Empirical Process Theory

Sara van de Geer

January 2020

Contents

1 Introduction **5**

2 Glivenko Cantelli classes **11**
2.1 Law of large numbers for real-valued random variables 11
2.2 Law of large numbers for $\mathbb{R}^d$-valued random variables. 12
2.3 Definition Glivenko Cantelli classes of sets 13
2.4 Convergence of averages to their expectations 13
2.5 ULLN and maximum likelihood 14

3 (Exponential) probability inequalities **17**
3.1 Chebyshev's inequality 17
3.2 Hoeffding's inequality 19
3.3 Bernstein's inequality 22
3.4 Exercises 24

4 ULLNs based on entropy with bracketing **25**
4.1 The classical Glivenko Cantelli Theorem 25
4.2 Entropy 26
4.3 Entropy with bracketing 27
4.4 The envelope function 29
4.5 Entropy with bracketing and compactness 30
4.6 Maximum likelihood for mixture models 30
4.7 Exercises 32

5 Symmetrization **33**
5.1 Intermezzo: some facts about (conditional) expectations 34
5.1.1 Suprema in-/outside the expectation 34
5.1.2 Iterated expectations 34
5.2 Symmetrization with means 34
5.3 Symmetrization with probabilities 36
5.4 Exercises 37

6 ULLNs based on symmetrization **39**
6.1 Classes of functions 39
6.2 Classes of sets 43
6.3 Vapnik Chervonenkis classes 45

6.4 VC graph classes of functions 46
6.5 Exercises 48

7 M-estimators **51**
7.1 What is an M-estimator? 51
7.2 Consistency 52
7.3 Exercises 54

8 Uniform central limit theorems **55**
8.1 Real-valued random variables 55
8.2 $\mathbb{R}^r$-valued random variables 56
8.3 Donsker's theorem 56
8.4 Donsker classes 57

9 Chaining and asymptotic continuity **61**
9.1 Chaining 61
9.2 Increments of the symmetrized process 62
9.3 De-symmetrizing 63
9.4 Asymptotic continuity of the empirical process 64
9.5 Application to VC graph classes 65
9.6 Exercises 65

10 Asymptotic normality of M-estimators **67**
10.1 Asymptotic linearity 67
10.2 Conditions a, b and c for asymptotic normality 68
10.3 Asymptotics for the median 70
10.4 Conditions aa, bb and cc for asymptotic normality 71
10.5 Exercises 72

11 Rates of convergence for LSEs **75**
11.1 Gaussian errors 76
11.2 Rates of convergence 76
11.3 Examples 77
11.4 Exercises 81

12 Regularized least squares **83**
12.1 Estimation and approximation error 84
12.2 Finite models 85
12.3 Nested finite models 86
12.4 General penalties 87
12.5 Application to the "classical" penalty 89
12.5.1 Fixed smoothing parameter 89
12.5.2 Overruling the variance in this case 90
12.6 Exercises 91

Chapter 1

Introduction

This introduction 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

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

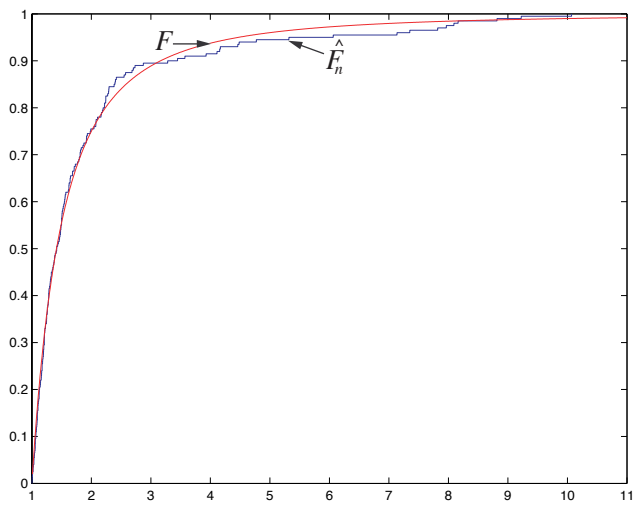

Figure 1

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 $\mathbb{R}$. 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 100th observation is equal to 1.4166 and the 101th 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 $\theta = (\theta_1, \cdots, \theta_r)$. 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.

Nonparametric models cannot be described by finitely many parameters.

Example: density estimation. An example of a nonparametric model is where one assumes that the density f of the distribution function F on $\mathbb{R}$ 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. Suppose 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.
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.

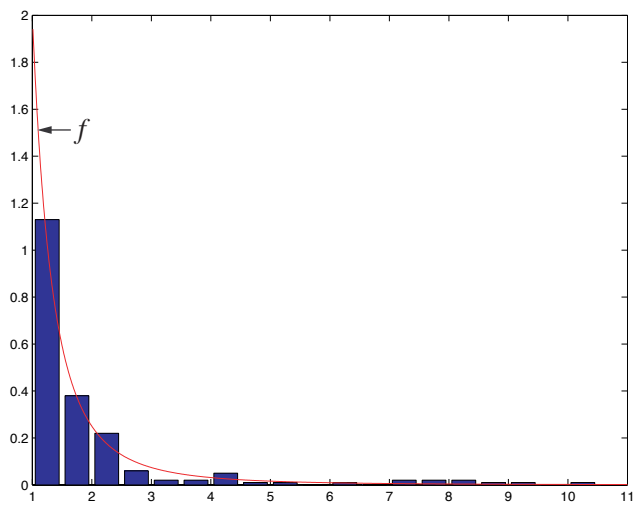

Figure 2

The question arises how to choose the bandwidth h? One may want to apply a data-dependent choice. For example, introduce for $\hat{f}_n = \hat{f}_{n,h}$ depending on h the risk function

$$R(\hat{f}_{n,h}) := \int (\hat{f}_{n,h}(t) - f(t))^2 dt.$$

Note that

$$R(\hat{f}_{n,h}) = \int \hat{f}_{n,h}^2(t)dt - 2\int \hat{f}_{n,h}(t)f(t)dt + \underbrace{\int f^2(t)dt}_{\text{does not depend on } h}.$$

The double product term in the middle can be written as

$$2\int \hat{f}_{n,h}(t)dF(t).$$

It can be estimated by

$$2\int \hat{f}_{n,h}(t)d\hat{F}_n(t) = \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,h}(X_i).$$

But using the data twice, both for estimation as well as for estimation of the performance of the estimator, is perhaps not a good idea as it may lead to "overfitting". Therefore, we propose here to use a fresh sample, that is, $\{X_i'\}_{i=1}^m$, i.i.d. copies from X independent of $\{X_i\}_{i=1}^n$, and apply the estimated bandwidth

$$\hat{h} := \arg\min_{h>0}\left\{\int \hat{f}_{n,h}^2(t)dt - \frac{2}{m}\sum_{i=1}^m \hat{f}_{n,h}(X_i')\right\}.$$

Note that the total sample is now

$$(X_1, \ldots, X_n, X_1', \ldots, X_m'),$$

with sample size $n+m$. In other words, for the selection of the bandwidth a sample splitting technique is used. In order to further improve performance, a common technique is "cross-validation" consisting of several sample splits (into training and test sets).

Example: the classification problem. Let $Y \in \{0,1\}$ be a response variable (a label) and $X \in \mathcal{X}$ a co-variable (or input). For all $x \in \mathcal{X}$ we define the probability that the label is $Y=1$ when the co-variable takes the value x:

$$\eta(x) := P(Y=1|X=x).$$

Our aim is now to predict the label given some input, say x_0. Call the predicted value y_0. Bayes rule says: predict the most likely label, that is predict

$$y_0 := \begin{cases} 1 & \text{if } \eta(x_0) > 1/2 \\ \text{undecided, say 1,} & \text{if } \eta(x_0) = 1/2 \\ 0 & \text{if } \eta(x_0) < 1/2 \end{cases}.$$

Note that $\eta(x) > 1/2$ if and only if the log-odds ratio

$$\log\left(\frac{\eta(x)}{1-\eta(x)}\right)$$

is strictly positive. When $\mathcal{X}$ is a subset of r-dimensional Euclidean space $\mathbb{R}^r$ one may want to assume the parametric logistic regression model, where the log-odds ratio is a linear function of the (row) vector x:

$$\log\left(\frac{\eta(x)}{1-\eta(x)}\right) = \alpha^0 + x\beta^0, \ x \in \mathcal{X},$$

with $\alpha^0 \in \mathbb{R}$ an unknown intercept and $\beta^0 \in \mathbb{R}^d$ an unknown (column) vector of coefficients. Bayes rule is now l_{A_0} where $A_0 := \{x : \ \alpha^0 + x\beta^0 \geq 0\}$ is a half-space. Maximum likelihood estimators $\hat{\alpha}_{\mathrm{MLE}}$ and $\hat{\beta}_{\mathrm{MLE}}$ of the unknown parameters can be obtained from labeled data $\{(X_i, Y_i)\}_{i=1}^n$, which are i.i.d. copies of (X,Y). This is called logistic regression. Then the estimated Bayes rule is $\mathrm{l}_{\hat{A}_{\mathrm{MLE}}}$, where $\hat{A}_{\mathrm{MLE}} := \{x : \ \hat{\alpha}_{\mathrm{MLE}} + x\hat{\beta}_{\mathrm{MLE}} \geq 0\}$.
The *risk* $R(A)$ of a predictor l_A is the probability that it makes a mistake:

$$R(A) = P(Y \neq \mathrm{l}_A(X)) = P(|Y - \mathrm{l}_A(X)| = 1).$$

Then one may verify that Bayes rule is

$$A_0 = \arg\min_{A \subset \mathcal{X}} R(A).$$

One my mimic this by replacing the unknown P by the empirical measure P_n. Then the empirical risk becomes

$$R_n(A) = \#\Big\{(X_i, Y_i) : \ Y_i \neq 1_A(X_i)\Big\} = \sum_{i=1}^n |Y_i - 1_A(X_i)|$$

and the empirical risk minimizer is

$$\hat{A}_{\text{ERM}} := \arg\min_{A \in \mathcal{A}} R_n(A)$$

where $\mathcal{A}$ is a given collection of subsets of $\mathcal{X}$, for example all half-spaces, such as is the case in logistic regression. In this notes, we will develop theory for deducing how close $\hat{A} = \hat{A}_{\text{ML}}$ or $\hat{A} = \hat{A}_{\text{ERM}}$ is to A_0, for example in terms of the measure of the symmetric difference $\hat{A} \Delta A_0$.

Sub-example. Here we give an illustration of what can be achieved, although the details will not be treated in these notes. Suppose $X \in \mathbb{R}$, that $A_0 = [\theta^0, \infty)$ and that $\mathcal{A}$ is the collection of all half-intervals (for example X could be the person's pitch of voice, and Y the person's gender). Let $\hat{A}_{\text{ERM}} =: [\hat{\theta}, \infty)$. Then, using arguments from empirical process theory, one can show that under certain conditions, $\hat{\theta}$ converges with rate $n^{-1/3}$ and that

$$n^{1/3}(\hat{\theta} - \theta^0)$$

converges in distribution to the maximum of a Brownian motion minus parabola.

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 is typically the case when the ordering in the data contains no information[1]), 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] When the sample is splitted in a training and a test set this symmetry is lost.

Chapter 2

Glivenko Cantelli classes

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 $\mathbb{P}$.

Let $\{T_n, T\}$ be a collection of real-valued random variables.

Definition 2.0.1 *The sequence T_n converges* in probability *to T, if for all $\epsilon > 0$,*

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

Notation: $T_n \xrightarrow{\mathbf{P}} T$ or $T_n = T + o_{\mathbf{P}}(1)$.
Moreover, T_n converges almost surely (a.s.) *to T if*

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

Remark. Convergence almost surely implies convergence in probability.

Definition 2.0.2 *We say that $\{T_n\}$ remains bounded in probability if*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(|T_n| > M) = 0.$$

Notation: $T_n = \mathcal{O}_{\mathbf{P}}(1)$.

2.1 Law of large numbers for real-valued random variables

Consider the case $\mathcal{X} = \mathbb{R}$. 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

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

be the theoretical distribution function, and

$$\hat{F}_n(t) := \frac{1}{n}\#\{X_i \leq t,\ 1 \leq i \leq n\},\ t \in \mathbb{R},$$

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

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

We will prove the Glivenko Cantelli Theorem, which says that

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

This is a **uniform** law of large numbers (ULLN).

Application: *Kolmogorov's goodness-of-fit test.* We want to test

$H_0 :\ F = F_0$.

Test statistic:

$$T_n := \sup_t |\hat{F}_n(t) - F_0(t)|.$$

Reject H_0 for large values of T_n.

2.2 Law of large numbers for $\mathbb{R}^d$-valued random variables.

Questions:

(i) What is a natural extension of half-intervals in $\mathbb{R}$ to higher dimensions?

(ii) Does Glivenko Cantelli hold for this extension?

2.3 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 2.3.1 *The collection $\mathcal{D}$ is called a* **Glivenko Cantelli** *(GC) class if*

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

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

$$\mathcal{D} = \{\mathrm{l}_{(-\infty,t]} :\ t \in \mathbb{R}\}$$

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.

2.4 Convergence of averages to their expectations

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

$$Pg := Eg(X),$$

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 2.4.1 *The class $\mathcal{G}$ is called a* **Glivenko Cantelli** *(GC) class if*

$$\sup_{g \in \mathcal{G}} |P_n(g) - P(g)| \xrightarrow{\mathbf{P}} 0.$$

We will often use the notation

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

[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.

2.5 ULLN and maximum likelihood

Let $\{P_\theta : \theta \in \Theta\}$ be a collection of probability measures dominated by a σ-finite measure μ. Suppose that $P = P_{\theta^0}$ for some $\theta^0 \in \Theta$. Define

$$p_\theta := \frac{dP_\theta}{d\mu}, \ \theta \in \Theta.$$

The maximum likelihood estimator (if it exists) is

$$\hat{\theta} := \arg\max_{\theta \in \Theta} P_n \log p_\theta.$$

We now let $\mathcal{P} := \{p_\theta : \theta \in \Theta\}$ be the collection of densities in our model, write $p_0 := p_{\theta^0}$ for the true density and $\hat{p} := p_{\hat{\theta}}$ for the estimated density. Thus

$$\hat{p} = \arg\max_{p \in \mathcal{P}} P \log p.$$

Definition 2.5.1 *The* Hellinger distance *between densities p and $\tilde{p}$ is*

$$h(p, \tilde{p}) := \left(\frac{1}{2} \int (\sqrt{p} - \sqrt{\tilde{p}})^2 d\mu \right)^{1/2}.$$

We define

$$\mathcal{G} := \left\{ \frac{1}{2} \log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} : \ p \in \mathcal{P} \right\}.$$

Theorem 2.5.1 *Suppose that $\mathcal{G}$ is GC. Then*

$$h(\hat{p}, p_0) \xrightarrow{\mathbf{P}} 0.$$

Proof. By the concavity of the log-function

$$\log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} \geq \frac{1}{2} \log \frac{p}{p_0} 1\!\!1\{p_0 > 0\}.$$

Thus

$$\begin{aligned} 0 &\leq P_n\left(\log \frac{\hat{p}}{p_0} 1\!\!1\{p_0 > 0\} \right) \\ &\leq 2P_n\left(\log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} \right) \\ &= 2(P_n - P)\left(\log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} \right) + 2P\left(\log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} \right). \end{aligned}$$

But

$$\begin{aligned} & \frac{1}{2} P\left(\log \frac{p + p_0}{2p_0} 1\!\!1\{p_0 > 0\} \right) \\ \leq \ & P\left(\sqrt{\frac{p + p_0}{2p_0}} 1\!\!1\{p_0 > 0\} \right) - 1 \\ = \ & \int \sqrt{\frac{p + p_0}{2}} \sqrt{p_0} d\mu - 1 \\ = \ & -h^2\left(\frac{p + p_0}{2}, p_0 \right). \end{aligned}$$

It follows that

$$h^2\left(\frac{\hat{p}+p_0}{2}, p_0\right) \leq \|P_n - P\|_{\mathcal{G}}.$$

Finally

$$\begin{aligned} h^2(p, p_0) &= \frac{1}{2}\int (\sqrt{p} - \sqrt{p_0})^2 \\ &= \frac{1}{2}\int 4\left(\frac{\sqrt{\frac{p+p_0}{2}} + \sqrt{p_0}}{\sqrt{p} + \sqrt{p_0}}\right)^2 \left(\sqrt{\frac{p+p_0}{2}} - \sqrt{p_0}\right)^2 \\ &\leq 16 h^2\left(\frac{\hat{p}+p_0}{2}, p_0\right). \end{aligned}$$

So we conclude

$$h^2(\hat{p}, p_0) \leq 16\|P_n - P\|_{\mathcal{G}}.$$

By assumption, $\mathcal{G}$ is GC, i.e., $\|P_n - P\|_{\mathcal{G}} \xrightarrow{\mathbf{P}} 0$. $\square$

Chapter 3

(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 Gaussian distributions.) 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.

