KÖFOP-2.1.2-VEKOP-15. „A jó kormányzást megalapozó közszolgálat-fejlesztés”

László Györfi

Nonparametric Estimations and Predictions

Nonparametric Estimations and Predictions

László Györfi [1]

August 15, 2018

[1]The work was created in commission of the National University of Public Service under the priority project KÖFOP-2.1.2-VEKOP-15-2016-00001 titled "Public Service Development Establishing Good Governance" in Ludovika Research Group.

Contents

Chapter 1

The regression estimation problem

In this chapter we introduce the problem of regression function estimation and describe important properties of regression estimates. Furthermore, provide an overview of various approaches to nonparametric regression estimates.

1.1 Why to estimate a regression function?

In regression analysis one considers a random vector $(\mathbf{X}, Y)$, where $\mathbf{X}$ is $\mathbb{R}^d$-valued and Y is $\mathbb{R}$-valued, and one is interested how the value of the so-called response variable Y depends on the value of the observation vector $\mathbf{X}$. This means that one wants to find a (measurable) function $f : \mathbb{R}^d \to \mathbb{R}$, such that $f(\mathbf{X})$ is a "good approximation of Y," that is, $f(\mathbf{X})$ should be close to Y in some sense, which is equivalent to making $|f(\mathbf{X}) - Y|$ "small." Since $\mathbf{X}$ and Y are random vectors, $|f(\mathbf{X}) - Y|$ is random as well, therefore it is not clear what "small $|f(\mathbf{X}) - Y|$" means. We can resolve this problem by introducing the so-called L_2 *risk* or *mean squared error* of f,

$$\mathbb{E}|f(\mathbf{X}) - Y|^2,$$

and requiring it to be as small as possible.

There are two reasons for considering the L_2 risk. First, as we will see in the sequel, this simplifies the mathematical treatment of the whole problem. For example, as is shown below, the function which minimizes the L_2 risk can be derived explicitly. Second, and more important, trying to minimize the L_2 risk leads naturally to estimates which can be computed rapidly.

So we are interested in a (measurable) function $m^* : \mathbb{R}^d \to \mathbb{R}$ such that

$$\mathbb{E}|m^*(\mathbf{X}) - Y|^2 = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}|f(\mathbf{X}) - Y|^2.$$

Such a function can be obtained explicitly as follows. Let

$$m(\mathbf{x}) = \mathbb{E}\{Y|\mathbf{X} = \mathbf{x}\}$$

be the *regression function*. We will show that the regression function minimizes the L_2 risk. Indeed, for an arbitrary $f : \mathbb{R}^d \to \mathbb{R}$, one has

$$\begin{aligned}\mathbb{E}|f(\mathbf{X}) - Y|^2 &= \mathbb{E}|f(\mathbf{X}) - m(\mathbf{X}) + m(\mathbf{X}) - Y|^2 \\ &= \mathbb{E}|f(\mathbf{X}) - m(\mathbf{X})|^2 + \mathbb{E}|m(\mathbf{X}) - Y|^2,\end{aligned}$$

where we have used

$$\begin{aligned}&\mathbb{E}\left\{(f(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)\right\} \\ &= \mathbb{E}\left\{\mathbb{E}\left\{(f(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - Y)\middle|\mathbf{X}\right\}\right\} \\ &= \mathbb{E}\left\{(f(\mathbf{X}) - m(\mathbf{X}))\mathbb{E}\{m(\mathbf{X}) - Y|\mathbf{X}\}\right\} \\ &= \mathbb{E}\left\{(f(\mathbf{X}) - m(\mathbf{X}))(m(\mathbf{X}) - m(\mathbf{X}))\right\} \\ &= 0.\end{aligned}$$

Hence,

$$\mathbb{E}|f(\mathbf{X}) - Y|^2 = \int_{\mathbb{R}^d} |f(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) + \mathbb{E}|m(\mathbf{X}) - Y|^2, \tag{1.1}$$

where μ denotes the distribution of $\mathbf{X}$. The first term is called the L_2 error of f. It is always nonnegative and is zero if $f(\mathbf{x}) = m(\mathbf{x})$. Therefore, $m^*(\mathbf{x}) = m(\mathbf{x})$, i.e., the optimal approximation (with respect to the L_2 risk) of Y by a function of $\mathbf{X}$ is given by $m(\mathbf{X})$.

In applications the distribution of $(\mathbf{X}, Y)$ (and hence also the regression function) is usually unknown. Therefore it is impossible to predict Y using $m(\mathbf{X})$. But it is often possible to observe data according to the distribution of $(\mathbf{X}, Y)$ and to estimate the regression function from these data.

To be more precise, denote by $(\mathbf{X}, Y)$, $(\mathbf{X}_1, Y_1)$, $(\mathbf{X}_2, Y_2), \ldots$ independent and identically distributed (i.i.d.) random variables with $\mathbb{E}Y^2 < \infty$. Let D_n be the set of *data* defined by

$$D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}.$$

In the regression function estimation problem one wants to use the data D_n in order to construct an estimate $m_n : \mathbb{R}^d \to \mathbb{R}$ of the regression function m. Here $m_n(\mathbf{x}) = m_n(\mathbf{x}, D_n)$ is a measurable function of $\mathbf{x}$ and the data. For simplicity, we will suppress D_n in the notation and write $m_n(\mathbf{x})$ instead of $m_n(\mathbf{x}, D_n)$.

