Lectures on statistical theory for high-dimensional and infinite-dimensional models

THIS IS A DRAFT
under major construction
certainly contains errors
use at your own risk!

Harry van Zanten

June 11, 2018 (this version)
Contents

1 High-dimensional, or nonparametric statistics 5
  1.1 Asymptotics in smooth, parametric models .............. 5
  1.2 High-dimensional and nonparametric models ............ 7
  1.3 Exercises ........................................... 12

2 The signal-in-white-noise model 15
  2.1 Definition of the model ................................ 15
  2.2 Brownian motion and Wiener integrals .................. 16
  2.3 Sequence formulation of the model ....................... 18
  2.4 Expressions for the likelihood ........................ 19
  2.5 Asymptotic equivalence of nonparametric models ........ 21
  2.6 Exercises ........................................... 25

3 Stein’s phenomenon, regularization 27
  3.1 The James-Stein theorem ................................ 27
  3.2 Regularization in the white noise model ............... 30
  3.3 Regularization in other models ......................... 35
  3.4 The bias-variance trade-off and convergence rates ....... 38
  3.5 Exercises ........................................... 47

4 Lower bounds for testing 51
  4.1 Statistical tests in the white noise model .............. 51
  4.2 Two simple hypotheses ............................... 52
  4.3 Composite alternative hypothesis ....................... 55
  4.4 Testing multiple hypotheses .......................... 59
  4.5 Exercises ........................................... 63

5 Lower bounds for estimation 65
  5.1 Minimax lower bounds for estimation .................... 65
  5.2 Connection to multiple hypotheses testing problems .... 66
  5.3 Minimax risk and Bayes risk .......................... 68
  5.4 Lower bounds for estimating smooth functions .......... 70
  5.5 Minimax lower bounds in other statistical settings .... 73
  5.6 Exercises ........................................... 76
## Adaptive estimation

6.1 Adaptation and data-driven tuning ........................................ 79
6.2 Lepski’s method ................................................................. 80
6.3 Exercises ........................................................................ 84

**Bibliography** ...................................................................... 85
Chapter 1

High-dimensional, or nonparametric statistics

1.1 Asymptotics in smooth, parametric models

In most introductory courses on mathematical statistics the focus is on classical statistical models in which the number of unknown parameters is small relative to the sample size. If such models depend smoothly on the parameter of interest, then under regularity conditions we typically have that as the sample size $n$ tends to infinity, maximum likelihood or Bayesian estimators converge at the rate $\sqrt{n}$ to the parameter corresponding to the true distribution that generated the data. Moreover, asymptotically such estimators are normally distributed and efficient, for instance in the sense that they have minimal asymptotic variance.

Consider for example an i.i.d. sample $X_1, \ldots, X_n$ from a positive density $p_\theta$ depending smoothly on a real-valued parameter $\theta$. Denote the true parameter by $\theta_0$, i.e. the true marginal density of the sample is $p_{\theta_0}$. Then the maximum likelihood estimator (MLE) $\hat{\theta}_n$ is the maximizer of the random function $\theta \mapsto -\sum (\ell_\theta - \ell_{\theta_0})(X_i)$, where $\ell_\theta(x) = \log p_\theta(x)$. By the law of large numbers this function converges pointwise to the corresponding expectation under $P_{\theta_0}$, which is minus the Kullback-Leibler (KL) divergence $-\text{KL}(p_{\theta_0}, p_\theta)$. Here the KL-divergence between two probability measures with positive densities $f$ and $g$ with respect to a dominating measure $\mu$ is defined as

$$\text{KL}(f, g) = \int f \log \frac{f}{g} d\mu.$$ 

If the model is identifiable the function $\theta \mapsto -\text{KL}(p_{\theta_0}, p_\theta)$ has a unique maximum at $\theta_0$ (see Exercise 1.1). Under regularity conditions it can be shown that the maximizer $\hat{\theta}_n$ of $\theta \mapsto -\sum (\ell_\theta - \ell_{\theta_0})(X_i)$ converges in probability under $P_{\theta_0}$ to the maximizer $\theta_0$ of the limiting function $\theta \mapsto -\text{KL}(p_{\theta_0}, p_\theta)$, i.e. that the maximum likelihood estimator is consistent.
To derive further asymptotic properties of the MLE we note that, again under some regularity conditions, the MLE $\hat{\theta}_n$ is also the (or a) zero of the score function $s_n$ given by

$$s_n(\theta) = \sum \dot{\ell}_\theta(X_i),$$

where $\dot{\ell}_\theta(x) = \frac{\partial}{\partial \theta}\ell_\theta(x)$. A Taylor expansion of the score function around the true parameter $\theta_0$ then shows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \left( -\frac{1}{n} \dot{s}_n(\hat{\theta}_n) \right)^{-1} \frac{1}{\sqrt{n}} s_n(\theta_0),$$

where $\hat{\theta}_n$ is a (random) point between $\theta_0$ and $\hat{\theta}_n$. If we have consistency then $\hat{\theta}_n$ will converge to $\theta_0$ in $P_{\theta_0}$-probability, so we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \left( -\frac{1}{n} \dot{s}_n(\theta_0) \right)^{-1} \frac{1}{\sqrt{n}} s_n(\theta_0).$$

By the law of large numbers, the central limit theorem and Slutsky’s lemma, we have under regularity conditions that the right-hand side converges in distribution to $i^{-1}\theta_0 N(0, i^{-1}\theta_0)$ as $n \to \infty$, where $i_\theta = \text{Var}_\theta \dot{\ell}_\theta(X_1) = -\mathbb{E}_\theta \dot{\ell}_\theta(X_1)$ is the Fisher information in a single observation. In other words, for large $n$ we have the approximation

$$\hat{\theta}_n \overset{d}{=} N(\theta_0, 1/(ni_\theta)).$$

In particular, in these regular models the asymptotic variance of the MLE equals the Cramér-Rao lower bound $1/(ni_\theta)$ for the variance of an unbiased estimator and the asymptotic bias is typically $o(1/\sqrt{n})$.

This sloppy asymptotic analysis of the MLE can be made precise and generalizes to the situation that the parameter $\theta$ belongs to $\mathbb{R}^k$ for some fixed $k > 1$. As a consequence of the Bernstein-von Mises theorem, Bayesian estimators typically have the same asymptotic behaviour as the MLE in regular models, under mild conditions. (Note that this means in particular that the influence of the prior disappears asymptotically.) Other estimators, such as moment estimators for instance, will usually converge at rate $1/\sqrt{n}$ and be asymptotically normal as well, but will typically have an asymptotic variance that is larger than the Cramér-Rao lower bound.

The rate $1/\sqrt{n}$ arises in the analysis of parametric testing problems as well. Suppose again that we have $n$ i.i.d. observations from a density $p_\theta$ parametrized by a parameter $\theta \in \Theta \subset \mathbb{R}$. Then for two different, fixed points $\theta_0, \theta_1 \in \Theta$, the likelihood ratio test for the hypotheses $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$ is typically consistent. Indeed, consider the test $\varphi_n = 1_{T_n \geq c}$ which rejects the null hypothesis if $T_n \geq c$, where

$$T_n = \sum \log \frac{p_{\theta_1}}{p_{\theta_0}}(X_i)$$

is the log-likelihood ratio statistic and $c \in \mathbb{R}$ is some threshold. By the law of large numbers we have the almost sure convergence

$$\frac{T_n}{n} \to \begin{cases} -\text{KL}(p_{\theta_0}, p_{\theta_1}) & \text{under } H_0, \\ \text{KL}(p_{\theta_1}, p_{\theta_0}) & \text{under } H_1. \end{cases}$$
It follows that for every value of $c \in \mathbb{R}$ the total error of the test satisfies $E_{\theta_0} \varphi_n + E_{\theta_1} (1 - \varphi_n) \to 0$ as $n \to \infty$, provided that the model is identifiable (check!).

Now a natural question to ask is how fast we can let $\theta_1$ tend to $\theta_0$ as $n$ grows, while still retaining consistency of the test, i.e. how far the hypotheses should be separated in order to be statistically distinguishable. Using the same notation as above, a Taylor expansion shows that

$$T_n = \sum (\ell_{\theta_1} - \ell_{\theta_0})(X_i) \approx (\theta_1 - \theta_0) \sum \dot{\ell}_{\theta_0}(X_i).$$

Hence if for some sequence $M_n \to \infty$ we have $\theta_1 - \theta_0 = M_n \sqrt{i_{\theta_0}}$, then by the central limit theorem we have that $T_n$ is approximately $N(0, M_n^2 i_{\theta_0})$-distributed under $H_0$. To understand the behaviour under $H_1$ we write

$$(\theta_1 - \theta_0) \sum \dot{\ell}_{\theta_0}(X_i) = (\theta_1 - \theta_0) \sum \dot{\ell}_{\theta_1}(X_i) + (\theta_1 - \theta_0) \sum (\dot{\ell}_{\theta_0} - \dot{\ell}_{\theta_1})(X_i).$$

Using another Taylor approximation we then get

$$T_n \approx (\theta_1 - \theta_0) \sum \dot{\ell}_{\theta_1}(X_i) - (\theta_1 - \theta_0)^2 \sum \ddot{\ell}_{\theta_1}(X_i).$$

So if $\theta_1 - \theta_0 = M_n \sqrt{i_{\theta_1}}$ again, then by the central limit theorem and the law of large numbers, $T_n$ is approximately $N(M_n^2 i_{\theta_1}, M_n^2 i_{\theta_1})$-distributed under $H_1$. By combining these facts we see that if in the definition of the test we take an appropriate level $c$, we still get a consistent test if the hypotheses are at distance of order larger than $1/\sqrt{n}$. Indeed, in view of the approximations for the distribution of $T_n$ under the respective hypotheses, we have, for $Z$ a standard normal variable,

$$E_{\theta_0} \varphi_n = P_{\theta_0}(T_n \geq c) \approx P\left(Z \geq \frac{c}{M_n \sqrt{i_{\theta_0}}}\right) \to 0$$

if $c/M_n \to \infty$ and

$$E_{\theta_1} (1 - \varphi_n) = P_{\theta_1}(T_n \leq c) \approx P\left(Z \leq \frac{c - M_n^2 i_{\theta_1}}{M_n \sqrt{i_{\theta_1}}}\right) \to 0$$

if $c/M_n - M_n i_{\theta_1} \to -\infty$. Hence is we set for instance $c = M_n \log M_n$, we get a consistent test. If $\theta_1 - \theta_0 = o(1/\sqrt{n})$ however, the test statistic $T_n$ converges to 0 at the same speed under both hypotheses and we loose consistency.

Precise results in the asymptotic theory of parametric estimators and tests, of various degrees of mathematical sophistication, can be found for instance in Ibragimov and Has'minskii (1981), Lehmann and Casella (1998), Van der Vaart (1998), and Lehmann and Romano (2005).

1.2 High-dimensional and nonparametric models

There are many situations in which it is natural, or desirable, to consider statistical models with an unknown parameter that is very high-dimensional compared to sample size, or even infinite-dimensional.
Example 1.2.1 (Nonparametric regression). Suppose we observe pairs \((X_1, Y_1), \ldots, (X_n, Y_n)\) and wish to model the relation between the \(x\) and \(y\)-variables, for instance with the goal to predict the \(y\)-values corresponding to new \(x\)-variables that we will get to observe later.

If we plot the data points in a scatter plot and we see a picture like the left-hand panel of Figure 1.1, then it makes perfect sense to postulate a linear model with just a few parameters, maybe two to describe the linear relation and one more to describe the variance of the fluctuations around the line. The standard linear model postulates in this case that

\[
Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \(\beta_0, \beta_1\) are real parameters and the \(\varepsilon_i\) are independent, \(N(0, \sigma^2)\)-distributed variables. We can then use maximum likelihood, for instance, to estimate the parameter \(\theta = (\beta_0, \beta_1, \sigma^2)\).

If the scatter plot looks like the right-hand panel however, it is less clear what a reasonable choice for a parametric model would be. Alternatively, we can choose to only assume some qualitative properties of the relation between the variables, for instance that the \(x\) and \(y\)-variables are, up to noise, related through a continuous, or differentiable function. We can postulate for instance that

\[
Y_i = f(X_i) + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \(f\) is a differentiable function and the \(\varepsilon_i\) are independent \(N(0, \sigma^2)\) variables. The full parameter of the model is then the pair \(\theta = (f, \sigma)\), which lives in \(\mathcal{F} \times (0, \infty)\), for some function space \(\mathcal{F}\).

Since this model is not described in terms of a small number of real parameters, it is called nonparametric. A model like the one just postulated is called a nonparametric regression model. See for instance Wasserman (2006) for examples of statistical methods for nonparametric regression.
Example 1.2.2 (High-dimensional linear regression). The basic Gaussian linear model considers observations $Y \in \mathbb{R}^n$ that satisfy the linear regression relation $Y = X \theta + \varepsilon$, where $\theta \in \mathbb{R}^p$ is an unknown parameter vector, $X$ is a fixed, known $n \times p$ design matrix, and $\varepsilon \sim N_n(0, \sigma^2 I)$ for some $\sigma^2 > 0$. If $p$ is small relative to $n$ this is a classical parametric model as considered in Section 1.1. If $p$ is fixed for instance and $n \rightarrow \infty$, then we have $\sqrt{n}$-consistency, asymptotic normality, and asymptotic efficiency of the maximum likelihood estimator (see for instance Chapter 8 of Davison (2003)).

There are however also situations in which $p$ is not small compared to $n$. In certain applications in genomics for instance it is not uncommon that $p$ is in the many thousands, while $n$ is only in the hundreds. In such cases we speak of a high-dimensional linear model. A simple special case is the normal means model, which is the case $p = n$ and $X = I$, so that $Y = \theta + \varepsilon$. This arises for instance in image denoising, with $\theta$ representing a noiseless image and $Y$ the noisy observation. Estimating $\theta$ then corresponds to removing the noise from the observed image.

Suppose for instance that we need to denoise a single noisy $200 \times 200$ pixel image. If we postulate the normal means model we then have $p = 40,000$ real parameters to estimate (plus one for the variance of the noise). Clearly this does not fall into the category of parametric problems as considered in Section 1.1. The maximum likelihood estimator for $\theta$ in the model $Y = \theta + \varepsilon$ is simply the data $Y$ itself, which is of course rather useless. This indicates that in models of this type, we need to consider alternative approaches to making inference about unknown parameters. Although high-dimensional models have finitely many parameters, the number is so high relative to sample size that the problem has much more in common with a fully nonparametric problem as considered in the preceding example, than with a classical parametric problems with just a relatively small number of real parameters. For much more on this, see for instance Bühlmann and van de Geer (2011).

Example 1.2.3 (Nonparametric density estimation). Suppose we observe a sample $X_1, \ldots, X_n$ from some unknown distribution. The classical approach to fitting a statistical model to such data, as considered in Section 1.1, is to postulate that the $X_i$ form a sample from some parametric family of densities $\{p_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$ and then to estimate the parameter $\theta$ from the data. Often it is however not clear what a suitable parametric family of densities is and we would prefer to avoid postulating a model that is potentially incorrect. As an example, see the data in Figure 1.2.

In such cases we might view the whole unknown density $p$ of the $X_i$ as parameter, and only assume that it belongs to some broad class $\mathcal{P}$ of densities, for instance the class of all continuous densities, or all densities with a certain degree of smoothness. In this case the whole function $p$ is the parameter of the statistical model and the parameter space $\mathcal{P}$ is an infinite-dimensional function space. The problem of estimating $p$ in such an infinite-dimensional space of densities is called nonparametric density estimation.
In Exercise 1.2 we consider some issues regarding parametric versus non-parametric maximum likelihood estimation for the density of the data in Figure 1.2.

Example 1.2.4 (High-dimensional classification). Consider a binary classification problem in which we observe independent pairs \((X_1, Y_1), \ldots, (X_n, Y_n)\), with \(X_i \in \mathcal{X} \subset \mathbb{R}^k\) and \(Y_i \in \{0, 1\}\). Think for example of a situation in which the \(X_i\) represent images that can have two different labels, 0 or 1. We want to infer from the data the binary regression function \(p: \mathcal{X} \to [0, 1]\) given by

\[
p(x) = P(Y_i = 1 \mid X_i = x).
\]

An estimator \(\hat{p}\) of this function can be used for instance to predict the label of new, unlabelled images. For a given \(x \in \mathcal{X}\) the usual prediction rule is to predict the label to be 1 if \(\hat{p}(x) > 1/2\) and 0 otherwise.

A popular parametric model for this setting is the logistic regression model, which postulates that the binary regression function is of the form

\[
p_\theta(x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}}
\]

for some \(\theta \in \mathbb{R}^k\). Note that for this model we have \(p_\theta(x) > 1/2\) if and only if \(\theta^T x > 0\). Hence, label prediction becomes straightforward once the parameter \(\theta\) has been estimated. If we have an estimator \(\hat{\theta}\) for \(\theta\) obtained from fitting the
logistic regression model to some training data, then a new, unlabelled \( x \in \mathcal{X} \) can be classified by simply considering the sign of \( x^T \hat{\theta} \).

If the \( X_i \) are deterministic, the log-likelihood is easily seen to be given by

\[
\theta \mapsto \sum Y_i \theta^T X_i \log(1 + e^{\theta^T X_i}).
\]

There exists no closed-form expression for the maximizer of this function, but if \( k \) is small relative to \( n \) then the MLE can be efficiently computed numerically. In the high-dimensional situation that \( k \geq n \) however, the MLE typically does not exist. Indeed, the \( j \)th partial derivative of the log-likelihood is given by

\[
\sum Y_i X_{ij} - \frac{X_{ij} e^{\theta^T X_i}}{1 + e^{\theta^T X_i}} = \sum X_{ij} \frac{Y_i(1 + e^{\theta^T X_i}) - e^{\theta^T X_i}}{1 + e^{\theta^T X_i}} = (X^T v_\theta)_j,
\]

where \( X = (X_{ij}) \) is an \( n \times k \) matrix and \( v_\theta \) is the \( n \)-dimensional vector with coordinates

\[
v_{\theta,i} = \frac{Y_i(1 + e^{\theta^T X_i}) - e^{\theta^T X_i}}{1 + e^{\theta^T X_i}}.
\]

If \( \theta \) is a stationary point of the log-likelihood, then \( X^T v_\theta = 0 \). But if \( k \geq n \), then \( X^T \) typically has full rank \( n \). If that is the case, then for a stationary point \( \theta \) it must hold that \( v_\theta = 0 \), i.e. that

\[
Y_i(1 + e^{\theta^T X_i}) = e^{\theta^T X_i}
\]

for every \( i \). Clearly this is impossible.

So also in high-dimensional logistic regression the usual parametric approach breaks down and alternative ideas are necessary. See Exercise 1.3 for a concrete example.

The examples we have considered in this section have in common that they involve statistical models with parameters that are very high-dimensional compared to sample size, or even infinite-dimensional. Naively using methods like maximum likelihood is typically either impossible or useless in these cases, so we will have to develop alternative ideas to tackle such problems.

We will see that in high-dimensional or nonparametric models we usually have completely different behaviour of statistical procedures. Convergence rates are typically slower than \( \sqrt{n} \), asymptotic normality is not guaranteed, and optimality of procedures can not be assessed in terms of minimal variance. Moreover, we will need to use other technical machinery for the mathematical analysis, since the ‘Taylor’s formula + Central Limit Theorem’ approach will not get us very far.
1.3 Exercises

**Exercise 1.1** (Kullback-Leibler divergence). Let \( f, g \) be positive probability densities with respect to some measure \( \mu \). Show that \( \text{KL}(f, g) \geq 0 \) and that \( \text{KL}(f, g) = 0 \) if and only if \( f = g, \mu \)-almost everywhere. (Hint: lower bound the KL-divergence by the squared Hellinger distance, i.e. the squared \( L^2 \)-distance between the square roots of the densities.)

**Exercise 1.2** (Parametric versus nonparametric mixture models in density estimation). A distribution on the line is called a (discrete) mixture of normals if its density \( f \) is of the form

\[
f(x) = \sum_{k=1}^{K} \frac{1}{\sigma_k} \varphi \left( \frac{x - \mu_k}{\sigma_k} \right),
\]

where \( \varphi \) is the standard normal density, \( K \in \mathbb{N} \) is the number of mixture components, \( \mu_1, \ldots, \mu_K \in \mathbb{R} \) are the locations of the components, \( \sigma_1, \ldots, \sigma_K > 0 \) are the standard deviations, and \( p_1, \ldots, p_K > 0 \) are the component weights. Clearly, we should have \( \sum p_k = 1 \) to ensure that \( f \) is a probability density.

(i) If we fix the number of mixture components at some level \( K \), a normal mixture model is just a parametric model with \( 3K - 1 \) real parameters. Computing the maximum likelihood estimator for these parameters is not entirely straightforward however. Closed-form expressions can not be derived and one has to resort to numerical methods. An effective approach in this setting is to introduce, in addition to the datapoints, a layer of unobserved auxiliary variables that describe the mixture components to which the datapoints belong. The EM algorithm can then be used to iteratively compute the MLE for the parameters. See for instance Section 5.5.2. of Davison (2003) for details.

In R this estimation procedure has been implemented for instance in the function `normalmixEM` in the package `mixtools`. Use this function to fit normal mixture models with \( K = 2, 3, 4 \) and \( 5 \) components to the galaxies dataset from Figure 1.2, which can be found in the R package `MASS`. Report the estimates for the weights, the locations, and variances. Plot the weighted densities of the components and the estimated overall density over the histogram of Figure 1.2.

(ii) We can also leave the number of components unspecified, leaving it as a parameter that is to be estimated from the data. This turns the model into an infinite-dimensional model.

Prove that the MLE does not exist in this case. More precisely, let \( \mathcal{F} \) be the class of all mixture densities of the form (1.1) and prove that if
$X_1, \ldots, X_n$ is a sample from any $f_0 \in \mathcal{F}$, then

$$\max_{f \in \mathcal{F}} \prod_{i=1}^{n} f(X_i) = \infty$$

almost surely.

**Exercise 1.3** (High-dimensional versus low-dimensional classification). The famous MNIST dataset contains a large number of $28 \times 28$ pixel, grayscale images of handwritten digits. These images are labeled, i.e. the true number that they depict is in the dataset as well. We have extracted from the MNIST database a file containing 2215 labeled images of 0’s and 1’s. From this file, 100 labeled images were selected at random, see Figure 1.3. In this exercise we explore the possibility of fitting a logistic regression model to this training dataset, with the goal of predicting the labels of the other 2115 images.

(i) The training data are available in the file `traindata01.csv` on the webpage. This file contains 100 lines, each corresponding to a labeled image. A line contains numeric values separated by commas. The first value on a line is the label of the image, which is a 0 or a 1. The other $28 \times 28 = 784$ values are integers between 0 and 255, which are the gray-scale values of the pixels in the image. The value 0 corresponds to white, the value 255 corresponds to black.

We can treat the data as independent pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ as in Example 1.2.4, with $n = 100$ and for every $i$, $X_i$ a 784-dimensional vector.

Show that the matrix $X = (X_{ij})_{i=1,\ldots,100;j=1,\ldots,784}$, as considered in Example 1.2.4, has rank 100 in this case. Hence, we are in the high-dimensional setting in which the MLE does not exist.

Nevertheless, the R function `glm` can still be used to fit a logistic regression model to the data. This does not produce an error message that the rank of $X$ is too low. Instead, R basically removes columns from $X$ that are linear combinations of other columns. For those columns, it produces an `NA` as estimate of the corresponding coordinate of the parameter vector $\theta$.

Fit the logistic regression model using `glm`. Determine which columns of $X$ were retained by R, i.e. which coordinates of the estimate for $\theta$ contain non-`NA`’s. How many such columns are there?

Let $\tilde{X}$ and $\tilde{\theta}$ be the matrix $X$ and the estimated vector $\theta$ from which the columns and coordinates corresponding to the `NA`’s have been removed. Show that by considering the sign of $\tilde{X}\tilde{\theta}$, we obtain predictions for the labels of the training images that are correct in all 100 cases.
(ii) Instead of using the complete images as explanatory vectors in the logistic regression model we can try to come up with reasonable low-dimensional feature vectors that also have good predictive power, so that we can fit an ordinary, low-dimensional logistic regression model and avoid the issues arising in the high-dimensional setting.

Show that if instead of with the complete training images we just work with the $2 \times 2$ pixel sub-images at the center of each image, we can fit a logistic regression model, do not obtain NA’s, and also predict all 100 training image labels correctly.

(iii) Parts (i) and (ii) give rise to two different predictors for labels of images which take a $28 \times 28$ image as input and produce a label in $\{0,1\}$ as output.