3.1 Chebyshev's inequality

Theorem 3.1.1 *(***Chebyshev's inequality***) Consider a random variable $X \in \mathbb{R}$ with distribution P, and an increasing function $\phi : \mathbb{R} \to [0, \infty)$. Then for all a with $\phi(a) > 0$, we have*

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

Proof.

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

□

Example. Let $X \in \mathbb{R}$, with mean $\mu := EX$ and finite variance $\sigma^2 := \text{var}(X) = E(X-\mu)^2$. Let $X_1, \ldots, X_n, \ldots$ be i.i.d. copies of X and $\bar{X}_n := \sum_{i=1}^n X_i/n$ be the sample average. Then by Chebyshev's inequality

$$\mathbb{P}\Big(|\bar{X}_n - \mu| \geq a\Big) \leq \frac{\mathbb{E}(\bar{X}_n - \mu)^2}{a^2} = \frac{\sigma^2}{na^2} \to 0 \ \forall \ a > 0.$$

Thus

$$\bar{X}_n \to^{\mathbf{P}} \mu$$

((weak) law of large numbers[1]). In a reformulation we put $a = \sigma\sqrt{t/n}$,

$$\mathbb{P}\Big(|\bar{X}_n - \mu| \geq \sigma\sqrt{\frac{t}{n}}\Big) \leq \frac{1}{t}, \ \forall \ t > 0.$$

In other words (see Definition 2.0.2)

$$\bar{X}_n = \mu + \mathcal{O}_{\mathbf{P}}(1/\sqrt{n}).$$

We say that $\bar{X}_n$ converges in probability to μ with rate $1/\sqrt{n}$.

Example. Let $\{g_1, \ldots, g_N\}$ be a collection of N real-valued functions with domain $\mathcal{X}$, such that for some fixed constant σ^2, $\text{var}(g_j(X)) \leq \sigma^2$ for all $j \in \{1, \ldots, N\}$. This class of functions is allowed to depend on n and N is allowed to grow with n. Then by the union bound[2]

$$\mathbb{P}\Big(\max_{1\leq j\leq N} |(P_n - P)g_j| \geq a\Big) \leq \sum_{j=1}^N \mathbb{P}\Big(|(P_n - P)g_j| \geq a\Big) \leq \frac{N\sigma^2}{na^2} \ \forall \ a > 0.$$

We may reformulate this to

$$\mathbb{P}\Big(\max_{1\leq j\leq N} |(P_n - P)g_j| \geq \sigma\sqrt{\frac{Nt}{n}}\Big) \leq \frac{1}{t} \ \forall \ t > 0.$$

Thus, with σ kept fixed, we conclude that

$$\max_{1\leq j\leq N} |(P_n - P)g_j| = \mathcal{O}_{\mathbf{P}}(\sqrt{N/n}).$$

Our aim is now to improve here the dependence on N.

Example. Let X be $\mathcal{N}(0,1)$-distributed. By Exercise 3.4.1

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

[1]This implies $\bar{X}_n \to \mu$ almost surely by a martingale argument (skipped here).
[2]The union bound says that $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for any two sets A and B.

Corollary 3.1.1 *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

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2b^2}\right] \ \forall\ a > 0,$$

or reformulated

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq b\sqrt{2t}\right) \leq \exp[-t] \ \forall\ t > 0.$$

Corollary 3.1.2 *Let $Z_1, \ldots, Z_N$ be possibly dependent real-valued random variables, and let Z_j have the $\mathcal{N}(0, \sigma_j^2)$-distribution, $j = 1, \ldots, N$. Define $\sigma^2 := \max_{1 \leq j \leq N} \sigma_j^2$. Then*

$$\mathbb{P}\left(\max_{1 \leq j \leq N} |Z_j| \geq \sigma\sqrt{2(t + \log N)}\right) \leq 2N \exp[-(t + \log N)] = 2\exp[-t] \ \forall\ t > 0.$$

3.2 Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent real-valued random variables.

Definition 3.2.1 *(**Hoeffding's condition**) Suppose that for all i, $\mathbb{E}X_i = 0$, and for certain constants $c_i > 0$,*

$$|X_i| \leq c_i.$$

Lemma 3.2.1 *Assume Hoeffding's condition. Let $b^2 := \sum_{i=1}^n c_i^2$. Then for all $\lambda > 0$*

$$\mathbb{E}\exp\left[\lambda \sum_{i=1}^n X_i\right] \leq \exp\left[\frac{\lambda^2 b^2}{2}\right].$$

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 $\mathbb{E}\alpha_i = 1/2$, we find

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

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

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

Therefore,

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

□

Theorem 3.2.1 *(***Hoeffding's inequality***) Assume Hoeffding's condition. Let* $b^2 := \sum_{i=1}^{n} c_i^2$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2\sum_{i=1}^{n} c_i^2}\right] \ \forall\ a > 0,$$

or reformulated

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq b\sqrt{2t}\right) \leq \exp[-t]\ \forall\ t > 0.$$

Proof. It follows from Chebyshev's inequality that

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

Take $\lambda = a/b^2$ to obtain the first part. Take $a = b\sqrt{2t}$ to arrive at the reformulation.

□

Corollary 3.2.1 *Consider N functions $g_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, N$, with, for some constant K,*

$$\forall\ j \in \{1, \ldots, N\} : \begin{cases} Eg_j(X) = 0 \\ \sup_{x \in \mathcal{X}} |g_j(x)| \leq K \end{cases} .$$

Then for all j

$$\mathbb{P}\left(|P_n g_j| \geq K \sqrt{\frac{2t}{n}} \right) \leq 2 \exp[-t],\ \forall\ t > 0.$$

Hence by the union bound

$$\mathbb{P}\left(\max_{1 \leq j \leq N} |P_n g_j| \geq K \sqrt{\frac{2(t + \log N)}{n}} \right) \leq 2 \exp[-t],\ \forall\ t > 0.$$

One can also derive an inequality for the expectation of a maximum (instead of a probability inequality).

Lemma 3.2.2 *Consider N functions $g_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, N$, with, for some constant K,*

$$\forall\ j \in \{1, \ldots, N\} : \begin{cases} Eg_j(X) = 0 \\ \sup_{x \in \mathcal{X}} |g_j(x)| \leq K \end{cases} .$$

Then

$$\mathbb{E}\left(\max_{1 \leq j \leq N} |P_n g_j| \right) \leq K \sqrt{\frac{2 \log(2N)}{n}}.$$

Proof. We have

$$\begin{aligned} \mathbb{E} \exp[\lambda |n P_n g_j|] &\overset{e^{|x|} \leq e^x + e^{-x}}{\leq} \mathbb{E} \exp[\lambda n P_n g_j] + \mathbb{E} \exp[\lambda n P_n g_j] \\ &\overset{\text{by Lemma 3.2.1}}{\leq} 2 \exp\left[\frac{\lambda^2 K^2}{2} \right]. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}\left(\max_{1 \leq j \leq N} n |P_n g_j| \right) &= \frac{1}{\lambda} \mathbb{E} \log \exp\left[\lambda \max_{1 \leq j \leq N} n |P_n g_j| \right] \\ &\overset{\text{Jensen}}{\leq} \frac{1}{\lambda} \log \mathbb{E} \exp\left[\lambda \max_{1 \leq j \leq N} n |P_n g_j| \right] \\ &= \frac{1}{\lambda} \log \mathbb{E}\left(\max_{1 \leq j \leq N} \exp\left[\lambda n |P_n g_j| \right] \right) \\ &\leq \frac{1}{\lambda} \log \sum_{j=1}^{N} \mathbb{E}\left(\exp\left[\lambda n |P_n g_j| \right] \right) \\ &\leq \frac{1}{\lambda} \log\left(2N \exp\left[\frac{\lambda^2 K^2}{2} \right] \right) \\ &= \frac{\log(2N)}{\lambda} + \frac{\lambda K^2}{2}. \end{aligned}$$

We minimize the last expression over λ. Take the derivative and put it to zero:

$$-\frac{\log(2N)}{\lambda^2} + \frac{K^2}{2} \stackrel{\Delta}{=} 0.$$

This gives

$$\lambda = \frac{\sqrt{2\log(2N)}}{K}$$

and with this value of λ

$$\frac{\log(2N)}{\lambda} + \frac{\lambda K^2}{2} = K\sqrt{\frac{\log(2N)}{2}} + K\sqrt{\frac{\log(2N)}{2}} = K\sqrt{2\log(2N)}.$$

□

3.3 Bernstein's inequality

Bernstein's inequality is a close relative of Hoeffding's inequality and an important tool in empirical process theory. We present it here. However, unfortunately we have no space in these lecture notes to show its power.

Let $X_1, \ldots, X_n$ be independent real-valued random variables.

Definition 3.3.1 *(***Bernstein's condition***) For all i,*

$$\mathbb{E}X_i = 0,$$

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

Lemma 3.3.1 *Suppose Bernstein's condition. Define $b^2 := \sum_{i=1}^n \sigma_i^2$. Then for all $0 < \lambda < 1/K$*

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

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

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

□

Theorem 3.3.1 (**Bernstein's inequality**) *Suppose Bernstein's condition. Define* $b^2 = \sum_{i=1}^n \sigma_i^2$. *We have*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq a\right) \leq \exp\left[-\frac{a^2}{2(aK+b^2)}\right] \ \forall\ a > 0.$$

Or reformulated

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq b\sqrt{2t} + Kt\right) \leq \exp[-t]\ \forall\ t > 0.$$

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

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

It follows that

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

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

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

Take

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

to complete the first part. For the reformulation, choose $a = b\sqrt{2t} + Kt$.

□

Corollary 3.3.1 *Consider N functions $g_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, N$, with, for some constants σ^2 and K*

$$\forall\ j \in \{1, \ldots, N\} : \begin{cases} Eg_j(X) = 0 \\ Eg_j^2(X) \leq \sigma^2 \\ \sup_{x\in\mathcal{X}} |g_j(x)| \leq K \end{cases}.$$

Then for all j

$$\mathbb{P}\left(|P_n g_j| \geq \sigma\sqrt{\frac{2t}{n}} + \frac{Kt}{n}\right) \leq 2\exp[-t],\ \forall\ t > 0.$$

Hence by the union bound

$$\mathbb{P}\left(\max_{1\leq j\leq N} |P_n g_j| \geq \sigma\sqrt{\frac{2(t+\log N)}{n}} + \frac{K(t+\log N)}{n}\right) \leq 2\exp[-t],\ \forall\ t > 0.$$

3.4 Exercises

Exercise 3.4.1 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

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

Exercise 3.4.2 Consider N functions $g_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, N$, with, for some constants σ^2 and K,

$$\forall\ j \in \{1, \ldots, N\} : \begin{cases} Eg_j(X) = 0 \\ Eg_j^2(X) \leq \sigma^2 \\ \sup_{x\in\mathcal{X}} |g_j(x)| \leq K \end{cases}.$$

Derive an inequality for $\mathbb{E}(\max_{1\leq j\leq N} |P_n g_j|)$ using Bernstein's inequality (Theorem 3.3.1) instead of Hoeffding's inequality (Theorem 3.2.1).

Exercise 3.4.3 Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable $X \in \mathbb{R}$ with mean zero. Suppose that for some constants λ and σ,

$$\mathbb{E}\exp[\lambda|X|] - 1 - \lambda|X| \leq \varsigma^2.$$

Show that Bernstein's condition (Condition 3.3.1) holds with appropriate constants σ^2 and K.

Exercise 3.4.4 Recall the class

$$\mathcal{G} := \left\{\frac{1}{2}\log\frac{p+p_0}{2p_0}1\{p_0 > 0\} : p \in \mathcal{P}\right\},\ p_0 := \frac{dP}{d\mu},$$

defined in Section 2.5. Show that for all $g \in \mathcal{G}$ the random variable $g(X) - Eg(X)$ satisfied Bernstein's condition (Condition 3.3.1). Hint: use Exercise 3.4.3.

