STATISTICAL THEORY

Richard Nickl *

Version: November 27, 2013

Contents

*Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge. Email: r.nickl@statslab.cam.ac.uk. These are informal notes for a lecture course of the same title held in Part III of the Mathematical Tripos 2011-2013. Please inform me of any typos/mistakes in these notes. I would like to thank Adam Bull, Ismaël Castillo, Philippe Charmoy, Yining Chen, Helge Dietert, Andre Kueh, Matthew Phinney, Kolyan Ray, Richard Samworth, Jakob Söhl, Aaron Yang Yu and Dominic Zedan for comments and/or corrections. All remaining errors are mine.

1 Introduction

Consider a real random variable Y with unknown distribution function

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

where P is a probability measure defined on the Borel sets of $\mathbb{R}$, and suppose one observes a *sample* of Y, that is, n independent and identically distributed copies $Y_1, \ldots, Y_n$ of Y. Probability theory provides the axiomatic definition of the mathematical objects P and F, and furnishes us with an exact notion of *independence* of random variables. But can the physically more plausible frequentist notion of probability be *derived* from these minimal axiomatic foundations? Or, in simpler words: does the sample $Y_1, \ldots, Y_n$ 'tell us' what the distribution F of Y is, at least approximately, for sufficiently large sample size n? We expect the answer to be 'yes' in light of the *law of large numbers*, and mathematical statistics is about developing a rigorous theory about the precise meaning of this question, and about the various complex issues at the heart of the possible answers one may give.

Statistical analysis starts with a specification of a 'model' for F. This means that we specify a subset $\mathcal{P}$ of the set of all probability distribution functions on $\mathbb{R}$. We shall encounter in this course models that range from the simplest model of univariate normal distributions

$$\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

to the exhaustive infinite-dimensional model

$$\mathcal{P} = \{\text{All probability distribution functions on } \mathbb{R}\}.$$

This varying complexity does not only occur in the problem of estimating a distribution function, but likewise in regression problems: Often the parameters specifying the distribution are modelled themselves by the statistician to explain certain functional (or even causal) relationships. In the simplest case this functional relation is modelled linearly, for instance one postulates

$$Y_i = \theta x_i + u_i$$

where x_i an explanatory variable/regressor, $\theta \in \mathbb{R}$ the parameter and $u_i \sim N(0, \sigma^2)$. Again, the set of all functional relationships

$$Y_i = g(x_i) + u_i$$

is infinite-dimensional, and the restriction to linear g, or to g a fixed known function, is not necessarily sensible.

These lecture notes try to give a mathematical introduction to some key aspects of statistical theory. An attempt is made to be mathematically as self-contained as possible without loosing focus over excessive technicalities. An emphasis is given to develop an understanding of the interplay of probabilistic properties of random samples with the analytic structures of the model $\mathcal{P}$. This approach goes at the expense of the breadth of topics that can be covered, but the hope is that some main ideas that are representative of the whole field can be developed rigorously. We shall analyse both finite and infinite dimensional models in this course, and we shall see that, much like in analysis, the statistical theory of finite-dimensional – or 'parametric' – models is distinctively different from the theory of infinite-dimensional – or 'nonparametric' – models, and this will introduce a natural division of this course into two parts, which require different mathematical techniques and statistical intuitions. Somewhat in between lies the family of 'high-dimensional' models, which shall be introduced as well.

The rough outline is as follows: We shall first review some basic results on probabilistic limit theorems, which will be at the heart of many results of this course, and we shall also prove a basic proposition from empirical process theory that will be useful throughout. We then develop the by now classical consistency and asymptotic normality theory in regular parametric models and explain how this leads to Le Cam's unifying notion of *local asymptotic normality (LAN)* of statistical experiments. We shall use Le Cam theory to prove the celebrated Bernstein-von Mises theorem about the frequentist interpretation of Bayes procedures in a locally asymptotically normal setting. We shall then develop some main ideas of the theory of 'large p - small n' problems in the setting of normal linear models, including the LASSO estimator and an analysis of the restricted isometry property for Gaussian design matrices. In the part concerning 'nonparametric models' the theory admits a less unified structure due to the absence of 'local asymptotic normality' of most of these models, but we shall nevertheless attempt to highlight some of the integral ideas behind this theory, in particular the minimax paradigm and the solution of the adaptation problem that arises from it.

We shall assume that the reader is familiar with basic measure theory and elementary stochastic convergence properties of random variables, see Dudley [29] and van der Vaart ([81], Chapter 2) for comprehensive accounts of these topics.

1.1 The Law of Large Numbers and the Central Limit Theorem

Under a random variable we shall understand a measurable mapping X from some probability space $(\Omega, \mathcal{A}, \mu)$ into some metric space (S, d). The *law* of X is the image measure $\mu \circ X^{-1}$. A sequence of random variables X_n taking values in (S, d)

converges to a random variable X *in probability*, or $X_n \to^P X$ in (S,d) if, for every $\varepsilon > 0$,

$$\mu(\omega \in \Omega : d(X_n(\omega), X(\omega)) > \varepsilon) \to 0 \text{ as } n \to \infty.$$

Likewise X_n converges to X *in distribution*, or *in law*, or $X_n \to^d X$, in the space (S,d), if

$$Ef(X_n) \to Ef(X)$$

as $n \to \infty$ for every bounded continuous function $f : S \to \mathbb{R}$. There is also the notion of almost sure convergence: $X_n \to X$ $\mu - a.s.$ if

$$\mu(\omega \in \Omega : \lim X_n(\omega) = X(\omega)) = 1.$$

Almost sure convergence, which is statistically less relevant, is stronger than convergence in probability of X_n to X, and is sometimes useful in proofs. Recall further that convergence in probability implies convergence in distribution, but that the converse is false.

If (S,d) equals $\mathbb{R}^p$ with the standard metric induced by the Euclidean norm $\|\cdot\|$, then this is the classical definition of a random variable/vector. In particular convergence in distribution is then equivalent to convergence of the distribution functions $F_{X_n}(t) = \mu(X_n \leq t)$ to $F_X(t) = \mu(X \leq t)$ at continuity points of F_X. We shall omit to mention (S,d) if it equals $\mathbb{R}^p$.

Let now $X, X_1, \ldots, X_n, \ldots$ be i.i.d. random vectors in $\mathbb{R}^p$, and write $P \equiv \mu \circ X^{-1}$ for their common law. By the symbol Pr we shall always mean the product probability measure $P^{\mathbb{N}}$ defined on the canonical product space $(\mathbb{R}^p)^\infty$ given by the joint law of $(X_1, ..., X_n, \ldots)$. This measure exists as the unique extension of the joint law P^n of $(X_1, \ldots, X_n)$ to $(\mathbb{R}^p)^\infty$ (see Chapter 8.2 in [29] for instance). By E we shall in such a case denote expectation with respect to Pr. This notation allows us, as will be convenient, to avoid mentioning the underlying probability space $(\Omega, \mathcal{A}, \mu)$. For instance, if we assume in addition that $E\|X\| < \infty$, the law of large numbers states that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to E(X) \quad \Pr - a.s. \tag{1}$$

and thus also in probability, as $n \to \infty$. The central limit theorem states that if X satisfies $E\|X\|^2 < \infty$ and has a positive definite covariance matrix Σ then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E(X)\right) \to^d N(0, \Sigma). \tag{2}$$

These two classical results are the pillars of much of asymptotic ('n large') statistical theory, as we shall see. They have a comparably less well known 'nonasymptotic' analogue, known as Hoeffding's inequality: If $X_1, ..., X_n$ are mean zero independent random variables taking values in $[b_i, c_i]$ for constants $b_i < c_i$, then, for

every $n \in \mathbb{N}, u > 0$,

$$\Pr\left\{\left|\sum_{i=1}^{n} X_i\right| > u\right\} \leq 2\exp\left(-\frac{2u^2}{\sum_{i=1}^{n}(c_i - b_i)^2}\right), \tag{3}$$

which should be compared to the tail of the limit distribution in (2). It shows that the normal approximation is valid in the tails, if the X_i's are bounded, for every sample size n. The proof is left as Exercise 1.

1.2 Uniform Laws of Large Numbers

The key results (1) and (2) are very useful in statistics. Some more subtle mathematical arguments will in fact require that the law of large numbers holds 'simultaneously' for many random variables – the way to make this precise is via laws of large numbers that are *uniform* in certain classes of functions.

Consider for the moment the even more general case where $X, X_1, X_2, \ldots$ are i.i.d. random variables taking values in the arbitrary measurable space T (typically $T = \mathbb{R}^d$, but other choices are possible) so that their joint law is the product probability measure Pr on T^∞. If $\mathcal{H}$ is a class of measurable real-valued functions defined on T and such that $E|h(X)| < \infty$ for each $h \in \mathcal{H}$, then

$$\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Eh(X)) \to 0 \quad \Pr - a.s$$

as $n \to \infty$, for every $h \in \mathcal{H}$ by (1). A law of large numbers holds *uniformly* over a class $\mathcal{H}$ of functions if also

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Eh(X))\right| \to 0 \quad \Pr - a.s. \tag{4}$$

as $n \to \infty$. The following general purpose result, which is based on a simple 'bracketing idea', gives a sufficient condition for $\mathcal{H}$ to satisfy such a uniform law of large numbers. Given two (measurable) real-valued functions $l(x), u(x)$ on the (measurable) space T, a 'bracket' is the set of functions

$$[l, u] := \{f : T \to \mathbb{R} : l(x) \leq f(x) \leq u(x) \text{ for all } x \in T\}.$$

Proposition 1. *Let $\mathcal{H}$ be a class of functions from T to $\mathbb{R}$. Assume that for every $\varepsilon > 0$ there exists a finite set of brackets $[l_j, u_j], j = 1, \ldots, N(\varepsilon)$, such that $E|l_j(X)| < \infty, E|u_j(X)| < \infty$ and $E|u_j(X) - l_j(X)| < \varepsilon$ for every j. Suppose moreover that for every $h \in \mathcal{H}$ there exists j with $h \in [l_j, u_j]$. Then (4) holds.*

Proof. Write shorthand $E_n g := n^{-1}\sum_{i=1}^n g(X_i)$ and $Eg := Eg(X)$ for functions $g : T \to \mathbb{R}$. Let $\varepsilon > 0$ be arbitrary and choose brackets $[l_j, u_j]$ such that

$$E|u_j - l_j|(X) < \varepsilon/2 \tag{5}$$

for every $j = 1, ..., N(\varepsilon/2)$, which is possible by hypothesis. Consider first the case where the u_j, l_j do not depend on ε. We claim that for every $\omega \in T^\infty \setminus A$ with $\Pr(A) = 0$ (A is called a 'null-set') there exists an index $n_0 := n_0(\omega, \varepsilon)$ such that $n \geq n_0$ implies

$$\max_{j=1,...,N(\varepsilon/2)} |E_n u_j - Eu_j| < \varepsilon/2 \tag{6}$$

as well as

$$\max_{j=1,...,N(\varepsilon/2)} |E_n l_j - El_j| < \varepsilon/2. \tag{7}$$

To see this observe the following: by the ordinary strong law of large numbers (and definition of almost sure convergence), there exist sets A_j independent of ε with $\Pr(A_j) = 0$ such that the limit of the j-th term in these maxima is zero for every $\omega \in T^\infty \setminus A_j$, so in particular each term is less than $\varepsilon/2$ for every $\omega \in T^\infty \setminus A_j$ and $n \geq n_0(j, \omega, \varepsilon)$. Then choose $n_0 := \max_j n_0(j, \omega, \varepsilon)$ which, being a maximum of finitely many integers, is again finite. The finite union $A := \cup_j A_j$ of null sets still satisfies

$$\Pr(A) = \Pr(\cup A_j) \leq \sum_{j=1}^{N(\varepsilon)} \Pr(A_j) = 0.$$

If the u_j, l_j depend on ε repeat the above argument with $\varepsilon_m = 1/m, m \in \mathbb{N}$, in which case the exceptional sets $A_{j,m}$ depend on m as well but still satisfy that $\sum_{j,m} \Pr(A_{j,m}) = 0$.

Now combining (5), (6), (7) we have for $h \in \mathcal{H}$ arbitrary, every $\omega \in T^\infty \setminus A$ and $n \geq n_0$ that

$$\frac{1}{n}\sum_{i=1}^n h(X_i) - Eh(X) = E_n h - Eh \leq E_n u_j - Eh = E_n u_j - Eu_j + E(u_j - h) < \varepsilon$$

and similarly

$$E_n h - Eh \geq E_n l_j - El_j + E(l_j - h) > -\varepsilon.$$

Hence for every $\omega \in T^\infty \setminus A$ and $n \geq n_0$, $|E_n h - Eh| < \varepsilon$. Since ε and $h \in \mathcal{H}$ were arbitrary, we have $\lim_n \sup_{h\in\mathcal{H}} |E_n h - Eh| = 0$ for every $\omega \in T^\infty \setminus A$ and hence the result. □

For measure theory enthusiasts: It should be noted that the supremum in (4) is not necessarily measurable (i.e., a proper random variable). The simplest way

to show that it is one is to show that the supremum can be realized as one over a countable subset of $\mathcal{H}$.

We shall use Proposition 1 at several key places in the rest of these lecture notes. For the interested reader let us mention the following consequence of it.

Example 1 (The Law of Large Numbers in Banach Spaces)**.** Together with some facts from functional analysis Proposition 1 can be used to prove the following: Let $(S, \|\cdot\|_S)$ be any separable Banach (i.e., complete normed linear) space, let $X_1, ..., X_n$ be i.i.d. random variables taking values in S, and assume that $E\|X\|_S < \infty$. Then $\left\|\frac{1}{n}\sum_{i=1}^n X_i - E(X)\right\|_S \to 0$ almost surely, and the moment condition is necessary, see [54], Corollary 7.10. For a proof that uses Proposition 1 and the Ascoli-Arzela theorem see Chapter 7.1 in Dudley [28], to whom this proof is due.

1.3 Exercises

Exercise 1. The following result is known as Hoeffding's inequality: If $X_1, ..., X_n$ are mean zero independent random variables taking values in $[b_i, c_i]$ for constants $b_i < c_i$, then for every $n \in \mathbb{N}, u > 0$,

$$\Pr\left\{\sum_{i=1}^n X_i > u\right\} \leq \exp\left(-\frac{2u^2}{\sum_{i=1}^n (c_i - b_i)^2}\right) \tag{8}$$

of which an obvious consequence is (why?)

$$\Pr\left\{\left|\sum_{i=1}^n X_i\right| > u\right\} \leq 2\exp\left(-\frac{2u^2}{\sum_{i=1}^n (c_i - b_i)^2}\right). \tag{9}$$

Provide a proof of this inequality. [You may find it useful to first prove the auxiliary result $E(\exp\{vX_i\}) \leq \exp\{v^2(c_i - b_i)^2/8\}$ for $v > 0$, and then use Markov's inequality in conjuction with a bound for the moment generating function of $v\sum X_i$.]

2 Parametric Models

Consider the situation where we are given an i.i.d. sample $Y_1, \ldots, Y_n$ with unknown distribution $F(t) = P(Y \leq t)$. Suppose we have reason to postulate, before the data was collected, that P belongs to a family of probability measures

$$\mathcal{P}_\Theta := \{P_\theta : \theta \in \Theta\}$$

where Θ is a subset of an Euclidean space $\mathbb{R}^p$. The set Θ is called the *parameter space*. This general setting entails all finite-dimensional models usually encountered in statistical inference: for instance it includes normal, Poisson, Beta, exponential, binomial etc., – indeed all parametric families of distributions. This setting also easily extends to important statistical problems where the sample is independent but not identically distributed, say $Y_i \sim F_i(\theta)$, where $\theta \in \Theta$ does not depend on i, which covers linear, generalised linear and nonlinear regression problems that are at the heart of much of statistical inference.

A crucial assumption that we shall impose on $\mathcal{P}_\Theta$ is that it is *correctly specified.* This means that we assume that there exists a point $\theta_0 \in \Theta$ such that the true distribution P of the sample equals P_{θ_0} – we shall usually refer to θ_0 as the true value of θ. We shall then often write, in slight abuse of notation, P_{θ_0} for Pr to indicate that we are computing probabilities under the law that generated the actual sample, and likewise E_{θ_0} for expectations under Pr.

The goal of statistical inference in these situations is typically not to estimate P_{θ_0}, but rather to estimate θ_0 (which in turn entails an estimate of P_{θ_0} as well). This is often achieved by defining estimators as solutions of maximisation/minimisation problems, and the resulting estimators $\hat{\theta}_n$ are thus often called M-estimators. Here are two leading examples.

Example 2 (Nonlinear Least Squares (NLS))**.** Consider the model

$$Z_i = g(x_i, \theta) + u_i, \quad i = 1, \ldots, n$$

where the x_i's are some design points/explanatory variables, where g is a known possibly nonlinear regression function, $\theta \in \Theta$ a parameter indexing the set of possible regression functions, and where the u_i's are random variables with $E(u_i|x_i) = 0$. Fixed design corresponds to x_i nonrandom whereas in the random design case it is often assumed that the $Y_i \equiv (Z_i, x_i)$ are jointly i.i.d. random vectors. The intuitive least squares estimator $\hat{\theta}_n$ of θ solves the minimization problem

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (Z_i - g(x_i, \theta))^2.$$

Example 3 (Maximum Likelihood Estimators (MLEs))**.** Suppose we have a family $\mathcal{P}_\Theta = \{f(\theta, \cdot) : \theta \in \Theta\}$ of probability densities $f(\theta)$, and suppose we have an i.i.d. sample $Y_1, \ldots, Y_n$ from one of these densities, still denoted by $f(\theta, \cdot)$. We shall write P_θ for the probability measure induced by the density $f(\theta)$. The joint distribution of the sample is $\prod_{i=1}^n f(\theta, y_i)$, and if we view this as a function of the parameter only and evaluate y_i at the sample points, this defines the likelihood function

$$L_n(\theta) = L(\theta; Y_1, \ldots, Y_n) = \prod_{i=1}^n f(\theta, Y_i).$$

It is often convenient to work with the log-likelihood function

$$l_n(\theta) = l(\theta; Y_1, \ldots, Y_n) = \sum_{i=1}^n \log f(\theta, Y_i),$$

with the convention that $\log 0 = -\infty$. The maximum likelihood estimator solves

$$\max_{\theta \in \Theta} l_n(\theta).$$

Many other examples can be given, such as method of moment estimators. In general the finite-sample properties of so-defined estimators are intractable, and one has to resort to asymptotic (n large) approximations of the distribution of $\hat{\theta}_n$. A first goal, however, is to show that these estimators make sense from a frequentist point of view in that $\hat{\theta}_n$ converges in probability to θ_0 for every possible true value θ_0, as sample size increases ($n \to \infty$).

2.1 Consistency of M-Estimators

We shall in this subsection adopt a general framework and study estimators based on a sample $Y_1, \ldots, Y_n$ that are obtained from minimising a criterion function $Q_n(\theta) \equiv Q_n(\theta; Y_1, \ldots, Y_n), \theta \in \Theta \subset \mathbb{R}^p$, over Θ. In the nonlinear least squares example this criterion function equals, recalling the notation $Y_i = (Z_i, x_i)$,

$$Q_n(\theta; Y_1, \ldots, Y_n) = \frac{1}{n} \sum_{i=1}^n (Z_i - g(x_i, \theta))^2$$

and in the maximum likelihood problem it equals

$$Q_n(\theta; Y_1, \ldots, Y_n) = -\frac{1}{n} \sum_{i=1}^n \log f(\theta, Y_i) = -\frac{1}{n} l_n(\theta),$$

but the results that follow do not require this sample average structure unless specifically mentioned.

2.1.1 A General Consistency Theorem

The statistical intuiting behind such procedures is that $Q_n(\theta)$ is close, with high probability, to some nonrandom function $Q(\theta)$, that this function is minimized at the true value θ_0, in a unique way, and that thus a minimizer of Q_n should be close to θ_0. In the above examples Q_n is based on a sample mean and if $EQ_n(\theta)$ exists this will define $Q(\theta)$as the limit in probability of Q_n, by the law of large numbers (1). For instance, in the case of MLEs,

$$Q(\theta) = -E \log f(\theta, Y).$$

In mathematical terms we are asking the following: If a sequence of random functions Q_n converges to Q, can we find weak conditions that ensure that the minimizers of Q_n converge to the minimizer of Q, if the latter exists? Here is a general result of this kind.

Theorem 1. *Suppose $\Theta \subset \mathbb{R}^p$ is compact (i.e., bounded and closed). Assume that $Q : \Theta \to \mathbb{R}$ is a (nonrandom) function that is continuous on Θ, and that θ_0 is the unique minimizer of Q. If*

$$\sup_{\theta \in \Theta} |Q_n(\theta; Y_1, \ldots, Y_n) - Q(\theta)| \to^P 0 \tag{10}$$

as $n \to \infty$, then any solution $\hat{\theta}_n$ of

$$\min_{\theta \in \Theta} Q_n(\theta, Y_1, \ldots, Y_n)$$

converges to θ_0 in probability as $n \to \infty$.

Proof. For every $\varepsilon > 0$, the set $\{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}$ is compact and Q is continuous on this set, so $\inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon} Q(\theta)$ is attained, and since θ_0 is a unique minimiser we necessarily have

$$c(\varepsilon) \equiv \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon} Q(\theta) > Q(\theta_0). \tag{11}$$

Choose $0 < \delta(\varepsilon) < (c(\varepsilon) - Q(\theta_0))/2$ so that $c(\varepsilon) - \delta(\varepsilon) > Q(\theta_0) + \delta(\varepsilon)$. On the event

$$A_n(\varepsilon) \equiv \left\{ \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| < \delta(\varepsilon) \right\}$$

we have

$$\begin{aligned} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon} Q_n(\theta) &\geq \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon} Q(\theta) - \delta(\varepsilon) = c(\varepsilon) - \delta(\varepsilon) \\ &> Q(\theta_0) + \delta(\varepsilon) \geq Q_n(\theta_0) \end{aligned}$$

so if $\hat{\theta}_n$ would lie in $\{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon\}$ then θ_0 would yield a strictly smaller value of Q_n, a contradiction to $\hat{\theta}_n$ being a minimiser. We conclude

$$\left\{\|\hat{\theta}_n - \theta_0\| < \varepsilon\right\} \supset A_n(\varepsilon)$$

but by (10) we have $\Pr(A_n(\varepsilon)) \to 1$ as $n \to \infty$ for every $\varepsilon > 0$ so

$$\Pr\left(\left\{\|\hat{\theta}_n - \theta_0\| < \varepsilon\right\}\right) \to 1$$

as well, hence consistency. □

We have again completely neglected measurability issues: It is not a fortiori clear that $\hat{\theta}_n$ is a measurable function of the Y_i's. Sufficient conditions that apply to most examples can be found, for instance, in Lemma A3 in [66], but we shall neglect this issue in what follows and will tacitly assume that $\hat{\theta}_n$ is a proper random variable.

The assumptions of uniqueness of θ_0, continuity of Q, and condition (10) will be discussed in the next two subsections. The only other assumption is compactness of Θ, which at first looks restrictive – for instance in linear regression $y_i = \theta x_i + u_i$ the parameter space is usually all of $\mathbb{R}$, so not compact. Compactness is only explicitly used in the proof Theorem 1 to establish (11), which can often be verified without compactness. However, as we shall see below, compactness (or at least boundedness) of Θ is often crucial in the verification of (10), where it is not as easy to relax. A better strategy is to first prove that the criterion function Q_n is uniformly large outside of a fixed compact set Θ^*, so that $\hat{\theta}_n \in \Theta^*$ on sets of probability approaching one, and then to apply Theorem 1 with Θ_0 in place of Θ. See Exercise 6 for how this applies in regression problems.

2.1.2 Identifiability

Theorem 1 can be used if the limiting criterion function Q is continuous and has a unique minimiser. This assumption, which depends on the analytic properties of the parameterisation $\theta \mapsto Q(\theta)$, is typically a natural one, as we show in the following examples.

Example 4. *(Nonlinear Least Squares)* Consider again $Z_i = g(x_i, \theta_0) + u_i$ where $Y_i = (Z_i, x_i)$ are i.i.d. random vectors, where g is a known regression function, and where u_i and X_i are independent, $E(u_i^2) = \sigma^2 < \infty$. Then

$$\begin{aligned} Q(\theta) &= E\left[(Z_i - g(x_i, \theta_0) + g(x_i, \theta_0) - g(x_i, \theta))^2\right] \\ &= E\left[(u_i + g(x_i, \theta_0) - g(x_i, \theta))^2\right] \\ &= \sigma^2 + E\left[(g(x_i, \theta_0) - g(x_i, \theta))^2\right]. \end{aligned}$$

Thus the identification assumption of uniqueness of θ_0 reduces to an assumption on the (known) regression function g, namely, whether $g(\cdot,\theta) = g(\cdot,\theta_0)$ holds in mean square if and only if $\theta = \theta_0$, a more than reasonable assumption for any regression model.

Example 5. *(Maximum Likelihood Estimators)* Consider next the case of maximum likelihood estimation where $Y_1,\ldots,Y_n$ come from some density $f(\theta_0,\cdot)$ on $\mathcal{Y} \subset \mathbb{R}^d$, and $\theta_0 \in \Theta$ so that the model $\{f(\theta,\cdot) : \theta \in \Theta\}$ is correctly specified. Assume $\int_{\mathcal{Y}} |\log f(\theta,y)| f(\theta_0,y)dy < \infty$ and $f(\theta,y) > 0$ for every $\theta \in \Theta$ and every $y \in \mathcal{Y}$. In this case the limiting criterion function is, by the law of large numbers, equal to

$$Q(\theta) = -E_{\theta_0}\left(\frac{1}{n}\sum_{i=1}^{n}\log f(\theta,Y_i)\right) = -E_{\theta_0}\log f(\theta,Y) = -\int_{\mathcal{Y}}\log f(\theta,y)f(\theta_0,y)dy$$

and here it is less obvious that the limiting minimizer is the true value θ_0. The difference $Q(\theta_0) - Q(\theta)$ equals the negative of the so-called Kullback-Leibler distance between $f(\theta)$ and $f(\theta_0)$, a concept of key importance in statistics and information theory: In fact

$$\begin{aligned} Q(\theta_0) - Q(\theta) &= \int\left[\log\frac{f(\theta,y)}{f(\theta_0,y)}f(\theta_0,y)\right]dy \qquad (12)\\ &\leq \log\int f(\theta,y)dy = \log 1 = 0. \end{aligned}$$

by Jensen's inequality. So $Q(\theta_0) \leq Q(\theta)$ for every $\theta \in \Theta$, i.e., θ_0 is a minimiser of the limiting function Q. If we impose further the natural identifiability assumption

$$f(\theta_0,\cdot) = f(\theta_1,\cdot) \text{ Lebesgue-almost everywhere} \Leftrightarrow \theta_0 = \theta_1 \qquad (13)$$

then the ratio $f(\theta,y)/f(\theta_0,y)$ is not identical one almost everywhere for every $\theta \neq \theta_0$. Since the logarithm is strictly concave the strict version of Jensen's inequality implies that (12) holds with strict inequality, that is under (13) we have $Q(\theta_0) < Q(\theta)$ for every $\theta \neq \theta_0$, so θ_0 is unique.

2.1.3 Verifying Uniform Convergence

A key step to making Theorem 1 applicable is to verify uniform convergence (10). Note first that without uniformity in θ in (10) the conclusion of Theorem 1 may be false, see Exercise 5 below.

If Q_n has the form of a sample mean such as

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} q(\theta,X_i),$$

as is the case of NLS and ML estimation, then uniform convergence can be established without too much difficulty using the 'bracketing' uniform law of large numbers from the introduction. The following proposition gives mild sufficient conditions under which Proposition 1 applies.

Proposition 2. *Suppose Θ is a bounded and closed subset of $\mathbb{R}^p$, and let $q(\theta, x) : \Theta \times \mathbb{R}^d \to \mathbb{R}$ be continuous in θ for each x and measurable in x for each θ. If $X_1, ..., X_n$ are i.i.d. in $\mathbb{R}^d$ and if*

$$E \sup_{\theta \in \Theta} |q(\theta, X)| < \infty \tag{14}$$

then, as $n \to \infty$,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} q(\theta, X_i) - Eq(\theta, X) \right| \to 0 \quad \Pr - a.s. \tag{15}$$

Proof. We apply Proposition 1, so (15) will be proved if we find suitable brackets for the class of functions

$$\mathcal{H} = \{q(\theta, \cdot) : \theta \in \Theta\},$$

which is done as follows: First define the open balls $B(\theta, \eta) = \{\theta' \in \Theta : \|\theta - \theta'\| < \eta\}$, and define, for *every* $\theta \in \Theta$, the auxiliary brackets

$$u(x, \theta, \eta) = \sup_{\theta' \in B(\theta, \eta)} q(\theta', x)$$

and

$$l(x, \theta, \eta) = \inf_{\theta' \in B(\theta, \eta)} q(\theta', x)$$

so that clearly $l(x, \theta, \eta) \leq q(\theta', x) \leq u(x, \theta, \eta)$ holds for every $x \in \mathbb{R}^d$ and every $\theta' \in B(\theta, \eta)$. By condition (14) we have

$$E|u(X, \theta, \eta)| < \infty, \quad E|l(X, \theta, \eta)| < \infty \tag{16}$$

for every $\theta \in \Theta$ and every η. Furthermore since $q(\cdot, x)$ is continuous, the suprema in the definition of $u(x, \theta, \eta)$ are attained at points $\theta^u(\theta)$ that satisfy $\|\theta^u(\theta) - \theta\| \leq \eta$, and likewise for the infimum in the definition of $l(x, \theta, \eta)$. Hence $\lim_{\eta \to 0} |u(x, \theta, \eta) - q(\theta, x)| \to 0$ for every x and every $\theta \in \Theta$, and an analogous result holds for the lower brackets. We can integrate this limit by using the dominated convergence theorem (cf. Exercise 2) together with (14), so that we conclude

$$\lim_{\eta \to 0} E|u(X, \theta, \eta) - q(\theta, X)| \to 0 \text{ and } \lim_{\eta \to 0} E|l(X, \theta, \eta) - q(\theta, X)| \to 0$$

for every $\theta \in \Theta$. Consequently, for $\varepsilon > 0$ arbitrary and every $\theta \in \Theta$ we can find $\eta := \eta(\varepsilon, \theta)$ small enough such that

$$E|u(X,\theta,\eta) - l(X,\theta,\eta)| \leq E|u(X,\theta,\eta) - q(\theta,X)| + E|q(\theta,X) - l(X,\theta,\eta)| < \varepsilon. \tag{17}$$

The open balls $\{B(\theta, \eta(\varepsilon,\theta))\}_{\theta\in\Theta}$ constitute an open cover of the compact set Θ in $\mathbb{R}^p$, so by compactness there exists a finite subcover with centers $\theta_1, ..., \theta_N$, $j = 1, ..., N$ (the Heine-Borel theorem). The functions $q(\theta', \cdot)$ for $\theta' \in B(\theta_j, \eta(\varepsilon, j))$ are bracketed between $u_j := u(\cdot, \theta_j, \eta(\varepsilon, j))$ and $l_j := l(\cdot, \theta_j, \eta(\varepsilon, j))$, $j = 1, ..., N$, so that (16) and (17) complete the proof of (15) by invoking Proposition 1. □

First, Condition (14) *can not* be weakened: This follows from the fact that the limit (15) is a law of large numbers in the separable Banach space of continuous functions on Θ (cf. Example 1). Second, exactly the same proof works if (Θ, d) is *any* compact metric space. Third, for maximum likelihood estimation often the i.i.d. assumption is inconvenient, but the same 'bracketing' techniques work for dependent data as well, we refer to [66].

2.1.4 Consistency of the Maximum Likelihood Estimator

Putting the previous general results together, we can now derive a generic consistency result for maximum likelihood estimators under assumptions on the parametric model $\{f(\theta, \cdot) : \theta \in \Theta\}$ only.

Theorem 2. *Consider the model $f(\theta, y), \theta \in \Theta \subset \mathbb{R}^p, y \in \mathcal{Y} \subset \mathbb{R}^d$. Assume $f(\theta, y) > 0$ for all $y \in \mathcal{Y}$ and all $\theta \in \Theta$, and that $\int_{\mathcal{Y}} f(\theta, y)dy = 1$ for every $\theta \in \Theta$. Assume further that Θ is compact and that the map $\theta \mapsto f(\theta, y)$ is continuous on Θ for every $y \in \mathcal{Y}$. Let $Y_1, \ldots, Y_n$ be i.i.d. with common density $f(\theta_0)$, where $\theta_0 \in \Theta$. Suppose finally that the identification condition (13) and the domination condition*

$$\int \sup_{\theta'\in\Theta} |\log f(\theta', y)| f(\theta_0, y)dy < \infty$$

hold. If $\hat{\theta}_n$ is the MLE in the model $\{f(\theta, \cdot) : \theta \in \Theta\}$ based on the sample $Y_1, \ldots, Y_n$, then $\hat{\theta}_n$ is consistent, i.e.,

$$\hat{\theta}_n \to^{P_{\theta_0}} \theta_0 \quad \text{as } n \to \infty. \tag{18}$$

Proof. Setting $q(\theta, y) = -\log f(\theta, y), Q(\theta) = E_{\theta_0} q(\theta, Y), Q_n(\theta) = n^{-1} \sum_{i=1}^n q(\theta, Y_i)$, this follows from combining Theorem 1, Proposition 2 with what has been said in Example 5, and noting that continuity of Q follows from continuity of $\log f(y, \cdot)$ combined with the domination condition and the dominated convergence theorem (see Exercise 2). □

The first general result of this kind goes back to Wald (1949), who also realised that $\theta \mapsto f(\theta, y)$ only has to be upper semicontinuous in θ for it to hold, at the expense of a slightly more technical proof. Theorem 2 (and its proof) applies to families of discrete probability distributions line by line if one replaces probability densities $f(\theta, y)$ by probability mass functions $p(\theta, y)$, and integrals by sums. Consequently it can be applied to most parametric models for which maximum likelihood can be used.

A similar consistency result for nonlinear least squares estimators is part of the exercises. The theory for more general M-estimation procedures follows the same patterns, see [66, 67, 80].

A simple example to which Theorem 2 applies is the following.

Example 6. *[Exponential Families]* Consider the classical exponential family of order 1

$$f(\theta, y) = e^{\theta y - K(\theta)} f_0(y), \theta \in \Theta,$$

generated by some fixed density f_0, where $K(\theta)$ is the cumulant generating function of the model. Assume K is continuous on the compact set Θ. For instance if f_0 is the standard normal density, so that we are modelling a $N(\theta, 1)$ family, then $K(\theta) = \theta^2/2$, or if f_0 is Poisson with parameter $\lambda = 1$ then $K(\theta) = e^\theta - 1$. Then $\theta \mapsto f(\theta, y)$ is continuous for every y, and the domination condition reduces to

$$\begin{aligned} \int \sup_{\theta' \in \Theta} |\log f(\theta', y)| f(\theta, y) dy &= \int \sup_{\theta' \in \Theta} |(\theta' y - K(\theta')) + \log f_0(y)| f(\theta, y) dy < \infty \\ &\leq \sup_{\theta'} |\theta'| E_\theta |Y| + \sup_{\theta' \in \Theta} K(\theta')| + E_\theta |\log f_0(Y)| \end{aligned}$$

which is finite if $f(\theta)$ has a first moment and integrates $\log f_0$, since continuous K is bounded on the compact set $K(\theta)$. Thus Theorem 2 applies for compact Θ, and non-compact Θ can be dealt along the lines of Exercise 6.

2.1.5 Uniform Consistency

The above results show that one can find estimators that are consistent for every $\theta \in \Theta$ in the parameter space. For instance in Theorem 2, if $Y_1, \ldots, Y_n$ come from density $f(\theta)$ then $\hat{\theta}_n \to \theta$ in P_θ-probability for every $\theta \in \Theta$. A stronger requirement is consistency of an estimator T_n *uniformly* in $\theta \in \Theta$, that is, for every $\delta > 0$

$$\sup_{\theta_0 \in \Theta} P_{\theta_0}(\|T_n - \theta_0\| > \delta) \to 0 \quad \text{as } n \to \infty. \tag{19}$$

Inspection of the proof of Theorem 1 shows that sufficient conditions for this to be the case are that, for every $\varepsilon > 0$,

$$\inf_{\theta_0 \in \Theta} \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \varepsilon} (Q(\theta) - Q(\theta_0)) > 0$$

is satisfied and that the uniform law of large numbers in (10) holds uniformly under the law P_θ, that is, if

$$\sup_{\theta_0 \in \Theta} P_{\theta_0} \left(\sup_{\theta \in \Theta} |Q_n(\theta; Y_1, \ldots, Y_n) - Q(\theta)| > \delta \right) \to 0 \text{ as } n \to \infty$$

where the Y_i inside the probability are drawn from law P_{θ_0}. While the first is a reasonable analytic condition on the limiting criterion function Q that requires the identifiability of θ_0 to be 'uniform' in Θ, the second requires a little more thought: careful inspection of the proofs above, combined with the proof of the weak law of large numbers by Chebyshev's inequality, shows that a sufficient condition is

$$\sup_{\theta_0 \in \Theta} E_{\theta_0} \sup_{\theta \in \Theta} |Q(\theta, X)|^2 < \infty,$$

but weaker assumptions are possible. In fact classes of functions for which the uniform law of large numbers holds uniformly in a set of indexing laws P_θ were completely characterised in [30], see also Chapter 2.8 in [82] .

2.1.6 Exercises

Exercise 2. *[Dominated Convergence Theorem.]* Let f_n and g be real-valued functions defined on some measure space $(T, \mathcal{A}, \mu)$, and suppose f_n, g are μ-integrable. Assume $|f_n(x)| \leq g(x)$ and $f_n(x) \to f(x)$ as $n \to \infty$ for μ-almost every $x \in T$. Then f is μ-integrable and $\lim_n \int_T f_n(x) d\mu(x) = \int_T \lim_n f_n(x) d\mu(x) = \int_T f(x) d\mu(x)$.

Exercise 3. Derive an analogue of the consistency result for the MLE (Theorem 2) for the nonlinear least squares estimator with random design, under the assumptions that the $Y_i = (Z_i, x_i)$ are i.i.d., that $E(Z_i|x_i) = 0$, and that the parameter space Θ is compact. Which further assumptions do you need (be as economical as you can)? Show that the general normal linear model

$$Y = X\theta + u$$

with X a $n \times p$ matrix, $\theta \in \mathbb{R}^p$, $u \sim N(0, \sigma^2 I_n), \sigma^2 > 0$, is a special case of the NLS model from Example 2, and show further that the uniqueness condition for θ_0 is satisfied if the $n \times p$ matrix X has full column rank.

Exercise 4. A class of model functions of the form

$$f(\mu, \sigma^2, y) = a(\sigma^2, y) \exp \left\{ \frac{\zeta(\mu) y - K(\zeta(\mu))}{\sigma^2} \right\}, \; y \in \mathcal{Y}, \mu \in \mathcal{M}, \sigma^2 \in \Phi \subseteq (0, \infty)$$

where $a(\sigma^2, y)$ is a known positive function, is called an *exponential dispersion family* (of order 1). The parameter μ is the mean of this distribution, ζ is a suitable

real function defined on $\mathcal{M}$, and the parameter σ^2 is called the *dispersion parameter*, so that the full parameter is $\theta = (\mu, \sigma^2)$. Find $\zeta(\mu)$ and $K(\zeta(\mu))$ for normal, Poisson and binomial models. Restricting to compact Θ, formulate conditions on K, ζ so that the maximum likelihood estimator based on an i.i.d. sample from f is consistent, and verify them for the normal, Poisson and binomial case.

Exercise 5. Give an example of criterion functions Q_n, Q such

$$Q_n(\theta) \to^P Q(\theta)$$

for every $\theta \in \Theta$ as $n \to \infty$, that further satisfy all the conditions of Theorem 1 except for the uniform convergence condition (10), and for which $\hat{\theta}_n$ converges to a value different from the true value θ_0.

Exercise 6. Consider the problem of Exercise 3 above, but now with $\Theta = \mathbb{R}$. Assuming that

$$E\left[\inf_\theta \left(\frac{\partial g(x_t, \theta)}{\partial \theta}\right)^2\right] > 0,$$

show that one can find a compact set $\Theta^* = [\theta_0 - M, \theta_0 + M]$ such that

$$\inf_{\theta \notin \Theta^*} Q_n(\theta) > Q_n(\theta_0)$$

with probability approaching one, and use this to prove consistency of the NLS estimator. How does the condition on the derivative of g simplify in linear regression where $g(x_i, \theta) = x_i\theta$?

2.2 Asymptotic Distribution Theory

2.2.1 Asymptotic Normality of Maximum Likelihood Estimators

While a consistency result like the one from the previous section appears to be a minimal requirement for an estimator $\hat{\theta}_n$, it is not useful for statistical inference as it stands. For this to be the case, the accuracy of estimation $\hat{\theta}_n - \theta_0$ has to be quantified, and since $\hat{\theta}_n$ is random, this means that we would like to derive or approximate the distribution of the random fluctuations $\hat{\theta}_n - \theta_0$. In general not much can be said about this distribution for fixed sample size, as $\hat{\theta}_n$ depends non-linearly on the sample. One can, however, often obtain reasonable approximations of the distribution of $\hat{\theta}_n - \theta_0$ whose accuracy increases with sample size n. The study of such approximations constitutes the field of *asymptotic statistics*, and one of the most remarkable result in this theory is the *asymptotic normality and optimality of maximum likelihood estimators*, which holds in some nontrivial

universality, and which was first observed by R.A. Fisher ([33, 34]) in the 1920s, with first rigorous proofs due to Cramér [15].

Let us first sketch the main ideas behind the proof of asymptotic normality of MLEs. Consider $\Theta \subset \mathbb{R}$ for the moment. The derivative of Q_n at the maximiser $\hat{\theta}_n$ should equal zero, and the mean value theorem implies, assuming necessary regularity properties of Q_n

$$0 = Q_n'(\hat{\theta}_n) = Q_n'(\theta_0) + Q_n''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

so

$$\hat{\theta}_n - \theta_0 = -\frac{Q_n'(\theta_0)}{Q_n''(\tilde{\theta}_n)}.$$

Typically $EQ_n(\theta) = Q(\theta)$ and under regularity conditions also $EQ_n'(\theta_0) = Q'(\theta_0) = 0$ since we are thinking of θ_0 being a minimizer of Q. If $Q_n(\theta_0)$ is of sample mean form $\frac{1}{n}\sum_i q(\theta_0, Y_i)$ as before Proposition 2, then we can informally conclude

$$\sqrt{n}Q_n'(\theta_0) = \frac{1}{\sqrt{n}}\sum_i (q'(\theta_0, Y_i) - Eq'(\theta_0, Y_i)) \to^d N(0, \sigma^2) \text{ as } n \to \infty$$

by the central limit theorem (2), where $\sigma^2 = E(q'(\theta_0, Y))^2$. If $Q_n''(\tilde{\theta}_n)$ converges also to some limit in probability (typically to $Q''(\theta_0)$ under consistency of $\hat{\theta}_n$), then we deduce that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a non-degenerate normal distribution.

Let us now investigate these ideas rigorously, which requires some care, and which leads to some assumptions on the model $\{f(\theta) : \theta \in \Theta\}$ that at first look technical, but that are natural and shall be shown to be satisfied in most relevant examples.