Compare the prediction performance of these two methods on the whole collection of 2215 images, available in the file `alldata01.csv`.

Can you explain why there is a difference in the generalization performance, i.e. in the quality of the label predictions for images of which the label has not been observed?
Chapter 2

The signal-in-white-noise model

2.1 Definition of the model

To explore some of the issues that arise in many high-dimensional or non-parametric statistical models it is useful to have a ‘canonical’ model which on the one hand really exhibits these nonparametric features in a non-trivial way, and on the other hand is tractable enough to allow for a detailed mathematical analysis. In these notes this role is played by the so-called signal-in-(Gaussian)-white-noise model.

In this model it is assumed that we observe a sample path \((X_t : t \in [0, 1])\) of a stochastic process \(X\) that satisfies the stochastic differential equation (SDE)

\[
dX_t = f(t) \, dt + \frac{1}{\sqrt{n}} \, dW_t, \quad X_0 = 0.
\]

Here \(W\) is a standard Brownian motion (we recall the definition in the next section) and \(f\) is an unknown square integrable function on \([0, 1]\). The SDE is just short-hand notation for the corresponding integral form of the equation:

\[
X_t = \int_0^t f(s) \, ds + \frac{1}{\sqrt{n}} W_t, \quad t \in [0, 1].
\]   (2.1)

So indeed, the model postulates that the observed data is a deterministic signal, corrupted with additive Gaussian noise. The noise is called ‘white’ because the increments of the Brownian motion are uncorrelated.

The unknown parameter in the white noise model is the function \(f \in L^2[0, 1]\), so it is a genuinely nonparametric model. The natural number \(n\) quantifies the signal-to-noise ratio. It plays the role of sample size in this model, in the sense that it should become easier to recover \(f\) as \(n\) gets larger. We will be interested in the asymptotic behaviour of statistical procedures as \(n \to \infty\).

Figure 2.1 shows a simulated data example. The true underlying signal \(f\) is the function on the left. The right-hand panel shows a realization of the process \(X\) given by (2.1). The statistical problem is to recover the function \(f\) from this noisy observation of its primitive function.
Figure 2.1: From left to right: a function $f$, its primitive $F(t) = \int_0^t f(s) \, ds$ and the noisy version $F + n^{-1/2}W$.

Clearly the signal-in-white-noise model is somewhat academic. It assumes for instance that we observe a continuous sample path, which is of course not possible in reality. Also, the assumption of purely Gaussian noise can be restrictive in realistic applied settings. Nevertheless, it is a very useful model to study. It allows a detailed analysis that gives insight into some important general phenomena that occur in some form in all high-dimensional or nonparametric models. It can in fact be shown that the model is in some sense very close to some of the more realistic models considered in Chapter 1, as we will briefly discuss in Section 2.5.

2.2 Brownian motion and Wiener integrals

We recall some basic facts about Brownian motion that we need to analyze the signal-in-white-noise model.

**Definition 2.2.1.** A (standard) Brownian motion is a stochastic process $W = (W_t : t \geq 0)$, defined on some probability space $(\Omega, \mathcal{F}, P)$, such that

(i) $W_0 = 0$;

(ii) For all $t \geq s$, the increment $W_t - W_s$ is independent of $(W_u : u \leq s)$;

(iii) For all $t \geq s$, we have $W_t - W_s \sim N(0, t - s)$;

(iv) Almost all sample paths $t \mapsto W_t$ are continuous functions.

Items (ii) and (iii) show that the Brownian motion is a process with independent and stationary increments. It is also a Gaussian process, meaning that for every finite number of time points $t_1, \ldots, t_n$, the vector $(W_{t_1}, \ldots, W_{t_n})$
has a multivariate normal distribution. Indeed, such a vector has a \( N(0, \Sigma) \)-distribution, where \( \Sigma \) is the covariance matrix with entries \( \Sigma_{ij} = t_i \wedge t_j \). It is not at all immediately clear that the Brownian motion process exists, this is a non-trivial theorem. See for instance Karatzas and Shreve (1991), Revuz and Yor (1991), or Mörters and Peres (2010) for a proof of this theorem and for many of the interesting properties of Brownian motion.

One of the properties of Brownian motion is that its sample paths \( t \mapsto W_t \) are very rough. They are nowhere differentiable functions that are not of bounded variation on finite intervals. This implies that integrals of the form \( \int f \, dW \) can not be defined simply \( \omega \)-by-\( \omega \) as Stieltjes integrals. There exists however a natural alternative way of defining integrals of deterministic functions in \( L^2[0, \infty) \) with respect to Brownian motion. Such integrals are called Wiener integrals.

For concreteness, let \((\Omega, \mathcal{F}, P)\) be the underlying probability space. For an indicator function \( f \) of the form \( f = 1_{(s,t]} \) for some \( t > s \geq 0 \) we simply define

\[
I(f) = W_t - W_s.
\]

We extend this definition by linearity to simple functions of the form \( f = \sum a_i 1_{(s_i,t_i]} \). It is straightforward to verify that this is a proper definition and that such integrals of simple functions are centered Gaussian random variables that satisfy the isometry relation

\[
E I(f)I(g) = \int_0^\infty f(t)g(t) \, dt
\]

(see Exercise 2.1). Since the simple functions are dense in \( L^2[0, \infty) \), this allows us to extend the definition of the integrals to all of \( L^2[0, \infty) \) (see Exercise 2.1 again). We end up with a linear map \( f \mapsto I(f) \) from \( L^2(\Omega, \mathcal{F}, P) \) that satisfies the isometry relation for all \( f \) and \( g \) and such that \( I(f) \sim N(0, \int f^2(t) \, dt) \). We note the ‘integral’ \( I(f) \) is defined as an \( L^2 \)-limit, not \( \omega \)-by-\( \omega \) on the underlying probability space. In particular, it is only unique almost surely.

The following theorem summarizes the properties of the Wiener integral.

**Theorem 2.2.2.** Let \( W \) be a Brownian motion on \((\Omega, \mathcal{F}, P)\). There exists a map \( I : L^2[0, \infty) \to L^2(\mathcal{P}) \) such that

(i) If \( f = \sum a_i 1_{[t_{i-1}, t_i)} \) for \( a_0, \ldots, a_n \in \mathbb{R} \) and \( 0 \leq t_0 \leq \cdots \leq t_n \), then

\[
I(f) = \sum a_i (W_{t_i} - W_{t_{i-1}});
\]

(ii) The map \( I \) is linear;

(iii) For all \( f \in L^2[0, \infty) \), the random variable \( I(f) \) is centered and Gaussian;
(iv) For all \( f, g \in L^2[0, \infty) \),
\[
E I(f)I(g) = \int f(t)g(t) \, dt. \tag{2.3}
\]

For \( f \in L^2[0, \infty) \) and \( t \geq 0 \) we write \( \int f \, dW \) for \( I(f) \) and we use the notation \( \int_0^t f(s) \, dX_s = \int_0^t f(s) \, dA_s + c \int_0^t f(s) \, dW_s. \)

We note that the integral notation is justifiable, in the sense that this stochastic integral has many properties similar to usual integrals. In particular, in view of the definition of the Wiener integrals and the well-known properties of Lebesgue integrals, the integral \( \int_0^t f(s) \, dX_s \) can be obtained as an \( L^2 \)-limit of Riemann-Stieltjes-type sums. Moreover, of course, the map \( f \mapsto \int_0^t f(s) \, dX_s \) is linear.

\section{2.3 Sequence formulation of the model}

In the signal-in-white-noise model (2.1) we assume that we observe a noisy version of the primitive of a function \( f \in L^2[0, 1] \). It turns out that equivalently, we can assume that we observe an infinite sequence of noisy Fourier coefficients of \( f \), relative to an arbitrary orthonormal basis of \( L^2[0, 1] \).

The space \( L^2[0, 1] \) of (equivalence classes of) square integrable functions on the unit interval is a separable Hilbert space with inner product
\[
\langle f, g \rangle = \int_0^1 f(s)g(s) \, ds
\]
and corresponding norm \( \| \cdot \|_2 \) given by \( \| f \|_2^2 = \langle f, f \rangle \). Let \( e_1, e_2, \ldots \) be an orthonormal basis of the space, for instance the classical Fourier basis. Then every \( f \in L^2[0, 1] \) can be expanded as \( f = \sum f_j e_j \), with convergence in \( L^2 \)-sense, where \( f_j = \langle f, e_j \rangle \). Slightly abusing terminology we call the \( f_j \) the Fourier coefficients of \( f \) relative to the basis \( e_j \). By the Plancherel relation the Fourier coefficients form a sequence in \( \ell^2 \), with squared \( \ell^2 \)-norm given by \( \sum f_j^2 = \| f \|_2^2 \). (We use the usual notation \( \ell^2 = \{(a_1, a_2, \ldots) \in \mathbb{R}^\infty : \sum a_i^2 < \infty \} \) for the Hilbert space of square summable sequences.)

If we have a process \( X \) satisfying (2.1) we can define the random variables
\[
Y_j = \int_0^1 e_j(s) \, dX_s, \quad j = 1, 2, \ldots,
\]
where the stochastic integral is defined as in the preceding section. If we then set \( Z_j = \int_0^1 e_j(s) \, dW_s \) and let \( f_j \) be the Fourier coefficients of \( f \) relative to the basis \( e_j \), we obtain

\[
Y_j = f_j + \frac{1}{\sqrt{n}} Z_j, \quad j = 1, 2, \ldots
\]  
(2.4)

Moreover, the properties of the Wiener integral and the fact that the \( e_j \) form an orthonormal basis imply that the \( Z_j \) form a sequence of independent, standard normal random variables.

Conversely, suppose that (2.4) holds for independent standard normal variables \( Z_j \) and a sequence \( f_j \) in \( \ell^2 \). Let \( e_j \) be an arbitrary orthonormal basis of \( L^2[0, 1] \). Then it can be shown that \( W_t = \sum Z_j \int_0^t e_j(s) \, ds \) defines a Brownian motion (see Exercise 2.2). Since \( \sum f_j \int_0^t e_j(s) \, ds = \int_0^t f(s) \, ds \), for \( f = \sum f_j e_j \), we then have that \( X_t = \sum Y_j \int_0^t e_j(s) \, ds \) is a well-defined stochastic process that satisfies (2.1).

We conclude that the statistical problem of inferring the function \( f \in L^2[0, 1] \) from observations \( (X_t : t \in [0, 1]) \) satisfying (2.1) is equivalent to the problem of inferring the sequence \( (f_j) \in \ell^2 \) from the observations \( Y_j \) satisfying (2.4).

### 2.4 Expressions for the likelihood

A stochastic process \( X = (X_t : t \in [0, 1]) \) satisfying (2.1) has continuous sample paths. Hence, it can be seen as a random element in the space \( C[0, 1] \) of continuous functions on \( [0, 1] \). More precisely, we view \( C[0, 1] \) as a Banach space with the uniform norm \( \|f\|_\infty = \sup_{t \in [0, 1]} |f(t)| \) and we endow it with the corresponding Borel \( \sigma \)-algebra \( \mathcal{B}(C[0, 1]) \). Then if \((\Omega, \mathcal{F}, P)\) is the probability space on which \( X \) is defined, the process \( X \) defines a measurable map

\[
(\Omega, \mathcal{F}) \rightarrow (C[0, 1], \mathcal{B}(C[0, 1]))
\]

\[
\omega \mapsto (t \mapsto X_t(\omega)).
\]

The **distribution**, or **law** of the process \( X \) is the image measure of the probability measure \( P \) under this map. For the process \( X \) defined by (2.1), we denote this distribution by \( P_{f,n} \). Concretely, for a Borel set \( B \subset C[0, 1] \) we have

\[
P_{f,n}(B) = P(X \in B),
\]

where \( X \) is the process given by (2.1). In particular, the measure \( P_{0,n} \) is the distribution of \( W/\sqrt{n} \), for \( W \) the standard Brownian motion. The measure \( P_{0,1} \), i.e. the distribution of Brownian motion, is commonly called the **Wiener measure**. When there is no risk of confusion we omit the index \( n \) and simply write \( P_f \) instead of \( P_{f,n} \).

The following theorem, which is a consequence of the more general **Girsanov theorem**, asserts that for every \( n \), all the laws \( P_{f,n}, f \in L^2[0, 1] \), are equivalent,
i.e. have the same null sets. Moreover, it gives an expression for the Radon-Nikodym derivatives, or likelihood ratios $dP_{f,n}/dP_{f_0,n}$. These derivatives are measurable maps on $C[0,1]$ and hence can be viewed as random variables defined on the probability space $(C[0,1], B(C[0,1]), P_{g,n})$ for every $g \in L^2[0,1]$. The theorem describes the distribution of the likelihoods when viewed in this manner.

**Theorem 2.4.1.** For every $n \in \mathbb{N}$ the laws $P_{f,n}$, $f \in L^2[0,1]$, are all equivalent measures on $C[0,1]$. Moreover,

$$
\frac{dP_{f,n}}{dP_{f_0,n}} = \exp \left( \frac{1}{\sqrt{n}} \int_0^1 (f(t) - f_0(t)) \, dW_t - \frac{1}{2} n \|f - f_0\|^2_2 \right),
$$

$P_{f_0,n}$-almost surely, where $W$ is a $P_{f_0,n}$-Brownian motion.

*Proof. The result is a straightforward consequence of Girsanov’s theorem. See for instance Section 3.5 of Karatzas and Shreve (1991). We show that for every pair $f, f_0 \in L^2[0,1]$ we have $P_{f,n} \ll P_{f_0,n}$ and that the expression for the Radon-Nikodym derivative given in the statement of the theorem holds. By reversing the roles of $f$ and $f_0$ we then obtain the statement of equivalence.

Let $X$ be the coordinate process on $C[0,1]$, i.e. $X_t(\omega) = \omega(t)$ for all $\omega \in C[0,1]$ and $t \in [0,1]$. Then the process $W$ defined by

$$
dX_t = f_0(t) \, dt + \frac{1}{\sqrt{n}} \, dW_t
$$

is a $P_{f_0,n}$-Brownian motion. The process $Z$ defined by

$$
Z_t = \exp \left( \frac{1}{\sqrt{n}} \int_0^t (f(s) - f_0(s)) \, dW_s - \frac{1}{2} n \int_0^t (f(s) - f_0(s))^2 \, ds \right)
$$

is a $P_{f_0,n}$-martingale (check!). Hence, by Girsanov’s theorem, the process $\tilde{W}$ defined by

$$
d\tilde{W}_t = dW_t - \sqrt{n} (f(t) - f_0(t)) \, dt
$$

is a Brownian motion under the new measure $P$ defined by $dP = Z_1 \, dP_{f_0,n}$. Observe that $X$ satisfies $dX = f \, dt + n^{-1/2} d\tilde{W}$. Hence, since $\tilde{W}$ is a $P$-Brownian motion, we have that $P = P_{f,n}$ (check!). This shows that $P_{f,n} \ll P_{f_0,n}$ and that $dP_{f,n}/dP_{f_0,n} = Z_1$, almost surely. 

If $X$ is the coordinate process on $C[0,1]$, then, by construction, under $P_{f_0}$ it is a process which satisfies $(2.1)$ with $f_0$ instead of $f$. Hence, another way to formulate the statement of this theorem is to say that for $p_f$ the density of
the law \( P_f \) with respect to the law \( P_0 \) of the normalized Brownian motion and \( X \) a process satisfying (2.1) with \( f_0 \) instead of \( f \), the densities a.s. satisfy

\[
\frac{P_f}{P_{f_0}}(X) = \exp \left( \sqrt{n} \int_0^1 (f(t) - f_0(t)) \, dW_t - \frac{1}{2} n \| f - f_0 \|_2^2 \right),
\]

(2.6)

where \( W \) is a \( P_{f_0} \)-Brownian motion. By setting \( f_0 = 0 \) and recalling (2.5) we find that if \( X \) satisfies (2.1), then for the density \( p_f \) we have

\[
p_f(X) = \exp \left( \int_0^1 f(t) \, dX_t - \frac{1}{2} \int_0^1 f^2(t) \, dt \right)
\]

(2.7)

a.s. (check!). Hence if we consider a statistical model where we observe a process \( X \) that satisfies (2.1) for some unknown signal \( f \) in a collection \( F \), then the MLE for \( f \), if it exists, is the maximizer of the likelihood (2.7) over \( F \). See Exercise 2.3 for a simple example.

2.5 Asymptotic equivalence of nonparametric models

We have already seen that the signal-in-white-noise model is equivalent to the sequence model (2.4). The latter is obviously very similar to the normal means model considered in Example 1.2.2.

In fact, the model is asymptotically close to a number of other nonparametric models as well. This can be made very precise in the context of Le Cam’s theory of limits of experiments. See for instance the papers Brown and Low (1996) and Nussbaum (1996) on the equivalence of the signal in white model to nonparametric regression and density estimation. These results are very technical and we will not go into the details here. We just indicate very briefly that there are indeed intimate connections between the models, in order to motivate the fact that the signal-in-white-noise model serves as a ‘benchmark model’ in this text. The idea is simply that two statistical models are (asymptotically) very similar if the corresponding likelihoods are (asymptotically) very similar.

Example 2.5.1 (Regression). Let \( Y_1, \ldots, Y_n \) be observations satisfying

\[
Y_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n,
\]

(2.8)

where \( f : [0, 1] \to \mathbb{R} \) is a continuous function and \( \varepsilon_1, \ldots, \varepsilon_n \) are i.i.d. random variables with mean 0 and variance 1. Observing the \( Y_i \) is clearly equivalent to observing the process \( X^{(n)} \) defined by

\[
X^{(n)}_t = \frac{1}{n} \sum_{i \leq \lfloor nt \rfloor} Y_i, \quad t \in [0, 1].
\]

This process can be decomposed as

\[
X^{(n)}_t = \frac{1}{n} \sum_{i \leq \lfloor nt \rfloor} f(i/n) + \frac{1}{n} \sum_{i \leq \lfloor nt \rfloor} \varepsilon_i.
\]
The first term is a Riemann sum that converges uniformly in $t$ to the integral $\int_0^t f(s) \, ds$ if $f$ is for instance Hölder continuous with some exponent $\alpha > 0$. By Donsker’s theorem we have the uniform weak convergence

$$\frac{1}{\sqrt{n}} \sum_{i \leq \lfloor n \cdot \epsilon \rfloor} \varepsilon_i \Rightarrow W,$$

where $W$ is a Brownian motion on $[0,1]$. We conclude that for large $n$, the statistical problem of inferring the regression function $f$ in the regression context (2.8) is very similar to inferring $f$ in the white noise model (2.1).

For a precise result in the context of Le Cam’s theory of limits of experiments, see for instance Brown and Low (1996) or Giné and Nickl (2016). The main assumption under which this result is derived is that $f$ is Hölder continuous with some exponent $\alpha > 1/2$.

*Example 2.5.2 (Density estimation). Let $X_1, \ldots, X_n$ be a sample from a density $f$ on $[0,1]$. If $F$ is the corresponding distribution function, then we can represent the observations as $X_i = F^{-1}(U_i)$, where $U_1, \ldots, U_n$ are independent, uniform variables on $(0,1)$. Hence, the likelihood ratio can be written as

$$\prod \frac{f}{f_0}(X_i) = e^{\sum \log \frac{f}{f_0}(F^{-1}_0(U_i))},$$

(2.9)

with the $U_i$ i.i.d. and uniform $[0,1]$ under the true distribution $P_{f_0}$. Let $\mathbb{G}_n(t) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n 1_{(0,t]}(U_i) - t \right)$ be the empirical process of the $U_i$ (the normalised empirical distribution function). Then for an arbitrary measurable function $h$ we have

$$\sqrt{n} \int h \, d\mathbb{G}_n = \sum_{i=1}^n h(U_i) - n \int h(t) \, dt.$$

Hence, setting $h = \log \frac{f}{f_0}(F^{-1}_0(\cdot))$ we can rewrite the likelihood ratio as

$$e^{\sqrt{n} \int h \, d\mathbb{G}_n} + n \int h(t) \, dt.$$

Observe that

$$\int h(t) \, dt = \int f_0(t) \log \frac{f}{f_0}(t) \, dt = -\text{KL}(f_0, f).$$

Moreover, by Donsker’s theorem the empirical process $\mathbb{G}_n$ converges weakly (for instance in the Skorohod space $D[0,1]$) to a standard Brownian bridge $B$. Hence, by the continuous mapping theorem, we have the weak convergence

$$\sqrt{n} \int h \, d\mathbb{G}_n \Rightarrow \sqrt{n} \int h(t) \, dB_t.$$
Since $B_t$ can be represented as $B_t = W_t - tW_1$ for a Brownian motion $W$, the right-hand side can be written as
\[
\sqrt{n} \int h(t) d(W_t - tW_1) = \sqrt{n} \int h(t) dW_t - W_1 \int h(t) dt
= \sqrt{n} \int_0^1 (h(t) - \text{KL}(f_0, f)) dW_t.
\]
So asymptotically the likelihood ratio looks like a constant times
\[\exp(\sqrt{n} \int_0^1 (h(t) + \text{KL}(f_0, f)) dW_t).\]
After normalizing we obtain
\[\exp \left( \sqrt{n} \int_0^1 (h(t) + \text{KL}(f_0, f)) dW_t - \frac{1}{2n} \int_0^1 (h(t) + \text{KL}(f_0, f))^2 dt \right)\]
as approximation to (2.9). By Theorem 2.4.1 this is the likelihood ratio for the model in which we have observations $(X_t : t \in [0, 1])$ satisfying
\[dX_t = \left( \log \frac{f}{f_0}(F^{-1}_0(t)) + \text{KL}(f_0, f) \right) dt + \frac{1}{\sqrt{n}} dW_t. \tag{2.10}\]
The derivations above indicate that the statistical problem of inferring the density $f$ from an i.i.d. sample is, for large $n$, similar to the problem of inferring $f$ from the observation of a stochastic process satisfying (2.10).

If we know in advance that $f$ is close to $f_0$ we can obtain further approximations. Indeed, in that case $\text{KL}(f_0, f)$ is negligible and
\[h(t) = 2 \log \sqrt{\frac{f}{f_0}(F^{-1}_0(t))} \approx 2 \left( \sqrt{\frac{f}{f_0}(F^{-1}_0(t))} - 1 \right).
\]
It follows that the stochastic integral in the likelihood is then approximately equal to
\[2 \int_0^1 \left( \sqrt{\frac{f}{f_0}(F^{-1}_0(t))} - 1 \right) dW_t = 2 \int_0^1 \left( \sqrt{f(t)} - \sqrt{f_0(t)} \right) dZ_t,
\]
where $Z$ is given by
\[Z_t = \int_0^t \frac{1}{\sqrt{f_0(s)}} dW_{f_0(s)} = \int_0^{F_0(t)} \frac{1}{\sqrt{f_0(F_0^{-1}(s))}} dW_s.
\]
The process $Z$ is a continuous martingale with quadratic variation
\[\langle Z \rangle_t = \int_0^{F_0(t)} \frac{1}{f_0(F_0^{-1}(s))} dF_0(s) = \int_0^t \frac{1}{f_0(s)} dF_0(s) = t,
\]
which implies that $Z$ is in fact a Brownian motion. So for $f$ close to $f_0$, the likelihood ratio is well approximated by a constant times $\exp(2 \int_0^1 (\sqrt{f(t)} -
\( \sqrt{f_0(t)} \, dW_t \), where \( W \) is a Brownian motion. By Theorem 2.4.1 again, this is the likelihood ratio for the statistical problem of inferring \( f \) from the observation of a process \( X \) satisfying the SDE

\[
dX_t = \sqrt{f(t)} \, dt + \frac{1}{2\sqrt{n}} \, dW_t.
\] (2.11)

With (considerably) more mathematical work the local asymptotic equivalence of the statistical problems of i.i.d. density estimation and the signal-in-white-noise model (2.11) can be proved rigorously. Moreover, the local result can be extended to a global result. The main assumptions that are needed to derive the result are that the densities are Hölder continuous of some order \( \alpha > 1/2 \) and that they are bounded away from 0 on \([0, 1]\). See Nussbaum (1996) for details.

Exercise 2.4 briefly explores one more example of asymptotic equivalence, relating drift estimation for a certain class of SDE’s to estimating a signal in Gaussian white noise.
2.6 Exercises

Exercise 2.1 (Isometry relation for Wiener integrals). Show that the relation (2.2) holds for simple functions $f$ and $g$ of the form $f = \sum a_i 1_{(t_{i-1}, t_i]}$, $g = \sum b_j 1_{(s_{j-1}, s_j]}$. Next, show that these simple functions are dense in $L^2[0, \infty)$ and use this to prove that (2.2) holds for all $f, g \in L^2[0, \infty)$.