Chapter 4

ULLNs based on entropy with bracketing

4.1 The classical Glivenko Cantelli Theorem

For the case $\mathcal{X} = \mathbb{R}$ we define the theoretical distribution function

$$F(x) := P(X \leq x),\ x \in \mathbb{R}$$

and the empirical distribution function

$$\hat{F}_n(x) := \frac{1}{n}\#\{X_i \leq x,\ 1 \leq i \leq n\}.$$

Theorem 4.1.1 *(***the classical Glivenko Cantelli Theorem***) It holds that*

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\mathbf{P}} 0.$$

Remark 4.1.1 *The convergence in probability can be strengthened to convergence almost surely.*

Proof of Theorem 4.1.1 for the case F continuous. We have for all x

$$|1_{(-\infty,x]} - F(x)| \leq 1$$

so that by Hoeffding's inequality (Theorem 3.2.1)

$$\mathbb{P}\left(|\hat{F}_n(x) - F(x)| \geq \sqrt{2t/n}\right) \leq 2\exp[-t]\ \forall\ t > 0.$$

Let now $0 < \delta < 1$ be arbitrary and take $a_0 < a_1 < \cdots < a_N$ such that $F(a_j) - F(a_{j-1}) = \delta$. Note that then $N \leq 1 + 1/\delta$. If $x \in (a_{j-1}, a_j]$ we clearly have

$$(-\infty, a_{j-1}] \subset (-\infty, x] \subset (-\infty, a_j],$$

that is, we can "bracket" $\mathrm{l}_{(-\infty,x]}$ between a lower function $\mathrm{l}_{(-\infty,a_{j-1}]}$ and an upper function $\mathrm{l}_{(-\infty,a_j]}$. This gives for $x \in (a_{j-1}, a_j]$

$$\begin{aligned} \hat{F}_n(x) - F(x) &\leq \hat{F}_n(a_j) - F(x) \\ &\leq \hat{F}_n(a_j) - F(a_{j-1}) \\ &= \hat{F}_n(a_j) - F(a_j) + \delta \end{aligned}$$

and

$$\begin{aligned} \hat{F}_n(x) - F(x) &\geq \hat{F}_n(a_{j-1}) - F(x) \\ &\geq \hat{F}_n(a_{j-1}) - F(a_j) \\ &= \hat{F}_n(a_{j-1}) - F(a_{j-1}) - \delta \end{aligned}$$

It follows that

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \max_{j=1,\ldots,N} |\hat{F}_n(a_j) - F(a_j)| + \delta.$$

By Hoeffding's inequality and the union bound

$$\mathbb{P}\left(\max_{j=1,\ldots,N} |\hat{F}_n(a_j) - F(a_j)| \geq \sqrt{\frac{2(t+\log N)}{n}} \right) \leq 2\exp[-t], \ \forall\ t > 0.$$

Take

$$t = \frac{n\delta^2}{4}.$$

Take n sufficiently large:

$$n \geq \frac{4}{\delta^2} \log N.$$

Then

$$\sqrt{\frac{2(t+\log N)}{n}} \leq \delta.$$

and we find

$$\mathbb{P}\left(\max_{j=1,\ldots,N} |\hat{F}_n(a_j) - F(a_j)| \geq \delta \right) \leq 2\exp[-n\delta^2/4].$$

Hence for $n \geq 4\log(1+1/\delta)/\delta^2$

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \geq 2\delta \right) \leq 2\exp[-n\delta^2/4].$$

□

4.2 Entropy

Definition 4.2.1 *Consider a subset S of a metric space with metric d. For any $\delta > 0$, let $N(\delta, S, d)$ be the minimum number of balls with radius δ, necessary to cover S. Then $N(\delta, S, d)$ is called the* δ-covering number *of S. Moreover*

$$H(\cdot, S, d) := \log N(\cdot, S, d)$$

is called the entropy *of S.*

Remark. We sometimes require the centres of the balls to lie in S.

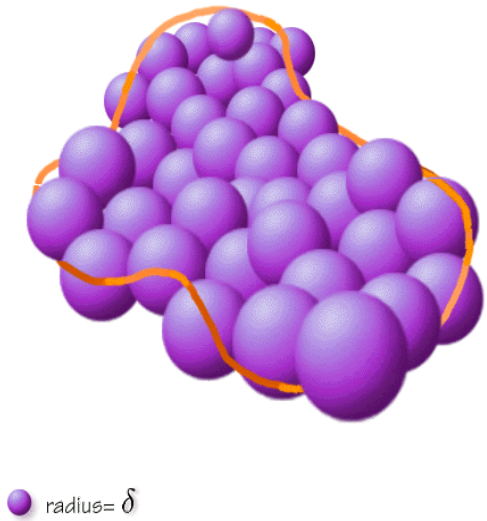

Figure 3

Example 4.2.1 *Let $S = [-1,1]^r$ be the r-dimensional hypercube. Take as metric*

$$d(x,y) := \max_{1\le k\le r} |x_k - y_k| =: \|x-y\|_\infty,\ x \in \mathbb{R}^r,\ y \in \mathbb{R}^r.$$

Then for all $0 < \delta \le 1$ and for $a > 0$, $\lceil a \rceil := \min m \in \mathbb{N} : m \ge a$,

$$\begin{aligned} N(\delta, S, d) &\le \lceil 1/\delta \rceil^r \\ &\le (1 + 1/\delta)^r, \\ H(\delta, S, d) &\le r \log(1 + 1/\delta). \end{aligned}$$

4.3 Entropy with bracketing

We define for $q \ge 1$

$$L_q(P) := \{g :\ \mathcal{X} \to \mathbb{R} :\ P|g|^q < \infty\}.$$

Endow this space with the norm

$$\|g\|_q := (P|g|^q)^{1/q},\ g \in L_q(P).$$

We let

$$L_\infty = \{g :\ \|g\|_\infty < \infty\},$$

where

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

is the sup-norm.

Definition 4.3.1 *Consider a class $\mathcal{G} \subset L_q(P)$. For any $\delta > 0$, let $\{[g_j^L, g_j^U]\}_{j=1}^N \subset L_q(P)$ be such that*
(i) $\forall\ j,\ g_j^L \le g_j^U$ and $\|g_j^U - g_j^L\|_q \le \delta$,
(ii) $\forall\ g \in \mathcal{G}\ \exists\ j$ such that $g_j^L \le g \le g_j^U$.
*We then call $\{[g_j^L, g_j^U]\}_{j=1}^N$ a $\delta-$*bracketing set*.*
*The δ-*covering number with bracketing *is*

$$N_{q,B}(\delta, \mathcal{G}, P) := \min\left\{N :\ \exists\ \delta-\text{bracketing set } \{[g_j^L, g_j^U]\}_{j=1}^N\right\}.$$

The entropy with bracketing *is*

$$H_{q,B}(\cdot, \mathcal{G}, P) := \log N_{q,B}(\cdot, \mathcal{G}, P).$$

Remark 4.3.1 *Note that*

$$H_{q,B}(\delta, \mathcal{G}, P) \le H_\infty(\delta/2, \mathcal{G})$$

where $H_\infty(\cdot, \mathcal{G}) := H(\cdot, \mathcal{G}, \|\cdot\|_\infty)$ is the entropy of $\mathcal{G}$ endowed with the metric induced by the sup-norm $\|\cdot\|_\infty$.

Example 4.3.1 *A function g on a subinterval $\mathcal{X}$ of $\mathbb{R}$ is called L-Lipschitz if*

$$|g(x) - g(\tilde{x})| \le L|x - \tilde{x}|\ \forall\ x, \tilde{x} \in \mathcal{X}.$$

Let now $\mathcal{X} = (0, 1]$ and

$$\mathcal{G} := \{g :\ (0, 1] \to [0, 1] :\ g\ 1-\text{Lipschitz}\}.$$

We partition the interval $[0, 1]$ into $m \le 1 + 1/\delta$ intervals $(a_{j-1}, a_j]$ with length at most δ, $0 = a_0 < \cdots < a_N = 1$. For $g \in \mathcal{G}$ and $x \in (a_{j-1}, a_j]$ we let

$$\tilde{g}(x)/\delta = \lfloor g(a_j)/\delta \rfloor$$

be the integer part of $g(a_j)/\delta$. Then, since $g(a_j) - \delta < \lfloor g(a_j)/\delta \rfloor \delta \le g(a_j)$,

$$|g(x) - \tilde{g}(x)| \le |g(x) - g(a_j)| + \delta \le 2\delta.$$

We have at most $1 + 1/\delta$ choices for $\lfloor g(a_1)/\delta \rfloor$. Given $\lfloor g(a_j)/\delta \rfloor$ we have

$$\left|\lfloor g(a_{j+1})/\delta \rfloor - \lfloor g(a_j)/\delta \rfloor\right| \le 2 + |g(a_{j+1}) - g(a_j)|/\delta \le 3$$

so there are at most 7 choices for $\lfloor g(a_{j+1})/\delta \rfloor$. The total number of functions $\tilde{g}$ as g varies is thus at most

$$(1 + 1/\delta) \times \underbrace{7 \times \cdots \times 7}_{m-1 \text{ times}} \le (1 + 1/\delta)7^{1/\delta}.$$

Thus

$$H_\infty(\delta, \mathcal{G}) \le \log(1 + 1/\delta) + \frac{\log 7}{\delta}\ \forall\ 0 < \delta < 1.$$

Example 4.3.2 *Let* $\mathcal{X} = \mathbb{R}$ *and* $\mathcal{G} := \{1_{(-\infty,x]} : \ x \in \mathbb{R}\}$. *Let* $F := P(X \leq \cdot)$ *be continuous (say). Then*

$$H_{q,B}(\delta, \mathcal{G}, P) \leq q \log(1 + 1/\delta), \ 0 < \delta \leq 1.$$

In this case $H_\infty(\cdot, \mathcal{G}) = \infty$ *for all* $0 < \delta < 1$.

Recall the notation

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

Lemma 4.3.1 *(***ULLN based on entropy with bracketing***)*
Suppose

$$H_{1,B}(\cdot, \mathcal{G}, P) < \infty.$$

Then

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

Proof. We use the same arguments as in the proof of Theorem 4.1.1: let $\delta > 0$ be arbitrary and let $\{[g_j^L, g_j^U]\}_{j=1}^N \subset L_1(P)$ be a minimal δ-covering set with bracketing ($N = N_B(\delta, \mathcal{G}, P)$). Because this is a finite set, we know that[1]

$$\left(\max_{1 \leq j \leq N} |(P_n - P)g_j^L| \right) \vee \left(\max_{1 \leq j \leq N} |(P_n - P)g_j^U| \right) \xrightarrow{\mathbf{P}} 0.$$

Now use that when $g_j^L \leq g \leq g_j^U$, $P(g_j^U - g_j^L) \leq \delta$,

$$\begin{aligned} (P_n - P)g &\leq P_n g_j^U - P g_j^L \\ &\leq (P_n - P)g_j^U + \delta \\ (P_n - P)g &\geq P_n g_j^L - P g_j^U \\ &\geq (P_n - P)g_j^L - \delta. \end{aligned}$$

□

4.4 The envelope function

Let $\mathcal{G} \subset L_q(P)$.

Definition 4.4.1 *The* envelope function *(or* envelope*) is*

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

We call $\mathcal{G}$ uniformly bounded *if*

$$\|G\|_\infty < \infty.$$

Note that

$$H_{q,B}(\cdot, \mathcal{G}, P) < \infty \ \Rightarrow \ G \in L_q(P).$$

Similarly

$$H_\infty(\cdot, \mathcal{G}) < \infty \ \Rightarrow \ \|G\|_\infty < \infty.$$

[1] for a and b in $\mathbb{R}$ we let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

4.5 Entropy with bracketing and compactness

Lemma 4.5.1 *Let $\mathcal{G} := \{g_\theta : \ \theta \in \Theta\}$ where (Θ, d) is a compact metric space. Suppose*
- $\theta \mapsto g_\theta(x)$ is continuous for all $x \in \mathcal{X}$,
- $G(\cdot) := \sup_{\theta\in\Theta} |g_\theta(\cdot)| \in L_q(P)$.
Then

$$H_{q,B}(\cdot, \mathcal{G}, P) < \infty.$$

Proof. Define for $\theta \in \Theta$ and $\rho > 0$

$$w_{\theta,\rho}(\cdot) := \sup_{\vartheta:\ d(\theta,\vartheta)<\rho} |g_\theta - g_\vartheta|.$$

Then by continuity

$$\lim_{\rho\to 0} w_{\theta,\rho} = 0.$$

Since $G \in L_q(P)$, by dominated convergence,

$$\lim_{\rho\to 0} P w^q_{\theta,\rho} = 0.$$

Let $\delta > 0$ be arbitrary and ρ_θ be such that

$$P w^q_{\theta,\rho_\theta} \le \delta^q.$$

Define

$$B_\theta := \{\vartheta : \ d(\theta, \vartheta) < \rho_\theta\}.$$

Then $\{B_\theta : \ \theta \in \Theta\}$ is an open cover of Θ. Since Θ is compact there exists a finite sub-cover, say

$$\{B_j := \{\vartheta : \ d(\theta_j, \vartheta) < \rho_{\theta_j}\}\}_{j=1}^N.$$

For $\theta \in B_j$ we have

$$\begin{aligned} g_\theta &\le g_{\theta_j} + w_{\theta_j,\rho_{\theta_j}} =: g_j^U, \\ g_\theta &\ge g_{\theta_j} - w_{\theta_j,\rho_{\theta_j}} =: g_j^L \end{aligned}$$

with

$$P(g_j^U - g_j^L)^q = P(2w_{\theta_j,\rho_{\theta_j}})^q \le (2\delta)^q.$$

□

4.6 Maximum likelihood for mixture models

Here is an example where the ULLN is used for proving consistency of certain maximum likelihood estimators. Let $Y \in \mathbb{R}$ be a random variable with unknown

distribution function F_0. Given $Y = y$, let X have a known distribution with density $k(x|y)$. Then X has mixture density

$$p_{F_0} := \int k(\cdot|y) dF_0(y).$$

WE observe n i.i.d. copies $X_1, \ldots, X_n$ of X. Let $\mathcal{F}$ be the collection of all distribution functions. The maximum likelihood estimator of F_0 is

$$\hat{F} := \arg\max_{F \in \mathcal{F}} P_n \log p_F.$$

Recall the definition of Hellinger distance (Definition 2.5.1):

$$h(p, \tilde{p}) = \left(\frac{1}{2} \int (\sqrt{p} - \sqrt{\tilde{p}})^2 \right)^{1/2}.$$

Lemma 4.6.1 *It holds that*

$$h(p_{\hat{F}}, p_{F_0}) \xrightarrow{\mathbf{P}} 0.$$

Proof. We have $\mathcal{F} \subset (\mathcal{F}^*, d)$ where d is the metric induced by the "vague topology". The space $(\mathcal{F}^*, d)$ is compact.[2] The map

$$F \mapsto p_F = \int k(\cdot|y) dF_0(y)$$

is continuous for d, so

$$F \mapsto \frac{2p_F}{p_F + p_{F_0}}$$

is continuous for d as well. Let

$$\mathcal{G} := \left\{ \frac{2p_F}{p_F + p_{F_0}} : F \in \mathcal{F}^* \right\}.$$

This class has envelope

$$G \leq 2.$$

It follows from Lemma 4.5.1 that $\mathcal{G}$ is GC:

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

We have

$$\begin{aligned} 0 &\leq P_n \left(\log \frac{2p_{\hat{F}}}{p_{\hat{F}} + p_{F_0}} \right) \\ &\leq P_n \left(\frac{2p_{\hat{F}}}{p_{\hat{F}} + p_{F_0}} \right) - 1 \\ &= (P_n - P) \left(\frac{2p_{\hat{F}}}{p_{\hat{F}} + p_{F_0}} \right) + P \left(\frac{2p_{\hat{F}}}{p_{\hat{F}} + p_{F_0}} \right) - 1. \end{aligned}$$

[2] We skip details.

But

$$
\begin{aligned}
h^2(p,p_0) &= \frac{1}{2}\int(\sqrt{p}-\sqrt{p_0})^2 \\
&= \frac{1}{2}\int\frac{(p-p_0)^2}{(\sqrt{p}+\sqrt{p_0})^2} \\
&\leq \frac{1}{2}\int\frac{(p-p_0)^2}{p+p_0} \\
&= \frac{1}{2}\int\frac{(p_0-p)^2}{p_0+p}+\frac{1}{2}\underbrace{\int\frac{p_0-p}{p_0+p}(p_0+p)}_{=0} \\
&= \int\frac{p_0-p}{p_0+p}p_0 \\
&= \int\left(1-\frac{2p}{p+p_0}\right)p_0 \\
&= 1-P\left(\frac{2p}{p+p_0}\right).
\end{aligned}
$$

Thus we have shown that

$$h^2(p_{\hat{F}},p_{F_0}) \leq \|P_n - P\|_{\mathcal{G}}$$

whence the result. □

4.7 Exercises

Exercise 4.7.1 Complete the proof of Theorem 4.1.1 allowing for distribution functions F with jumps.