In general, estimates will not be equal to the regression function. To compare different estimates, we need an error criterion which measures the difference between the regression function and an arbitrary estimate m_n. One of the key points we would like to make is that the motivation for introducing the regression function leads naturally to an L_2 error criterion for measuring the performance of the regression function estimate. Recall that the main goal was to find a function f such that the L_2 risk $\mathbb{E}|f(\mathbf{X}) - Y|^2$ is small. The minimal value of this L_2 risk is $\mathbb{E}|m(\mathbf{X}) - Y|^2$, and it is achieved by the regression function m. Similarly to (1.1), one can show that the L_2 risk $\mathbb{E}\{|m_n(\mathbf{X}) - Y|^2|D_n\}$ of an estimate m_n satisfies

$$\mathbb{E}\left\{|m_n(\mathbf{X}) - Y|^2|D_n\right\} = \int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) + \mathbb{E}|m(\mathbf{X}) - Y|^2. \tag{1.2}$$

Thus the L_2 risk of an estimate m_n is close to the optimal value if and only if the L_2 error

$$\|m_n - m\|^2 = \int_{\mathbb{R}^d} |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}) \tag{1.3}$$

is close to zero. Therefore we will use the L_2 error (1.3) in order to measure the quality of an estimate and we will study estimates for which this L_2 error is small.

The classical approach for estimating a regression function is the so-called parametric regression estimation. Here one assumes that the structure of the regression function is known and depends only on finitely many parameters, and one uses the data to estimate the (unknown) values of these parameters.

The linear regression estimate is an example of such an estimate. In linear regression one assumes that the regression function is a linear combination of the components of $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^T$, i.e.,

$$m(x^{(1)}, \ldots, x^{(d)}) = a_0 + \sum_{i=1}^{d} a_i x^{(i)} \quad ((x^{(1)}, \ldots, x^{(d)})^T \in \mathbb{R}^d)$$

for some unknown $a_0, \ldots, a_d \in \mathbb{R}$. Then one uses the data to estimate these parameters, e.g., by applying the principle of least squares, where one chooses the coefficients $a_0, \ldots, a_d$ of the linear function such that it best fits the given data:

$$(\hat{a}_0, \ldots, \hat{a}_d) = \arg \min_{a_0, \ldots, a_d \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{j=1}^{n} \left| Y_j - a_0 - \sum_{i=1}^{d} a_i X_j^{(i)} \right|^2 \right\}.$$

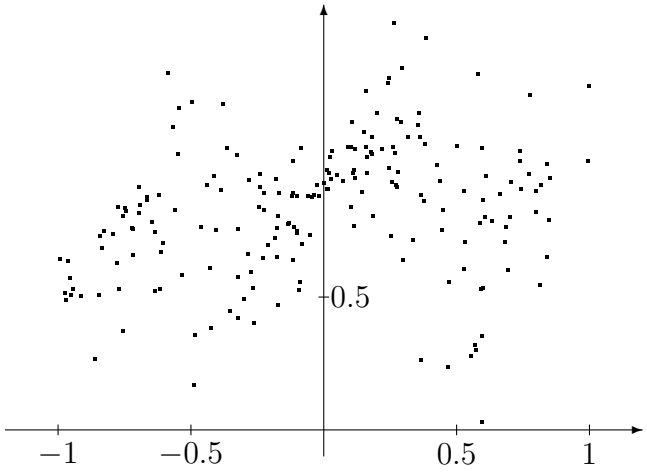

Figure 1.1: Simulated data points.

Here $X_j^{(i)}$ denotes the ith component of $\mathbf{X}_j$ and $\mathbf{z} = \arg\min_{\mathbf{x}\in D} f(\mathbf{x})$ is the abbreviation for $\mathbf{z} \in D$ and $f(\mathbf{z}) = \min_{\mathbf{x}\in D} f(\mathbf{x})$. Finally one defines the estimate by

$$\hat{m}_n(\mathbf{x}) = \hat{a}_0 + \sum_{i=1}^{d} \hat{a}_i x^{(i)}.$$

Parametric estimates usually depend only on a few parameters, therefore they are suitable even for small sample sizes n, if the parametric model is appropriately chosen. Furthermore, they are often easy to interpret. For instance in a linear model (when $m(\mathbf{x})$ is a linear function) the absolute value of the coefficient $\hat{a}_i$ indicates how much influence the ith component of $\mathbf{X}$ has on the value of Y, and the sign of $\hat{a}_i$ describes the nature of this influence (increasing or decreasing the value of Y).

However, parametric estimates have a big drawback. Regardless of the data, a parametric estimate cannot approximate the regression function better than the best function which has the assumed parametric structure. For example, a linear regression estimate will produce a large error for every sample size if the true underlying regression function is not linear and cannot be well approximated by linear functions.