Exercise 2.2 (Series expansions of Brownian motion). Let $e_j$ be the standard Fourier basis. Show that if $Z_j$ are independent standard normal variables, then the process $W = (W_t : t \in [0, 1])$ given by

$$W_t = \sum Z_j \int_0^t e_j(s) \, ds, \quad t \in [0, 1],$$

is well defined and satisfies items (i)–(iii) of Definition 2.2.1. (It can be shown that $W$ satisfies the continuity (iv) of the definition as well and that this is true for any orthonormal basis $e_j$ of $L^2[0, 1]$, but this is outside the scope of this lecture.)

Exercise 2.3 (Parametric signal-in-white-noise). Suppose we observe a stochastic process $X = (X_t : t \in [0, 1])$ satisfying

$$dX_t = \theta \, dt + \frac{1}{\sqrt{n}} \, dW_t,$$

where $W$ is a Brownian motion and $\theta \in \mathbb{R}$ is an unknown parameter. Give an expression for the MLE for $\theta$ and derive its asymptotic behaviour.

Exercise 2.4 (Local equivalence of periodic drift estimation and signal-in-white-noise). Suppose that we observe a process $X = (X_t : t \in [0, T])$ satisfying the SDE

$$dX_t = b(X_t) \, dt + dW_t,$$

where $b$ is belongs to the space $C^1_o$ of 1-periodic, continuously differentiable functions. The statistical goal is to estimate the function $b$. For such a process we have a law of large numbers which asserts that if $f \in C^1_o$, then as $T \to \infty$

$$\frac{1}{T} \int_0^T f(X_t) \, dt \to \int_0^1 f(x) \rho(x) \, dx,$$

almost surely, where $\rho$ is the probability density on $[0, 1]$ given by

$$\rho(x) = Ce^{2\int_0^x b(y) \, dy},$$
where $C$ is the appropriate normalizing constant. Moreover, we have a central limit theorem which asserts that for $f \in C^1_0$

$$
\frac{1}{\sqrt{T}} \int_0^T f(X_t) \, dW_t \xrightarrow{d} N \left(0, \int_0^1 f^2(x) \rho(x) \, dx \right)
$$
as $T \to \infty$.

Let $P_{b,T}$ be the law that the process $X$ induces on the space $C[0,T]$ of continuous functions on $[0,T]$. By a slight extension of Theorem 2.4.1, the likelihood ratio $dP_{b,T}/dP_{b_0,T}$ for two functions $b, b_0 \in C^1_0$ is given by

$$
dP_{b,T}/dP_{b_0,T} = \exp \left( \int_0^T (b - b_0)(X_t) \, dW_t - \frac{1}{2} \int_0^T (b - b_0)^2(X_t) \, dt \right),
$$
where $W$ is a $P_{b_0,T}$-Brownian motion. Use this in combination with the LLN and CLT given above to convince yourself that if we know in advance that $b$ is close to another function $b_0$ in $C^1_0$, then for $T \to \infty$, the statistical problem under consideration is close to a signal-in-white-noise problem. Which one exactly?
Chapter 3

Stein’s phenomenon, regularization

3.1 The James-Stein theorem

Consider the normal means model where we observe $Y \sim N_n(\theta, \sigma^2 I)$, for some $n \in \mathbb{N}$ and $\sigma > 0$, and the aim is to estimate the mean $\theta$ (see Example 1.2.2). In the case $n = 1$ it seems rather clear that if we do not know anything about the parameter $\theta$, we can not do much better than estimating it by the observation $Y$. Proving this rigorously is actually not completely trivial, see Exercise 3.5. For larger $n$ it is in fact also not immediately clear whether if we assume no further structure on $\theta$, we can do better than simply using the maximum likelihood estimator for $\hat{\theta}_{\text{MLE}} = Y$. It turns out however that it is possible to perform strictly better, in the sense of expected quadratic error.

To get a first indication of this fact, note that for any estimator $\hat{\theta}$ with a finite covariance we have the bias-variance decomposition

$$E_\theta \|\hat{\theta} - \theta\|^2 = \|E_\theta \hat{\theta} - \theta\|^2 + \text{tr} \text{Cov}_\theta \hat{\theta}$$

(check!). Here $\| \cdot \|$ is the Euclidean norm, $\text{Cov}X = E(X - E(X)(X - E(X)^T)$ is the covariance matrix of a random vector $X$ and $\text{tr} A$ is the trace of the matrix $A$. If we apply this to $\hat{\theta}_c = cY$ we find that $E_\theta \|cY - \theta\|^2 = (c-1)^2 \|\theta\|^2 + c^2 \sigma^2 n$, which, for given $\theta$, is minimal for $c$ equal to

$$c_\theta = \frac{\|\theta\|^2}{\|\theta\|^2 + \sigma^2 n},$$

and the minimal value is

$$E_\theta \|\hat{\theta}_{c_\theta} - \theta\|^2 = \frac{\sigma^2 n \|\theta\|^2}{\|\theta\|^2 + \sigma^2 n} = \frac{\|\theta\|^2}{\|\theta\|^2 + \sigma^2 n} E_\theta \|\hat{\theta}_{\text{MLE}} - \theta\|^2.$$  (3.1)

Since $c_\theta < 1$, this indicates that it might be advantageous to shrink the estimator $Y$ towards 0, that is, to multiply it by a factor strictly smaller than 1.
Multiplying by $c_\theta$ does not lead to a valid estimator however, since $c_\theta$ depends on the unknown parameter $\theta$. It turns out that for $n \geq 3$, shrinking by an appropriate data-dependent constant leads to an estimator with an expected squared error that is strictly smaller than that of the MLE.

**Theorem 3.1.1** (James-Stein). Define

$$\hat{\theta}_{JS} = \left(1 - \frac{\sigma^2(n-2)}{\|Y\|_2^2}\right)Y.$$ 

For $n \geq 3$, we have $E_\theta \|\hat{\theta}_{JS} - \theta\|^2 < E_\theta \|\hat{\theta}_{MLE} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$.

**Proof.** For the bias and variance of the $i$th component of the JS estimator we have

$$E_\theta \hat{\theta}_{JS,i} - \theta_i = -\sigma^2(n-2)E_\theta \frac{Y_i}{\|Y\|_2^2}$$

and

$$\text{Var}_\theta \hat{\theta}_{JS,i} = \sigma^2 + \sigma^4(n-2)^2 \text{Var}_\theta \frac{Y_i}{\|Y\|_2^2} - 2\sigma^2(n-2)\left(E_\theta \frac{Y_i^2}{\|Y\|_2^2} - E_\theta \frac{\theta_i Y_i}{\|Y\|_2^2}\right),$$

respectively. (Note that since $E_\theta 1/\|Y\|^p$ is finite if and only if $n > p$, all expectations here are finite for $n \geq 3$. See Exercise 3.1.) It follows that the mean squared error of the estimator is given by

$$\sigma^2 n + \sigma^4(n-2)^2 E_\theta \frac{1}{\|Y\|_2^2} - 2\sigma^2(n-2)\left(\sum E_\theta \frac{Y_i(Y_i - \theta_i)}{\|Y\|_2^2}\right)$$

(check!). By Lemma 3.1.2 below,

$$E_\theta Y_i(Y_i - \theta_i) = E_\theta \frac{\sigma^2}{\|Y\|_2^2} - 2E_\theta \frac{\sigma^2 Y_i^2}{\|Y\|_2^4}.$$

Hence, the mean squared error (MSE) $E_\theta \|\hat{\theta}_{JS} - \theta\|^2$ equals

$$\sigma^2 n - \sigma^4(n-2)^2 E_\theta \frac{1}{\|Y\|_2^2}.$$ (3.2)

Since the MSE of the MLE $Y$ equals $n$, this completes the proof. 

**Lemma 3.1.2.** Let $Y \sim N_n(\theta, I)$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be an almost everywhere differentiable function such that $E_\theta |(\partial f/\partial x_i)(Y)| < \infty$ for $i = 1, \ldots, n$. Then for $i = 1, \ldots, n$,

$$E_\theta (Y_i - \theta_i) f(Y) = E_\theta \frac{\partial f}{\partial x_i}(Y).$$
Proof. Integration by parts, see Exercise 3.2.

The James-Stein theorem gives a number of very interesting insights in statistics for high-dimensional models. It shows that by shrinking the MLE towards zero, thereby reducing the variance at the cost of increasing the bias, we obtain an estimator with a strictly better risk $E_\theta \| \hat{\theta} - \theta \|^2$. Moreover, although the observed $Y_i$ are independent by assumption, the shrinkage factor depends on all the observations. Hence, to estimate the $i$th component $\theta_i$, we do not only use the information in $Y_i$, but we also borrow strength from the other observations.

One argument that Stein (1956) used to intuitively justify the concept of shrinkage is the observation that if $Y \sim N_n(\theta, I)$, then by the law of large numbers it holds for large $n$ that $\|Y\|^2 \approx \|\theta\|^2 + n$. So the norm of the MLE $Y$ is typically substantially larger than the norm of the parameter $\theta$ it is supposed to estimate. Therefore, it may be beneficial to shrink the vector $Y$ so that the norm is reduced.

Alternatively, we may argue that shrinking reduces the contributions of outliers, i.e. relatively large observations $Y_i$, on the squared estimation error. This possibly comes at the cost of increasing the error made in the other coordinates, but the net effect is that shrinking improves the total squared error $\|\hat{\theta}_{JS} - \theta\|^2$ of the estimator on average. Observe that this reasoning indicates that it is essential that the we assess the quality of the estimator using a norm that simultaneously takes all coordinates of $\theta$ into account. This allows us to trade off gains in one coordinate with losses in others.

The James-Stein theorem can be generalized in many directions, for instance away from the normal distribution with unit variance, using other norms, other statistical models, et cetera. The precise form of the shrinking is not crucial either. Shrinking towards a fixed point $v \in \mathbb{R}^d$ other than 0 works just as well for instance (see Exercise 3.3). The general message is always that in high-dimensional settings it is typically advantageous to somehow reduce the variance by shrinking, or otherwise regularizing. We explore this further in the next section.

Theorem 3.1.1 shows that for $n \geq 3$, the MLE $\hat{\theta}_{\text{MLE}} = Y$ is inadmissible in the model $Y \sim N_n(\theta, I)$, with respect to the squared Euclidean risk. By definition this means that there exists another estimator $\hat{\theta}$ such that $E_\theta \| \hat{\theta} - \theta \|^2 \leq E_\theta \| \hat{\theta}_{\text{MLE}} - \theta \|^2$ for all $\theta \in \mathbb{R}^n$, with strict inequality for at least one $\theta \in \mathbb{R}^n$. The theorem asserts that the James-Stein estimator is such an estimator. It can be shown however that the James-Stein estimator itself is inadmissible as well. For example the positive part Stein estimator

$$\hat{\theta}_{JS^+} = \left( 1 - \frac{n - 2}{\| Y \|^2} \right)_+ Y$$

is an estimator with strictly smaller risk for all $\theta \in \mathbb{R}^n$. See for instance Section 3.4 of Tsybakov (2009). Unfortunately, $\hat{\theta}_{JS^+}$ is not admissible either.
Admissible estimators exist, but have more complicated forms. See for instance Section 5.5 of Lehmann and Casella (1998).

### 3.2 Regularization in the white noise model

The main message that we should pick up from the James-Stein theorem is that if we are estimating a high-dimensional or infinite-dimensional parameter and we want good performance relative to a global norm on the parameter that somehow takes all dimensions into account simultaneously, then we should use some form of shrinking. In Theorem 3.1.1 the shrinking is done quite literally. The MLE is multiplied by a factor smaller than 1, which shrinks the estimator towards 0.

In this section we explore similar methods in the context of the white noise model. It is convenient to work in the sequence formulation (2.4) in which we want to infer the sequence of coefficients \((f_j)\) in \(\ell^2\) from the observations \(Y_j\) satisfying

\[
Y_j = f_j + \frac{1}{\sqrt{n}}Z_j, \quad j = 1, 2, \ldots,
\]

where \(Z_1, Z_2, \ldots\) are independent standard Gaussians. We assess the quality of an estimator \(\hat{f}\) by the mean squared error, or squared \(\ell^2\)-risk \(E_f \| \hat{f} - f \|^2 = E_f \sum (\hat{f}_j - f_j)^2\). The MLE in this model is simply the sequence of observations \((Y_j)\). This is clearly not a very useful estimator. It a.s. does not belong to \(\ell^2\) (why?) and in particular its mean squared error is equal to \(\infty\). As in the preceding section, we can improve the situation by using one of several possible methods to ‘shrink’ the MLE. In this case simply multiplying by a single factor less than 1 will not work, since this would lead to an estimator which still does not belong to \(\ell^2\). There are various other methods however, which can all be viewed as some form of shrinking.

The key in every method is to optimize the risk by changing the bias-variance trade-off. The risk of an estimator \(\hat{f}\) can in the present situation be decomposed as

\[
E_f \| \hat{f} - f \|^2 = \sum (E_f \hat{f}_j - f_j)^2 + \sum \text{Var}_f \hat{f}_j.
\]

As usual, we call the first term the squared bias term and the second one the variance term. In the white noise model the MLE \(Y\) has squared bias \(0\) and variance \(\infty\). The examples below give methods that reduces the variance, at the cost of introducing some bias. The net effect can then be that the risk of the estimator improves.

**Example 3.2.1** (Simple projection estimators). A straightforward method for shrinking the MLE towards 0 is to choose a ‘dimension’ \(J \in \mathbb{N}\) and to truncate all observations above that dimension. In other words, we define the estimator \(\hat{f}\) by setting

\[
\hat{f}_j = Y_j 1_{j \leq J}, \quad j = 1, 2, \ldots
\]
This is called a projection estimator, because it projects the MLE $Y$ onto a finite-dimensional subspace of $\mathbb{R}^\infty$. The risk of the estimator is given by

$$E_f \| \hat{f} - f \|_2^2 = \sum_{j>J} f_j^2 + \frac{J}{n},$$

where the first term is the squared bias and the second one is the variance. For every choice of $J$ the projection estimator beats the MLE in terms of risk. We see from the expression for the risk that the optimal choice of $J$ depends on the rate of decay of the squared coefficients $f_j^2$.

**Example 3.2.2** (Smoothing). Recall that the $f_j$ in the sequence formulation of the white noise model can be interpreted as the Fourier coefficients of the signal $f$ in the model (2.1) relative to some orthonormal basis $e_j$ of $L^2[0,1]$. In the case of the ordinary Fourier basis, the rate of decay of the Fourier coefficients describes the ‘smoothness’, or ‘regularity’ of the signal. Roughly speaking: the faster the rate of decay, the less high frequencies the signal contains, the smoother it is (see Exercise 3.6). Viewed from this perspective, the projection estimator can be seen as performing some kind of smoothing. Instead of the MLE, which corresponds to a very rough signal containing all frequencies at the same energy level, it returns a signal in which all Fourier coefficients except the first $J$ are 0. In other words, it projects the MLE onto a space of smoother signals.

We can generalize this idea and consider other methods that somehow project onto spaces of ‘smooth’ signals. For instance, we can fix a sequence of numbers $c \in \ell^2$ and consider the estimator $\hat{f}$ defined by

$$\hat{f}_j = Y_j c_j.$$ 

If $c_j = 1_{j\leq J}$ we recover the simple projection estimator. In other cases it is not exactly equal to the simple projection estimator, but it reduces the Fourier coefficients in a similar way and can be seen as a form of smoothing as well. The risk of the estimator is given by

$$E_f \| \hat{f} - f \|_2^2 = \sum_{j} (1 - c_j)^2 f_j^2 + \frac{\|c\|_2^2}{n}. \quad (3.3)$$

Again, the size of the risk is determined by the combination of the choice of $c$ and the rate of decay of the Fourier coefficients.

**Example 3.2.3** (Restricting the parameter space). We can view the shrinkage estimators we have just seen as restricted maximum likelihood estimators, where we maximize the likelihood over a particular subset of (Fourier coefficients of) signals. Indeed, in the sequence model the likelihood (relative to an infinite product of $N(0,1/n)$-distributions) is proportional to

$$f \mapsto \prod e^{-\frac{1}{2} (Y_j - f_j)^2}.$$
Maximizing this over all possible sequences $f$ gives the MLE $Y$. But if instead we fix $J \in \mathbb{N}$ and consider the set $F_J = \{ f \in \mathbb{R}^\infty : f_j = 0 \text{ for all } j > J \}$ and maximize the likelihood over $F_J$, then we obtain as maximizer the simple projection estimator of Example 3.2.1.

More generally, we can define an estimator by choosing some set of sequences $F \subset \ell^2$ and defining $\hat{f}$ to be the (or a) point in $F$ where the likelihood is maximal, provided that it exists. Since maximizing the likelihood is equivalent to minimizing the function $f \mapsto \sum (Y_j - f_j)^2$, this estimator is given by

$$\hat{f} = \arg\min_{f \in F} \sum (Y_j - f_j)^2. \quad (3.4)$$

Different choices for $F$ correspond to different ways of shrinking or smoothing. In the next example we will see for instance that if $F$ is an ellipsoid of the form $F = \{ f \in \ell^2 : \sum a_i^2 f_i^2 \leq r^2 \}$, then the restricted MLE coincides with a smoothing estimator as considered in Example 3.2.2. The size of the set $F$ determines the degree of shrinking in some sense. The smaller the $\ell^2$-diameter of the set $F$ for instance, the more the MLE is shrunk towards 0.

**Example 3.2.4** (Penalization). Consider again the estimator $\hat{f}$ defined as the maximizer of the likelihood over a restricted set of (Fourier coefficients of) signals $F$. In many situations the minimization problem (3.4) has a dual formulation that can be useful and that gives yet another perspective on shrinkage. To indicate how this works, suppose for concreteness that the set is of the form $F = \{ f \in \ell^2 : p(f) \leq c \}$ for some function $p : \ell^2 \to \mathbb{R}$ and $c \in \mathbb{R}$. So the restricted MLE is given by

$$\hat{f} = \arg\min_{f \in \ell^2, p(f) \leq c} \sum (Y_j - f_j)^2. \quad (3.5)$$

Then setting $h = p - c$ this constrained minimization problem is equivalent to the global minimization problem

$$\hat{f} = \arg\min_{f \in \ell^2} \sum (Y_j - f_j)^2 + \infty \cdot 1_{h(f) > 0}.$$ 

Now the idea is to replace the function $x \mapsto \infty \cdot 1_{x > 0}$ by the linear function $x \mapsto \lambda x$ for some $\lambda \geq 0$ and to consider instead the minimization of the Lagrangian

$$L(f, \lambda) = \sum (Y_j - f_j)^2 + \lambda h(f).$$

Observe that for every $f \in \ell^2$ we have

$$\sup_{\lambda \geq 0} L(f, \lambda) = \sum (Y_j - f_j)^2 + \infty \cdot 1_{h(f) > 0},$$

hence finding $\hat{f}$ means solving the optimization problem $\inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda)$. 
We might hope that we can interchange the inf and the sup in this problem and that it is equivalent to solve the dual problem

$$\sup_{\lambda \geq 0} \inf_{f \in \ell^2} L(f, \lambda).$$

This is indeed the case under some conditions. It is easy to see that we always have the inequality

$$\sup_{\lambda \geq 0} \inf_{f \in \ell^2} L(f, \lambda) \leq \inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda),$$

which is called weak duality (see Exercise 3.7). If $h$ is convex, then the function $L$ is convex in $f$ and linear in $\lambda$. An appropriate version of the minimax theorem (see for instance Section 45 of Strasser (1985)) then implies that for all compact $\Lambda \subset [0, \infty)$,

$$\sup_{\lambda \in \Lambda} \inf_{f \in \ell^2} L(f, \lambda) = \inf_{f \in \ell^2} \sup_{\lambda \in \Lambda} L(f, \lambda).$$

Thus we have the desired strong duality if

$$\inf_{f \in \ell^2} \sup_{\lambda \in [0, n]} L(f, \lambda) \to \inf_{f \in \ell^2} \sup_{\lambda \geq 0} L(f, \lambda).$$

This is true for instance if the so-called Slater condition holds, which postulates that there exists an $f_0 \in \ell^2$ such that $h(f_0) < 0$. In our case this corresponds to the condition that there exists an $f_0 \in \ell^2$ such that $p(f_0) < c$, which is typically very mild. See for instance Section 5.3.2 of Boyd and Vandenberghe (2004).

If we indeed have strong duality and the estimator $\hat{f}$ can be obtained by solving the dual optimization problem, then for some $\lambda > 0$, called the Lagrange multiplier, we have

$$\hat{f} = \arg\min_{f \in \ell^2} \sum (Y_j - f_j)^2 + \lambda p(f).$$

The multiplier $\lambda$ will obviously depend on the constant $c$ in the primal formulation (3.5) of the problem.

An estimator like (3.7) is called a penalized maximum likelihood estimator and the term $\lambda p(f)$ that is added to the likelihood is called a penalty term. The idea is that functions $f$ for which the penalty $p(f)$ is large are considered to be undesirable as estimators. Minimizing the penalized least squared criterion on the right of (3.7) means balancing the likelihood term and the penalty term, i.e. finding functions $f$ for which the the likelihood is large, but at the same time the penalty is small. The tuning parameter $\lambda$ controls the relative importance of the likelihood and the penalty. For $\lambda = 0$ the penalty disappears and we just have the MLE. As $\lambda$ becomes larger, the penalty becomes more important, which corresponds to more and more regularization or shrinking.
Penalties that are routinely used include various kinds of norms on the function $f$ or its derivatives.

As a concrete example, consider a penalty of the form

$$p(f) = \sum a_i^2 f_i^2$$

for some sequence $a_i$. (If $a_i = j^\beta$ for some $\beta > 0$, then we can think of $p(f)$ as the squared $L^2$-norm of the $\beta$th derivative of the function $f$ with Fourier coefficients $f_k$, cf. Exercise 3.6.) Then the minimization in (3.7) can be done coordinatewise and we find that the penalized MLE $\hat{f}$ is in this case given by

$$\hat{f}_j = \frac{1}{1 + \lambda a_j^2} Y_j$$

(check!). So we see that the smoothing estimators considered in Example 3.2.2 can also be viewed as particular penalized MLEs. Note that the tuning parameter $\lambda$ indeed controls the degree of shrinkage.

Example 3.2.5 (Bayesian methods). The expression (3.7) for the penalized MLE points at a connection to Bayesian methods. Doing Bayesian inference in the white noise model involves putting some prior distribution on $f$ and computing the corresponding posterior. As noted in Example 3.2.3 the likelihood is a multiple of $\prod \exp\left(-\frac{n}{2}(Y_j - f_j)^2\right)$. Hence if the prior has density $\pi$ relative to some dominating measure, then the posterior has density proportional to

$$f \mapsto \pi(f) \prod e^{-\frac{n}{2}(Y_j - f_j)^2}.$$  

In particular, we see that the posterior mode, or maximum a-posteriori (MAP) estimator, is given by

$$\hat{f}_{\text{MAP}} = \arg\min_f \sum (Y_j - f_j)^2 - \frac{2}{n} \log \pi(f).$$

(3.8)

This shows that the MAP can be viewed as a penalized MLE which shrinks the estimator towards the mode of the prior. The smaller the prior density $\pi$ at a function $f$, the more $f$ is penalized.

As a concrete example, suppose we put a product of normals as prior on $f$, with under the prior $f_j \sim N(0, \tau_j^2)$ for some sequence of variances $\tau_j^2$. Then since both the likelihood and the prior factorize, the posterior factorizes as well and we can compute the the posterior coordinatewise. The posterior density for $f_j$ is proportional to

$$e^{-\frac{n}{2}(Y_j - f_j)^2} e^{-\frac{1}{2}f_j^2/\tau_j^2}.$$  

Completing the square we find that the posterior is a product of normals

$$\bigotimes N\left(\frac{n\tau_j^2 Y_j}{1 + n\tau_j^2}, \frac{\tau_j^2}{1 + n\tau_j^2}\right).$$
(check!). In particular, the posterior mean is the estimator $\hat{f}$ given by

$$\hat{f}_j = \frac{n\tau_j^2}{1 + n\tau_j^2} Y_j, \quad j = 1, 2, \ldots. \tag{3.9}$$

Hence, the posterior mean is a shrinkage, or smoothing estimator and the prior variances $\tau_j^2$ determine the degree of smoothing. We note that since the posterior is Gaussian in this case, its mean is in fact equal to its mode. Indeed, in this example it can be verified directly that the estimator defined by (3.8) coincides with the estimator $\hat{f}$ given above (Exercise 3.8).

\[\square\]

### 3.3 Regularization in other models

In these notes we mostly focus on the analysis of the canonical signal-in-white-noise model. The idea of regularization is general however and plays a role in some form in every high-dimensional or nonparametric problem. In this section we give a few more well-known examples.