Exercise 4.7.2 Let

$$d^2(x,y) := \sum_{k=1}^{r}(x_k-y_k)^2 =: \|x-y\|_2^2,\ x\in\mathbb{R}^r,\ y\in\mathbb{R}^r$$

and let

$$S := \{x:\ \|x\|_2 \leq 1\}$$

be the r-dimensional ball. Show that

$$H(\delta,S,d) \leq \text{const. } r\log(1/\delta),\ \forall\ \delta>0$$

(where the "const." does not depend on r or δ).

Exercise 4.7.3 Extend the result of Example 4.3.1 to L-Lipschitz functions with L possibly unequal to 1.

Chapter 5

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 related: it splits the sample of size $2n$ randomly in two subsamples of size n.

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

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

and likewise

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

and

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

5.1 Intermezzo: some facts about (conditional) expectations

5.1.1 Suprema in-/outside the expectation

Let $Z \in \mathcal{Z}$ be random variable and for a set T, let be defined a function $f_t : \mathcal{Z} \to \mathbb{R}$ for all $t \in T$. Then

$$\mathbb{E} \sup_{t \in T} |f_t(Z)| \geq \sup_{t \in T} \mathbb{E}|f_t(Z)| \geq \sup_{t \in T} |\mathbb{E} f_t(Z)|.$$

5.1.2 Iterated expectations

Let X and Y be random variables, say in $\mathbb{R}$. Then

$$EE(Y|X) = EY$$

and for functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$,

$$E[f(X)g(X,Y)|X] = f(X)E[g(X,Y)|X].$$

Moreover, for sets $A \subset \mathbb{R}$ and B of $\mathbb{R}^2$

$$P(Y \in A|X) = E[1_A(Y)|X],$$

and

$$\begin{aligned} P(X \in A,\ (X,Y) \in B|X) &= E[1_A(X)1_B(X,Y)|X] \\ &= 1_A(X)E[1_B(X,Y)|X] \\ &= 1_A(X)P((X,Y) \in B|X). \end{aligned}$$

5.2 Symmetrization with means

Lemma 5.2.1 *We have*

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

Proof. Obviously,

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

and

$$\mathbb{E}(P_n' g|\mathbf{X}) = Pg.$$

So

$$(P_n - P)g = \mathbb{E}[(P_n - P_n')g|\mathbf{X}].$$

Hence

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

Now use the observation of Subsection 5.1.1. We see that

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

So we now showed that

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

Finally, use iterated expectations (Subsection 5.1.2):

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

□

Definition 5.2.1 *A* Rademacher *sequence* $\{\sigma_i\}_{i=1}^n$ *is a sequence of independent random variables* σ_i*, with*

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

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 5.2.2 *We have*

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

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

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

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

□

5.3 Symmetrization with probabilities

Lemma 5.3.1 *Let $\delta > 0$. Suppose that for all $g \in \mathcal{G}$,*

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

Then

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

Proof. Let $\mathbb{P}_{\mathbf{X}}$ 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 $\mathbf{X}$,

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

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

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

Thus,

$$\begin{aligned}
& \mathbb{P}\left(|(P_n - P)g_*| > \delta \text{ and } |(P_n' - P)g_*| \leq \frac{\delta}{2}\right) \\
= \; & \mathbb{E}\mathbb{P}_{\mathbf{X}}\left(|(P_n - P)g_*| > \delta \text{ and } |(P_n' - P)g_*| \leq \frac{\delta}{2}\right) \\
= \; & \mathbb{E}\mathbb{P}_{\mathbf{X}}\left(|(P_n' - P)g_*| \leq \frac{\delta}{2}\right) 1\!\!1\{|(P_n - P)g_*| > \delta\} \\
\geq \; & \frac{1}{2}\mathbb{E}1\!\!1\{(|(P_n - P)g_*| > \delta\} \\
= \; & \frac{1}{2}\mathbb{P}\left(|(P_n - P)g_*| > \delta\right).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > \delta\right) &\leq \mathbb{P}\left(|(P_n - P)g_*| > \delta\right) \\
&\leq 2\mathbb{P}\left(|(P_n - P)g_*| > \delta \text{ and } |(P_n' - P)g_*| \leq \frac{\delta}{2}\right) \\
&\leq 2\mathbf{P}\left(|(P_n - P_n')g_*| > \frac{\delta}{2}\right).
\end{aligned}$$

□

Corollary 5.3.1 *Let $\delta > 0$. Suppose that for all $g \in \mathcal{G}$,*

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

Then

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

5.4 Exercises

Exercise 5.4.1 Use the same arguments as in the proof of Lemma 3.2.2 to show that

$$\mathbb{E}\left[\max_{1\leq j\leq N}|P_n^\sigma g_j|\middle|\mathbf{X}\right]\leq R_n\sqrt{\frac{2\log(2N)}{n}},$$

with

$$R_n^2:=\max_{1\leq j\leq N}\|g_j\|_n^2,$$

where for a function $g:\ \mathcal{X}\to\mathbb{R}$, $\|\cdot\|_n$ is a short hand (abuse of) notation for the empirical norm

$$\|g\|_n:=\sqrt{P_n g^2},\ g:\ \mathcal{X}\to\mathbb{R}.$$

Chapter 6

ULLNs based on symmetrization

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 minimal 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 $\|P_n - P\|_{\mathcal{G}} \xrightarrow{\mathbf{P}} 0$. If this is the case, we call $\mathcal{G}$ a Glivenko Cantelli (GC) class.

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

6.1 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

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

Lemma 6.1.1 *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 we have

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > K\sqrt{\frac{2(\log N + t)}{n}}\right) \leq 2\exp\left[-t\right] \ \forall\ t > 0.$$

and for all $t > 0$ such that $\log N + t \geq 1$

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > 4K\sqrt{\frac{2(\log N + t)}{n}}\right) \leq 8\exp\left[-t\right].$$

Proof.

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

$$\mathbb{P}\left(|P_n^\sigma(g)| > K\sqrt{2t}\right) \leq 2\exp\left[-t\right] \ \forall\ t > 0.$$

• Use the elementary observation to conclude that

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > K\sqrt{\frac{2(\log N + t)}{n}}\right) \quad \leq \quad 2\exp\left[-t\right] \ \forall\ t > 0.$$

• By Chebyshev's inequality, for each $g \in \mathcal{G}$ and for $K^2/(n\delta^2) \leq 1/2$

$$\begin{aligned}\mathbb{P}\left(|(P_n - P)g| > \delta\right) &\leq \frac{\mathrm{var}(g(X))}{n\delta^2} \\ &\leq \frac{K^2}{n\delta^2} \leq \frac{1}{2}.\end{aligned}$$

• Hence, by symmetrization with probabilities when $\log N + t \geq 1$

$$\begin{aligned}\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > 4K\sqrt{\frac{2(\log N + t)}{n}}\right) &\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > K\sqrt{\frac{2(\log N + t)}{n}}\right) \\ &\leq 8\exp\left[-t\right].\end{aligned}$$

□

We recall some definitions from Chapter 4.

Definition 6.1.1 *The* envelope G *of a collection of functions* $\mathcal{G}$ *is defined by*

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

Definition 6.1.2 *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 minimum 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*[1] 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 minimal δ-covering *of S. The logarithm $H(\cdot, S, d) := \log N(\cdot, S, d)$ of the covering number is called the* entropy *of S.*

Notation. We let $N_q(\cdot, \mathcal{G}, P_n)$ be the covering numbers of $\mathcal{G}$ endowed with the metric corresponding to the empirical norm

$$(P_n|g|^q)^{1/q}, \ g: \ \mathcal{X} \to \mathbb{R},$$

and let $H_q(\cdot, \mathcal{G}, P_n)$ be the entropy for this metric.

Theorem 6.1.1 *Suppose*

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

Assume moreover that

$$\frac{1}{n} H_1(\delta, \mathcal{G}, P_n) \xrightarrow{\mathbf{P}} 0.$$

Then

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

Proof. Let $\delta > 0$. Let $g_1, \ldots, g_N$, with $N = N_1(\delta, \mathcal{G}, P_n)$, be a minimal 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

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

• Let $\mathbb{P}_{\mathbf{X}}$ dente conditional probability given $\mathbf{X}$. By Hoeffding's inequality and the elementary observation, we have

$$\mathbb{P}_{\mathbf{X}} \left(\max_{j=1,\ldots,N} |P_n^\sigma(g_j)| > K \sqrt{\frac{2(\log N + t)}{n}} \right) \leq 2 \exp[-t] \ \forall \ t > 0.$$

[1] We will sometimes require that $\{s_j\}_{j=1}^N \in S$

•

Conclude that

$$\mathbb{P}_{\mathbf{X}}\left(\|P_n^\sigma\|_{\mathcal{G}} > \delta + K\sqrt{\frac{2(\log N + t)}{n}}\right) \leq 2\exp[-t] \ \forall\ t > 0.$$

• But then

$$\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > 2\delta + K\sqrt{\frac{2t}{n}}\right)$$

$$\leq 2\exp[-t] + \mathbb{P}\left(K\sqrt{\frac{2H_1(\delta, \mathcal{G}, P_n)}{n}} > \delta\right) \ \forall\ t > 0.$$

• Now, use the symmetrization with probabilities to conclude that

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > 8\delta + 4K\sqrt{\frac{2t}{n}}\right)$$

$$\leq 8\exp[-t] + 4\mathbb{P}\left(K\sqrt{\frac{2H_1(\delta, \mathcal{G}, P_n)}{n}} > \delta\right) \ \forall\ t \geq 1.$$

Since δ is arbitrary, this concludes the proof.

□

Example 6.1.1 *Consider the case $\mathcal{X} = \mathbb{R}$ and*

$$\mathcal{G} := \{g \uparrow,\ 0 \leq g \leq 1\}.$$

Let $N_\infty(\cdot, \mathcal{G}, P_n)$ be the covering number of $\mathcal{G}$ for the (pseudo-)metric induced by the (pseudo-)norm

$$\max_{1 \leq i \leq n} |g(X_i)|,\ g : \mathcal{X} \to \mathbb{R}.$$

Then $N_1(\cdot, \mathcal{G}, P_n) \leq N_\infty(\cdot, \mathcal{G}, P_n)$. Let $\delta > 0$. We approximate $g \in \mathcal{G}$ by

$$\tilde{g}(x) := \lceil g(x)/\delta \rceil \delta,\ x \in \mathbb{R},$$

where $\lceil a \rceil := \min\{b \in \mathbb{N} :\ b \geq a\}$, $a \geq 0$. Now we count how many such functions $\tilde{g}$ we obtain as g varies. The function $\tilde{g}$ has at most $m \leq 1 + 1/\delta$ jumps. The jumps are counted at $X_1, \ldots, X_n$. So we can represent $\tilde{g}$ by a sequence of $n-1$ zeroes and m ones. The number of such sequences is

$$\binom{m+n-1}{m}.$$

We conclude that

$$N_\infty(\delta, \mathcal{G}, P_n) \leq \binom{m+n-1}{m},$$

and hence

$$\begin{aligned} H_\infty(\delta, \mathcal{G}, P_n) &:= \log N_\infty(\delta, \mathcal{G}, P_n) \\ &\leq \log \binom{m+n-1}{m} \\ &\leq m \log(m+n-1) \\ &\leq (1+1/\delta)\log(n+1/\delta). \end{aligned}$$

By Theorem 6.1.1 we conclude that the class of monotone functions $\mathcal{G}$ is GC.

We next replace the assumption that $\mathcal{G}$ is uniformly bounded by the weaker assumption that its envelope G is P-integrable.

Theorem 6.1.2 *Suppose*

$$G \in L_1(P),$$

and

$$\frac{1}{n} H_1(\delta, \mathcal{G}, P_n) \xrightarrow{\mathbf{P}} 0.$$

Then

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

Proof. It holds that for all $g \in \mathcal{G}$ and any $K > 0$

$$|(P_n - P)g| \leq \underbrace{|(P_n - P)g\mathrm{l}\{G \leq K\}}_{:=(i)} + \underbrace{|(P_n - P)g\mathrm{l}\{G > K\}}_{:=(ii)}.$$

(i) Let $\mathcal{G}_K := \{g\mathrm{l}\{G \leq K\}\}$. This class is uniformly bounded by K and since $H_1(\cdot, \mathcal{G}_K, P_n) \leq H_1(\cdot, \mathcal{G}, P_n)$ we conclude from Theorem 6.1.1 that

$$\|P_n - P\|_{\mathcal{G}_K} \xrightarrow{\mathbf{P}} 0.$$

(ii) We have

$$\begin{aligned} \sup_{g \in \mathcal{G}} |(P_n - P)g\mathrm{l}\{G > K\} &\leq (P_n + P)G\mathrm{l}\{G > K\} \\ &= \underbrace{(P_n - P)G\mathrm{l}\{G > K\}}_{\xrightarrow{\mathbf{P}} 0 \text{ as } n \to \infty \ \forall\ K} + \underbrace{2PG\mathrm{l}\{G > K\}}_{\to 0 \text{ as } K \to \infty}. \end{aligned}$$

□

6.2 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 6.2.1 *We write*

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

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 $\mathcal{X}=\mathbb{R}$ and

$$\mathcal{D}=\{\mathrm{l}_{(-\infty,t]}:\ t\in\mathbb{R}\}.$$

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

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

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 6.2.1 *(***Vapnik and Chervonenkis (1971)***) We have*

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

if and only if

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

Proof of the only if part. This follows from applying Theorem 6.1.1 to

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

To see this, note first that a class of indicator functions is uniformly bounded by 1. This is also true for the centred version, i.e. we can take $K=1$ in Theorem 6.1.1. Moreover, writing $N_\infty(\cdot,\mathcal{G},P_n)$ for the covering number of $\mathcal{G}$ for the (pseudo-)metric induced by the (pseudo-)norm

$$\max_{1\leq i\leq n}|g(X_i)|,\ g:\mathcal{X}\to\mathbb{R},$$

we see that

$$N_1(\cdot,\mathcal{G},P_n)\leq N_\infty(\cdot,\mathcal{G},P_n).$$

But for $0 < \delta < 1$,

$$N_\infty(\delta, \{\mathrm{l}_D : \ D \in \mathcal{D}\}, P_n) = \Delta^{\mathcal{D}}(X_1, \ldots, X_n).$$

So indeed, if $\frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \xrightarrow{\mathbf{P}} 0$, then also

$$\frac{1}{n} H_1(\delta, \{\mathrm{l}_D : \ D \in \mathcal{D}\}, P_n) \leq \frac{1}{n} \log \Delta^{\mathcal{D}}(X_1, \ldots, X_n) \xrightarrow{\mathbf{P}} 0.$$

□

6.3 Vapnik Chervonenkis classes

Definition 6.3.1 *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 $\boxed{\text{VC} \Rightarrow \text{GC}}$.