For univariate $X = \mathbf{X}$ one can often use a plot of the data to choose a proper parametric estimate. But this is not always possible, as we now illustrate using simulated data. These data will be used throughout the book. They consist of $n = 200$ points such that X is standard normal restricted to $[-1, 1]$, i.e., the density of X is proportional to the standard normal density on $[-1, 1]$ and is zero elsewhere. The regression function is

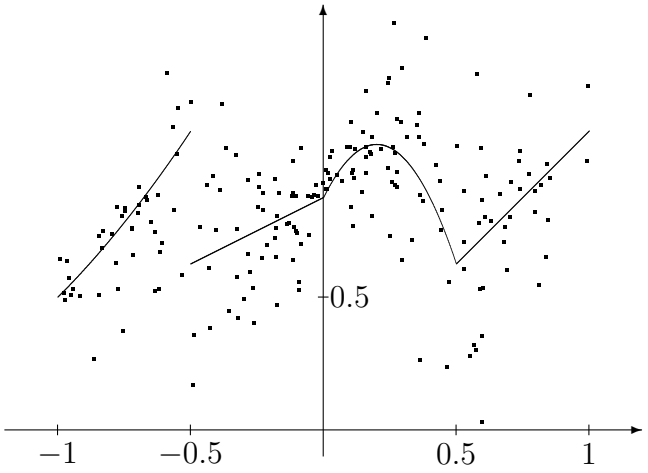

Figure 1.2: Data points and regression function.

piecewise polynomial:

$$m(x) = \begin{cases} (x+2)^2/2 & \text{if} \quad -1 \le x < -0.5, \\ x/2 + 0.875 & \text{if} \quad -0.5 \le x < 0, \\ -5(x-0.2)^2 + 1.075 & \text{if} \quad 0 < x \le 0.5, \\ x + 0.125 & \text{if} \quad 0.5 \le x < 1. \end{cases}$$

Given X, the conditional distribution of $Y - m(X)$ is normal with mean zero and standard deviation

$$\sigma(X) = 0.2 - 0.1 \cos(2\pi X).$$

Figure 1.1 shows the data points. In this example the human eye is not able to see from the data points what the regression function looks like. In Figure 1.2 the data points are shown together with the regression function.

In Figure 1.3 a linear estimate is constructed for these simulated data. Obviously, a linear function does not approximate the regression function well.

Furthermore, for multivariate $\mathbf{X}$, there is no easy way to visualize the data. Thus, especially for multivariate $\mathbf{X}$, it is not clear how to choose a proper form of a parametric estimate, and a wrong form will lead to a bad estimate. This inflexibility concerning the structure of the regression function is avoided by so-called nonparametric regression estimates.

We will now define the modes of convergence of the regression estimates that we will study in this book.

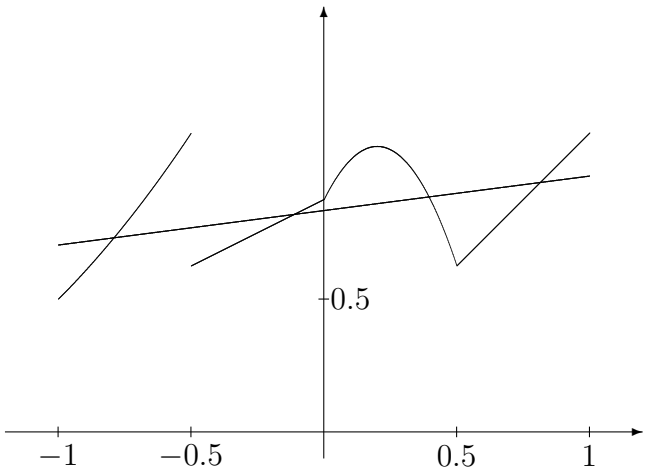

Figure 1.3: Linear regression estimate.

The first and weakest property an estimate should have is that, as the sample size grows, it should converge to the estimated quantity, i.e., the error of the estimate should converge to zero for a sample size tending to infinity. Estimates which have this property are called consistent.

To measure the error of a regression estimate, we use the L_2 error

$$\int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}).$$

The estimate m_n depends on the data D_n, therefore the L_2 error is a random variable. We are interested in the convergence of the expectation of this random variable to zero as well as in the almost sure ($a.s.$) convergence of this random variable to zero.