**Example 3.3.1 (Kernel density estimation).** Consider the density estimation problem of Example 1.2.3. So we have an i.i.d. sample $X_1, \ldots, X_n$ from a positive density $p$ on the line. In this model too, it is necessary to use some form of regularization. The likelihood is given by

$$p \mapsto \prod \rho(X_j).$$

By considering a sequence of densities with higher and higher spikes at the observations it is seen that the maximum over all densities is $\infty$, hence the unrestricted MLE does not exist.

Although strictly speaking it does not exist as a function, the MLE can be thought of as a sum of delta functions at the observations. The classical kernel density estimator can be seen as a way of regularizing this ‘estimator’. Instead of putting a delta function at each observation we distribute mass 1 over the line by putting a smooth ‘bump’ of mass $1/n$ at each observation. To do so, we fix a kernel function $K$, which is a function that integrates to 1 and has mean 0 (usually a probability density) and a so-called bandwidth parameter $h > 0$. Then at every observation $X_i$ we put the ‘bump’ $x \mapsto (nh)^{-1}K((x - X_j)/h)$, which is a function that integrates to $1/n$. The kernel estimator $\hat{p}_h$ for $p$ is thus defined as

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right), \quad x \in \mathbb{R}.$$  

Note that $h$ controls the smoothness of the estimator. If $h$ is very small each bump looks like a delta function and $\hat{p}_h$ is very spiky. If $h$ becomes very large the bumps look more and more constant and $\hat{p}_h$ ‘shrinks’ towards a ‘uniform’ distribution.

In Exercise 3.9 you are asked to explore kernel density estimation in R. [\square]
Example 3.3.2 (Estimating a monotone density). Kernel density estimators provide smooth estimators of the unknown density, the degree of smoothness being controlled in some sense by the bandwidth. If the true density itself is believed to be smooth, this kind of regularization makes perfect sense. There are other methods of regularizing as well, for instance by restricting shape instead of smoothness.

To illustrate this, consider an i.i.d. sample $X_1, \ldots, X_n$ from a density $p$ on $[0, \infty)$. If the true density is believed to be decreasing, it makes sense to consider the class $\mathcal{P}_{\text{dec}}$ of decreasing densities on $[0, \infty)$ and to consider the restricted MLE over that class, i.e.

$$\hat{p} = \arg\max_{p \in \mathcal{P}_{\text{dec}}} \prod p(X_i).$$

It can be seen that this criterion is maximized at densities that are left-continuous, locally constant, with jumps only at the observations (check!). For such a maximizer the constant value between 0 and $X_{(1)}$ can be at most $1/X_{(1)}$, since the density must integrate to 1. Hence, if we denote the function values of a potential maximizer from left to right by $h_1, \ldots, h_n$, then $h = (h_1, \ldots, h_n)$ must belong to the set

$$\mathcal{H} = \{h \in \mathbb{R}^n : 0 \leq h_1 \leq \cdots \leq h_n \leq 1/X_{(1)}; \sum h_j (X_{(j)} - X_{(j-1)}) = 1\}.$$ 

Maximizing the likelihood over $\mathcal{P}_{\text{dec}}$ then boils down to maximizing the continuous function $h \mapsto \prod h_j$ over the compact set $\mathcal{H}$. This shows that the restricted MLE $\hat{p}$ exists.

It can be shown that the solution is unique and that it is the left derivative of the so-called least concave majorant of the empirical distribution function. As the name suggests, the latter is the smallest concave function that lies pointwise above the empirical distribution function. This is a piecewise linear function with kinks at the observations, hence its derivative is indeed a piecewise constant function with jumps at the observations. This monotone density estimator is known as the Grenander estimator.

In Exercise 3.10 you are asked to compare the Grenander estimator to kernel density estimation on simulated data.

\[\blacksquare\]

Example 3.3.3 (Ridge regression). Consider a linear model $Y = X\theta + \varepsilon$, with $X$ a fixed $n \times p$ matrix, $\theta \in \mathbb{R}^p$ and $\varepsilon \sim N_n(0, \sigma^2 I)$. The MLE for $\theta$ is the least squares estimator, which minimizes the residual sum of squares $\theta \mapsto ||Y - X\theta||^2$. Equivalently, it is the solution of the linear system $X^T X\theta = X^T Y$. In the high-dimensional setting $p > n$ this system does not have a unique solution however, since $X^T X$ is $p \times p$ matrix with rank at most $n < p$. To remedy the situation a form of regularization can be employed, for instance by adding a penalty term to the least squares criterion. One method that is often used is adding a penalty proportional to the squared Euclidean norm of $\theta$. This is called the ridge penalty and the resulting procedure is called ridge regression. Specifically,
the ridge regression estimator is defined by choosing a tuning parameter $\lambda > 0$ and defining

$$\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|^2.$$

A simple calculation shows that the minimizer satisfies the equation $X^T Y - X^T X \hat{\theta} = \lambda \hat{\theta}$, i.e. the ridge regression estimator is given by

$$\hat{\theta} = \left(X^T X + \lambda I\right)^{-1} X^T Y$$

(check!). We see that adding the ridge penalty indeed results in a degree of shrinkage towards 0, controlled by the tuning parameter $\lambda$.

Example 3.3.4 (The lasso). Consider again the high-dimensional linear model as in the preceding example. In ridge regression the penalty is the Euclidean, or $\ell^2$-norm of the parameter $\theta$. Another very popular method is adding an $\ell^1$-penalty. That is, we define $\|\theta\|_1 = \sum |\theta_j|$, choose some tuning parameter $\lambda > 0$ and define the estimator for $\theta$ by

$$\hat{\theta} = \arg\min_{\theta} \|Y - X\theta\|^2 + \lambda \|\theta\|_1.$$

This estimator is called the lasso (least absolute shrinkage and selection operator) estimator. Contrary to the ridge regression estimator, the lasso can in general not be expressed explicitly. Computing it is still a convex optimization problem however and there exist efficient algorithms to compute it numerically. An important feature of the lasso is that since it minimizes a convex function over an $\ell^1$-ball (see the dual perspective explained in Example 3.2.4), it typically produces estimators with many 0 coordinates. This is a desirable feature in many applications in which the true parameter is believed to be sparse, i.e. to have many 0 coordinates.

This is most easily seen in the case $X = I$, i.e. in the normal means model. It can be verified that in that case the lasso defined above is given explicitly by

$$\hat{\theta}_i = h_\lambda(Y_i),$$

where $h_\lambda$ is the function defined by

$$h_\lambda(y) = \begin{cases} y + \lambda/2 & \text{if } y < -\lambda/2, \\ 0 & \text{if } -\lambda/2 \leq y \leq \lambda/2, \\ y - \lambda/2 & \text{if } y > \lambda/2 \end{cases}$$

(see Exercise 3.11). The lasso is therefore called a soft thresholding procedure: if $|Y_i|$ is below the threshold $\lambda/2$, then the estimator for $\theta_i$ is 0. For larger values the estimator moves away from 0 linearly. For $|Y_i|$ very large the impact of $\lambda$ becomes negligible and we have $\hat{\theta}_i \approx Y_i$. ■
Example 3.3.5 ($\ell^0$-penalization). Another penalty that is used in the high-dimensional linear model setting is the $\ell^0$-norm. The $\ell^0$-norm $||\theta||_0$ of a vector in $\mathbb{R}^n$ is defined as the number of non-zero coordinates: $||\theta||_0 = |\{i : \theta_i \neq 0\}|$. This penalty penalizes non-sparse vectors even more clearly than the $\ell^1$-penalty of the lasso. The corresponding penalized MLE for $\theta$ is defined by

$$\hat{\theta} = \arg\min_{\theta} ||Y - X\theta||^2 + \lambda||\theta||_0,$$

(3.10)

for some tuning parameter $\lambda > 0$. Computing this estimator is in general computationally hard if $p$ is large, since it is not a convex optimization problem. The computation essentially involves searching over all possible sparsity patterns. The lasso can be viewed as a convex relaxation of this problem. It still has much of the same flavour, but is computationally much more feasible.

Also in this case we can compute the estimator explicitly in the normal means model, i.e. when $X = I$. It can be seen that the $\ell^0$-penalized estimator is then given by

$$\hat{\theta}_j = Y_j 1_{|Y_j| > \sqrt{\lambda}}$$

(see Exercise 3.12). For obvious reasons it is therefore called a hard thresholding estimator.

The examples we gave in this section treat some important classes of regularization methods in the context of several classical statistical models. It should be clear that this list is by no means complete and that many variations on this theme exist. See exercises 3.13 and 3.14 for some more examples.

3.4 The bias-variance trade-off and convergence rates

We have seen in the James-Stein theorem and in the examples in this chapter that in high-dimensional or nonparametric models, naively applying methods like maximum likelihood is either simply not possible, or leads to procedures with sub-optimal performance in terms of mean squared error, for instance. Regularization methods improve the situation by changing the bias-variance trade-off. They yield estimators with reduced variance, at the cost of introducing some bias. The net effect is generally an estimator with better overall performance.

The variance of an estimator quantifies how much the estimator depends on the particular dataset. Of course the estimator should depend on the dataset, but if the variance is too large it will depend too much on the particular realization that has been observed and will not provide a stable estimator of the underlying truth. On the other hand if the variance is too small then the bias will be relatively large which results in an estimator which is quite stable, but which makes a large structural error. If the variance is too large, we also say that the estimator is overfitting, reflecting the idea that it follows the fluctuations in the particular dataset too closely. If the variance is too small, hence the bias too large, we say we are underfitting.
Example 3.4.1 (over- and undersmoothing in kernel density estimation). To illustrate the issue of over- and underfitting we simulate i.i.d. data from a known density and try to recover it using a kernel density estimator as in Example 3.3.1. Figure 3.1 shows kernel density estimates for 6 different samples from the same density, namely a mixture of two normals $N(\pm 1, 0.5)$, with equal weights $1/2$.

![Kernel density estimates for different samples of size 200 from a normal mixture. Top row: normal kernel with bandwidth $h = 0.05$. Bottom row: normal kernel with bandwidth $h = 1$.](image)

The performance depends crucially on the bandwidth parameter $h$ that is chosen. In the top row $h$ was chosen too small on purpose. The estimator then closely follows the particular data that is observed. As a result the variance of the estimator becomes too large, i.e. we are overfitting. In this context this is also called undersmoothing, meaning that we do not regularize enough, the estimator is too ‘rough’ compared to the truth. In the lower row the bandwidth is chosen too large. The estimator is very stable, but we introduce a large bias and the estimator remains far away from the truth. This is called oversmoothing: the estimator is too ‘smooth’ relative to the truth. Choosing of a good bandwidth means finding a comprise between these two extremes. ■
As we have seen either explicitly or implicitly in the examples, all regularization methods have some tuning mechanism to control the bias-variance trade-off. For the simple projection estimator it is the dimension $J$ at which the MLE is truncated, for the penalization methods it is the tuning parameter $\lambda$ which multiplies the penalty, for the restricted MLE it is the size of the set over which the likelihood is maximized, for the kernel density estimator it is the bandwidth $h$, et cetera. What we can also see from the examples, for instance from expressions for the risk like (3.3) or from numerical examples like Example 3.4.1, is that the optimal choice of the tuning parameters will depend on properties of the true parameter. Unfortunately, these are usually properties like some form of ‘smoothness’, or another measure of the ‘complexity’ of the true parameter that we don’t have direct access to, since the true parameter is unknown.

It is useful however to understand how the tuning parameters of a given procedure should be set if properties like smoothness were in fact known. This optimal setting is sometimes called the oracle setting of a tuning parameter, reflecting the idea that an oracle that knows aspects of the truth that are inaccessible to us would know how to set it. Knowing this oracle choice of a tuning parameter gives insight in the best performance that can in principle be achieved with a given method.

One way to assess the performance of a statistical procedure is to take an asymptotic perspective and study how the risk of the estimator behaves as $n \to \infty$, where $n$ denotes the ‘amount’ of data that is available. In the case of i.i.d. observations $n$ is the sample size, in the case of the white noise model it is the signal-to-noise ratio $n$ in (2.1). Typically the risk of an estimator will tend to 0 as $n \to \infty$. Of course the rate at which this happens depends on the particular risk that it is considered. In these notes we often consider risks of the form $E_\theta \| \hat{\theta}_n - \theta \|^p$, for some norm $\| \cdot \|$ on the parameter of interest and $p = 1$ or $p = 2$. If the risk tends to 0 as $n \to \infty$, then by Markov’s inequality the distance $\| \hat{\theta}_n - \theta \|$ of the estimator to the true parameter is of stochastic order $(E_\theta \| \hat{\theta}_n - \theta \|^p)^{1/p}$. This is then called the rate of convergence of the estimator (relative to the norm $\| \cdot \|$). More generally, if $\| \hat{\theta}_n - \theta \| = O_{P_\theta}(r_n)$ for some $r_n \to 0$, then $r_n$ is called a rate of convergence of the estimator $\hat{\theta}_n$.

As discussed in Section 1.1, estimators in smooth parametric models typically have rate of convergence $1/\sqrt{n}$ (relative to the Euclidean norm). Rates in nonparametric or high-dimensional problems are usually slower.

**Example 3.4.2** (Simple projection estimator, ellipsoids). Consider the situation of Example 3.2.1 again. We can optimize the risk by tuning the truncation point $J$ to the rate of decay of the true sequence $f_j$. Indeed, suppose that for some $\beta, r > 0$, $f$ belongs to the ellipsoid

$$\mathcal{F} = \{ f : \sum j^{2\beta} f_j^2 \leq r^2 \}.$$
For \( f \in \mathcal{F} \) the squared bias satisfies
\[
\sum_{j > J} f_j^2 = \sum_{j > J} j^{-2\beta} j^{2\beta} f_j^2 \leq J^{-2\beta} r^2.
\]
This is balanced with the variance term \( J/n \) for \( J \sim n^{1/(1+2\beta)} \). The risk of the estimator is then of the order \( n^{-2\beta/(1+2\beta)} \), hence the rate of convergence is of the order \( n^{-\beta/(1+2\beta)} \).

Recall that the sequence formulation (2.4) of the white noise model can be seen as the ‘frequency domain’ view, which corresponds to the ‘time domain’ view (2.1). In the time domain, the assumption that the sequence of Fourier coefficients belongs to \( \mathcal{F} \) can be viewed as a ‘smoothness’ condition on the function in (2.1). As explained in Exercise 3.6, the condition that the Fourier coefficients of a function \( f \) belong to the ellipsoid \( \mathcal{F} \) corresponds to the smoothness condition that the signal \( f \) has \( \beta \) square integrable derivatives, with \( L^2 \)-norms bounded by a common constant. Such a space of functions is called a Sobolev ball of order \( \beta \) (a precise definition is given ahead).

Using this terminology, we see in this example that if the signal \( f \) has Sobolev smoothness \( \beta \), then the projection estimator with \( J \sim n^{1/(1+2\beta)} \) converges in \( \ell^2 \) to the true \( f \) at the rate \( n^{-\beta/(1+2\beta)} \). For very small \( \beta \), i.e. very rough signals, this rate is very slow. For \( \beta \to \infty \) the rate comes closer and closer to the parametric rate \( 1/\sqrt{n} \). We will see in Chapter 5 that the rate \( n^{-\beta/(1+2\beta)} \) is in some sense the best we can get if we want to do well for all possible signals in a Sobolev ball of order \( \beta \).

For kernel density estimation, as considered in Examples 3.3.1 and 3.4.1, results similar those for the projection estimator can be obtained both for the pointwise risk \( \mathbb{E}_p (\hat{p}_h(x) - p(x))^2 \), with \( x \) a fixed point, and for the integrated risk \( \mathbb{E}_p \int (\hat{p}_h(x) - p(x))^2 \, dx \). A good bias-variance trade-off can be achieved by appropriately tuning the bandwidth parameter \( h \) to the smoothness of the unknown density \( p \), see Exercise 3.15.

The following two examples consider kernel smoothing in the context of the white noise model. We investigate in particular how the rate changes if we consider uniform risk instead of pointwise risk. In this setting the natural notion of smoothness to consider is Hölder regularity. We say a function \( f \) on \([0, 1]\) has Hölder regularity \( \beta > 0 \) if, for \( \beta \) the largest integer strictly smaller than \( \beta \), it has \( \beta \) continuous derivatives and the highest derivative \( f^{(\beta)} \) satisfies
\[
|f^{(\beta)}(t) - f^{(\beta)}(s)| \leq C|t - s|^{(\beta - \beta)}
\]
for a constant \( C > 0 \) and all \( s, t \in [0, 1] \). If this holds we write \( f \in C^\beta[0, 1] \).

**Example 3.4.3** (Kernel smoothing in the white noise model, pointwise risk). Consider the white noise model (2.1). To estimate the signal \( f \) we fix a symmetric, smooth, compactly supported probability density \( K \) and a bandwidth
\( h > 0 \). We define the estimator \( \hat{f}_h \) by

\[
\hat{f}_h(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dX_s, \quad t \in [0, 1], \tag{3.11}
\]

where the integral is a Wiener integral as defined in Section 2.2.

Now fix a point \( t \in (0, 1) \). Using the SDE for \( X \) we see that

\[
\hat{f}_h(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) f(s) \, ds + \frac{1}{\sqrt{n}} \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dW_s.
\]

Hence, it follows from the properties of the Wiener integral that the bias and the variance of the estimator \( \hat{f}_h(t) \) are given by

\[
\mathbb{E}_f \hat{f}_h(t) - f(t) = \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) f(s) \, ds - f(t),
\]

\[
\text{Var}_f \hat{f}_h(t) = \frac{1}{n} \int_0^1 \frac{1}{h^2} K^2\left(\frac{t-s}{h}\right) \, ds,
\]

respectively. Note that the variance does not depend on the true signal \( f \). A change of variables shows it is of the order \( 1/(nh) \). To bound the bias, first note that by a change of variables and using that the support of \( K \) is contained in a compact interval \([-a, a]\) for some \( a > 0 \), we have, for \( h \) small enough,

\[
\mathbb{E}_f \hat{f}_h(t) - f(t) = \int_{-a}^a K(u)(f(t-hu) - f(t)) \, du
\]

(check!). Now suppose that \( f \in C^\beta[0, 1] \) for some \( \beta \in (0, 2] \). Let \( \hat{\beta} \) be the largest integer strictly smaller than \( \beta \). Then by Taylor’s formula,

\[
f(t-hu) = \sum_{k \leq \hat{\beta}} f^{(k)}(t)(-hu)^k + r,
\]

where \(|r| \leq \text{const} \times |hu|^\beta\). Since \( K \) integrates to 1 and its first moment is 0, we conclude that

\[
\left| \mathbb{E}_f \hat{f}_h(t) - f(t) \right| \leq \text{const} \times h^\beta \int_{-a}^a |K(u)||u|^\beta \, du.
\]

So the squared bias is bounded by a constant times \( h^{2\beta} \) if \( f \in C^\beta \). Taking the bandwidth \( h \) of the order \( n^{-1/(1+2\beta)} \) then balances squared bias and variance and leads to a mean squared error, or pointwise risk \( \mathbb{E}_f (\hat{f}_h(t) - f(t))^2 \) of the order \( n^{-2\beta/(1+2\beta)} \). Hence, the pointwise rate of convergence is of the order \( n^{-\beta/(1+2\beta)} \).

It is also possible to get the rate \( n^{-\beta/(1+2\beta)} \) if \( f \in C^\beta[0, 1] \) for \( \beta > 2 \), but then the probability kernel \( K \) has to be replaced by a so-called higher order kernel, which has 0 moments of sufficiently high order.
In the preceding example we only examined the pointwise risk for points $t$ strictly inside $(0, 1)$. This is because for $t = 0$ or $t = 1$ some boundary issues arise with the kernel estimator (3.11). The variance is still of the same order (but the constants are different), but the behaviour of the bias changes. Indeed, for $t = 0$ for instance, we get that the bias is

$$E_f f\hat{h}(0) - f(0) = \int_0^a K(u)(f(-hu) - f(0)) du$$

for $h$ small enough (check!). For $\beta \in (0, 1]$ this is still of the order $h^\beta$, but for $\beta \in (1, 2]$ this is no longer necessarily true. The intuitive reason for the different behaviour of the estimator at the boundaries is that for $t$ in the interior, (3.11) in some sense uses observations $X_s$ both to the left and to the right of $t$ in a symmetric manner. At the boundaries, one could say the estimator somehow uses only half the information it uses in the interior.

In the next example we investigate how the rate of the kernel estimator changes if we replace the pointwise risk by uniform risk. To avoid the technical boundary issues just indicated, we consider the uniform distance over a compact subinterval of $(0, 1)$. Since uniform convergence is stronger than pointwise convergence, we can expect the rate for the uniform risk to be worse than the one obtained in the preceding example. It turns out that the the difference in the rate is only a logarithmic factor.

*Example 3.4.4 (Kernel smoothing in the white noise model, uniform risk). We consider the same estimator $\hat{f}_h$ as in Example 3.4.3, but now we study how the uniform risk $E_f \sup_{u \leq t \leq v} |\hat{f}_h(t) - f(t)|$ behaves for fixed $0 < u < v < 1$. We assume again that $f \in C^\beta[0, 1]$ for $\beta \in (0, 2]$.

For every $t \in [0, 1]$ we have

$$|\hat{f}_h(t) - f(t)| \leq \left| \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) f(s) ds - f(t) \right| + \frac{1}{\sqrt{n}} \left| \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dW_s \right|.$$

By the bias considerations in Example 3.4.3 the first term is bounded by a constant times $h^\beta$, uniformly for $t \in [u, v]$. Hence, it remains to investigate the quantity

$$E_f \sup_{u \leq t \leq v} \left| \int_0^1 \frac{1}{h} K\left(\frac{t-s}{h}\right) dW_s \right|.$$

This is the expectation of the supremum of a centered Gaussian process and we can use some general bounds that are available for such quantities. For instance, Dudley’s maximal inequality for (sub-)Gaussian processes asserts that for a centered Gaussian process $Z$ indexed by a set $T$ and any $t_0 \in T$ it holds that

$$E \sup_{t \in T} |Z_t| \leq E|Z_{t_0}| + C \int_0^{\text{diam}(T)} \sqrt{\log N(\varepsilon, T, d)} d\varepsilon,$$
where $C > 0$ is a universal constant, $d$ is the semimetric on $T$ given by $d^2(s, t) = E(Z_t - Z_s)^2$ and $N(\varepsilon, T, d)$ is the minimal number of balls of $d$-radius $\varepsilon$ that is needed to cover $T$ (see for instance Van der Vaart and Wellner (1996)). We apply this inequality with $Z_t = \int_0^1 h K\left(\frac{t - s}{h}\right) dW_s$,

$T = [u, v]$ and $t_0 = u$. In Example 3.4.3 we already saw that the variance of $Z_{t_0}$ is bounded by $1/h$. By Gaussianity, this implies that $E|Z_{t_0}|$ is of the order $1/\sqrt{h}$. The semimetric $d$ is in this case given by

$$d^2(t_1, t_2) = \int_0^{t_1} \left(\frac{1}{h} K\left(\frac{t_1 - s}{h}\right) - \frac{1}{h} K\left(\frac{t_2 - s}{h}\right)\right)^2 ds.$$ 

If we assume that $K$ has support $[-1, 1]$ for concreteness, we see from this expression that two points $t_1, t_2 \in T$ such that $|t_1 - t_2| > 2h$ are at $d$-distance of the order $1/\sqrt{h}$. It follows that the diameter of $(T, d)$ is of the order $1/\sqrt{h}$ as well. Next, we have

$$K\left(\frac{t_1 - s}{h}\right) - K\left(\frac{t_2 - s}{h}\right) = \int_{t_1}^{t_2} \frac{1}{h} K'(\frac{x - s}{h}) dx.$$ 

By Jensen’s inequality and Fubini, it follows that

$$d^2(t_1, t_2) \leq \frac{t_2 - t_1}{h^3} \int_{t_1}^{t_2} \left(\frac{1}{h} (K')^2\left(\frac{x - s}{h}\right) ds\right) dx.$$ 

Since $K$ has compact support the inner integral is bounded by a constant, hence $d^2(t_1, t_2) \leq C(t_1 - t_2)^2/h^3$ for a constant $C > 0$. This implies that the covering number $N(\varepsilon, T, d)$ is bounded by a constant times $1/(\varepsilon h^{3/2})$ and therefore

$$\int_0^{\text{diam}(T)} \sqrt{\log N(\varepsilon, T, d)} d\varepsilon = \frac{1}{h^{3/2}} \int_0^{h} \sqrt{\log \frac{1}{x}} dx \lesssim \frac{\sqrt{\log 1/h}}{\sqrt{h}}.$$ 

All together, we conclude that

$$E \sup_{u \leq t \leq v} |\hat{f}_h(t) - f(t)| \lesssim h^\beta + \frac{\sqrt{\log 1/h}}{\sqrt{n}h}.$$ 

The two terms on the right are balanced for $h$ of the order $(n/\log n)^{-1/(1+2\beta)}$ and the resulting convergence rate relative to the uniform norm is of the order $(n/\log n)^{-\beta/(1+2\beta)}$. 

\[\blacksquare\]

The examples we have seen show that in the white noise model we have an estimator for the unknown signal $f$ that attains the rate $n^{-\beta/(1+2\beta)}$ relative to
the $L^2$-norm if $f$ belongs to a Sobolev space of order $\beta$, we have an estimator that attains the same rate pointwise if $f$ belongs to a Hölder space of order $\beta$ and under the same Hölder condition we have an estimator that attains the slightly slower rate $(n/\log n)^{-\beta/(1+2\beta)}$ relative to the uniform norm. Closer inspection of the examples shows that the bounds on the different risks actually hold uniformly for $f$ in Sobolev or Hölder balls of order $\beta$.