Examples.

a) $\mathcal{X} = \mathbb{R}$, $\mathcal{D} = \{\mathrm{l}_{(-\infty,t]} : \ t \in \mathbb{R}\}$. Since $m^{\mathcal{D}}(n) \leq n + 1$, $\mathcal{D}$ is VC.

b) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{D} = \{\mathrm{l}_{(-\infty,t]} : \ t \in \mathbb{R}^d\}$. Since $m^{\mathcal{D}}(n) \leq (n+1)^d$, $\mathcal{D}$ is VC.

c) $\mathcal{X} = \mathbb{R}^d$, $\mathcal{D} = \{\{x : \ \theta^T x > t\}, \begin{pmatrix} \theta \\ t \end{pmatrix} \in \mathbb{R}^{d+1}\}$. 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 6.3.1 *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 half-spaces,

- all ellipsoids,

- all half-ellipsoids,

- in $\mathbb{R}$, the class $\left\{\{x:\ \theta_1 x + \ldots + \theta_r x^r \leq t\}:\ \begin{pmatrix}\theta\\ t\end{pmatrix} \in \mathbb{R}^{r+1}\right\}$.

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 6.3.2 *The VC dimension of $\mathcal{D}$ is*

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

The following lemma is beautiful, but to avoid digressions, we will not provide a proof.

Lemma 6.3.2 *(***Sauer-Shelah Lemma***) We have that $\mathcal{D}$ is VC if and only if $V(\mathcal{D}) < \infty$. In fact, we have for $V = V(\mathcal{D})$, $m^{\mathcal{D}}(n) \leq \sum_{k=0}^{V} \binom{n}{k}$.*

□

6.4 VC graph classes of functions

Definition 6.4.1 *The* subgraph *of a function $g : \mathcal{X} \to \mathbb{R}$ is*

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

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

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

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

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

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

Definition 6.4.2 *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*

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

Lemma 6.4.1 *For all $\delta > 0$*
(i) $N(\delta, S, d) \leq D(\delta, S, d)$,
(ii) $D(\delta, S, d) \leq N(\delta/2, S, d)$.

Proof. Let $s_1, \ldots, s_N$ in S with $d(s_k, s_j) > \delta, \ \forall \ k \neq j$. Let it be a maximal such set.
(i) Then for any $s \in S$ there is a $j \in \{1, \ldots, N\}$ such that $d(s, s_j) \leq \delta$ (since otherwise we could add s to the packing).
(ii) Since $\{s_1, \ldots, s_N\} \subset S$

$$N(\delta/2, S, d) \geq N(\delta/2, \{s_1, \ldots, s_N\}, d).$$

But

$$N(\delta/2, \{s_1, \ldots, s_N\}, d) = N.$$

□

Theorem 6.4.1 *Let Q be any probability measure on $\mathcal{X}$ and let $N_1(\cdot, \mathcal{G}, Q)$ be the covering number of $\mathcal{G}$ endowed with the metric corresponding to the $L_1(Q)$ norm. For a VC class $\mathcal{G}$ with VC dimension V, we have for a constant A depending only on V,*

$$N_1(\delta QG, \mathcal{G}, Q) \leq \max(A\delta^{-2V}, \mathrm{e}^{\delta/4}), \ \forall \ \delta > 0$$

.

Proof. Without loss of generality, assume $Q(G) = 1$. Choose $S \in \mathcal{X}$ with distribution $dQ_S = GdQ$. 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

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

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 $N \geq \exp[\delta/4]$,

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

So

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

□

Corollary 6.4.1 *By Theorem 6.1.2 and Theorem 6.4.1 we arrive at the following important conclusion:*

$$\boxed{\mathcal{G}\ VC\ \&\ PG < \infty \Rightarrow \mathcal{G}\ GC}$$

6.5 Exercises

Exercise 6.5.1 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

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

Exercise 6.5.2 Let $\mathcal{G}_K := \{gl\{G \leq K\} : \ g \in \mathcal{G}\}$. Show that

$$H_q(\cdot, \mathcal{G}_K, P_n) \leq H_q(\cdot, \mathcal{G}, P_n).$$

Exercise 6.5.3 Present a proof of Lemma 6.3.1.

Exercise 6.5.4 Are the following classes of sets (functions) VC? Why (not)?

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

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

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 $\mathbb{R}^2$ (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 $(a_1, a_2) \in \mathbb{R}^2$, some $r > 0$, and some $0 \leq \theta_1 \leq \theta_2 \leq 2\pi$).

5) The class of all star-shaped sets in $\mathbb{R}^2$ (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 6.5.5 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.

Chapter 7

M-estimators

7.1 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 with metric d) and let for $\theta \in \Theta$,

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

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. $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$, 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. a 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:

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

7.2 Consistency

Define for $\theta \in \Theta$,

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

and

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

Definition 7.2.1 *We say that θ_0 is well-separated if for all $\eta > 0$*

$$\inf\{R(\theta): \ d(\theta,\theta_0) > \eta\} > R(\theta_0).$$

We first present an easy proposition with a too stringent condition ($\bullet$).

Proposition 7.2.1 *Suppose that*

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

i.e., that $\{\gamma_\theta: \ \theta \in \Theta\}$ is a GC class. Then $R(\hat\theta_n) \xrightarrow{\mathbf{P}} R(\theta_0)$. If moreover θ_0 is well-separated, also $\hat\theta_n \xrightarrow{\mathbf{P}} \theta_0$.

Proof. We have

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

So $R(\hat\theta_n) \xrightarrow{\mathbf{P}} R(\theta_0)$ and hence, if θ_0 is well-separated, $\hat\theta_n \xrightarrow{\mathbf{P}} \theta_0$.

□

The assumption ($\bullet$) is rather severe. It is close to requiring compactness of Θ.

Lemma 7.2.1 *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

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

Proof. This is Lemma 4.5.1. □

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

Lemma 7.2.2 *Suppose that Θ is a convex subset of a normed vector space with norm $\|\cdot\|$ and that $\theta \mapsto \gamma_\theta$, $\theta \in \Theta$ is convex. Suppose that for some $\epsilon > 0$,*
- $\Theta_\epsilon := \{\theta \in \Theta : \ \|\theta - \theta_0\| \leq \epsilon\}$ is compact,
- $P(G_\epsilon) < \infty$ where $G_\epsilon := \sup_{\theta \in \Theta_\epsilon} |\gamma_\theta|$.
Then $\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta_0$.

Proof. By Lemma 4.5.1

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

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 7.2.1 that $R(\tilde{\theta}_n) \xrightarrow{\mathbf{P}} R(\theta_0)$. The convexity and the uniqueness of θ_0 implies now that

$$\|\tilde{\theta}_n - \theta_0\| \xrightarrow{\mathbf{P}} 0.$$

But then also

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

□

Example 7.2.1 *Let $X \in \mathbb{R}$, $\Theta = \mathbb{R}$ and for some $q \geq 1$*

$$\gamma_\theta(x) := |x - \theta|^q, \ x \in \mathbb{R}.$$

Assume $E|X - \theta_0|^q < \infty$. Then

$$G_\epsilon := \sup_{\|\theta - \theta_0\| \leq \epsilon} |\gamma_\theta| \leq 2^{q-1}(|X - \theta_0|^q + \epsilon^q) \in L_1(P).$$

Hence (assuming uniqueness of θ_0)

$$\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta_0.$$

We may extend this to the situation where there are co-variables: replace X by (X, Y) where $X \in \mathbb{R}^r$ is a row-vector containing the co-variables and $Y \in \mathbb{R}$ is the response variable. The loss function is

$$\gamma_\theta(x, y) = |y - x\theta|^q.$$

Then $E|Y - X\theta_0|^q < \infty$ together with uniqueness of θ_0 yields consistency.

Example 7.2.2 *Replace X by (X, Y) where $X \in [0, 1]$ is a co-variable and*

$$Y = \theta_0(X) + \xi,$$

with the noise $\xi \sim \mathcal{N}(0, \sigma^2)$. We use least squares loss:

$$\gamma_\theta(x, y) := (y - \theta(x))^2, \ x \in [0, 1], \ y \in \mathbb{R}.$$

We assume that $\theta_0 \in \Theta$ where, for a given $m \in \mathbb{N}$, Θ is the "Sobolev" space

$$\theta := \left\{\theta : \ [0, 1] \to \mathbb{R}, \ \int_0^1 |\theta^{(m)}(x)|^2 dx \leq 1\right\}.$$

Here $\theta^{(m)}$ denotes the m-th derivative of the function θ. We endow the space Θ with the sup-norm

$$\|\theta\| := \|\theta\|_\infty = \sup_{x \in [0,1]} |\theta(x)|.$$

Then one can show (we omit the details here) that for any ϵ

$$\Theta_\epsilon := \left\{\theta : \ [0, 1] \to \mathbb{R}, \ \|\theta - \theta_0\| \leq \epsilon, \ \int_0^1 |\theta^{(m)}(x)|^2 dx \leq 1\right\}$$

is compact. It follows that under the assumption that θ_0 is unique (which is true for example if the distribution of X is absolutely continuous with a density that stays away from zero), the (non-parametric) least squares estimator $\hat{\theta}_n$ is consistent in sup-norm.

7.3 Exercises

Exercise 7.3.1 Let $Y \in \{0, 1\}$ be a binary response variable and $X \in \mathbb{R}$ be a co-variable. Assume the logistic regression model

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

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

Exercise 7.3.2 Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued random variables with density $p_0 = dP/d\mu$ on [0,1]. Here, μ is Lebesgue measure on $[0, 1]$. Suppose it is given that $p_0 \in \mathcal{P}$, with $\mathcal{P}$ the set of all decreasing densities bounded from above by 2 and from below by 1/2. Let $\hat{p}_n$ be the MLE. Can you show consistency of $\hat{p}_n$? For what metric? Hint: use the result of Section 2.5 and Example 6.1.1.

Chapter 8

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 this Chapter we will therefore only briefly indicate the results, and not present any proofs. These sections only reveal a glimpse 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.*

8.1 Real-valued random variables

Let $\mathcal{X} = \mathbb{R}$.

Theorem 8.1.1 *(***Central limit theorem in $\mathbb{R}$***)* *Suppose $EX = \mu$, and* $\mathrm{var}(X) = \sigma^2$ *exist. Then*

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

where Φ is the standard normal distribution function.

□

Notation.

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

or

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

8.2 $\mathbb{R}^r$-valued random variables

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

Theorem 8.2.1 *(***Central limit theorem in $\mathbb{R}^r$***) We have*

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

i.e.

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

□.

8.3 Donsker's theorem

Let $\mathcal{X} = \mathbb{R}$. Recall the definition of the distribution function F and the empirical distribution function $\hat{F}_n$:

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

$$\hat{F}_n(t) = \frac{1}{n}\#\{X_i \leq t, \ 1 \leq i \leq n\}, \ t \in \mathbb{R}.$$

Define

$$W_n(t) := \sqrt{n}(\hat{F}_n(t) - F(t)), \ t \in \mathbb{R}.$$

By the central limit theorem in $\mathbb{R}$ (Section 8.1), for all t

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

Also, by the central limit theorem in $\mathbb{R}^2$ (Section 8.2), for all $s < t$,

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

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** $W_n = \{W_n(t) : \ t \in \mathbb{R}\}$. The process W_n is called the (classical) empirical process.

Definition 8.3.1 *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 next consider the process W_F defined as

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

Thus, $W_F = B \circ F$.

Theorem 8.3.1 *(***Donsker's theorem***) Consider W_n and W_F as elements of the space $\mathcal{K}$ of bounded functions on $\mathcal{R}$. We have*

$$W_n \xrightarrow{\mathcal{L}} W_F,$$

that is,

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

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,

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

This is called **asymptotic continuity**.

8.4 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 $g : \mathcal{X} \to \mathbb{R}$. The (theoretical) mean of a function g is

$$Pg := Eg(X),$$

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 8.4.1 *The* empirical process *indexed by $\mathcal{G}$ is*

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

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

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

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

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

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

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

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

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

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 8.4.2 *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 8.4.3 *Consider ν_n and ν as bounded functions on $\mathcal{G}$. We call $\mathcal{G}$ a P-**Donsker class** if*

$$\nu_n \xrightarrow{\mathcal{L}} \nu,$$

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

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

Definition 8.4.4 *The process ν_n on $\mathcal{G}$ is called* asymptotically continuous *at $g_0 \in \mathcal{G}$ if for all (possibly random) sequences $\{g_n\} \subset \mathcal{G}$ with $\sigma(g_n - g_0) \xrightarrow{\mathbf{P}} 0$, we have*

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

If this is true for all $g_0 \in \mathcal{G}$ we call ν_n on $\mathcal{G}$ asymptotically continuous.

We will use the notation

$$\|g\|^2 := Pg^2, g \in L_2(P),$$

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

Remark. Note that $\sigma(g) \leq \|g\|$.

Definition 8.4.5 *The class $\mathcal{G}$ is called* totally bounded *for the metric induced by $\|\cdot\|$, if its entropy $H_2(\cdot, \mathcal{G}, P)$ is finite.*

Theorem 8.4.1 *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.*

□

Chapter 9

Chaining and asymptotic continuity

9.1 Chaining

We consider in this section the symmetrized empirical process and work conditionally on $\mathbf{X} = (X_1, \ldots, X_n)$. We let $\mathbb{P}_{\mathbf{X}}$ be the conditional distribution given $\mathbf{X}$. We describe the chaining technique in this context

Define the empirical norm

$$\|g\|_n := \sqrt{P_n g^2}$$

and the empirical radius

$$R_n := \sup_{g \in \mathcal{G}} \|g\|_n.$$

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 $s = 0, 1, 2, \ldots$, $\{g_j^s\}_{j=1}^{N_s} \subset \mathcal{G}$ be a minimal $2^{-s}R_n$-covering set of $(\mathcal{G}, \|\cdot\|_n)$. So $N_s = N_2(2^{-s}R_n, \mathcal{G}, P_n)$, and for each θ, there exists a $g_\theta^s \in \{g_1^s, \ldots, g_{N_s}^s\}$ such that $\|g_\theta - g_\theta^s\|_n \leq 2^{-s}R_n$. 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 $\|g_\theta\|_n \leq R_n$. We let $H_s := \log N_s$ for all s.

Let $S \in \mathbb{N}$ to be fixed later and let for $g_j^{S+1} \in \{g_1^{S+1}, \ldots, g_{N_{S+1}}^{S+1}\}$,

$$g_{j,S}^S := \arg\min \left\{ \|g_j^{S+1} - g_k^S\|_n : \ g_k^S \in \{g_1^S, \ldots, g_{N_S}^S\} \right\}$$

and for $s \in \{0, \ldots, S-1\}$,

$$g_{j,S}^s := \arg\min \left\{ \|g_j^{s+1} - g_k^s\|_n : \ g_k^s \in \{g_1^s, \ldots, g_{N_s}^s\} \right\}.$$

Then we can write for $g_\theta^{S+1} = g_j^{S+1}$

$$g_\theta = \sum_{s=0}^{S} (g_{j,S}^{s+1} - g_{j,S}^s) + (g_\theta - g_\theta^{S+1}).$$

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

9.2 Increments of the symmetrized process

We define

$$J_n := \sum_{s=0}^S 2^{-s} R_n \sqrt{2H_{s+1}}.$$

Remark We may use the bound

$$J_n \le 2 \int_{2^{-(S+2)} R_n}^{R_n} \sqrt{2H_2(u, \mathcal{G}, P_n)} du.$$

The right-hand side is called (up to constants) Dudley's entropy integral.