Definition 1.1. *A sequence of regression function estimates* $\{m_n\}$ *is called* **weakly consistent for a certain distribution of** $(\mathbf{X}, Y)$*, if*

$$\lim_{n\to\infty} \mathbb{E}\left\{\int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x})\right\} = 0.$$

Definition 1.2. *A sequence of regression function estimates* $\{m_n\}$ *is called* **strongly consistent for a certain distribution of** $(\mathbf{X}, Y)$*, if*

$$\lim_{n\to\infty} \int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) = 0 \quad \textit{with probability one.}$$

It may be that a regression function estimate is consistent for a certain class of distributions of $(\mathbf{X}, Y)$, but not consistent for others. It is clearly desirable to have estimates that are consistent for a large class of distributions. In the next chapters we are interested in properties of m_n that are valid for all distributions of $(\mathbf{X}, Y)$, that is, in distribution-free or universal properties. The concept of universal consistency is important in nonparametric regression because the mere use of a nonparametric estimate is normally a consequence of the partial or total lack of information about the distribution of $(\mathbf{X}, Y)$. Since in many situations we do not have any prior information about the distribution, it is essential to have estimates that perform well for *all* distributions. This very strong requirement of universal goodness is formulated as follows:

Definition 1.3. *A sequence of regression function estimates $\{m_n\}$ is called* **weakly universally consistent** *if it is weakly consistent for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}\{Y^2\} < \infty$.*

Definition 1.4. *A sequence of regression function estimates $\{m_n\}$ is called* **strongly universally consistent** *if it is strongly consistent for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}\{Y^2\} < \infty$.*

We will later give many examples of estimates that are weakly and strongly universally consistent.

If an estimate is universally consistent, then, regardless of the true underlying distribution of $(\mathbf{X}, Y)$, the L_2 error of the estimate converges to zero for a sample size tending to infinity. But this says nothing about how fast this happens. Clearly, it is desirable to have estimates for which the L_2 error converges to zero as fast as possible.

To decide about the rate of convergence of an estimate m_n, we will look at the expectation of the L_2 error,

$$\mathbb{E} \int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x}). \tag{1.4}$$

A natural question to ask is whether there exist estimates for which (1.4) converges to zero at some fixed, nontrivial rate for all distributions of $(\mathbf{X}, Y)$. Unfortunately, such estimates do not exist, i.e., for any estimate the rate of convergence may be arbitrarily slow. In order to get nontrivial rates of convergence, one has to restrict the class of distributions, e.g., by imposing some smoothness assumptions on the regression function.

1.2 How to estimate a regression function?

In this section we describe two principles of nonparametric regression: **local averaging** and **empirical error minimization**.

Recall that the regression function is defined by a conditional expectation

$$m(\mathbf{x}) = \mathbb{E}\{Y \mid \mathbf{X} = \mathbf{x}\}.$$

If $\mathbf{x}$ is an atom of $\mathbf{X}$, i.e., $\mathbb{P}\{\mathbf{X} = \mathbf{x}\} > 0$ then the conditional expectation is defined by the conventional way:

$$\mathbb{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \frac{\mathbb{E}\{Y\, \mathbb{I}_{\{\mathbf{X}=\mathbf{x}\}}\}}{\mathbb{P}\{\mathbf{X} = \mathbf{x}\}},$$

where $\mathbb{I}_A$ denotes the indicator function of set A. In this definition one can estimate the numerator by

$$\frac{1}{n}\sum_{i=1}^{n} Y_i\, \mathbb{I}_{\{\mathbf{X}_i=\mathbf{x}\}},$$

while the denominator's estimate is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i=\mathbf{x}\}},$$

so the obvious regression estimate can be

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i\, \mathbb{I}_{\{\mathbf{X}_i=\mathbf{x}\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i=\mathbf{x}\}}}.$$

In the general case of $\mathbb{P}\{\mathbf{X} = \mathbf{x}\} = 0$ we can refer to the measure theoretic definition of conditional expectation (cf. Appendix of Devroye, Györfi, and Lugosi (1996)). However, this definition is useless from the point of view of statistics. One can derive an estimate from the property

$$\mathbb{E}\{Y \mid \mathbf{X} = \mathbf{x}\} = \lim_{h\to 0} \frac{\mathbb{E}\{Y\, \mathbb{I}_{\{\|\mathbf{X}-\mathbf{x}\|\le h\}}\}}{\mathbb{P}\{\|\mathbf{X} - \mathbf{x}\| \le h\}}$$

so the following estimate can be introduced:

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i\, \mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{x}\|\le h\}}}{\sum_{i=1}^{n} \mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{x}\|\le h\}}}.$$

This estimate is called naive kernel estimate.

We can generalize this idea by *local averaging*, i.e., estimation of $m(\mathbf{x})$ is the average of those Y_i, where $\mathbf{X}_i$ is "close" to $\mathbf{x}$. Such an estimate can be written as

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} W_{n,i}(\mathbf{x}) \cdot Y_i,$$

where the weights $W_{n,i}(\mathbf{x}) = W_{n,i}(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}$ depend on $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Usually the weights are nonnegative and $W_{n,i}(\mathbf{x})$ is "small" if $\mathbf{X}_i$ is "far" from $\mathbf{x}$.

Examples of such an estimates are the *partitioning estimate*, the *kernel estimate* and the *k-nearest neighbor estimate*.