Concretely, for $f \in C^\beta[0,1]$, we define the Hölder norm of order $\beta$ by

$$
\|f\|_{C^\beta} = \max_{k \leq 2} \|f^{(k)}\|_{\infty} + \sup_{s \neq t} \frac{|f^{(2)}(t) - f^{(2)}(s)|}{|t - s|^{\beta - \beta}}.
$$

The Hölder ball of order $\beta$ and radius $R$ is then defined as $C^\beta_R[0,1] = \{f \in C^0[0,1] : \|f\|_{C^\beta} \leq R\}$. Inspection of Example 3.4.3 shows that for the appropriate choice of the bandwidth $h$, the kernel estimator $\hat{f}_h$ satisfies

$$
\sup_{f \in C^\beta[0,1]} E_f (\hat{f}_h(t) - f(t))^2 \leq \text{const.} \times n^{-2\beta/(1+2\beta)}
$$

(3.12) for all $t \in (0,1)$. Similarly, Example 3.4.4 in fact shows that for an appropriate (different!) choice of the bandwidth, we have

$$
\sup_{f \in C^\beta[0,1]} \sup_{u \leq t \leq v} |\hat{f}_h(t) - f(t)| \leq \text{const.} \times \left(\frac{n}{\log n}\right)^{-\beta/(1+2\beta)}
$$

(3.13) for all $0 < u < v < 1$. The result of Example 3.4.2 is uniform over balls as well. For $\beta > 0$ and $e_j$ an orthonormal basis of $L^2[0,1]$, we define the associated Sobolev space of order $\beta$ by

$$
H^\beta[0,1] = \{f = \sum f_j e_j : \sum j^{2\beta} f_j^2 < \infty\}.
$$

The Sobolev norm of a function $f = \sum f_j e_j$ in $H^\beta[0,1]$ is defined by

$$
\|f\|_{H^\beta}^2 = \sum j^{2\beta} f_j^2.
$$

(Recall that if $e_j$ is the standard Fourier basis and $\beta$ is an integer, then this is essentially the $L^2$-norm of the $\beta$th derivative of $f$, see Exercise 3.6.) The Sobolev ball of order $\beta$ and radius $R$ is $H^\beta_R[0,1] = \{f \in H^\beta[0,1] : \|f\|_{H^\beta} \leq R\}$. The result of Example 3.4.2 is that for the appropriate truncation level $J$, the projection estimator $\hat{f}_J$ satisfies

$$
\sup_{f \in H^\beta[0,1]} E_f \|\hat{f}_J - f\|_2^2 \leq \text{const.} \times n^{-2\beta/(1+2\beta)}.
$$

(3.14)

We will see in Chapter 5 that the rates (3.12)–(3.14) we found are in fact the best possible, in the sense that no other estimators can achieve better risks uniformly over these balls.

The examples considered so far give convergence rates of estimators under smoothness conditions. We end the section with an example of a risk bound under a sparsity condition.
Example 3.4.5 (\(\ell_0\)-penalization, sparse signals). Consider the normal means model \(Y = \theta + \varepsilon\), with \(\varepsilon \sim N_n(0, I)\). Then the \(\ell_0\)-penalized estimator (3.10) is given by \(\hat{\theta}_j = Y_j 1_{|Y_j| \geq \sqrt{\lambda}}\), see Example 3.3.5. In this case the bias-variance trade-off can be optimized by tuning \(\lambda\) to the sparseness of the true signal \(\theta\). Let \(\theta\) be \(s\)-sparse, meaning that it has \(s\) non-zero components. Denote its support set, which is the set of indices of the non-zero components, by \(S\).

Note that \(\hat{\theta}_j\) is symmetric and hence unbiased if \(j \notin S\). If \(j \in S\), then

\[
(E_{\theta}(\hat{\theta}_j - \theta_j))^2 = (E_{\theta}(Y_j - \theta_j - E_{\theta}Y_j 1_{|Y_j| < \sqrt{\lambda}}))^2 \leq \lambda.
\]

Hence, the squared bias term in the risk \(E_{\theta}\|\hat{\theta} - \theta\|^2\) is bounded by \(s\lambda\). As for the variance, note that for \(j \notin S\) we have

\[
\text{Var}_{\theta}\hat{\theta}_j = E_{\theta}Y_j^2 1_{|Y_j| \geq \sqrt{\lambda}} \lesssim \sqrt{\lambda}e^{-\lambda/2}
\]

for \(\lambda \geq 1\) (see Exercise 3.17). For \(j \in S\),

\[
\text{Var}_{\theta}\hat{\theta}_j \leq E_{\theta}(\hat{\theta}_j - \theta_j)^2 = E_{\theta}((Y_j - \theta_j) - Y_j 1_{|Y_j| < \sqrt{\lambda}})^2 \\
\leq 2E_{\theta}(Y_j - \theta_j)^2 + 2E_{\theta}Y_j^2 1_{|Y_j| < \sqrt{\lambda}} \leq 2 + 2\lambda.
\]

Together, we find that

\[
E_{\theta}\|\hat{\theta} - \theta\|^2 \lesssim s(1 + \lambda) + (n - s)\sqrt{\lambda}e^{-\lambda/2}.
\]

For \(\lambda = 2\log(n/s)\) we see that the risk is bounded by a multiple of \(s\log(n/s)\), provided \(s = o(n)\).

Note that through the tuning parameter the estimator uses the knowledge of the number \(s\) of non-zero coefficients, but not their locations. If we know which \(s\) of the \(\theta_j\)'s are nonzero, we can estimate them by \(Y_j\), while we estimate all others by 0. The resulting estimator is unbiased, and has variance \(s\). Hence, the risk of that estimator is of the order \(s\). The additional factor \(\log(n/s)\) is the price that we pay for not knowing the support set of the signal. The rate \(s\log(n/s)\) is in fact the best possible rate that can be attained uniformly over the set so-called nearly black vectors, i.e. the set of \(s\)-sparse vectors with \(s = o(n)\), cf. Donoho et al. (1992). ■

In Exercise 3.18 you are asked to work out the rate of convergence of the lasso in the normal means model.
3.5 Exercises

Exercise 3.1 (Negative moments of the multivariate Gaussian). Let $Z \sim N_n(0, I)$. Show that $E(1/\|Z\|^p) < \infty$ if and only if $n > p$.

Exercise 3.2 (Proof of the James-Stein lemma). Prove Lemma 3.1.2.

Exercise 3.3 (Shrinking towards another point). Let $Y \sim N_n(\theta, I)$ and $v \in \mathbb{R}^n$. Define the estimator
\[
\hat{\theta}_{JS} = v + \left(1 - \frac{n-2}{\|Y-v\|^2}\right)(Y-v).
\]
Prove that for $n \geq 3$, this estimator also satisfies $E_\theta\|\hat{\theta}_{JS} - \theta\|^2 < E_\theta\|\hat{\theta}_{MLE} - \theta\|^2$ for all $\theta \in \mathbb{R}^n$.

Exercise 3.4 (Oracle version of James-Stein). Use the expression (3.2) for the risk of the James-Stein estimator to prove that if $Y \sim N(\theta, \sigma^2 I)$, then for every $\theta \in \mathbb{R}^n$ and $n \geq 3$,
\[
E_\theta\|\hat{\theta}_{JS} - \theta\|^2 \leq 4\sigma^2 + \inf_{c \in \mathbb{R}} E_\theta\|cY - \theta\|^2.
\]
This is a so-called oracle inequality that asserts that up to a constant, the risk of the James-Stein estimator is as good as the risk that could be achieved by an oracle that may use its knowledge of the true parameter $\theta$ to choose the degree of shrinking.

Exercise 3.5 (Admissibility of the MLE in the normal mean model for $n = 1$). For $n = 1$ and $n = 2$ the MLE $Y$ is admissible in the model $Y \sim N_n(\theta, I)$. This exercise deals with the case $n = 1$. So we assume that $Y \sim N(\theta, 1)$. The goal is to prove that there exists no other estimator $\hat{\theta}$ such that $E_\theta(\hat{\theta} - \theta)^2 \leq E_\theta(Y - \theta)^2$ for all $\theta \in \mathbb{R}$, with strict inequality for some $\theta \in \mathbb{R}$.

For $\tau > 0$, consider the $N(0, \tau)$ prior on the parameter $\theta$. Denote the corresponding prior density by $\pi_\tau$.

(i) Show that if an estimator $\hat{\theta}$ as described above would exist, then there would exist an $\varepsilon > 0$ and $\theta_0 < \theta_1$ such that
\[
1 - \int E_\theta(\hat{\theta} - \theta)^2 \pi_\tau(\theta) d\theta \geq \varepsilon \int_{\theta_0}^{\theta_1} \pi_\tau(\theta) d\theta.
\]
(ii) Let $\tilde{\theta}_\tau$ be the posterior mean corresponding to the prior $\pi_\tau$. Compute the corresponding Bayes risk

$$\int E_\theta(\tilde{\theta}_\tau - \theta)^2 \pi_\tau(\theta) d\theta.$$

(iii) Using the results of (i) and (ii), show that if an estimator $\hat{\theta}$ as described above would exist, then

$$\frac{1 - \int E_\theta(\hat{\theta} - \theta)^2 \pi_\tau(\theta) d\theta}{1 - \int E_\theta(\tilde{\theta}_\tau - \theta)^2 \pi_\tau(\theta) d\theta} \to \infty$$

as $\tau \to \infty$. Derive a contradiction.

Admissibility of the MLE in the case $n = 2$ can also be proved using this approach via the Bayes risk. The analysis is more involved however, since using conjugate Gaussian priors as in the case $n = 1$ does not work. See Problem 4.5 on p. 398 of Lehmann and Casella (1998).

**Exercise 3.6** (Smoothness and decay of Fourier coefficients). Let $f$ be a function that has compact support in $(0, 1)$ and that is square integrable on $[0, 1]$. Let $e_k$, $k \in \mathbb{Z}$, be the standard (complex) Fourier basis of $L^2[0, 1]$, defined by $e_k(x) = \exp(i2\pi kx)$. For $\beta \in \mathbb{N}$, show that $f$ has $\beta$ square integrable, almost everywhere defined derivatives if and only if its Fourier coefficients $f_k$ satisfy $\sum k^{2\beta}|f_k|^2 < \infty$.

**Exercise 3.7** (Weak duality). In Example 3.2.4, show that the weak duality (3.6) holds.

**Exercise 3.8** (MAP estimator versus posterior mean). Show by direct computation that for the prior $\pi$ considered in the last part of Example 3.2.5, the MAP estimator (3.8) coincides with the posterior mean (3.9).

**Exercise 3.9** (Kernel density estimation). Using the R function `density`, fit a kernel density estimator to the galaxy dataset of Figure 1.2 (which can be found in the package `MASS`). Plot the result over a histogram of the data. Experiment with different bandwidths and kernels. Do the same with the dataset `faithful$eruptions`. 
Exercise 3.10 (Grenander estimator). The Grenander estimator is implemented for instance in the R function `grenander` in the package `fdrtool`. Try it out on a simulated sample of size 200 from the standard exponential distribution. Compare to kernel density estimates for the same dataset.

Exercise 3.11 (The lasso). Verify the expression for the lasso estimator in the normal means model, given in Example 3.3.4. Plot the soft thresholding function $h_\lambda$.

Exercise 3.12 ($\ell^0$-penalization). Verify the expression for the $\ell^0$-penalized estimator in the normal means model, given in Example 3.3.5. Plot the corresponding hard thresholding function.

Exercise 3.13 (Data reduction as regularization). Consider a general statistical experiment with an observation of the form $(X,Y)$ in some Euclidean space $\mathbb{R}^k \times \mathbb{R}^m$ (for simplicity). Suppose we have a parametrized statistical model $(P_\theta : \theta \in \Theta)$ for some parameter space $\Theta$ and assume that every $P_\theta$ admits a Lebesgue density. Suppose that under each $P_\theta$ both the distribution of $Y$ and the conditional distribution of $X$ given $Y$ have positive densities on, respectively, $\mathbb{R}^m$ and $\mathbb{R}^k$ as well.

Show that the MLE for $\theta$ based on the reduced data $Y$ can be viewed as a penalized MLE for the full data $(X,Y)$. Show that the penalization is non-trivial if and only if $Y$ is not a sufficient statistic for $\theta$.

Exercise 3.14 (MNIST digit prediction using penalized logistic regression). In Exercise 1.3 we considered the problem of predicting the correct label of images of handwritten 0’s and 1’s from the MNIST dataset. We fitted two types of logistic regression models to the training dataset shown in Figure 1.3. Then we used these models to predict the labels of the remaining 2115 images in the dataset. We compared logistic regression using the whole images and using just the $2 \times 2$ sub-images in the center. It turned out that the latter works better.

By Exercise 3.13 we can view using the sub-images instead of the full images as a form of regularization. We can also combine logistic regression with other penalties, for instance ridge, i.e. $\ell^2$, or lasso-type, i.e. $\ell^1$, penalties. The function `glmnet` in the R package with the same name implements logistic regression with these penalties. Use this to fit an $\ell^1$-penalized logistic regression model to the training data of Figure 1.3, using the default choice of the tuning parameter. Use this model to predict the labels of the remaining 2115 images in the dataset. How does the prediction performance compare to that of the earlier methods?
Exercise 3.15 (Rates for kernel density estimation). Mimic the approach of Example 3.4.3 to derive the best rate of convergence for the kernel density estimator under the condition that the true density belongs to the Hölder space $C^\beta[0, 1]$ for $\beta \in (0, 2]$. Do it both for the pointwise risk and for the $L^2$-risk, or mean integrated squared error (MISE)

$$E_p \int_0^1 (\hat{p}_h(x) - p(x))^2 \, dx.$$  

What is the optimal choice of the bandwidth?

Exercise 3.16 (Rates for Bayesian procedures). Consider the Bayes procedure of Example 3.2.5, with prior variances given by $\tau_j^2 = j^{-1-2\alpha}$ for some $\alpha > 0$. Show that for this choice, the squared $L^2$-risk of the posterior mean (3.9) satisfies, for $\beta, R > 0$,

$$\sup_{f \in H^\beta_R[0, 1]} E_f \| \hat{f} - f \|_2^2 \lesssim$$

Which choice of $\alpha$ gives the best upper bound?

Exercise 3.17 (Gaussian tail bound). Let $Z$ be a standard normal random variable. Show that for all $p > 0$ there exists a constant $C_p > 0$ such that for all $a \geq 1$,

$$EZ^p 1_{Z > a} \leq C_p a^{p-1} e^{-\frac{1}{2} a^2}.$$  

Exercise 3.18 (Rate of convergence for the lasso). Investigate which rate the lasso can achieve for $s$-sparse vectors in the normal means model. What is the appropriate oracle choice of the tuning parameter?
Chapter 4

Lower bounds for testing

4.1 Statistical tests in the white noise model

In this chapter we consider testing problems in the context of the signal in white noise model (2.1). Specifically, we assume that we observe a stochastic process \( X = (X_t : t \in [0, 1]) \) satisfying

\[
\mathrm{d}X_t = f(t) \, \mathrm{d}t + \frac{1}{\sqrt{n}} \, \mathrm{d}W_t
\]

for some \( f \in L^2[0, 1] \). We mostly consider tests for a simple null hypothesis of the form \( H_0 : f = f_0 \) for some fixed \( f_0 \in L^2[0, 1] \), against an alternative \( H_1 \) which can be either simple or composite. Formally, a (non-randomized) test in this context is a \( \{0, 1\} \)-valued random variable \( \varphi \) that is a function of the observation \( X \). The variable \( \varphi = \varphi(X) \) should simply be viewed as a decision rule. If we observe \( X \), then \( \varphi(X) \) says whether we should reject the null or not. If \( \varphi(X) = 0 \) we don’t reject \( H_0 \), and if \( \varphi(X) = 1 \), then we reject \( H_0 \) and accept \( H_1 \). If the hypotheses are of the form \( H_0 : f = f_0 \) and \( H_1 : f \in \mathcal{F} \), then in this formalism for describing tests, the error of the first kind, i.e. the probability that \( H_0 \) is rejected while it is true, is equal to \( \mathbb{E}_{f_0} \varphi \). Similarly, the error of the second kind is \( \sup_{f \in \mathcal{F}} \, \mathbb{E}_f (1 - \varphi) \). The total error of the test, or simply the error of the test \( \varphi \) is defined as the sum \( \mathbb{E}_{f_0} \varphi + \sup_{f \in \mathcal{F}} \, \mathbb{E}_f (1 - \varphi) \) of the two types of errors. Recall that for \( \alpha \in (0, 1) \) a test is said to be of level \( \alpha \) if the error of the first kind is bounded by \( \alpha \). In the case of a simple null as above this means that \( \mathbb{E}_{f_0} \varphi \leq \alpha \).

We will be interested in the existence of consistent tests. A sequence of tests \( \varphi_n \) is called consistent if the total error of tests, as just defined, vanishes as \( n \to \infty \). Existence of a consistent test means that asymptotically, we can distinguish observations from the null and from the alternative hypothesis without error. If, on the other hand, the error of a sequence of tests is bounded away from 0, then using these tests it is not possible to distinguish the hypotheses. If that holds uniformly for all tests \( \varphi \), then there does not exists any test that can asymptotically discriminate between the hypotheses without error.
The fundamental difficulty of the testing problem \( H_0 : f = f_0 \) against \( H_1 : f \in \mathcal{F} \) can therefore be assessed by studying lower bounds for the minimax risk for testing \( f_0 \) against \( \mathcal{F} \), defined as

\[
\inf_\varphi \left( E_{f_0} \varphi + \sup_{f \in \mathcal{F}} E_f (1 - \varphi) \right),
\]

where the infimum is over all tests \( \varphi \). The larger \( \mathcal{F} \) is, the larger this quantity becomes. It is also intuitively clear that the ‘closer’ \( \mathcal{F} \) is to \( f_0 \), the more difficult the testing problem is. In this chapter we derive results that make these intuitive statements mathematically precise and that allow us to study the difficulty of certain specific testing problems in the white noise model.

Testing problems of the form \( H_0 : f = 0 \) against \( H_1 : f \in \mathcal{F} \) can be viewed as signal detection problems. We will see that for a consistent test to exist for such a problem, the set of signals \( \mathcal{F} \) has to be sufficiently separated from 0. The idea is simply that for consistent testing for the presence of a signal to be possible, the signal should be ‘large’ or ‘strong’ enough in some sense. The concrete form that this requirement can take depends on the type of signals that are considered. In this chapter we will consider detection of Hölder or Sobolev smooth signals. It turns out that the strength of the signal that is necessary for consistent testing to be possible in these cases, depends on the degree of smoothness and on the type of norm that is considered to quantify the strength.

### 4.2 Two simple hypotheses

Suppose we want to test two simple hypotheses \( H_0 : f = f_0 \) and \( H_1 : f = f_1 \) against each other. There exists a consistent sequence of tests for these hypotheses if and only if the infimum over all possible tests \( \varphi \) of the total error \( E_{f_0} \varphi + E_{f_1} (1 - \varphi) \) vanishes as \( n \to \infty \). Using the Neyman-Pearson lemma we can actually compute the minimax risk for testing \( f_0 \) against \( f_1 \) explicitly.

We first recall the Neyman-Pearson lemma, which takes the following form in the white noise model.

**Theorem 4.2.1 (Neyman-Pearson).** Consider an observation \( X \) from the white noise model (2.1) and two simple hypotheses \( H_0 : f = f_0, H_1 : f = f_1 \), for \( f_0, f_1 \in L^2[0,1] \). For \( \alpha \in (0,1) \), the most powerful test \( \varphi \) of level \( \alpha \), i.e. the test of level \( \alpha \) for which \( E_{f_1} \varphi \) is maximal, is the likelihood ratio test

\[
\varphi = \frac{1}{p_{f_1}(X)/p_{f_0}(X) > c_\alpha},
\]

where \( p_f \) is the density given by (2.7) and

\[
c_\alpha = \exp \left( \sqrt{n} \| f_1 - f_0 \|_2 \xi_{1-\alpha} - \frac{1}{2} n \| f_1 - f_0 \|_2^2 \right),
\]

with \( \xi_{1-\alpha} \) the \( 1 - \alpha \) quantile of the standard normal distribution.
Proof. We first note that the test indeed has level $\alpha$. To see this, we use the expression (2.6) for the likelihood ratio. It implies that for $W$ a $P_{f_0}$-Brownian motion,

$$E_{f_0} \varphi = P_{f_0} \left( e^{\sqrt{n} \int (f_1 - f_0) dW - \frac{1}{2} n \|f_1 - f_0\|^2} > c_\alpha \right)$$

$$= P_{f_0} \left( \int (f_1 - f_0) dW > \|f_1 - f_0\|_2 \xi_{1-\alpha} \right)$$

$$= \alpha,$$

where the last equality follows from the properties of the Wiener integral (see Theorem 2.2.2).

Now let $\varphi'$ be another test of level $\alpha$. Then by separately considering the events that $\varphi = 1$ and $\varphi = 0$ we see that it almost surely holds that

$$(\varphi' - \varphi)(c_\alpha p_{f_0}(X) - p_{f_1}(X)) \geq 0.$$  

By taking the expectation under $P_0$ and using the fact that $p_f = dP_f/dP_0$ it follows that

$$c_\alpha (E_{f_0} \varphi' - E_{f_0} \varphi) - (E_{f_1} \varphi' - E_{f_1} \varphi) \geq 0.$$  

Rearranging this and using that $E_{f_0} \varphi' \leq \alpha = E_{f_0} \varphi$ we obtain the desired inequality $E_{f_1} \varphi' \leq E_{f_1} \varphi$. \hfill \Box

We can now compute the minimax risk for testing two simple hypothesis against each other. As usual, $\Phi$ denotes the standard normal distribution function.

**Proposition 4.2.2.** For all $f_0, f_1 \in L^2[0,1]$,

$$\frac{1}{2} \inf_{\varphi} (E_{f_0} \varphi + E_{f_1} (1 - \varphi)) = 1 - \Phi \left( \frac{1}{2} \sqrt{n} \|f_1 - f_0\|_2 \right),$$

where the infimum is over all tests $\varphi$.

**Proof.** By writing the collection of all tests as a union of collections of tests of a fixed level we see that

$$\inf_{\varphi} (E_{f_0} \varphi + E_{f_1} (1 - \varphi)) = \inf_{\alpha \in (0,1)} \left( \alpha + \inf_{\varphi \text{ of level } \alpha} E_{f_1} (1 - \varphi) \right).$$

By the Neyman-Pearson lemma, the most powerful test of level $\alpha$ is the likelihood ratio test (4.1). For that test we have, by (2.6) and with $W$ a $P_{f_1}$-Brownian motion, that

$$E_{f_1} (1 - \varphi) = P_{f_1} (p_{f_0}(X) / p_{f_1}(X) \geq 1/c_\alpha)$$

$$= P_{f_1} (\log(p_{f_0}(X) / p_{f_1}(X)) \geq -\log c_\alpha)$$

$$= P_{f_1} \left( \sqrt{n} \int (f_0 - f_1) dW \geq n \|f_1 - f_0\|^2_2 - \xi_{1-\alpha} \sqrt{n} \|f_1 - f_0\|_2 \right)$$

$$= 1 - \Phi \left( \sqrt{n} \|f_1 - f_0\|_2 - \xi_{1-\alpha} \right).$$
Hence we obtain
\[ \inf_{\varphi}(E_{f_0}\varphi + E_{f_1}(1 - \varphi)) = \inf_{\alpha \in (0,1)} \left( \alpha + 1 - \Phi \left( \sqrt{n} \| f_1 - f_0 \|_2 - \xi_1 - \alpha \right) \right). \]

The proof is completed by minimizing over \( \alpha \) (see Exercise 4.1).