Theorem 9.2.1 *We have*

$$\mathbb{P}_{\mathbf{X}} \left(\max_j |\sum_{s=0}^S P_n^\sigma (g_{j,S}^{s+1} - g_{j,S}^s)| \ge \frac{J_n}{\sqrt{n}} + 4R_n \sqrt{\frac{1+t}{n}} \right) \le 2 \exp[-t] \ \forall \ t > 0.$$

Proof. By Hoeffding's inequality (Theorem 3.2.1), for all j and s

$$\mathbb{P}_{\mathbf{X}} \left(|P_n^\sigma (g_{j,S}^{s+1} - g_{j,S}^s)| \ge 2^{-s} R_n \sqrt{\frac{2t}{n}} \right) \le 2 \exp[-t] \ \forall \ t > 0.$$

Therefore (by the union bound), for all s

$$\mathbb{P}_{\mathbf{X}} \left(\max_j |P_n^\sigma (g_{j,S}^{s+1} - g_{j,S}^s)| \ge 2^{-s} R_n \sqrt{\frac{2(H_{s+1} + t)}{n}} \right) \le 2 \exp[-t] \ \forall \ t > 0.$$

Fix t and let for $s = 0, \ldots, S$

$$\alpha_s := 2^{-s} R_n \left(\sqrt{2H_{s+1}} + \sqrt{2(1+s)(1+t)} \right).$$

Then

$$\sum_{s=0}^S \alpha_s = J_n + \sum_{s=0}^S 2^{-s} R_n \sqrt{2(1+s)(1+t)} \le J_n + 4R_n \sqrt{1+t}.$$

Therefore

$$
\begin{aligned}
& \mathbb{P}_{\mathbf{X}}\left(\max_j |\sum_{s=0}^{S} P_n^\sigma(g_{j,S}^{s+1} - g_{j,S}^s)| \geq \frac{J_n}{\sqrt{n}} + 4R_n\sqrt{\frac{1+t}{n}}\right) \\
\leq \quad & \sum_{s=0}^{S} \mathbb{P}_{\mathbf{X}}\left(\max_j |P_n^\sigma(g_{j,S}^{s+1} - g_{j,S}^s)| \geq \alpha_s\right) \\
\leq \quad & 2\sum_{s=0}^{S} \exp[-(1+s)(1+t)] \\
\leq \quad & 2\exp[-t]
\end{aligned}
$$

□

9.3 De-symmetrizing

Recall the (theoretical) $L_2(P)$-norm

$$\|g\| = \sqrt{Pg^2},\ g \in L_2(P).$$

We let

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

be the diameter of $\mathcal{G}$.

For any probability measure Q, we let $H_2(\cdot, \mathcal{G}, Q)$ be the entropy of $\mathcal{G}$ endowed with the metric induced by the $L_2(Q)$-norm $\|\cdot\|_Q$.

Condition 9.3.1 *For all probability measures Q it holds that*

$$H_2(u\|G\|_Q, \mathcal{G}, Q) \leq \mathcal{H}(u),\ \forall\ u > 0$$

for some decreasing function $\mathcal{H}(\cdot)$ satisfying

$$\mathcal{J}(\mathcal{R}) := \int_0^1 \sqrt{\mathcal{H}(u)}du < \infty.$$

We then define

$$\mathcal{J}(\rho) := 2\int_0^\rho \sqrt{2\mathcal{H}(u)}du,\ \rho > 0.$$

Theorem 9.3.1 *Suppose Condition 9.3.1 is met and $G \in L_2(P)$. Then*

$$\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > \frac{4\|G\|\mathcal{J}(4R\|G\|)}{\sqrt{n}} + 32R\sqrt{\frac{1+t}{n}}\right)$$

$$\leq 8\exp[-t] + 4\mathbb{P}\left(\sup_{g \in \mathcal{G}\cup\{G\}} |(P_n - P)g^2| > R^2\right)\ \forall\ t > 0$$

Proof. Take S as the smallest value in $\mathbb{N}$ such that $2^{-S} \leq 1/\sqrt{n}$. Then $\|g_\theta - g_\theta^{S+1}\|_n \leq 2^{-(S+1)}R_n \leq R_n/(2\sqrt{n})$. On the set where $R_n \leq 2R$ and $\|G\|_n \leq 2\|G\|$ we have

$$2\int_{2^{-(S+2)}R_n}^{R_n} \sqrt{2H_2(u,\mathcal{G},P_n)}du \leq \|G\|\mathcal{J}(4R\|G\|)$$

so for $\mathbf{X}$ in this set

$$\mathbb{P}_{\mathbf{X}}\left(\|P_n^\sigma\|_{\mathcal{G}} \geq \frac{\|G\|\mathcal{J}(4R\|G\|)}{\sqrt{n}} + 8R\sqrt{\frac{1+t}{n}}\right) \leq 2\exp[-t] \ \forall\ t > 0.$$

We can then de-symmetrize

$$\begin{aligned}
&\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}} > \frac{4\|G\|\mathcal{J}(4R\|G\|)}{\sqrt{n}} + 32R\sqrt{\frac{1+t}{n}}\right) \\
&\leq 4\mathbb{P}\left(\|P_n^\sigma\|_{\mathcal{G}} > \frac{\|G\|\mathcal{J}(4R\|G\|)}{\sqrt{n}} + 8R\sqrt{\frac{1+t}{n}}\right). \\
&\leq 8\exp[-t] + 4\mathbb{P}\left(\sup_{g\in\mathcal{G}\cup\{G\}} |(P_n - P)g^2| > R^2\right).
\end{aligned}$$

□

9.4 Asymptotic continuity of the empirical process

Theorem 9.4.1 *Assume Condition 9.3.1 and that $\mathcal{G}$ has envelope G, with $P(G^2) < \infty$. Then ν_n is asymptotically continuous.*

Proof. Define for $\delta > 0$ and $g_0 \in \mathcal{G}$

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

By Theorem 9.3.1

$$\begin{aligned}
&\mathbb{P}\left(\|P_n - P\|_{\mathcal{G}(\delta)} > \frac{8\|G\|\mathcal{J}(8\delta\|G\|)}{\sqrt{n}} + 32\delta\sqrt{\frac{1+t}{n}}\right) \\
&\leq 8\exp[-t] + 4\mathbb{P}\left(\sup_{g\in\mathcal{G}(\delta)\cup\{2G+g_0\}} |(P_n - P)(g-g_0)^2| > \delta^2\right) \ \forall\ t > 0.
\end{aligned}$$

Take $\epsilon > 0$ arbitrary and $t = \log(8/\epsilon)$. Since $\mathcal{J}(8\delta\|G\|) \downarrow 0$ as $\delta \downarrow 0$, we see that there is a $\delta > 0$ such that

$$\mathbb{P}\left(\sqrt{n}\|P_n - P\|_{\mathcal{G}(\delta)} > \epsilon\right) \leq \epsilon + 4\mathbb{P}\left(\sup_{g\in\mathcal{G}(\delta)\cup\{2G+g_0\}} |(P_n - P)(g-g_0)^2| > \delta^2\right).$$

By the ULLN for $\{(g-g_0)^2 :\ g \in \mathcal{G}(\delta) \cup \{2G+g_0\}\}$,

$$\mathbb{P}\left(\sup_{g\in\mathcal{G}(\delta)\cup\{2G+g_0\}} |(P_n - P)(g-g_0)^2| > \delta^2\right) \to 0.$$

□

9.5 Application to VC graph classes

Theorem 9.5.1 *Suppose that $\mathcal{G}$ is a VC class with envelope $G \in L_2(P)$. Then $\{\nu_n(g) : \ g \in \mathcal{G}\}$ is asymptotically continuous.*

Proof. We recall Theorem 6.2.1:

$$N_1(\delta QG, \mathcal{G}, Q) \leq \max(A\delta^{-2V}, \mathrm{e}^{\delta/4}), \ \forall \ \delta > 0.$$

Now note that for any two functions g and $\tilde{g}$ in $\mathcal{G}$,

$$\int (g - \tilde{g})^2 dQ \leq \int |g - \tilde{g}| d\bar{Q},$$

where $d\bar{Q} := 2GdQ$. This gives

$$N_2(\delta \|G\|_Q, \mathcal{G}, Q) \leq N_1(\delta^2 \|G\|_Q^2, \mathcal{G}, \bar{Q}).$$

The measure $\bar{Q}$ is perhaps not a probability measure, but since it is a finite measure this only effects the constants. In other words, Condition 9.3.1 is met. Apply Theorem 9.4.1 to finish the proof. □

Remark. In particular, suppose that a VC class $\mathcal{G}$ with square integrable envelope G is parametrized by θ in some parameter space $\Theta \subset \mathbb{R}^r$, 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

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

Indeed, if $\|g_\theta - g_{\theta_0}\| \stackrel{\mathbf{P}}{\to} 0$ as θ converges to θ_0, the answer is yes. In other words, we need here mean square continuity of the map $\theta \mapsto g_\theta$.

9.6 Exercises

Exercise 9.6.1 Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. $\mathcal{N}(0,1)$ random variables, independent of $X_1, \ldots, X_n$. Define

$$(\epsilon, g)_n := \frac{1}{n} \sum_{i=1}^n \epsilon_i g(X_i), \ g \in \mathcal{G}.$$

Let

$$J(R_n) := \sum_{s=0}^\infty 2^{-s} \sqrt{2H_2(u, \mathcal{G}, P_n) du},$$

where we assume that the sum converges. Show that

$$\mathbb{P}_{\mathbf{X}}\left(\sup_{g \in \mathcal{G}} (\epsilon, g)_n \geq \frac{J(R_n)}{\sqrt{n}} + 4R_n \sqrt{\frac{1+t}{n}} \right) \leq \exp[-t] \ \forall \ t > 0.$$

Chapter 10

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 $\Theta \subset \mathbb{R}^r$. Moreover, we assume that we already showed that $\hat{\theta}_n$ is consistent.

10.1 Asymptotic linearity

Definition 10.1.1 *The (sequence of) estimator(s) $\hat{\theta}_n$ of θ_0 is called* asymptotically linear *if we may write*

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

where

$$\ell = \begin{pmatrix} \ell_1 \\ \vdots \\ \ell_r \end{pmatrix} : \mathcal{X} \to \mathbb{R}^r,$$

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

Definition 10.1.2 *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 $\hat{\theta}_{n,2}$).*

10.2 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 aa, bb and cc.

Condition a. There exists an $\epsilon > 0$ such that $\theta \mapsto \gamma_\theta(x)$ 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 $V \in \mathbb{R}^{r \times r}$ is a positive definite matrix.

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

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

is asymptotically continuous at ψ_{θ_0}. Moreover,

$$\lim_{\theta \to \theta_0} \|\psi_\theta - \psi_{\theta_0}\| = 0.$$

Lemma 10.2.1 *Suppose conditions a,b and c. Then $\hat{\theta}_n$ is asymptotically linear with influence function*

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

so

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

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

$$\begin{aligned} 0 &= P_n\psi_{\hat\theta_n} = P_n(\psi_{\hat\theta_n} - \psi_{\theta_0}) + P_n\psi_{\theta_0} \\ &= (P_n - P)(\psi_{\hat\theta_n} - \psi_{\theta_0}) + P\psi_{\hat\theta_n} + P_n\psi_{\theta_0}. \end{aligned}$$

Now, use condition b and the asymptotic continuity of $\{\psi_\theta : |\theta - \theta_0| \le \epsilon\}$, 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 10.2.1 *(***Huber estimator***) Let $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$. The Huber estimator corresponds to the loss function*

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

with

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

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

a)

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

b) We have

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

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

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

c) Clearly $\psi_\theta : \theta \in \mathbb{R}$ is a VC graph class, with envelope $\Psi \le 2k$. The asymptotic continuity follows from Theorem 9.5.1.

So the Huber estimator $\hat\theta_n$ has influence function

$$\ell(x) = \begin{cases} \frac{-k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 \le -k \\ \frac{x-\theta_0}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } |x - \theta_0| \le k \\ \frac{k}{F(k+\theta_0)-F(-k+\theta_0)} & \text{if } x - \theta_0 \ge k \end{cases}.$$

The asymptotic variance is

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

10.3 Asymptotics for the median

The sample median 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* $\hat{F}_n(\hat{\theta}_n) - \frac{1}{2} = 0$. Let us investigate this closer.

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

$$F(\theta_0) = \frac{1}{2}.$$

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

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

Then we get

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

where $W_n := \sqrt{n}(\hat{F}_n - F)$ 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 (use the VC-property of intervals) that $W_n(\hat{\theta}_n) = W_n(\theta_0) + o_{\mathbf{P}}(1)$. We thus arrive at

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

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

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

and the asymptotic variance is

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

We are now in the position to 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

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

10.4 Conditions aa, bb and cc for asymptotic normality

We are now going to relax the condition of differentiability of γ_θ.

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

$$\lim_{\theta \to \theta_0} \frac{\left\|\gamma_\theta - \gamma_{\theta_0} - (\theta - \theta_0)^T \psi_0\right\|}{|\theta - \theta_0|} = 0.$$

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

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

with $V \in \mathbb{R}^{r \times r}$ a positive definite matrix.

Condition cc. Define

$$g_\theta = \begin{cases} \frac{\gamma_\theta - \gamma_{\theta_0} - (\theta - \theta_0)^T \psi_0}{|\theta - \theta_0|} & \theta \neq \theta_0 \\ 0 & \theta = \theta_0 \end{cases}.$$

Suppose that for some $\epsilon > 0$, the class $\mathcal{G} := \{g_\theta : \ 0 < |\theta - \theta_0| < \epsilon\}$ is asymptotically continuous at 0.

Lemma 10.4.1 *Suppose conditions aa, bb and cc are met. Then $\hat{\theta}_n$ has influence function*

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

and so

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

where $J = P\psi_0\psi_0^T$.

Proof. Since $\{g_\theta : |\theta-\theta_0| \leq \epsilon\}$ is asymptotically continuous we have as $\theta \to \theta_0$

$$\begin{aligned}
& P_n(\gamma_\theta - \gamma_{\theta_0}) \\
= \; & (P_n - P)(\gamma_\theta - \gamma_{\theta_0}) + P(\gamma_\theta - \gamma_{\theta_0}) \\
= \; & (P_n - P)g_\theta|\theta - \theta_0| + (\theta - \theta_0)^T P_n\psi_0 + P(\gamma_\theta - \gamma_{\theta_0}) \\
= \; & o_{\mathbf{P}}(1/\sqrt{n})|\theta - \theta_0| + (\theta - \theta_0)^T P_n\psi_0 + P(\gamma_\theta - \gamma_{\theta_0}) \\
= \; & o_{\mathbf{P}}(1/\sqrt{n})|\theta - \theta_0| + (\theta - \theta_0)^T P_n\psi_0 + \frac{1}{2}(\theta - \theta_0)^T V(\theta - \theta_0) + o(|\theta - \theta_0|^2) \\
= \; & \frac{1}{2}\left|V^{1/2}(\theta - \theta_0) + o(|\theta - \theta_0|) + \mathcal{O}_{\mathbf{P}}(1/\sqrt{n})\right|^2
\end{aligned}$$

Because $P_n(\gamma_{\hat\theta_n} - \gamma_{\theta_0}) \leq 0$ the previous applied with $\theta = \hat\theta_n$ gives $|\hat\theta_n - \theta_0| = \mathcal{O}_{\mathbf{P}}(1/\sqrt{n})$. The previous applied to the sequence $\tilde\theta_n := \theta_0 - V^{-1/2}P_n\psi_0$ gives

$$P_n(\gamma_{\tilde\theta_n} - \gamma_{\theta_0}) = -\frac{1}{2}|V^{-1/2}P_n\psi_0|^2 + o_{\mathbf{P}}(1/n).$$

Because $P_n(\gamma_{\hat\theta_n} - \gamma_{\theta_0}) \leq P_n(\gamma_{\tilde\theta_n} - \gamma_{\theta_0})$ we get

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

or

$$\frac{1}{2}\left|V^{1/2}(\hat\theta_n - \theta_0) + V^{-1/2}P_n\psi_0\right|^2 = o_{\mathbf{P}}(1/n).$$

Thus

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

or

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

□

10.5 Exercises

Exercise 10.5.1 Suppose X has the logistic distribution with location parameter θ_0. 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 10.5.2 Let (X_i, Y_i), $i = 1, \ldots, n, \ldots$ be i.i.d. copies of (X, Y), where $X \in \mathbb{R}^r$ and $Y \in \mathbb{R}$. Suppose that the conditional distribution of Y given $X = x$ has median $m(x) = \beta_0^0 + \beta_1^0 x_1 + \ldots \beta_r^0 x_r$, with

$$\beta^0 = \begin{pmatrix} \beta_0^0 \\ \vdots \\ \beta_r^0 \end{pmatrix} \in \mathbb{R}^{r+1}.$$

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

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

exists and is invertible. Let

$$\hat{\beta}_n = \arg \min_{b \in \mathbb{R}^{r+1}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - b_0 - b_1 X_{i,1} - \ldots - b_r X_{i,r}| \,,$$

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

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

by verifying conditions aa, bb and cc.