When deriving the asymptotic distribution of MLEs, it is common to *assume* consistency of $\hat{\theta}_n$ in the formal statements. The reason is the following: Consistency is a statement about the global behaviour of $\hat{\theta}_n$ as an estimator of θ_0, and can be established under assumptions discussed in the prevous section. Asymptotic normality of the MLE is, in contrast, only a 'local' statement about the random fluctuations of $\hat{\theta}_n$ in $1/\sqrt{n}$-neighborhoods of θ_0, and it makes sense to separate 'local' and 'global' statements mathematically, as they require different (but compatible) sets of assumptions.

In the following theorem $\|\cdot\|$ stands, depending on the context, either for the Euclidean norm on $\mathbb{R}^p$ or for any matrix norm, and we shall write

$$\frac{\partial \log f(\theta_0, Y)}{\partial \theta} \text{ for } \frac{\partial \log f(\theta, Y)}{\partial \theta}_{|\theta=\theta_0},$$

and likewise for second derivatives, to simplify notation throughout.

Theorem 3. *Consider the model $f(\theta, y), \theta \in \Theta \subset \mathbb{R}^p, y \in \mathcal{Y} \subset \mathbb{R}^d$. Assume $f(\theta, y) > 0$ for all $y \in \mathcal{Y}$ and all $\theta \in \Theta$, and that $\int_{\mathcal{Y}} f(\theta, y)dy = 1$ for every $\theta \in \Theta$. Let $Y_1, \ldots, Y_n$ be i.i.d. from density $f(\theta_0, y)$ for some $\theta_0 \in \Theta$. Assume moreover*
i) that θ_0 is an interior point of Θ,
ii) that there exists an open set U satisfying $\theta_0 \in U \subset \Theta$ such that $f(\theta, y)$ is, for every $y \in \mathcal{Y}$, twice continuously differentiable w.r.t. θ on U,
iii) $E_{\theta_0}[\partial^2 \log f(\theta_0, Y)/\partial\theta\partial\theta^T]$ is nonsingular and

$$E_{\theta_0} \left\| \frac{\partial \log f(\theta_0, Y)}{\partial \theta} \right\|^2 < \infty,$$

iv) there exists a compact ball $K \subset U$ (with nonempty interior) centered at θ_0 s.t.

$$E_{\theta_0} \sup_{\theta \in K} \left\| \frac{\partial^2 \log f(\theta, Y)}{\partial\theta\partial\theta^T} \right\| < \infty,$$

$$\int_{\mathcal{Y}} \sup_{\theta \in K} \left\| \frac{\partial f(\theta, y)}{\partial \theta} \right\| dy < \infty \quad \text{and} \quad \int_{\mathcal{Y}} \sup_{\theta \in K} \left\| \frac{\partial^2 f(\theta, y)}{\partial\theta\partial\theta^T} \right\| dy < \infty.$$

Let $\hat{\theta}_n$ be the MLE in the model $\{f(\theta, \cdot); \theta \in \Theta\}$ based on the sample $Y_1, \ldots, Y_n$, and assume $\hat{\theta}_n \to^{P_{\theta_0}} \theta_0$ as $n \to \infty$. Define the Fisher information

$$i(\theta_0) := E_{\theta_0} \left[\frac{\partial \log f(\theta_0, Y)}{\partial \theta} \frac{\partial \log f(\theta_0, Y)}{\partial \theta}^T \right]. \tag{20}$$

Then $i(\theta_0) = -E_{\theta_0}[\partial^2 \log f(\theta_0, Y)/\partial\theta\partial\theta^T]$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to^d N(0, i^{-1}(\theta_0)), \quad \text{as } n \to \infty. \tag{21}$$

Proof. Let us note first that under the maintained assumptions

$$Q_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log f(\theta, Y_i)$$

is twice continuously differentiable on Θ and by using Exercise 7 and the first part of condition iv) we can differentiate under the integral sign to deduce that

$$Q(\theta) = -E_{\theta_0} \log f(\theta, Y)$$

is twice continuously differentiable in the interior of K.

We need another preliminary remark: $\int f(\theta, y)dy = 1$ for every $\theta \in \Theta$ implies

$$\frac{\partial}{\partial \theta} \int f(\theta, y)dy = 0 \quad \text{for every } \theta \in U$$

and by the second part of condition iv) and Exercise 7 we can interchange integration and differentiation with respect to θ in the interior of K to conclude

$$0 = \int \frac{\partial f(\theta, y)}{\partial \theta} dy = \int \frac{\partial \log f(\theta, y)}{\partial \theta} f(\theta, y) dy \quad \text{for every } \theta \in int(K). \tag{22}$$

Since $\theta_0 \in int(K)$ we thus have

$$E_{\theta_0}\left[\frac{\partial \log f(\theta_0, Y)}{\partial \theta}\right] = 0. \tag{23}$$

Since $\Pr(X_n \leq t) = \Pr(\{X_n \leq t\} \cap A_n) + \Pr(\{X_n \leq t\} \cap A_n^c)$ it suffices to prove a distributional limit theorem for X_n on events A_n whose probability approaches one. Since $\hat{\theta}_n \to^P \theta_0$ we infer that $\hat{\theta}_n$ is an interior point of Θ on events of probability approaching one as $n \to \infty$, and thus we must have

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = 0$$

on these sets (which we shall not repeat to mention in the proof). We can apply the mean value theorem to each component of this vector to obtain

$$0 = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + \bar{A}_n \sqrt{n}(\hat{\theta}_n - \theta_0) \tag{24}$$

where $\bar{A}_n$ is the matrix $(\partial^2/(\partial\theta\partial\theta^T))Q_n(\theta)$ of second derivatives of Q_n with the j-th row evaluated at a mean value $\bar{\theta}_{nj}$ on the line segment between θ_0 and $\hat{\theta}_n$.

Let us first study

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(\theta_0, Y_i)}{\partial \theta},$$

which is centred in view of (23). The central limit theorem (2) and existence of second moments from condition iii) thus imply

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \to^d N(0, i(\theta_0)) \quad \text{as } n \to \infty. \tag{25}$$

We next consider $\bar{A}_n$, and show that this matrix converges in probability to $-E[\partial^2 \log f(\theta_0, Y)/\partial\theta\partial\theta^T]$ as $n \to \infty$, for which it suffices to show convergence of each matrix component. The k-th entry in the j-th row of $\bar{A}_n$ is

$$\frac{1}{n} \sum_{i=1}^{n} h_{jk}(\bar{\theta}_{nj}, Y_i)$$

where h_{jk} is the second mixed partial derivative of $-\log f(\theta, Y_i)$ with respect to θ_j and θ_k, and we wish to show that each of these components converges to $Eh_{jk}(\theta_0, Y)$ in probability. We can write, noting that $\hat{\theta}_n \to^{P_{\theta_0}} \theta_0$ implies $\bar{\theta}_{nj} \to^{P_{\theta_0}} \theta_0$ and that $\bar{\theta}_{nj} \in K$ on events of probability approaching one,

$$\left| \frac{1}{n} \sum_{i=1}^{n} h_{jk}(\bar{\theta}_{nj}, Y_i) - Eh_{jk}(\bar{\theta}_{nj}, Y) + Eh_{jk}(\bar{\theta}_{nj}, Y) - Eh_{jk}(\theta_0, Y) \right|$$
$$\leq \sup_{\theta \in K} \left| \frac{1}{n} \sum_{i=1}^{n} h_{jk}(\theta, Y_i) - Eh_{jk}(\theta, Y) \right| + \left| Eh_{jk}(\bar{\theta}_{nj}, Y) - Eh_{jk}(\theta_0, Y) \right|,$$

where E denotes expectation w.r.t. Y only (and not w.r.t. $\bar{\theta}_{nj}$). Note next that twofold continuous differentiability of $\theta \mapsto f(\theta, y)$ implies, by the dominated convergence theorem and the first part of condition iv) (cf. Exercise 7), that the expected partial second derivatives $Eh_{jk}(\cdot, Y)$ are continuous on K. Using again the first part of condition iv) then verifies the conditions of the uniform (over K) law of large numbers in Proposition 2, so that the first quantity in the last display converges to zero. The continuity of $Eh_{jk}(\cdot, Y)$ and the fact that $\bar{\theta}_{nj} \to \theta_0$ in probability as $n \to \infty$ imply that the second quantity converges to zero in probability as well. [For the pedantic reader: One may show that the $\bar{\theta}_{nj}$ are measurable so that the above probability statements make sense. This extra argument is however not necessary, since measurable upper bounds for the expression in the last display converge to zero by the arguments just employed.] Thus

$$-\bar{A}_n \to^{P_{\theta_0}} E_{\theta_0} \left[\frac{\partial^2 \log f(\theta_0, Y)}{\partial \theta \partial \theta^T} \right] \equiv \Sigma(\theta_0) \text{ as } n \to \infty. \tag{26}$$

Since the limiting matrix is invertible we infer that $\bar{A}_n$ is invertible on sets whose probability approaches one, and we can thus rewrite (24) on these sets as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\bar{A}_n^{-1} \sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta} \to^d N\left(0, \Sigma^{-1}(\theta_0) i(\theta_0) \Sigma^{-1}(\theta_0)\right)$$

the limit following from (25), (26) and from Slutsky's lemma. This completes the proof if we show $\Sigma(\theta_0) = i(\theta_0)$, which is done as follows: Note that (22) implies

$$\frac{\partial}{\partial \theta^T} \int \frac{\partial f(\theta, y)}{\partial \theta} dy = 0 \quad \text{for every } \theta \in int(K)$$

and by the third part of condition iv) and Exercise 7 we can interchange integration and differentiation to deduce

$$\int \frac{\partial^2 f(\theta, y)}{\partial \theta \partial \theta^T} dy = 0 \quad \text{for every } \theta \in int(K). \tag{27}$$

The chain rule implies, for every $\theta \in U$,

$$\begin{aligned}\frac{\partial^2 \log f(\theta,y)}{\partial\theta\partial\theta^T} &= \frac{1}{f(\theta,y)}\frac{\partial^2 f(\theta,y)}{\partial\theta\partial\theta^T} - \frac{1}{f^2(\theta,y)}\frac{\partial f(\theta,y)}{\partial\theta}\frac{\partial f(\theta,y)}{\partial\theta}^T \\ &= \frac{1}{f(\theta,y)}\frac{\partial^2 f(\theta,y)}{\partial\theta\partial\theta^T} - \frac{\partial \log f(\theta,y)}{\partial\theta}\frac{\partial \log f(\theta,y)}{\partial\theta}^T .\end{aligned}$$

Using this identity at θ_0 and integrating it with respect to $f(\theta_0, y)$ combined with (27) implies $\Sigma(\theta_0) = i(\theta_0)$, and thus completes the proof. □

Theorem 3 can be readily used for asymptotic inference with the maximum likelihood estimator, as soon as we can estimate the Fisher information $i(\theta_0)$ consistently. This is discussed in Exercises 9 and 10.

2.2.2 Asymptotic Efficiency, Plug-in MLEs and the Delta-Method

Theorem 3 establishes a way to use the maximum likelihood estimator for asymptotic inference. The question arises as to whether one can find estimators for θ that are better than the MLE. While an improvement of the rate of convergence $1/\sqrt{n}$ cannot be expected in regular parametric models, one can still ask whether the asymptotic variance of the estimator is the smallest possible one. This question has more than one answer, and we do not attempt to provide a rigorous derivation of these results, but rather discuss one of them briefly.

A first basic observation is the following result, known as the Cramèr-Rao lower bound, which we give, for simplicity, in the one-dimensional situation $p = 1$.

Proposition 3. *In the framework of Theorem 3 with $p = 1$ and for $n \in \mathbb{N}$ fixed, let $\tilde{\theta} = \tilde{\theta}(Y_1, \ldots, Y_n)$ be any unbiased estimator of θ, i.e., one that satisfies $E_\theta\tilde{\theta} = \theta$ for all $\theta \in \Theta$. Then*

$$Var_\theta(\tilde{\theta}_n) \geq \frac{1}{ni(\theta)} \quad \forall\theta \in int(\Theta).$$

Proof. Write $y = (y_1, \ldots, y_n)$ in slight abuse of notation and set

$$l'(\theta, Y) \equiv \frac{d}{d\theta}\log\prod_{i=1}^n f(\theta, Y_i) = \sum_{i=1}^n \frac{d}{d\theta}\log f(\theta, Y_i).$$

We have, by the Cauchy-Schwarz inequality and (23),

$$Var_\theta(\tilde{\theta}) \geq \frac{Cov_\theta^2(\tilde{\theta}, l'(\theta, Y))}{Var_\theta(l'(\theta, Y))} = \frac{1}{ni(\theta)},$$

since, interchanging differentiation and integration,

$$Cov_\theta(\tilde{\theta}, l'(\theta, Y)) = \int \tilde{\theta}(y) l'(\theta, y) \prod_{i=1}^{n} f(\theta, y_i) dy$$
$$= \int \tilde{\theta}(y) \frac{d}{d\theta} f(\theta, y) dy = \frac{d}{d\theta} E_\theta \tilde{\theta} = \frac{d}{d\theta} \theta = 1.$$

□

A multi-dimensional extension ob the above proposition is straightforward to obtain if one agrees to say that a symmetric matrix A is greater than or equal to another symmetric matrix B if $A - B$ is positive semi-definite.

The above result can be used to informally justify the maximum likelihood estimator among all asymptotically unbiased estimators, a class very similar in nature to all consistent estimators. The following example however shows that difficulties can be expected when attempting to make this intuition rigorous.

Example 7. *[Hodges' estimator.]* In a parametric model satisfying the conditions of Theorem 3, and with $\Theta = \mathbb{R}$ for simplicity, let $\hat{\theta}_n$ be the maximum likelihood estimator. Consider the alternative estimator

$$\tilde{\theta}_n = \hat{\theta}_n 1\{|\hat{\theta}_n| \geq n^{-1/4}\}$$

that is thresholded to zero whenever the MLE does not exceed $n^{-1/4}$ in absolute value. Suppose first the Y_i's are drawn from $P_\theta, \theta \neq 0$. Let n large enough such that $|\theta| - n^{-1/4} \geq |\theta|/2 > 0$, then by the triangle inequality and consistency of $\hat{\theta}_n$ we have

$$P_\theta(\tilde{\theta}_n \neq \hat{\theta}_n) \leq P_\theta(|\hat{\theta}_n - \theta + \theta| < n^{-1/4})$$
$$\leq P_\theta(|\theta| - n^{-1/4} < |\hat{\theta}_n - \theta|)$$
$$\leq P_\theta(|\hat{\theta}_n - \theta| > |\theta|/2) \to 0$$

as $n \to \infty$, and hence $\sqrt{n}(\tilde{\theta} - \theta) \to^d N(0, i^{-1}(\theta))$ under $P_\theta, \theta \neq 0$. When $\theta = 0$, however, for any $t \in \mathbb{R}$,

$$P_0\left(\sqrt{n}(\tilde{\theta}_n - \theta) \leq t\right) = P_0\left(0 \leq t, \tilde{\theta}_n = 0\right) + P_0\left(\sqrt{n}(\tilde{\theta}_n - \theta) \leq t, \tilde{\theta}_n \neq 0\right).$$

We have

$$P_0(\tilde{\theta}_n \neq 0) = P_0(|\hat{\theta}_n| \geq n^{-1/4}) = P_0(\sqrt{n}|\hat{\theta}_n - \theta| \geq n^{1/4}) \to 0$$

as $n \to \infty$ since $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal (and hence stochastically bounded), in view of Theorem 3. Conclude from the last but one display that under P_0 we have $\sqrt{n}(\tilde{\theta}_n - \theta) \to^d \delta_0 = N(0, 0)$, hence the estimator $\tilde{\theta}$ has an asymptotic covariance that strictly dominates, at $\theta = 0$, the asymptotic variance of the maximum likelihood estimator.

To rule out estimators as the one above we can either restrict to 'regular' estimators, or invoke the minimax principle. Consider a parametric model $\{f(\theta) : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p$, that satisfies the regularity conditions of Theorem 3, and suppose we are interested in making inference on $\Phi(\theta)$, where $\Phi : \Theta \to \mathbb{R}^m$ is differentiable at θ, possibly $\Phi = id, m = p$. Let $Y_1, \ldots, Y_n$ be i.i.d. from density $f(\theta)$ and let T_n be *any* estimator for $\Phi(\theta)$, i.e., any measurable function of a sample of size n. Then, for any bowl-shaped loss function ℓ (i.e., any nonnegative function for which the sets $\{x : \ell(x) \leq c\}$ are convex and symmetric about the origin), and for every $\theta \in int(\Theta)$, one can show

$$\sup_I \liminf_n \sup_{h \in I} E_{\theta + \frac{h}{\sqrt{n}}} \ell\left(\sqrt{n}(T_n - \Phi\left(\theta + \frac{h}{\sqrt{n}}\right)\right) \geq E_{N(0,\Sigma(\Phi,\theta))}\ell(X) \tag{28}$$

where the first supremum runs over all finite subsets $I \subset \mathbb{R}^p$, and where the asymptotic variance equals the $m \times m$ matrix

$$\Sigma(\Phi, \theta) = \frac{\partial \Phi(\theta)}{\partial \theta} i^{-1}(\theta) \frac{\partial \Phi(\theta)}{\partial \theta}^T.$$

This means that if we scale the risk of any estimator T_n for $\Phi(\theta)$ by $\sqrt{n}$, then T_n cannot have smaller asymptotic risk than $E_{N(0,\Sigma(\Phi,\theta))}\ell(X)$ uniformly in neighborhoods of θ that shrink at rate $1/\sqrt{n}$, so in particular not uniformly in all of Θ. See Chapter 8.7 in [81] for a proof. In the case where $\Phi = id$ the asymptotic covariance $\Sigma(\Phi, \theta)$ simplifies to $\Sigma(\theta)$ from Theorem 3 and shows that the maximum likelihood estimator is asymptotically efficient from a local minimax point of view in that it attains the above lower bound. Moreover the Hodges' estimator from above can be shown to have minimax risk equal to infinity, see Example 8 below.

If Φ is not equal to the identity function this gives a lower bound on the behaviour of any estimator of the functional $\Phi(\theta)$ defined on Θ, and it is a natural question to compare this lower bound to the asymptotic performance of the plug-in maximum likelihood estimator $\Phi(\hat{\theta}_n)$. The following general result, which we shall prove in Section 3.7.1 below in even more generality, is known as the *Delta-method*. It implies in particular that a plug-in MLE is also asymptotically efficient for estimating $\Phi(\theta)$.

Proposition 4. *Let Θ be an open subset of $\mathbb{R}^p$ and let $\Phi : \Theta \to \mathbb{R}^m$ be differentiable at $\theta \in \Theta$, with derivative $D\Phi_\theta$. Let r_n be a divergent sequence of positive real numbers and let X_n be random variables taking values in Θ such that $r_n(X_n - \theta) \to^d X$ as $n \to \infty$. Then*

$$r_n(\Phi(X_n) - \Phi(\theta)) \to^d D\Phi_\theta(X)$$

as $n \to \infty$. If $X \sim N(0, i^{-1}(\theta))$ then

$$D\Phi_\theta(X) \sim N(0, \Sigma(\Phi, \theta)).$$

The result should not come as a surprise, in particular since one can show that the maximum likelihood estimator is preserved under transformations, that is, $\Phi(\hat{\theta}_n)$ equals the MLE in the model $\{\Phi(\theta) : \theta \in \Theta\}$ based on the sample, a result that even holds when Φ is not differentiable. See Exercise 12 for this fact.

2.2.3 Parametric Testing Theory

Suppose we observe $Y_1, \ldots, Y_n$ from a density $f(\theta, \cdot)$ and consider the testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta$$

where $\Theta_0 \subset \Theta \subset \mathbb{R}^p$. The Neyman-Pearson theory suggests to test these hypotheses by the likelihood ratio test statistic

$$\Lambda_n(\Theta, \Theta_0) := 2 \log \frac{\sup_{\theta \in \Theta} \prod_{i=1}^n f(\theta, Y_i)}{\sup_{\theta \in \Theta_0} \prod_{i=1}^n f(\theta, Y_i)} \tag{29}$$

which in terms of the maximum likelihood estimators $\hat{\theta}_n, \hat{\theta}_{n,0}$ of the models Θ, Θ_0 equals

$$\Lambda_n(\Theta, \Theta_0) = 2 \log \frac{\prod_{i=1}^n f(\hat{\theta}_n, Y_i)}{\prod_{i=1}^n f(\hat{\theta}_{n,0}, Y_i)} = -2 \sum_{i=1}^n (\log f(\hat{\theta}_{n,0}, Y_i) - \log f(\hat{\theta}_n, Y_i)).$$

If the null-hypothesis is simple, $\Theta_0 = \{\theta_0\}$, then $\hat{\theta}_{n,0} = \theta_0$, and a first key result is the following.

Theorem 4. *Consider a parametric model $f(\theta, y), \theta \in \Theta \subset \mathbb{R}^p$, that satisfies the assumptions of Theorem 3. Consider the simple null hypothesis $\Theta_0 = \{\theta_0\}, \theta_0 \in \Theta$. Then, under H_0, the likelihood ratio test statistic is asymptotically chi-square distributed, i.e.,*

$$\Lambda_n(\Theta, \Theta_0) \to^d \chi^2_p \quad \text{as } n \to \infty \text{ under } P_{\theta_0} \tag{30}$$

where χ^2_p is a chi-square random variable with p degrees of freedom.

Proof. Using the notation from the proof of Theorem 3 we see that $\Lambda_n(\Theta, \Theta_0) = 2nQ_n(\theta_0) - 2nQ_n(\hat{\theta}_n)$, which we can expand into a Taylor series about $\hat{\theta}_n$, up to second order,

$$\begin{aligned} \Lambda_n(\Theta, \Theta_0) &= 2nQ_n(\theta_0) - 2nQ_n(\hat{\theta}_n) \\ &= 2n \frac{\partial Q_n(\hat{\theta}_n)^T}{\partial \theta} (\theta_0 - \hat{\theta}_n) + n(\theta_0 - \hat{\theta}_n)^T \frac{\partial^2 Q(\bar{\theta}_n)}{\partial \theta \partial \theta^T} (\theta_0 - \hat{\theta}_n) \end{aligned}$$

for a vector $\bar{\theta}_n$ on the line segment between $\hat{\theta}_n$ and θ_0. The first term in the last line equals zero on sets of probability approaching one since $\hat{\theta}_n$ is eventually an

interior minimizer of Q_n. Using the uniform law of large numbers in Proposition 2 one shows, as in (26), that the matrix $\bar{A}_n$ of second negative log-likelihood derivatives evaluated at $\bar{\theta}_n$ converges to $i(\theta_0)$ in probability. By Theorem 3 and Slutsky's lemma we deduce that $\sqrt{n}(\hat{\theta}_n - \theta_0)^T(\bar{A}_n - i(\theta_0))$ converges to zero in distribution and then also in probability (as the limit is constant), and by repeating this argument we deduce that

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T(\bar{A}_n - i(\theta_0))\sqrt{n}(\hat{\theta}_n - \theta_0) \to^{P_{\theta_0}} 0 \text{ as } n \to \infty.$$

Consequently $\Lambda_n(\Theta, \Theta_0)$ has the same limiting distribution under P_{θ_0} as the random variable

$$\sqrt{n}(\hat{\theta}_n - \theta_0)^T i(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0).$$

Since the mapping $x \mapsto x^T i(\theta_0)x$ is continuous from $\mathbb{R}^p$ to $\mathbb{R}$ we obtain from Theorem 3 and the continuous mapping theorem that this limiting distribution equals $X^T i(\theta_0)X$ where $X \sim N(0, i^{-1}(\theta_0))$. This equals the squared Euclidean norm of a multivariate standard normal $N(0, I_p)$ vector, which has a χ^2_p distribution, and completes the proof. □

Note that the above result can be extended to composite null hypotheses Θ_0 with dimension $p_0 < p$, the limit being χ^2_q with q degrees of freedom, where $q = p - p_0$. We do not pursue this further here, as the modifications are mostly of a technical nature. See Chapter 16 in [81] for details.

Let us instead ask a more abstract question about testing parametric hypotheses, that will be useful later on, but is also of separate interest. By definition we shall say that a test ϕ_n is a random indicator function depending on the sample $Y_1, \ldots, Y_n$ that takes values in $\{0, 1\}$, so accepts a null hypothesis H_0 if $\phi_n = 0$ and rejects it otherwise. Theorem 4 implies that we can design a test that is consistent under the null-hypothesis, i.e., $E_{\theta_0}\phi_n \to 0$ as $n \to \infty$. However, for the test to be informative one also has to ask for the behaviour of the test under alternatives $\theta \neq \theta_0, \theta \in \Theta$. Ideally we would want a test such that, for every $\delta > 0$,

$$E_{\theta_0}\phi_n \to 0 \text{ and } \sup_{\theta:\|\theta-\theta_0\|>\delta} E_\theta(1 - \phi_n) \to 0 \text{ as } n \to \infty. \tag{31}$$

The existence of such tests can be verified in any parametric model $\{f(\theta) : \theta \in \Theta\}$ in which uniformly consistent estimators T_n for θ exist.

Lemma 1. *Let $\{f(\theta) : \theta \in \Theta\}$ be a parametric model in which uniformly consistent estimators $T_n = T(Y_1, \ldots Y_n)$ in the sense of (19) exist. Then there exist tests $\phi_n \equiv \phi(Y_1, \ldots, Y_n, \theta_0)$ for $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \setminus \{\theta_0\}$ such that for every θ satisfying $\|\theta - \theta_0\| > \delta > 0$, for some universal constant C,*

$$\max\left(E_{\theta_0}\phi_n, E_\theta(1 - \phi_n)\right) \leq e^{-Cn}.$$

Proof. We can assume n large enough as otherwise the bound is trivial for a sufficiently small constant C. We first show that consistent tests exist: Set

$$\psi_n = 1\{\|T_n - \theta_0\| \geq \delta/2\}$$

which converges to 0 under P_{θ_0} and also satisfies, using the triangle inequality,

$$\begin{aligned}\sup_{\theta:\|\theta-\theta_0\|\geq\delta} E_\theta(1-\psi_n) &\leq \sup_{\theta:\|\theta-\theta_0\|\geq\delta} P_\theta^n(\|T_n-\theta\| > \|\theta-\theta_0\| - \delta/2) \\ &\leq \sup_{\theta:\|\theta-\theta_0\|\geq\delta} P_\theta^n(\|T_n-\theta\| > \delta/2) \to 0 \quad \text{as } n\to\infty.\end{aligned}$$

To establish the exponential bound, fix $k \in \mathbb{N}$ such that $E_{\theta_0}\psi_k$ and $E_\theta(1-\psi_k)$ are both less than $1/4$ for every θ that satisfies $\|\theta-\theta_0\| > \delta$. For $n = mk+r, m \in \mathbb{N}, 0 \leq r < k$ order the sample into groups of size k, $(X_1,\ldots,X_k),(X_{k+1},\ldots,X_{2k}),\ldots$, set $Y_{nj} = \psi_k(X_{k(j-1)+1},\ldots,X_{kj})$ and define the sample average $\bar{Y}_{nm} = m^{-1}\sum_{j=1}^m Y_{nj}$. Define new tests $\phi_n = 1\{\bar{Y}_{nm} \geq 1/2\}$. Since $E_\theta Y_{nj} \geq 3/4$ we can use Hoeffding's inequality (3) to obtain

$$E_\theta(1-\phi_n) = P_\theta(\bar{Y}_{nm} < 1/2) \leq e^{-2m\left(\frac{3}{4}-\frac{1}{2}\right)^2} \leq e^{-m/8}.$$

Since $m \sim n$ this proves the desired exponential decay under the alternative. Since $E_{\theta_0}Y_{nj} \leq 1/4$ the same proof applies under the null hypothesis, which establishes the same exponential bound for $E_{\theta_0}\phi_n$. □

While in (31) we asked for tests that are globally consistent for alternatives bounded away from θ_0, one can further ask the 'local' question about the performance of tests against local alternatives $\theta_0 + M/\sqrt{n}$ that may be as close to θ_0 as a multiple of $1/\sqrt{n}$. At least for M large enough such local alternatives can still be distinguished consistently, as the following result shows. We only consider alternatives that are 'close', $\|\theta-\theta_0\| < \delta$ for $\delta > 0$ small as $\|\theta-\theta_0\| > \delta > 0$ has been dealt with above.

Lemma 2. *Let $\{f(\theta) : \theta \in \Theta\}$ be a parametric model that satisfies the conditions of Theorem 4. Let $M_n \to \infty$ as $n \to \infty$. Then there exist tests ϕ_n for $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \setminus \{\theta_0\}$ such that*

$$E_{\theta_0}\phi_n \to 0 \quad \text{as } n \to \infty$$

and, for every θ satisfying $M_n/\sqrt{n} < \|\theta-\theta_0\| \leq \delta$ for some $\delta < 1$, we have for some universal constant D,

$$E_\theta(1-\phi_n) \leq \frac{1}{D}e^{-Dn\|\theta-\theta_0\|^2} \to 0 \quad \text{as } n\to\infty.$$

Proof. We shall again use the notation from the proof of Theorem 3, particularly $q(\theta, y) = -\log f(\theta, y)$, and we shall prove the result under the additional assumption that $\partial q(\theta_0, y)/\partial\theta$ is a bounded function in y (otherwise a simple truncation argument can be used). Define

$$\phi_n = 1\left\{\left\|\frac{\partial Q_n(\theta_0)}{\partial\theta}\right\| \geq \sqrt{M_n/n}\right\},$$

a quantity known as the *score test*. We shall use here for simplicity the maximum norm

$$\|v\| = \max_{i=1,\ldots,p} |v_i|,$$

but any other (equivalent) norm on $\mathbb{R}^p$ works as well. By (25), Prohorov's theorem and since $M_n \to \infty$, this quantity converges to zero in probability under P_{θ_0}, so verifies the first conclusion of the lemma. Under any alternative, using the triangle inequality, (23), a Taylor expansion up to second order and the 'regularity' conditions of Theorem 3, including invertibility of $i(\theta_0)$, we can lower bound $\left\|\frac{\partial Q_n(\theta_0)}{\partial\theta}\right\|$ by

$$\begin{aligned}
&\left\|E_\theta \frac{\partial q(\theta_0, Y)}{\partial\theta} - E_{\theta_0}\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| - \left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| \\
&= \left\|\int\left(\frac{\partial q(\theta_0, y)}{\partial\theta}\frac{\partial f(\theta_0, y)^T}{\partial\theta} + o(\|\theta - \theta_0\|)\right) dy(\theta - \theta_0)\right\| - \left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| \\
&= \|(i(\theta_0) + o(\|\theta - \theta_0\|))(\theta - \theta_0)\| - \left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| \\
&\geq c\|\theta - \theta_0\| - \left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\|
\end{aligned}$$

where $c > 0$ is some constant. Therefore, since $\|\theta - \theta_0\| \geq M_n/\sqrt{n}$ we have for n large enough

$$\begin{aligned}
E_\theta(1 - \phi_n) &= P_\theta\left(\left\|\frac{\partial Q_n(\theta_0)}{\partial\theta}\right\| < \sqrt{M_n/n}\right) \\
&\leq P_\theta\left(\left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| > c\|\theta - \theta_0\| - \sqrt{\frac{M_n}{n}}\right) \\
&\leq P_\theta\left(\left\|\frac{\partial Q_n(\theta_0)}{\partial\theta} - E_\theta\frac{\partial q(\theta_0, Y)}{\partial\theta}\right\| > \frac{c}{2}\|\theta - \theta_0\|\right) \leq D^{-1}e^{-Dn\|\theta - \theta_0\|^2}
\end{aligned}$$

by Hoeffding's inequality (3), recalling the assumption that $\partial q(\theta_0, y)/\partial\theta$ is bounded and using $\Pr(\max_{i=1,\ldots,p}|v_i| > u) \leq p\max_{i=1,\ldots,p}\Pr(|v_i| > u)$. □

While the previous lemmata only prove existence of such tests, some more detailed analysis shows that the score test used in the previous proof also works in Lemma 1, under suitable conditions on $\{f(\theta) : \theta \in \Theta\}$ like the ones discussed after (19), so that this gives a concrete example of a test that works. Likelihood ratio tests can also be used, but we do not pursue this further here.

2.2.4 Local Asymptotic Normality and Contiguity

The results from the previous subsection on parametric testing theory allow to reduce the search for the true value θ_0 of θ to a $M/\sqrt{n}$ neighborhood of θ_0. Once we have 'zoomed in' to this neighborhood, inference on θ_0 starts to resemble the structure of a simple Gaussian experiment, and this phenomenon is often referred to as 'local asymptotic normality'.

Suppose we observe a 'Gaussian shift experiment' given by the single normal observation $X \sim N(g, i^{-1}(\theta))$ with unknown shift $g \in \mathbb{R}^p$, and consider the likelihood ratio between a $N(h, i^{-1}(\theta))$ model and a $N(0, i^{-1}(\theta))$ model

$$\log \frac{dN(h, i^{-1}(\theta))}{dN(0, i^{-1}(\theta))}(X) = h^T i(\theta) X - \frac{1}{2} h^T i(\theta) h. \tag{32}$$

So the local difference between the relevant likelihood ratios is again a normally distributed random variable. This motivates the following definition.

Definition 1 (Local Asymptotic Normality (LAN))**.** *Consider a parametric model*

$$f(\theta) \equiv f(\theta, \cdot), \theta \in \Theta \subset \mathbb{R}^p,$$

and let $q(\theta, y) = -\log f(\theta, y)$*. Suppose* $(\partial/\partial\theta) q(\theta_0, y)$ *and the Fisher information* $i(\theta_0)$ *exist at the interior point* $\theta_0 \in \Theta$*. We say that the model* $\{f(\theta) : \theta \in \Theta\}$ *is locally asymptotically normal at* θ_0 *if for every convergent sequence* $h_n \to h$ *and for* $Y_1, \ldots, Y_n$ *i.i.d.*$\sim f(\theta_0)$ *we have, as* $n \to \infty$*,*

$$\log \prod_{i=1}^{n} \frac{f\left(\theta_0 + h_n/\sqrt{n}\right)}{f(\theta_0)}(Y_i) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \frac{\partial q(\theta_0, Y_i)}{\partial \theta} - \frac{1}{2} h^T i(\theta_0) h + o_{P_{\theta_0}}(1).$$

We say that the model $\{f(\theta) : \theta \in \Theta\}$ *is locally asymptotically normal if it is locally asymptotically normal for every* $\theta \in int(\Theta)$*.*

In words local asymptotic normality means the following: If the Y_i are drawn from $f(\theta)$ for some θ in the interior of Θ, then the ratio between the likelihood $\Pi_{i=1}^n f(\theta + h/\sqrt{n}, Y_i)$ of the local 'alternative' $\theta + h/\sqrt{n}$ and the likelihood $\Pi_{i=1}^n f(\theta, Y_i)$ of the true parameter θ admits an asymptotic approximation by a

random variable whose limit distribution under P_θ is the Gaussian variable occurring on the right hand side in (32) when $g = 0$. 'Local' refers here to $1/\sqrt{n}$ neighborhoods of θ.

Intuitively speaking, statistical inference in a LAN model is asymptotically locally equivalent to inference in a Gaussian shift experiment with shift g, and the parametric models we have dealt with so far are locally asymptotically normal, as the following proposition shows.

Proposition 5. *Consider a parametric model $\{f(\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}^p$, that satisfies the assumptions of Theorem 3. Then $\{f(\theta) : \theta \in \Theta_0\}$ is locally asymptotically normal for every open subset Θ_0 of Θ.*

Proof. We only prove $h_n = h$ fixed, the proof for $h_n \to h$ follows analogously. Expanding $\log f(\theta_0 + h/\sqrt{n})$ about $\log f(\theta_0)$ up to second order as in the proof of Theorem 4 we see that the likelihood ratio equals

$$-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} h^T \frac{\partial q(\theta_0, Y_i)}{\partial \theta} - \frac{1}{2n} h^T \sum_{i=1}^{n} \frac{\partial^2 q(\bar{\theta}_n, Y_i)}{\partial \theta \partial \theta^T} h$$

for some vector $\bar{\theta}_n$ lying on the line segment between θ_0 and $\theta_0 + h/\sqrt{n}$. Using the uniform law of large numbers Proposition 2 one shows, as in (26), that

$$\frac{1}{2n} h^T \sum_{i=1}^{n} \frac{\partial^2 q(\bar{\theta}_n, Y_i)}{\partial \theta \partial \theta^T} h - \frac{1}{2} h^T i(\theta_0) h \to^{P_{\theta_0}} 0 \text{ as } n \to \infty$$

which completes the proof. □

Le Cam [52, 53] developed the notion of local asymptotic normality as a unifying notion of much of asymptotic parametric statistics. It can be seen as the 'statistical equivalent' of the regularity assumptions from Theorem 3. Assuming local asymptotic normality together with 'differentiability in quadratic mean' of $\theta \mapsto f(\theta)$ is an alternative route to derive the asymptotic distribution of maximum likelihood estimators, see Section 7.4 in [81]. The assumption of differentiability in quadratic mean is slightly weaker than the assumptions we imposed in Theorem 3, but for most relevant parametric models these assumptions are equivalent, and then the present approach avoids some technicalities. Proposition 5 highlights, however, the close connection between these approaches.

Local asymptotic normality is not only another assumption, but also a key concept to derive further properties in parametric models, often in conjunction with the following concept.

Definition 2 (Contiguity.)**.** *Let P_n, Q_n be two sequences of probability measures. We say that Q_n is contiguous with respect to P_n if for every sequence of measurable*

sets A_n the hypothesis $P_n(A_n) \to 0$ as $n \to \infty$ implies $Q_n(A_n) \to 0$ as $n \to \infty$, and write $Q_n \lhd P_n$. The sequences P_n, Q_n are mutually contiguous if both $Q_n \lhd P_n$ and $P_n \lhd Q_n$, which is denoted by $P_n \lhd\rhd Q_n$.

A useful way to take advantage of contiguity is the following lemma, which is due to Le Cam. For two probability measures P, Q the ratio dP^a/dQ is the Radon-Nikodym density of the absolutely continuous part P^a of P with respect to Q, and we can write

$$P^a(A) = \int_A \frac{dP^a}{dQ}(x) dQ(x)$$

for every measurable set A. In statistics the random variable $(dP/dQ)(X) \equiv (dP^a/dQ)(X)$ (with a suppressed) where X has law Q is usually referred to as the likelihood ratio, and by the usual convention in measure theory ($\infty \cdot 0 = 0$) we may leave the quotient undefined when $dQ = 0$. Moreover the notation $dP_n/dQ_n \to^d_{Q_n} U$ shall be used to denote $dP_n/dQ_n(X_n) \to^d U$ for random variables X_n that have distribution Q_n.

Lemma 3. *[Le Cam's first lemma] Let P_n, Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Then the following are equivalent.*
i) $Q_n \lhd P_n$.
ii) If $dP_n/dQ_n \to^d_{Q_n} U$ along a subsequence of n, then $P(U > 0) = 1$.
iii) If $dQ_n/dP_n \to^d_{P_n} V$ along a subsequence of n, then $EV = 1$.
iv) For any sequence of statistics (measurable functions) $T_n : \Omega_n \to \mathbb{R}^k$ we have: $T_n \to^{P_n} 0$ as $n \to \infty$ implies $T_n \to^{Q_n} 0$ as $n \to \infty$.

Proof. The equivalence of (i) and (iv) is clear: Assuming contiguity we can take sets $A_n = \{\|T_n\| > \varepsilon\}$ so that $Q_n(A_n) \to 0$ means $T_n \to^{Q_n} 0$. Conversely given sets A_n take statistics $T_n = 1_{A_n}$.
i) $\to$ ii): In abuse of notation denote the subsequence of n again by n. Define $g_n(\varepsilon) = Q_n(dP_n/dQ_n < \varepsilon) - P(U < \varepsilon)$. By the portmanteau theorem $\liminf_n g_n(\varepsilon) \geq 0$, and we can find $\varepsilon_n \searrow 0$ such that also $\liminf_n g_n(\varepsilon_n) \geq 0$, so

$$P(U = 0) = \lim_n P(U < \varepsilon_n) \leq \liminf_n Q_n \left(\frac{dP_n}{dQ_n} < \varepsilon_n \right).$$

On the other hand

$$P_n \left(\frac{dP_n}{dQ_n} \leq \varepsilon_n, dQ_n > 0 \right) = \int_{dP_n/dQ_n \leq \varepsilon_n} \frac{dP_n}{dQ_n} dQ_n \leq \int \varepsilon_n dQ_n \to 0.$$

By contiguity and equivalence of (i) and (iv) the P_n probability on the left in the last line thus converges to zero also under Q_n, so that the right hand side in the last but one display equals zero, so $P(U = 0) = 0$.

(iii) $\to$ (i): If $P_n(A_n) \to 0$ then $1_{\Omega_n \setminus A_n}$ converges to one under P_n. By Prohorov's theorem we can find a subsequence of n along which the random vector $(dQ_n/dP_n, 1_{\Omega_n - A_n}) \to^d (V, 1)$ under P_n. The mapping $(v,t) \to vt$ is continuous and nonnegative on the set $[0,\infty) \times \{0,1\}$, so the portmanteau lemma gives

$$\liminf_n Q_n(\Omega_n \setminus A_n) \geq \liminf \int 1_{\Omega_n \setminus A_n} \frac{dQ_n}{dP_n} dP_n \geq E1 \cdot V = 1,$$

so the left hand side has to converge to one, i.e., $Q_n(A_n) \to 0$ as well.
ii) $\to$ iii): Define the probability measures

$$\mu_n = \frac{P_n + Q_n}{2}$$

which dominate both P_n and Q_n in the sense that $\mu_n(A) = 0$ implies $P_n(A) = Q_n(A) = 0$. If p_n, q_n are the densities of P_n and Q_n, respectively, with respect to μ_n, then $p_n + q_n = 2$ μ_n-almost everywhere and so p_n and q_n take values in the interval $[0,2]$. Conclude that $dP_n/d\mu_n$ is uniformly tight with respect to μ_n, and using the hypotheses of ii), iii) we can thus find a subsequence of n that we still denote by n such that

$$\frac{dP_n}{dQ_n} \to^d_{Q_n} U, \quad \frac{dQ_n}{dP_n} \to^d_{P_n} V, \quad W_n \equiv \frac{dP_n}{d\mu_n} \to^d_{\mu_n} W$$

for some random variables U, V, W and W_n with $E_{\mu_n} W_n = \int dP_n = 1$ for every n. Since the densities p_n are uniformly bounded we can use dominated convergence to infer $EW_n \to EW$ as $n \to \infty$, so $EW = 1$ as well. For given bounded and continuous f define $g : [0,2] \to \mathbb{R}$ as $g(w) = f(w/(2-w))(2-w)$ if $0 \leq w < 2$ and $g(2) = 0$, which is again bounded and continuous. Now clearly $dP_n/dQ_n = W_n/(2 - W_n)$ and $dQ_n/d\mu_n = 2 - W_n$ so by the portmanteau lemma

$$E_{Q_n} f\left(\frac{dP_n}{dQ_n}\right) = E_{\mu_n} f\left(\frac{dP_n}{dQ_n}\right)\frac{dQ_n}{d\mu_n} = E_{\mu_n} g(W_n) \to Ef\left(\frac{W}{2-W}\right)(2-W).$$

By hypothesis the left hand side converges to $Ef(U)$, which thus equals the right hand side for every bounded continuous function f. Take a sequence of such f satisfying $1 \geq f_m \searrow 1_{\{0\}}$, so that dominated convergence implies

$$P(U=0) = E1_{\{0\}}(U) = E1_{\{0\}}\left(\frac{W}{2-W}\right)(2-W) = 2P(W=0).$$

By analogous arguments, $Ef(V) = Ef((2-W)/W)W$ for every continuous bounded f, and taking a sequence $0 \leq f_m(x) \nearrow x$ we conclude from monotone convergence

$$EV = E\left(\frac{2-W}{W}\right)W = E(2-W)1\{W>0\} = 2P(W>0) - 1$$

so $P(U=0) + EV = 1$, which conlcudes the proof. $\square$

The key application of this lemma in our context is the following corollary, which we shall use repeatedly in the next section.