The proposition implies that two simple hypotheses corresponding to functions \( f_0 \) and \( f_1 \) are asymptotically distinguishable by a consistent sequence of tests if and only if \( \sqrt{n} \| f_1 - f_0 \|_2 \to \infty \). The next examples treat some concrete testing problems.

**Example 4.2.3** (Detecting a spike at a fixed point). Suppose we want to test \( H_0 : f = 0 \) against \( H_1 : f = f_\sigma \), where \( f_\sigma \) is a ‘spike’ at 1/2 given by
\[ f_\sigma(x) = K \left( \frac{x - 1/2}{\sigma} \right), \quad x \in [0, 1], \]
where \( K \) is the tent-shaped function \( K(x) = (1 - |x|)_{|x| \leq 1} \) and \( \sigma > 0 \) is a parameter that determines how narrow the spike is. For the \( L^2 \)-norm of \( f_\sigma \) we have
\[ \| f_\sigma \|_2^2 = \int_0^1 K^2 \left( \frac{x - 1/2}{\sigma} \right) \, dx = \sigma \int_{-1/(2\sigma)}^{1/(2\sigma)} K^2(x) \, dx. \]
Due to the compactness of the support of \( K \), \( \| f_\sigma \|_2^2 \) behaves like a constant times \( \sigma \) for \( \sigma \to 0 \), hence \( \sqrt{n} \| f_\sigma \|_2 \sim \text{const} \times \sqrt{n}\sigma \). For the existence of a consistent test for the presence of the spike this quantity should tend to +\( \infty \). We see that this is the case if and only if \( n\sigma \to \infty \). Hence, a spike with compact support is only detectable in Gaussian white noise if its width is of larger order than the square \( 1/n \) of the ‘magnitude’ of the noise.

Whether certain smooth bumps can be detected in Gaussian white noise depends on the relation between the smoothness and the size of the bump. The smoother the bump, the easier it is to detect it.

**Example 4.2.4** (Detecting a smooth bump at a fixed point). Consider the function
\[ f_\sigma(x) = \sigma^\beta K \left( \frac{x - 1/2}{\sigma} \right), \quad x \in [0, 1], \]
where \( K \) is \( C^\infty \) and has compact support, and \( \beta, \sigma > 0 \). This function has \( C^\beta \)-norm of the order 1, in the sense that the supremum of the \( C^\beta \)-norm over bounded \( \sigma \) is bounded, and its squared \( L^2 \)-norm of the order \( \sigma^{1+2\beta} \) (see Exercise 4.2). So the spike is consistently detectable if and only if \( n\sigma^{1+2\beta} \to \infty \), that is, if and only the height \( f_\sigma(1/2) \) of the bump is of of larger order than \( n^{-\beta/(1+2\beta)} \).
The last example implies that we are testing $H_0 : f = 0$ against an alternative that contains a smooth bump at a fixed location with height bounded by $n^{-\beta/(1+2\beta)}$, then no consistent test exists. See Exercise 4.3 for an example.

4.3 Composite alternative hypothesis

In this section we consider the problem of testing a simple null hypothesis $H_0 : f = f_0$ against a composite alternative of the form $H_1 : f \in \{f_1, \ldots, f_k\}$. The error of a test $\varphi$ is in this case given by $E_{f_0}\varphi + \max_{i=1,\ldots,k} E_{f_i} (1-\varphi)$. Again we are interested in conditions for the existence or non-existence of consistent tests. It is in the present setting no longer possible to derive an explicit expression for the minimax testing risk. It is however still possible to derive useful sufficient conditions.

It is intuitively clear perhaps that it should be possible to relate the difficulty of the testing problem to the average of the likelihood ratios $dP_{f_i}/dP_{f_0}$. Indeed, if this average is ‘large’, then one of the likelihoods $dP_{f_i}/dP_{f_0}$ must be large as well. This means that data arising from the model with signal $f_0$ can with substantial probability be generated by the model with signal $f_i$ too, and hence the two hypotheses will be difficult to distinguish.

The following lemma make this reasoning precise and asserts that consistent testing is impossible if under the null, the average likelihood is bounded away from 0 with sufficient probability.

**Lemma 4.3.1.** Let $f_0, f_1, \ldots, f_k \in L^2[0,1]$. If for some $c, \varepsilon \in (0,1)$, independent of $n$, it holds that

$$P_{f_0}\left(\frac{1}{k} \sum_{i=1}^k \frac{dP_{f_i}}{dP_{f_0}} > c\right) \geq \varepsilon,$$  \hspace{1cm} (4.2)

then there exists no consistent test for the hypotheses $H_0 : f = f_0$ against $H_1 : f \in \{f_1, \ldots, f_k\}$. A sufficient condition for (4.2) is that

$$E_{f_0} \left(\frac{1}{k} \sum_{i=1}^k \frac{dP_{f_i}}{dP_{f_0}} - 1\right)^2 \leq \delta$$

for some $\delta \in (0,1)$, independent of $n$.

**Proof.** Observe that for an arbitrary test $\varphi$,

$$\max_{i=1,\ldots,k} E_{f_i} (1-\varphi) \geq \frac{1}{k} \sum_{i=1}^k E_{f_i} (1-\varphi) = \frac{1}{k} \sum_{i=1}^k E_{f_0} \frac{dP_{f_i}}{dP_{f_0}} (1-\varphi) = E_{f_0} L(1-\varphi),$$
where $L$ is the average likelihood defined by $L = k^{-1} \sum_{i=1}^{k} dP_{f_i}/dP_{f_0}$. Hence for $c \in (0,1)$ the testing error $E_{f_0}(\varphi + \max_{i=1, \ldots, k} E_{f_i}(1 - \varphi))$ is bounded from below by

$$E_{f_0}(\varphi + L(1 - \varphi)) = E_{f_0}(1_{\varphi=1} + L1_{\varphi=0}) \geq E_{f_0}(1_{\varphi=1} + L1_{\varphi=0})1_{L>c} \geq cP_{f_0}(L > c).$$

It follows that under the condition of the lemma, the error of any test is bounded away from 0.

To prove the second statement of the lemma note that $P_{f_0}(L > c) \geq 1 - P_{f_0}(|L - 1| \geq 1 - c) \geq 1 - \frac{E_{f_0}(L - 1)^2}{(1 - c)^2}$, by Markov’s inequality. If $\delta \in (0,1)$, there is a $c \in (0,1)$ such that $\delta < (1 - c)^2$. This completes the proof. \qed

The preceding lemma is general in that it does not use the fact that we are in the setting of the white noise model. The following proposition gives conditions specific to our situation.

**Proposition 4.3.2.** Let $f_0, f_1, \ldots, f_k \in L^2[0,1]$. If for some $\delta \in (0,1)$, independent of $n$,

$$\frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} e^{n\langle f_i - f_0, f_j - f_0 \rangle} \leq 1 + \delta,$$

then there exists no consistent test for the hypotheses $H_0 : f = f_0$ against $H_1 : f \in \{f_1, \ldots, f_k\}$.

**Proof.** Denote the average likelihood by $L$ again. By (2.6) we have

$$L^2 = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} e^{\sqrt{n} \int (f_i - f_0) dW - \frac{1}{2} n\|f_i - f_0\|^2} e^{\sqrt{n} \int (f_j - f_0) dW - \frac{1}{2} n\|f_j - f_0\|^2},$$

where $W$ is a $P_{f_0}$-Brownian motion. It follows from the properties of the Wiener integral that the $P_{f_0}$-expectation of term $(i, j)$ in the sum equals $\exp(n \langle f_i - f_0, f_j - f_0 \rangle)$ (check!). Hence

$$E_{f_0}(L - 1)^2 = E_{f_0}L^2 - 1 = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} e^{n\langle f_i - f_0, f_j - f_0 \rangle} - 1,$$

and the proposition follows from Lemma 4.3.1. \qed
Observe that for \( k = 1 \) the proposition implies that there exists no consistent test for the simple hypotheses \( H_0 : f = f_0 \) against \( H_1 : f = f_1 \) if \( n\|f_1 - f_0\|^2 \leq \log(1 + \delta) \). This matches the sharp result that we obtained above from Proposition 4.2.2.

The proof of Proposition 4.3.2 shows that the sum in (4.3) is always bounded from below by 1 (this can also be seen directly using Jensen’s inequality, for instance). The sum can be decomposed as

\[
\frac{1}{k^2} \sum_{i=1}^{k} e^{n\|f_i - f_0\|^2} + \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j \neq i} e^{n(f_i - f_0, f_j - f_0)}.
\]

The first term quantifies the \( L^2 \)-distance of the functions in the alternative to \( f_0 \). It is small if all the \( f_i \) are close to \( f_0 \). The second term somehow quantifies the ‘dimension’ of the alternative. It is small if many of the ‘directions’ \( f_i - f_0 \) are (nearly) orthogonal to each other. So we see that the testing problem \( H_0 : f = f_0 \) against \( H_1 : f \in F \) becomes hard if the alternative \( F \) contains many functions that are in \( L^2 \)-sense close to \( f_0 \) and that are at orthogonal directions from \( f_0 \). See Exercise 4.4 for a precise result of this type.

**Example 4.3.3** (Detecting a smooth bump at one of many locations). Fix \( 0 < u < v < 1 \). For \( x_1, \ldots, x_k \in (u, v) \), consider the bump functions \( f_{\sigma,i} \) defined by

\[
f_{\sigma,i}(x) = \varepsilon \sigma^\beta K\left(\frac{x - x_i}{\sigma}\right), \quad x \in [0, 1],
\]

where \( K \) is a \( C^\infty \) function with compact support and \( \beta, \sigma, \varepsilon > 0 \). As in Example 4.2.4, the \( C^\beta \)-norm of these functions is uniformly bounded by 1 if \( \varepsilon \) is small enough and their squared \( L^2 \)-norm is bounded by \( C\varepsilon^2 \sigma^{-1+2\beta} \) for some \( C > 0 \). Since \( K \) has compact support, we can position \( k \sim 1/\sigma \) points \( x_i \) in \( (u, v) \) in such a way that the \( k \) resulting bump functions have disjoint supports, hence are orthogonal in \( L^2[0,1] \). It follows from Proposition 4.3.2, see also Exercise 4.4, there exists no consistent test for the hypotheses \( H_0 : f = 0 \) against \( H_1 : f \in \{f_{\sigma,1}, \ldots, f_{\sigma,k}\} \) if

\[
s\varepsilon e^{nC\varepsilon^2 \sigma^{-1+2\beta}} \leq \delta
\]

for some \( \delta \in (0,1) \). This condition is fulfilled for \( \varepsilon \) small enough and \( \sigma = (n/\log n)^{-1/(1+2\beta)} \), which corresponds to \( k \sim (n/\log n)^{1/(1+2\beta)} \).

So to be able to detect a bump at one of the locations \( x_i \) in the interval \([u, v]\), the height \( f_{\sigma,i}(x_i) \) of the bumps has to be a large enough multiple of \( (n/\log n)^{-\beta/(1+2\beta)} \). This is a (slightly) more stringent condition than the condition that we obtained in Example 4.2.4 for detecting a single bump, reflecting the fact that statistically it is more difficult to detect a bump at one of many possible locations than at a single fixed location.

Using this last example we can obtain the uniform norm analogue of the result of Exercise 4.3, see Exercise 4.5. Our next goal is to understand how
the results of Examples 4.2.4 and 4.3.3 and Exercises 4.3 and 4.5 change if we replace Hölder regularity by Sobolev regularity and the pointwise and uniform risks by $L^2$-risk. It will turn out that the result looks rather different in this case. To prove this we need a different construction than the orthogonal bump functions.

For concreteness we fix an orthonormal basis $e_j$ of $L^2[0,1]$. The Fourier coefficients of a function $f \in L^2[0,1]$ are denoted by $f_j$, i.e. $f_j = \langle f, e_j \rangle$. For $\beta > 0$, the associated Sobolev norm $\|f\|_{H^\beta}$ of $f$ is defined by

$$\|f\|_{H^\beta}^2 = \sum_j f_j^2 j^{2\beta}.$$  

Recall that the Sobolev space $H^\beta[0,1]$ is the space of all functions in $L^2[0,1]$ for which this norm is finite and that the unit ball of $H^\beta[0,1]$ is denoted by $H^\beta_1[0,1]$. Following the same line of reasoning as before we now construct a finite collection of functions in $H^\beta_1[0,1]$ that are difficult to distinguish from 0.

**Example 4.3.4 (Detecting Sobolev smooth signals).** Let $k \sim n^{1/(1/2+2\beta)}$. For $\varepsilon > 0$ and $\theta \in \{-1,+1\}^k$, define

$$f_\theta = \varepsilon k^{-1/2-\beta} \sum_{j=1}^k \theta_j e_j.$$  

We have $\langle f_\theta, f_\psi \rangle = \varepsilon^2 k^{-1-2\beta} \sum \theta_j \psi_j$. In particular, $\|f_\theta\|_2^2 = \varepsilon^2 k^{-2\beta} \sim \varepsilon^2 n^{-3\beta/(1/2+2\beta)}$. For the Sobolev norm we have

$$\|f_\theta\|_{H^\beta}^2 = \varepsilon^2 k^{-1-2\beta} \sum_{j=1}^k j^{2\beta} \theta_j^2 = \text{const.} \times \varepsilon^2.$$  

So for $\varepsilon$ small enough all functions $f_\theta$ belong to the Sobolev ball $H^\beta_1[0,1]$ and their squared $L^2$-norm is of the order $n^{-3\beta/(1/2+2\beta)}$.

The condition (4.3) for non-testability of $H_0 : f = 0$ against $H_1 : f \in \{f_\theta : \theta \in \{-1,+1\}^k\}$ reads in this case

$$\frac{1}{4^k} \sum_{\theta} \sum_{\psi} e^{\pi \varepsilon^2 k^{-1-2\beta} \sum \theta_j \psi_j} \leq 1 + \delta$$  

for some $\delta \in (0,1)$. Now let $R_1,\ldots,R_k$ and $S_1,\ldots,S_k$ be two independent sequences of $k$ independent Rademacher variables, each variable having values $\pm 1$ with probability 1/2. Then the quantity on the left-hand side of the display equals

$$\mathbb{E} e^{\pi \varepsilon^2 k^{-1-2\beta} \sum R_j S_j} = \left(\mathbb{E} e^{\pi \varepsilon^2 k^{-1-2\beta} R_1 S_1}\right)^k$$  

(check!). But $R_1 S_1$ is again a Rademacher variable, so there exists no consistent test if

$$k \log \mathbb{E} e^{\pi \varepsilon^2 k^{-1-2\beta}} \leq \delta'$$
for some $\delta' \in (0, 1/2)$, where $R$ is a Rademacher variable. It is easily verified that this condition is fulfilled (see Exercise 4.6) if $\varepsilon$ is small enough. We conclude that it is not possible to distinguish whether the signal is 0 or one of the $f_\theta \in H^2_\beta[0, 1]$ if $\varepsilon$ is small enough. 

In Example 4.3.4 we obtain a different rate than in the preceding Examples 4.2.4 and 4.3.3, namely $n^{-\beta/(1/2+2\beta)}$ instead of $n^{-\beta/(1+2\beta)}$ (with or without a logarithmic factor). As a result, the separation relative to the $L^2$-norm that is necessary for consistent testing of $f = 0$ against Sobolev smooth alternatives is different from the pointwise or uniform separation that is necessary for consistent testing of $f = 0$ against H"older smooth alternatives. See Exercise 4.7.

4.4 Testing multiple hypotheses

So far we have been considering tests for two hypotheses $H_0$ and $H_1$, where $H_0$ is a simple hypothesis and $H_1$ might be composite. In this section we study tests for the situation that we have multiple simple hypotheses $H_0, H_1, \ldots, H_k$ of the form $H_i : f = f_i$, for some collection of functions $\{f_0, f_1, \ldots, f_k\} \subset L^2[0, 1]$. Now the problem is not to only decide whether $f_0$ is the truth or not, but which of the $f_i$’s is in fact the true parameter. In this setting a test is formally a $\{0, 1, \ldots, k\}$-valued measurable function $\varphi = \varphi(X)$ of the observation $X$ and we want all of the error probabilities $P_{f_i}(\varphi \neq i)$ to be small. In this situation we say that there exists a consistent test if

$$\inf_{\varphi} \max_i P_{f_i}(\varphi \neq i) \to 0$$

as $n \to \infty$, where the infimum is over all possible tests. It is intuitively clear that this might be more demanding than testing $f_0$ against the union of the other $f_i$, and that we might get different results than before.

The following lemma gives a sufficient condition for non-testability in terms of likelihood ratios, in the spirit of Lemma 4.3.1.

Lemma 4.4.1. Let $f_0, f_1, \ldots, f_k \in L^2[0, 1]$. If for some $i \in \{0, 1, \ldots, k\}$ and $\varepsilon \in (0, 1)$, independent of $n$, and $c > 0$ such that $ck$ is bounded away from 0 it holds that

$$\frac{1}{k} \sum_{j \neq i} P_{f_j} \left( \frac{dP_{f_i}}{dP_{f_j}} \geq c \right) \geq \varepsilon,$$

then there exists no consistent multiple hypothesis test for the hypotheses $H_0 : f = f_0, H_1 : f = f_1, \ldots, H_k : f = f_k$.
Proof. For \( j = 1, \ldots, k \), define the events \( A_j = \{ dP_{f_0}/dP_{f_j} \geq c \} \). Then we have

\[
P_{f_0}(\varphi \neq 0) = \sum_{j \neq 0} P_{f_0}(\varphi = j) \\
= \sum_{j \neq 0} E_{f_j} \frac{dP_{f_0}}{dP_{f_j}} 1_{\varphi = j} \\
\geq c \sum_{j \neq 0} P_{f_j}(\varphi = j, A_j) \\
\geq c \sum_{j \neq 0} \left( P_{f_j}(\varphi = j) - P_{f_j}(A^c_j) \right).
\]

Writing

\[
s = \frac{1}{k} \sum_{j=1}^k P_{f_j}(\varphi = j), \quad t = \frac{1}{k} \sum_{j=1}^k P_{f_j}\left( \frac{dP_{f_0}}{dP_{f_j}} < c \right),
\]

it follows that

\[
\max_j P_{f_j}(\varphi \neq j) = \max \left\{ P_{f_0}(\varphi \neq 0), \max_{j \neq 0} P_{f_j}(\varphi \neq j) \right\} \\
\geq \max \left\{ P_{f_0}(\varphi \neq 0), \frac{1}{k} \sum_{j=1}^k P_{f_j}(\varphi \neq j) \right\} \\
\geq \max \left\{ ck(s-t), 1-s \right\}.
\]

By considering the functions \( s \mapsto ck(s-t) \) and \( s \mapsto 1-s \) on \((0,1)\) we see that the maximum is bounded from below by \((1-t)ck/(1+ck)\) (check!), i.e.

\[
\max_j P_{f_j}(\varphi \neq j) \geq \frac{ck}{1+ck} \frac{1}{k} \sum_{f \neq f_0} P_f\left( \frac{dP_{f_0}}{dP_f} \geq c \right)
\]

The proof of the lemma is completed by noting that we can replace \( f_0 \) by any other of the \( f_i \)'s.

The following proposition translates the requirements of the general lemma into specific conditions in the setting of the white noise model. Note that the condition is very close to the condition obtained Exercise 4.4 for non-testability of 0 against a composite alternative consisting of \( k \) orthogonal functions. In the present setting of multiple hypothesis testing, there is no orthogonality requirement.
Proposition 4.4.2. Let $f_0, f_1, \ldots, f_k \in L^2[0,1]$. If there exist a constant $c > 0$, independent of $n$, such that

$$n \max_{i,j} \|f_i - f_j\|_2^2 \leq c + 2 \log k$$

for $n$ large enough, then there exists no consistent multiple hypothesis test for the hypotheses $H_0 : f = f_0$, $H_1 : f = f_1, \ldots, H_k : f = f_k$.

Proof. By lower bounding the maximum of a sequence by the mean we see from the preceding lemma that a sufficient condition for non-testability is that for some $c > 0$ such that $ck$ is bounded away from 0 it holds that

$$\frac{1}{k(k+1)} \sum_i \sum_{j \neq i} P_{f_i} \left( \frac{dP_{f_j}}{dP_{f_i}} \geq c \right)$$

is bounded away from 0. By the properties of the Wiener integral (see Theorem 2.2.2),

$$P_{f_i} \left( \frac{dP_{f_j}}{dP_{f_i}} \geq c \right) = P_{f_i} \left( \log \frac{dP_{f_j}}{dP_{f_i}} \geq \log c \right) = 1 - \Phi \left( \frac{\frac{1}{n} \|f_i - f_j\|_2^2 + \log c}{\sqrt{n} \|f_i - f_j\|_2} \right).$$

For $\log c = -\max_{i,j} n \|f_i - f_j\|_2^2/2$ this is bounded from below by 1/2, and therefore also

$$\frac{1}{k(k+1)} \sum_i \sum_{j \neq i} P_{f_i} \left( \frac{dP_{f_j}}{dP_{f_i}} \geq c \right) \geq 1/2.$$

For $c$ as defined above, the condition that $ck$ is bounded away from 0 is exactly the condition of the proposition. \qed

The following example shows that the condition for non-testability in the present multiple hypothesis setting is indeed less restrictive than what we had before. Compare with Example 4.3.4.

Example 4.4.3 (Distinguishing Sobolev smooth functions). Set $k = n^{1/(1+2\beta)}$ and consider the first $k$ elements $e_1, \ldots, e_k$ of an orthonormal basis $e_i$ of $L^2[0,1]$. For $\varepsilon > 0$ and $\theta \in \{-1,+1\}^k$, define

$$f_{\theta} = \varepsilon k^{-1/2-\beta} \sum_{i=1}^k \theta_i e_i.$$

We have

$$\|f_{\theta} - f_{\psi}\|_2^2 = \varepsilon^2 k^{-1-2\beta} \|\theta - \psi\|_2^2.$$

Since $\theta, \psi$ are vectors of $\pm 1$’s, their squared distance equals 4 times the number $d_{\text{ham}}(\theta, \psi)$ of indices at which they differ. In other words, $\|f_{\theta} - f_{\psi}\|_2^2 = \varepsilon^2 k^{-1-2\beta} d_{\text{ham}}(\theta, \psi)^2$. 

We also see that \( \| f_\theta \|_2^2 = \varepsilon^2 k^{-2\beta} = \varepsilon^2 n^{-2\beta/(1+2\beta)} \).
Moreover, for the \( \beta \)-Sobolev norm (relative to the basis \( e_i \)) we have
\[
\| f_\theta \|_{H^\beta}^2 = \varepsilon^2 k^{-1-2\beta} \sum_{i=1}^k i^{2\beta} \theta_i^2 = \text{const.} \times \varepsilon^2.
\]
We have constructed \( 2^k \) different functions and
\[
\max_{\theta, \psi} \| f_\theta - f_\psi \|_2^2 = 4\varepsilon^2 nk^{-2\beta} = 4\varepsilon^2 k.
\]
This implies that the condition of Proposition 4.4.2 is fulfilled if \( \varepsilon \) is small enough.

It is interesting to compare this example with the result of Exercise 4.7. By part (ii) of that exercise, we can consistently test whether there is a non-zero signal that belongs to the Sobolev ball \( H^\beta_1[0, 1] \) as soon as that signal has \( L^2 \)-norm of larger order than \( n^{-\beta/(1/2+2\beta)} \). Example 4.4.3 shows however that there are signals in \( H^\beta_1[0, 1] \) with \( L^2 \)-norm between \( n^{-\beta/(1/2+2\beta)} \) and \( n^{-\beta/(1+2\beta)} \) such that we can not consistently distinguish between them. In other words, there are collections of signals for which we can consistently decide that one of them generated the data, but for which we can not say which one of them did so. This hints at the fact that in this setting, signal detection is in some sense easier than signal estimation. The latter will be studied in the next chapter.
4.5 Exercises

Exercise 4.1. Complete the proof of Proposition 4.2.2.