Chapter 11

Rates of convergence for LSEs

Probability inequalities for the least squares estimator (LSE) 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, and image restoration. Results for the 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

$$Y_i = g_0(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

with $x_1, \ldots, x_n$ (fixed) covariates in a space $\mathcal{X}$, $\epsilon_1, \ldots, \epsilon_n$ 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

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

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

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

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

$$\|g\|_n^2 := \frac{1}{n} \sum_{i=1}^n g^2(x_i).$$

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

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

Finally, we let for $\delta > 0$

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

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

Lemma 11.0.1 *(***Basic inequality***). It holds that*

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

Proof. This is rewriting the inequality

$$\sum_{i=1}^n (Y_i - \hat{g}_n(x_i))^2 \leq \sum_{i=1}^n (Y_i - g_0(x_i))^2.$$

□

The main idea to arrive at rates of convergence for $\hat{g}_n$ is to invoke the basic inequality. The modulus of continuity of the process $\{(\epsilon, g - g_0)_n : \ g \in \mathcal{G}(\delta)\}$ can be derived from the entropy $H_2(\cdot, \mathcal{G}(\delta), Q_n)$ of $\mathcal{G}(\delta)$, endowed with the metric induced the the norm $\|\cdot\|_n$.

Condition 11.0.1 *For all $\delta > 0$, the entropy integral*

$$J(\delta) := 2 \int_0^\delta \sqrt{2H_2(u, \mathcal{G}(\delta), Q_n)} du$$

exists (i.e. is finite).

11.1 Gaussian errors

To simplify the exposition, will assume that

$$\epsilon_1, \ldots, \epsilon_n \text{ are i.i, d, } \ \mathcal{N}(0,1)-\text{distributed.}$$

Then, as in Theorem 9.2.1

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}(\delta)} (\epsilon, g - g_0)_n \geq \frac{J(\delta)}{\sqrt{n}} + 4\delta(1+t) \right) \leq \exp[-t] \ \forall \ t > 0.$$

11.2 Rates of convergence

Theorem 11.2.1 *Assume Condition 11.0.1 and that $J(\delta)/\delta^2$ is decreasing in δ. Then for all $t > 0$ and for*

$$\sqrt{n}\delta_n^2 \geq 8\left(J(\delta_n) + 4\delta_n \sqrt{1+t} \right)$$

it holds that

$$\mathbb{P}(\|\hat{g} - g_0\|_n > \delta_n) \leq \frac{\mathrm{e}}{\mathrm{e}-1} \exp[-t].$$

Proof. We use the "peeling device"

$$\mathbb{P}(\|\hat{g}-g_0\|_n > \delta_n) \leq \sum_{j=1}^{\infty} \mathbb{P}\Big(\sup_{g \in \mathcal{G}(2^j \delta_n)} 2(\epsilon, g-g_0)_n \geq (2^{j-1}\delta_n)^2 \Big).$$

The function

$$j \mapsto \frac{J(2^j\delta_n) + 42^j \delta \sqrt{1+t+j}}{(2^j\delta)^2}$$

is the sum of two decreasing functions and hence is decreasing. So for all $j \in \mathbb{N}$

$$\frac{J(2^j\delta_n) + 42^j \delta \sqrt{1+t+j}}{(2^j\delta_n)^2} \leq \frac{J(\delta_n) + 4\delta_n\sqrt{1+t}}{\delta_n^2} \leq \frac{\sqrt{n}}{8}$$

so that

$$\frac{1}{2}(2^{j-1}\delta_n)^2 = \frac{1}{8}(2^j\delta_n)^2 \geq \frac{J(2^j\delta_n)}{\sqrt{n}} + 42^j\delta_n\sqrt{\frac{1+t+j}{n}}$$

It follows that

$$\begin{aligned} & \mathbb{P}(\|\hat{g}-g_0\|_n > \delta_n) \\ \leq \; & \sum_{j=1}^{\infty} \mathbb{P}\Big(\sup_{g \in \mathcal{G}(2^j\delta_n)} (\epsilon, g-g_0)_n \geq \frac{J(2^j\delta)}{\sqrt{n}} + 42^j\delta\sqrt{\frac{1+t+j}{n}} \Big) \\ \leq \; & \exp[-(t+j)] = \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t]. \end{aligned}$$

□

11.3 Examples

Example 11.3.1 Linear regression *Let*

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

One may verify

$$H_2(u, \mathcal{G}(\delta), Q_n) \leq r\log(\frac{\delta+4u}{u}), \text{ for all } 0 < u < \delta, \; \delta > 0.$$

So

$$\begin{aligned} J(\delta) = 2\int_0^{\delta} \sqrt{2H_2(u, \mathcal{G}(\delta), Q_n)}du &\leq 2\sqrt{2r}\int_0^{\delta} \log^{1/2}(\frac{\delta+4u}{\delta})du \\ &= 2\sqrt{2r}\delta\int_0^1 \log^{1/2}(1+4v)dv := A_0\sqrt{r}\delta. \end{aligned}$$

So Theorem 11.2.1 can be applied with

$$\delta_n \geq 8\Big(A_0\sqrt{\frac{r}{n}} + 4\sqrt{\frac{1+t}{n}}\Big).$$

It yields that

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 8\left(A_0\sqrt{\frac{r}{n}} + 4\sqrt{\frac{1+t}{n}}\right)\right) \leq \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t] \ \forall\ t > 0.$$

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

Example 11.3.2 Smooth functions *Let*

$$\mathcal{G} = \left\{g : [0,1] \to \mathbb{R},\ \int_0^1 |g^{(m)}(x)|^2 dx \leq 1\right\}.$$

Let $\psi_k(x) = x^{k-1}$, $k = 1, \ldots, m$, $\psi(x) = (\psi_1(x), \ldots, \psi_m(x))^T$ 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

$$H_2(\delta, \mathcal{G}(\delta), Q_n) \leq A\delta^{-\frac{1}{m}}, \text{ for small } \delta > 0,$$

where the constant A depends on λ. Thus

$$J(\delta) \leq A_0^{1/2}\delta^{1-\frac{1}{2m}}$$

for some constant A_0. For

$$\delta_n \geq 16\left(A_0^{\frac{m}{2m+1}} n^{-\frac{m}{2m+1}} + 4\sqrt{\frac{1+t}{n}}\right)$$

we get

$$\sqrt{n}\delta_n^2 \geq 8\left(J(\delta_n) + 4\delta_n\sqrt{1+t}\right).$$

Hence, we find from Theorem 11.2.1 that

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 16\left(A_0^{\frac{m}{2m+1}} n^{-\frac{m}{2m+1}} + 4\sqrt{\frac{1+t}{n}}\right)\right)$$

$$\leq \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t] \ \forall\ t > 0.$$

Example 11.3.3 Functions of bounded variation in $\mathbb{R}$ *Let*

$$\mathcal{G} = \left\{g : \mathbb{R} \to \mathbb{R},\ \int |dg| \leq 1\right\}.$$

Without loss of generality, we may assume that $x_1 \leq \ldots \leq x_n$. The derivative should be understood in the generalized sense:

$$\int |dg| := \sum_{i=2}^n |g(x_i) - g(x_{i-1})|.$$

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

$$\bar{g} := \int g dQ_n.$$

Then it is easy to see that,

$$\max_{i=1,\ldots,n} |g(x_i)| \leq \bar{g} + 1.$$

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

$$H_2(\delta, \mathcal{G}(\delta), Q_n) \leq A\delta^{-1}, \text{ for small } \delta > 0,$$

and therefore, for some constant A_0

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 16\left(A_0^{\frac{1}{3}} n^{-\frac{1}{3}} + 4\sqrt{\frac{1+t}{n}} \right) \right) \leq \frac{\mathrm{e}}{\mathrm{e}-1} \exp[-t] \ \forall \ t > 0.$$

Example 11.3.4 Concave functions *Let*

$$\mathcal{G} = \{g : [0,1] \to \mathbb{R}, \ 0 \leq \dot{g} \leq 1, \ \dot{g} \text{ decreasing}\}.$$

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

$$\left\{ g : [0,1] \to \mathbb{R}, \ \int_0^1 |d\dot{g}| \leq 2 \right\}.$$

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

$$H_\infty\left(\delta, \left\{ g : [0,1] \to [0,1] : \ \int_0^1 |g^{(m)}(x)| dx \leq 1 \right\} \right) \leq A\delta^{-\frac{1}{m}}, \text{ for all } \delta > 0.$$

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

$$g = g_1 + g_2,$$

with $g_1(x) := \theta_1 + \theta_2 x$ and $\|g_2\|_\infty \leq 2$. Assume now that $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$ stays away from 0. Then, we obtain for a constant A_0

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 16\left(A_0^{\frac{2}{5}} n^{-\frac{2}{5}} + 4\sqrt{\frac{1+t}{n}} \right) \right) \leq \frac{\mathrm{e}}{\mathrm{e}-1} \exp[-t] \ \forall \ t > 0.$$

Example 11.3.5 Image restoration

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

$$Y_{kl} = g_0(x_{kl}) + \epsilon_{kl},$$

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

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

be the closed convex hull of $\mathcal{K}$*, where where*

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

Assume that

$$N_2(\delta, \mathcal{K}, Q_n) \leq c\delta^{-w}, \text{ for all } \delta > 0.$$

Then from Ball and Pajor(1990),

$$H_2(\delta, \mathcal{G}, Q_n) \leq A\delta^{-\frac{2w}{2+w}}, \text{ for all } \delta > 0.$$

It follows from Theorem 11.2.1 for a constant A_0

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 16\left(A_0^{\frac{2+w}{4+4W}} n^{-\frac{2+w}{4+4w}} + 4\sqrt{\frac{1+t}{n}}\right)\right)$$

$$\leq \frac{\text{e}}{\text{e}-1} \exp[-t] \ \forall \ t > 0.$$

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

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

The black part of the image is

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

We observe

$$Y_{kl} = g_0(x_{kl}) + \epsilon_{kl},$$

with $x_{kl} = (u_k, v_l)$*,* $x_k = k/m$*,* $x_l = l/m$*,* $k, l \in \{1, \ldots, m\}$*. 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

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

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

$$H_2(\delta, \mathcal{G}, Q_n) \leq A\delta^{-\frac{1}{2}},$$

so that for a constant A_0

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n > 16\left(A_0^{\frac{2}{5}} n^{-\frac{2}{5}} + 4\sqrt{\frac{1+t}{n}}\right)\right)$$
$$\leq \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t] \ \forall\ t > 0.$$

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 $Q_n(D) = \|1_D\|_n^2$, *we find*

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

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

$$H_2(\delta, \mathcal{G}, Q_n) \leq A\delta^{-\frac{r-1}{2}},\ \delta \downarrow 0,$$

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

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

For $r \geq 5$, *the rate of convergence can still be shown to be* $O_{\mathbf{P}}(n^{-\frac{4}{r+3}})$. *One needs to refine Theorem 11.2.1 for the case of a diverging entropy integral.*

11.4 Exercises

Exercise 11.4.1 Let $Y_1, \ldots, Y_n$ be independent, 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 $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(1/\sqrt{n})$, and $|\hat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}(1/\sqrt{n})$, $|\hat{\beta}_n - \beta_0| = O_{\mathbf{P}}(1/\sqrt{n})$, and $|\hat{\gamma}_n - \gamma_0| = O_{\mathbf{P}}(1/n)$. If $(\alpha_0, \beta_0, \gamma_0)$ is not identifiable, show that $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(\sqrt{\log\log n/n})$.

Exercise 11.4.2 Let $x_i = i/n$, $i = 1, \ldots, n$, and let $\mathcal{G}$ consist of the functions

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

Suppose $g_0 \in \mathcal{G}$ 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 $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(1/\sqrt{n})$, and that $|\hat{\alpha}_n - \alpha_0| = O_{\mathbf{P}}(1/\sqrt{n})$, $|\hat{\beta}_n - \beta_0| = O_{\mathbf{P}}(1/\sqrt{n})$ and $|\hat{\gamma}_n - \gamma_0| = O_{\mathbf{P}}(1/n^{1/3})$.

Exercise 11.4.3 If $\mathcal{G}$ is a uniformly bounded class of increasing functions, show that it follows from Theorem 11.2.1 that $\|\hat{g}_n - g_0\|_n = O_{\mathbf{P}}(n^{-1/3}(\log n)^{1/3})$. (By a more tight bound on the entropy one finds the rate $O_{\mathbf{P}}(n^{-1/3})$ as in Example 11.3.3).

Chapter 12

Regularized least squares

We revisit the regression problem of the previous chapter. One has observations $\{Y_i\}_{i=1}^n$, and fixed co-variables $x_1, \ldots, x_n$, where the response variables satisfy the regression

$$Y_i = g_0(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and centred noise variables, and where g_0 is an unknown function on $\mathcal{X}$. The errors are assumed to be $\mathcal{N}(0, \sigma^2)$-distributed.

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

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

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

$$g_* = \arg\min_{g \in \bar{\mathcal{G}}} \left\{ \|g - g_0\|_n^2 + \text{pen}(g) \right\}.$$

Our aim is to show that

$$(*) \qquad \mathbb{E}\|\hat{g}_n - g_0\|_n^2 \leq \text{const.} \left\{ \|g_* - g_0\|_n^2 + \text{pen}(g_*) \right\}.$$

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

In Section 12.1, we recall the definitions of estimation and approximation error. Section 12.2 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 12.3 shows that when considering a collection of nested finite models, a penalty $\text{pen}(g)$ proportional to the log-cardinality of the smallest class containing g will indeed mimic the oracle over this collection of models. In Section 12.4, 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. When the penalty is taken "too small" there will appear an additional term showing that not all variance was "killed". Section 12.5 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.