For nonparametric regression estimation, the other principle is the *empirical error minimization estimates*, where there is a class $\mathcal{F}_n$ of functions, and the estimate is defined by.

$$m_n(\cdot) = \arg\min_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2 \right\}. \tag{1.5}$$

Hence it minimizes the empirical L_2 risk

$$\frac{1}{n} \sum_{i=1}^{n} |f(\mathbf{X}_i) - Y_i|^2 \tag{1.6}$$

over $\mathcal{F}_n$. Observe that it doesn't make sense to minimize (1.6) over all (measurable) functions f, because this may lead to a function which interpolates the data and hence is not a reasonable estimate. Thus one has to restrict the set of functions over which one minimizes the empirical L_2 risk. Examples of possible choices of the set $\mathcal{F}_n$ are sets of piecewise polynomials or sets of smooth piecewise polynomials (splines). The use of spline spaces ensures that the estimate is a smooth function. An important member of least squares estimates is the *generalized linear estimates*. Let $\{\phi_j\}_{j=1}^{\infty}$ be real-valued functions defined on $\mathbb{R}^d$ and let $\mathcal{F}_n$ be defined by

$$\mathcal{F}_n = \left\{ f;\ f = \sum_{j=1}^{\ell_n} c_j \phi_j \right\}.$$

Then the generalized linear estimate is defined by

$$\begin{aligned} m_n(\cdot) &= \arg\min_{f \in \mathcal{F}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{X}_i) - Y_i)^2 \right\} \\ &= \arg\min_{c_1, \ldots, c_{\ell_n}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{\ell_n} c_j \phi_j(\mathbf{X}_i) - Y_i \right)^2 \right\}. \end{aligned}$$

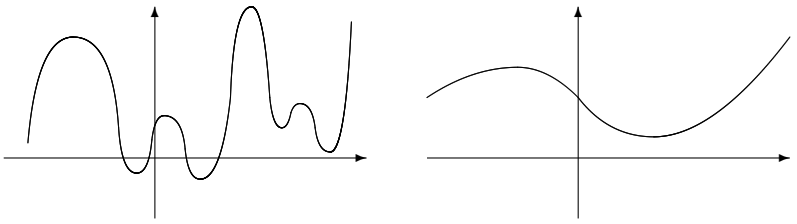

Figure 1.4: The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

For least squares estimates, other example can be the neural networks or radial basis functions or orthogonal series estimates.

Let m_n be an arbitrary estimate. For any $\mathbf{x} \in \mathbb{R}^d$ we can write the expected squared error of m_n at $\mathbf{x}$ as

$$\begin{aligned}
&\mathbb{E}\{|m_n(\mathbf{x}) - m(\mathbf{x})|^2\} \\
&= \mathbb{E}\{|m_n(\mathbf{x}) - \mathbb{E}\{m_n(\mathbf{x})\}|^2\} + |\mathbb{E}\{m_n(\mathbf{x})\} - m(\mathbf{x})|^2 \\
&= \mathbb{V}ar(m_n(\mathbf{x})) + |bias(m_n(\mathbf{x}))|^2.
\end{aligned}$$

Here $\mathbb{V}ar(m_n(\mathbf{x}))$ is the variance of the random variable $m_n(\mathbf{x})$ and $bias(m_n(\mathbf{x}))$ is the difference between the expectation of $m_n(\mathbf{x})$ and $m(\mathbf{x})$. This also leads to a similar decomposition of the expected L_2 error:

$$\begin{aligned}
&\mathbb{E}\left\{\int |m_n(\mathbf{x}) - m(\mathbf{x})|^2 \mu(d\mathbf{x})\right\} \\
&= \int \mathbb{E}\{|m_n(\mathbf{x}) - m(\mathbf{x})|^2\}\mu(d\mathbf{x}) \\
&= \int \mathbb{V}ar(m_n(\mathbf{x}))\mu(d\mathbf{x}) + \int |bias(m_n(\mathbf{x}))|^2 \mu(d\mathbf{x}).
\end{aligned}$$

The importance of these decompositions is that the integrated variance and the integrated squared bias depend in opposite ways on the wiggliness of an estimate. If one increases the wiggliness of an estimate, then usually the integrated bias will decrease, but the integrated variance will increase (so-called **bias–variance tradeoff**).

In Figure 1.5 this is illustrated for the kernel estimate, where one has, under some regularity conditions on the underlying distribution and for the naive kernel,

$$\int_{\mathbb{R}^d} \mathbb{V}ar(m_n(\mathbf{x}))\mu(d\mathbf{x}) = c_1 \frac{1}{nh^d} + o\left(\frac{1}{nh^d}\right)$$

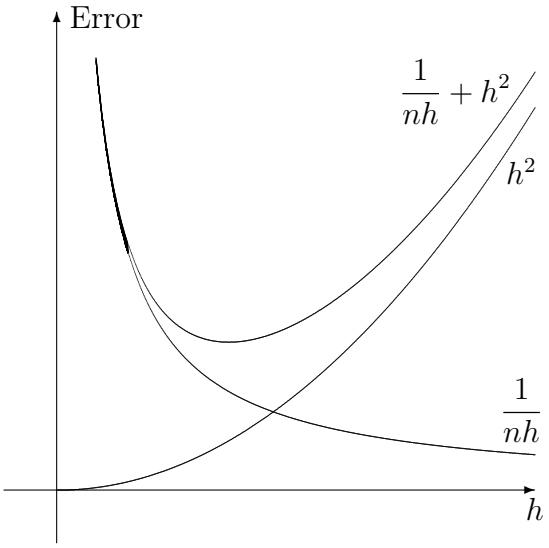

Figure 1.5: Bias–variance tradeoff.