Exercise 4.2 ($L^2$- and $C^\beta$-norms of smooth bump functions). Prove the claims about the $C^\beta$- and $L^2$-norms of the function $f_\sigma$ in Example 4.2.4.

Exercise 4.3 (Testing against Hölder alternatives, pointwise separation). Let $\beta > 0$. Let $C^\beta_1[0, 1]$ be the unit ball of the Hölder space $C^\beta[0, 1]$.

(i) Show that in the white noise model, there exists no consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in C^\beta_1[0, 1] \setminus \{0\}$.

(ii) Let $t_0 \in (0, 1)$. Show that if $r_n$ is of smaller order than $n^{-\beta/(1+2\beta)}$, then there exists no consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{f \in C^\beta_1[0, 1] : |f(t_0)| \geq r_n\}$.

(iii) Let $t_0 \in (0, 1)$. Show that if $r_n$ is of larger order than $n^{-\beta/(1+2\beta)}$, then there does exist a consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{f \in C^\beta_1[0, 1] : |f(t_0)| \geq r_n\}$. (Hint: use the estimator of Example 3.4.3 as test statistic and use the result (3.12) to bound the errors of the test.)

Exercise 4.4. Show that in the white noise model there exists no consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in F_n$ if $F_n$ contains $k_n$ different orthogonal functions $f_i$ such that $k_n^{-1} \exp(n \max_i \|f_i\|^2_2) \leq \delta$ for some $\delta \in (0, 1)$.

Exercise 4.5 (Testing against Hölder alternatives, uniform separation). Let $\beta > 0$. Let $C^\beta_1[0, 1]$ be the unit ball of the Hölder space $C^\beta[0, 1]$.

(i) Show that if $r_n$ is of smaller order than $(n/\log n)^{-\beta/(1+2\beta)}$, then there exists no consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{f \in C^\beta_1[0, 1] : \sup_{u < t < v} |f(t)| \geq r_n\}$.

(ii) Show that if $r_n$ is of larger order than $(n/\log n)^{-\beta/(1+2\beta)}$, then there does exist a consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{f \in C^\beta_1[0, 1] : \sup_{u < t < v} |f(t)| \geq r_n\}$ (Hint: use the estimator of Example 3.4.4 as test statistic and use the result (3.13) to bound the errors of the test.)

Explain the differences with the results of Exercise 4.3.
**Exercise 4.6.** Prove the last statements in Example 4.3.4.

**Exercise 4.7** (Testing against Sobolev alternatives, $L^2$-separation). Let $\beta > 0$. Let $H^\beta_1[0, 1]$ be the unit ball of the Sobolev space $H^\beta[0, 1]$.

(i) Show that if $r_n$ is of smaller order than $n^{-\beta/(1/2+2\beta)}$, then there exists no consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{ f \in H^\beta_1[0, 1] : \|f\|_2 \geq r_n \}$.

(ii) Show that if $r_n$ is of larger order than $n^{-\beta/(1/2+2\beta)}$, then there does exist a consistent test for the hypotheses $H_0 : f = 0$ against $H_1 : f \in \{ f \in H^\beta_1[0, 1] : \|f\|_2 \geq r_n \}$. (Hint: use a test that rejects $H_0$ for large values of $T = \sum_{j \leq J}(Y_j^2 - 1/n)$, for $Y_j = \int e_j \, dX$ and $J \sim n^{1/(1/2+2\beta)}$.)
Chapter 5

Lower bounds for estimation

5.1 Minimax lower bounds for estimation

After studying the fundamental possibilities and limitations in various testing problems in the white noise model in Chapter 4 we return to estimation problems in this chapter. In Chapter 3 we introduced and studied various regularization methods for nonparametric estimation of the signal $f$ from observations $X_t = (X_t : t \in [0, 1])$ satisfying

$$dX_t = f(t) \, dt + \frac{1}{\sqrt{n}} dW_t$$

We found in particular that under the assumption that $f$ is $\beta$-smooth in Hölder or Sobolev sense, there exist estimators that, perhaps up to a logarithmic factor, achieve a rate of convergence of the order $n^{-\beta/(1+2\beta)}$ relative to the pointwise, supremum or $L^2$-norm. Moreover, such results hold uniformly over Hölder or Sobolev balls. Specifically, we have the upper bounds (3.12), (3.13), (3.14) for appropriately tuned kernel or projection estimators.

The cited rate of convergence results are all risk upper bounds of the form

$$\sup_{f \in \mathcal{F}} E_f \ell(\hat{f}, f) \leq r_n,$$

where $\hat{f}$ is a specific estimator, $\mathcal{F} \subset L^2[0, 1]$ is a class of signals of interest, $\ell$ is a choice of loss function, such as pointwise loss, uniform loss, or squared $L^2$-loss, and $r_n$ is a sequence converging to 0 as $n \to \infty$. Such results quantify how well a specific estimator performs for a certain class of truths. As we have seen, the bounds typically depend both on the class $\mathcal{F}$ and on the type of loss $\ell$ that is considered.

In this chapter we consider the converse question. We ask what the best possible performance of any estimator is over a given class of signals $\mathcal{F}$. More precisely, we are interested in proving so-called minimax lower bounds for estimation of the form

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_f \ell(\hat{f}, f) \geq r_n,$$
where \( \ell \) is a again a loss function and the infimum is over all possible estimators. Such a result quantifies the fundamental difficulty of estimating a signal in the class \( \mathcal{F} \), as it implies that no estimator can have a risk of smaller order than \( r_n \), uniformly over the class \( \mathcal{F} \). We will study in particular the converse of the results (3.12), (3.13), (3.14) to investigate whether these estimator are rate-optimal in the minimax sense.

5.2 Connection to multiple hypotheses testing problems

The problem of estimating a signal in the white noise model is clearly closely related to the problem of testing multiple hypotheses as studied in Section 4.4. If the class of signals \( \mathcal{F} \) of interest is finite the problem is in fact exactly the same. But also if \( \mathcal{F} \) is not finite the connection to testing provides a useful approach to proving minimax lower bounds for estimation.

The following proposition make this precise. It assumes that the loss is of the form \( \ell = d^p \), for \( p > 0 \) and \( d \) a semimetric on \( \mathcal{F} \subset L^2[0,1] \) (i.e. a map \( d: \mathcal{F} \times \mathcal{F} \to [0,\infty) \) that is symmetric and that satisfies the triangle inequality, but not necessarily \( d(f,g) = 0 \) implies \( f = g \) for all \( f,g \in \mathcal{F} \)).

**Proposition 5.2.1.** Consider a class \( \mathcal{F} \subset L^2[0,1] \) a semimetric \( d \) on \( \mathcal{F} \) and \( p > 0 \). If there exist \( f_0, f_1, \ldots, f_k \in \mathcal{F} \) such that

(i) for all \( i \neq j \), we have \( d(f_i, f_j) \geq 2r_n \),

(ii) there exists no consistent test for the multiple hypotheses \( H_0 : f = f_0 \),

\( H_1 : f = f_1 \), \ldots , \( H_k : f = f_k \),

then there exists a constant \( c > 0 \), independent of \( n \), such that

\[
\inf_{\hat{f}} \sup_{f \in \mathcal{F}} E_f d^p(\hat{f}, f) \geq cr_n^p. \tag{5.1}
\]

**Proof.** First let \( \hat{f} \) be an arbitrary \( \mathcal{F} \)-valued estimator. Consider the \( \{0,1,\ldots,k\} \)-valued test that selects the \( f_i \) that is closest to the estimator with respect to the loss \( \ell \), i.e.

\[
\varphi = \arg\min_i d(\hat{f}, f_i),
\]

where in the case of multiple minima we select an arbitrary minimizer. By assumption (i) and the triangle inequality it holds for every \( i \) that

\[
P_{f_i}(\varphi \neq i) \leq P_{f_i}(d(\hat{f}, f_i) \geq r_n),
\]

and hence

\[
\sup_{f \in \mathcal{F}} P_f(d(\hat{f}, f) \geq r_n) \geq \max_i P_{f_i}(\varphi \neq i) \geq \inf_{\varphi} \max_i P_{f_i}(\varphi \neq i).
\]
Then by assumption (ii) it follows that there exists a constant \( c > 0 \), independent of \( n \), such that
\[
\inf_{f} \sup_{f \in \mathcal{F}} \mathbb{P}_f(d(\hat{f}, f) \geq r_n) \geq c.
\]
The observation that
\[
E_f d^p(\hat{f}, f) \geq E_f d^p(\hat{f}, f)1_{d(\hat{f}, f) \geq r_n} \geq r_n^p \mathbb{P}_f(d(\hat{f}, f) \geq r_n)
\]
completes the proof. \( \square \)

Note that the proof of the proposition shows that the power function \( x \mapsto x^p \) appearing on both sides of (5.1) may actually be replaced by any increasing function \( w : [0, \infty) \to [0, \infty) \) such that \( w(0) = 0 \). We will however only apply the proposition in the form (5.1), with \( p = 1 \) or \( p = 2 \).

If we combine the preceding proposition with the sufficient condition for non-existence of a consistent test given by Proposition 4.4.2 we obtain the following useful result.

**Proposition 5.2.2.** Suppose that there exists a finite subset \( \mathcal{G}_n \subset \mathcal{F} \) whose elements that are \( 2r_n \)-separated with respect to \( d \), with cardinality \( |\mathcal{G}_n| \) and \( L^2 \)-diameter \( \text{diam}(\mathcal{G}_n) \). If for some \( C > 0 \),
\[
e^{\frac{1}{2} n \text{diam}^2(\mathcal{G}_n)} \leq C |\mathcal{G}_n|
\]
for \( n \) large enough, then for some constant \( c > 0 \),
\[
\inf_{f} \sup_{f \in \mathcal{F}} E_f d^p(\hat{f}, f) \geq cr_n^p
\]
for all \( p > 0 \).

As a first example we consider classes of signals smoothly parametrized by a Euclidean parameter. As expected the minimax rate is of the order \( 1/\sqrt{n} \) in that case.

**Example 5.2.3** (Smooth parametric models). Suppose \( \mathcal{F} = \{ f_\theta : \theta \in \Theta \} \), where \( \Theta \subset \mathbb{R}^k \) is a set with nonempty interior. Assume the model is smooth in the sense that for every \( \theta \in \Theta \), there exist constants \( 0 < c_\theta \leq C_\theta < \infty \) such that
\[
c_\theta \| \theta - \psi \| \leq \| f_\theta - f_\psi \|_2 \leq C_\theta \| \theta - \psi \|
\]
for all \( \psi \) in a neighborhood of \( \theta \). We are interested in a lower bound for estimators \( \hat{\theta} \) relative to the Euclidean distance, so we consider the semimetric
\[
d(f_\theta, f_\psi) = \| \theta - \psi \|.
\]
Now let \( \theta_0 \) be the center of a ball that is entirely contained in \( \Theta \). Let \( \theta_1 \) be an element at Euclidean distance \( 2/(c_{\theta_0}\sqrt{n}) \) of \( \theta_0 \), so that \( f_{\theta_0} \) and \( f_{\theta_1} \) are \( 2/\sqrt{n} \)-separated with respect to \( d \). Consider the finite set \( G = \{f_{\theta_0}, f_{\theta_1}\} \). We have \( |G| = 2 \) and by the smoothness condition, the \( L^2 \)-diameter of \( G \) is bounded by \( 2C_{\theta_0}/(c_{\theta_0}\sqrt{n}) \). Hence, the condition of Proposition 5.2.2 is fulfilled for \( r_n = 1/\sqrt{n} \) and \( C \) large enough. We conclude that for some \( c > 0 \),

\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{\theta} \| \hat{\theta} - \theta \|^p \geq \frac{c}{np^{p/2}}
\]

for all \( p > 0 \). So we see that in smooth parametric models it is not possible to construct estimators that have a convergence rate that is better than \( 1/\sqrt{n} \), uniformly over the parameter space.

In parametric models in which the dependence on the parameter is not smooth such as in (5.2) it is possible to have convergence rates different from the usual \( 1/\sqrt{n} \). Exercise 5.2 considers an example in the context of the white noise model.

### 5.3 Minimax risk and Bayes risk

An alternative approach to obtaining minimax lower bounds for estimation starts from the simple but useful observation that the maximum risk of an estimator over a class of functions \( \mathcal{F} \subset L^2[0,1] \) is lower bounded by the Bayes risk of that estimator relative to any prior \( \Pi \) on \( \mathcal{F} \). Indeed, assuming the appropriate measurability, we have

\[
\sup_{f \in \mathcal{F}} \mathbb{E}_{f} \ell(\hat{f}, f) \geq \int_{\mathcal{F}} \mathbb{E}_{f} \ell(\hat{f}, f) \Pi(df).
\]

The estimator minimizing the Bayes risk on the right is, by definition, the Bayes estimator \( \bar{f} \) with respect to the loss \( \ell \) and the prior \( \Pi \). Hence, we have the bound

\[
\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_{f} \ell(\hat{f}, f) \geq \int_{\mathcal{F}} \mathbb{E}_{f} \ell(\bar{f}, f) \Pi(df) \tag{5.3}
\]

for the minimax risk. For every estimator \( \hat{f} = \hat{f}(X) \) the Bayes risk can be written as

\[
\int_{\mathcal{F}} \mathbb{E}_{f} \ell(\hat{f}(X), f) \Pi(df) = \int_{\mathcal{F}} \int \ell(\hat{f}(x), f) P_f(dx) \Pi(df)
\]

\[
= \int \int \ell(\hat{f}(x), f) \Pi(df \mid x) P(dx),
\]

where \( \Pi(df \mid x) \) is the posterior distribution and \( P \) is the marginal distribution of \( X \) in the Bayesian setup in which \( f \sim \Pi \) and \( X \mid f \sim P_f \). The inner integral in the last line of the display typically has a unique minimizer for every fixed
Given a function $x$, which is then necessarily the Bayes estimate relative to the loss $\ell$. So in that case the Bayes estimator is given by

$$\hat{f} = \arg\min_{g \in \mathcal{F}} \int_{\mathcal{F}} \ell(g, f) \Pi(df | X)$$  \hfill (5.4)

For certain combinations of loss functions $\ell$ and priors $\Pi$ the Bayes estimator can be explicitly determined in this way and the lower bound (5.3) can be used to obtain a bound for the minimax risk. We can for instance give an alternative proof of Proposition 5.2.2 using this approach.

**Alternative proof of Proposition 5.2.2.** Consider the loss function $\ell(f, g) = \frac{1}{d(f, g)} \geq r_n$. As before, observe that for the risk of the estimator $\hat{f}$ we have

$$E_f d^p(\hat{f}, f) \geq r_n E_f \ell(\hat{f}, f).$$

Hence in view of (5.3) is suffices to show that for a particular prior $\Pi$ on $\mathcal{F}$, it holds that the Bayes risk

$$\int_{\mathcal{F}} E_f \ell(\hat{f}, f) \Pi(df)$$  \hfill (5.5)

is bounded away from 0, with $\hat{f}$ the Bayes estimator for the loss $\ell$.

Denote the functions in $\mathcal{G}_n \subset \mathcal{F}$ by $f_1, f_2, \ldots, f_k$, with $k = |\mathcal{G}_n|$. We consider a discrete prior $\Pi$ on $\mathcal{F}$ that assigns mass $\pi_j$ to the function $f_j$ for $j = 1, 2, \ldots, k$. The posterior is then a discrete measure on the same points, with $f_j$ receiving posterior mass proportional to $\pi_j p_{f_j}(X)$. Moreover, we see from (5.4) that the Bayes estimator $\hat{f}$ for the loss $\ell$ is the (random) function in $\mathcal{G}_n$ which receives the most posterior mass (check!). Since the $f_j$ are $2r_n$-separated, it follows that the Bayes risk (5.5) is given by

$$\sum \pi_j P_{f_j}(d(\hat{f}, f_j) \geq r_n) = \sum \pi_j P_{f_j} (\max_i \pi_i p_{f_i}(X) > \pi_j p_{f_j}(X)).$$

Now we consider the particular prior weights

$$\pi_1 = \frac{1}{1 + (k-1)e^{-\frac{1}{2}nD^2}}, \quad \pi_2 = \cdots = \pi_k = \frac{e^{-\frac{1}{2}nD^2}}{1 + (k-1)e^{-\frac{1}{2}nD^2}},$$

where $D = \max_{i,j} ||f_i - f_j||$ is the $L^2$-diameter of $\mathcal{G}_n$. For this choice is it easily verified that the Bayes risk is bounded away from 0 under the conditions of the proposition (Exercise 5.3).

We note that this proof only depends on the specific aspects of the white noise model in a few places. The crucial elements that make the proof work
are first of all the existence a ‘large’ set of well-separated elements $f_j$ in the parameter space that are ‘difficult’ to estimate and second the ability to appropriately bound the tail probability of the likelihood ratio $(p_{f_i}/p_{f_j})(X)$ under the measure $P_{f_j}$. Using the same approach we can also obtain minimax lower bounds in other statistical settings, as we illustrate in Section 5.5.

5.4 Lower bounds for estimating smooth functions

In this section we use the developed machinery to prove the converse of the results (3.12), (3.13) and (3.14). The theorems below imply that the rates obtained in these results are optimal. Appropriately tuned kernel estimators are rate-optimal with respect to pointwise and uniform risk over Hölder balls and an appropriately tuned projection estimator is rate-optimal with respect to the $L^2$-norm, uniformly over a Sobolev ball with a given regularity.

Theorem 5.4.1 (Minimax lower bound over Hölder balls for pointwise risk). Let $t \in (0, 1)$ be fixed. For $\beta, R, p > 0$ we have

$$\inf_{\hat{f}} \sup_{f \in C_\beta^R[0,1]} \mathbb{E}_f |\hat{f}(t) - f(t)|^p \gtrsim n^{-p\beta/(1+2\beta)},$$

where the infimum is over all estimators.

Proof. Consider the semimetric $d(f, g) = |f(t) - g(t)|$ on $C_\beta^R[0,1]$. We know from (a minor adaptation of) Example 4.2.4 that for every $\sigma > 0$ there exists a function $f_\sigma \in C_\beta^R[0,1]$ with $f(t)$ of the order $\sigma^\beta$ and squared $L^2$-norm of the order $\sigma^{1+2\beta}$. Now for $\sigma = n^{-1/(1+2\beta)}$ consider the set $G = \{0, f_\sigma\}$. We have $d(0, f_\sigma) = |f_\sigma(t)| \sim n^{-\beta/(1+2\beta)}$, so the functions in $G$ are $2cn^{-\beta/(1+2\beta)}$-separated with respect to $d$, for some constant $c > 0$. Since the number of elements of $G$ is 2 and the $L^2$-diameter is of the order $1/\sqrt{n}$, an application of Proposition 5.2.2 proves the statement of the theorem.

A similar argument proves the following minimax lower bound relative to the uniform distance.

Theorem 5.4.2 (Minimax lower bound over Hölder balls for uniform risk). Let $0 < u < v < 1$ be fixed. For $\beta, R, p > 0$ we have

$$\inf_{\hat{f}} \sup_{f \in C_\beta^R[0,1]} \mathbb{E}_f \sup_{u < t < v} |\hat{f}(t) - f(t)|^p \gtrsim \left( \frac{n}{\log n} \right)^{-p\beta/(1+2\beta)},$$

where the infimum is over all estimators.
Exercise 5.4.

To prove a minimax lower bound that complements the upper bound (3.14) we want to follow the same strategy as above, using Proposition 5.2.2 in combination with the examples from Chapter 4 of functions in a Sobolev ball that are difficult to distinguish. A complication is that in this case the functions that are for instance constructed in Example 4.4.3 do not fulfil the separation condition required by Proposition 5.2.2. To remedy this we will consider smaller collections of functions, obtained by deleting functions from the class constructed in Example 4.4.3 that are too close together in $L^2$-sense.

The following result will ensure that we will still have a sufficiently large class of separated ‘difficult’ functions. The lemma is a so-called packing bound for the space $\{-1,1\}^k$ of $k$-dimensional vectors of signs, endowed with the Hamming distance. In general, an $\varepsilon$-packing of a metric space is a collection of disjoint balls of radius $\varepsilon$ in that space. The associated $\varepsilon$-packing number is the maximum cardinality of an $\varepsilon$-packing, i.e. the maximum number of disjoint balls that fit into the space. This number somehow measures how ‘massive’, or ‘complex’ the space is.

**Lemma 5.4.3** (Varshamov-Gilbert packing bound for vectors of signs). Let $k \geq 8$ be a natural number. There exists a subset of $\{-1,1\}^k$ whose elements are $k/8$-separated for the Hamming distance and whose cardinality is larger than $a^k$ for a universal constant $a > 1$.

**Proof.** Let $S = \{-1,1\}^k$ and let $d_{\text{ham}}$ be the Hamming distance on $S$. Fix a point $s_1 \in S$ and consider the set

$$S_1 = \{s \in S : d_{\text{ham}}(s,s_1) > k/8\}.$$

Next, fix a point $s_2 \in S_1$ and define

$$S_2 = \{s \in S_1 : d_{\text{ham}}(s,s_2) > k/8\}.$$

This procedure is repeated $m$ times, until the set $S_m$ is empty.

By construction the points $s_1, s_2, \ldots, s_m$ are $k/8$-separated with respect to $d_{\text{ham}}$. To determine the number $m$, let $n_j = \text{the number of points that are removed in step } j$ to construct the set $S_j$ from $S_{j-1}$. Clearly $n_1 + n_2 + \cdots + n_m = 2^k$. Moreover, every $n_j$ is bounded by the maximal number of points in a Hamming ball of radius $k/8$, hence

$$n_j \leq \sum_{i \leq k/8} \binom{k}{i}.$$
Together, we obtain the bound

\[ 2^k \leq m \sum_{i \leq k/8} \binom{k}{i}. \]

The last inequality can be written as \( P(X \leq k/8) \geq 1/m \), where \( X \) is a binomial random variable with parameters \( 1/2 \) and \( k \). By Hoeffding’s inequality,

\[ P(X \leq k/8) \leq e^{-\frac{9}{32}k} \quad (5.6) \]

(see Exercise 5.5). We conclude that \( m \geq a^k \), with \( a = \exp(9/32) \approx 1.32 \). \( \Box \)

We can now prove the minimax lower bound that complements the upper bound (3.14), showing that an appropriately tuned projection estimator is indeed rate-optimal with respect to the \( L^2 \)-norm, uniformly over a Sobolev ball with a given regularity.

**Theorem 5.4.4** (Minimax lower bound over Sobolev balls for \( L^2 \)-risk). For all \( \beta, R, p > 0 \) we have

\[ \inf_{\hat{f}} \sup_{f} \mathbb{E}_f \| \hat{f} - f \|_2^p \gtrsim n^{-p\beta/(1+2\beta)}, \]

where the infimum is over all estimators.

**Proof.** In Example 4.4.3 we constructed for every \( \varepsilon > 0 \) a set of \( 2^k \) functions \( f_\theta, \theta \in \{-1,1\}^k \), for \( k = n^{-1/(1+2\beta)} \), satisfying

\[ \| f_\theta - f_\psi \|_2^2 = 4\varepsilon^2 k^{-1-2\beta}d_{\text{ham}}(\theta, \psi) \]

and

\[ \| f_\theta \|_{H^{\beta}}^2 = \text{const.} \times \varepsilon^2. \]

By Lemma 5.4.3 there we can find a subset \( \Theta \subset \{-1,1\}^k \) that is \( k/8 \)-separated for the Hamming distance and whose cardinality is of the order at least \( a^k \) for some \( a > 1 \). Then for \( \varepsilon > 0 \) small enough the functions in the set \( G_n = \{ f_\theta : \theta \in \Theta \} \) are contained in \( H^{\beta}_{R_n}[0,1] \), are \( 2r_n \)-separated for \( r_n \) a multiple of \( n^{-\beta/(1+2\beta)} \), and the \( L^2 \)-diameter of \( G_n \) is bounded by a constant times \( \varepsilon n^{-\beta/(1+2\beta)} \) (check!). It follows that all conditions of Proposition 5.2.2 are fulfilled if \( \varepsilon > 0 \) is small enough. \( \Box \)
5.5 Minimax lower bounds in other statistical settings

The ideas and techniques we have developed in this chapter are not only useful for deriving minimax lower bounds in the specific context of the white noise model. One relatively straightforward example is considered in Exercise 5.6, which asks to derive a minimax lower bound for a nonparametric regression model. As another illustration, we consider the problem of estimating a sparse vector in the normal means model in this section.