Corollary 1. *i) Let P_n, Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$ such that $\frac{dP_n}{dQ_n} \to^d_{Q_n} e^X$ where $X \sim N(-\frac{1}{2}\sigma^2, \sigma^2)$ for some $\sigma^2 > 0$ as $n \to \infty$. Then $P_n \lhd\rhd Q_n$.*
ii) If $\{f(\theta) : \theta \in \Theta\}$ is locally asymptotically normal and if $h_n \to h \in \mathbb{R}^p$, then the product measures $P^n_{\theta+h_n/\sqrt{n}}$ and P^n_θ corresponding to samples $X_1, \ldots, X_n$ from densities $f(\theta + h_n/\sqrt{n})$ and $f(\theta)$, respectively, are mutually contiguous. In particular if a statistic $T(Y_1, \ldots, Y_n)$ converges to zero in probability under P^n_θ then it also converges to zero in $P^n_{\theta+h_n/\sqrt{n}}$-probability.

Proof. i) Since $P(e^X > 0) = 1$ for every normal random variable, part ii) of Lemma 3 implies $Q_n \lhd P_n$, and since $Ee^{N(\mu,\sigma^2)} = 1$ if and only if $\mu = -\sigma^2/2$ the converse follows from part iii) of the same lemma. Part ii) now follows immediately from Proposition 5 and the fact that the asymptotic expansion there converges in distribution to $N(-h^T i(\theta)h/2, h^T i(\theta)h)$ under P_θ. The last claim follows from the last part of Lemma 3. □

Example 8. *[Hodges' estimator, continued.]* Let us return to Example 7 and apply the above ideas to detect the flaw in the construction of the Hodges' estimator $\tilde{\theta}$ in the setting of locally asymptotically normal models. Intuitively speaking, the improvement of $\tilde{\theta}$ over $\hat{\theta}$ occurs at the origin $\theta = 0$, but this comes at the price of unbounded minimax risk 'near' the origin. To make this precise consider a 'local alternative' $0 + h/\sqrt{n}$ where $h \in \mathbb{R}$ is arbitrary, let P^n_θ be the product measure representing the law of a sample of size n from P_θ, and let E^n_θ be the expectation under this product measure. The minimax quadratic risk for fixed n is bounded from below by

$$\begin{aligned}\sup_{\theta\in\Theta} E^n_\theta(\sqrt{n}(\tilde{\theta}(X_1, \ldots, X_n) - \theta))^2 &\geq E^n_{h/\sqrt{n}} n(\tilde{\theta} - h/\sqrt{n})^2 \\ &\geq h^2 E^n_{h/\sqrt{n}} 1\{\tilde{\theta} = 0\} \\ &\geq h^2(1 - P^n_{h/\sqrt{n}}(\tilde{\theta} \neq 0)).\end{aligned}$$

We know from Example 7 above that $P^n_0(\tilde{\theta} \neq 0) \to 0$ as $n \to \infty$. By Corollary 1 the product measures $P^n_0, P^n_{h/\sqrt{n}}$ are contiguous, hence the quantity in the last line of the last display converges to h^2, in particular it exceeds $h^2/2$ for all $n \geq n_0(h)$ large enough. Since h was arbitrary we conclude that

$$\lim_n \sup_{\theta\in\Theta} E^n_\theta(\sqrt{n}(\tilde{\theta} - \theta))^2 = \infty.$$

This is in contrast to the MLE whose asymptotic minimax risk

$$\limsup_n \sup_{\theta\in\Theta} E_\theta^n(\sqrt{n}(\hat{\theta}-\theta))^2 < \infty,$$

see the discussion after Example 7 above. In this sense the Hodges' estimator is *not* a uniform improvement over the maximum likelihood estimator.

2.2.5 Bayesian Inference and the Bernstein - von Mises Theorem

Bayesian parametric inference also starts with the specification of a model of probability distributions $P_\theta, \theta \in \Theta \subset \mathbb{R}^p$, but it views the unknown parameter θ itself as a random variable. This means that the Bayesian has prior beliefs about the value of θ which are encoded in a random variable $\bar{\Theta}$ that has law, or *prior* distribution, Π on Θ. The observations $X_1, \ldots, X_n$ are viewed as being realisations of the conditional law $X|\theta \sim P_\theta$ given that $\bar{\Theta} = \theta$ has occurred. The conditional distribution of $\bar{\Theta}$ given $X_1, \ldots, X_n$ is called the *posterior* distribution, which can be computed by Bayes' formula. If the parametric model consists of probability densities $f(\theta)$, and if Π possesses a probability density π, then the posterior equals

$$\pi(\theta|X_1,\ldots,X_n) = \frac{\prod_{i=1}^n f(\theta,X_i)\pi(\theta)}{\int_\Theta \prod_{i=1}^n f(\theta,X_i)d\Pi(\theta)}, \tag{33}$$

see Exercise 13. The posterior can thus be interpreted as the weighted (and renormalised) likelihood of the model $\{f(\theta) : \theta \in \Theta\}$. We denote the corresponding posterior probability distribution by

$$\Pi(B|X_1,\ldots,X_n) = \int_B \pi(\theta|X_1,\ldots,X_n)d\theta, \quad B \text{ a Borel subset of } \Theta.$$

The posterior is the main tool for Bayesian inference: it gives rise to point estimates for θ by taking, for instance, the posterior mean

$$E(\bar{\Theta}|X_1,\ldots,X_n) = \int_\Theta \theta\pi(\theta|X_1,\ldots,X_n)d\theta,$$

and can also be used directly to construct 'credibility regions' (the 'Bayesian version' of confidence sets) for θ based on the quantiles of the posterior distribution.

While for a Bayesian statistician the analysis ends in a certain sense with the posterior, one can ask interesting questions about the the properties of posterior-based inference from a frequentist point of view. This means that one assumes that $X_1, \ldots, X_n$ are realisations from a fixed density $f(\theta_0)$ with law P_{θ_0} and studies the behaviour of the posterior, which is a random probability measure that depends on $X_1, \ldots, X_n$, under P_{θ_0}. The hope is that the information contained in the

sample eventually dominates the influence of the prior. For instance one can ask for frequentist consistency in the sense that, for every $\delta > 0$, as $n \to \infty$,

$$\Pi\left(\{\theta : \|\theta - \theta_0\| > \delta\} | X_1, \ldots, X_n\right) \to 0 \; in \; P_{\theta_0} - \text{probability},$$

which means that the posterior distribution asymptotically collapses to a point mass at θ_0. In fact we shall prove the more general result that under the assumption that the prior Π has a positive density in a neighbourhood of θ_0, and if $\hat{\theta}_n$ is the maximum likelihood estimator of $\{f(\theta) : \theta \in \Theta\}$ based on the sample $X_1, \ldots, X_n$ from density $f(\theta_0)$, then the posterior distribution is asymptotically equal to a normal distribution centred at $\hat{\theta}_n$ with covariance $i^{-1}(\theta_0)/n$. This entails that for purposes of frequentist asymptotic inference typical Bayesian procedures give the same results as if one uses the asymptotic distribution of the maximum likelihood estimator. This remarkable result is known as the Bernstein-von Mises theorem, see [3, 83], and its origins date back as far as to Laplace [51].

Before we give a proof of the general Bernstein-von Mises theorem, let us first examine a simple special case to gain some intuition. Consider observing $X, X_1, \ldots, X_n$ from P_θ equal to a $N(\theta, 1)$ distribution where $\theta \in \Theta = \mathbb{R}$, and take as prior Π a standard $N(0, 1)$ distribution on the θ's. Assuming $X|\theta \sim N(\theta, 1)$ one easily deduces from (33) that (see Exercise 13)

$$\theta | X_1, \ldots, X_n \sim N\left(\frac{\sum_{i=1}^n X_i}{n+1}, \frac{1}{n+1}\right). \tag{34}$$

We see that for Gaussian observations and Gaussian prior, the posterior is also Gaussian, a special case of a *conjugate* situation. The frequentist asymptotics of the normal distribution on the right hand side of (34) are easily obtained: If $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ is the sample mean we have

$$\sqrt{n}([\theta | X_1, \ldots, X_n] - \bar{X}_n) \sim N(Z_n, n/(n+1)),$$

where, assuming we are sampling from a fixed $N(\theta_0, 1)$ distribution,

$$Z_n = \sqrt{n}\left(E[\theta | X_1, \ldots X_n] - \bar{X}_n\right) = \frac{-\sqrt{n}}{n(n+1)} \bar{X}_n \to^{P_{\theta_0}} 0$$

as $n \to \infty$ by the law of large numbers. We can conclude that

$$\sqrt{n}([\theta | X_1, \ldots, X_n] - \bar{X}_n) \to^d N(0, 1) \;\text{ in } P_{\theta_0} - \text{probability},$$

in particular posterior-based inference coincides asymptotically with the standard frequentist inference based on the sample mean.

The remarkable fact is that the above phenomenon is not tied to the Gaussian conjugate situation, but that it is entirely universal for any prior that charges a

neighborhood of the true parameter θ_0 with positive probability. We now give Le Cam's [52] proof of the general Bernstein-von Mises theorem for parametric models under our standard regularity conditions, provided the model allows for the existence of uniformly consistent estimators, sufficient conditions for the existence of which were given after (19). Recall that

$$\|P - Q\|_{TV} := \sup_{B \in \mathcal{B}(\mathbb{R}^p)} |P(B) - Q(B)|$$

is the total variation distance on the set of probability measures on the Borel σ-algebra $\mathcal{B}(\mathbb{R}^p)$ of $\mathbb{R}^p$.

Theorem 5. *Consider a parametric model $\{f(\theta), \theta \in \Theta\}, \Theta \subset \mathbb{R}^p$, that satisfies the assumptions of Theorem 3. Suppose that the model admits a uniformly consistent estimator T_n in the sense of (19). Let $X_1, \ldots, X_n$ be i.i.d. from density $f(\theta_0)$, let $\hat{\theta}_n$ be the MLE based on the sample, assume the prior measure Π is defined on the Borel sets of $\mathbb{R}^p$ and that Π possesses a Lebesgue-density π that is continuous and positive in a neighbourhood of θ_0. Then, if $\Pi(\cdot|X_1, \ldots, X_n)$ is the posterior distribution given the sample, we have*

$$\left\| \Pi(\cdot|X_1, \ldots, X_n) - N\left(\hat{\theta}_n, \frac{1}{n} i^{-1}(\theta_0)\right) \right\|_{TV} \to^{P_{\theta_0}} 0 \quad \text{as } n \to \infty. \tag{35}$$

Proof. We shall assume, for notational simplicity, that the X_i are univariate random variables. We shall write, to ease notation, π_n and Π_n for the posterior density and distribution given the priors π and Π, respectively.

Step 1: We shall first translate the problem into a LAN setting. The total variation norm on densities equals half the L^1-norm and thus the quantity in (35) equals, using the change of variables $h = \sqrt{n}(\theta - \theta_0)$, and writing $\sigma_0 := \sqrt{2\pi \det(i^{-1}(\theta_0))}$,

$$\begin{aligned} &\frac{1}{2} \int \left| \pi_n(\theta) - \frac{n^{p/2}}{\sigma_0} \exp\left\{ -\frac{1}{2} n(\theta - \hat{\theta}_n)^T i(\theta_0)(\theta - \hat{\theta}_n) \right\} \right| d\theta = \\ &\frac{1}{2} \int \left| \frac{\pi_n(\theta_0 + h/\sqrt{n})}{n^{p/2}} - \frac{1}{\sigma_0} e^{-\frac{1}{2}(h - \sqrt{n}(\hat{\theta}_n - \theta_0))^T i(\theta_0)(h - \sqrt{n}(\hat{\theta}_n - \theta_0))} \right| dh \end{aligned} \tag{36}$$

where the normal density subtracted in the second line is the one of a $N(\sqrt{n}(\hat{\theta}_n - \theta_0), i^{-1}(\theta_0))$ distribution. Note moreover that, using the above change of variables again in the denominator,

$$\frac{\pi_n(\theta_0 + h/\sqrt{n})}{n^{p/2}} = \frac{\prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, X_i)\pi(\theta_0 + h/\sqrt{n})}{\int \prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, X_i)\pi(\theta_0 + h/\sqrt{n})dh} \equiv \tilde{\pi}_n(\theta),$$

which equals the posterior density arising from the sample and prior density $\tilde{\pi}(h) := \pi(\theta_0 + h/\sqrt{n})/n^{p/2}$, positive in a neighbourhood of zero. Define by $\tilde{\Pi}_n, \tilde{\Pi}$ the corresponding posterior and prior probability measures concentrated on $\{h = \sqrt{n}(\theta - \theta_0) : \theta \in \Theta\}$, respectively, and deduce that it suffices to prove that

$$\|\tilde{\Pi}_n - N(\sqrt{n}(\hat{\theta}_n - \theta_0), i^{-1}(\theta_0))\|_{TV} \to^{P_{\theta_0}} 0 \tag{37}$$

as $n \to \infty$.

Define next

$$\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} i^{-1}(\theta_0) \frac{\partial Q(\theta_0, X_i)}{\partial \theta}.$$

The proof of Theorem 3 implies that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) - \Delta_{n,\theta_0} \to^{P_{\theta_0}} 0 \text{ as } n \to \infty,$$

and moreover that both terms in the difference are uniformly tight (Prohorov's theorem) and hence concentrate with probability arbitrarily close to one one a bounded subset of $\mathbb{R}^p$. The total variation difference between two normal densities $dN(u, Z^{-1})$ and $dN(v, Z^{-1})$ can be written, up to multiplicative constants, as

$$\int e^{-h^T Z h} \left| e^{2u^T Z h} e^{-u^T Z u} - e^{2v^T Z h} e^{-v^T Z v} \right| dh$$

and since the exponential map and $x \mapsto x^2$ are Lipschitz on bounded sets, the previous remark implies, for some universal constant C,

$$\|N(\sqrt{n}(\hat{\theta}_n - \theta_0), i^{-1}(\theta_0)) - N(\Delta_{n,\theta_0}, i^{-1}(\theta_0))\|_{TV} \leq C \|\sqrt{n}(\hat{\theta}_n - \theta_0) - \Delta_{n,\theta_0}\|$$

so that this quantity converges to zero in P_{θ_0}-probability as $n \to \infty$. These observations combined with (38) above imply that in order to prove (35) it suffices to prove

$$\|\tilde{\Pi}_n - N(\Delta_{n,\theta_0}, i^{-1}(\theta_0))\|_{TV} \to^{P_{\theta_0}} 0 \tag{38}$$

as $n \to \infty$.

To proceed we introduce some simplifying notation. Define, for any measurable set $C \subset \mathbb{R}^p$, the probability measure

$$\tilde{\Pi}^C(B) = \frac{\tilde{\Pi}(B \cap C)}{\tilde{\Pi}(C)}$$

obtained from restricting (and renormalising) the prior $\tilde{\Pi}$ to C, let $\tilde{\pi}^C$ be its density, and set $\tilde{\Pi}_n^C = \tilde{\Pi}^C(\cdot | X_1, \ldots, X_n)$ to be the posterior obtained under the

restricted prior $\tilde{\Pi}^C$. Write further P_h^n for the distribution of the sample $X_1, \ldots, X_n$ under $\theta_0 + h/\sqrt{n}$ and define

$$P_C^n = \int_C P_h^n d\tilde{\Pi}^C(h),$$

the $\tilde{\Pi}^C$-mean of P_h^n over C, which should be understood in the sense that whenever we compute a probability / take expectation with respect to P_C^n then we compute a probability / take an expectation under $P_n^h, h \in C$, first and then integrate the result over C with respect to $\tilde{\Pi}^C$.

Throughout the remainder of the proof we shall use the contiguity relation

$$P_U^n \lhd\rhd P_0^n \tag{39}$$

for U any closed ball of fixed radius around zero. To verify this, let first A_n be a sequence of sets for which $P_0^n(A_n) \to 0$. Then Corollary 1 implies $P_h^n(A_n) \to 0$ for every $h \in U$ and since probabilities are bounded by one the dominated convergence implies $P_U^n(A_n) = \int_U P_h^n(A_n) d\tilde{\Pi}^U(h) \to 0$ as $n \to \infty$. The converse also follows, noting $P_U^n(A_n) \geq \inf_{h \in U} P_h^n(A_n)$, passing to a convergent subsequence $h_n \in U$ approaching the infimum, and using Proposition 5 combined with Corollary 1. We conclude that when showing convergence to zero of a random variable we may interchange P_U^n and P_0^n as we wish in view of Corollary 1.

Step 2: We first show that the difference between the full posterior and the posterior obtained from the prior restricted to C converges to zero where C is a ball of radius M sufficiently large about zero. For arbitrary measurable set A, B, writing $J(A) = \int_A \prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, X_i) d\tilde{\Pi}(h)$,

$$\begin{aligned}
\left|\tilde{\Pi}_n(B) - \tilde{\Pi}_n^C(B)\right| &= \left|\frac{J(B)}{J(\mathbb{R}^p)} - \frac{J(B \cap C)}{J(C)}\right| \\
&= \left|\frac{J(C^c \cap B)}{J(\mathbb{R}^p)} + \frac{J(C \cap B)}{J(\mathbb{R}^p)} - \frac{J(C \cap B)}{J(C)}\right| \\
&= \left|\frac{J(C^c \cap B)}{J(\mathbb{R}^p)} + \frac{J(C \cap B)J(C) - J(C \cap B)J(\mathbb{R}^p)}{J(\mathbb{R}^p)J(C)}\right| \\
&= \left|\frac{J(C^c \cap B)}{J(\mathbb{R}^p)} - \frac{J(\mathbb{R}^p \setminus C)}{J(\mathbb{R}^p)}\frac{J(C \cap B)}{J(C)}\right| \\
&= \left|\tilde{\Pi}_n(B \cap C^c) - \tilde{\Pi}_n(C^c)\tilde{\Pi}_n^C(B)\right| \leq 2\tilde{\Pi}_n(C^c),
\end{aligned}$$

a bound that is uniform in B and hence applies to the total variation distance between $\tilde{\Pi}_n$ and $\tilde{\Pi}_n^C$. We now show that the E_U^n - expectation of this bound converges to zero, so that it also converges to zero in P_U^n-probability by Markov's

inequality. Define ϕ_n to be the test for $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta \setminus \{\theta_0\}$ obtained from taking the maximum of the tests ϕ_n in Lemmata 1, 2. Recalling (39) and since $\phi_n \to 0$ under P_0^n as $n \to \infty$ we deduce

$$E_U^n \tilde{\Pi}_n(C^c) = E_U^n[\tilde{\Pi}_n(C^c)(1-\phi_n)] + o_{P_U^n}(1). \tag{40}$$

The quantity $E_U^n \tilde{\Pi}_n(C^c)(1-\phi_n)$ equals by definition and Fubini's theorem

$$\int_U \int_{\mathbb{R}^n} \int_{C^c} (1-\phi_n) \frac{\prod_{i=1}^n f(\theta_0 + \frac{g}{\sqrt{n}}, x_i)}{\int \prod_{i=1}^n f(\theta_0 + \frac{m}{\sqrt{n}}, x_i) d\tilde{\Pi}(m)} d\tilde{\Pi}(g) \prod_{i=1}^n f\left(\theta_0 + \frac{h}{\sqrt{n}}, x_i\right) dx_i \frac{d\tilde{\Pi}(h)}{\tilde{\Pi}(U)}$$

$$= \frac{\tilde{\Pi}(C^c)}{\tilde{\Pi}(U)} \int_{C^c} \int_{\mathbb{R}^n} \int_U (1-\phi_n) \frac{\prod_{i=1}^n f(\theta_0 + \frac{h}{\sqrt{n}}, x_i)}{\int \prod_{i=1}^n f(\theta_0 + \frac{m}{\sqrt{n}}, x_i) d\tilde{\Pi}(m)} d\tilde{\Pi}(h) dP_g^n(x) d\tilde{\Pi}^{C^c}(g)$$

$$= \frac{\tilde{\Pi}(C^c)}{\tilde{\Pi}(U)} E_{C^c}^n \tilde{\Pi}_n(U)(1-\phi_n),$$

so we have exchanged the 'centered' expectation over U with expectation under 'distant alternatives' C^c. Under these alternatives the type two errors in (40) converge to zero exponentially fast, and we exploit this now: using the transformation theorem for probability measures, Lemma 2, the inequality $\tilde{\Pi}(U) = \Pi(\theta_0 + U/\sqrt{n}) \geq 1/(cn^{p/2})$ for some $c > 0$ by positivity and continuity of the density π in a neighbourhood of θ_0, and for $D' \leq 1$ small enough so that π is bounded by $c'(\Pi)$ on $\{\theta : \|\theta - \theta_0\| \leq D'\}$ that

$$\begin{aligned}
\frac{\tilde{\Pi}(C^c)}{\tilde{\Pi}(U)} E_{C^c}^n \tilde{\Pi}_n(U)(1-\phi_n) &\leq \frac{1}{\tilde{\Pi}(U)} \int_{C^c} E_h^n (1-\phi_n) d\tilde{\Pi}(h) \\
&= \frac{1}{\tilde{\Pi}(U)} \int_{\theta : \|\theta-\theta_0\| \geq M/\sqrt{n}} E_\theta^n (1-\phi_n) d\Pi(\theta) \\
&\leq c'(\Pi) c n^{p/2} \int_{\theta : M/\sqrt{n} \leq \|\theta-\theta_0\| \leq D'} E_\theta^n (1-\phi_n) d\theta \\
&\quad + c n^{p/2} \int_{\theta : \|\theta-\theta_0\| > D'} E_\theta^n (1-\phi_n) d\Pi(\theta) \\
&\leq c'' \int_{\theta : \|\theta-\theta_0\| \geq M/\sqrt{n}} e^{-Dn\|\theta-\theta_0\|^2} n^{p/2} d\theta + 2cn^{p/2} e^{-Cn} \\
&= c'' \int_{h : \|h\| \geq M} e^{-D\|h\|^2} dh + 2cn^{p/2} e^{-Cn}
\end{aligned}$$

which can be made smaller than $\varepsilon > 0$ arbitrary by choosing M large enough and thus implies, as $M \to \infty$

$$\sup_B \left| \tilde{\Pi}_n(B) - \tilde{\Pi}_n^C(B) \right| \to^{P_U^n} 0. \tag{41}$$

By (39) this quantity then also converges to zero in P_0^n-probability.

To complete the second step, note that for any sequence $N(\mu_n, i)$ of normal probability measures such that $\sup_n |\mu_n| < \infty$ and for every Borel set B in $\mathbb{R}^p$ we have

$$|N(\mu_n, i)(B) - N^C(\mu_n, i)(B)| \leq 2N(\mu_n, i)(C^c),$$

where N^C is N restricted to C and renormalised, and since Δ_{n,θ_0} is uniformly tight by Prohorov's theorem we conclude that for C_n a sufficiently large ball of radius $M(\varepsilon)$ around zero

$$\|N(\Delta_{n,\theta_0}, i^{-1}(\theta_0)) - N^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))\|_{TV} < \varepsilon$$

for every $\varepsilon > 0$ on a set of P_0^n probability as close to one as desired. This leaves us with having to prove

$$\left\| N^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0)) - \tilde{\Pi}_n^C \right\|_{TV} \to^{P_0^n} 0 \text{ as } n \to \infty \tag{42}$$

for every ball C about zero of fixed radius M.

Step 3: We prove that (42) holds under the law P_C^n, which is sufficient as it is contiguous to P_0^n by (39). The total variation norm of two probability measures P, Q can be expressed as $\|P-Q\|_{TV} = 2^{-1}\int(1-dP/dQ)^+dQ$ so we bound, writing $f_{h,n}$ for $\prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, X_i)$ and λ_C for Lebesgue measure on C,

$$\begin{aligned}
&\frac{1}{2}\|N^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0)) - \tilde{\Pi}_n^C\|_{TV} \\
&= \int \left(1 - \frac{dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(h)}{1_C f_{h,n} d\tilde{\Pi}(h) / \int_C f_{g,n} d\tilde{\Pi}(g)}\right)^+ d\tilde{\Pi}_n^C(h) \\
&\leq \int\int \left(1 - \frac{f_{g,n} d\tilde{\Pi}(g) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(h)}{f_{h,n} d\tilde{\Pi}(h) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(g)}\right)^+ dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(g) d\tilde{\Pi}_n^C(h) \\
&\leq c\int\int \left(1 - \frac{f_{g,n} d\tilde{\Pi}(g) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(h)}{f_{h,n} d\tilde{\Pi}(h) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(g)}\right)^+ d\lambda_C(g) d\tilde{\Pi}_n^C(h)
\end{aligned}$$

where we used $(1-EY)^+ \leq E(1-Y)^+$ in the first inequality. The P_C^n-expectation of this quantity equals the expectation of the integrand

$$\left(1 - \frac{f_{g,n} d\tilde{\Pi}(g) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(h)}{f_{h,n} d\tilde{\Pi}(h) dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(g)}\right)^+$$

under

$$\tilde{\Pi}_n^C(dh) P_C^n(dx) \lambda_C(dg) = P_h^n(dx) \tilde{\Pi}^C(dh) \lambda_C(dg)$$

the latter identity following from Fubini's theorem and

$$\int_C \int_{\mathbb{R}^n} \prod_{i=1}^n f(\theta_0 + k/\sqrt{n}, x_i) \frac{\prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, x_i) d\tilde{\Pi}^C(h)}{\int \prod_{i=1}^n f(\theta_0 + m/\sqrt{n}, x_i) d\tilde{\Pi}^C(m)} dx d\tilde{\Pi}^C(k)$$
$$= \int_{\mathbb{R}^n} \prod_{i=1}^n f(\theta_0 + h/\sqrt{n}, x_i) dx d\tilde{\Pi}^C(h).$$

Since $\tilde{\Pi}^C(dh)$ has a bounded density it suffices to prove convergence to zero under $P_h^n(dx)\lambda_C(dh)\lambda_C(dg)$ which is contiguous to $P_0^n(dx)\lambda_C(dh)\lambda_C(dg)$ by (39). By the dominated convergence theorem it thus suffices to prove that the integrand converges to zero under P_0^n for every h, g, which follows from continuity of π at θ_0 and the fact that Proposition 5 implies that the likelihood ratios $f_{g,n}/f_{h,n} = (f_{g,n}/f_{0,n}) \cdot (f_{0,n}/f_{h,n})$ admit, under P_0^n, the LAN expansion

$$\exp\left\{ -\frac{1}{2} g^T i(\theta_0) g + \frac{1}{2} h^i(\theta_0) h - \frac{1}{\sqrt{n}} \sum_{i=1}^n g^T \frac{\partial Q(\theta_0, X_i)}{\partial \theta} + \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \frac{\partial Q(\theta_0, X_i)}{\partial \theta} \right\}$$

which exactly cancels with the ratio

$$\frac{dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(h)}{dN^C(\Delta_{n,\theta_0}, i^{-1}(\theta_0))(g)}.$$

□

2.2.6 Exercises

Exercise 7. *[Differentiating under an Integral Sign.]* Let V be an open subset of $\mathbb{R}^p$ and let $(S, \mathcal{A}, \mu)$ be a measure space. Suppose the function $f : V \times S \to \mathbb{R}$ is μ-integrable for every $v \in V$, and assume that for every v, s the derivative $(\partial/\partial v) f(v, s)$ exists and is continuous as a mapping from V to $\mathbb{R}^p$ for every s. Suppose further there exists a μ-integrable function $g : S \to \mathbb{R}$ such that $\|(\partial/\partial v) f(v, s)\| \leq g(s)$ for every $v \in V, s \in S$. Prove that the mapping $\phi : v \mapsto \int_S f(v, s) d\mu(s)$ from V to $\mathbb{R}$ is differentiable with derivative $(\partial/\partial v)\phi(v) = \int_S (\partial/\partial v) f(v, s) d\mu(s)$. [Hint: use the pathwise mean value theorem and dominated convergence.]

Exercise 8. Formulate mild conditions on $K(\theta)$ such that the conditions of Theorem 3 are satisfied for the exponential family from Example 6.

Exercise 9. *[Estimation of the Fisher Information.]* Let the assumptions of Theorem 3 be satisfied. Assuming consistency of $\hat{\theta}_n$, prove that

$$\hat{i}_n = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 Q(\hat{\theta}_n, Y_i)}{\partial \theta \partial \theta^T} \to^{P_{\theta_0}} i(\theta_0) \text{ as } n \to \infty.$$

Exercise 10. *[Confidence Sets and the Wald test.]* Working in the framework and under the assumptions of Theorem 3, and using Exercise 9, construct a random set $C_n \in \mathbb{R}^p$ (a 'confidence region') that depends only on α and $Y_1, \ldots, Y_n$ such that

$$\lim_n P_{\theta_0}(\theta_0 \in C_n) = 1 - \alpha.$$

If $\hat{\theta}_n$ is the MLE, derive further the asymptotic distribution of the Wald statistic

$$n(\hat{\theta}_n - \theta_0)^T \hat{i}_n (\hat{\theta}_n - \theta_0)$$

under P_{θ_0}, and use it to design an asymptotic level α test for the null hypothesis $H_0 : \theta = 0$ against $H_1 : \theta \in \Theta, \theta \neq 0$.

Exercise 11. Consider $Y_1, \ldots, Y_n$ i.i.d. Poisson random variables with parameter λ. Derive explicit formulas for the MLE and for the likelihood ratio test statistic for testing $H_0 : \lambda = \lambda_0$ against $H_1 : \lambda \neq \lambda_0$. Deduce the asymptotic distribution of $\sqrt{n}(\hat{\lambda}_n - \lambda)$ directly, and verify that it agrees with what the general asymptotic theory predicts.

Exercise 12. Let $\{f(\theta) : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p$, be a parametric model and let $\Phi : \Theta \to \mathbb{R}^m$ be measurable function. Let $\hat{\theta}_n$ be the MLE in the model Θ. Show that the maximum likelihood estimator $\hat{\phi}_n$ in the model $\{f(\phi) : \phi = \phi(\theta) : \theta \in \Theta\}$ equals $\Phi(\hat{\theta}_n)$.

Exercise 13. Use Bayes' rule to derive the expressions (33) and (34).

Exercise 14. In the setting of Theorem 5, let C_n be an Euclidean ball in $\mathbb{R}^p$ centred at the MLE $\hat{\theta}_n$ such that $\Pi(C_n | X_1, \ldots, X_n) = 1 - \alpha$ for all n (C_n is a *credible set* for the posterior distribution). Show that $P_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha$ as $n \to \infty$ (that is, C_n is a frequentist confidence set).

2.3 High Dimensional Linear Models

In this section we consider a response variable $Y = (Y_1, \dots, Y_n)^T$ in vector notation, and study linear models

$$Y = X\theta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \theta \in \Theta = \mathbb{R}^p, \sigma^2 > 0, \tag{43}$$

where X is a $n \times p$ design matrix, and where ε is a standard Gaussian noise vector in $\mathbb{R}^n$. Throughout we denote the resulting $p \times p$ *Gram matrix* by

$$\hat{\Sigma} = \frac{1}{n} X^T X,$$

which is symmetric and positive semi-definite. We denote by $\|\theta\|_2 = \sqrt{\sum_j \theta_j^2}$ and $\|\theta\|_1 = \sum_j |\theta_j|$ the usual ℓ_2 and ℓ_1 norm of a vector in Euclidean space. Moreover $a \lesssim b$ will mean that $a \leq Cb$ for some fixed (ideally numerical, or otherwise at least 'harmless') constant $C > 0$.

2.3.1 Beyond the standard linear model

In the model (43) with $p \leq n$ the classical *least squares estimator* introduced by Gauß solves the minimisation problem

$$\min_{\theta \in \mathbb{R}^p} \frac{\|Y - X\theta\|^2}{n} = \min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (Y_i - (X\theta)_i)^2,$$

whose solution has the well known form, in matrix notation,

$$\hat{\theta} = (X^T X)^{-1} X^T Y \sim N(\theta, \sigma^2 (X^T X)^{-1}), \tag{44}$$

where we assume that X has full column rank so that $X^T X$ is invertible. Even without assuming Gaussianity for ε the normal distribution of $\hat{\theta}$ is still approximately true under some mild conditions on ε, X, as can be shown using the central limit theorem (for triangular arrays of independent random variables). Thus (44) can be used to conduct inference on θ following the principles laid out in the previous section, including the construction of confidence sets and tests of hypotheses.

Let us note that the performance of the least squares estimator depends strongly on the dimensionality of Θ. Indeed, assuming the simplest case of orthogonal design $X^T X/n = I_p$ for instance, we see

$$\frac{1}{n} E_\theta \|X(\hat{\theta} - \theta)\|_2^2 = E_\theta \|\hat{\theta} - \theta\|_2^2 = \frac{\sigma^2}{n} tr(I_p) = \frac{\sigma^2 p}{n}. \tag{45}$$

We can conclude from these heuristics that for 'reasonable' design matrices X the quadratic risk of the least squares estimator of θ is, for $p \leq n$, of order of magnitude

$$E_\theta \|\hat{\theta} - \theta\|_2^2 = \text{error variance } \sigma^2 \times \frac{\text{model dimension } p}{\text{sample-size } n},$$

and the prediction risk $E_\theta \|X(\hat{\theta} - \theta)\|_2^2/n$ is of the same order. See Exercise 15 for some details.

Recent advances in science and information processing have generated complex data sets that have to be thought of as *high-dimensional* in the sense that the number p of possible explanatory variables $\mathbf{x}_j = (x_{ij} : i = 1, \ldots, n), j = 1, \ldots, p$ *exceeds* the number n of observed responses $Y_i, i = 1, \ldots, n$. Moreover a priori selection of the relevant variables $\mathbf{x}_j$ is often impossible, in fact we may even think of many $\mathbf{x}_j$'s chosen at random by the scientist to 'sense' a high-dimensional signal θ, without requiring a scientific interpretation of the influence of the $\mathbf{x}_j$'s. Such models are fundamentally ill-posed unless one believes in *sparsity of the signal* θ in the sense that *most of the coefficients* $\theta_j, j = 1, \ldots, p$, are zero. Assuming sparsity the challenge for the statistician starts from the fact that one does not know which ones the nonzero coefficients are.

A basic setting to study such 'large p-small n' problems is to assume that a 'true' low-dimensional submodel $Y = X\theta^0 + \varepsilon$ sits within the linear model (43), where one assumes that

$$\theta^0 \in B_0(k) \equiv \{\theta \in \mathbb{R}^p \text{ has at most } k \text{ nonzero entries}\}, \quad k \leq p. \tag{46}$$

The parameter k is called the *sparsity level* of θ^0 which itself is called a *k-sparse* vector or signal. For $\theta^0 \in B_0(k), k \leq p$, we call

$$S_0 = \{j : \theta_j^0 \neq 0\}$$

the *active set* of θ^0, pertaining to those indices j that have nonzero coefficients θ_j^0. Moreover for arbitrary $\theta \in \mathbb{R}^p$ denote by θ_{S_0} the vector obtained from setting all $\theta_j, j \in S_0^c$, equal to zero and leaving the remaining θ_j's unchanged.

When thinking of a sparse vector we think of k much smaller than p, in fact typically even much smaller than n. In this situation we cannot use the least squares estimator $\hat{\theta}$ since $X^T X$ is never invertible for $p > n$. We may still hope to achieve a perfomance that improves on the (in the high-dimensional setting useless) bound p/n from (45), ideally in a way that would reflect the bound k/n corresponding to the low-dimensional submodel $Y = X\theta^0 + \varepsilon$. In other words we are trying to find an estimator that 'mimics the oracle' that would fit a least squares procedure on the k-dimensional 'true' submodel, with all the unnecessary

covariates removed. Such a question in fact already arises when $p \leq n$, and has been studied in the area of statistical model or variable selection.

A natural attempt to deal with such problems is to consider a modified criterion function that penalises 'too many nonzero' estimated coefficients;

$$Q_n(\theta) = \frac{\|Y - X\theta\|_2^2}{n} + \lambda \sum_{j=1}^{p} 1\{\theta_j \neq 0\}, \ \theta \in \mathbb{R}^p, \tag{47}$$

where λ is a penalisation parameter, paralleling the weights occuring in standard model-selection criteria, such as AIC, BIC or Mallow's C_p. For instance, restricting attention to $p \leq n$ and least squares estimators, we can minimise $crit_{C_p}(M)$ over all candidate submodels M of $\mathbb{R}^p$, where $crit_{C_p}(M)$ is defined and derived in Exercise 15, and fit least squares in the selected model $\hat{M}$. Even for k fixed minimising such a penalised least squares criterion functions over all $\binom{p}{k}$ submodels of $\mathbb{R}^p$ is combinatorially hard and practically not feasible for large p. We hence have to search for an alternative method that is computationally tractable in high dimensions but still incorporates the same penalisation ideas.

2.3.2 The LASSO

The main idea to resolve the above problems is to search for a *convex relaxation* of the (non-convex) problem of optimising Q_n from (47). If we consider the scale of 'complexity penalisation' functionals

$$\|\theta\|_q^q = \sum_{j=1}^{p} |\theta_j|^q, \ q > 0,$$

we would want to take q as close to zero as possible to mimic the 'ℓ_0-penalty' in (47). On the other hand the minimisation problem will only be convex in θ if $p \geq 1$, and hence the boundary value $p = 1$ becomes a natural choice that accommodates both practical feasibility and the attempt to penalise non-sparse models. Let us thus define the estimator

$$\tilde{\theta} = \tilde{\theta}_{LASSO} = \arg\min_{\theta \in \mathbb{R}^p} \left[\frac{\|Y - X\theta\|_2^2}{n} + \lambda \|\theta\|_1 \right], \tag{48}$$

where $\lambda > 0$ is a scalar penalisation parameter, known as the LASSO ('Least Absolute Shrinkage and Selection Operator'). The above minimisation problem may have several solutions, and the theory below holds for any selection from the set of its minimisers. We note that the fitted values $X\tilde{\theta}$ as well as the estimated ℓ_1-norm $\|\tilde{\theta}\|_1$ coincide for all solutions of (48), see Exercise 17.

Algorithms for efficient computation of this estimator exist, and we now investigate the theoretical properties of the LASSO. We will prove that if the design matrix X has some specific structure compatible with the solution path of $\tilde{\theta}_{LASSO}$, and if the true θ^0 generating the high-dimensional linear model is k-sparse, then the LASSO performs almost as well as the least squares estimator $\hat{\theta}$ in the k-dimensional submodel, that is,

$$\sup_{\theta \in B_0(k)} E_\theta \frac{\|X(\tilde{\theta} - \theta)\|_2^2}{n} \lesssim \log p \times \frac{k}{n}$$

where we recall that $B_0(k)$ denotes all k-sparse vectors in $\mathbb{R}^p$.

Next to the crucial Condition (49) that we discuss in detail in the next section we will assume in the following theorem that the error variance σ^2 is known and in fact standardised to one. If σ is unknown we need to multiply our choice of λ below by an estimate $\hat{\sigma}$ of it (see Exercise 18).

Theorem 6. *Let $\theta^0 \in B_0(k)$ be a k-sparse vector in $\mathbb{R}^p$ with active set S_0. Suppose*

$$Y = X\theta_0 + \varepsilon$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N(0, I_n)$, let $\tilde{\theta}$ be the LASSO estimator with penalisation parameter

$$\lambda = 4\bar{\sigma}\sqrt{\frac{t^2 + 2\log p}{n}}, \quad \bar{\sigma}^2 = \max_{j=1,\ldots,p} \hat{\Sigma}_{jj},$$

and assume the $n \times p$ matrix X is such that, for some $r_0 > 0$,

$$\|\tilde{\theta}_{S_0} - \theta^0\|_1^2 \leq k r_0 (\tilde{\theta} - \theta^0)^T \hat{\Sigma} (\tilde{\theta} - \theta^0) \tag{49}$$

on an event of probability at least $1 - \beta$. Then with probability at least $1 - \beta - \exp\{-t^2/2\}$ we have

$$\frac{1}{n}\|X(\tilde{\theta} - \theta^0)\|_2^2 + \lambda\|\tilde{\theta} - \theta^0\|_1 \leq 4\lambda^2 k r_0 \lesssim \frac{k}{n} \times \log p. \tag{50}$$

Proof. We first note that the definition of $\tilde{\theta}$ implies

$$\frac{1}{n}\|Y - X\tilde{\theta}\|_2^2 + \lambda\|\tilde{\theta}\|_1 \leq \frac{1}{n}\|Y - X\theta^0\|_2^2 + \lambda\|\theta^0\|_1$$

or equivalently, inserting the model equation $Y = X\theta^0 + \varepsilon$,

$$\frac{1}{n}\|X(\theta^0 - \tilde{\theta}) + \varepsilon\|_2^2 + \lambda\|\tilde{\theta}\|_1 \leq \frac{1}{n}\|\varepsilon\|_2^2 + \lambda\|\theta^0\|_1,$$

hence

$$\frac{1}{n}\|X(\theta^0 - \tilde{\theta})\|_2^2 + \lambda\|\tilde{\theta}\|_1 \leq \frac{2}{n}\varepsilon^T X(\tilde{\theta} - \theta^0) + \lambda\|\theta^0\|_1. \tag{51}$$

Lemma 4. *Let $\lambda_0 = \lambda/2$. Then for all $t > 0$*

$$\Pr\left(\max_{j=1,\dots,p} \frac{2}{n}|(\varepsilon^T X)_j| \le \lambda_0\right) \ge 1 - \exp\{-t^2/2\}.$$

Proof. The variables $(\varepsilon^T X)/\sqrt{n}$ are $N(0, \hat{\Sigma})$-distributed. Note that $\bar{\sigma}^2 \ge \hat{\Sigma}_{jj}$ for all j. For $Z \sim N(0,1)$ the probability in question therefore exceeds one minus

$$\begin{aligned}\Pr\left(\max_{j=1,\dots,p} \frac{1}{\sqrt{n}}|(\varepsilon^T X)_j| > \bar{\sigma}\sqrt{t^2 + 2\log p}\right) &\le \sum_{j=1}^{p} \Pr(|Z| > \sqrt{t^2 + 2\log p}) \\ &\le p e^{-t^2/2} e^{-\log p} = e^{-t^2/2}\end{aligned}$$

where we used Exercise 16 in the last inequality. □

We hence have on the event inside of the probability of the last lemma – call it A – the inequality

$$|2\varepsilon^T X(\tilde{\theta} - \theta^0)/n| \le \max_{j=1,\dots,p} |2(\varepsilon^T X)_j/n| \|\tilde{\theta} - \theta^0\|_1 \le (\lambda/2)\|\tilde{\theta} - \theta^0\|_1 \tag{52}$$

which combined with (51) gives, on that event,

$$\frac{2}{n}\|X(\theta^0 - \tilde{\theta})\|_2^2 + 2\lambda\|\tilde{\theta}\|_1 \le \lambda\|\tilde{\theta} - \theta^0\|_1 + 2\lambda\|\theta^0\|_1. \tag{53}$$

Now using

$$\|\tilde{\theta}\|_1 = \|\tilde{\theta}_{S_0}\|_1 + \|\tilde{\theta}_{S_0^c}\|_1 \ge \|\theta^0_{S_0}\|_1 - \|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 + \|\tilde{\theta}_{S_0^c}\|_1$$

we obtain, on the event A and noting $\theta^0_{S_0^c} = 0$ by definition of S_0,

$$\begin{aligned}\frac{2}{n}\|X(\theta^0 - \tilde{\theta})\|_2^2 + 2\lambda\|\tilde{\theta}_{S_0^c}\|_1 &\le \frac{2}{n}\|X(\theta^0 - \tilde{\theta})\|_2^2 + 2\lambda\|\tilde{\theta}\|_1 - 2\lambda\|\theta^0_{S_0}\|_1 + 2\lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 \\ &\le \lambda\|\tilde{\theta} - \theta^0\|_1 + 2\lambda\|\theta^0\|_1 - 2\lambda\|\theta^0_{S_0}\|_1 + 2\lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 \\ &= 3\lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 + \lambda\|\tilde{\theta}_{S_0^c}\|_1\end{aligned}$$

so that after subtracting

$$\frac{2}{n}\|X(\theta^0 - \tilde{\theta})\|_2^2 + \lambda\|\tilde{\theta}_{S_0^c}\|_1 \le 3\lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 \tag{54}$$

holds on the event A. Now (50) follows since, on the event A, using the last inequality, (49) and $4ab \le a^2 + 4b^2$,

$$\begin{aligned}\frac{2}{n}\|X(\tilde{\theta} - \theta^0)\|_2^2 + \lambda\|\tilde{\theta} - \theta^0\|_1 &= \frac{2}{n}\|X(\tilde{\theta} - \theta^0)\|_2^2 + \lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 + \lambda\|\tilde{\theta}_{S_0^c}\|_1 \\ &\le 4\lambda\|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1 \\ &\le 4\lambda\sqrt{kr_0/n}\|X(\tilde{\theta} - \theta^0)\|_2 \\ &\le \frac{1}{n}\|X(\tilde{\theta} - \theta^0)\|_2^2 + 4\lambda^2 k r_0.\end{aligned}$$

□

The above result gives a bound for the prediction error $\|X(\tilde{\theta}-\theta)\|_2$ that 'almost' matches the one of the 'oracle' least squares estimator in the k-sparse submodel, with the typically mild 'penalty' $\log p$ for not knowing the position of the active set S_0. At least when $\beta = 0$ the above bounds can be integrated to give bounds for the expectations of the above errors (and thus for the risk) too, using the inequality $E(X) = K + \int_K^\infty P(X > u)du$. One can also deduce from Theorem 6 a result for the estimation error $\|\tilde{\theta} - \theta\|_2$, see Exercise 19.