and

$$\int_{\mathbb{R}^d} |bias(m_n(\mathbf{x}))|^2 \mu(d\mathbf{x}) = c_2 h^2 + o\left(h^2\right).$$

Here h denotes the bandwidth of the kernel estimate which controls the wiggliness of the estimate, c_1 is some constant depending on the conditional variance $\mathbb{V}ar\{Y|\mathbf{X} = \mathbf{x}\}$, the regression function is assumed to be Lipschitz continuous, and c_2 is some constant depending on the Lipschitz constant.

The value h^* of the bandwidth for which the sum of the integrated variance and the squared bias is minimal depends on c_1 and c_2. Since the underlying distribution, and hence also c_1 and c_2, are unknown in an application, it is important to have methods which choose the bandwidth automatically using only the data D_n.

Chapter 2

Partitioning estimates

2.1 Introduction

In the next chapters we briefly review the most important local averaging regression estimates. Concerning further details see Györfi *et al.* (2002).

The partitioning estimate, called a regressogram, was introduced by Tukey (1947; 1961) and studied by Collomb (1977), Bosq and Lecoutre (1987), and Lecoutre (1980). Concerning its consistency, see Devroye and Györfi (1983) and Györfi (1991). Beirlant and Györfi (1998) proved the asymptotic normality of the L_2 error, while Györfi, Schäfer, and Walk (2002) showed its relative stability.

Let $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots\}$ be a partition of $\mathbb{R}^d$ and for each $\mathbf{x} \in \mathbb{R}^d$ let $A_n(\mathbf{x})$ denote the cell of $\mathcal{P}_n$ containing $\mathbf{x}$. The partitioning estimate (histogram) of the regression function is defined as

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}$$

with $0/0 = 0$ by definition. This means that the partitioning estimate is a local averaging estimate such for a given $\mathbf{x}$ we take the average of those Y_i's for which $\mathbf{X}_i$ belongs to the same cell into which $\mathbf{x}$ falls.

The simplest version of this estimate is obtained for $d = 1$ and when the cells $A_{n,j}$ are intervals of size $h = h_n$. Figures 2.1 – 2.3 show the estimates for various choices of h for our simulated data introduced in Chapter 1. In the first figure h is too small (undersmoothing, large variance), in the second choice it is about right, while in the third it is too large (oversmoothing, large bias).

For $d > 1$ one can use, e.g., a cubic partition, where the cells $A_{n,j}$ are cubes of volume h_n^d, or a rectangle partition which consists of rectangles $A_{n,j}$ with side lengths

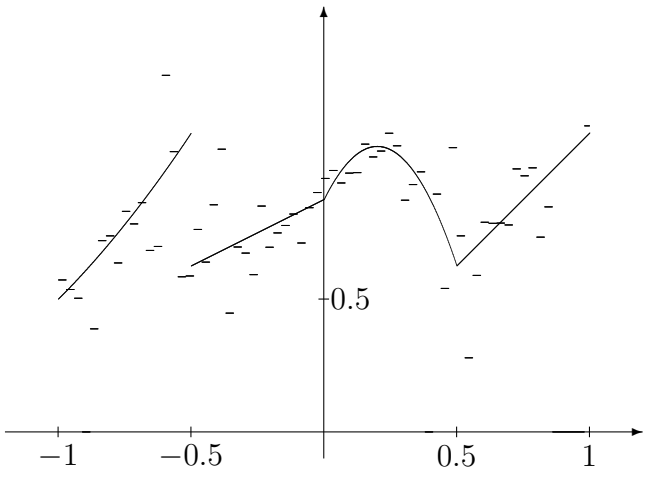

Figure 2.1: Undersmoothing: $h = 0.03$, L_2 error $= 0.062433$.

$h_{n1}, \ldots, h_{nd}$. For the sake of illustration we generated two-dimensional data when the actual distribution is a correlated normal distribution. The partition in Figure 2.4 is cubic, and the partition in Figure 2.5 is made of rectangles.

Cubic and rectangle partitions are particularly attractive from the computational point of view, because the set $A_n(\mathbf{x})$ can be determined for each $\mathbf{x}$ in constant time, provided that we use an appropriate data structure. In most cases, partitioning estimates are computationally superior to the other nonparametric estimates, particularly if the search for $A_n(\mathbf{x})$ is organized using binary decision trees (cf. Friedman (1977)).

The partitions may depend on the data. Figure 2.6 shows such a partition, where each cell contains an equal number of points. This partition consists of so-called statistically equivalent blocks.

Another advantage of the partitioning estimate is that it can be represented or compressed very efficiently. Instead of storing all data D_n, one should only know the estimate for each nonempty cell, i.e., for cells $A_{n,j}$ for which $\mu_n(A_{n,j}) > 0$, where μ_n denotes the empirical distribution. The number of nonempty cells is much smaller than n. (Cf. Lugosi, Nobel (1996).)

