f}_n)\} \leq 2 \min_f \{\|f - f_0\|_n^2 + \lambda^2 I^p(f)\}$$

$$+c_1 \left(\left(\frac{1}{n\lambda^{\frac{2}{ps}}} \right)^{\frac{2ps}{2ps+p-2}} + \frac{\log\left(\frac{1}{\lambda} \vee 1\right)}{n} \right).$$

9.5.2. Overruling the variance in this case. For choosing the smoothing parameter λ, the above suggests the penalty

$$\text{pen}(f) = \min_\lambda \left\{ \lambda^2 I^p(f) + + \left(\frac{C_0}{n\lambda^{\frac{2}{ps}}} \right)^{\frac{2ps}{2ps+p-2}} \right\},$$

with C_0 a suitable constant. The minimization within this penalty yields

$$\text{pen}(f) = C_0' n^{-\frac{2s}{2s+1}} I^{\frac{2}{2s+1}}(f),$$

where C_0' depends on C_0 and s. From the computational point of view (in particular, when $p = 2$), it may be convenient to carry out the penalized least squares as in the previous subsection, for all values of λ, yielding the estimators

$$\hat{f}_n(\cdot, \lambda) = \arg\min_f \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - f(x_i)|^2 + \lambda^2 I^p(f) \right\}.$$

Then the estimator with the penalty of this subsection is $\hat{f}_n(\cdot, \hat{\lambda}_n)$, where

$$\hat{\lambda}_n = \arg\min_{\lambda > 0} \left\{ \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{f}_n(x_i, \lambda)|^2 + \left(\frac{C_0}{n\lambda^{\frac{2}{ps}}} \right)^{\frac{2ps}{2ps+p-2}} . \right\}.$$

From the same calculations as in the proof of Lemma 9.5.1, one arrives at the following corollary.

Corollary 9.5.3 *For an appropriate, large enough, choice of C_0' (or C_0), depending on c, p and s, we have for a constant c_0' depending on c, c', C_0' (C_0), p and s.*

$$\mathbf{E}\left\{\|\hat{f}_n - f_0\|_n^2 + C_0' n^{-\frac{2s}{2s+1}} I^{\frac{2}{2s+1}}(\hat{f}_n)\right\}$$

$$\leq 2 \min_f \left\{\|f - f_0\|_n^2 + C_0' n^{-\frac{2s}{2s+1}} I^{\frac{2}{2s+1}}(f)\right\} + \frac{c_0'}{n}.$$

Thus, the estimator adapts to small values of $I(f_0)$. For example, when $s = 1$ and $I(f_0) = 0$ (i.e., when f_0 is the constant function), the excess risk of the estimator converges with parametric rate $1/n$. If we knew that f_0 is constant, we would of course use the $\sum_{i=1}^n Y_i/n$ as estimator. Thus, this penalized estimator mimics an oracle.

9.6. Exercise.

Exercise 9.1. Using Lemma 9.2.1, and the formula

$$EZ = \int_0^\infty \mathbf{P}(Z \geq t) dt$$

for a non-negative random variable Z, derive bounds for the average excess risk $\mathbf{E}\|\hat{f}_n - f_0\|_n^2$ of the estimator considered there.