While the above theorem is a neat result about the performance of the LASSO, inference based on $\tilde{\theta}$ is not a straightforward task. For instance, unlike in the standard linear model, or in the parametric models dealt with above, the distribution of $\tilde{\theta}$ is not known, and it is not obvious at all how to construct a confidence set for θ^0. In fact inference based on *any* sparse estimator is fundamentally different from the standard theory developed above, and nontrivial issues arise. See the article [62] where it is shown that a basic confidence set of diameter at best of the order $n^{-1/4}$ can be constructed, and that uniform improvements on such a confidence set are impossible without further restrictions on the parameter θ^0.

2.3.3 Coherence conditions for design matrices

We now turn to discuss the crucial Condition (49) which requires

$$\|\tilde{\theta}_{S_0} - \theta^0\|_1^2 \leq k r_0 (\tilde{\theta} - \theta^0)^T \hat{\Sigma} (\tilde{\theta} - \theta^0)$$

to hold true with high probability. One way to verify this condition is to verify it with $\tilde{\theta}$ replaced by an arbitrary $\theta \in \mathcal{V}$ where $\mathcal{V}$ is a subset of $\mathbb{R}^p$ on which $\tilde{\theta}$ concentrates with high probability. Taking note of (54) the proof of the last theorem implies that the solution path of the LASSO satisfies (on the event A from the proof) $\|\tilde{\theta}_{S_0^c}\|_1 \leq 3\lambda \|\tilde{\theta}_{S_0} - \theta^0_{S_0}\|_1$.

Corollary 2. *Theorem 6 remains true with Condition (49) replaced by the following condition: For S_0 the active set of $\theta^0 \in B_0(k), k \leq p$, assume the $n \times p$ matrix X satisfies, for all θ in*

$$\{\theta \in \mathbb{R}^p : \|\theta_{S_0^c}\|_1 \leq 3\|\theta_{S_0} - \theta^0_{S_0}\|_1\}$$

and some universal constant r_0,

$$\|\theta_{S_0} - \theta^0\|_1^2 \leq k r_0 (\theta - \theta^0)^T \hat{\Sigma} (\theta - \theta^0). \tag{55}$$

An extensive discussion of this condition and the many variants of it can be found in Sections 6.12 and 6.13 in [7].

There is no space here to treat this subject in full, but let us investigate a key aspect of it in more detail. We can first note that, since $|S_0| = k$, one clearly has

$$\|\tilde{\theta}_{S_0} - \theta^0\|_1^2 \leq k\|\tilde{\theta}_{S_0} - \theta^0\|_2^2$$

so that (49) can be verified by requiring

$$\|\theta_{S_0} - \theta^0\|_2^2 \leq r_0(\theta - \theta^0)^T\hat{\Sigma}(\theta - \theta^0) \tag{56}$$

to hold for $\theta = \tilde{\theta}$ or for all $\theta \in \mathcal{V}$, with a uniform constant $r_0 > 0$. Condition (56) cannot be true without restrictions. For instance for $\mathcal{V} = \mathbb{R}^p, S_0 = \{1, \ldots, p\}$, the above condition effectively requires

$$\inf_{\theta \in \mathbb{R}^p} \frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} \geq r_0^{-1},$$

hence that $\hat{\Sigma} = X^TX/n$ has a minimal eigenvalue bounded away from zero, which is impossible for the case $p > n$ relevant here. But under suitable restrictions on the sparsity of θ^0 there may still be some hope. For instance if the LASSO concentrates on k'-sparse solutions – as can be shown under suitable conditions (see, e.g., Section 2.6 in [7]) – it suffices to verify (56) for all $\theta \in B_0(k')$. Since the difference $\theta - \theta^0$ is a $\bar{k} = k + k'$-sparse vector this leads to a basic mathematical question whether a noninvertible Gram matrix $\hat{\Sigma} = X^TX/n$ can have '*a smallest eigenvalue bounded away from zero along sparse subspaces*', that is, whether

$$\inf_{\theta \in B_0(\bar{k})} \frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} \geq r_0^{-1}, \quad \bar{k} \leq n, \tag{57}$$

can hold true.

From a deterministic point of view checking (57) when p is large will be hard if not impossible. But random matrix theory can come to our aid to provide some intuitions, particularly relevant if we think of $\hat{\Sigma} = X^TX/n$ as a sampled correlation or covariance matrix from population analogue Σ. Theorem 7 below will show that (57) does hold true for design matrices X whose entries are drawn i.i.d. from a standard Gaussian distribution, *and with high probability*. Results of this kind are related to the so-called *restricted isometry property* of high-dimensional random matrices which requires (57) and a corresponding, easier, upper bound too: One assumes for some $\epsilon > 0$ (typically desired to be as small as possible) that

$$(1 - \epsilon)\|\theta\|_2^2 \leq \|\hat{\Sigma}\theta\|_2^2 \leq (1 + \epsilon)\|\theta\|_2^2 \quad \forall \theta \in B_0(k). \tag{58}$$

Such conditions and their verification have been key topics in the related area of *compressed sensing*, see the papers Candès, Romberg and Tao (2006a,b), Candès and Tao (2007) for instance. They can also be used directly to verify the conditions in Corollary 2, as discussed in Sections 6.12 and 6.13 in [7].

For sake of exposition we formulate the following result as a 'large enough sample size n' result. It proves the lower bound in (58) with $\epsilon = 1/2$. The proof in fact gives a result for every $\epsilon > 0$ if one carefully tracks the dependence of all constants on ϵ, but the case $\epsilon < 1$ already makes the main point. Moreover, the easier right hand side inequality in (58) follows from the proof as well. A non-asymptotic version of the proof that holds for every $n \in \mathbb{N}$ can be proved too at the expense of slightly more tedious expressions – see Corollary 1 in [62], from where the proof of the following result is taken.

Theorem 7. *Let the $n \times p$ matrix X have entries $(X_{ij}) \sim^{i.i.d.} N(0,1)$, and let $\hat{\Sigma} = X^T X/n$. Suppose $n/\log p \to \infty$ as $\min(p,n) \to \infty$. Then for every $k \in \mathbb{N}$ fixed and every $0 < C < \infty$ there exists n large enough such that*

$$\Pr\left(\theta^T\hat{\Sigma}\theta \geq \frac{1}{2}\|\theta\|_2^2 \;\; \forall\theta \in B_0(k)\right) \geq 1 - 2\exp\left\{-Ck\log p\right\}.$$

Proof. The result is clearly true when $\theta = 0$. Hence it suffices to bound

$$\begin{aligned}
&\Pr\left(\theta^T\hat{\Sigma}\theta \geq \frac{\|\theta\|_2^2}{2} \;\; \forall\theta \in B_0(k)\setminus\{0\}\right)\\
&= \Pr\left(\frac{\theta^T\hat{\Sigma}\theta}{\|\theta\|_2^2} - 1 \geq -\frac{1}{2} \;\; \forall\theta \in B_0(k)\setminus\{0\}\right)\\
&\geq \Pr\left(\sup_{\theta\in B_0(k),\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \leq 1/2\right)
\end{aligned}$$

from below by $1 - 2\exp\{-Ck\log p\}$. To achieve this, fix a set $S \subset \{1,\ldots,p\}$ of cardinality $|S| = k$ and let $\mathbb{R}^p_S$ denote the corresponding k-dimensional subspace of $\mathbb{R}^p$. By the union bound for probabilities we see

$$\Pr\left(\sup_{\theta\in B_0(k),\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \geq \frac{1}{2}\right) \leq \sum_{S\subset\{1,\ldots,p\}} \Pr\left(\sup_{\theta\in\mathbb{R}^p_S,\|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \geq \frac{1}{2}\right).$$

If we can bound each of the probabilities in the last sum by $2e^{-(C+1)k\log p} = 2e^{-Ck\log p}p^{-k}$ then the proof is complete since there are $\binom{p}{k} \leq p^k$ subsets S of cardinality k in $\{1,\ldots,p\}$. The required bounds follow from Lemma 5 below upon

taking $t = (C+1)k\log p$, and noting that $(k\log p)/n \to 0$ as $n \to \infty$ implies that, for any C we have

$$18\left(\sqrt{\frac{(C+1)k\log p + c_0 k}{n}} + \frac{(C+1)k\log p + c_0 k}{n}\right) < 1/2$$

whenever n is large enough. □

The proof of Theorem 7 relied on the following key lemma, which is an application of basic ideas from 'empirical process theory'. It starts with a concentration inequality for single random variables (Lemma 6 given below), and deduces a concentration inequality that is uniform in many variables indexed by a set whose 'degree of compactness' can be controlled in a quantitative way – in the present case this set is the unit ball in a finite-dimensional space.

Lemma 5. *Under the conditions of Theorem 7 we have for some universal constant $c_0 > 0$, every $S \subset \{1,\ldots,p\}$ such that $|S| = k$ and every $t > 0$,*

$$\Pr\left(\sup_{\theta\in\mathbb{R}^p_S, \|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| \geq 18\left(\sqrt{\frac{t+c_0k}{n}} + \frac{t+c_0k}{n}\right)\right) \leq 2e^{-t}.$$

Proof. We note

$$\sup_{\theta\in\mathbb{R}^p_S, \|\theta\|_2^2\neq 0}\left|\frac{\theta^T\hat{\Sigma}\theta}{\theta^T\theta} - 1\right| = \sup_{\theta\in\mathbb{R}^p_S, \|\theta\|_2^2\neq 0}\left|\frac{\theta^T(\hat{\Sigma}-I)\theta}{\theta^T\theta}\right| = \sup_{\theta\in\mathbb{R}^p_S, \|\theta\|_2^2\leq 1}\left|\theta^T(\hat{\Sigma}-I)\theta\right|.$$

The unit ball

$$B(S) \equiv \{\theta \in \mathbb{R}^p_S : \|\theta\|_2 \leq 1\}$$

of $\mathbb{R}^p_S$ is compact and hence for any $0 < \delta < 1$ we can cover $B(S)$ by a net of points $\theta^l \in B(S), l = 1,\ldots,N(\delta)$, such that for every $\theta \in B(S)$ there exists l for which $\|\theta - \theta^l\|_2 \leq \delta$. Writing $\Phi = \hat{\Sigma} - I$ we have

$$\theta^T\Phi\theta = (\theta-\theta^l)^T\Phi(\theta-\theta^l) + (\theta^l)^T\Phi\theta^l + 2(\theta-\theta^l)^T\Phi\theta^l.$$

Given any $\theta \in B(S)$ and fixing $\delta = 1/3$ we can find θ^l such that

$$|(\theta-\theta^l)^T\Phi(\theta-\theta^l)| \leq \frac{1}{9}\sup_{v\in B(S)}|v^T\Phi v|.$$

Also, for ϕ_i the eigenvalues of the symmetric matrix Φ acting on $\otimes_{j\in S}\mathbb{R}$ and $\phi^2_{\max} = \max_i \phi_i^2$, by the Cauchy-Schwarz inequality,

$$|(\theta-\theta^l)^T\Phi\theta^l| \leq \delta\|\Phi\theta^l\|_2 \leq \delta|\phi_{\max}| \leq \frac{1}{3}\sup_{v\in B(S)}|v^T\Phi v|,$$

so that

$$\sup_{\theta\in B(S)} |\theta^T\Phi\theta| \le (9/2) \max_{l=1,\dots,N(1/3)} |(\theta^l)^T\Phi\theta^l|, \tag{59}$$

reducing the supremum over the unit ball to a maximum over $N(1/3)$ points.

For each fixed $\theta^l \in B(S)$ we have

$$(\theta^l)^T\Phi\theta^l = \frac{1}{n}\sum_{i=1}^{n}((X\theta^l)_i^2 - E(X\theta^l)_i^2)$$

and the random variables $(X\theta^l)_i$ are independent $N(0, \|\theta^l\|_2^2)$ distributed with variances $\|\theta^l\|_2^2 \le 1$. Thus, for $g_i \sim^{i.i.d.} N(0,1)$, using the union bound for probabilities,

$$\begin{aligned}
&\Pr\left((9/2)\max_{l=1,\dots,N(1/3)} |(\theta^l)^T\Phi\theta^l| > 18\left(\sqrt{\frac{t+c_0k}{n}} + \frac{t+c_0k}{n}\right)\right)\\
&\le \sum_{l=1}^{N(1/3)} \Pr\left(|(\theta^l)^T\Phi\theta^l| > 4\|\theta^l\|_2^2\left(\sqrt{\frac{t+c_0k}{n}} + \frac{t+c_0k}{n}\right)\right)\\
&= \sum_{l=1}^{N(1/3)} \Pr\left(\left|\sum_{i=1}^{n}(g_i^2-1)\right| > 4\left(\sqrt{n(t+c_0k)} + t + c_0k\right)\right)\\
&\le 2N(1/3)e^{-t}e^{-c_0k} \le 2e^{-t},
\end{aligned}$$

where we used the second inequality in Lemma 6 below with $z = t + c_0k$, that the covering numbers of the unit ball in k-dimensional Euclidean space satisfy $N(\delta) \le (A/\delta)^k$ for some universal constant $A > 0$ (see Exercise 21), and where we have chosen c_0 large enough in dependence of A only. □

The final ingredient is a basic concentration inequality for sums of centred squared Gaussian random variables. The inequality combines two concentration 'regimes', pertaining to product measure concentration (for n large it gives a Gaussian tail suggested by the central limit theorem) and to exponential concentration (for t large but n fixed it gives the tail of a squared standard normal variable).

Lemma 6. *Let $g_i, i = 1,\dots,n$, be i.i.d. $N(0,1)$ and set*

$$X = \sum_{i=1}^{n}(g_i^2 - 1).$$

Then for all $t \ge 0$ and every $n \in \mathbb{N}$,

$$\Pr(|X| \ge t) \le 2\exp\left\{-\frac{t^2}{4(n+t)}\right\}. \tag{60}$$

Moreover, for every $z \geq 0$ and every $n \in \mathbb{N}$,

$$\Pr(|X| \geq 4(\sqrt{nz} + z)) \leq 2e^{-z}. \tag{61}$$

Remark 1. *One may replace 4 by 2 in (61) at the expense of a slightly longer proof.*

Proof. For λ satisfying $|\lambda| < 1/2$ and g standard normal,

$$Ee^{\lambda(g^2-1)} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda(x^2-1)-x^2/2} dx = e^{-\lambda}/\sqrt{1-2\lambda} = e^{\frac{1}{2}[-\log(1-2\lambda)-2\lambda]}.$$

By Taylor development, for $|\lambda| < 1/2$,

$$\frac{1}{2}[-\log(1-2\lambda) - 2\lambda] = \lambda^2 \left(1 + \frac{2}{3}2\lambda + \cdots + \frac{2}{k+2}(2\lambda)^k + \dots\right) \leq \frac{\lambda^2}{1-2\lambda}.$$

Hence for all $0 < |\lambda| < 1/2$ and since the g_i's are i.i.d.,

$$\log Ee^{\lambda X} = \log \left(Ee^{\lambda(g^2-1)}\right)^n \leq \frac{n\lambda^2}{1-2\lambda} \tag{62}$$

so that

$$Ee^{\lambda X} \leq e^{n\lambda^2/(1-2\lambda)}$$

follows. Now using Markov's inequality gives, for all $t > 0$ and $0 < \lambda < 1/2$,

$$\Pr(X > t) \leq Ee^{\lambda X - \lambda t} \leq e^{n\lambda^2/(1-2\lambda)-\lambda t} = \exp\left\{-\frac{t^2}{4(n+t)}\right\}$$

after taking $\lambda = t/(2n+2t)$. Repeating the above argument with $-X, -\lambda$ and using the union bound $\Pr(|X| > t) \leq \Pr(X > t) + \Pr(-X > t)$ gives the first inequality of the lemma. The second inequality now follows from the first after substituting $t = 4(\sqrt{nz} + z)$ into the first inequality. □

The above proof relies on the assumption that the X_i's are Gaussian only through the last lemma. In [62] it is shown that one can treat more general 'sub-gaussian' designs if one replaces Lemma 6 by Bernstein's inequality (see p.486 in [7], for instance, for the relevant version of that inequality). Also, instead of i.i.d. X_{ij}'s one could have considered correlated designs that allow for $E(X^TX/n) = \Sigma$ where Σ is not necessarily the identity matrix but is invertible with minimal eigenvalue bounded away from zero – we again refer to [62] for these facts.

2.3.4 Exercises

Exercise 15. Derive the formula $\hat{\theta} = (X^T X)^{-1} X^T Y$ for the least squares estimator in the standard Gaussian linear model

$$Y = X\theta + \varepsilon,$$

when $p \leq n$, X is a $n \times p$ matrix of full column rank p, and $\varepsilon \sim N(0, \sigma^2 I_p), \sigma > 0$. Show that $X\hat{\theta} = PY$ where P is the projection matrix that projects onto the span of the column vectors of X and deduce $E\|X\hat{\theta} - X\theta\|^2 = \sigma^2 p$. Now let X be partitioned as (X^M, X^{M^c}) where X^M is a $n \times k$ matrix, $k < p$, and consider the least squares predictor $P_M Y = X\hat{\theta}^M$ from sub-model M, where P_M projects onto the linear span of the column vectors of X_M. For

$$\hat{\sigma}^2 = (n-p)^{-1}\|Y - PY\|^2$$

show that Mallow's C_p criterion

$$crit_{C_p}(M) = \|Y - P_M Y\|^2 + 2\hat{\sigma}^2 k - n\hat{\sigma}^2,$$

is an unbiased estimator of the prediction risk

$$E\|X\hat{\theta}^M - X\theta\|^2$$

of the least squares predictor from the restricted model M.

Exercise 16. Prove that Gaussian random variables are subgaussian, that is, for $Z \sim N(0,1)$ prove that for all $x > 0$,

$$\Pr(|Z| > x) \leq e^{-x^2/2}.$$

Exercise 17. Prove that every solution $\tilde{\theta}_{LASSO}$ of the LASSO criterion function generates the same fitted value $X\tilde{\theta}_{LASSO}$ and the same ℓ_1-norm $\|\tilde{\theta}_{LASSO}\|_1$.

Exercise 18. In the linear model (43) generated from θ^0, the 'signal to noise ratio' is defined as $SNR = \|X\theta^0\|_2/\sqrt{n}\sigma$. If $\hat{\sigma}^2 = Y^T Y/n$ (and assuming $EY = 0$ for simplicity), show that for all $t > 0$ and with probability at least $1 - \exp\{-t^2/2\}$ we have

$$\frac{\hat{\sigma}^2}{\sigma^2} \in \left[1 + SNR(SNR \pm 2t/\sqrt{n}) \pm b_n\right], \quad b_n \equiv \left|\frac{\varepsilon^T\varepsilon}{n\sigma^2} - 1\right|.$$

Exercise 19. In the setting of Corollary 2, prove that with probability at least $1 - e^{-t^2/2}$ one has

$$\|\tilde{\theta} - \theta^0\|_2^2 \lesssim \frac{k}{n}\log p,$$

assuming in addition, for the θ's relevant in (55), that

$$\|\theta_{\mathcal{N}} - \theta^0\|_2^2 \leq r_1(\theta - \theta^0)^T \hat{\Sigma}(\theta - \theta^0)$$

for some $r_1 > 0$, where $\theta_{\mathcal{N}}$ is the vector consisting of zeros except for those θ_j's for which $j \in S_0$ joined by those θ_j's with indices corresponding to the k largest $|\theta_j|'s$ for $j \notin S_0$.

Exercise 20. For a $p \times p$ symmetric matrix Φ, show that the maximal absolute eigenvalue $\phi_{\max} = \max_i |\phi_i|$ is equal to $\sup_{\|v\|_2 \leq 1} |v^T \Phi v|$. Show further that the minimal absolute eigenvalue corresponds to $\inf_{\|v\|_2 \leq 1} |v^T \Phi v|$.

Exercise 21. Let B be the unit ball in a k-dimensional Euclidean space. Let $N(\delta), \delta > 0$ be the minimal number of closed balls of radius δ with centers in B that are required to cover B. Show that for some constant $A > 0$ and every $0 < \delta < A$ we have

$$N(\delta) \leq (A/\delta)^k.$$

3 Nonparametric Models

We shall in the third part of these notes consider statistical models that are infinite-dimensional. There is at first no reason to call these models 'nonparametric', since one could easily think of a parametric model $\{f(\theta) : \theta \in \Theta\}$ where Θ is an infinite-dimensional set, but if one thinks of the infinite-dimensional models

$$\{\text{All probability distribution functions}\} \quad \text{or} \quad \{\text{All probability density functions}\}$$

then the parameter is the probability distribution / density itself, so that speaking of a parameter is not necessarily natural.

We shall see, however, that the differences between finite and infinite dimensional models are not only of a semantic nature, and that asymptotic theory in infinite dimensional models is distinctively different.

To ease the transition we shall start with a review of some classical nonparametric problems where the theory is similar to the 'parametric' case.

3.1 Classical Empirical Processes

3.1.1 Empirical Distribution Functions

Suppose we are given a random variable X that has unknown law P and distribution function $F(t) = P(X \leq t)$, and suppose we obtain n independent and identically distributed copies $X_1, ..., X_n$ from X. Suppose we want to estimate the distribution function F at the point t. The obvious estimator is to count the proportion of observations that are smaller or equal to t, namely

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \leq t\} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(X_i). \tag{63}$$

The function $t \mapsto F_n(t)$ is called the *empirical distribution function*, which is a random step function defined on the real line and taking values in $[0, 1]$.

Is $F_n(t)$ a good estimator of $F(t)$? Defining $Z_i(t) = 1_{(-\infty,t]}(X_i)$, these are again i.i.d. random variables, and their expectation is $E|1_{(-\infty,t]}(X)| = P(X \leq t) = F(t) \leq 1$. Consequently we have

$$|F_n(t) - F(t)| = \left|\frac{1}{n}\sum_{i=1}^{n}(Z_i(t) - EZ_i(t))\right| \to 0 \quad \Pr - a.s. \tag{64}$$

as $n \to \infty$, by the law of large numbers. This already tells us that we can estimate consistently *an arbitrary* distribution function at any given point t. Moreover, this law of large numbers holds uniformly in t, a result that is sometimes called the *fundamental theorem of mathematical statistics*, namely

Theorem 8. *(Glivenko (1933), Cantelli (1933)). Let $X_1, ..., X_n$ be i.i.d. random variables with arbitrary distribution function F. Then*

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to 0 \quad \Pr -a.s.$$

as $n \to \infty$.

Proof. We use Proposition 1, and have to find a suitable bracketing for the class of functions

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

For any grid of points $-\infty < t_0 < t_1 < ... < t_k < \infty$ the brackets $[l_0, u_0] = [0, 1_{(-\infty,t_0)}]$, $[l_i, u_i] = [1_{(-\infty,t_{i-1}]}, 1_{(-\infty,t_i)}]$ for $i = 1, ..., k$, $[l_{k+1}, u_{k+1}] = [1_{(-\infty,t_k]}, 1]$ cover $\mathcal{H}$, and clearly $E|l_i(X)| < \infty, E|u_i(X)| < \infty$ for all i. It remains to choose the grid such that $E|u_i(X) - l_i(X)| < \varepsilon$. If F is continuous, it takes $\mathbb{R}$ onto $(0, 1)$, so divide $(0, 1)$ into $[1/\varepsilon] + 1$ pieces with breakpoints $a_i, i = 1, ..., [1/\varepsilon]$, $|a_{i+1} - a_i| < \varepsilon$ and choose the t_i's such that $F(t_i) = a_i$, so that $F(t_{i+1}) - F(t_i) < \varepsilon$ for every i, which completes the proof since

$$E|u_i(X) - l_i(X)| = \int_{t_i}^{t_{i+1}} dP = F(t_{i+1}) - F(t_i).$$

If F is not continuous, take t_i as before, but if F has a jump at t so that it 'skips' the level a_j, then add the point t (without counting multiplicities). The brackets still have size $F(t_i-) - F(t_{i-1}) < \varepsilon, F(t_{i+1}) - F(t_i+) < \varepsilon$ by construction. □

In higher dimensions we have the following analogue of the Glivenko-Cantelli theorem. Let us write, for $t \in \mathbb{R}^d$, in abuse of notation, $1_{(-\infty,t]} = 1_{(-\infty,t_1]}...1_{(-\infty,t_d]}$.

Theorem 9. *Let $X_1, ..., X_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with common distribution function $F(t) = P(X \le t), t \in \mathbb{R}^d$. Define further $F_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{(-\infty,t]}(X_i)$. Then*

$$\sup_{t\in\mathbb{R}^d} |F_n(t) - F(t)| \to 0 \quad \Pr -a.s.$$

as $n \to \infty$.

The proof, which is only a little more involved than the one of Theorem 8, is left as Exercise 26.

One might then ask for probabilistic statements for $F_n(t) - F(t)$ that are more exact than just a law of large numbers, e.g., for a central limit theorem. Such questions have been at the heart of mathematical statistics and probability theory from the 1940s onwards: Several deep and fundamental results have been obtained for what is known as the *empirical process*, namely the stochastic process

$$t \mapsto \sqrt{n}(F_n(t) - F(t)), \qquad t \in \mathbb{R} \tag{65}$$

Note first that, similar to (64), we have for each given t that

$$\sqrt{n}(F_n(t) - F(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i(t) - EZ_i(t)) \to^d N(0, F(t)(1 - F(t)))$$

from the central limit theorem. Moreover, from the multivariate central limit theorem we have (Exercise 22) for any finite set of points $t_1, ..., t_k \in \mathbb{R}$

$$[\sqrt{n}(F_n(t_1) - F(t_1)), ..., \sqrt{n}(F_n(t_k) - F(t_k))] \to^d N(0, \Sigma) \tag{66}$$

as $n \to \infty$ where the limit is multivariate normal, and the covariance matrix has (i, j)-th entry $F(t_i \wedge t_j) - F(t_i)F(t_j)$.

To make this result 'uniform in t' is much more involved than in Theorem 8, as it essentially amounts to proving a central limit theorem in the infinite-dimensional space of bounded functions on $\mathbb{R}$. While a full understanding of the mathematics behind such results was not achieved before the 1990s ('empirical process theory', cf. Dudley (1999)), the following remarkable result was proved already in the 1950s. Denote by L^∞ the space of bounded functions on $\mathbb{R}$ equipped with the usual uniform norm $\|f\|_\infty := \sup_{t \in \mathbb{R}} |f(t)|$. We can view $F_n - F$ as random variables in the metric space L^∞. But what about the normal limiting variable suggested in (66)? Here is the relevant definition:

Definition 3 (Brownian Bridge)**.** The *F-Brownian bridge* process $\mathbb{G}_F$ is the mean-zero Gaussian process indexed by $\mathbb{R}$ that has the covariance

$$E\mathbb{G}_F(t_i)\mathbb{G}_F(t_j) = F(t_i \wedge t_j) - F(t_i)F(t_j)$$

for any $t_i, t_j \in \mathbb{R}$.

For F equal to the uniform distribution on $[0, 1]$ this process equals the *standard* Brownian bridge process $\mathbb{G}$. See Exercise 22 for more facts. There always exists a version of $\mathbb{G}_F$ which is sample bounded *almost surely*, that is $\sup_{t \in \mathbb{R}} |\mathbb{G}_F(t)| < \infty$ holds *almost surely*, a non-trivial fact that follows, e.g., from existence of sample-continuous versions of Brownian motion as proved in Theorem 12.1.5 in [29]. Hence the trajectories of (a suitable version of) the Brownian bridge are almost surely in the space L^∞.

The following result can be thought of as a central limit theorem in infinite dimensions.

Theorem 10. *(Doob (1949), Donsker (1952), Skorohod (1956), Dudley (1966)). Let $X_1, ..., X_n$ be i.i.d. random variables with arbitrary distribution function F. Then the random functions $\sqrt{n}(F_n - F)$ converge in distribution in the space L^∞ to the F-Brownian bridge process $\mathbb{G}_F$ as $n \to \infty$.*

One of the delicate points here is to actually show that $\sqrt{n}(F_n - F)$ is a proper random variable in (measurable mapping to) L^∞. Dudley (1966) gave an appropriate treatment of this point (which circumvents the somewhat involved alternative approach via the 'Skorohod topology', but is now also outdated), and he also proved the multi-dimensional analogue of Theorem 10 (i.e., for i.i.d. random vectors $X_1, ...X_n$). The proof of this theorem belongs to a course on empirical process theory and will not be given in these notes: The first (somewhat incorrect) proof is by Donsker (1952), a classical proof using the Skorohod topology can be found, e.g., in [6], Chapter 14, and a proof using more modern techniques is in [28].

3.1.2 Finite-sample error bounds and Minimaxity

Whereas Theorem 10 tells us that the stochastic behaviour of $\sqrt{n}(F_n - F)$ is approximately the one of a F-Brownian bridge, it is still a limit theorem, and so it is not clear what 'approximately' means for given sample size n, and what error we make by using this approximation. The following classical inequality shows quite remarkably that the normal approximation is *effective for every sample size*, as it shows that the probability of $\sqrt{n}\|F_n - F\|_\infty$ to cross the level λ is bounded by the tail of a normal distribution.

Theorem 11. *(Dvoretzky, Kiefer, Wolfowitz (1956), Massart (1990)) Let $X_1, ..., X_n$ be i.i.d. random variables with arbitrary distribution function F. Then, for every $n \in \mathbb{N}$ and every $\lambda \geq 0$,*

$$\Pr\left(\sqrt{n}\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| > \lambda\right) \leq 2\exp\{-2\lambda^2\}.$$

This inequality was proved by Dvoretzky, Kiefer and Wolfowitz (1956), with a larger leading constant. The sharp constants in this inequality were not obtained until Massart (1990). A proof of this inequality for fixed t will be Exercise 23.

Dvoretzky, Kiefer and Wolfowitz (1956) moved on and used their inequality to prove the following result, which establishes the 'asymptotic minimax optimality' of the empirical distribution function.

Theorem 12. *(Dvoretzky, Kiefer, Wolfowitz (1956)) Let $X_1, ..., X_n$ be i.i.d. random variables with arbitrary distribution function F. Denote by $\mathcal{P}$ the set of all probability distribution functions on $\mathbb{R}$, and by $\mathcal{T}_n$ the set of all estimators for F. Then*

$$\lim_n \frac{\sup_{F\in\mathcal{P}} \sqrt{n}E_F\|F_n - F\|_\infty}{\inf_{T_n\in\mathcal{T}_n}\sup_{F\in\mathcal{P}}\sqrt{n}E_F\|T_n - F\|_\infty} = 1$$

This result shows that if nothing is known a priori about the underlying distribution F, then, for large samples, the empirical distribution function is the best possible estimator for F in a minimax sense. This is still true if one has some a priori information on F, such as knowing that F is concave or convex, cf. Kiefer and Wolfowitz (1976). For a fresh view at the optimality of F_n (including the construction of an estimator that is uniformly better than F_n in a certain sense) see [40], Theorem 2.

3.1.3 Some Applications

The Kolmogorov-Smirnov Statistic.

We know from Theorem 10 that $\sqrt{n}(F_n - F)$ behaves approximately like a F-Brownian bridge $\mathbb{G}_F$, so that $\sqrt{n}\|F_n - F\|_\infty$ should behave approximately as the maximum (over $\mathbb{R}$) of $\mathbb{G}_F$. Whereas the limit process $\mathbb{G}_F$ does still depend on F, the maximum of its absolute value actually *does not depend on F anymore.* Using results of Kolmogorov, Smirnov proved the following result even before Theorem 10 was known. It follows quite easily from a 'continuous mapping' argument once one has proved Theorem 10.

Theorem 13. *(Kolmogorov (1933), Smirnov (1939)) Let $X_1, ..., X_n$ be i.i.d. random variables with continuous distribution function F. Then*

$$\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \to^d \sup_{t\in[0,1]} |\mathbb{G}(t)|$$

as $n \to \infty$ where $\mathbb{G}$ is a standard Brownian bridge. Furthermore the distribution of the limit is given by

$$\Pr\left(\sup_{t\in[0,1]} |\mathbb{G}(t)| > \lambda\right) = 2\sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2\lambda^2}.$$

Proof. To prove the first claim, consider first an i.i.d. sample $U_1, ..., U_n$ drawn from the uniform distribution $G(x) = x$ on $[0, 1]$, and denote by $G_n(x)$ the empirical distribution function of the uniform sample. Then, if F is a continuous distribution function, F takes $\mathbb{R}$ onto $(0, 1)$, hence

$$\sup_{x\in[0,1]} |G_n(x) - G(x)| = \sup_{t\in\mathbb{R}} |G_n(F(t)) - G(F(t))|.$$

[The boundary values $0, 1$ are negligible since $G_n(0) - 0 = G_n(1) - 1 = 0$ almost surely in view of absolute continuity of G.] Now clearly $G(F(t)) = F(t)$ and the distribution of $G_n(F(t))$ is the same as the distribution of $F_n(t)$ (in view of

$\{0 \le U_i \le F(t)\} = \{-\infty < F^{-1}(U_i) \le t\}$ and since $F^{-1}(U_i)$ has the same distribution as X_i drawn from the distribution F, see Exercise 25). We conclude that $\sqrt{n}\sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$ has the same distribution as $\sqrt{n}\sup_{x\in[0,1]} |G_n(x) - G(x)|$. Since $\sqrt{n}(G_n(x) - G(x))$ converges in law in $L^\infty(\mathbb{R})$ to the standard Brownian bridge $\mathbb{G}$ by Theorem 10, the first result follows from continuity of the mapping $f \mapsto \|f\|_\infty$ on $L^\infty(\mathbb{R})$ and the continuous mapping theorem for weak convergence.

The second result is a nice exercise in probability theory, and follows, e.g., from the reflection principle for Brownian motion, see [29], Chapter 12.3. □

This result, in particular the fact that the limit distribution does not depend on F, is extremely useful in statistical applications. Suppose for instance we conjecture that F is standard normal, or any other distribution function H, then we just have to compute the maximal deviation of F_n to H and compare it with the limiting distribution of Theorem 13, whose quantiles can be easily tabulated (e.g., Shorack and Wellner (1986, p.143)). Another application is in the construction of confidence bands for the unknown distribution function F, see Exercise 24.

Estimation of the Quantile Function.

Often the object of statistical interest is not the distribution function $F(t)$, but the quantile function $F^{-1} : (0, 1] \to \mathbb{R}$ which, when F has no inverse, is defined as the generalized inverse

$$F^{-1}(p) = \inf\{x : F(x) \ge p\}.$$

The natural estimator is to take the generalized inverse F_n^{-1} of the empirical distribution function F_n, namely

$$F_n^{-1}(p) = X_{(i)} \qquad \text{for } p \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$$

where $X_{(i)}$ is the i-th order statistic of the sample. (The order statistic $X_{(1)} < X_{(2)} < ... < X_{(n)}$ is the sample ordered beginning with the smallest observation and ending with the largest.) An application of Theorems 8 and 10 gives:

Theorem 14. *Let $0 < p < 1$. If F is differentiable at $F^{-1}(p)$ with positive derivative $f(F^{-1}(p))$, then*

$$|F_n^{-1}(p) - F^{-1}(p)| \to 0 \quad a.s.$$

as well as

$$\sqrt{n}(F_n^{-1}(p) - F^{-1}(p)) \to^d N\left(0, \frac{p(1-p)}{f^2(F^{-1}(p))}\right)$$

as $n \to \infty$.

We will prove this result later in this course as a corollary to Theorem 10 using what is known as the 'functional' or infinite-dimensional delta-method, see Section 3.7.1.

Estimation of Cumulative Hazard Functions.

A natural object in survival analysis and insurance mathematics is the hazard rate function

$$\frac{f(t)}{1-F(t-)}$$

of a nonnegative random variable X with distribution function $F : [0,\infty) \to \mathbb{R}$ and density f. The cumulative hazard function is

$$\Lambda_F(t) = \int_0^t \frac{f(u)}{1-F(u-)} du,$$

which can be estimated by the empirical cumulative hazard function

$$\Lambda_{F_n}(t) = \frac{1}{n}\sum_{i=1}^{n} (1-F_n(X_i-))^{-1} 1_{[0,t]}(X_i).$$

Theorem 15. *Let t be such that $1-F(t) > 0$. Then*

$$|\Lambda_{F_n}(t) - \Lambda_F(t)| \to 0 \quad a.s.$$

as well as

$$\sqrt{n}(\Lambda_{F_n}(t) - \Lambda_F(t)) \to^d N\left(0, \frac{F(t)}{1-F(t)}\right)$$

as $n \to \infty$.

Similar to Theorem 25, we will derive this result as a corollary to Theorems 8 and 10 later in this course, see Example 10.

3.1.4 Exercises

Exercise 22. Let T be a nonempty set and let $(W, \mathcal{W}, \mu)$ be a probability space. A *Gaussian process* G indexed by T is a mapping $G : T \times (W, \mathcal{W}, \mu) \to \mathbb{R}$ such that the vector $(G(t_1), ..., G(t_k))$ has a multivariate normal distribution for every finite set of elements $t_1, ..., t_k$ of T. A *Brownian motion* or *Wiener* process is the Gaussian process $B(t)$ on $T = [0,\infty)$ with mean zero and covariance $EB(t)B(s) = \min(s,t)$. For $T = [0,1]$ the *Brownian bridge process* is defined as $\mathbb{G}(t) = B(t) - tB(1)$. Find the covariance $E\mathbb{G}(t)\mathbb{G}(s)$ of this process. Show that the F-Brownian bridge $\mathbb{G}_F$ can be obtained from $\mathbb{G} \circ F$ by showing that the covariances coincide. Let $(\sqrt{n}(F_n(t_1) - F(t_1)), ..., \sqrt{n}(F_n(t_k) - F(t_k))$ be the empirical process indexed

by the finite set of points $t_1, ..., t_k$. Use the multivariate central limit theorem to derive its limiting distribution. How does the covariance of the limiting normal distribution relate to the covariance of the F-Brownian bridge?

Exercise 23. Use Hoeffding's inequality to deduce the Dvoretzky-Kiefer-Wolfowitz inequality for a fixed t, that is, without the $\sup_{t\in\mathbb{R}}$.

Exercise 24. Consider $X_1, ..., X_n$ independent and identically distributed random variables with continuous distribution function F. A level $1-\alpha$ confidence band for F centered at the empirical distribution function F_n is a family of random intervals $C_n(x) = [F_n(x)-d_n, F_n(x)+d_n], x \in \mathbb{R}, d_n > 0$, whose coverage probability satisfies $\Pr(F(x) \in C_n(x) \text{ for every } x \in \mathbb{R}) \geq 1-\alpha$. The band C_n has *asymptotic* level α if coverage holds in the limit, that is, $\liminf_n \Pr(F(x) \in C_n(x) \text{ for every } x \in \mathbb{R}) \geq 1-\alpha$. Argue that an asymptotic level 0.95 confidence band can be constructed by choosing $d_n = 1.359/\sqrt{n}$. [You may use that $P(\sup_{t\in[0,1]} |\mathbb{G}(t)| \leq 1.359) \simeq 0.95$, where $\mathbb{G}$ is the standard Brownian bridge.] Show further that the choice 1.36 in place of 1.359 implies the same coverage result for every fixed sample size n. Compute the diameter of these bands when $n = 100, n = 1000, n = 10000$.

Exercise 25. Quantile Transform. If $F : \mathbb{R} \to [0, 1]$ is a continuous distribution function and U a random variable that is uniformly distributed on $[0, 1]$, show that the new random variable $F^{-1}(U)$ has distribution function F.

Exercise 26. Let $X_1, ..., X_n$ be independent and identically distributed random vectors in $\mathbb{R}^2$ with continuous distribution function $F : \mathbb{R}^2 \to \mathbb{R}$. Define the empirical distribution function F_n of the sample, and prove that $\sup_{t\in\mathbb{R}^2} |F_n(t) - F(t)| \to 0$ almost surely as $n \to \infty$. What about $d > 2$?

3.2 Minimax Lower Bounds

We shall develop in this subsection some simple lower bound techniques from minimax theory, that shed light on the class of infinite-dimensional models that we will study subsequently. They will show that once we leave the simple setting of estimating a distribution function, statistical inference in infinite dimensional models may become qualitatively more complex.