2.2 Stone's Theorem

In the next section we will prove the weak universal consistency of partitioning estimates. In the proof we will use Stone's theorem (Theorem 2.1 below) which is a powerful tool

for proving weak consistency for local averaging regression function estimates. It will also be applied to prove the weak universal consistency of kernel and nearest neighbor estimates in Chapters 3 and 4.

Local averaging regression function estimates take the form

$$m_n(\mathbf{x}) = \sum_{i=1}^{n} W_{ni}(\mathbf{x}) \cdot Y_i,$$

where the weights $W_{n,i}(\mathbf{x}) = W_{n,i}(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}$ are depending on $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Usually the weights are nonnegative and $W_{n,i}(\mathbf{x})$ is "small" if $\mathbf{X}_i$ is "far" from $\mathbf{x}$. The next theorem states conditions on the weights which guarantee the weak universal consistency of the local averaging estimates.

Theorem 2.1. (STONE'S THEOREM, STONE (1977)). *Assume that the following conditions are satisfied for any distribution of* $\mathbf{X}$*:*

(i) There is a constant c *such that for every nonnegative measurable function* f *satisfying* $\mathbb{E}f(\mathbf{X}) < \infty$ *and any* n*,*

$$\mathbb{E}\left\{\sum_{i=1}^{n} |W_{n,i}(\mathbf{X})| f(\mathbf{X}_i)\right\} \leq c\mathbb{E}f(\mathbf{X}).$$

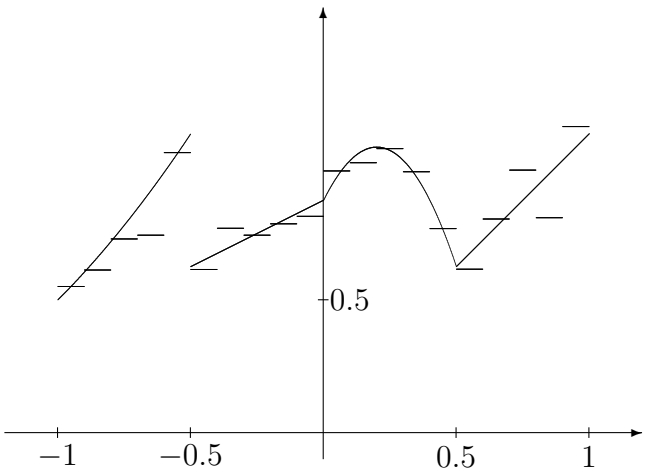

Figure 2.2: Good choice: $h = 0.1$, L_2 error $= 0.003642$.

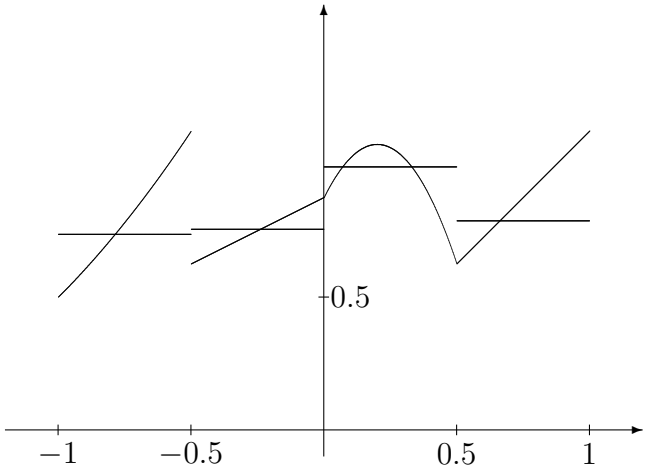

Figure 2.3: Oversmoothing: $h = 0.5$, L_2 error $= 0.013208$.

(ii) There is a $D \geq 1$ such that

$$\mathbb{P}\left\{\sum_{i=1}^{n} |W_{n,i}(\mathbf{X})| \leq D\right\} = 1,$$

for all n.

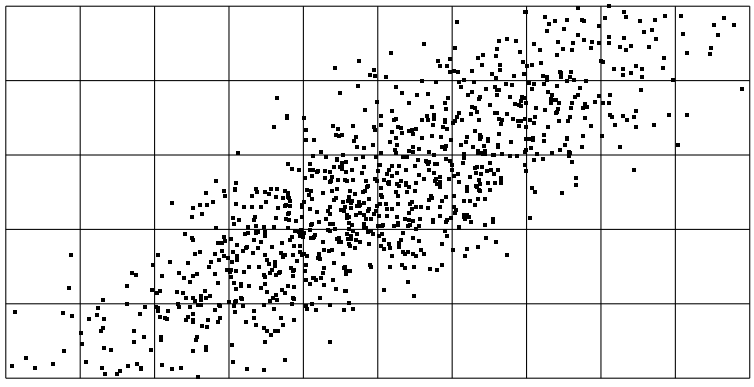

Figure 2.4: Cubic partition.