Recall that in the normal means model we observe a vector \( Y \) satisfying
\[
Y = \theta + \varepsilon,
\]
where \( \theta \in \mathbb{R}^n \) is the parameter of interest and \( \varepsilon \sim N_n(0, I) \). We have seen in Example 3.4.5 and Exercise 3.18 that if the vector \( \theta \) is assumed to be \( s \)-sparse and the degree of sparsity satisfies \( s = o(n) \), then there exist estimators which have rate of convergence \( \sqrt{s \log(n/s)} \) with respect to the Euclidean distance. Both the \( \ell^0 \)-penalized MLE and the lasso estimator achieve this rate, provided they are appropriately tuned. The theorem below gives a minimax lower bound which complements these results.

In the proof of the lower bound we use the following lemma, which gives a packing bound for the space of \( s \)-sparse vectors in \( \mathbb{R}^n \). In the proof we use, as before, the notation \( \| x \|_0 = |\{ i : x_i \neq 0 \}| \) for the \( \ell^0 \)-norm of a vector \( x \), which is its number of non-zero coordinates. A vector \( x \) is \( s \)-sparse if \( \| x \|_0 \leq s \).

**Lemma 5.5.1** (Packing bound for sparse vectors).

(i) Let \( s \leq n \) be natural numbers and \( p \in (0, 1) \). There exists a set of \( s \)-sparse vectors in the unit ball of \( \mathbb{R}^n \) whose elements are \( p \)-separated for the Euclidean distance and whose cardinality is larger than \( \exp(c_p s \log(n/C_p s)) \) for constants
\[
c_p = \frac{1 - p^2}{4}, \quad C_p = (1 - p^2)^2 \left( \frac{e}{p^2} \right)^{2p^2/(1-p^2)}.
\]

(ii) Let \( r \in (0, 1) \). There exist constants \( \varepsilon, c > 0 \) such that if \( s \) and \( n \) are natural numbers satisfying \( s \leq rn \), then there exists a set of \( s \)-sparse vectors in the unit ball of \( \mathbb{R}^n \) whose elements are \( \varepsilon \)-separated for the Euclidean distance and whose cardinality is larger than \( n \vee \exp(cs \log(n/s)) \).

**Proof.** (i). Let \( S \) be the set of \( s \)-sparse vectors with exactly \( s \) non-zero coordinates, all equal to \( 1/\sqrt{s} \). Note that these vectors belong to the unit ball of \( \mathbb{R}^n \). Let \( X \) and \( Y \) be independent random vectors in \( \mathbb{R}^n \), uniformly distributed on \( S \). Since \( \| X - Y \|^2 = \| X - Y \|_0/s \), we have
\[
P(\| X - Y \| \leq p) = P(\| X - Y \|_0 \leq p^2 s).
\]
Now given $X$, we have (check!)

$$P(\|X - Y\|_0 \leq p^2s \mid X) = \frac{|\{y \in S : \|y - X\|_0 \leq p^2s\}|}{|S|} \leq \binom{n}{s}^{-1} \left( \binom{s}{(1 - p^2)s} \right) \left( \frac{n - (1 - p^2)s}{p^2s} \right) = \left( \frac{n}{(1 - p^2)s} \right)^{-1} \left( \frac{s}{p^2s} \right)^2.$$

Using the bounds $(n/k)^k \leq \binom{n}{k} \leq (ne/k)^k$ for binomial coefficients it follows that

$$P(\|X - Y\|_0 \leq p^2s) \leq \left( \frac{e}{p^2s} \right)^{2p^2s} \left( \frac{(1 - p^2)s}{n} \right)^{(1 - p^2)s} = \left( \frac{C_{p^2s}^{(1-p^2)s}}{n} \right)^{(1-p^2)s}.$$

It follows that if $X_1, \ldots, X_N$ are $N$ independent vectors that are uniformly distributed on $S$, then

$$P\left(\|X_i - X_j\|_2 \leq p \text{ for some } i \neq j\right) \leq N^2 e^{-(1-\delta^2)s \log \frac{n}{p^2s}}.$$

For $N = e^{s \log \frac{n}{p^2s}}$ this is strictly less than 1, so that

$$P\left(\|X_i - X_j\|_2 > p \text{ for all } i \neq j\right) > 0.$$

Since the distribution of $(X_1, \ldots, X_N)$ is discrete, this implies that there exists at least one realization consisting of $p$-separated vectors.

(ii). Follows from part (ii), see Exercise 5.7. \hfill \Box

**Theorem 5.5.2** (Minimax lower bound for sparse signals in the normal means model). Let $r \in (0, 1)$ and let $s, n$ be natural numbers such that $n \geq 2$ and $s \leq rn$. Then for the normal means model we have, for every $p > 0$,

$$\inf_{\hat{\theta}} \sup_{\theta: \|\theta\|_0 \leq s} \mathbb{E}_\theta \|\hat{\theta} - \theta\|^p \gtrsim \left( s \log \frac{n}{s} \right)^{p/2},$$

where the infimum is over all estimators.

**Proof.** Fix $\delta > 0$. By multiplying the vectors exhibited in part (ii) of the preceding lemma by $\delta$ we see that there exist $N \geq 2 \vee \exp(cs \log(n/s))$ $s$-sparse vectors $\theta_1, \theta_2, \ldots, \theta_N$ with $L^2$-norm $\delta$ which are $2\varepsilon \delta$-separated for some $\varepsilon, c > 0$. Let $\Pi$ be a discrete prior on the set of $s$-sparse vectors which assigns prior probability $\pi_j$ to $\theta_j$ for every $j$. Then by following the line of reasoning in the alternative proof of Proposition 5.2.2 in Section 5.3, we see that to lower
bound the minimax risk by a constant times a particular choice of \( \delta \), it suffices to show that for that \( \delta \) and for an appropriate choice \( \Pi \),
\[
\sum \pi_j P_{\theta_j} \left( \max_i \pi_i p_{\theta_i}(Y) > \pi_j p_{\theta_j}(Y) \right)
\] (5.7)
is bounded away from 0, where \( p_{\theta}(Y) \) is the likelihood in the model. As before we use the simple lower bound
\[
P_{\theta_j} \left( \max_i \pi_i p_{\theta_i}(Y) > \pi_j p_{\theta_j}(Y) \right) \geq P_{\theta_j} \left( \frac{p_{\theta_j}}{p_{\theta_j}}(Y) > \frac{\pi_j}{\pi_1} \right).
\]
Observe that for every \( j \), the likelihood ratio satisfies
\[
\frac{p_{\theta_1}}{p_{\theta_j}}(Y) = e^{(\theta_1 - \theta_j, Y - \theta_j) - \frac{1}{2} \|\theta_1 - \theta_j\|^2} = e^{\|\theta_1 - \theta_j\| Z - \frac{1}{2} \|\theta_1 - \theta_j\|^2},
\]
where \( Z \) is standard normal under \( P_{\theta_j} \) (check!). It follows that the sum (5.7) is bounded from below by
\[
\sum_{j \neq 1} \pi_j \left( 1 - \Phi \left( \frac{\log \pi_j/\pi_1 + \frac{1}{2} \|\theta_1 - \theta_j\|^2}{\|\theta_1 - \theta_j\|^2} \right) \right).
\]
For all \( j \) we have that \( \|\theta_1 - \theta_j\| \leq 2\delta \), by the triangle inequality. Hence if we consider the prior given by
\[
\pi_1 = \frac{1}{1 + (N - 1)e^{-2\delta^2}},
\pi_2 = \cdots = \pi_N = \frac{e^{-2\delta^2}}{1 + (N - 1)e^{-2\delta^2}},
\]
then (5.7) is further lower bounded by \( 1/2 \) times
\[
\frac{(N - 1)e^{-2\delta^2}}{1 + (N - 1)e^{-2\delta^2}}.
\]
Since \( N \geq 2 \), this is bounded away from 0 as soon as \( Ne^{-2\delta^2} \) is bounded away from 0. Using the other lower bound on \( N \) we see that is the case for the choice
\[
\delta = \sqrt{\frac{1}{2} cs \log \frac{n}{s}}.
\]
This completes the proof of the theorem. \( \Box \)
5.6 Exercises

Exercise 5.1 (Minimax risk over large classes of signals).

(i) Prove that for every $p > 0$,
\[ \inf_{\hat{f}} \sup_{f \in L^2[0,1]} E_f \| \hat{f} - f \|_2^p \to \infty. \]

(ii) Let $L^2_1[0,1]$ be the unit ball in $L^2[0,1]$, i.e. $L^2_1[0,1] = \{ f \in L^2[0,1] : \| f \|_2 \leq 1 \}$. Prove that for every $p > 0$, there exists a constant $c > 0$ such that
\[ \inf_{\hat{f}} \sup_{f \in L^2_1[0,1]} E_f \| \hat{f} - f \|_2^p \geq c. \]

(iii) Recall that a subset of a metric space is called totally bounded if for all $\varepsilon > 0$ it can be covered by finitely many balls of radius $\varepsilon$.

Let $F \subset L^2[0,1]$ be a set of that is bounded but not totally bounded. Prove that also in this case it holds that for every $p > 0$, there exists a constant $c > 0$ such that
\[ \inf_{\hat{f}} \sup_{f \in F} E_f \| \hat{f} - f \|_2^p \geq c. \]

What do these statements mean in terms of convergence rates?

Exercise 5.2 (Minimax estimation in a non-regular parametric model). Consider the (parametric) model
\[ dX_t = \text{sign}(t - \theta) \, dt + \frac{1}{\sqrt{n}} \, dW_t, \]
where $\text{sign}(x) = 1_{x \geq 0} - 1_{x < 0}$ and $\theta \in (0,1)$ is an unknown parameter.

(i) Prove the minimax lower bound
\[ \inf_{\theta} \sup_{\theta \in (0,1)} E_\theta |\hat{\theta} - \theta| \gtrsim \frac{1}{n}. \]

(ii) Describe the MLE $\hat{\theta}_{\text{MLE}}$ in this model and show that it achieves the rate $1/n$, uniformly over $(0,1)$, i.e.
\[ \sup_{\theta \in (0,1)} E_\theta |\hat{\theta}_{\text{MLE}} - \theta| \lesssim \frac{1}{n}. \]

(iii) Find the limiting distribution of $n(\hat{\theta}_{\text{MLE}} - \theta)$. 
**Exercise 5.3** (Alternative proof of Proposition 5.2.2). Complete the alternative proof of Proposition 5.2.2 given in Section 5.3.

**Exercise 5.4** (Proof of Theorem 5.4.2). Give a proof of Theorem 5.4.2.

**Exercise 5.5** (Special case of Hoeffding’s inequality). Hoeffding’s inequality implies that if \( X_1, \ldots, X_k \) are independent Rademacher variables (±1 with probability 1/2) and \( a_1, \ldots, a_j \) are real numbers, then for all \( x > 0 \),

\[
P\left( \sum_{j=1}^{k} a_j X_j \geq x \right) \leq e^{-\frac{x^2}{2 \sum_{j} a_j^2}}.
\]

(i) Prove this inequality. (Hint: first show that for every \( \lambda > 0 \) we have \( E \exp(\lambda X_j) \leq \exp(\lambda^2/2) \).)

(ii) Use the inequality to prove (5.6).

**Exercise 5.6** (Minimax lower bound for nonparametric regression). Consider the fixed design regression model in which we observe \( Y_1, \ldots, Y_n \) satisfying the relation

\[ Y_i = f(i/n) + \varepsilon_i, \quad i = 1, \ldots, n, \]

where \( \varepsilon_1, \ldots, \varepsilon_n \) are independent standard normal variables and \( f : [0, 1] \to \mathbb{R} \) is the function of interest. In this statistical setting, derive a (non-trivial) minimax lower bound for the pointwise risk \( E(f(\hat{f}(t)) - f(t))^2 \) over the Hölder ball \( C^\beta([0, 1]) \) for \( \beta > 0 \), where \( t \) is a fixed point in \((0, 1)\).

**Exercise 5.7** (\( \varepsilon \)-separated sparse vectors). Verify the details of part (i) of the proof of Lemma 5.5.1 and provide a proof of part (ii).
Chapter 6

Adaptive estimation

6.1 Adaptation and data-driven tuning

So far we have studied various high- or infinite-dimensional statistical problems and methods in some detail. For various problems we have derived minimax lower bounds for the risk that estimators can achieve uniformly over certain classes of underlying truths. Moreover, we have seen that different estimators that we have considered in Chapter 3 achieve these optimal rates.

An important observation is that for all methods that we have seen so far it holds that to achieve an optimal convergence rate uniformly over a certain class of truths, it is required to choose a specific setting of a tuning parameter that depends on some feature of the class. This feature however is an aspect of the unknown truth, and it is typically not very realistic to assume that we actually have access to it. The kernel estimators or projection estimator for instance can achieve optimal rates over Hölder or Sobolev balls, but the rate-optimal estimators use the knowledge of the degree of smoothness $\beta$ of the unknown function to set the appropriate bandwidth or point of truncation, see Examples 3.4.2, 3.4.3 and 3.4.4. Similarly, the rate-optimal estimators that we considered for the estimation of sparse vectors in the normal means model use the knowledge of the degree of sparsity $s$ of the unknown signal, cf. Example 3.4.5 and Exercise 3.18. These examples illustrate that the rate-optimality results we have proved so far should really be seen as oracle results: there exists rate-optimal estimators for various problems, but they can not be practically implemented unless you have access to an oracle that has information about certain properties of the unknown truth.

The question that we address in this chapter is whether it is possible to construct estimators that do not use knowledge of aspects of the truth such as its degree of smoothness or sparsity, but that still achieve optimal convergence rates over smoothness or sparsity classes, for instance. If an estimator has this property, it is said to be adaptive with respect to the specific feature under consideration. Adaptive methods somehow manage, either explicitly or implicitly, to automatically choose the correct value of their tuning parameters
in a purely data-driven way.

6.2 Lepski’s method

The first method for adaptation that we consider tries to select the best estimator among a scale of non-adaptive estimators by trading off bias and variance in a purely data-driven manner. Of course this is not entirely straightforward, since we do not have direct access to these quantities in general. Still, it turns out that by considering certain cleverly chosen statistics, we can get sufficient information to make an optimal choice. For concreteness, we study the method in the context of the simple projection estimator of Example 3.2.1. However, as we will repeat at the end of this section, the method is much more general in nature and can be adapted to work in many more situations.

So we assume we are in the setting of the sequence formulation (2.4) of the white noise model. For every \( k \in \mathbb{N} \) we can then consider the projection estimator

\[
\hat{f}_k = (Y_1, Y_2, \ldots, Y_k, 0, 0, \ldots).
\]

Lepski’s method for choosing the appropriate truncation point \( k \) is based on an analysis of the statistics

\[
\|\hat{f}_k - \hat{f}_l\|_2^2
\]

for different \( k \) and \( l \). For \( l > k \) we can write

\[
\|\hat{f}_k - \hat{f}_l\|_2^2 = \|\hat{f}_k - E_f \hat{f}_k - (\hat{f}_l - E_f \hat{f}_l) + E_f \hat{f}_l - E_f \hat{f}_k\|_2^2
\]

\[
= \frac{1}{n} \sum_{k < j \leq l} Z_j^2 + \sum_{k < j \leq l} f_j^2 - 2 \frac{1}{\sqrt{n}} \sum_{k < j \leq l} f_j Z_j.
\]

(6.1)

Note that the first term in the last line of the display is of stochastic order \( l/n \), the second term is a lower bound for the squared bias of \( \hat{f}_k \) (see Example 3.2.1). If \( f \in H^\beta_R[0,1] \) for some \( \beta, R > 0 \), then the third term is of lower order than the first two (see the proof of Lemma 6.2.1 ahead). Hence if \( \|\hat{f}_k - \hat{f}_l\|_2^2 \gg l/n \) for some \( l \geq k \), then the squared bias of \( \hat{f}_k \) must be of larger order than \( l/n \), hence the bias is too large relative to the variance. If, on the other hand, \( \|\hat{f}_k - \hat{f}_l\|_2^2 \ll l/n \) for all \( l \geq k \), then this holds in particular for \( l \) a very large multiple of \( k \), and hence the squared bias of \( \hat{f}_k \), which is approximately the second term in that case, is of smaller order than \( k/n \). So in that case the bias is too small. This reasoning indicates that the ‘right’ choice of \( k \), which balances squared bias and variance, is the choice that ensures that \( \|\hat{f}_k - \hat{f}_l\|_2^2 \) is of the order \( l/n \) for \( l \geq k \).

These considerations motivate the following procedure. For \( \tau > 0 \), define

\[
\hat{k} = \min \left\{ k \leq n : \|\hat{f}_k - \hat{f}_l\|_2^2 \leq \frac{\tau l}{n} \text{ for all } k \leq l \leq n \right\}.
\]

We use this data-driven truncation level and define the estimator for \( f \) by

\[
\hat{f} = \hat{f}_{\hat{k}} = (Y_1, Y_2, \ldots, Y_{\hat{k}}, 0, 0, \ldots).
\]

A crucial step in the proof of the fact that the estimator \( \hat{f} \) is rate-optimal over Sobolev balls is the following lemma, which asserts that with probability tending to 1, the truncation level \( \hat{k} \) is upper bounded by the oracle choice (see Example 3.4.2).
Lemma 6.2.1. Suppose that $\beta, R > 0$ and $\tau \geq 4$. There exist universal constants $c_1, c_2 > 0$ such that

$$\sup_{f \in H^\beta_R[0,1]} P_f\left(\hat{k} > (nR^2)^{1/(1+2\beta)}\right) \leq c_1 e^{-c_2 \tau (nR^2)^{1/(1+2\beta)}}.$$ 

Proof. Set $k = (nR^2)^{1/(1+2\beta)}$. By definition of $\hat{k}$ and a union bound,

$$P_f(\hat{k} > k) \leq \sum_{l \geq k} P_f\left(\|\hat{f}_k - \hat{f}_l\|^2 > \frac{\tau l}{n}\right).$$

For $f \in H^\beta_R[0,1]$ we have the bound

$$\sum_{k < j \leq l} f_j^2 \leq R^2 k^{-2\beta} \leq l/n$$

for all $l \geq k$. In view of the expansion (6.1) it follows that

$$P_f\left(\|\hat{f}_k - \hat{f}_l\|^2 > \frac{\tau l}{n}\right) \leq P\left(\frac{1}{n} \sum_{k < j \leq l} Z_j^2 > \frac{(\tau - 1)l}{2n}\right) + P\left(\frac{1}{\sqrt{n}} \sum_{k < j \leq l} f_j Z_j > \frac{(\tau - 1)l}{2n}\right),$$

where the $Z_j$ are independent standard normal variables. The variance of the random variable in the second probability on the right is

$$\frac{4}{n} \sum_{k < j \leq l} f_j^2 \leq \frac{4l}{n^2}.$$ 

Hence, the probability is bounded by

$$1 - \Phi\left(\frac{(\tau - 1)\sqrt{l}}{4}\right) \leq e^{-\frac{(\tau - 1)^2 l}{32}}.$$ 

By Markov’s inequality the first probability is bounded by

$$e^{-\frac{(\tau - 1)^2 l}{8}}\left(1 + E e^{\frac{4}{l} Z_j^2}\right)^l = e^{-\frac{(\tau - 1)^2 - \log \sqrt{2}}{8} l}$$

(check!). For $\tau \geq 4$ it holds that $(\tau - 1)/8 - \log \sqrt{2} > c_0 \tau$ for some $c_0 > 0$. We conclude that there exists a constant $c_2 > 0$ such that for all $\tau \geq 6$,

$$P_f(\hat{k} > k) \leq 2 \sum_{l > k} e^{-c_2 \tau l} = \frac{2e^{-c_2 \tau}}{1 - e^{-c_2 \tau}} e^{-c_2 \tau k}.$$ 

Since $x \mapsto x/(1 - x)$ is increasing on $(0,1)$, taking

$$c_1 = \frac{2e^{-2c_2}}{1 - e^{-2c_2}}$$

completes the proof. \qed
So with probability tending to 1 the data-driven truncation point \( \hat{k} \) is not too large, i.e. we are not undersmoothing. The same cannot be proved for the possible problem of oversmoothing, but it turns out that we can have adaptive estimation nonetheless.

**Theorem 6.2.2.** Suppose that \( \beta, R > 0 \) and \( \tau \geq 4 \). Then

\[
\sup_{f \in H_0^\beta[0,1]} E_f \| \hat{f} - f \|_2 \lesssim n^{-\beta/(1+2\beta)}.
\]

**Proof.** Again let \( k = (nR^2)^{1/(1+2\beta)} \). By Cauchy-Schwarz,

\[
E_f \| \hat{f} - f \|_2 1_{k > k} = \sum_{k < j \leq n} E_f \| \hat{f}_j - f \|_2 1_{k = j} \leq \sum_{k < j \leq n} \sqrt{E_f \| \hat{f}_j - f \|^2} \sqrt{P_f(\hat{k} = j)}.
\]

For the risk of the projection estimator \( \hat{f}_j \) we have, for \( k < j \leq n \) and \( f \in H_0^\beta[0,1] \),

\[
E_f \| \hat{f}_j - f \|^2 = \sum_{i > j} f_i^2 + \frac{j}{n} \leq R^2 j^{-2\beta} + 1 \leq R^2 + 1.
\]

Hence, by the preceding lemma,

\[
\sup_{f \in H_0^\beta[0,1]} E_f \| \hat{f} - f \|_2 1_{k > k} \leq n \sqrt{1 + R^2} \sqrt{P_f(\hat{k} > k)} = o\left( n^{-\beta/(1+2\beta)} \right).
\]

Next, we note that

\[
E_f \| \hat{f} - f \|_2 1_{k \leq k} \leq E_f \| \hat{f} - \hat{f}_k \|_2 1_{k \leq k} + E_f \| \hat{f}_k - f \|_2 1_{k \leq k}.
\]

The second term is bounded by \( (E_f \| \hat{f}_k - f \|^2)^{1/2} \), which is bounded by a constant times \( n^{-\beta/(1+2\beta)} \), uniformly for \( f \in H_0^\beta[0,1] \) (see (3.14)). For the first term we note that by definition of \( \hat{k} \),

\[
E_f \| \hat{f} - \hat{f}_k \|_2 1_{k \leq k} \leq \sqrt{\tau k} \leq n^{-\beta/(1+2\beta)}.
\]

This completes the proof. \( \square \)

So indeed, we see that the estimator \( \hat{f} \) attains the optimal rate \( n^{-\beta/(1+2\beta)} \) uniformly over Sobolev balls with regularity \( \beta \). Moreover, the estimator does not use knowledge of the parameter \( \beta \). Hence, the procedure automatically adapts to the smoothness of the unknown signal!

For concreteness we have worked out the performance of Lepski’s method in the context of the white noise model, using the projection estimators as class of non-adaptive estimators that serve as input for the method and obtaining
optimal rates relative to the $L^2$-norm. All of this can be generalized. The proofs show that the essence of the argument goes through as soon as we have a scale of estimators with a one-dimensional tuning parameter such that the bias and the variance in the relevant norm are tractable and depend in a monotone way on the tuning parameter. See Exercise 6.1 for a version of Lepski’s method for the kernel estimator of Example 3.4.3.
6.3 Exercises

Exercise 6.1 (Lepski’s method for kernel estimators in the white noise model).
For a fixed point $t \in (0, 1)$ consider the kernel estimator $\hat{f}_h(t)$ for $f(t)$ in the white noise model, as studied in Example 3.4.3. For $\tau > 0$ and $H_n = \{1/n, 2/n, \ldots, 1\}$, define

$$\hat{h} = \sup \left\{ h_1 \in H_n : \left| \hat{f}_{h_1}(t) - \hat{f}_{h_2}(t) \right| \leq \sqrt{\frac{\tau}{nh_2}} \text{ for all } h_2 \in H_n \text{ such that } h_2 < h_1 \right\}.$$

Define the estimator $\hat{f}(t) = \hat{f}_{\hat{h}}(t)$ obtained by using the data-driven choice $\hat{h}$ as bandwidth.

(i) Prove the analogue of Lemma 6.2.1 for the estimator $\hat{h}$, where $f$ is assumed to be in a Hölder ball $C^\beta_R[0, 1]$ for some $\beta \in (0, 1]$ and $R > 0$.

(ii) Prove the analogue of Theorem 6.2.2, where again $f$ is assumed to be in a Hölder ball $C^\beta_R[0, 1]$ for some $\beta \in (0, 1]$ and $R > 0$ and pointwise risk is considered.
Bibliography