Suppose we are given a model $\mathcal{P}$ of probability densities, and a random sample $X_1, \ldots, X_n$ from $f \in \mathcal{P}$, where $\mathcal{P}$ is equipped with some (pseudo-)metric d (satisfying the triangle inequality). Let $P_f \equiv P_f^n$ be probability law of the $X_1, \ldots, X_n$ and denote by E_f expectation with respect to P_f. Consider *any* estimator $f_n(x) = f(x; X_1, \ldots, X_n)$ for $f(x)$. The best performance in terms of risk under d-loss we can expect in this estimation problem is the minimax risk

$$\inf_{f_n} \sup_{f\in\mathcal{P}} r_n^{-1} E_f d(f_n, f), \tag{67}$$

where r_n is a sequence of positive real numbers. If r_n is chosen in such a way that this quantity is bounded away from zero and infinity, then we speak of r_n as being the minimax rate of convergence in the metric d over the class $\mathcal{P}$. In parametric problems this rate was seen to equal $r_n = 1/\sqrt{n}$. Likewise, when estimating a distribution function in sup-norm loss, the rate is $1/\sqrt{n}$. We shall see that in a variety of other relevant nonparametric problems, the rate is strictly slower than $1/\sqrt{n}$. The general conclusion will be that the minimax rate depends both on the complexity of the set $\mathcal{P}$ and on the loss function d.

3.2.1 A Reduction to Testing Problems

We shall show in this section how minimax lower bounds can be reduced to simple testing problems, for which lower bounds can be obtained in an effective way. We have from Markov's inequality

$$E_f r_n^{-1} d(f_n, f) \geq P_f(d(f_n, f) \geq r_n). \tag{68}$$

Let $n > 2$ and consider testing the simple hypothesis

$$H_0 : f = f_0 \ \text{ against } \ H_1 : f = f_1, \quad f_0, f_1 \in \mathcal{P}, \ d(f_0, f_1) \geq 2r_n.$$

Define the test $\Psi_n = 0$ if $d(f_n, f_0) < d(f_n, f_1)$ and $\Psi_n = 1$ otherwise. Then if $\Psi_n \neq 1$ we necessarily have

$$d(f_n, f_1) \geq d(f_1, f_0) - d(f_n, f_0) \geq 2r_n - d(f_n, f_1)$$

by the triangle inequality and repeating the argument for $\Psi_n \neq 0$ we conclude

$$P_{f_j}(d(f_n, f_j) \geq r_n) \geq P_{f_j}(\Psi_n \neq j), \ j = 0, 1.$$

We thus deduce, for such f_0, f_1,

$$\inf_{f_n \in \mathcal{F}} \sup_{f \in \mathcal{P}} P_f(d(f_n, f) \geq r_n) \geq \inf_{\Psi} \max_{j \in \{0,1\}} P_{f_j}(\Psi \neq j) \tag{69}$$

where the infimum extends over all tests based on the sample. One of the appeals of this lower bound is that it is independent of the metric d and, since these bounds hold for every n, we can let f_0, f_1 depend on n as long as they stay in $\mathcal{P}$ for every n.

We now need a lower bound on tests. A simple one is the following.

Lemma 7. *We have, for every $\eta > 0$,*

$$\inf_{\Psi} \max_{j \in \{0,1\}} P_{f_j}(\Psi \neq j) \geq \frac{1 - \eta}{2} \left(1 - \frac{E_{f_0}|Z - 1|}{\eta}\right)$$

where Z equals the likelihood ratio dP_{f_1}/dP_{f_0}.

Proof. We have, for any test Ψ,

$$\begin{aligned} 2\max_{j=0,1} P_{f_j}(\Psi \neq j) &\geq (E_{f_0}\Psi + E_{f_1}(1-\Psi)) = E_{f_0}(\Psi + (1-\Psi)Z) \\ &\geq E_{f_0}[(\Psi(1-\eta) + (1-\Psi)(1-\eta))1\{Z \geq 1-\eta\}] \\ &\geq (1-\eta)(1 - P_{f_0}(|Z-1| > \eta)) \\ &\geq (1-\eta)\left(1 - \frac{E_{f_0}|Z-1|}{\eta}\right) \end{aligned}$$

where we used Markov's inequality in the last step. □

This bound is useful if the likelihood ratio Z is close to one with large probability. If P_{f_0}, P_{f_1} correspond to product measures from samples from f_0, f_1, respectively, then closeness of Z to one means that the 'data' coming from P_{f_0} 'look similar' to data coming from P_{f_1}, which makes the testing problem more difficult. Quantitative estimates depend on concrete examples, and we study a key case in what follows. While Lemma 7 will be seen to be useful in this particular example, we should note that in many other minimax problems a reduction to a two-hypothesis testing problem can be too crude, and one has to resort to lower bounds for multiple hypothesis testing problems. We refer to [79] for an excellent introduction into the field of minimax lower bounds in nonparametric statistics.

3.2.2 Lower Bounds for Estimating a Differentiable Density

We shall now apply the previous techniques to show that the minimax risk for estimating a differentiable density at a point depends on the number of existing derivatives, and that the risk is always of slower order than $1/\sqrt{n}$.

On the space of m-times differentiable real-valued functions on $[0,1]$ define the norm

$$\|f\|_{m,\infty} := \max_{0 \leq \alpha \leq m} \|D^\alpha f\|_\infty \tag{70}$$

with the convention $D^0 f = f$, and with derivatives being understood one-sided at $0, 1$.

Theorem 16. *Let $m > 0, B > 1$ and let*

$$\mathcal{C}(m,B) = \left\{ f : [0,1] \to [0,\infty) : \int_0^1 f(x)dx = 1, \|f\|_{m,\infty} \leq B \right\}.$$

Let $\mathcal{F}$ be the class of all possible density estimators (i.e., all measurable functions of $X_1, \ldots, X_n$), and let $x_0 \in [0,1]$ be arbitrary. Then

$$\liminf_n \inf_{f_n \in \mathcal{F}} \sup_{f \in \mathcal{C}(m,B)} n^{m/(2m+1)} E_f |f_n(x_0) - f(x_0)| \geq c > 0$$

for some constant c that depends on B only.

Proof. We shall only prove the case where $B \geq 2$ and $x_0 \in (0,1)$ for simplicity. Let f_0 be identically 1 on $[0,1]$, which clearly belongs to $\mathcal{C}(m,B)$ since $B \geq 1$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a function of compact support that is m-times continuously differentiable, $\int_{\mathbb{R}} \psi^2(x)dx = 1, \int_{\mathbb{R}} \psi(x)dx = 0$ and such that $|\psi(0)| > 0$. Such functions exist, for instance suitably translated Daubechies wavelets that we shall encounter later on. Define the alternative

$$f_1(x) = 1 + \epsilon 2^{-j_n(m+1/2)} \psi_{j_n}(x), \ x \in [0,1], \tag{71}$$

where $\psi_j(x) = 2^{j/2}\psi(2^j x - 2^j x_0)$, where $\epsilon > 0$ will be chosen below, and where $j \in \mathbb{N}$ is such that

$$2^{-j_n m} \simeq r_n(m), \quad r_n(m) = n^{-m/(2m+1)}. \tag{72}$$

For n and thus 2^{j_n} large enough the function ψ_{j_n} is supported in the interior of $[0,1]$, and since $\int_{\mathbb{R}} \psi = 0$ we infer $\int f_1 = 1$. Since ψ is bounded we can choose ϵ small enough depending only on $\|\psi\|_\infty, B$ such that $0 \leq f_1 \leq B$, so f_1 is a probability density bounded by B. Moreover, for $0 < \alpha \leq m$, $D^\alpha f_1 = \epsilon 2^{-j_n m} 2^{j_n \alpha}(D^\alpha \psi)(2^{j_n} x - 2^{j_n} x_0)$ so

$$\|D^\alpha f_1\|_\infty \leq \epsilon \|\psi\|_{m,\infty} \leq B$$

for ϵ small enough, so that $f_1 \in \mathcal{C}(m,B)$ for every $n \in \mathbb{N}$ large enough.

We now apply the arguments from the previous subsection with $d(f_0, f_1) = |f_0(x_0) - f_1(x_0)|$, $\mathcal{P} = \mathcal{C}(m,B)$, with P_{f_0} the product probability measure of a sample of size n from f_0 (so the uniform distribution on $[0,1]^n$) and with P_{f_1} the product probability measure on $[0,1]$ with density

$$\prod_{i=1}^n (1 + \epsilon 2^{-j_n(m+1/2)} \psi_{j_n}(x_i)).$$

The pointwise distance between f_0 and f_1 equals $|f_0(x_0) - f_1(x_0)| = \epsilon 2^{-j_n m}|\psi(0)| \geq c r_n(m)$, the constant c depending only on ϵ, ψ, so that the result will follow from (69) and Lemma 7 if we can show that for every $\delta > 0$ we can choose $\epsilon > 0$ small enough such that

$$(E_{f_0}|Z-1|)^2 \leq E_{f_0}(Z-1)^2 < \delta$$

for Z the likelihood ratio between P_{f_0} and P_{f_1}.

Writing (in slight abuse of notation) $j = j_n, \gamma_j = \epsilon 2^{-j_n(m+1/2)}$, using $\int \psi_{j_n}^2 =$

$1, \int \psi_{j_n} = 0$ repeatedly as well as $(1+x) \leq e^x$ we see

$$\begin{aligned} E_{f_0}(Z-1)^2 &= \int_{[0,1]^n} ((\Pi_{i=1}^n (1+\gamma_j \psi_j(x_i)) - 1))^2 \, dx \\ &= \int_{[0,1]^n} \Pi_{i=1}^n (1+\gamma_j \psi_j(x_i))^2 dx - 1 \\ &= \left(\int_{[0,1]} (1+\gamma_j \psi_j(x))^2 dx \right)^n - 1 = (1+\gamma_j^2)^n - 1 \leq e^{n\gamma_j^2} - 1. \end{aligned}$$

Now by (72) we see $n\gamma_j^2 = c\epsilon^2$ for some constant $c > 0$ and by choosing ϵ small enough we can make $e^{c\epsilon^2} - 1 < \delta$, which completes the proof by virtue of Lemma 7. □

The rate of convergence $n^{-m/(2m+1)}$ is strictly slower than $1/\sqrt{n}$, for $m = 1$ for instance we only have the rate $n^{-1/3}$. As $m \to \infty$ the model $\mathcal{C}(m, B)$ approaches a finite-dimensional model and the rate approaches the parametric rate $1/\sqrt{n}$.

Similar results can be proved in different metrics, such as $d^p(f,g) = \int_0^1 |f - g|(x)^p dx$ or $d(f,g) = \|f - g\|_\infty$, but the proofs require more refined lower bounds on multiple testing problems than the one in Lemma 7, see [79] for such results.

Intuitively the reason behind Theorem 16 is that two functions f_0, f_1 that differ by a large 'peak' on a very small neighborhood of the point x_0 may give rise to samples that look very much alike, so that such peaks are hard to detect statistically. A rigorous analysis of this intuition showed that the maximal size of an 'undetectable' peak depends on the smoothness bounds m, B for f_0, f_1, and on sample size, and it resulted in a minimax lower bound for the accuracy of estimating a density at a fixed point. We shall see in the next sections that this lower bound can be attained by concrete nonparametric estimators, but for this we have to review some basic approximation theory first.

3.3 Approximation of Functions

In contrast to estimation of the distribution function F, estimation of the density function f of F is a more intricate problem. Similar difficulties arise in non-parametric regression problems. One way to approach these difficulties is to first approximate f by a simpler function, $K(f)$ say, and then to estimate the simpler object $K(f)$ by $K_n(f)$ based on the sample. Evidently, to achieve a good overall performance we will have to 'balance' the approximation error $|K(f) - f|$ with the estimation error $|K_n(f) - K(f)|$, and we therefore first review some techniques from analysis on how to approximate arbitrary functions by simpler functions. Since a probability density is by definition an integrable function, we will focus here on

approximation of integrable functions, and – to obtain some stronger results – also on continuous functions.

Some notation: We recall that

$$L^p = \left\{ f : \int_{\mathbb{R}} |f(x)|^p dx < \infty \right\},$$

is the space of p-fold Lebesgue-integrable functions, $1 \leq p < \infty$, normed by $\|f\|_p := \left(\int_{\mathbb{R}} |f(x)|^p dx\right)^{1/p}$, and that

$$L^2 = \left\{ f : \int_{\mathbb{R}} f^2(x) dx < \infty \right\}$$

is the space of square Lebesgue-integrable functions normed by $\|f\|_2^2 := \int_{\mathbb{R}} f^2(x) dx$. Finally, a *locally integrable* function $f : \mathbb{R} \mapsto \mathbb{R}$ is a function that satisfies $\int_C |f(x)| dx < \infty$ for every bounded (Borel-) set $C \subset \mathbb{R}$.

3.3.1 Regularisation by Convolution

For two real valued functions f, g defined on $\mathbb{R}$, define their *convolution*

$$f * g(x) = \int_{\mathbb{R}} f(x-y) g(y) dy$$

if the integral exists. It is obvious that $f * g = g * f$. A simple way to approximate an arbitrary integrable, or a bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ is by the convolution $K_h * f$ of f with a suitably 'localized' kernel function

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right),$$

where K is integrable and satisfies $\int_{\mathbb{R}} K(x) dx = 1$ and where $h > 0$ governs the degree of 'localization'. (Informally speaking, a positive function $f : \mathbb{R} \to \mathbb{R}$ is 'localized around a point x' if most of the area under its graph is concentrated above a 'small interval centered at x'.) In the language of functional analysis, the functions K_h, as $h \to 0$, furnish an approximation of the identity operator on certain spaces of integrable functions. Furthermore, the quality of approximation increases if f is smoother (and if K is 'suitably regular'). We summarize some of these facts in the following proposition.

Proposition 6. *Let $f : \mathbb{R} \to \mathbb{R}$ be a (measurable) function and let $K : \mathbb{R} \to \mathbb{R}$ be an integrable function (the 'kernel') that satisfies $\int_{\mathbb{R}} K(u) du = 1$.*
i.) If f is bounded on $\mathbb{R}$ and continuous at x, then $K_h * f(x)$ converges to $f(x)$ as $h \to 0$.

ii.) If $\int_{\mathbb{R}} |f(x)|^p dx < \infty$ *for some* $1 \leq p < \infty$ *then* $\int_{\mathbb{R}} |K_h * f(x) - f(x)|^p dx \to 0$ *as* $h \to 0$.
Assume further that K *is a function symmetric about the origin that satisfies* $\int_{\mathbb{R}} |K(u)| u^2 du < \infty$ *and define* $\kappa(m) = \int_{\mathbb{R}} |K(u)||u|^m du$ *for* $m \leq 2$. *Suppose* f *is* m*-times differentiable on* $\mathbb{R}$, $m = 1, 2$.
iii.) If f *is bounded on* $\mathbb{R}$, *and if the* m*-th derivative of* f *is bounded on* $\mathbb{R}$, *by* D *say, then for every* $x \in \mathbb{R}$ *we have*

$$|K_h * f(x) - f(x)| \leq h^m 2^{1-m} D\kappa(m).$$

iv.) If $\int_{\mathbb{R}} |f(x)|^p dx$ *and* $D' := \int_{\mathbb{R}} |D^m f(x)|^p dx$ *both are finite, then*

$$\int_{\mathbb{R}} |K_h * f(x) - f(x)|^p dx \leq h^{pm} 2^{p(1-m)} D' \kappa(m)^p.$$

Proof. This proof is neither short nor difficult. We assume for simplicity that K is bounded, symmetric and has compact support, say in $[-a, a]$, and we also restrict ourselves to the case $p = 1$ as the general case is only slightly more involved but notationally inconvenient.

To prove i.), note first that boundedness of f implies that the integral $K_h * f$ converges, and we have

$$\begin{aligned} |K_h * f(x) - f(x)| &= \left| \int_{\mathbb{R}} h^{-1} K((x-y)/h) f(y) dy - f(x) \right| \\ &= \left| \int_{\mathbb{R}} K(u) f(x - uh) du - f(x) \right| \\ &= \left| \int_{\mathbb{R}} K(u)(f(x - uh) - f(x)) du \right| \qquad (73) \\ &\leq \int_{-a}^{a} |K(u)||f(x - uh) - f(x)| du \end{aligned}$$

where we have used the substitution $(x - y)/h \mapsto u$ and that $\int_{\mathbb{R}} K(u) du = 1$ by assumption. Let now $\varepsilon > 0$ arbitrary be given, and let $\delta > 0$ be such that $|f(x+v) - f(x)| < \varepsilon / \int_{-a}^{a} |K(u)| du$ for $|v| < \delta$. Such δ exists since f is continuous at x and since $\int_{-a}^{a} |K(u)| du$ is finite. Then $|uh| \leq ah < \delta$ for h small enough, so that the last expression in the last display is less than ε, proving this claim.

To prove ii.) ($p = 1$), we note first that the integral $K_h * f$ converges in view of boundedness of K and $f \in L^1$. We integrate the last inequality in the last display so that

$$\begin{aligned} \int_{\mathbb{R}} |K_h * f(x) - f(x)| dx &\leq \int_{\mathbb{R}} \int_{\mathbb{R}} |K(u)||f(x - uh) - f(x)| du dx \\ &= \int_{-a}^{a} |K(u)| \int_{\mathbb{R}} |f(x - uh) - f(x)| dx du \end{aligned}$$

using Fubini's theorem for the last identity. The last expression in the above display now converges to 0 as $h \to 0$ since $\sup_{u\in[-a,a]} \int_{\mathbb{R}} |f(x-uh) - f(x)|dx$ does in view of Exercise 27.

We next prove iv.), again only for $p = 1$. If $m = 1$, we write (understanding $\int_0^v$ as $-\int_v^0$ if $v < 0$)

$$f(x-uh) - f(x) = \int_0^{-uh} Df(x+t)dt \tag{74}$$

by the fundamental theorem of calculus if Df is continuous (and by absolute continuity of f otherwise, cf. Corollary 3.33 in [35]) and then we have from integrating (73), Fubini's theorem and change of variables that

$$\begin{aligned}
&\int_{\mathbb{R}} |K_h * f(x) - f(x)|dx = \int_{\mathbb{R}} \left| \int_{\mathbb{R}} K(u) \int_0^{-uh} Df(x+t)dtdu \right| dx \\
&\leq \int_{\mathbb{R}} |K(u)| \int_0^{-uh} \int_{\mathbb{R}} |Df(x)|dxdtdu \leq h\kappa(1)\|Df\|_1
\end{aligned}$$

which proves the case $m = 1$. If $m = 2$, use again (73) and expand f into a Taylor series up to second order about x with Laplacian representation of the remainder (e.g., (8.14.3) in [20]) to obtain

$$\begin{aligned}
&\int_{\mathbb{R}} |K_h * f(x) - f(x)|dx \\
&= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} K(u)[Df(x)uh + (uh)^2 \int_0^1 D^2 f(x - tuh)(1-t)dt]du \right| dx \\
&\leq h^2 \int_{\mathbb{R}} |K(u)|u^2 du \int_{\mathbb{R}} |D^2 f(x)|dx \int_0^1 (1-t)dt = h^2\kappa(2)\|D^2 f\|_1 2^{-1}
\end{aligned}$$

where we used that $\int K(u)udu = 0$ since K is symmetric around 0, Fubini's theorem and a change of variables. The proof of iii.), which is simpler than (and implicit in) iv.), is left to the reader. □

The above proposition allows to approximate functions that are at most twice differentiable in a good way, but one would expect an error bound of magnitude h^m as in iii.) even for $m > 2$. This can indeed be achieved by using 'higher order' kernels K, see Exercises 28, 29.

3.3.2 Approximation by Basis Functions

Another approach to approximate an arbitrary function $f : \mathbb{R} \to \mathbb{R}$ is to decompose it into sufficiently many linear combinations of much simpler basis functions e_k, so

that the approximation of $f(x)$ is of the form $\sum_k c_k(f)e_k(x)$ where $c_k(f)$ are some coefficients.

To understand this approach better we review here briefly some facts about Hilbert spaces. A complete normed linear space H whose norm $\|\cdot\|_H$ is given by $\|h\|_H^2 =< h, h >_H$, where $< \cdot, \cdot >_H$ is an inner product (i.e., a symmetric bilinear mapping from $H \times H$ to $\mathbb{R}$ or $\mathbb{C}$), is called a (real or complex) *Hilbert space*. An example is $\mathbb{R}^d$ with $< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^d x_i y_i$, but for us more important is the space L^2 which has inner product

$$< f, g >:= \int_{\mathbb{R}} f(x)g(x)dx. \tag{75}$$

In analogy to the Euclidean case we say that an element $h \in H$ is orthogonal to $h' \in H$ if $< h, h' >= 0$. If M is any closed linear subspace of H, its orthogonal complement is

$$M^- := H \ominus M = \{h \in H :< h, x >_H = 0 \text{ for every } x \in M\},$$

and H equals the direct sum $M^- \oplus M = \{x + y : x \in M, y \in M^-\}$, the sum being 'direct' because its summands are orthogonal to each other. If $\{e_k\}$ is an orthonormal basis for M, then the orthogonal projection $\pi_M(h)$ of any element $h \in H$ onto M has the representation

$$\pi_M(h) = \sum_k < h, e_k > e_k.$$

See Chapters 5.3 and 5.4 in [29] or Chapter 5.5 in [35] for more details and facts.

The classical example is the reconstruction of a periodic function on $(0, 1]$ by the partial sum of its *Fourier-series*, namely by

$$S_n(f)(x) = \sum_{k \in \mathbb{Z}: |k| \leq n} c_k(f)e^{2\pi i x k}$$

where $c_k(f) = (2\pi)^{-1} \int_0^1 f(x)e^{-2\pi i x k}dx$ are the Fourier-coefficients of f. Whereas this approximation is optimal in the space L^2, it can be very *bad* at any given point. In particular, the Fourier-series $S_n(f)(x)$ of 'most' continuous functions diverges at some $x \in (0, 1]$. (To be precise, the set of continuous periodic functions on $(0, 1]$ for which the Fourier series converges at all points can be shown to be a 'meagre' subset - in the sense of Baire categories - of the Banach space of continuous periodic functions on $(0, 1]$.) Can we find orthonormal bases for L^2 where these pathologies do not occur?

Another way to decompose a function into linear combinations of 'atoms' is by piecewise constant functions: The *Haar basis* (named after its inventor Haar (1910)) is the set of functions

$$\left\{\phi(\cdot - k), 2^{l/2}\psi(2^l(\cdot) - k), k \in \mathbb{Z}, l \in \mathbb{N} \cup \{0\}\right\}$$

where $\phi(y) = 1_{(0,1]}(y)$ and $\psi = 1_{[0,1/2]} - 1_{(1/2,1]}$, and where we write shorthand $\phi_k(x) = \phi(x-k)$ and $\psi_{lk} = 2^{l/2}\psi(2^l x - k)$. It is easily checked that this basis is orthogonal with respect to the L^2-inner product (75), in fact its linear span is dense in the space L^2 (see Exercise 30). Moreover, a locally integrable function $f : \mathbb{R} \to \mathbb{R}$ can be approximated by its Haar-projection, i.e., by the piecewise constant function

$$H_j(f)(x) = \sum_{k\in\mathbb{Z}} < f, \phi_k > \phi_k(x) + \sum_{l=0}^{j-1}\sum_{k\in\mathbb{Z}} < f, \psi_{lk} > \psi_{lk}(x) \tag{76}$$

For the Haar basis, one can prove the following analogue to Proposition 6:

Proposition 7. *Let $f : \mathbb{R} \to \mathbb{R}$ be a locally integrable function.*
i.) If f is continuous at $x \in \mathbb{R}$, then $H_j(f)(x)$ converges to $f(x)$ as $j \to \infty$.
ii.) If $\int_{\mathbb{R}} |f(x)|^p dx < \infty$ for some $1 \leq p < \infty$ then

$$\int_{\mathbb{R}} |H_j(f)(x) - f(x)|^p dx \to 0$$

as $j \to \infty$.
Suppose further that f is differentiable on $\mathbb{R}$.
iii.) If the derivative of f is bounded in absolute value on $\mathbb{R}$, by D say, then we have for every $x \in \mathbb{R}$ that

$$|H_j(f)(x) - f(x)| \leq 2^{-j} D.$$

iv.) If $\int_{\mathbb{R}} |f(x)|^p dx$ and $D' := \int_{\mathbb{R}} |Df(x)|^p dx$ both are finite then

$$\int_{\mathbb{R}} |H_j(f)(x) - f(x)|^p dx \leq 2^{-pj} D'.$$

We will prove this result as a special case of Proposition 9 below. The approximation by the Haar-basis is very simple and useful, but also has limitations. Comparing Proposition 7 with Proposition 6 for the approximation $K_h * f$, the question arises whether $m = 2$ could be considered in the case of the Haar basis as well. It turns out that the second part of Proposition 7 can only be proved with $m = 1$, that is, an analogue of Parts iii.) and iv.) of Proposition 6 with $m = 2$ does *not* exist for the Haar basis. The intuitive reason is that one can not approximate a smooth function *very well* by unsmooth functions. Roughly speaking one can say that if we want to approximate a m-times differentiable function in an optimal way, we should take the basis function of our approximation to be at least $m-1$ times differentiable. The Haar basis functions are not differentiable and so cannot 'detect' differentiability of order higher than one. Can this shortcoming of the Haar basis be circumvented, by considering 'smoother basis functions'?

From the Haar basis to B-splines$^+$*.*

One might be tempted to suggest to replace the simple function $1_{(0,1]}$ by, say, the basis function $N^{(2)}(x)$ given by x on $[0,1]$ and by $1-x$ on $[1,2]$. These 'hat functions' can indeed be used to approximate functions, and the family

$$\left\{N^{(2)}(\cdot - k), k \in \mathbb{Z}\right\}$$

is known as the *linear* B-spline basis (with integer breakpoints). To arrive at even smoother basis functions, it is useful to note that the hat function can be obtained from the Haar basis by virtue of

$$N^{(2)} = 1_{(0,1]} * 1_{(0,1]},$$

which motivates to define the B-spline of degree r iteratively by

$$N^{(r)} = 1_{(0,1]} * 1_{(0,1]} * .... * 1_{(0,1]} \qquad r - \text{times}. \tag{77}$$

For $r = 3$ these functions are called quadratic B-splines (with integer breakpoints), and the case $r = 4$ corresponds to cubic B-splines. It is easily checked that $N^{(r)}$ is $r-2$ times differentiable (and in fact $r-1$ times 'weakly' differentiable, which is the relevant notion here).

Theorem 17 (Curry-Schoenberg (1966))**.** *The dyadic B-spline family of degree r,*

$$\left\{N_{lk}^{(r)} := N^{(r)}(2^l(\cdot) - k), l \in \mathbb{N} \cup \{0\}, k \in \mathbb{Z}\right\}$$

is a basis for the linear ('Schoenberg'-) space of piecewise polynomials of order $r-1$ *with dyadic breakpoints* $\{2^{-j}k\}_{k\in\mathbb{Z}}$*.*

The space of piecewise polynomials is here (and usually, but not necessarily) taken to consist of functions that are $r-2$ times continuously differentiable at the breakpoints. One can also choose a grid of breakpoints different from the integers and obtain the same result, but in this case the family $N_{lk}^{(r)}$ cannot be described in simple terms of translates and dilates of the basic function $N^{(r)}$.

We will not prove this theorem (which belongs in a course on approximation theory) and we refer to p.141 in [18] for a proof and the exact definition of the *Schoenberg* spaces. The theorem tells us that every piecewise polynomial P of degree $r-1$ with dyadic breakpoints $\{2^{-j}k\}_{k\in\mathbb{Z}}$ can be written as

$$P(x) = \sum_k c_{k,j}(P) N^{(r)}(2^j x - k)$$

for some suitable coefficients $c_{k,j}(P)$. This is a 'sparse' representation of P since we only need a few translates and dilates of the fixed basis function $N^{(r)}(x)$ to

reconstruct $P(x)$ at a given point $x \in \mathbb{R}$ (note that $N_{jk}^{(r)}$ is supported in $[k/2^j, (k+r)/2^j]$). This can be used in various simple and computationally efficient ways to approximate integrable and/or continuous functions as in Proposition 6, and these approximations can be shown to improve in quality if we refine the partition of breakpoints, that is, if we increase j. For example, one can prove the following proposition:

Proposition 8. *Let $f : \mathbb{R} \to \mathbb{R}$ be bounded, continuous at x and suppose f is r-times differentiable, with r-th derivative uniformly bounded by D. Then there exist coefficients $\{c_k(f,r)\}_{k\in\mathbb{Z}}$ such that for $P_f(x) = \sum_k c_k(f,r)N_{jk}^{(r)}(x)$ we have*

$$|P_f(x) - f(x)| \leq c2^{-jr}$$

where the constant c depends only on r and D.

Again, this result belongs to approximation theory. It can be deduced without too many complications from Theorem 12.2.4 in [18].

It is not clear from the proposition how the coefficients $c_k(f,r)$ should be chosen in practice. A good way to do this is to choose them by *projection* from the space of square-integrable functions onto the Schoenberg-space. This projection inherits the approximation properties from Proposition 8 and can be computed easily (by simple linear algebra), cf., e.g., p.401f. in [18].

Whereas the B-spline basis give us a simple and localized way to approximate many functions f by piecewise polynomials, it is still not satisfactory for all purposes since, when compared to the Haar basis, one looses 'orthogonality' of the translates. One verifies easily that for every k (and $r > 1$)

$$< N^{(r)}(\cdot - k), N^{(r)}(\cdot - k - 1) > \neq 0,$$

which can be inconvenient.

3.3.3 Orthornormal Wavelet Bases

The question remains whether one can find a set of basis functions that is *simultaneously*
a) *orthogonal* with respect to the L^2-inner product $< \cdot, \cdot >$,
b) *localized* in the sense that $f(x)$ is approximated, in a sparse way, by just a few basis functions with support close to x and
c) a good *'approximator'* in the sense that an analogue of Proposition 6 can be proved, possibly even with $m \in \mathbb{N}$ arbitrary.

For some time it was thought that this was impossible, but this question was eventually answered in the positive by the advent of *wavelets* around 1990 (Meyer

(1992), Daubechies (1992) and others). In fact, the first wavelets were constructed by 'orthogonalizing' the B-spline bases from Theorem 17 (by virtue of a 'Gram-Schmidt' orthogonalization procedure), but the most striking breakthrough was the construction of *compactly supported* wavelets that have an arbitrary degree of smoothness: so called *Daubechies' wavelets*.

The general definition of a wavelet basis (whose existence has to be ensured) is the following: Let $\phi \in L^2(\mathbb{R})$ be a *scaling function* ('father wavelet'), that is, ϕ is such that

$$\{\phi(\cdot - k) : k \in \mathbb{Z}\}$$

is an orthonormal system in L^2, and moreover the linear spaces

$$V_0 = \left\{ f(x) = \sum_k c_k \phi(x-k), \{c_k\}_{k\in\mathbb{Z}} : \sum_k c_k^2 < \infty \right\},$$

$$V_1 = \{h(x) = f(2x) : f \in V_0\},$$

$$..., V_j = \{h(x) = f(2^j x) : f \in V_0\}, ...,$$

are nested ($V_{j-1} \subset V_j$ for $j \in \mathbb{N}$) and such that $\cup_{j\geq 0} V_j$ is dense in L^2. Such ϕ exists, e.g., the Haar function $\phi = 1_{(0,1]}$ (see Exercise 30), but other examples will be discussed below.

Moreover, since the spaces V_j are nested, there are nontrivial subspaces of L^2 obtained from taking the orthogonal complements $W_l := V_{l+1} \ominus V_l$, indeed, we can 'telescope' these orthogonal complements to see that the space V_j can be written as

$$\begin{aligned} V_j &= V_0 \oplus V_1 \ominus V_0 \oplus V_2 \ominus V_1 \oplus ..... \oplus V_{j-1} \ominus V_{j-2} \oplus V_j \ominus V_{j-1} \\ &= V_0 \oplus \left(\bigoplus_{l=0}^{j-1} W_l \right). \end{aligned} \tag{78}$$

If we want to find the othogonal projection of $f \in L^2$ onto V_j, then the above orthogonal decomposition of V_j tells us that we can describe this projection as the projection of f onto V_0 plus the sum of the projections of f onto W_l from $l = 0$ to $j-1$.

Now clearly the projection of f onto V_0 is

$$K_0(f)(x) = \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x),$$

where we write $\phi_k = \phi(\cdot - k)$. To describe the projections onto W_l, we would like to find basis functions that span the spaces W_l, and this is where the 'mother' *wavelet*

enters the stage: Assume that there exists a fixed function ψ, the ('mother') wavelet, such that, for every $l \in \mathbb{N} \cup \{0\}$,

$$\left\{\psi_{lk} := 2^{l/2}\psi(2^l(\cdot) - k) : k \in \mathbb{Z}\right\}$$

is an orthonormal set of functions that spans W_l. Again, such ψ exists (take the Haar-wavelet $\psi = 1_{[0,1/2]} - 1_{(1/2,1]}$, cf. Exercise 30), but other examples can be constructed, and they will be discussed below. The projection of f onto W_l is

$$\sum_k < \psi_{lk}, f > \psi_{lk}$$

and, adding things up, we see that the projection $K_j(f)$ of f onto V_j is given by

$$K_j(f)(x) = \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x) + \sum_{l=0}^{j-1}\sum_{k\in\mathbb{Z}} < \psi_{lk}, f > \psi_{lk}(x). \tag{79}$$

It should be clear that for ϕ equal to the Haar wavelet, this projection exactly corresponds to the quantity $H_j(f)$ from (76).

This projection is the partial sum of what is called the *wavelet series* of a function $f \in L^2$: To understand this, note first that, since $\cup_{j\geq 0}V_j$ is dense in L^2 we necessarily conclude from (78) that the space L^2 can be decomposed into the direct sum

$$L^2 = V_0 \oplus \left(\bigoplus_{l=0}^{\infty} W_l\right),$$

so that the set of functions

$$\{\phi(\cdot - k), 2^{l/2}\psi(2^l(\cdot) - k) : k \in \mathbb{Z}, l \in \mathbb{N} \cup \{0\}\} \tag{80}$$

is an orthonormal basis of the Hilbert space L^2. As a consequence, every $f \in L^2$ has the wavelet series expansion

$$f(x) = \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi_k(x) + \sum_{l=0}^{\infty}\sum_{k\in\mathbb{Z}} < \psi_{lk}, f > \psi_{lk}(x) \tag{81}$$

where convergence is guaranteed at least in the space L^2.

Now the question arises whether functions ϕ and ψ besides the Haar basis exist such that the class of functions (80) is an orthonormal basis in L^2 (and such that the associated spaces V_j are nested). In fact, for several reasons, we would like the basis functions ϕ and ψ to be both smooth and compactly supported. The following assumption will formalize this desire. The symbols ϕ_k, ψ_{lk} and the spaces V_j and W_l are defined as above.

Condition 1. (S) *We say that an orthonormal system* $\{\phi_k, \psi_{lk} : k \in \mathbb{Z}, l \in \mathbb{N} \cup \{0\}\}$ *is an* S*-regular compactly supported wavelet basis if the following holds:*
a) ϕ *and* ψ *are bounded and have compact support,*
b) $\{\phi_k\}_{k\in\mathbb{Z}}$ *spans* V_0, $\{\psi_{lk}\}_{k\in\mathbb{Z}}$ *spans* W_l*, the associated spaces* $\{V_j\}_{j\geq 0}$ *are nested and* $\cup_{j\geq 0} V_j$ *is dense in* L^2*,*
c) one of the following two conditions is satisfied: Either i) ϕ *has bounded derivatives up to order* S*; or ii)* ψ *satisfies* $\int_{\mathbb{R}} x^i \psi(x)dx = 0$*,* $i = 0, ..., S$*.*

The Haar wavelets, corresponding to $\phi = 1_{(0,1]}$ and $\psi = 1_{(0,1/2]} - 1_{(1/2,1]}$, satisfy this condition only for $S = 0$. The following fundamental result is due to Daubechies (1988).

Theorem 18 (Daubechies (1988))**.** *For any given* S *there exist wavelets* ϕ *and* ψ *that satisfy Condition (S).*

These wavelet bases also always satisfy $\int_{\mathbb{R}} \phi(x)dx = 1, \int_{\mathbb{R}} \psi(x)dx = 0$. The proof of Theorem 18 uses nontrivial Fourier analysis, and can be found, e.g., in Daubechies (1992) or Meyer (1992). (The one in Daubechies (1988) is not recommended for first reading.)

It remains to obtain an analogue of Propositions 6, 7 for wavelets. The first two claims of the following result show that the wavelet series (81) converges in L^p for every $1 \leq p < \infty$ and pointwise for any continuous function, so remarkably outperforming Fourier series (and other orthonormal bases of L^2).

Proposition 9. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a locally integrable function and let* ϕ, ψ *be a wavelet basis that satisfies Condition 1(S) for some* $S \geq 0$*. Denote by* $K_j(f)$ *the wavelet projection (79).*
i.) If f *is continuous at* x*, then* $K_j(f)(x)$ *converges to* $f(x)$ *as* $j \to \infty$*.*
ii.) If $\int_{\mathbb{R}} |f(x)|^p dx < \infty$ *for some* $1 \leq p < \infty$ *then* $\int |K_j(f)(x) - f(x)|^p dx \to 0$ *as* $j \to \infty$*.*
Suppose further that f *is* m*-times differentiable on* $\mathbb{R}$*,* $m \leq S + 1$*.*
iii.) If f *is bounded on* $\mathbb{R}$*, and if the* m*-th derivative of* f *is bounded on* $\mathbb{R}$*, by* D *say, then for every* $x \in \mathbb{R}$ *we have*

$$|K_j(f)(x) - f(x)| \leq C 2^{-jm}$$

for some constant C *that depends only on* D *and* ϕ*.*
iv.) If $\int_{\mathbb{R}} |f(x)|^p dx$ *and* $D' := \int_{\mathbb{R}} |D^m f(x)|^p dx$ *both are finite then*

$$\int_{\mathbb{R}} |K_j(f)(x) - f(x)|^p dx \leq C' 2^{-jpm}$$

where C' *depends only on* D', ϕ *and* p*.*

Proof. We will prove the result for the special case of the Haar wavelet (and hence only for $m = 1$) to illustrate the main ideas, and we also restrict ourselves to $p = 1$. The general case will be discussed at the end.

It follows from the definitions that

$$\{\phi_{jk} := 2^{j/2}\phi(2^j x - k) : k \in \mathbb{Z}\}$$

is an orthonormal basis for the space V_j. Consequently the projection $K_j(f)$ can also be written as

$$K_j(f)(x) = \sum_{k\in\mathbb{Z}} < \phi_{jk}, f > \phi_{jk}(x) = 2^j \int_{\mathbb{R}} K(2^j x, 2^j y) f(y) dy \tag{82}$$

where

$$K(x, y) = \sum_{k\in\mathbb{Z}} 1_{(0,1]}(x - k) 1_{(0,1]}(y - k)$$

(note that, for fixed x, all sums here are finite, so no convergence issues of these series have to be addressed). Moreover, since $(k, k + 1]$, $k \in \mathbb{Z}$, forms a partition of $\mathbb{R}$ and since $\int_{\mathbb{R}} 1_{(0,1]}(y)dy = 1$ we have

$$\int_{\mathbb{R}} K(x, y)dy = \sum_{k\in\mathbb{Z}} 1_{(0,1]}(x - k) = 1 \tag{83}$$

for every $x \in \mathbb{R}$, and by substitution we also have $2^j \int_{\mathbb{R}} K(2^j x, 2^j y)dy = 1$. Furthermore, since the support of ϕ is $(0, 1]$, we have, for every x,

$$0 \le K(2^j x, 2^j x - u) \le 1_{[-1,1]}(u). \tag{84}$$

Using these facts, we obtain, substituting $2^j y \mapsto 2^j x - u$

$$\begin{aligned}
|K_j(f)(x) - f(x)| &= \left| 2^j \int_{\mathbb{R}} K(2^j x, 2^j y) f(y) dy - f(x) \right| \\
&= \left| 2^j \int_{\mathbb{R}} K(2^j x, 2^j y)(f(y) - f(x)) dy \right| \\
&= \left| \int_{\mathbb{R}} K(2^j x, 2^j x - u)(f(x - 2^{-j}u) - f(x)) du \right| \\
&\le \int_{\mathbb{R}} |K(2^j x, 2^j x - u)||f(x - 2^{-j}u) - f(x)| du \qquad (85) \\
&\le \sup_{u\in[-1,1]} |f(x - 2^{-j}u) - f(x)| \int_{-1}^{1} du \to 0
\end{aligned}$$

as $j \to \infty$ by continuity of f at x. This proves i.). Part ii.) is proved by integrating (85) and using again (84) together with continuity of translation in L^1, see Exercise

Q8. To prove part iii.), we apply the mean value theorem to (85) and use (84) to obtain

$$|K_j(f)(x) - f(x)| \leq 2^{-j}\|Df\|_\infty \int_{-1}^{1} |u| du = 2^{-j} D.$$

Part iv.) is proved by again integrating (85) with respect to x, using the fundamental theorem of calculus as in (74), and then proceeding as in the proof of Part iv.) of Proposition 6), using also (84).

The proof for general wavelets essentially reduces to obtaining properties analogous to (83) and (84) for general scaling functions ϕ. If ϕ is bounded and compactly supported, in $[0, a]$ say (up to a translation), and if $K(x, y) = \sum_k \phi(x-k)\phi(y-k)$, then clearly

$$|K(2^j x, 2^j x - u)| \leq c\|\phi\|_\infty^2 1_{[-a,a]}(u) \tag{86}$$

for some fixed constant c that depends only on a, which can be used to replace (84) in the above argument. To deal with $m > 1$, one also needs the property that the wavelet projection kernel $K(x, y)$ reproduces polynomials, i.e., $\int_{\mathbb{R}} K(x, y) y^\alpha = x^\alpha$ for every integer $0 \leq \alpha \leq S$. This nontrivial result can be established for the wavelet bases from Condition 1(S) using Fourier analytic methods. They can be found, e.g., in [58] or Chapters 8 and 9 of [47]. □

3.3.4 Exercises

Exercise 27. Use the fact that the set of continuous functions with compact support is dense in the normed space L^1 of integrable functions to show that the mapping $h \mapsto \int |f(x+h) - f(x)| dx$ is continuous at 0 for every integrable function f.

Exercise 28. Higher order kernels. A kernel is said to be of order l if it integrates to one, if $\int_{\mathbb{R}} K(u) u^m du = 0$ for every $m = 1, ..., l$, and if

$$\kappa(l) = \int_{\mathbb{R}} |K(u)| u^{l+1} du < \infty.$$

Any compactly supported symmetric probability density $K : \mathbb{R} \to \mathbb{R}$ is a kernel of order 1 (why?). To construct kernels of higher order than 1, let $\{\phi_m\}_{m \in \mathbb{N}}$ be the orthonormal basis in $L^2([-1, 1])$ of Légendre polynomials defined by

$$\phi_0(x) := 2^{-1/2}, \quad \phi_m(x) = \sqrt{\frac{2m+1}{2}} \frac{1}{2^m m!} \frac{d^m}{dx^m}[(x^2 - 1)^m]$$

for $x \in [-1, 1]$ and $m \in \mathbb{N}$. Define, for $l \in \mathbb{N}$,

$$K^{(l)}(u) = \sum_{m=0}^{l} \phi_m(0)\phi_m(u) 1\{|u| \leq 1\}.$$

Given an exact formula for $K^{(2)}(x)$, and show that it is a kernel of order 2. Is it of order 3? Of order 4? Show in general that $K^{(l)}$ defines a kernel of order l. (In doing so, you may use the fact that $\{\phi_m\}_{m\in\mathbb{N}}$ is orthonormal w.r.t. the inner product $< f, g >_{[-1,1]} := \int_{-1}^{1} fg$.)