12.1 Estimation and approximation error

Let $\mathcal{G}$ be a model class. Consider first the least squares estimator without penalty

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

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

$$\text{pen}(g) = \min_{\mathcal{G}:\ g \in \mathcal{G}} \text{pen}(\mathcal{G}).$$

An estimator that makes a data-dependent choice among the possible models is

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

where $\hat{g}_n(\cdot, \mathcal{G})$ is the least squares estimator over $\mathcal{G}$. Let

$$g_*(\cdot, \mathcal{G}) := \arg\min_{g \in \mathcal{G}} \|g - g_0\|_n$$

be the best approximation of g_0 within the model $\mathcal{G}$. Then $\|g_*(\cdot, \mathcal{G}) - g_0\|_n^2$ is the (squared) approximation error if the model $\mathcal{G}$ is used. We define

$$g_* = \arg\min_{\mathcal{G} \in \{\mathcal{G}\}} \{\|g_*(\cdot, \mathcal{G}) - g_0\|_n^2 + \text{pen}(\mathcal{G})\},$$

which trades off approximation error $\|g_*(\cdot, \mathcal{G}) - g_0\|_n^2$ against complexity $\text{pen}(\mathcal{G})$.

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

12.2 Finite models

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

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

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

$$\|g_* - g_0\|_n = \min_{g \in \mathcal{G}} \|g - g_0\|_n.$$

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

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

The result of Lemma 12.2.1 below implies that the estimation error is proportional to $\log |\mathcal{G}|/n$, 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 12.6.1).

Lemma 12.2.1 *We have for all $t > 0$ and $0 < \delta < 1$,*

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n^2 \geq \frac{1}{1-\delta} \left\{ (1+\delta)\|g_* - g_0\|_n^2 + \frac{4(\log|\mathcal{G}| + t)}{n\delta} \right\} \right) \leq \exp[-t].$$

Proof. We have the basic inequality

$$\|\hat{g}_n - g_0\|_n^2 \leq 2(\epsilon, \hat{g}_n - g_*)_n + \|g_* - g_0\|_n^2.$$

For all $t > 0$, by the union bound

$$\mathbb{P}\left(\sup_{g \in \mathcal{G},\ \|g - g_*\|_n > 0} \frac{(\epsilon, g - g_*)_n}{\|g - g_*\|_n} > \sqrt{\frac{2(\log|\mathcal{G}| + t)}{n}} \right) \leq \exp[-t]\ \forall\ t > 0.$$

If $(\epsilon, \hat{g}_n - g_*)_n \leq \sqrt{\frac{2(\log|\mathcal{G}|+t)}{n}} \|\hat{g}_n - g_*\|_n$, we have, using $2\sqrt{ab} \leq a + b$ for all non-negative a and b,

$$\begin{aligned} \|\hat{g}_n - g_0\|_n^2 &\leq 2\sqrt{\frac{2(\log|\mathcal{G}| + t)}{n}} \|\hat{g}_n - g_*\|_n + \|g_* - g_0\|_n^2 \\ &\leq 2\sqrt{\frac{2(\log|\mathcal{G}| + t)}{n}} \Big(\|\hat{g}_n - g_0\|_n + \|g_* - g_0\|_n \Big) + \|g_* - g_0\|_n^2 \\ &\leq \delta \|\hat{g}_n - g_0\|_n^2 + \frac{4(\log|\mathcal{G}|/(n\delta) + t)}{n\delta} + (1+\delta)\|g_* - g_0\|_n^2. \end{aligned}$$

□

12.3 Nested finite models

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

As indicated in Section 12.1, it is a good strategy to take the penalty proportional to the estimation error. In the present context, this works as follows. Define

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

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

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

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

Let

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

and

$$g_* = \arg\min_{g \in \bar{\mathcal{G}}} \left\{ \|g - g_0\|_n^2 + \text{pen}(g) \right\}.$$

Lemma 12.3.1 *We have, for all $t > 0$ and $0 < \delta < 1$,*

$$\mathbb{P}\left(\|\hat{g}_n - g_0\|_n^2 > \frac{1}{1-\delta} \left\{ (1+\delta)\|g_* - g_0\|_n^2 + \text{pen}(g_*) + \frac{2t}{n\delta} \right\} \right) \leq \exp[-t].$$

Proof. Write down the basic inequality

$$\|\hat{g}_n - g_0\|_n^2 + \text{pen}(\hat{g}_n) \leq 2(\epsilon, \hat{g}_n - g_*)_n + \|g_* - g_0\|_n^2 + \text{pen}(g_*).$$

We invoke a peeling device. Define $\bar{\mathcal{G}}_j = \{g : \ 2^j < |\log \mathcal{G}(g)| \leq 2^{j+1}\}$, $j = 0, 1, \ldots$. We have for all $t > 0$,

$$\begin{aligned}
& \mathbb{P}\left(\exists \ g \in \bar{\mathcal{G}} : \ (\epsilon, g - g_*)_n > \sqrt{\frac{2(4 \log |\mathcal{G}(g)| + t)}{n}} \|g - g_*\|_n \right) \\
\leq \ & \sum_{j=0}^{\infty} \mathbb{P}\left(\exists g \in \bar{\mathcal{G}}_j, \ (\epsilon, g - g_*)_n > \sqrt{\frac{2(42^j + t)}{n}} \|g - g_*\|_n \right) \\
\leq \ & \sum_{j=0}^{\infty} \mathbb{P}\left(\exists g \in \bar{\mathcal{G}}_j, \ (\epsilon, g - g_*)_n > \sqrt{\frac{2(\log |\bar{\mathcal{G}}_j| + 2^{j+1} + t)}{n}} \|g - g_*\|_n \right) \\
\leq \ & \sum_{j=0}^{\infty} \exp[-(2^{j+1} + t)] \\
\leq \ & \sum_{j=0}^{\infty} \exp[-(j + 1 + t)] \leq \int_0^{\infty} \exp[-(x + t)] = \exp[-t].
\end{aligned}$$

But if

$$(\epsilon, \hat{g}_n - g_*)_n \leq \sqrt{\frac{2(4\log|\mathcal{G}(\hat{g}_n)| + t)}{n}} \|\hat{g}_n - g_*\|_n$$

the Basic inequality gives

$$\begin{aligned}
& \|\hat{g}_n - g_0\|_n^2 \\
\leq\ & 2\sqrt{\frac{2(4\log|\mathcal{G}(\hat{g}_n)| + t)}{n}} \|\hat{g}_n - g_*\|_n \\
+\ & \|g_* - g_0\|_n^2 + \text{pen}(g_*) - \text{pen}(\hat{g}_n) \\
\leq\ & 2\sqrt{\frac{2(4\log|\mathcal{G}(\hat{g}_n)| + t)}{n}} \Big(\|\hat{g}_n - g_0\|_n + \|g_* - g_0\|_n \Big) \\
+\ & \|g_* - g_0\|_n^2 + \text{pen}(g_*) - \text{pen}(\hat{g}_n) \\
\leq\ & \delta \|\hat{g}_n - g_0\|_n^2 + \frac{4(4\log|\mathcal{G}(\hat{g}_n)| + t)}{n\delta} - \text{pen}(\hat{g}_n) \\
+\ & (1+\delta)\|g_* - g_0\|_n^2 + \text{pen}(g_*) \\
=\ & \delta \|\hat{g}_n - g_0\|_n^2 + (1+\delta)\|g_* - g_0\|_n^2 + \text{pen}(g_*) + \frac{4t}{n\delta},
\end{aligned}$$

by the definition of pen(g). □

12.4 General penalties

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

Recall the definition of the estimator

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

and of the noiseless version

$$g_* = \arg\min_{g \in \bar{\mathcal{G}}} \left\{ \|g - g_0\|_n^2 + \text{pen}(g) \right\}.$$

We moreover define

$$\tau^2(g) := \|g - g_0\|_n^2 + \text{pen}(g),\ g \in \bar{\mathcal{G}},$$

and

$$\mathcal{G}(\delta) = \{g \in \bar{\mathcal{G}} :\ \tau^2(g) \leq \delta^2\},\ \delta > 0.$$

Consider the entropy $H(\cdot, \mathcal{G}(\delta), Q_n)$ of $\mathcal{G}(\delta)$. Suppose it is finite for each δ, and in fact that the square root of the entropy is integrable:

Condition 12.4.1 *One has*

$$J(\delta) := 2\int_0^{2\delta} \sqrt{2H(u, \mathcal{G}(\delta), Q_n)}du < \infty,\ \forall\ \delta > 0. \qquad (**)$$

This means that near $u = 0$, the entropy $H(u, \mathcal{G}(\delta), Q_n)$ is not allowed to grow faster than $1/u^2$.

Theorem 12.4.1 *Assume Condition 12.4.1. Suppose that $J(\delta)/\delta^2$ is decreasing function of δ. Then for all $t > 0$ and*

$$(\bullet) \qquad \delta_n^2 \geq 4\tau^2(g_*) + 8\left(\frac{J(\delta_n)}{\sqrt{n}} + 8\delta_n\sqrt{\frac{1+t}{n}}\right)$$

we have

$$\mathbb{P}(\tau(\hat{g}_n) > \delta_n) \leq \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t].$$

Proof. Since $\delta_n \geq \tau(g_*)$ we know that when $g \in \mathcal{G}(2^j\delta_n)$

$$\|g - g_*\|_n \leq \|g - g_0\|_n + \|g_* - g_0\|_n \leq \tau(g) + \tau(g_*) \leq 2^j\delta_n + \delta_n \leq 2^{j+1}\delta_n.$$

By the Basic inequality

$$\mathbb{P}(\tau(\hat{g}_n) > \delta_n) \leq$$

$$\sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{g\in\mathcal{G}(2^j\delta_n)} 2(\epsilon, g - g_*)_n \geq (2^{j-1}\delta_n)^2 - \tau^2(g_*)\right).$$

We can apply the same arguments as in the proof of Theorem 11.2.1: since

$$\delta_n^2 \geq 4\tau^2(g_*) + 8\left(\frac{J(\delta_n)}{\sqrt{n}} + 8\delta_n\sqrt{\frac{1+t}{n}}\right)$$

and $\delta \mapsto J(\delta)/\delta^2$ is decreasing, it holds for all $j \in \mathbb{N}$

$$(2^j\delta_n)^2 \geq 4\tau^2(g_*) + 8\left(\frac{J(2^j\delta_n)}{\sqrt{n}} + 8 2^j\delta_n\sqrt{\frac{1+t+j}{n}}\right)$$

so that

$$(2^{j-1}\delta_n)^2 - \tau^2(g_*) \geq 2\left(\frac{J(2^j\delta_n)}{\sqrt{n}} + 8 2^j\delta_n\sqrt{\frac{1+t+j}{n}}\right)$$

and hence

$$\begin{aligned}
& \sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{g\in\mathcal{G}(2^j\delta_n)} 2(\epsilon, g - g_*)_n \geq (2^{j-1}\delta_n)^2 - \tau^2(g_*)\right) \\
\leq \; & \sum_{j=1}^{\infty}\mathbb{P}\left(\sup_{g\in\mathcal{G}(2^j\delta_n)} 2(\epsilon, g - g_*)_n \geq 2\left(\frac{J(2^j\delta_n)}{\sqrt{n}} + 8 2^j\delta_n\sqrt{\frac{1+t+j}{n}}\right)\right) \\
\leq \; & \sum_{j=1}^{\infty}\exp[-(t+j)] = \frac{\mathrm{e}}{\mathrm{e}-1}\exp[-t].
\end{aligned}$$

□

12.5 Application to the "classical" penalty

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

$$I^p(g) = \int_0^1 |g^{(m)}(x)|^p dx, \ g \in \bar{\mathcal{G}}.$$

We consider two cases. In Subsection 12.5.1, we fix a tuning parameter $\lambda > 0$ and take the penalty $\text{pen}(g) = \lambda^2 I^p(g)$. 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 12.5.2 to a data dependent choice for λ, or alternatively viewed, a penalty of the form $\text{pen}(g) = \tilde{\lambda}^2 I^{\frac{2}{2m+1}}(g)$, with $\tilde{\lambda} > 0$ depending on m and n. This penalty allows one to adapt to small values for $I(g_0)$.

12.5.1 Fixed smoothing parameter

For a function $g \in \bar{\mathcal{G}}$, we define the penalty

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

with a given $\lambda > 0$.

Lemma 12.5.1 *The entropy integral J can be bounded by*

$$J(\delta) \leq A_0 \left(\delta^{\frac{2pm+2-p}{2pm}} \lambda^{-\frac{1}{pm}} + \delta \sqrt{\log(\frac{1}{\lambda} \vee 1)} \right) \ \delta > 0.$$

Here, A_0 is a constant depending on m and p.

Proof. This follows from the fact that

$$H_\infty(u, \{g \in \bar{\mathcal{G}} : \ I(g) \leq 1, \ |g| \leq 1\}) \leq A u^{-1/m}, \ u > 0$$

where the constant A depends on m and p (see Birman and Solomjak (1967)). For $g \in \mathcal{G}(\delta)$, we have

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

and

$$\|g - g_*\|_n \leq \delta.$$

We therefore may write $g \in \mathcal{G}(\delta)$ as $g = g_1 + g_2$, with $|g_1| \leq I(g_1) = I(g)$ and $\|g_2 - g_*\|_n \leq \delta + I(g)$. It is now not difficult to show that for some constant A_1

$$H_2(u, \mathcal{G}(\delta), Q_n) \leq A_1 \left(\left(\frac{\delta}{\lambda} \right)^{\frac{2}{pm}} u^{-\frac{1}{m}} + \log\left(\frac{\delta}{(\lambda \wedge 1)u} \right) \right), \ 0 < u < \delta.$$

□

Corollary 12.5.1 *By applying Lemma 12.5.1 , we find that for some constant* c_1,

$$\|\hat{g}_n - g_0\|_n^2 + \lambda^2 I^p(\hat{g}_n) \leq 4 \min_g \{\|g - g_0\|_n^2 + \lambda^2 I^p(g)\}$$

$$+\mathcal{O}_{\mathbf{P}}\left(\left(\frac{1}{n\lambda^{\frac{2}{pm}}}\right)^{\frac{2pm}{2pm+p-2}} + \frac{\log\left(\frac{1}{\lambda} \vee 1\right)}{n}\right).$$

12.5.2 Overruling the variance in this case

For choosing the smoothing parameter λ, the above suggests the penalty

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

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

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

where C_0' depends on C_0 and m. 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

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

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

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

One arrives at the following corollary.

Corollary 12.5.2 *For an appropriate, large enough, choice of* C_0' *depending on* p *and* m, *we have*

$$\|\hat{g}_n - g_0\|_n^2 + C_0' n^{-\frac{2m}{2m+1}} I^{\frac{2}{2m+1}}(\hat{g}_n)$$

$$\leq 4 \min_g \left\{\|g - g_0\|_n^2 + C_0' n^{-\frac{2m}{2m+1}} I^{\frac{2}{2m+1}}(g)\right\} + \mathcal{O}_{\mathbf{P}}(1/n).$$

Thus, the estimator adapts to small values of $I(g_0)$. For example, when $m = 1$ and $I(g_0) = 0$ (i.e., when g_0 is the constant function), the excess risk of the estimator converges with parametric rate $1/n$. If we knew that g_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.

12.6 Exercises

Exercise 12.6.1 Using the formula

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

for a non-negative random variable Z, derive bounds for the average excess risk $\mathbb{E}\|\hat{g}_n - g_0\|_n^2$ of the estimator considered in this chapter.