Exercise 29. Approximation with higher order kernels. Suppose $f : \mathbb{R} \to \mathbb{R}$ is a bounded function that is m-times differentiable, with m-th derivative bounded by D. If m is any positive integer, show that one can devise a kernel K such that $|K_h * f(x) - f(x)| \leq Ch^m$ where the constant C depends only on K and D. (Use the previous exercise.)

Exercise 30. *[Haar basis.]* If

$$V_0 = \left\{ f(x) = \sum_k c_k \phi(x-k), \{c_k\}_{k\in\mathbb{Z}} : \sum_k c_k^2 < \infty \right\},$$

$$V_j = \{h(x) = f(2^j x) : f \in V_0\},$$

give a description of the spaces V_j, $j \geq 0$, and

$$W_j = V_{j+1} \ominus V_j$$

if ϕ, ψ is the Haar wavelet basis. Why are the V_j's nested? Verify that $\psi_{lk} := 2^{l/2}\psi(2^l(\cdot) - k)$ is orthogonal to $\psi_{l'k}$ for $l \neq l'$. Recalling the construction of the Lebesgue integral, why is it obvious that the union of the V_j's is dense in the normed space L^1 of Lebesgue-integrable functions?

3.4 Density Estimation on $\mathbb{R}$

We now return to the situation where we observe an i.i.d. sample $X_1, ..., X_n$ of a random variable X with distribution function F. Suppose we know further that F is absolutely continuous with probability density function $f : \mathbb{R} \to \mathbb{R}$, so that

$$P(X \in A) = \int_A f(x)dx$$

for every Borel set A. Can we estimate the object f in a similar way as we were able to estimate F? Certainly, we cannot use the empirical distribution function F_n to do this, since F_n does not possess a density. Are there any 'natural estimators' for a density f? At first sight the answer might be no: indeed, in contrast to distribution functions, which are nondecreasing right-continuous bounded functions, a probability density is potentially a very erratic object, that can be

unbounded, have wild oscillations, cusps, uncountably many jumps, etc. These facts just reflect the richness of the infinite-dimensional set

$$\left\{ f : \mathbb{R} \to \mathbb{R}, f \geq 0, \int_{\mathbb{R}} f(x)dx = 1 \right\}$$

of densities in L^1. On the other hand, the results from Section 3.3 taught us that arbitrary integrable functions f can be approximated by much simpler functions $K(f)$, and this applies to densities as well. The fact that approximations $K(f)$ of the type introduced in the last section can usually be estimated from the sample $X_1, ..., X_n$ in a reasonable way nourishes our hope to construct a reasonable estimator for the ultimate object of interest, f. The different types of approximation $K(f)$ then give rise to different choices for 'density estimators'.

Another question is the 'loss function' in which we would like to assess the performance of an estimator. Whereas sup-norm loss seems natural to estimate a function, this introduces a priori restrictions in the case of densities (which can be unbounded so that $\|f\|_\infty = \infty$!). At first sight then L^1-loss $\|f - g\|_1 = \int_{\mathbb{R}} |f(x) - g(x)|dx$ seems natural, as it is defined for all densities. If one is willing to assume more on the density f, one can also consider L^2-loss and pointwise loss, where the theory is simpler. If one assumes that f is uniformly continuous, then the stronger sup-norm loss $d(f, g) = \sup_{x \in \mathbb{R}} |f(x) - g(x)|$ is also of interest.

3.4.1 Kernel Density Estimators

From Proposition 6 one candidate to approximate f is its convolution with a kernel K_h, i.e.,

$$K_h * f(x) = \frac{1}{h} \int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f(y)dy.$$

This is an integral of a fixed function against a probability density, and it can be estimated in a natural, unbiased way by

$$f_n^K(h, x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{87}$$

which is known as the *kernel density estimator*, introduced by Akaike (1954), Rosenblatt (1956) and Parzen (1962) and much studied since then.

We start with a positive result, which implies that the kernel density estimator is consistent in L^1-loss for *any* density f.

Theorem 19. *Let $X_1, ..., X_n$ be i.i.d. with arbitrary density f and let f_n^K be the kernel density estimator from (87), where $K \in L^2$ is nonnegative and satisfies*

$\int_{\mathbb{R}} K(u)du = 1$. *If* $h := h_n$ *is chosen in dependence of* n *such that* $h_n \to 0$ *and* $nh_n \to \infty$, *then*

$$E\|f_n^K(h) - f\|_1 \to 0$$

as $n \to \infty$.

Proof. We have the 'variance-bias' decomposition

$$\int_{\mathbb{R}} |f_n^K(h,x) - f(x)|dx \leq \int_{\mathbb{R}} |f_n^K(h,x) - K_h * f(x)|dx + \int_{\mathbb{R}} |K_h * f(x) - f(x)|dx,$$

where the first term is random and the second is not. In fact, the second term converges to zero as $n \to \infty$ by Proposition 6. For the first term, note that $h^{-1}\int_{\mathbb{R}} K((x-y)/h)dx = 1$ for every y and Fubini's theorem imply

$$\int_{\mathbb{R}} (f_n^K(h,x) - K_h * f(x))dx = 1 - \int_{\mathbb{R}}\int_{\mathbb{R}} h^{-1}K((x-y)/h)dx f(y)dy = 0,$$

and hence the integral of the positive part of $(f_n^K(h,x) - K_h * f(x))$ has to equal the integral of the negative part, so that, using again Fubini's theorem and since $E|X| \leq (EX^2)^{1/2}$ for any random variable X we have

$$\begin{aligned}
E\int_{\mathbb{R}} |f_n^K(h,x) - K_h * f(x)|dx &= 2E\int_{\mathbb{R}} (K_h * f(x) - f_n^K(h,x))_+ dx \\
&= 2\int_{\mathbb{R}} E(K_h * f(x) - f_n^K(h,x))_+ dx \\
&\leq 2\int_{\mathbb{R}} \min(K_h * f(x), E(K_h * f(x) - f_n^K(h,x))_+)dx \\
&\leq 2\int_{\mathbb{R}} \min(f(x), (E(K_h * f(x) - f_n^K(h,x))^2)^{1/2})dx \\
&\quad +2\int_{\mathbb{R}} |K_h * f(x) - f(x)|dx.
\end{aligned}$$

The second term in the last expression is identical to twice the bias term and hence again converges to zero as $n \to \infty$ by Proposition 6. For the first term note that $(K_h * f(x) - f_n^K(h,x)) = n^{-1}\sum_{i=1}^n (EZ_i - Z_i)$ where $Z_i = h^{-1}K((x-X_i)/h)$ are i.i.d. random variables, so that

$$[E(K_h * f(x) - f_n^K(h,x))^2]^{1/2} \leq \sqrt{1/n}(EZ_i^2)^{1/2}. \tag{88}$$

To proceed, we consider first the simpler case where $K = 1_{[-1/2,1/2]}$ so that $K^2 = K$. We then see that $EZ_i^2 = \frac{1}{h}K_h * f(x)$ and

$$K_{h_n} * f(x) = \frac{1}{h_n}\int_{x-(h_n/2)}^{x+(h_n/2)} f(y)dy \to f(x)$$

as $n \to \infty$ for almost every $x \in \mathbb{R}$ by Lebesgue's differentiation theorem (e.g., Theorem 3.21 in [35]). Consequently

$$\frac{1}{nh_n} K_{h_n} * f(x) = \frac{1}{nh_n}(K_{h_n} * f(x) - f(x)) + \frac{1}{nh_n} f(x) \to 0$$

as $n \to \infty$ for almost every $x \in \mathbb{R}$ since $nh_n \to \infty$ by assumption. We conclude that

$$\min(f(x), (E(K_{h_n} * f(x) - f_n^K(h_n, x))^2)^{1/2}) \leq \min\left(f(x), \sqrt{\frac{1}{nh_n} K_{h_n} * f(x)}\right) \to 0$$

as $n \to \infty$ for almost every $x \in \mathbb{R}$, and since this quantity is (pointwise) bounded from above by the integrable function f, its integral also converges to zero in view of the dominated convergence theorem (Exercise 2). This completes the proof for this choice of K, and the case of general K follows from a simple reduction from general kernels to $K = 1_{[-1/2,1/2]}$, see Theorem 9.2 in Devroye and Lugosi (2001), from where this proof is taken. □

This theorem shows that one can estimate an unknown density consistently by a kernel estimator. One may be tempted to view Theorem 19 as a density-analogue of the Glivenko-Cantelli theorem (Theorem 8), and go further and ask for the rate of convergence to zero in Theorem 19. The following proposition contains some sobering facts, that we will not prove here (see Devroye and Lugosi (2001, p.85)):

Proposition 10. *Let $X_1, ..., X_n$ be i.i.d. random variables with density f, and let f_n^K be the kernel estimator from (87), where $K \in L^1$ is nonnegative and integrates to one. Then*

$$\sup_{f: f \geq 0, \int f = 1} \inf_{h > 0} E\|f_n^K(h) - f\|_1 = 2,$$

and for any sequence $a_n \to 0$ there exists a density f such that for all n large enough

$$\inf_h E\|f_n^K(h) - f\|_1 > a_n.$$

So in general there is 'no rate of convergence' in Theorem 19, and the kernel estimator is not uniformly consistent for the model of all probability densities.

Pointwise Risk Bounds.

Proposition 10 is related to the minimax lower bounds from Section 3.2. In the lower bound from Theorem 16 quantitative assumptions were made on the existence and size of derivatives of f, and the question arises how the kernel estimator performs under such assumptions.

Proposition 11. *Let $X_1, ..., X_n$ be i.i.d. random variables with bounded density f, and let f_n^K be the kernel estimator from (87). Suppose the kernel $K \in L^2$ satisfies the conditions of Part iii.) of Proposition 6. If the m-th derivative of f, $m \leq 2$, is bounded in absolute value on $\mathbb{R}$, by D say, then for every $x \in \mathbb{R}$, every $n \in \mathbb{N}$ and every $h > 0$ we have*

$$E|f_n^K(h,x) - f(x)| \leq \sqrt{\frac{1}{nh}}\|f\|_\infty^{1/2}\|K\|_2 + h^m 2^{1-m} D\kappa(m).$$

Proof. As usual

$$E|f_n^K(h,x) - f(x)| \leq E|f_n^K(h,x) - K_h * f(x)| + |K_h * f(x) - f(x)|.$$

The second term is bounded by using Proposition 6. To bound the first term, we have from (88)

$$\begin{aligned} E(f_n^K(h,x) - K_h * f(x))^2 &\leq \frac{1}{h^2 n} E K^2((x-X)/h) \\ &= \frac{1}{nh}\int_{\mathbb{R}} K^2(u) f(x-uh)du \leq \frac{1}{nh}\|K\|_2^2\|f\|_\infty \end{aligned}$$

which completes the proof. □

It follows directly from the proof of Proposition 6 that, if K is supported in $[-a, a]$, then there is a 'local' version of the above result, where f is only required to be m-times differentiable in a neighborhood of x, and where D has to be replaced by $\sup_{y\in[x-ha,x+ha]} |D^m f(y)|$.

A result similar to Proposition 11 can be obtained for $m > 2$, using the same proof and Exercises 28, 29. The first term in the above decomposition is often called the 'variance term', whereas the second one is called the 'bias term'. We see that, as h decreases, the first term increases whereas the second one decreases. Hence it makes sense to choose h such that the two terms are of the same size. Straightforward algebra shows that this happens when

$$h \simeq \left(\frac{1}{n}\right)^{\frac{1}{2m+1}} \left(\frac{\|f\|_\infty^{1/2}\|K\|_2}{D\kappa(m)}\right)^{\frac{1}{m+1/2}}. \tag{89}$$

This choice is statistically not feasible: The constants D and $\|f\|_\infty$ are generally not known to the statistician, but they can be estimated (D only under additional assumptions). A more fundamental problem is that m is unknown, and this parameter can in fact not be reliably estimated. However, it is interesting to consider

for the moment the case where m *is* known: then *any* choice $h_n^* \simeq n^{-1/(2m+1)}$ would produce an estimator with the *rate of convergence*

$$\sup_{f:\|f\|_\infty + \|D^m f\|_\infty \leq B} E|f_n^K(h_n^*, x) - f(x)| \leq C \left(\frac{1}{n}\right)^{m/(2m+1)} \tag{90}$$

where the constant C does only depend on the number B and the kernel K. Thus the kernel estimator with this bandwidth choice achieves the lower bound for the minimax rate of convergence in pointwise loss derived in Theorem 16. This implies in particular that $n^{-m/(2m+1)}$ is the minimax rate of convergence in this problem. A rate of convergence is a relatively crude probabilistic result. However, the moment bound (90) is quite precise for every $n \in \mathbb{N}$: the probabilistic fluctuations of the estimation error measured at the scale $n^{-m/(2m+1)}$ satisfy a good exponential inequality, see Exercise 32.

Another question is whether this result can be made uniform in all $x \in \mathbb{R}$, similar to the results for the empirical distribution function. This can be done, but it requires substantially more techniques, mostly from empirical process theory, and was not done until recently, see [38].

Pointwise Asymptotic Distribution of Kernel Estimators

Whereas the results from the previous section have shown us that $f_n^K(x)$ converges to $f(x)$ in probability under certain assumptions, we cannot straightforwardly use this for statistical inference. Ideally, if we want to estimate f at the point x, we would like to have exact confidence statements of the form

$$\Pr\left(f(x) \in [f_n^K(h,x) - c(n,\alpha,x,K), f_n^K(h,x) + c(n,\alpha,x,K)\right) \geq 1 - \alpha$$

where α is some significance level and where $c(n,\alpha,x,K)$ is a sequence of constants that one would like to be as small as possible (given α). The kernel estimator from the previous section can be used for this, if one 'undersmooths' slightly, by virtue of the following exact distributional limit theorem.

Proposition 12. *Let $f_n^K(h,x)$ be the kernel density estimator from (87), where K is bounded and satisfies the conditions from Part iii.) of Proposition 6. Suppose f is m-times differentiable, $m = 1, 2$, and that f and $D^m f$ are bounded. If $h_n \to 0$ is chosen such that $nh_n \to \infty$ but $\sqrt{n} h_n^{m+1/2} \to 0$ as $n \to \infty$, then*

$$\sqrt{nh_n}\left(f_n^K(h_n,x) - f(x)\right) \to^d N(0, f(x)\|K\|_2^2)$$

as $n \to \infty$.

Proof. We have to prove this limit theorem only for $\sqrt{nh}\left(f_n^K(h_n,x) - K_h * f(x)\right)$ since

$$\sqrt{nh}\,|K_h * f(x) - f(x)| \leq C\sqrt{nh}h^m \to 0$$

for some fixed constant C by Proposition 6 and assumption on h_n.

To prove the limit for the 'variance term', we use the Lindeberg-Feller central limit theorem for triangular arrays of i.i.d. random variables, which reads as follows (e.g., [81], p.20): For each n let $Y_{n1}, ..., Y_{nn}$ be i.i.d. random variables with finite variances. If, as $n \to \infty$, i) $nEY_{ni}^2 1\{|Y_{ni}| > \varepsilon\} \to 0$ for every $\varepsilon > 0$ and ii) $nE(Y_{ni} - EY_{ni})^2 \to \sigma^2$, then $\sum_{i=1}^n (Y_{ni} - EY_{ni}) \to^d N(0, \sigma^2)$ as $n \to \infty$.

We apply this theorem with

$$Y_{ni} = \sqrt{nh_n} \frac{1}{nh_n} K\left(\frac{x - X_i}{h_n}\right) = \sqrt{\frac{1}{nh_n}} K\left(\frac{x - X_i}{h_n}\right)$$

so that we have, similar as before (88),

$$\sqrt{nh_n}\left(f_n^K(h_n, x) - K_{h_n} * f(x)\right) = \sum_{i=1}^n (Y_{ni} - EY_{ni}),$$

and it remains to verify the hypotheses from above. Clearly,

$$\begin{aligned} nEY_{ni}^2 &= \frac{1}{h_n} \int_{\mathbb{R}} K^2\left(\frac{x - y}{h_n}\right) f(y) dy \\ &= \int_{\mathbb{R}} K^2(u) f(x - uh_n) du \to f(x) \|K\|_2^2 \end{aligned}$$

as $n \to \infty$ by the dominated convergence theorem, since f is continuous at x and bounded on $\mathbb{R}$, and since $K \in L^1 \cap L^\infty \subset L^2$. Furthermore, $|Y_{ni}| \leq (nh_n)^{-1/2} \|K\|_\infty \to 0$ as $n \to \infty$ by assumption on h_n, so also

$$1\{|Y_{ni}| > \varepsilon\} \to 0$$

for every $\varepsilon > 0$. This already verifies Condition i), and ii) follows from $E(Y_{ni} - EY_{ni})^2 = EY_{ni}^2 - (EY_{ni})^2$, the limit $nEY_{ni}^2 \to f(x)\|K\|_2^2$ as established above and since

$$\begin{aligned} n(EY_{ni})^2 &= \frac{1}{h_n} \left(\int_{\mathbb{R}} K\left(\frac{x - y}{h_n}\right) f(y) dy \right)^2 \\ &\leq h_n \|f\|_\infty^2 \|K\|_1^2 \to 0 \end{aligned}$$

as $n \to \infty$. □

See Exercise 31 for how to apply this result to obtain confidence intervals for $f(x)$. The theory for confidence 'bands', where coverage is simultaneously in all points x in a given interval $[a, b]$, is more complicated. One can use extreme value theory and some sophisticated Gaussian approximation arguments to obtain

these as well, as was shown in remarkable work by Smirnov (1950) and Bickel and Rosenblatt (1973). See [39] and [42] for recent work on this.

The kernel density estimator as an estimator of F

We have seen now that, at least under certain assumptions, the kernel density estimator $f_n^K(x)$ has a reasonable performance as an estimator of the unknown density $f(x)$. Moreover, it also gives us a natural estimator of the unknown distribution function $F(t) = P(X \leq t)$, namely

$$F_n^K(t,h) = \int_{-\infty}^{t} f_n^K(h,y)dy,$$

and it would be reassuring to see that this estimator is not worse than the very good and simple estimator F_n. Recall the notion of a kernel of order l from Exercise 28.

Theorem 20. *Let $X_1, ..., X_n$ be i.i.d. with bounded density f. Suppose f is m-times continuously differentiable, $m \geq 0$, with m-th derivative bounded by D, assume $h \geq d(\log n/n)$ for some constant d, and that the kernel K is integrable and of order m. If F_n^K is the distribution function of f_n^K and if F_n is the empirical distribution function, then, for every $n \in \mathbb{N}$*

$$E \sup_{t\in\mathbb{R}} |F_n^K(t,h) - F_n(t)| \leq c\sqrt{\frac{h\log(1/h)}{n}} + c'h^{m+1}$$

for some constants c and c' depending only on $K, D, \|f\|_\infty$.

Furthermore, if $h_n \to 0$ as $n \to \infty$ is chosen such that $\sqrt{n}h_n^{m+1} \to 0$, then

$$\sqrt{n}(F_n^K - F) \to^d \mathbb{G}_F \quad \text{in } L^\infty$$

where $\mathbb{G}_F$ is the F-Brownian bridge.

Proof. The proof of the first claim with the supremum over $t \in \mathbb{R}$ uses empirical process methods that we have not developed so far, and follows from Theorem 1 in [40].

To illustrate the main ideas, we prove the first claim for a fixed t (in which case the $\log(1/h)$-term disappears), and for compactly supported K. We decompose

$$F_n^K(t,h) - F_n(t) = (F_n^K(t,h) - F_n(t) - EF_n^K(t,h) + EF_n(t)) + (EF_n^K(t,h) - EF_n(t))$$

and we deal with these two terms separately. The second term equals

$$\begin{aligned} EF_n^K(t,h) - EF_n(t) &= E\frac{1}{nh}\sum_{i=1}^n \int_{-\infty}^t K((x-X_i)/h)dx - F(t) \\ &= \int_{-\infty}^t \int_{\mathbb{R}} h^{-1}K((x-y)/h)f(y)dydx - F(t) \\ &= \int_{-\infty}^t \int_{\mathbb{R}} K(u)f(x-uh)dudx - F(t) \\ &= \int_{\mathbb{R}} K(u)\int_{-\infty}^{t-uh} f(v)dvdu - F(t) \\ &= \int_{\mathbb{R}} K(u)F(t-uh)du - F(t) \\ &= K_h * F(t) - F(t) \end{aligned}$$

where we have used Fubini's theorem and the change of variables $(x-y)/h \mapsto u$ in the dy integral. The absolute value of the last term is bounded by $c'h^{m+1}$ by Proposition 6 (and using Exercise 29 if $m+1 \geq 2$), observing that the fundamental theorem of calculus implies that F is $m+1$ times differentiable when f is m-times continuously differentiable.

To bound the first term, define the functions $g_h(x) = K_h * 1_{(-\infty,t]}(x) - 1_{(-\infty,t]}(x)$ and note that

$$\begin{aligned} F_n^K(t,h) - F_n(t) &= \frac{1}{n}\sum_{i=1}^n \left(\int_{\mathbb{R}} 1_{(-\infty,t]}(x)\frac{1}{h}K\left(\frac{x-X_i}{h}\right)dx - 1_{(-\infty,t]}(X_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^n g_h(X_i) \end{aligned}$$

(using also the symmetry of K). Hence the first term is bounded by

$$\begin{aligned} & E|F_n^K(t,h) - F_n(t) - EF_n^K(t,h) + EF_n(t)| \\ &= E\left|\frac{1}{n}\sum_{i=1}^n (g_h(X_i) - Eg_h(X))\right| \\ &\leq \sqrt{\frac{1}{n}}(Eg_h^2(X))^{1/2} \\ &= \sqrt{\frac{1}{n}}\left[E\left(\int_{\mathbb{R}} (1_{(-\infty,t]}(X+y) - 1_{(-\infty,t]}(X))K_h(y)dy\right)^2\right]^{1/2} \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{\frac{1}{n}} \int_{\mathbb{R}} [E(1_{(-\infty,t]}(X+y) - 1_{(-\infty,t]}(X))^2]^{1/2} |K_h(y)| dy \\
&= \sqrt{\frac{1}{n}} \int_{\mathbb{R}} \left[\int_{t-y}^{t} f(x) dx \right]^{1/2} |K_h(y)| dy \\
&\leq \sqrt{\frac{1}{n}} \|f\|_\infty^{1/2} \int_{\mathbb{R}} |y|^{1/2} |K_h(y)| dy \\
&= \sqrt{\frac{h}{n}} \|f\|_\infty^{1/2} \int |v|^{1/2} |K(v)| dv,
\end{aligned}
$$

where we have used $(E|X|)^2 \leq EX^2$, that K_h integrates to one and Minkowski's inequality for integrals (e.g., p.194 in [35]). This proves the first claim (for fixed $t \in \mathbb{R}$).

The second claim of the theorem is proved as follows: Markov's inequality ($\Pr(|X| > t) \leq E|X|/t$) implies that for this choice of h_n we have $\|\sqrt{n}(F_n^K - F_n)\|_\infty \to 0$ in probability as $n \to \infty$, so that $\sqrt{n}(F_n^K - F)$ has the same limit in law as $\sqrt{n}(F_n - F)$, which is the F-Brownian bridge by Theorem 10. □

This theorem can be used to derive analogues of the results in Section 3.1.3 with F_n replaced by a suitable kernel density estimator, see Section 3.7.1 for more details.

We see that the choice for h that yields a good pointwise density estimator in (90) *does satisfy* the hypotheses of the second claim of the last theorem. So, if m is known, this kernel density estimator simultaneously estimates the underlying distribution F function efficiently, and the density consistently at a point, with a rate of convergence depending on m. This simultaneous property (which is not shared by the empirical distribution function F_n) is often useful, see Proposition 17 for an example, and [4, 40] for further details.

One can also obtain analogues of the Dvoretzky-Kiefer-Wolfowitz Theorem 11 for the distribution function of the kernel density estimator, and other properties, but this requires a more exact probabilistic analysis of F_n^K. See [40] for further details. The paper [13] contains some refinements for the case where X is only required to possess a continuous distribution function, rather than a bounded density, where the kernel estimator can be shown to be still consistent.

3.4.2 Histogram Density Estimators

An intuitive way to estimate a density, which was also the first to be used in practice, is the following: take a partition of $\mathbb{R}$ into intervals $I_k = (a_k, a_{k+1}]$, count the number of observations in each I_k, divide this number by the length of I_k, multiply the indicator function $1_{I_k}(x)$ with the resulting number, and sum over

all k to obtain the estimator. For example, if the partition is dyadic, so that $(a_k)_{k\in\mathbb{Z}} = (2^{-j}k)_{k\in\mathbb{Z}}$, then the estimator has the simple form

$$f_n^H(j,x) = \sum_{k\in\mathbb{Z}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{(k/2^j,(k+1)/2^j]}(X_i) \right) 2^j 1_{(k/2^j,(k+1)/2^j]}(x). \tag{91}$$

Using (82) one sees that the expectation of this estimator $Ef_n^H(j,x) = H_j(f)(x)$ equals the Haar-projection (76) of f. Hence Proposition 7 can be used to control the approximation error when using the histogram density estimator (with 'dyadic bins'), and the role of the 'localization' parameter h for kernel estimators is the 'binsize' 2^{-j}, driven by the parameter j.

One can prove analogues of Theorem 19 and Proposition 11 for the dyadic histogram estimator. To introduce some variation, we now consider the so called *mean integrated squared error* (instead of the pointwise loss) of this estimator, namely

$$E\|f_n^H(j) - f\|_2^2 = E\int_{\mathbb{R}} (f_n^H(j,x) - f(x))^2 dx.$$

Proposition 13. *Let $X_1, ..., X_n$ be i.i.d. random variables with density $f \in L^2$, and let f_n^H be the dyadic histogram estimator. If the derivative Df of f exists and satisfies $\|Df\|_2 < \infty$, then for every $j, n \in \mathbb{N}$ we have*

$$E\|f_n^H(j) - f\|_2^2 \leq \frac{2^{j+1}}{n} + 2^{-2j}\|Df\|_2^2.$$

Proof. We use here the fact that $H_j(f)$ can be viewed as the Haar-wavelet projection and recall the notation from Section 3.3.3. Since $f_n^H(j) - H_j(f) \in V_j$ and $H_j(f) - f \in L^2 \ominus V_j$ we have

$$< f_n^H(j) - H_j(f), H_j(f) - f >= 0 \tag{92}$$

and hence the orthogonal decomposition

$$E\|f_n^H(j) - f\|_2^2 = E\int_{\mathbb{R}} (f_n^H(j,x) - H_j(f)(x))^2 dx + \|H_j(f) - f\|_2^2.$$

The second quantity is bounded using Proposition 7. To bound the first quantity, we write

$$f_n^H(j,x) - H_j(f)(x) = \frac{1}{n}\sum_{i=1}^{n} (Z_i(x) - EZ(x))$$

with

$$Z_i(x) = \sum_k 2^j 1_{(k/2^j,(k+1)/2^j]}(x) 1_{(k/2^j,(k+1)/2^j]}(X_i).$$

Now by Fubini's theorem, recalling the notation $K(x,y) = \sum_{k\in\mathbb{Z}} 1_{(0,1]}(x-k)1_{(0,1]}(y-k)$, using a substitution together with (84) and that f is a density we have

$$\begin{aligned} E\int_{\mathbb{R}}(f_n^H(j,x) - H_j(f)(x))^2dx &= \int_{\mathbb{R}} E(f_n^H(j,x) - H_j(f)(x))^2dx \\ &\leq \frac{1}{n}\int_{\mathbb{R}} EZ_i^2(x)dx \\ &= \frac{2^{2j}}{n}\int_{\mathbb{R}}\int_{\mathbb{R}} K^2(2^jx, 2^jy)f(y)dydx \\ &= \frac{2^j}{n}\int_{\mathbb{R}}\int_{\mathbb{R}} K^2(2^jx, 2^jx-u)f(x-2^{-j}u)dudx \\ &\leq \frac{2^j}{n}\int_{\mathbb{R}} 1_{[-1,1]}(u)\int_{\mathbb{R}} f(x-2^{-j}u)dxdu = \frac{2^{j+1}}{n}, \end{aligned}$$

which completes the proof. □

Balancing the two terms gives the choice

$$2^j \simeq \left(\frac{n\|Df\|_2^2}{2}\right)^{1/3}$$

which parallels the convolution kernel case after Proposition 11 with $m = 1$ (and after the 'conversion' $h \to 2^{-j}$), and in this case we would have the error bound

$$E\|f_n^H(j) - f\|_2 \leq \left(\frac{2}{n}\right)^{1/3}\|Df\|_2^{1/3}. \tag{93}$$

Although this is a very neat risk bound with explicit constants, we cannot improve this rate further if f is more than once differentiable, because of the limitations of Proposition 7. The histogram estimator is simple and useful in practice, even if it has theoretical limitations. These limitations can be overcome by replacing the Haar basis by more general wavelet bases.

3.4.3 Wavelet Density Estimators

Next to approximation by convolution with 'approximate identities' K_h, it was shown in the last section that one can approximate functions f by their wavelet series (81). Suppose again $X_1, ..., X_n$ are i.i.d. random variables from the density $f : \mathbb{R} \to \mathbb{R}$. If ϕ and ψ are the generating functions of a wavelet basis satisfying Condition 1, then recall from (79) that the projection of an arbitrary density f onto the space V_j spanned by this wavelet basis is given by

$$K_j(f)(x) = \sum_{k\in\mathbb{Z}} < \phi_k, f > \phi(x-k) + \sum_{l=0}^{j-1}\sum_{k\in\mathbb{Z}} < \psi_{lk}, f > \psi_{lk}(x),$$

where

$$< \phi_k, f >= \int_{\mathbb{R}} \phi(y-k)f(y)dy, \quad < \psi_{lk}, f >= \int_{\mathbb{R}} 2^{l/2}\psi(2^l y - k)f(y)dy$$

are the corresponding wavelet coefficients, which can be naturally estimated by the sample means

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i - k), \quad \hat{\beta}_{lk} = \frac{2^{l/2}}{n}\sum_{i=1}^{n} \psi(2^l X_i - k). \tag{94}$$

Hence the linear *wavelet density estimator* $f_n^W(j,x)$ *at resolution level* j is

$$\begin{aligned} f_n^W(j,x) &= \sum_{k\in\mathbb{Z}} \hat{\alpha}_k \phi(x-k) + \sum_{l=0}^{j-1}\sum_{k\in\mathbb{Z}} \hat{\beta}_{lk}\psi_{lk}(x) \\ &= \sum_{k\in\mathbb{Z}} \hat{\alpha}_{jk} 2^{j/2}\phi(2^j x - k) \end{aligned} \tag{95}$$

where the second identity follows from arguments similar to those leading to (82), and where

$$\hat{\alpha}_{jk} = \frac{2^{j/2}}{n}\sum_{i=1}^{n} \phi(2^j X_i - k).$$

One can prove results similar to Theorem 19 and Propositions 11 and 13 for the wavelet density estimator, as the next proposition shows.

Proposition 14. *Let $X_1, ..., X_n$ be i.i.d. random variables with density f, and let f_n^W be the wavelet density estimator from (95). Suppose the wavelet basis satisfies Condition 1 for some $S \geq 0$, and let $m \leq S+1$.*
i.) If f is bounded and if the m-th derivative of f is bounded on $\mathbb{R}$, by D say, then for every $x \in \mathbb{R}$, every $n \in \mathbb{N}$ and every $j > 0$ we have

$$E|f_n^W(j,x) - f(x)| \leq C\left(\sqrt{\frac{2^j}{n}} + 2^{-jm}\right)$$

where the constant C depends only on $\|f\|_\infty, \phi$ and on D.
ii.) If $f \in L^2$ and if $D' := \|D^m f\|_2 < \infty$ then

$$E\int_{\mathbb{R}} (f_n^W(j,x) - f(x))^2 dx \leq C'\left(\frac{2^j}{n} + 2^{-2jm}\right)$$

where the constant C' depends only on ϕ and on D'.
iii.) If $f \in L^1$, $D'' := \|D^m f\|_1 < \infty$ and if $L(\kappa) := \int_{\mathbb{R}}(1+|x|)^\kappa f(x)dx < \infty$ for

some $\kappa > 1$, then

$$E\int_{\mathbb{R}} |f_n^W(j,x) - f(x)|dx \leq C''\left(\sqrt{\frac{2^j}{n}} + 2^{-jm}\right)$$

where the constant C'' depends only on $\phi, D'', L(\kappa)$.

Proof. The variance-bias decomposition is ($p \in \{1,2\}$)

$$\int_{\mathbb{R}} |f_n^W(j,x) - f(x)|^p dx \leq \int_{\mathbb{R}} |f_n^W(j,x) - K_j(f)(x)|^p dx + \int_{\mathbb{R}} |K_j(f)(x) - f(x)|^p dx, \tag{96}$$

(in case $p = 2$ because of orthogonality as in (92)), and for Part i.) the same holds without integrals. To deal with the integrand of the first term, define the random variables $Z_i(x) = 2^j K(2^j x, 2^j X_i)$ where $K(x,y) = \sum_{k\in\mathbb{Z}} \phi(x-k)\phi(y-k)$, then, using the second characterisation of f_n^W in (95) we see that

$$\begin{aligned}
E(f_n^W(j,x) - K_j(f)(x))^2 &= E\left(\frac{1}{n}\sum_{i=1}^n (Z_i(x) - EZ_i(x))\right)^2 \\
&\leq \frac{1}{n} EZ_i^2(x) \\
&= \frac{2^{2j}}{n}\int_{\mathbb{R}} K^2(2^j x, 2^j y) f(y) dy \\
&= \frac{2^j}{n}\int_{\mathbb{R}} K^2(2^j x, 2^j x - u) f(x - 2^{-j}u) du \\
&\leq \frac{2^j}{n}\int_{\mathbb{R}} c^2\|\phi\|_\infty^2 1_{[-a,a]}(u) f(x - 2^{-j}u) du. \qquad (97)
\end{aligned}$$

where we have used (86).

To prove Part i.) we can use Proposition 9 to bound the second term from (96). For the first term $E|X| \leq \sqrt{EX^2}$ and (97) give the desired bound since

$$\int_{\mathbb{R}} c\|\phi\|_\infty^2 1_{[-a,a]}(u) f(x - 2^{-j}u) du \leq 2ac\|\phi\|_\infty^2 \|f\|_\infty.$$

Part ii.) follows from (96), Proposition 9 and from

$$\begin{aligned}
E\int_{\mathbb{R}} (f_n^W(j,x) - K_j(f)(x))^2 dx &= \int_{\mathbb{R}} E(f_n^W(j,x) - K_j(f)(x))^2 dx \\
&\leq \frac{2^j}{n}\int_{\mathbb{R}}\int_{\mathbb{R}} c\|\phi\|_\infty^2 1_{[-a,a]}(u) f(x - 2^{-j}u) du dx \\
&= \frac{2^{j+1}}{n} ac\|\phi\|_\infty^2
\end{aligned}$$

where we have used (97), Fubini's theorem and that f is a density.

To prove Part iii), (99) and Proposition 9 leaves us with the 'variance term': Setting $a_j(x) = 2^j 1_{[-a,a]}(2^j x)$ we obtain

$$\begin{aligned} E\int_{\mathbb{R}} |f_n^W(j,x) - K_j(f)(x)|dx &= \int_{\mathbb{R}} E|f_n^W(j,x) - K_j(f)(x)|dx \\ &\leq \sqrt{\frac{2^j}{n}} c\|\phi\|_\infty \int_{\mathbb{R}} \sqrt{\int_{\mathbb{R}} 1_{[-a,a]}(u) f(x - 2^{-j}u)du}dx \\ &= \sqrt{\frac{2^j}{n}} c\|\phi\|_\infty \int_{\mathbb{R}} \sqrt{a_j * f(x)}dx \\ &\leq \sqrt{\frac{2^j}{n}} cd\|\phi\|_\infty \int_{\mathbb{R}} (1+|x|)^\kappa f(x)dx \end{aligned}$$

using Exercise 36. □

Similar to the discussion after Proposition 11, one can obtain a 'local' version of Part i) of the above result. Moreover one can ask whether Part i.) of Proposition 14 can be made uniform in $x \in \mathbb{R}$ or whether the integrated wavelet density estimator has properties similar to Theorem 20. See [41] for such results.

The last proposition implies that we can estimate a density of arbitrary smoothness m at the minimax rate of convergence $n^{-m/(2m+1)}$ (cf. Theorem 16), at least *if we would know* m, and if we use wavelets of regularity $S \geq m-1$. As $m \to \infty$ this recovers the 'parametric' rate $1/\sqrt{n}$ from finite-dimensional models.

3.4.4 Application to Inverse Problems

An enormous field within applied mathematics is concerned with so-called 'inverse problems', which are strongly motivated by applications in industrial and medical engineering as well as in image and signal processing. The main problem can be described as follows: one would like to observe a signal or function f which, however, has been corrupted for some external reason (systematic measurement error, experimental setup of the measurement device, too high cost of full measurement of f etc.). Mathematically this is often modeled as the observed function being $K(f)$ instead of f, where K is some operator on some function space. If the operator K is invertible this poses no serious problems, but the interesting case is usually where K is not invertible, for instance if K is a compact operator on L^2. If K is not invertible these problems are called 'ill-posed' inverse problems, and the 'ill-posedness' is usually measured in terms of the spectral properties of the operator K (usually the rate of decay of its eigenvalues at infinity).

A special case of an ill-posed inverse problem is the *statistical deconvolution problem:* Suppose we would like to estimate a probability density $f : \mathbb{R} \to \mathbb{R}$ from a sample $X_1, ..., X_n$ of f, but we do not observe the X_i's, but rather $Y_i = X_i + \varepsilon_i$, where ε_i is a random error term with known probability distribution φ, independent of X_i. In the language of inverse problems this means that K is an integral operator that arises from convolving f with φ, and that we do not observe a sample from f, but from $K(f)$ which has density $g := f * \varphi = \int f(\cdot - y)d\varphi(y)$, and would like to estimate f. If φ is pointmass at some point x then K is invertible, but as soon as φ has a density the convolution operator is compact and thus K is not invertible, so that this is an ill-posed inverse problem.

Minimax Results for Deconvolution Density Estimation.

To understand how nonparametric density estimation changes in the deconvolution problem, let us first consider the minimax risk as in Subsection 3.2. Define, for $m \in \mathbb{N}, 0 < B < \infty$, the class of densities

$$\mathcal{W}(m, B) = \left\{ f : f \geq 0, \int_{\mathbb{R}} f(x)dx = 1, \|f\|_2 + \|D^m f\|_2 \leq B \right\}$$

and define the minimax L^2-risk

$$R_n(m, B) = \inf_{\tilde{f}_n} \sup_{f \in W(m,B)} E\|\tilde{f}_n - f\|_2, \tag{98}$$

where the infimum is taken over all possible estimators $\tilde{f}_n$. Note that an estimator in the deconvolution problem means any measurable function of a sample $Y_1, \ldots, Y_n$ from density $f * \varphi$.

The spectrum of the convolution operator K, in particular the decay of its eigenvalues at infinity, is linked to the decay at infinity of the Fourier transform

$$F[\varphi](u) = \int_{\mathbb{R}} e^{-ixu} d\varphi(x)$$

of φ. This decay measures, in a way, the regularity properties of the measure φ – for instance if φ has an infinitely differentiable density, then $F[\varphi]$ decays faster than any polynomial.

The following theorem distinguishes the so-called 'moderately ill-posed' case, where $F[\varphi]$ decays polynomially at infinity (as is the case, for instance, with Laplace errors), and the 'severely ill-posed' case where $F[\varphi]$ may decay exponentially fast (including, for instance, the Gaussian or Cauchy densities). Note also that if there is no measurement error present so that φ equals pointmass δ_0 at 0, then $c_0 = w = 0$ in the following theorem which retrieves the minimax rates from the usual density estimation problem.

Theorem 21. *Suppose φ has Fourier transform $F[\varphi](u) = C(1+u^2)^{-\frac{w}{2}} e^{-c_0|u|^\alpha}$, $u \in \mathbb{R}$, for some constants $C, \alpha > 0$ and $w, c_0 \geq 0$. Then for any $m, B > 0$ there exists a constant $c := c(m, B, C, \alpha, w, c_0) > 0$ such that for every n we have*

$$R_n(m, B)) \geq c \begin{cases} \left(\frac{1}{\log n}\right)^{\frac{m}{\alpha}} & \text{if } c_0 > 0 \\ \left(\frac{1}{n}\right)^{\frac{m}{2m+2w+1}} & \text{if } c_0 = 0. \end{cases}$$

For a proof see [32] and also [64]. This shows that for smoother error densities the problem of estimation of f becomes more difficult. In particular, if ε is standard normal then the best rate of convergence for estimating f is only of logarithmic order in n, instead of polynomial rates in the 'direct' density estimation problem. If the error density is the Laplace density $e^{-|u|}$, then the Fourier transform decays like $(1 + x^2)^{-1}$, so that the best possible rate is polynomial, but deteriorates by an exponent of $w = 2$. This is the price to be paid for the non-invertibility of the convolution operator $f \mapsto f * \varphi$.

Wavelet Deconvolution.

The above results show that deconvolution is a 'harder' problem than regular density estimation, and that there are some natural restrictions to the performance of any estimator. But can we find *one* estimator that attains the minimax risk from Theorem 21?

One idea is to 'deconvolve' $g = f * \varphi$ by 'Fourier inversion', as follows. We recall some properties of the Fourier transform F, which for $f \in L^1$ is defined as $F[f](u) = \int_{\mathbb{R}} f(x)e^{-ixu}dx$ and can be naturally extended to L^2. First

$$F[f * \mu] = F[f] \cdot F[\mu], \quad F[f * g] = F[f] \cdot F[g]$$

for any probability measure μ and $f, g \in L^1$, second the Plancherel identity

$$\langle g, h\rangle = \frac{1}{2\pi}\langle \overline{F[g]}, F[h]\rangle,$$

for $g, h \in L^2$, where $\langle \cdot, \cdot \rangle$ is the L^2-inner product; and third, for $h \in L^1$ and $\alpha \in \mathbb{R} \setminus \{0\}$ the function $h_\alpha(x) := h(\alpha x)$ has Fourier transform

$$F[h_\alpha](u) = \alpha^{-1} F[h](\alpha^{-1}u).$$

Denote finally the inverse Fourier(-Plancherel) transform by F^{-1}, $F^{-1}h(x) = (1/2\pi)\int_{\mathbb{R}} h(u)e^{ixu}du$, so that

$$F^{-1}Fh = h \quad \textit{for} \quad h \in L^2.$$

When estimating f we can, following the ideas from the previous sections, estimate its wavelet projection $K_j(f) = \sum_k \langle \phi_{jk}, f \rangle \phi_{jk}$ first and then balance estimation with approximation error. To estimate $K_j(f)$, recall (82), write $\phi_{0k} = \phi(\cdot - k)$, assume $|F[\varphi]| > 0$ on $\mathbb{R}$ and observe

$$\begin{aligned} K_j(f)(x) &= 2^j \sum_k \phi(2^j x - k) \int_{\mathbb{R}} \phi(2^j y - k) f(y) dy \\ &= \sum_k \phi(2^j x - k) \frac{1}{2\pi} \int_{\mathbb{R}} \overline{F[\phi_{0k}](2^{-j}u)} F[f](u) du \\ &= \sum_k \phi(2^j x - k) \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\overline{F[\phi_{0k}](2^{-j}u)}}{F[\varphi](u)} F[g](u) du \\ &= 2^j \sum_k \phi(2^j x - k) \int_{\mathbb{R}} \tilde{\phi}_{jk}(y) g(y) dy \\ &= \int_{\mathbb{R}} K_j^*(x, y) g(y) dy, \end{aligned} \tag{99}$$

where the (nonsymmetric) kernel K_j^* is given by

$$K_j^*(x, y) = 2^j \sum_{k \in \mathbb{Z}} \phi(2^j x - k) \tilde{\phi}_{jk}(y)$$

with

$$\tilde{\phi}_{jk}(x) = F^{-1}\left[2^{-j} \frac{F[\phi_{0k}](2^{-j}\cdot)}{\overline{F[\varphi]}} \right](x) = \phi_{0k}(2^j \cdot) * F^{-1}\left[1_{[-2^j a, 2^j a]} \frac{1}{\overline{F[\varphi]}} \right](x).$$

The interchange of summation and integration in (99) can be justified by using the dominated convergence theorem.

Since we have a sample $Y_1, ..., Y_n$ from the density g the identity (99) suggests a natural estimator of f, namely the *wavelet deconvolution density estimator*

$$f_n(x, j) = \frac{1}{n} \sum_{m=1}^{n} K_j^*(x, Y_m), \tag{100}$$

which is an unbiased estimate of $K_j(f)(x)$. This estimator was studied, for instance, in [64] and [48]. It turns out that both in practice and for the theoretical development it is advisable to choose wavelets that satisfy Condition 1 with one crucial difference: the requirement that ϕ has compact support should be replaced by the assumption that $F[\varphi]$ has compact support (to have both is not possible by Heisenberg's 'uncertainty principle'). Such wavelets exist, for instance 'Meyer wavelets'. One can then show that the estimator (100) attains the minimax rates of convergence in Theorem 21. See the exercises for a consistency result.

3.4.5 Exercises

Exercise 31. Suppose you are given an i.i.d. sample from a bounded density f with bounded derivative. Suppose $c(\alpha, x)$ is such that $\Pr(-c(\alpha, x) \leq Z \leq c(\alpha, x)) = 1 - \alpha$ where $Z \sim N(0, f(x))$. Use a kernel density estimator (with a suitable kernel) to obtain a 95 percent confidence interval for $f(x)$ in such a way that the size of the interval shrinks at rate $1/\sqrt{nh_n}$ as $n \to \infty$, and that h can be chosen so that this rate is 'almost' (say, up to a $\log n$ term) of order $n^{-1/3}$. Use an auxiliary estimator of $f(x)$ to construct a similar confidence interval if you are given only the quantiles of the standard normal distribution, thereby circumventing that the variance of the $N(0, f(x))$ distribution is unknown.

Exercise 32. The following result is known as Bernstein's inequality: If $X_1, ..., X_n$ are mean zero independent random variables taking values in $[-c, c]$ for some constant $0 < c < \infty$, then

$$\Pr\left\{\left|\sum_{i=1}^{n} X_i\right| > u\right\} \leq 2\exp\left(-\frac{u^2}{2nEX_i^2 + (2/3)cu}\right). \tag{101}$$

i) If $K = 1_{[-1/2,1/2]}$, use this inequality to prove for the kernel density estimator that

$$\Pr\left\{\left|f_n^K(x,h) - K_h * f(x)\right| > t\right\} \leq 1\exp\left\{-\frac{nht^2}{2\|f\|_\infty + (2/3)t}\right\}.$$

Now choose $t = x\sqrt{1/nh}$ and describe the tail of this inequality as x varies over $(0, \infty)$. In which range of ts can one deduce a similar inequality for $\{|f_n^K(x,h) - f(x)| \geq t\}$? ii)$^+$ Prove Bernstein's inequality.

Exercise 33. Let $f_n^K(h)$ be a kernel density estimator with compactly supported symmetric kernel function $K : \mathbb{R} \to \mathbb{R}$. Suppose $f \in L^2$ is twice differentiable with $\int_{\mathbb{R}}(D^2 f(x))^2 dx < \infty$. Bound the mean-integrated squared error $E\int_{\mathbb{R}}(f_n^K(h,x) - f(x))^2 dx$ by

$$\frac{1}{nh}\|K\|_2^2 + (1/3)h^4\|D^2 f\|_2^2\left(\int_{\mathbb{R}} u^2 K(u)du\right)^2.$$

Find the choice h that balances these antagonistic terms.

Exercise 34. Let $f_n^H(j)$ be the Haar wavelet density estimator. Assume that f is once differentiable with bounded derivative. Show that $E|f_n^H(j,x) - f(x)| \leq Cn^{-1/3}$ for every $x \in \mathbb{R}$, some constant C independent on n, and if one chooses $2^{j_n} \simeq n^{1/3}$.

Exercise 35. $^+$ Consider the statistical deconvolution problem $Y = X + \epsilon$ where ϵ is distributed as a standard Cauchy random variable, and where X is independent

of ϵ and has unknown density f. Let $\hat{f}_n$ be the Meyer-wavelet deconvolution density estimator at resolution level j_n based on Meyer wavelets ϕ whose Fourier transform $F[\phi]$ is supported in the compact interval $[-a, a]$. Assuming that f is bounded and continuous on $\mathbb{R}$ and that $j := j_n \to \infty$ is chosen in such a way that $e^{2^{j+1}a}/n \to 0$, show that $\hat{f}_n$ is pointwise consistent for f, i.e., show that $E|\hat{f}_n(x) - f(x)| \to 0$ as $n \to \infty$. [You may use freely facts from Fourier analysis, that the Meyer wavelets form an orthonormal basis of L^2, that $|K_j(f) - f|(x) \to 0$ as $j \to \infty$ if $K_j(f)$ is the wavelet projection of f, as well as the fact that $\sup_{x\in\mathbb{R}} \sum_k |\phi(x-k)| < \infty$.]

Exercise 36. $^+$ If f and g are two nonnegative functions on $\mathbb{R}$, such that $\int_{\mathbb{R}} f(x)(1+|x|)^s dx = c(f)$ is finite for some $s > 1$ and if g has compact support, in $[-a, a]$ say, prove that $\int_{\mathbb{R}} \sqrt{g_h * f}(x)dx \leq d\sqrt{c(g)c(f)}$ where the constant d depends only on s and a. (Hint: use Jensen's inequality.)

3.5 Nonparametric Regression

The typical regression problem is the following: Suppose we are given n pairs of observations $(X_1, Y_1), ..., (X_n, Y_n)$, and suppose the *response variable* Y is related to the *covariate* (or 'feature') X by a functional relationship of the form

$$Y_i = m(X_i) + \epsilon_i, \qquad E(\epsilon_i) = 0, \quad i = 1, ...n \tag{102}$$

where $m : \mathbb{R} \to \mathbb{R}$ is some unknown function. In simple linear regression models we assume $m(X_i) = a + bX_i$ for some unknown a, b, or $m(X_i) = m(X_i, \theta), \theta \in \Theta \subset \mathbb{R}^p$ in more general parametric models, but nonparametric regression models try to make as few assumptions on m as possible. If we view the X_i's as fixed numerical values, we speak of a 'fixed design' regression model (and one usually writes x_i for X_i then). Often it is reasonable to treat the covariates also as random, in which case we speak of a 'random design' regression model.

Given the random design situation, we usually assume that the (X_i, Y_i) are jointly i.i.d., and there are at least two ways to view the regression problem. For the first, suppose we have 'unobserved' errors ϵ_i that are i.i.d. mean zero with variance σ^2, independent of the X_i's . In this case we neccesarily have $m(x) = E(Y|X = x)$ in (102) since

$$E(Y|X = x) = E(m(X)|X = x) + E(\epsilon|X = x) = m(x).$$

The second approach to this problem avoids an explicit additive structure of the errors ϵ_i and views this problem as a mere prediction problem: Given the random variable X, we would like to predict the value of Y, and the function $m(x) = E(Y|X = x)$ is always the best predictor (in a mean-square sense), see Exercise 17. So, viewed as a prediction problem, it is of independent interest to estimate this conditional expectation nonparametrically. All results below apply to both these 'philosophies' behind the regression problem.

3.5.1 Nonparametric regression based on kernel methods

We will show that the methods from kernel density estimation from the last section can be adapted to estimation of a regression function, both in the random design case (prediction problem) as well as in the fixed design case.

The 'Nadaraya-Watson' estimator

We first consider the random design model: Let (X_i, Y_i) be an i.i.d. sample from (X, Y) with joint density $f(x, y)$ and denote the marginal density of X by $f^X(x)$. Recall that by definition

$$E(Y|X = x) = \int_{\mathbb{R}} y \frac{f(x, y)}{f^X(x)} dy.$$

If K is a nonnegative kernel function with $\int_{\mathbb{R}} K(u)du = 1$, then the proposed estimator for $m(x)$ is

$$\hat{m}_n(h, x) = \frac{\sum_{i=1}^n Y_i K((x - X_i)/h)}{\sum_{i=1}^n K((x - X_i)/h)} \tag{103}$$

if the denominator is nonzero, and zero otherwise. This estimator was first studied by Nadaraya (1964) and Watson (1964). We will now prove the following theorem for the case where m is twice and f^X once differentiable. Different sets of assumptions are possible, but at the expense of a somewhat unreasonable notational complexity.

Theorem 22. *Suppose $m(x) = E(Y|X = x)$ is bounded and twice continuously differentiable at $x \in \mathbb{R}$, that the conditional variance function $V(x) = Var(Y|X = x)$ is bounded on $\mathbb{R}$ and continuous at x, and that f^X is bounded on $\mathbb{R}$, continuously differentiable at x, and satisfies $f^X(x) > 0$. Suppose further that K is positive, symmetric, bounded, compactly supported and integrates to one. If $h \to 0$ as $n \to \infty$ satisfies $nh/\log n \to \infty$, then*

$$E|\hat{m}_n(h, x) - m(x)| \leq \frac{L}{\sqrt{nh}} + L'h^2 + Z_n \tag{104}$$

where $Z_n = o\left(h^2 + (nh)^{-1/2}\right)$,

$$L^2 = \frac{V(x)\|K\|_2^2}{f^X(x)}$$

and

$$L' = \kappa(2)\left(\frac{Dm(x)Df^X(x)}{f^X(x)} + \frac{D^2m(x)}{2}\right)$$

and where $\kappa(2)$ was defined in Proposition 6.

In particular, if $h_n \simeq n^{-1/5}$, then, as $n \to \infty$ we have

$$E|\hat{m}_n(h,x) - m(x)| = O(n^{-2/5}). \tag{105}$$

Proof. We take w.l.o.g. $K = 1_{[-1/2,1/2]}$ and note in advance that the denominator of $\hat{m}_n(h,x)$ is (up to the $1/nh$-term) just the kernel density estimator

$$\hat{f}^X(x) := f_n^K(h,x) = \frac{1}{nh}\sum_{i=1}^n K((x-X_i)/h)$$

based on the covariates. We will also use repeatedly that $E(Z) = \int E(Z|X = u)f^X(u)du$ for any absolutely continuous pair (Z,X) of random vectors with marginal density f^X.

Step 1. This step is important, but a little technical and can be skipped at first reading. We wish to first bound $E(\hat{m}_n(h,x) - m(x))^2$ from above by a constant. We have

$$E(\hat{m}_n(h,x))^2 = E\left(\frac{\sum_{i=1}^n Y_i K((x-X_i)/h)}{\sum_{i=1}^n K((x-X_i)/h)}\right)^2$$
$$= \int_{\mathbb{R}^n\setminus U} E\left[\left(\frac{\sum_{i=1}^n Y_i K((x-u_i)/h)}{\sum_{i=1}^n K((x-u_i)/h)}\right)^2 |X_1 = u_1, ..., X_n = u_n\right] \Pi_{i=1}^n f^X(u_i)du_i$$

where $U = \{(u_i)_{i=1}^n : \sum_i K((x-u_i)/h) = 0\}$, on which the random variable $\hat{m}_n(x)|\{X_i = u_i \; for \; all \; i\}$ equals zero by definition of $\hat{m}_n$. Now the integrand equals, using vector notation,

$$E\left[\left(\frac{\sum_{i=1}^n (Y_i - E[Y_i|\mathbf{X}=\mathbf{u}])K((x-u_i)/h)}{\sum_{i=1}^n K((x-u_i)/h)}\right)^2 |\mathbf{X}=\mathbf{u}\right]$$
$$+\left(\frac{\sum_{i=1}^n E[Y_i|\mathbf{X}=\mathbf{u}]K((x-u_i)/h)}{\sum_{i=1}^n K((x-u_i)/h)}\right)^2,$$

where the second summand is bounded by $\|m\|_\infty^2$ since $E[Y_i|\mathbf{X}=\mathbf{u}] = E[Y|X = u_i] = m(u_i)$ and since K is positive. Similarly, the first summand equals, by (conditional) independence,

$$\frac{\sum_{i=1}^n Var(Y|X=u_i])K^2((x-u_i)/h)}{\left(\sum_{i=1}^n K((x-u_i)/h)\right)^2} \leq \|V\|_\infty$$

since, setting $K_i = K((x-u_i)/h)$, we have

$$\frac{\sum K_i^2}{\left(\sum K_i\right)^2} \leq 1 \iff 2\sum_{i\neq j} K_i K_j \geq 0$$

which is satisfied by positivity of the kernel. Since $\Pi_{i=1}^n f^X$ is integrable on $\mathbb{R}^n$ we conclude

$$E(\hat{m}_n(h,x) - m(x))^2 \leq 4\|m\|_\infty^2 + 2\|V\|_\infty = d^2(\|m\|_\infty, \|V\|_\infty) := d^2. \tag{106}$$

Furthermore, since $f^X(x) > 0$ and since f^X is continuous at x, there exists a neighborhood of x where f^X is greater than or equal to 2δ for some $\delta > 0$. Then

$$K_h * f^X(x) = h^{-1} \int_{\mathbb{R}} K\left(\frac{x-y}{h}\right) f^X(y)dy = \int_{-1/2}^{1/2} f^X(x - hu)du \geq 2\delta \tag{107}$$

for h small enough and since K is positive and integrates to one. Now using the Cauchy-Schwarz and Bernstein's inequality (Exercise 32), (106), (107) and $nh/\log n \to \infty$, we have for some divergent sequence a_n and some fixed constant $c := c(\delta, \|f^X\|_\infty)$ that

$$\begin{aligned}
E|\hat{m}_n(h,x) - m(x)|1\{\hat{f}^X(x) \leq \delta\} &\leq \left(E(\hat{m}_n(h,x) - m(x))^2\right)^{1/2} \sqrt{\Pr\{\hat{f}^X(x) \leq \delta\}} \\
&\leq d\sqrt{\Pr\{|\hat{f}^X(x) - K_h * f^X(x)| \geq \delta\}} \\
&\leq \sqrt{2}d \exp\left\{-\frac{nh\delta^2}{4\|f^X\|_\infty + (2/6)\delta}\right\} \\
&\leq \sqrt{2}d \exp\left\{-\frac{nh\delta^2}{4\|f^X\|_\infty + (2/6)\delta}\right\}. \\
&\leq \sqrt{2}d \exp\{-a_n c \log n\} = O(n^{-a_n c})
\end{aligned}$$

which is $o\left(h^2 + (nh)^{-1/2}\right)$ in view of the maintained assumptions on h.

Step 2. We now proceed with the main proof, and only have to consider

$$E|\hat{m}_n(h,x) - m(x)|1\{\hat{f}^X(x) > \delta\},$$

so can work on the event $\{\hat{f}^X(x) > \delta\}$ for some $\delta > 0$. On this event, setting shorthand

$$\hat{g}(x) = \frac{1}{nh}\sum_{i=1}^n Y_i K((x - X_i)/h),$$

and $g(x) = m(x)f^X(x)$ we can write

$$\begin{aligned}
\hat{m}_n(h,x) - m(x) &= \frac{\hat{g}(x)}{\hat{f}^X(x)} - \frac{g(x)}{f^X(x)} \\
&= \frac{\hat{g}(x)f^X(x) - g(x)\hat{f}^X(x)}{\hat{f}^X(x)f^X(x)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\hat{g}(x)f^X(x) - g(x)\hat{f}^X(x)}{(f^X(x))^2} \\
&\quad + \frac{\hat{g}(x)f^X(x) - g(x)\hat{f}^X(x)}{(f^X(x))^2}\left(\frac{f^X(x)}{\hat{f}^X(x)} - 1\right) \\
&:= M_n + M_n\left(\frac{f^X(x)}{\hat{f}^X(x)} - 1\right), \qquad (108)
\end{aligned}
$$

and we treat the expectation of M_n first. Define the random variables

$$
W_i = Y_i K\left(\frac{x - X_i}{h}\right) f^X(x) - K\left(\frac{x - X_i}{h}\right) g(x)
$$

so that

$$
\begin{aligned}
(E|M_n|)^2 &\leq EM_n^2 \\
&= (f^X(x))^{-4} E\left(\hat{g}(x)f^X(x) - g(x)\hat{f}^X(x)\right)^2 \\
&= (f^X(x))^{-4} n^{-2}h^{-2} E\left(\sum_{i=1}^n W_i\right)^2 \\
&= (f^X(x))^{-4} n^{-2}h^{-2}\left(Var\left(\sum_{i=1}^n W_i\right) + \left(\sum_{i=1}^n EW_i\right)^2\right). \qquad (109)
\end{aligned}
$$

We bound the variances

$$
\begin{aligned}
Var(W_i) &\leq E(W_i^2) \\
&= E\left(YK\left(\frac{x - X}{h}\right) f^X(x) - K\left(\frac{x - X}{h}\right) m(x)f^X(x)\right)^2 \\
&= (f^X(x))^2 E\left((Y - m(x))K\left(\frac{x - X}{h}\right)\right)^2 \\
&= (f^X(x))^2 \int_{\mathbb{R}} E((Y - m(x))^2 | X = u)K^2\left(\frac{x - u}{h}\right) f^X(u)du \\
&= (f^X(x))^2 h \int_{\mathbb{R}} E((Y - m(x))^2 | X = x - vh)K^2(v)f^X(x - vh)dv \\
&= (f^X(x))^3 h Var(Y|X = x)\|K\|_2^2 + o(h)
\end{aligned}
$$

using that $m(x)$ is the conditional expectation $E(Y|X = x)$, and continuity of the functions $Var(Y|X = (\cdot))$, m and f^X at x, inserting and subtracting

$$
E((Y - m(x - vh))^2 | X = x - vh) = V(x - vh)
$$

in the last integrand. Since the (Y_i, X_i) are jointly i.i.d. we have $Var(\sum W_i) = nVar(W_i)$, and plugging this bound into the first part of (109) yields the first term in (104).

We next bound the means EW_i. By applying Taylor expansions similar as in the proof of Proposition 6, using that m is twice and f^X once continuously differentiable at x, we have

$$\begin{aligned}
EW_i &= E\left(YK\left(\frac{x-X}{h}\right)f^X(x) - K\left(\frac{x-X}{h}\right)m(x)f^X(x)\right) \\
&= f^X(x)\int_{\mathbb{R}}(E(Y|X=u)-m(x))K\left(\frac{x-u}{h}\right)f^X(u)du \\
&= f^X(x)h\int_{\mathbb{R}}(m(x-vh)-m(x))K(v)f^X(x-vh)dv \\
&= (f^X(x))^2 h\int_{\mathbb{R}}(m(x-vh)-m(x))K(v)dv \\
&\quad + f^X(x)h\int_{\mathbb{R}}(m(x-vh)-m(x))(f^X(x-vh)-f^X(x))K(v)dv \\
&\leq h^3\kappa(2)\left((f^X(x))^2 2^{-1}D^2m(x) + f^X(x)Dm(x)Df^X(x)\right) + o(h^3).
\end{aligned}$$

Feeding this bound into the second part of (109) yields the second term in (104). It remains to prove that the expectation of the second term in (108) is of smaller order than the first. Clearly

$$E\left|M_n\left(\frac{f^X}{\hat{f}^X(x)}-1\right)\right| \leq (EM_n^2)^{1/2}\left(E\left(\frac{f^X(x)}{\hat{f}^X(x)}-1\right)^2\right)^{1/2}$$

by the Cauchy-Schwarz inequality which completes the proof of the theorem by what was established about EM_n^2 above and since, on the event $\{\hat{f}_n^X(x) > \delta\}$,

$$E\left(\left(\frac{f^X(x)}{\hat{f}^X(x)}-1\right)\right)^2 \leq \delta^{-2}E(f^X(x)-\hat{f}^X(x))^2 \to 0$$

as $n \to \infty$ by the same arguments as in the proof of Proposition 11. □

This result shows that the very simple estimator (103) has a reasonable performance for estimating an arbitrary twice differentiable regression function. We should also note that the hypotheses on the error term ϵ are rather weak here, and essentially implicit in the assumptions on m and V.

Similar to Section 3.2, one can show that the rate of convergence in (105) is best possible over the class of twice differentiable regression functions, so that this theorem is optimal in this respect.

We have also seen that roughly the same techniques needed to derive properties of kernel density estimators are needed in regression. One can then proceed and prove results similar to those in Section 3.4 for regression estimators, e.g., a pointwise limit theorem analogous to Proposition 12.

3.5.2 Local polynomial estimators.

A more general class of regression estimators is given by *local polynomial estimators*: Define the $(\ell+1)\times 1$ vectors

$$U(t) = (1, t, t^2/2!, ..., t^\ell/\ell!)^T, \;\; M(x) = \left(m(x), Dm(x)h, D^2m(x)h^2, ..., D^\ell m(x)h^\ell\right)^T.$$

If $K \in L^1$ is a positive kernel that integrates to one and $\ell \geq 0$ an integer, define an estimator of $M(x)$ given by

$$\hat{M}_n(h,x) = \arg\min_{M\in\mathbb{R}^{\ell+1}} \sum_{i=1}^n \left[Y_i - M^T U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right). \tag{110}$$

The statistic $\hat{m}_n^\ell(h,x) = U(0)^T \hat{M}_n(h,x)$, which picks the first component out of the vector $\hat{M}_n(h,x)$, is called the *local polynomial estimator of order* ℓ of $m(x)$. If $\ell \geq 1$ the successive components of the vector $\hat{M}_n(h,x)$ furnish us with estimates of the derivatives of m as well.

For x fixed $\hat{m}_n^\ell(h,x)$ is a weighted least squares estimator, and this fact is investigated further in Exercise 38. In particular, the case $\ell = 0$ will be seen to correspond to the Nadaraya-Watson estimator (103), as one can show that

$$\hat{m}_n^\ell(h,x) = \sum_{i=1}^n Y_i W_{ni}(x) \tag{111}$$

where (see Exercise 38)

$$W_{ni}(x) = \frac{1}{nh} U^T(0) B^{-1} U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

if the matrix

$$B = \frac{1}{nh}\sum_{i=1}^n U\left(\frac{X_i - x}{h}\right) U^T\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

is invertible . In the fixed design case with equal spacings one can show that the matrix B has a lower bound $\lambda_0 > 0$ on its smallest eigenvalue that is uniform in x and $n \geq n_0$ for n_0 large enough, see p.40 in [79], so that the latter assumption is mild.

The following proposition gives a risk bound for the local polynomial estimator similar to the one in Theorem 22, but for fixed design.

Proposition 15. *Let $x_i = i/n$, $i = 1, ..., n$, be fixed design points on the interval $[0,1]$ and suppose $Y_i = m(x_i) + \epsilon_i$ where ϵ_i are mean zero i.i.d. random variables with finite variance σ^2. Let $\hat{m}_n(x) := \hat{m}_n^\ell(h,x)$ be the local polynomial estimator of order ℓ with compactly supported kernel $K \in L^\infty$, and assume that the smallest eigenvalue $\lambda_{\min}$ of B is greater than or equal to $\lambda_0 > 0$ for every $n \in \mathbb{N}$, $x \in [0,1]$. Suppose $m : \mathbb{R} \to \mathbb{R}$ is s-times differentiable with $m, D^s m \in L^\infty$, $s = \ell + 1$. Then there exists a constant $L := L(\|D^s m\|_\infty, \lambda_0, \sigma^2, \ell)$ such that for every $h > 0$, every $n \in \mathbb{N}$ such that $nh \geq 1$ and every $x \in [0,1]$ we have*

$$E\left|\hat{m}_n(x) - m(x)\right| \leq L\left(\frac{1}{\sqrt{nh}} + h^s\right).$$

Proof. Consider without loss of generality the case where K is supported in $[-1/2, 1/2]$. We need a few preliminary results: First, if Q is any polynomial of degree less than or equal to ℓ, then

$$\sum_{i=1}^n Q(x_i) W_{ni}(x) = Q(x). \tag{112}$$

To see this write

$$Q(x_i) = Q(x) + DQ(x)(x_i - x) + \cdots + \frac{D^\ell Q(x)(x_i - x)^\ell}{\ell!} =: q^T(x) U\left(\frac{x_i - x}{h}\right)$$

where $q(x) = (Q(x), DQ(x)h, \cdots, D^\ell Q(x) h^\ell)$. Setting $Y_i = Q(x_i)$ we obtain

$$\begin{aligned}
\hat{M}_n(h,x) &= \arg\min_{M \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n \left(Q(x_i) - M^T U\left(\frac{x_i - x}{h}\right)\right)^2 K\left(\frac{x_i - x}{h}\right) \\
&= \arg\min_{M \in \mathbb{R}^{\ell+1}} \sum_{i=1}^n \left((q(x) - M)^T U\left(\frac{x_i - x}{h}\right)\right)^2 K\left(\frac{x_i - x}{h}\right) \\
&= \arg\min_{M \in \mathbb{R}^{\ell+1}} (q(x) - M)^T B (q(x) - M)
\end{aligned}$$

which for invertible B is minimised at $M = q(x)$, and we obtain $\hat{m}_n(x) = Q(x)$. On the other hand by (111) we have $\hat{m}_n(x) = \sum_{i=1}^n Q(x_i) W_{ni}(x)$ for this Y_i, which establishes (112).

Moreover we have, using $\|U(0)\| = 1$ ($\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^{\ell+1}$)

$$\begin{aligned}
|W_{ni}(x)| &\leq \frac{1}{nh}\left\|B^{-1} U\left(\frac{x_i - x}{h}\right) K\left(\frac{x_i - x}{h}\right)\right\| \\
&\leq \frac{\|K\|_\infty}{\lambda_0 nh}\left\|U\left(\frac{x_i - x}{h}\right)\right\| I\left\{\left|\frac{x_i - x}{h}\right| \leq 1/2\right\} \leq \frac{C(\lambda_0, K)}{nh}
\end{aligned} \tag{113}$$

for some constant $C(\lambda_0)$, using the eigenvalue assumption on B. Likewise, using $nh \geq 1$ when nh is moderate and a Riemann sum approximation when $n >> h$ one has

$$\sum_{i=1}^{n} |W_{ni}(x)| \leq \frac{C(\lambda_0, K)}{nh} \sum_{i=1}^{n} I\left\{x - h/2 \leq i/n \leq x + h/2\right\} \leq C'(\lambda_0, K). \quad (114)$$

We now proceed to prove the proposition: (112) implies $\sum_i W_{ni}(x) = 1$ as well as $\sum_i (x_i - x)^k W_{ni}(x) = 0$ for $k = 1, ..., \ell$, so using also (114) and a Taylor expansion

$$\begin{aligned} |E\hat{m}_n(x) - m(x)| &= \left| \sum_{i=1}^{n} (m(x_i) - m(x)) W_{ni}(x) \right| \\ &= \left| \sum_{i=1}^{n} \frac{D^\ell m(x - \theta(x - x_i)) - D^\ell m(x)}{\ell!} (x_i - x)^\ell W_{ni}(x) \right| \\ &\leq \frac{\|D^s m\|_\infty}{\ell!} \sum_{i=1}^{n} |x_i - x|^s |W_{ni}(x)| I\left\{ \frac{|x_i - x|}{h} \leq 1/2 \right\} \leq Ch^s. \end{aligned}$$

where $C := C(\|D^s m\|_\infty, \ell, \lambda_0, K)$. Furthermore, using (113), (114), $E|X| \leq (EX^2)^{1/2}$ and independence

$$\begin{aligned} \left(E\left|\hat{m}_n(x) - E\hat{m}_n(x)\right|\right)^2 &= \left(E \left| \sum_{i=1}^{n} (Y_i - EY_i) W_{ni}(x) \right| \right)^2 \\ &\leq E\left(\sum_i \epsilon_i W_{ni}(x) \right)^2 \\ &= \sum_i W_{ni}^2(x) E(\epsilon_i^2) \\ &\leq \sigma^2 \sup_i |W_{ni}(x)| \sum_i |W_{ni}(x)| \\ &\leq \frac{C(\sigma^2, \lambda_0, K)}{nh} \end{aligned}$$

which completes the proof after taking square roots. □

3.5.3 More Regression Methods

Penalized nonparametric regression and cubic splines.

Consider again the fixed design regression model $Y_i = m(x_i) + \epsilon_i$, where we assume for simplicity that the design points are equally spaced on $[0, 1]$, say $x_i =$

$i/(n+1)$ for $i = 1, ..., n$. The estimators in (110) were obtained by a weighted least squares procedure. An alternative approach is to penalize the objective function to be minimized for complexity of the function m, measured, e.g., by the size

$$J(m) = \int_{\mathbb{R}} (D^2 m(x))^2 dx$$

of its second derivative. To be precise, we would like to minimize the objective function

$$Q(m, \lambda) = \sum_{i=1}^{n} (Y_i - m(x_i))^2 + \lambda J(m) \tag{115}$$

over all twice differentiable functions m (or, to be precise, all differentiable functions m with absolutely continuous Dm and $D^2 m \in L^2$). For each λ this minimizer is unique and can be explicitly characterised by a cubic spline $\hat{m}_n^S$ with breakpoints at the x_i's, see Schoenberg (1964) and Exercise 39. Similar to Theorem 17 every cubic spline can be uniquely decomposed into a linear combination of (suitably rescaled) cubic B-splines N_l, i.e.

$$\hat{m}_n^S(x) = \sum_{l=1}^{n+4} \hat{c}_l N_l(x).$$

Denote then by N the $n \times (n+4)$ matrix with entries $n_{kl} = N_l(k/(n+1))$ and by C the transposed $(n+4) \times 1$ vector of $\hat{c}_l$'s, then we can rewrite the minimization problem involving (115) as

$$\arg\min_C \left[(Y - NC)^T (Y - NC) + \lambda C^T \Omega C \right]$$

where the $(n+4) \times (n+4)$ matrix Ω has entry $\omega_{lk} = \int_{\mathbb{R}} D^2 N_l(x) D^2 N_k(x) dx$. This is now a simple linear algebra problem (similar to ridge-regression), and we obtain (cf. Exercise 40) that

$$C = (N^T N + \lambda \Omega)^{-1} N^T Y.$$

It should be clear the role of the 'bandwidth' h in local polynomial regression is paralleled by the parameter λ in penalized regression. In particular, one can show that these penalized spline estimates are equivalent to certain kernel estimators with a fixed kernel choice, see Silverman (1984), and then the techniques from kernel estimation can be applied here as well. For more details on cubic splines in regression estimation see Green and Silverman (1994).

Wavelet Regression.

Consider again the fixed design regression model $Y_i = m(x_i) + \epsilon_i$ with $x_i = i/n$ for $i = 1, ..., n$. Another approach to estimate m is to first approximate m by the

partial sum

$$K_j(m) = \sum_k < \phi_k, m > \phi_k + \sum_{l=0}^{j-1} \sum_k < \psi_{lk}, m > \psi_{lk}$$

of its wavelet series, cf. (79). To make this work we have to estimate the coefficients $< \phi_k, m >$ and $< \psi_{lk}, m >$ from the sample. A sensible choice is

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} Y_i \phi_k \left(\frac{i}{n} \right), \quad \hat{\beta}_{lk} = \frac{1}{n} \sum_{i=1}^{n} Y_i \psi_{lk} \left(\frac{i}{n} \right)$$

with expectations

$$E\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} m(i/n)\phi_k(i/n) \simeq \int m\phi_k, \quad E\hat{\beta}_{lk} = \frac{1}{n} \sum_{i=1}^{n} m(i/n)\psi_{lk}(i/n) \simeq \int m\psi_{lk}.$$

The *wavelet regression estimator* is then

$$\hat{m}_n^W(j, x) = \sum_k \hat{\alpha}_k \phi_k(x) + \sum_{l=0}^{j-1} \sum_k \hat{\beta}_{lk} \psi_{lk}(x). \tag{116}$$

To derive theoretical properties of these estimators one can proceed similar as in Proposition 14. See Chapter 10.8 in [47] for more details.

3.5.4 Exercises

Exercise 37. Suppose we are given two random variables Y and X, and given an observation from X, we want to predict the value of Y. Let $g(X)$ denote any predictor, i.e., g is any measurable function. Prove $E[(Y - g(X))^2] \geq E[(Y - E(Y|X))^2]$.

Exercise 38. Show that the local polynomial regression estimator $\hat{m}_n^\ell(h, x)$ equals the Nadaraya-Watson estimator for $\ell = 0$. Derive an explicit formula for the local polynomial estimator of the form $\hat{m}_n^\ell(h, x) = \sum_{i=1}^n Y_i W_{ni}(x)$ for some weight function $W_{ni}(x)$.

Exercise 39. A cubic spline r on $[0, 1]$ with breakpoints $0 < x_1 < x_2 < ... < x_n < 1$ is a continuous function that is a cubic polynomial over $(0, x_1), (x_i, x_{i+1})_{i=1}^{n-1}, (x_n, 1)$ and that has continuous first and second order derivatives at the knots. A cubic spline is called natural if $D^2 g(0) = D^2 g(1) = D^3 g(0) = D^3 g(1) = 0$, where third derivatives are understood one-sided. Let m be any minimizer of $Q(m) = \sum_{i=1}^n (Y_i - m(x_i))^2 + \lambda J(m)$ over the set of twice differentiable functions m defined

on $[0,1]$, where $x_i = i/(n+1)$, $i = 1, ..., n$, $Y_1, ..., Y_n$ are real numbers, $\lambda \in \mathbb{R}$ and $J(m) = \|m''\|_2^2$. Show that m can be taken to be a natural cubic spline . [You may use the fact that for any set of numbers $z_1, ..., z_n$ we can find a unique natural cubic spline g such that $g(x_i) = z_i$.]

Exercise 40. Let $p \geq n$ and let N be $n \times p$ and Ω a symmetric $p \times p$ matrix, C a $p \times 1$ and Y a $n \times 1$ vector. Consider the minimization problem

$$\arg\min_C \left[(Y - NC)^T(Y - NC) + \lambda C^T \Omega C\right]$$

where λ is a fixed scalar. Show that the minimzer satisfies $C = (N^T N + \lambda\Omega)^{-1} N^T Y$, assuming invertibility of the involved matrices when necessary

3.6 Choosing the Tuning Parameters

The risk bounds obtained in the last two sections did all depend on a 'tuning parameter', either the bandwidth h or the resolution level j. For instance, the optimal choices necessary to obtain the results (90), (93), (105) all depended on several unknown constants, most notably the unknown degree of differentiability of the unknown functions f or m. In practice then, how do we choose the bandwidth?

Another problem is the following: Suppose we know that f is once differentiable at x_0 but ten times differentiable at x_1. Then at the point x_1, we could use Daubechies wavelets of regularity 9 and estimate $f(x_1)$ very well with $2^{j_n} \simeq n^{1/21}$, achieving the optimal rate of convergence $n^{-10/21}$ in the pointwise risk. On the other hand, to estimate $f(x_0)$, this choice of j_n fails badly, having rate of convergence $n^{-1/21}$ in the pointwise risk, and we should rather take $2^{j_n} \simeq n^{1/3}$ to obtain the optimal rate $n^{-1/3}$. But then this estimator will be suboptimal for estimating $f(x_1)$. Can we find a single estimator that is optimal at both points?

At least two paradigms exist to address these questions. One is motivated from practice and tries to choose h and j_n in a way that depends both on the data and on the point x where we want to estimate f. While many procedures can be proposed, there are generally no theoretical justifications for or comparisons between these procedures.

The other paradigm could be called 'adaptive estimation', where one wants to devise estimators that, at least for large samples sizes, are *as good* (sometimes only 'nearly' as good) as the generally infeasible procedure that would be chosen if the smoothness of f were known. Adaptation can be 'spatial' (to the different smoothness degrees at different points in space), 'to the unknown smoothness of f', or to some other unknown structural property of f. Here some deep mathematical results can be obtained. The results are often only asymptotic (for large n), and performance in small samples is generally not well-understood, with some notable

exceptions. Furthermore, the adaptation paradigm has a concept of 'statistical performance' which derives from the minimax paradigm (cf. Section 3.2), and this itself can be questioned as being too pessimistic. In any case, adaptive estimation gives final theoretical answers of a certain kind to many nonparametric estimation problems.

3.6.1 Some Heuristic Methods

Cross Validation.

We start with a very simple yet practically effective method, and we present it in the framework of regression. Let $\hat{m}_n(h, x)$ be the Nadaraya Watson estimator from (103). Define

$$\hat{m}_{n,-i}(h, x) = \frac{\sum_{j=1, j\neq i}^{n} Y_j K((x - X_j)/h)}{\sum_{j=1, j\neq i}^{n} K((x - X_j)/h)},$$

the same estimator obtained from leaving the i-th observation out. Then the *leave-one-out cross validation score* is defined as

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}_{n,-i}(h, X_i))^2. \tag{117}$$

The idea would now be to choose h such that $CV(h)$ is minimized. For estimators that are 'linear' (as the Nadaraya-Watson estimator), one can compute this quantity in a numerically effective way. There are also generalized cross-validation ideas. One can then proceed heuristically with this choice of h to make statistical inference, see Chapters 5.2 to 5.9 in Wasserman (2006) for these and more facts and details.

Variable Bandwidth Choice.

Another way to choose h is dependent on the point at which one estimates f. Take for instance the kernel density estimator, then we propose

$$\hat{f}_n^K(x) = \sum_{i=1}^{n} \frac{1}{nh(x, i)} K\left(\frac{x - X_i}{h(x, i)}\right),$$

where now the bandwidth $h(x, i)$ may depend both on x and i. For example one can take $h(x, i) = h_k(x)$ to be the distance of x to the k-th nearest sample point. Or one takes $h(x, i) = h_k(i)$ the distance from X_i to the k-th nearest sample point. Another choice is $h_i \simeq f(X_i)^{-1/2}$, where of course f has to be replaced by a preliminary estimate. These methods are particularly designed to estimate the density in a 'localized' way, meaning that the bandwidth depends on the point

x. There seem to be no theoretical results confirming that these choices lead to good overall statistical procedures. But they are heuristically plausible and used in practice. We refer to the paper Terrell and Scott (1992) for more discussion and details.

Hard Thresholding.

If the nonparametric estimator comes from a series expansion, such as in the case of the Haar basis or wavelets, there is another simple heuristic to construct an estimator which, in a certain way, circumvents the problem of bandwidth-choice. Take, for instance, a general wavelet density estimator

$$f_n^W(j,x) = \sum_{k\in\mathbb{Z}} \hat{\alpha}_k \phi(x-k) + \sum_{l=0}^{j_0-1} \sum_{k\in\mathbb{Z}} \hat{\beta}_{lk}\psi_{lk}(x) + \sum_{l=j_0}^{j_1-1} \sum_{k\in\mathbb{Z}} \hat{\beta}_{lk}\psi_{lk}(x).$$

where the coefficients $\hat{\alpha}_k$ and $\hat{\beta}_{lk}$ are the empirical wavelet coefficients defined in (94). The idea of *hard thresholding* is to i) first choose a rather small level j_0 and a very large level j_1 (not depending on any unknown constants), and then ii) keep only those $\hat{\beta}_{lk}\psi_{lk}(x)$'s between the resolution levels j_0 and $j_1 - 1$ where $|\hat{\beta}_{lk}| > \tau$ for some *threshold* τ. More precisely, the thresholded wavelet density estimator is, for given j_0, j_1, τ.

$$f_n^T(x) = \sum_{k\in\mathbb{Z}} \hat{\alpha}_k \phi(x-k) + \sum_{l=0}^{j_0-1} \sum_{k\in\mathbb{Z}} \hat{\beta}_{lk}\psi_{lk}(x) + \sum_{l=j_0}^{j_1-1} \sum_{k\in\mathbb{Z}} \hat{\beta}_{lk} 1\{|\hat{\beta}_{lk}| > \tau\}\psi_{lk}(x). \quad (118)$$

This estimator was introduced and studied in Donoho and Johnstone (1995), Donoho, Johnstone, Kerkyacharian and Picard (1996). Of course, at first sight this just transfers the problem to the appropriate choice of the threshold τ. However, if $< \psi_{lk}, f >= 0$ so that the basis function ψ_{lk} has no significance in reconstructing f, then $\hat{\beta}_{lk} = n^{-1}\sum_{i=1}^n \psi_{lk}(X_i)$ is a centred i.i.d. sum of random variables, and if f is bounded the corresponding variances are uniformly bounded in l. By Bernstein's inequality (Exercise 32) such sample means make excursions of size $\sqrt{(\log n)/n}$ with probability vanishing polynomially in n, and this motivates the universal choice $\tau = C\sqrt{\log n/n}$, where C is some numerical constant which can be chosen to depend only on ψ and $\|f\|_\infty$ (the latter constant being estimable from the data in a simple way). It should be noted that, since ψ has compact support, this procedure effectively gives rise to a 'spatially variable resolution level choice', in the sense that the number of basis functions used to estimate f at x_0 can be very different from the number used at x_1.

3.6.2 Adaptive Estimation by Wavelet Thresholding

The heuristic procedures from the previous chapter all have their merits, but it is difficult to assess the statistical performance of the final estimator obtained, at least from a theoretical point of view. For example, we would like to have explicit risk bounds or rates of convergence for the regression estimator $\hat{m}_n(\hat{h}_n, x)$ where $\hat{h}_n$ was chosen by one of these procedures. Since $\hat{h}_n$ is now random, the proofs from the previous sections do not apply here.

Quite remarkably, the wavelet thresholding procedure can be shown to be *adaptive*: It has risk properties that are (almost) as good as an estimator built with the knowledge of the smoothness of f.

The main ideas of the proof of the following deep result are from Donoho, Johnstone, Kerkyacharian and Picard (1996). It shows that the estimator f_n^T with purely data-driven choice of τ, j_0, j_1 estimates a m-times differentiable density at the minimax rate of convergence in pointwise loss from Theorem 16, up to a term of logarithmic order in n, without requiring the knowledge of m. In the theorem below we require the density to have a *globally* bounded m-th derivative, so that the theorem does not prove spatial adaptivity. However, a refinement of this proof shows that this is also the case, i.e., one can relax the assumption of a globally bounded m-th derivative to existence of $D^m f$ in a neighborhood of x only, and to $D^m f$ being bounded in this neighborhood. We comment on the role of the constant κ after the proof of the theorem.

Theorem 23. *Let $X_1, ..., X_n$ be i.i.d. with bounded density f. Let f_n^T be the thresholding wavelet density estimator from (118) based on wavelets satisfying Condition 1 for some S. Choose $j_0 < j_1$ in such a way that $2^{j_0} \simeq n^{1/(2S+3)}$ and $n/\log n \leq 2^{j_1} \leq 2n/\log n$, and set $\tau := \tau_n = \kappa\sqrt{(\log n)/n}$. If f is m-times differentiable, $0 < m \leq S+1$, with $D^m f \in L^\infty$, then there exists a choice of κ depending only on $\|f\|_\infty$ and ψ such that*

$$E|f_n^T(x) - f(x)| = O\left(\left(\frac{\log n}{n}\right)^{m/(2m+1)}\right).$$

Proof. Writing β_{lk} for $< f, \psi_{lk} >$, we have

$$\begin{aligned} |f_n^T(x) - f(x)| \;\leq\; & |f_n^W(j_0)(x) - K_{j_0}(f)(x)| + |K_{j_1}(f)(x) - f(x)| \\ & + \left|\sum_{l=j_0}^{j_1-1}\sum_k (\hat{\beta}_{lk} 1\{|\hat{\beta}_{lk}| > \tau\} - \beta_{lk})\psi_{lk}(x)\right|. \end{aligned}$$

Using (97), the expectation of the first term on the r.h.s. of the inequality is of order

$$\sqrt{\frac{2^{j_0}}{n}} \simeq \left(\frac{1}{n}\right)^{\frac{S+1}{2(S+1)+1}} = O\left(\left(\frac{1}{n}\right)^{\frac{m}{2m+1}}\right)$$

since $m \leq S+1$, which is of smaller order than the bound required in the theorem. Similarly, using Proposition 9iii, the second term is of order

$$|K_{j_1}(f)(x) - f(x)| \leq C2^{-j_1 m} \simeq \left(\frac{\log n}{n}\right)^m$$

which is of smaller order than the bound required in the theorem in view of $m > m/(2m+1)$.

It remains to control the third term. We write

$$\sum_{l=j_0}^{j_1-1} \sum_k (\hat{\beta}_{lk} - \beta_{lk})\psi_{lk}(x) \left(1\{|\hat{\beta}_{lk}| > \tau, |\beta_{lk}| > \tau/2\} + 1\{|\hat{\beta}_{lk}| > \tau, |\beta_{lk}| \leq \tau/2\}\right)$$

$$- \sum_{l=j_0}^{j_1-1} \sum_k \beta_{lk}\psi_{lk}(x) \left(1\{|\hat{\beta}_{lk}| \leq \tau, |\beta_{lk}| > 2\tau\} + 1\{|\hat{\beta}_{lk}| \leq \tau, |\beta_{lk}| \leq 2\tau\}\right)$$

$$= I + II + III + IV,$$

and we treat these four terms separately.

We mention in advance some preliminary facts that we shall use repeatedly. First, for fixed x and l, all sums over k are finite and consist of at most $2a+1$ terms due to the compact support of ψ. (To be precise, only the k's with $2^l x - a \leq k \leq 2^l x + a$ are nonzero, where ψ is supported in $[-a, a]$.) Furthermore, we have

$$E(\hat{\beta}_{lk} - \beta_{lk})^2 \leq \frac{2^l}{n} \int \psi^2(2^l x - k) f(x) dx \leq \frac{1}{n}\|f\|_\infty, \tag{119}$$

recalling $\|\psi\|_2^2 = 1$. Finally, a bounded function with bounded m-th derivative can be shown to satisfy the estimate

$$\sup_k |\beta_{lk}| \leq d2^{-l(m+1/2)} \tag{120}$$

for some constant d: To see this note that $\int \psi(u)u^\alpha = 0$ for $0 \leq \alpha \leq S$ in view of Condition 1 allows us to write, using a Taylor expansion,

$$\begin{aligned} \beta_{lk} &= \int_{\mathbb{R}} \psi_{lk}(y)(f(y) - f(k2^{-l}))dy = 2^{-l/2} \int_{\mathbb{R}} \psi(u)(f((u+k)2^{-l}) - f(k2^{-l}))du \\ &= 2^{-l(m-1/2)} \int_{\mathbb{R}} \psi(u)u^{m-1} \left(\frac{D^{m-1}f((\zeta u + k)2^{-l}) - D^{m-1}f(k2^{-l})}{(m-1)!}\right) du \end{aligned}$$

which gives (120) by compact support of ψ.

About term (I): Let $j_1(m)$ be such that $j_0 \leq j_1(m) \leq j_1 - 1$ and

$$2^{j_1(m)} \simeq n^{1/(2m+1)}$$

(such $j_1(m)$ exists by the definitions). Using (119) and compact support of ψ, we have, for some constant that depends only on ψ and $\|f\|_\infty$ that

$$\begin{aligned}
&E\left|\sum_{l=j_0}^{j_1(m)-1}\sum_k(\hat\beta_{lk}-\beta_{lk})\psi_{lk}(x)I_{[|\hat\beta_{lk}|>\tau,|\beta_{lk}|>\tau/2]}\right| \\
&\le \sum_{l=j_0}^{j_1(m)-1}\sum_k\sqrt{E(\hat\beta_{lk}-\beta_{lk})^2}|\psi_{lk}(x)| \\
&\le C\sum_{l=j_0}^{j_1(m)-1}\sqrt{\frac{2^l}{n}} = O\left(\sqrt{\frac{2^{j_1(m)}}{n}}\right) = o\left(\left(\frac{\log n}{n}\right)^{m/(2m+1)}\right).
\end{aligned}$$

For the second part of (I), using (119), (120), the definition of τ and again compact support of ψ, we have

$$\begin{aligned}
&E\left|\sum_{l=j_1(m)}^{j_1-1}\sum_k(\hat\beta_{lk}-\beta_{lk})\psi_{lk}(x)I_{[|\hat\beta_{lk}|>\tau,|\beta_{lk}|>\tau/2]}\right| \\
&\le \sum_{l=j_1(m)}^{j_1-1}\sum_k\sqrt{E\left(|\hat\beta_{lk}-\beta_{lk}|\right)^2}\frac{2}{\kappa}\sqrt{\frac{n}{\log n}}\sup_k|\beta_{lk}||\psi_{lk}(x)| \\
&\le C(\log n)^{-1/2}\sum_{l=j_1(m)}^{j_1-1}2^{-lm} = o\left(\left(\frac{\log n}{n}\right)^{m/(2m+1)}\right).
\end{aligned}$$

For (II) we have, using (119) and the Cauchy-Schwarz inequality

$$\begin{aligned}
&E\left|\sum_{l=j_0}^{j_1-1}\sum_k(\hat\beta_{lk}-\beta_{lk})\psi_{lk}(x)I_{[|\hat\beta_{lk}|>\tau,|\beta_{lk}|\le\tau/2]}\right| \\
&\le \sum_{l=j_0}^{j_1-1}\sum_k\sqrt{E(\hat\beta_{lk}-\beta_{lk})^2}\Pr\{|\hat\beta_{lk}|>\tau,|\beta_{lk}|\le\tau/2\}^{1/2}|\psi_{lk}(x)| \\
&\le \frac{\|\psi\|_2\|\psi\|_\infty\|f\|_\infty^{1/2}}{\sqrt{n}}\sum_{l=j_0}^{j_1-1}2^{l/2}\sum_{k\in[2^jx-a,2^jx+a]}\Pr\{|\hat\beta_{lk}|>\tau,|\beta_{lk}|\le\tau/2\}^{1/2}
\end{aligned}$$

and we next analyse the probability appearing in the square root: Bernstein's

inequality (Exercise 32) gives, for $l \le j_1 - 1$ (and $n \ge e^2$),

$$\begin{aligned}
&\Pr\{|\hat{\beta}_{lk}| > \tau, |\beta_{lk}| \le \tau/2\} \\
&\le \Pr\{|\hat{\beta}_{lk} - \beta_{lk}| > \tau - |\beta_{lk}|, |\beta_{lk}| \le \tau/2\} \\
&\le \Pr\{|\hat{\beta}_{lk} - \beta_{lk}| > \tau/2\} \\
&= \Pr\left\{\left|\sum_{i=1}^{n}(\psi(2^l X_i - k) - E\psi(2^l X - k))\right| > 2^{-1}\kappa\sqrt{n\log n/2^l}\right\} \\
&\le 2\exp\left(-\frac{\kappa^2 \log n}{8\|\psi\|_2^2\|f\|_\infty + \frac{8}{3}\kappa\|\psi\|_\infty\sqrt{2^l \log n/n}}\right) \\
&\le 2\exp\left(-\frac{\kappa^2 \log n}{8\|f\|_\infty + \frac{8}{3\sqrt{\log 2}}\kappa\|\psi\|_\infty}\right), \qquad (121)
\end{aligned}$$

a bound which is independent of k. Consequently, we have the overall bound for (II)

$$C'\frac{1}{\sqrt{n}}\sum_{l=j_0}^{j_1-1} 2^{l/2}\exp\left(-\frac{\kappa^2 \log n}{16\|f\|_\infty + \frac{16}{3\sqrt{\log 2}}\kappa\|\psi\|_\infty}\right), \qquad (122)$$

which can be made as small as desired by choosing κ large enough. For term (III), using compact support of ψ, (120) and (121)

$$\begin{aligned}
&E\left|\sum_{l=j_0}^{j_1-1}\sum_k \beta_{lk}\psi_{lk}(x) I_{[|\hat{\beta}_{lk}|\le\tau, |\beta_{lk}|>2\tau]}\right| \\
&\le \sum_{l=j_0}^{j_1-1}(2a+1)2^{l/2}\|\psi\|_\infty \sup_k |\beta_{lk}| \sup_k \Pr\{|\hat{\beta}_{lk}| \le \tau, |\beta_{lk}| > 2\tau\} \\
&\le c\sum_{l=j_0}^{j_1-1} 2^{-lm}\sup_k \Pr\{|\hat{\beta}_{lk} - \beta_{lk}| > \tau\} \\
&\le c\sum_{l=j_0}^{j_1-1} 2^{-lm}\exp\left(-\frac{\kappa^2 \log n}{8\|p_0\|_\infty + \frac{8}{3\sqrt{\log 2}}\kappa\|\psi\|_\infty}\right) = o\left(\left(\frac{\log n}{n}\right)^{m/(2m+1)}\right)
\end{aligned}$$

for κ large enough. Finally, for term (IV) we have, using compact support of ψ and (120), that

$$\begin{aligned}
\left|\sum_{l=j_0}^{j_1-1}\sum_k \beta_{lk}\psi_{lk}(x) I_{[|\hat{\beta}_{lk}|\le\tau, |\beta_{lk}|\le 2\tau]}\right| &\le (2a+1)\|\psi\|_\infty \sum_{l=j_0}^{j_1-1}\sup_k 2^{l/2}|\beta_{lk}| I_{[|\beta_{lk}|\le 2\tau]} \\
&\le c\sum_{l=j_0}^{j_1-1}\min(2^{l/2}\tau, 2^{-lm}),
\end{aligned}$$

Let $\bar{j}_1(m) \in \{j_0, \ldots, j_1 - 1\}$ such that $2^{\bar{j}_1(m)} \simeq (n/\log n)^{1/(2m+1)}$ and estimate the last quantity by

$$c\sqrt{\frac{\log n}{n}} \sum_{j_0 \le l \le \bar{j}_1(m)-1} 2^{l/2} + c \sum_{\bar{j}_1(m) \le l \le j_1 - 1} 2^{-lm}$$

both of which are of order

$$O\bigg(\Big(\frac{\log n}{n} \Big)^{m/(2m+1)} \bigg),$$

completing the proof. □

To reduce technicalities, we did not specify the constant κ in the above theorem, but this can be done easily by tracking the constants explicitly in the proof, see Exercise 41. One can replace $\|f\|_\infty$ by $\|f_n^W(j_1)\|_\infty$ and the proof goes through as well, using results in [41].

We say that the hard thresholding estimator is *rate-adaptive within a logarithmic factor*, because it achieves the minimax rate of convergence from Theorem 16 up to a factor of order a power of $\log n$. Remarkably, an analogue of this theorem can be proved, for the same estimator (and compactly supported densities), for *all* L^p-loss functions. For $1 \le p < \infty$, this was proved in Donoho, Johnstone, Kerkyacharian and Picard (1996), and for $p = \infty$ in [41] (where no compact support of the density is needed, and no logarithmic penalty has to be paid). So the hard thresholding wavelet density estimator is rate adaptive within a logarithmic factor for all these loss functions simultaneously.

It should be noted that estimators that are *adaptive* in the sense of Theorem 23 have been studied extensively in the last 15 years, starting with path-breaking work by Lepski (1991), and later Donoho and Johnstone (1995), Donoho, Johnstone, Kerkyacharian and Picard (1996), Lepski and Spokoyni (1997), Lepski, Mammen and Spokoyni (1997), Barron, Birgé and Massart (1999), Tsybakov (1999), and many others. This has then led to many challenging new questions in recent years, such as the construction of adaptive confidence sets. The latter turns out to be a particularly intricate subject, see, for instance, Cai and Low (2004), Robins and van der Vaart (2006), Genovese and Wasserman (2008) and Giné and Nickl (2010).

3.6.3 Exercises

Exercise 41. Suppose you know $\|f\|_\infty$. Give an admissible choice for κ in the theorem on the pointwise risk of the hard thresholding wavelet density estimator.

Exercise 42. $^+$ Suppose the unknown density f is bounded, compactly supported, and m-times differentiable $D^m f \in L^\infty$. Prove that the hard thresholding density estimator based on Daubechies' wavelets of regularity S, and with $\tau = \kappa\sqrt{(\log n)/n}$ for suitable choice of κ, satisfies

$$E\|f_n^T - f\|_2^2 = O\left((n/\log n)^{-2m/(2m+1)}\right)$$

for $m \leq S+1$. (Hint: It is useful to apply Parseval's identity to the L^2-norm of the thresholded window, and then work in the sequence space ℓ^2 of square-summable sequences.)

3.7 Functional Estimation and Applications

So far we were able to come up with reasonable estimators of fundamental statistical quantities such as the distribution function F of a random variable, its density f, or a regression function m. The final aim of statistical inference is often to estimate some *aspect* of F, f, or m. These 'aspects' are usually simple functionals of F, f, or m. For example, we might be interested in the quantile function $\Phi(F) = F^{-1}$, the entropy $\Phi(f) = \int f \log f$, or a maximizer $\Phi(m) = \arg\max_{x\in[0,1]} m(x)$. Once we have phrased our statistical problem in that way, we can use the natural 'plug-in' estimators $\Phi(X_n)$ where X_n is one of the nonparametric estimators introduced in the previous sections.

Following the spirit of Proposition 4, a powerful and elegant way to derive properties of these plug-in estimators is to use continuity and differentiability properties of Φ on certain function spaces, typically spaces whose norms or metrics are given by the loss functions considered in the previous sections. We have already seen one instance of this approach in the proof of the Kolmogorov-Smirnov limit theorem (Theorem 13): There the continuous mapping $\Phi = \|\cdot\|_\infty$ was applied to the whole process $\sqrt{n}(F_n - F)$ (so that we obtained the limit of $\Phi(\sqrt{n}(F_n - F))$ from Theorem 10), but often the quantity of interest is $\Phi(F_n) - \Phi(F)$, and if Φ is nonlinear, one has to proceed differently.

3.7.1 The 'von Mises' or 'Functional Delta-Method'

Suppose we have a good estimate X_n for an object s_0 in a normed space $(S, \|\cdot\|_S)$, and we would like to estimate $\Phi(s_0)$ by $\Phi(X_n)$ where Φ is a real-valued mapping defined on S (or possibly a subset of it). Recalling the 'delta' method, our intuition would be to differentiate Φ around s_0, so we have to come up with a proper notion of differentiation on general normed spaces S. This approach was pioneered by von Mises (1947).

The classical notion of strong (or Fréchet-) differentiability of a real-valued mapping Φ defined on an open subset S_D of a normed space S at the point $s_0 \in S_D$

requires the existence of a linear continuous mapping $D\Phi_{s_0}[\cdot] : S \to \mathbb{R}$ such that

$$|\Phi(s_0 + h) - \Phi(s_0) - D\Phi_{s_0}[h]| = o(\|h\|_S).$$

See Dieudonné (1960) for a treatment of differentiation in general Banach spaces. However, this notion of differentiability is sometimes too strong for many statistical applications, and Reeds (1976) showed (in his PhD dissertation) that a weaker notion of differentiability still suffices for statistical applications.

Definition 4. *If S_D is a subset of a normed space S, then a mapping $\Phi : S_D \to \mathbb{R}$ is said to be Hadamard- (or compactly) differentiable at $s_0 \in S_D$ if there exists a linear continuous mapping $D\Phi_{s_0}[\cdot] : S \to \mathbb{R}$ such that*

$$\left| \frac{\Phi(s_0 + th_t) - \Phi(s_0)}{t} - D\Phi_{s_0}[h] \right| \to 0$$

for every $h \in S$, every $t \to 0$ and every h_t with $\|h_t - h\|_S \to 0$ and $s_0 + th_t \in S_D$. Furthermore, Φ is said to be Hadamard-differentiable tangentially to a set $S_0 \subset S$ by requiring $h_t \to h$ with $h \in S_0$.

It should be noted that, if S is finite-dimensional, then this notion can be shown to be equivalent to Fréchet differentiability, but in the infinite-dimensional case it is not. For statistical applications, the following result is central.

Proposition 16. *Let $(S, \|\cdot\|)$ be a normed space, let $S_D \subset S$ and let $\Phi : S_D \to \mathbb{R}$ be Hadamard-differentiable at $s_0 \in S_D$ tangentially to S_0, with derivative $D\Phi_{s_0}$. Let r_n be a real sequence such that $r_n \to \infty$ and let X_n be random variables taking values in S_D such that $r_n(X_n - s_0)$ converges in law to some random variable X taking values in S_0, as $n \to \infty$. Then*

$$r_n(\Phi(X_n) - \Phi(s_0)) \to^d D\Phi_{s_0}(X)$$

as $n \to \infty$.

Proof. We use here a theorem of Skorohod on 'almost surely convergent realizations of weakly convergent sequences of random variables': If Y_n, $n = 1, 2, ...$ are random variables taking values in a metric space (S_D, d) such that Y_n converges to Y_0 in law, then there exist a probability space $(W, \mathcal{W}, \mu)$ and random variables $\tilde{Y}_n : (W, \mathcal{W}, \mu) \to S_D$, $n = 0, 1, 2, ...$ such that $Y_n = \tilde{Y}_n$ in distribution and $d(\tilde{Y}_n, \tilde{Y}_0)$ converges to zero almost surely as $n \to \infty$. See [29], Theorem 11.7.2 for a proof (and also [28], Theorem 3.5.1, for the nonseparable case).

To prove the proposition, we apply this result and construct random variables $\tilde{X}_n$, $\tilde{X}$ such that $r_n(\tilde{X}_n - s_0)$ converges to $\tilde{X}$ almost surely. But now

$$\left| r_n(\Phi(\tilde{X}_n) - \Phi(s_0)) - D\Phi_{s_0}[\tilde{X}] \right| = \left| \frac{\Phi(s_0 + r_n^{-1} r_n(\tilde{X}_n - s_0)) - \Phi(s_0)}{r_n^{-1}} - D\Phi_{s_0}[\tilde{X}] \right|$$

converges to zero almost surely in view of Hadamard differentiability of Φ, since on a set of probability one we have $r_n(\tilde{X}_n - s_0) \to \tilde{X}$. Since $r_n(\Phi(X_n) - \Phi(s_0)) - D\Phi_{s_0}(X)$ has – by construction – the same distribution as $r_n(\Phi(\tilde{X}_n) - \Phi(s_0)) - D\Phi_{s_0}(\tilde{X})$, and since almost sure convergence implies convergence in law, the result follows. □

Classically one tries to combine this proposition with Theorem 10, so the choice for $r_n(X_n - s_0)$ is $\sqrt{n}(F_n - F)$, the empirical process, and $S = L^\infty$. Since L^∞ is a very large space, complications can arise, some from the fact that the functional Φ might not be defined on all of L^∞, others related to measurability issues. The latter can be circumvented by introducing a slightly different definition of convergence in law (e.g., Chapter 3 in [28]), and an analogue of the above proposition can still be proved in this case, see, e.g., van der Vaart (1998, Theorem 20.8). However, these problems partially disappear if one uses Theorem 20, which allows one to take $X_n = F_n^K$ the distribution function of a kernel density estimator, which is a continuous function, so that S can simply be taken to be the space of bounded continuous functions (or some other subspace of it).

The functional delta method has many applications to statistics, and we shall show how Theorems 14 and 15 can be proved using it, but there are many other applications, see, e.g., Gill (1989), Pitts (1994) and Chapter 20 in [81]. We also refer to Dudley (1992, 1994), who takes a different approach (using Fréchet differentiability with 'stronger' norms). Next to delivering elegant proofs one of the main appeals of the functional delta-method is that it separates the analysis from the probability part in a given statistical problem.

Example 9 (Differentiability of the quantile transform)**.** Recall the notation and assumptions from Theorem 25. Set $\phi(F) = F^{-1}(p)$, and consider the domain of definition $S_D \subset L^\infty$ to be the set of increasing functions G taking values between $[0, 1]$ for which the inequalities

$$G(G^{-1}(p)-) \leq p \leq G(G^{-1}(p)) \tag{123}$$

hold. Clearly $F \in S_D$ since its inverse exists, and also $F_n \in S_D$. Choose ξ_p such that $F(\xi_p) = p$ (which is possible by assumption on F) and set $\xi_{tp} = \phi(F + tH_t)$, where H_t is a sequence of functions in S_D that converges uniformly to H, which we take to be continuous at ξ_p. We want to differentiate $\phi(F + tH_t) - \phi(F) = \xi_{tp} - \xi_p$ tangentially to the subspace S_0 of functions that are continuous at ξ_p. If we knew that $G(G^{-1}(p)) = p$ for every $G \in S_D$, then applying the mean value theorem to the identity $(F + tH_t)(\xi_{tp}) = p$ and noting $F(\xi_p) = p$ would give

$$F'(\xi_p)(\xi_{tp} - \xi_p) + tH_t(\xi_{tp}) = o(|\xi_p - \xi_{tp}|),$$

and, since $\xi_{tp} \to \xi_p$ as $t \to 0$ as is easy to prove, one concludes

$$\left| \frac{\xi_{tp} - \xi_p}{t} - \frac{H}{F'}(\xi_p) \right| \to 0.$$

This already suggests that the Hadamard derivative should be

$$D\phi_F(H) = \frac{H}{F'}(F^{-1}(p)),$$

but does not quite work for us because F_n does not have a proper inverse. Even if we take F_n^K instead of F_n, and K such that F_n^K is invertible, then $F_n^K - F$ is not necessarily invertible. To overcome this problem, one has to be a little more careful and work with the inequalities (123), and this is left as Exercise 43, which completes the proof of Theorem 25 using Theorem 10 and Proposition 16, and noting that we can differentiate tangentially to the space of continuous functions since the F-Brownian bridge is sample-continuous almost surely.

Example 10 (Differentiability of Cumulative Hazard Functions)**.** If $D([0,\infty)$ is the space of bounded right-continuous real-valued functions on $[0,\infty)$ with left limits equipped with the supremum norm, and if we show that the mapping

$$\Lambda := \Lambda_u : F \mapsto \int_0^u (1 - F(x-))^{-1} dF(x)$$

is Hadamard-differentiable on a suitable subset of $D([0,\infty))$, then we have proved Theorem 15 by virtue of Theorem 10 and Proposition 16. Alternatively, one can combine Theorem 20 and Proposition 16 to prove an analogue of Theorem 15 for the kernel density estimator with continuous kernel K, and assuming that F has a continuous density $f = DF$. In this case we can replace $D([0,\infty))$ by the space $C^1(\mathbb{R})$ of bounded continuously differentiable functions on $\mathbb{R}$, still equipped with the supnorm. We give the details for this simpler case, and leave the case of $D([0,\infty))$, which applies to the empirical distribution function, to Exercise 44.

Theorem 24. *Suppose X is a nonnegative random variable with distribution function $F : [0,\infty) \to \mathbb{R}$ that has a continuous density DF, and let t be a point such that $1 - F(t) > 0$. Let further $F_n^K(t, h_n)$ be as in Theorem 20 where $h_n = o(n^{-1/2})$ and where the kernel $K : \mathbb{R} \to \mathbb{R}$ is a continuous function. Then $\sqrt{n}(\Lambda(F_n^K) - \Lambda(F))$ is asymptotically normally distributed with limit as in Theorem 15.*

Proof. On $S := (C^1(\mathbb{R}), \|\cdot\|_\infty)$ the functional $\Lambda : S \to \mathbb{R}$ is given by

$$\Lambda(F) = \int_0^u (1 - F(x))^{-1} DF(x) dx,$$

and can be viewed as the composition of the mappings

$$F \mapsto (F, 1-F) \mapsto (F, (1-F)^{-1}) \mapsto \int_0^u (1-F(x))^{-1} DF(x)dx. \tag{124}$$

If $C(\mathbb{R})$ is the space of bounded continuous functions on $\mathbb{R}$, then the first mapping is bi-linear and continuous, hence Fréchet differentiable from $C(\mathbb{R})$ to $C(\mathbb{R})^2$ (meaning the Cartesian product here), and then also Fréchet-differentiable at any point in $S \subset C(\mathbb{R})$. Next, let $B(\zeta) = \{f \in C(\mathbb{R}) : \sup_{0\le x\le u} f(x) < 1-\zeta\}$, which is open in $C(\mathbb{R})$. The second mapping in (124) is then Fréchet differentiable for every $(F, G) \in C(\mathbb{R}) \times B(\zeta) \subset C(\mathbb{R})^2$ and every $\zeta > 0$, by the chain rule (since $x \mapsto 1/x$ is differentiable on (ζ, ∞)). By the chain rule for Hadamard-differentiable mappings (Theorem 20.9 in [81]) it remains to prove Hadamard-differentiability of the third mapping in (124), namely the mapping

$$\Phi : (F, G) \mapsto \int_0^u G(x)DF(x)dx,$$

defined on $S_D \times S_D \subset S \times S$ where $S_D = \{F \in C^1(\mathbb{R}) : \|DF\|_1 \le 1\}$. To achieve this, recall Definition 4 above and take uniformly convergent sequences $h_{1t} \to h_1 \in S$ and $h_{2t} \to h_2 \in S$ such that $G + th_{1t}$ and $F + th_{2t}$ are all in S_D, which implies that both th_{1t} and th_{2t} are contained in $2S_D$ for every t. Then, for $s_0 = (F, G)$ and $h_t = (h_{2t}, h_{1t})$ we have, using integration by parts that

$$\begin{aligned}
\frac{\Phi(s_0 + th_t) - \Phi(s_0)}{t} &= \frac{\int_0^u (G + th_{1t})(DF + tDh_{2t}) - \int_0^u GDF}{t} \\
&= \int h_{1t}DF + \int GDh_{2t} + t\int h_{1t}Dh_{2t} \\
&= \int h_{1t}DF - \int h_{2t}DG + Gh_{2t}|_0^u + o(1) \\
&\to \int h_1 DF - \int h_2 DG + Gh_2|_0^u =: D\phi_{s_0}(h)
\end{aligned}$$

as $t \to 0$ since (recalling that $h_1 \in S = C^1(\mathbb{R})$, and using integration by parts again)

$$\begin{aligned}
\left|t\int h_{1t}Dh_{2t}\right| &\le t\left|\int h_1 Dh_{2t}\right| + \left|\int (h_{1t} - h_1)tDh_{2t}\right| \\
&\le t\left|\int Dh_1 h_{2t} - h_1 h_{2t}|_0^u\right| + 2\|h_{1t} - h_1\|_\infty = O(t) + o(1).
\end{aligned}$$

This completes the proof of Hadamard-differentiability of Λ on the domain $S_D \cap B(\zeta) \subset S$. To apply Proposition 16, note that $F = s_0$ and $X_n = F_n^K$ are both

contained in S_D: Clearly $\|DF\|_1 = 1, \|DF_n^K\|_1 = 1$ (since DF_n^K and DF are densities) and also $F(t) \leq 1 - \zeta/2$ for some $\zeta > 0$ which implies $F_n^K(t) \leq 1 - \zeta$ on a set whose probability approaches one by combining Theorem 20 with Theorem 8. Finally, the second part of Theorem 20 implies that $\sqrt{n}(X_n - s_0) = \sqrt{n}(F_n^K - F) \to \mathbb{G}_F$ in S, so that Proposition 16 gives the result. □

3.7.2 The 'Plug-in' property of density estimators

In Proposition 16 we have assumed that $r_n(X_n - s_0)$ converges in law in the normed space S. This is useful when convergence in law actually holds, such as in Theorems 3, 10 and 20. However, in other cases, in particular in density estimation, convergence in law might not hold, and one might just have a rate of convergence in some norm. Another problem with using the empirical distribution function F_n as a plug-in estimator is that some functionals are only defined for densities. For instance, if Φ is defined on a set of square-integrable densities, and $\Phi : L^2 \to \mathbb{R}$ is Fréchet differentiable, then we can estimate $\Phi(f)$ by the kernel-plug-in-estimator $\Phi(f_n^K)$, but $\Phi(F_n)$ does not necessarily make sense here.

Clearly under Fréchet-differentiability the derivative $D\Phi_f$ is a bounded linear map, so

$$|\Phi(f_n^K) - \Phi(f)| = D\Phi_f[f_n^K - f] + o(\|f_n^K - f\|_2) = O(\|f_n^K - f\|_2),$$

and a rate of convergence of f_n^K to f in L^2 carries over to $\Phi(f_n^K) - \Phi(f)$.

However we might be missing a substantial point here! On the one hand, the linear term $D\Phi_f[f_n^K - f]$ might have a much faster convergence rate (the functional-analytic intuition is that the 'topology' of convergence of linear functionals is 'weaker' than the norm topology), and the remainder in the linear approximation might be much smaller than just $o(\|f_n^K - f\|_2)$, so that $\Phi(f_n^K) - \Phi(f)$ could actually converge at a much faster rate than $\|f_n^K - f\|_2$, potentially in law, with a nice limiting distribution. An example for this is the functional $\Phi(f) = \int_{\mathbb{R}} f^2(x)dx$, where we can obtain an exact limit theorem for $\Phi(f_n^K) - \Phi(f)$ at rate $1/\sqrt{n}$ in a situation where $\|f_n^K - f\|_2$ is only of order $n^{-1/3}$.

Proposition 17. *Let $X_1, ..., X_n$ be i.i.d. with density f, and let $f_n^K(x)$ be the kernel density estimator from (87) with $h_n \simeq n^{-1/3}$ and with bounded, symmetric and compactly supported kernel K. Suppose f is continuously differentiable with derivative $Df \in L^1 \cap L^\infty$. Let $\Phi : L^2 \to \mathbb{R}$ be the mapping*

$$f \mapsto \Phi(f) = \int_{\mathbb{R}} f^2(x)dx.$$

Then

$$\sqrt{n}(\Phi(f_n^K) - \Phi(f)) \to^d N(0, \sigma(f))$$

where $\sigma(f) = 4\left[\int_{\mathbb{R}} f^3(x)dx - \left(\int_{\mathbb{R}} f^2(x)dx\right)^2\right]$.

Proof. The fundamental theorem of calculus and $Df \in L^1$ implies $f \in L^\infty$, which in turn implies, for densities f, that $f \in L^2$ and that $\lim_{|x|\to\infty} f(x) = 0$ (the latter fact is not necessary in this proof but useful). Moreover, for $D\Phi_f[h] = 2 < f, h >$, we have

$$\begin{aligned} |\Phi(f+h) - \Phi(f) - D\Phi_f[h]| &= | < f+h, f+h > - < f, f > -2 < f, h > | \\ &= < h, h >= O(\|h\|_2^2). \end{aligned}$$

Hence, using Exercise 33 to control the remainder

$$E|\Phi(f_n^K) - \Phi(f) - D\Phi_f[f_n^K - f]| = E\|f_n^K - f\|_2^2 = O(n^{-2/3}) = o(n^{-1/2})$$

so that $\sqrt{n}(\Phi(f_n^K) - \Phi(f))$ has the same limiting distribution as

$$\sqrt{n} D\Phi_f[f_n^K - f] = \sqrt{n} \int_{\mathbb{R}} 2f(x)(f_n^K - f)(x)dx.$$

Furthermore, writing $f(x) = \int_{-\infty}^x Df(t)dt = \int_{\mathbb{R}} 1_{(-\infty,x)}(t)Df(t)dt$ and using Fubini's theorem

$$\begin{aligned} & E\left|\int_{\mathbb{R}} 2f(x)f_n^K(x)dx - \frac{2}{n}\sum_{i=1}^n f(X_i)\right| \\ &= E\left|\int_{\mathbb{R}} 2\left(\int_{\mathbb{R}} 1_{(-\infty,x)}(t)f_n^K(x)dx - \frac{1}{n}\sum_{i=1}^n 1_{(-\infty,X_i)}(t)\right) Df(t)dt\right| \\ &= E\left|\int_{\mathbb{R}} 2\left(\int_{\mathbb{R}} 1_{(t,\infty)}(x)f_n^K(x)dx - \frac{1}{n}\sum_{i=1}^n 1_{(t,\infty)}(X_i)\right) Df(t)dt\right| \\ &\le 2E\sup_{t\in\mathbb{R}}|F_n^K(t) - F_n(t)|\|Df\|_1 = O(n^{-2/3}\sqrt{\log n}) = o(n^{-1/2}) \end{aligned}$$

where we have used Theorem 20. We finally conclude that $\sqrt{n}(\Phi(f_n^K) - \Phi(f))$ has the same limiting distribution as

$$\sqrt{n}\left(\frac{2}{n}\sum_{i=1}^n f(X_i) - 2\int_{\mathbb{R}} f^2(x)dx\right) = \frac{2}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - Ef(X))$$

which converges to the required limit by the central limit theorem, completing the proof. □

While $\sqrt{n}$-consistent estimators for $\int f^2$ can be constructed by different means (e.g., Hall and Marron (1987)), the plug-in approach presented here has more the flavour of a refined functional delta method. More details on this and examples can be found in Section 3.3 in [61] and Exercise 45.

3.7.3 Exercises

Exercise 43. $^+$ Let $D([a,b])$ be the space of bounded right-continuous real-valued functions on $[a,b]$ with left limits, equipped with the supnorm. For $p \in [a,b]$, let S_D be the subset of $D([a,b])$ consisting of all nondecreasing functions F such that $F(F^{-1}(p)-) \leq p \leq F(F^{-1}(p))$ where F^{-1} is the generalized inverse function. Let $F_0 \in S_D$ be differentiable at x_p where $F(x_p) = p$, with positive derivative. Show that the mapping $\phi(F) = F^{-1}(p)$, $\phi : S_D \subset D([a,b]) \to \mathbb{R}$ is Hadamard-differentiable at F_0 tangentially to the set of functions in $D([a,b])$ that are continuous at x_p.

Exercise 44. $^+$ A function $f : [a,b] \to \mathbb{R}$ is of bounded variation if it can be written as $f(x) = f(a) + \int_a^x d\mu_f(u)$ for some finite (signed) measure μ_f. Consider the functional ϕ from $D([a,b]) \times BV_0([a,b]) \to \mathbb{R}$ given by $(F,G) \mapsto \int_0^y F(x) d\mu_G(x)$, where $BV_0([a,b])$ is the space of functions f of total variation on $[a,b]$ bounded by one satisfying also $f(a) = 0$. Prove that ϕ is Hadamard-differentiable at every pair of functions $(F_1, F_2) \in BV_0([a,b]) \times BV_0([a,b])$.

Exercise 45. Show that the mapping $\phi : f \mapsto \int_a^b f(x) \log f(x) dx$ is Fréchet differentiable as a mapping from $L^\infty([a,b])$ to $\mathbb{R}$ at any density $f \in L^\infty$ that is bounded away from zero on $[a,b]$. What is the derivative? How would you estimate it using a kernel (or wavelet) plug-in-density estimator $\phi(f_n^K)$, and what rate of convergence do you expect if you assume, e.g., that f is also once differentiable?

References

[1] Akaike, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math.* **6** 127-132.

[2] Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301-413.

[3] Bernstein, S., *Theory of Probability*, in Russian.

[4] Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be 'plugged-in'. *Ann. Statist.* **31** 1033-1053.

[5] Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. Correction *ibid.* **3** 1370.

[6] Billingsley, P. (1999). *Convergence of Probability Measures.* 2nd edition. Wiley, New York.

[7] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer Berlin, Heidelberg.

[8] Cai, T.T. and Low, M.G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805-1840.

[9] Candès, E., Romberg, J. and Tao, T. (2006a). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory.* **52** 489-509.

[10] Candès, E., Romberg, J. and Tao, T. (2006b). Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics.* **59** 1207-1223.

[11] Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313-2351.

[12] Cantelli, F.P. (1933). Sulla determiniazione empirica della leggi di probabilitá. *Giorn. Ist. Ital. Attuari* **4** 421-424.

[13] Chacon, J.E. (2010). A note on the universal consistency of the kernel distribution function estimator. *Statist. Probab. Letters,* to appear.

[14] Curry, H.B. and Schoenberg, I.J. (1966). On polya frequency functions IV: The fundamental spline functions and their limits. *J. d'Analyse Math.* **17** 71-107.

[15] Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

[16] Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure and Appl. Math.* **41** 909-996.

[17] Daubechies, I. (1992). *Ten lectures on wavelets.* SIAM, Philadelphia.

[18] DeVore, R.A. and Lorentz, G.G. (1993). *Constructive Approximation.* Springer, New York.

[19] Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation.* Springer, New York.

[20] Dieudonné, J. (1960). *Foundations of modern analysis.*, Academic Press, New York.

[21] Donsker, M. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23** 277-281.

[22] Donoho, D. L. and Johnstone, I. M. (1994). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200-1224.

[23] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508-539.

[24] Doob, J.L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20** 393-403.

[25] Dudley, R.M. (1966). Weak convergences of probabilities on nonseperable metric spaces and empirical measures on Euclidean spaces. *Illinois J. Math.* **10** 109-126.

[26] Dudley, R.M. (1992). Fréchet differentiability, p-variation and uniform Donsker classes. *Annals of Probability* **20**, 1968-1982.

[27] Dudley, R.M. (1994). The order of the remainder in derivatives of composition and inverse operators for p-variation norms. *Annals of Statistics* **22** 1-20.

[28] Dudley, R.M. (1999). *Uniform Central Limit Theorems.* Cambridge University Press, Cambridge, UK.

[29] Dudley, R.M. (2002). *Real Analysis and Probability.* Cambridge University Press, Cambridge, UK.

[30] Dudley, R.M., Giné, E. and Zinn, J. (1991). Uniform and universal Glivenko-Cantelli classes. *J. Theoret. Probab.* **4** 485-510.

[31] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. (1956) Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27** 642-669.

[32] Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.* **21** 600-610.

[33] Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, Series A.* **222** 309-368.

[34] Fisher, R.A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society.* **22** 700-725.

[35] Folland, G.B. (1999). *Real analysis.* Wiley, New York.

[36] Genovese, C. and Wasserman, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875-905.

[37] Gill, R.D. (1989). Non- and semiparametric maximum likelihood estimators and the von-Mises method. (part I). *Scandinavian Journal of Statistics* **16**, 97-128.

[38] Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann Inst. H. Poincaré* **38** 907-921.

[39] Giné, E., Koltchinskii, V. and Sakhanenko, L. (2004). Kernel density estimators: convergence in distribution for weighted sup-norms. *Probab. Theory Related Fields* **130** 167-198.

[40] Giné, E. and Nickl, R. (2009a). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory and Related Fields* **143** 569-596.

[41] Giné, E. and Nickl, R. (2009b). Uniform limit theorems for wavelet density estimators. *Annals of Probability* **37** 1605-1646.

[42] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.

[43] Glivenko, V.I. (1933). Sulla determiniazione empirica della leggi di probabilitá. *Giorn. Ist. Ital. Attuari* **4** 92-99.

[44] Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models. A roughness penalty approach.* Chapmann and Hall, London.

[45] Haar, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Math. Ann.* **69** 331-371.

[46] Hall, P. and Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109-115.

[47] Härdle, W.; Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, approximation, and statistical applications.* Lecture Notes in Statistics **129**. Springer, New York.

[48] Johnstone, I.M., Kerkyacharian, G.; Picard, D. and Raimondo, M. (2004). Wavelet deconvolution in a periodic setting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 547-573.

[49] Kiefer, J. and Wolfowitz, J. (1976). Asymptotically minimax estimation of concave and convex distribution functions. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **34** 73–85.

[50] Kolmogorov, A.N. (1933). Sulla determiniazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari* **4** 83-91.

[51] Laplace, P.S. (1810). Mémoire sur les formules qui sont fonctions des tres grand nombres et sur leurs application aux probabilités. *Oeuvres de Laplace* **12** 301-345, 357-412.

[52] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

[53] Le Cam, L. and Yang, G.L. (1990). *Asymptotics in Statistics. Some Basic Concepts.* Springer, New York.

[54] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces.* Springer, Berlin.

[55] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682-697

[56] Lepski, O.V.; Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929-947.

[57] Lepski, O.V. and Spokoiny, V.G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512-2546.

[58] Meyer, Y. (1992). *Wavelets and operators.* Cambridge University Press, Cambridge, UK.

[59] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283.

[60] Nadaraya, E.A. (1964). On estimating regression. *Theory Probab. Appl.*, **9**, 141-142.

[61] Nickl, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Related Fields* **138** 411-449.

[62] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist.*, to appear.

[63] Parzen, E. (1962). On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33** 1065-1076.

[64] Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27** 2033-2053.

[65] Pitts, S.M. (1994). Nonparametric estimation of the stationary waiting time distribution function for the GI/G/1 queue. *Annals of Statistics* **22** 1428-1446.

[66] Pötscher, B.M. and Prucha, I.R. (1997). *Dynamic nonlinear econometric models. Asymptotic theory.* Springer, Berlin.

[67] Pötscher, B.M. (2007). *Asymptotic properties of M-estimators.* unpublished lecture notes.

[68] Reeds, J.A. (1976). *On the definition of von Mises functionals.* PhD dissertation, Department of Statistics, Harvard University, Cambridge, MA.

[69] Robins, J. and van der Vaart, A.W. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229-253.

[70] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832-837.

[71] Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. U.S.A.* **52** 947-950.

[72] Shorack, G.R. and Wellner, J.A. (1986). *Empirical processes. With Applications to Statistics*. Wiley, New York.

[73] Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.* **12** 898-916.

[74] Skorohod, A.V. (1956). Limit theorems for stochastic processes [in Russian]. *Theory Probab. Appl.* **1** 261-290.

[75] Smirnov, N.V. (1939). Estimation of the deviation between empirical distribution curves of two independent samples [in Russian]. *Bull.Univ.Moscow* **2** 3-14.

[76] Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. (Russian) *Doklady Akad. Nauk SSSR* **74** 189-191.

[77] Terrell, G. R. and Scott, D.W. (1992). Variable kernel density estimation. *Ann. Statist.* **20** 1236-1265.

[78] Tsybakov, A.B. (1998). Pointwise and sup-norm sharp adaptive estimation of the functions on the Sobolev classes. *Ann. Statist.* **26** 2420-2469.

[79] Tsybakov, A.B. (2009). *Introduction to nonparametric estimation.* Springer, New York.

[80] VAN DE GEER, S. (2000). *Empirical processes in M-estimation*. Cambridge University Press, Cambridge.

[81] VAN DER VAART, A.W (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

[82] VAN DER VAART, A.W. and WELLNER, J.A. (1996). Weak convergence and empirical processes. Springer, New York.

[83] VON MISES, R. (1931). *Wahrscheinlichkeitsrechnung*, Springer.

[84] VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **20** 309-348.

[85] WASSERMAN, L. (2006). *All of nonparametric statistics*. Springer, New York.

[86] WALD, A. (1949). Note on the consistency of the Maximum Likelihood Estimate. *Ann. Math. Statist.* **20** 595-601

[87] WATSON, G.S. (1964). Smooth regression analysis. *Sankhya, Ser. A* **26**, 359-372.