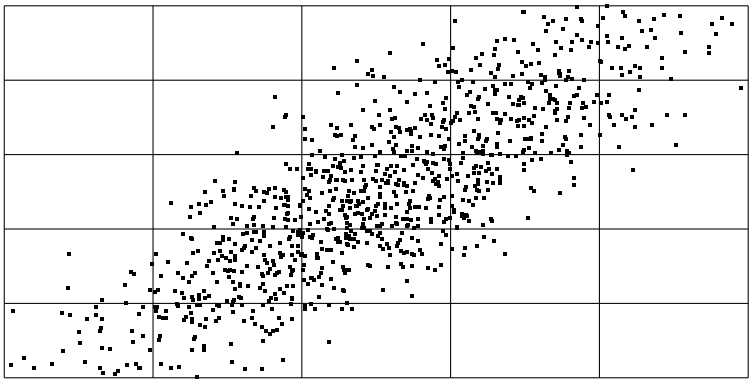

Figure 2.5: Rectangle partition.

(iii) For all $a > 0$,

$$\lim_{n\to\infty} \mathbb{E}\left\{\sum_{i=1}^{n} |W_{n,i}(\mathbf{X})| \mathbb{I}_{\{\|\mathbf{X}_i - \mathbf{X}\| > a\}}\right\} = 0.$$

(iv)

$$\sum_{i=1}^{n} W_{n,i}(\mathbf{X}) \to 1$$

in probability.

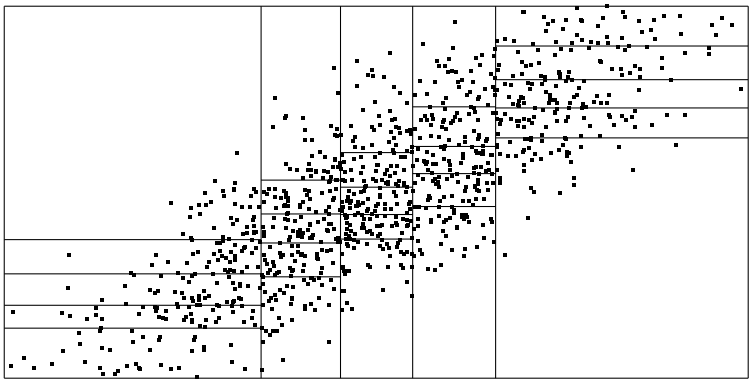

Figure 2.6: Statistically equivalent blocks.

(v)

$$\lim_{n\to\infty} \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{X})^2\right\} = 0.$$

Then the corresponding regression function estimate m_n is weakly universally consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left\{\int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x})\right\} = 0$$

for all distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$.

For nonnegative weights and noiseless data (i.e., $Y = m(\mathbf{X}) \geq 0$) condition (i) says that the mean value of the estimate is bounded above by some constant times the mean value of the regression function. Conditions (ii) and (iv) state that the sum of the weights is bounded and is asymptotically 1. Condition (iii) ensures that the estimate at a point $\mathbf{x}$ is asymptotically influenced only by the data close to $\mathbf{x}$. Condition (v) states that asymptotically all weights become small.

One can verify that under conditions (ii), (iii), (iv), and (v) alone weak consistency holds if the regression function is uniformly continuous and the conditional variance function $\sigma^2(\mathbf{x})$ is bounded. Condition (i) makes the extension possible. For nonnegative weights conditions (i), (iii), and (v) are necessary.

Definition 2.1. *The weights $\{W_{n,i}\}$ are called normal if $\sum_{i=1}^{n} W_{n,i}(\mathbf{x}) = 1$. The weights $\{W_{n,i}\}$ are called subprobability weights if they are nonnegative and sum up to ≤ 1. They are called probability weights if they are nonnegative and sum up to 1.*

Obviously for subprobability weights condition (ii) is satisfied, and for probability weights conditions (ii) and (iv) are satisfied.

PROOF OF THEOREM 2.1. Because of $(a+b+c)^2 \leq 3a^2+3b^2+3c^2$ we have that

$$\begin{aligned}
\mathbb{E}\{m_n(X)-m(X)\}^2 &\leq 3\mathbb{E}\left\{\left(\sum_{i=1}^n W_{n,i}(X)(Y_i-m(X_i))\right)^2\right\} \\
&\quad + 3\mathbb{E}\left\{\left(\sum_{i=1}^n W_{n,i}(X)(m(X_i)-m(X))\right)^2\right\} \\
&\quad + 3\mathbb{E}\left\{\left(\left(\sum_{i=1}^n W_{n,i}(X)-1\right)m(X)\right)^2\right\} \\
&= 3I_n + 3J_n + 3L_n.
\end{aligned}$$

By the Cauchy–Schwarz inequality, and condition (ii),

$$\begin{aligned}
J_n &\leq \mathbb{E}\left\{\left(\sum_{i=1}^n \sqrt{|W_{n,i}(X)|}\sqrt{|W_{n,i}(X)|}\,|m(X_i)-m(X)|\right)^2\right\} \\
&\leq \mathbb{E}\left\{\left(\sum_{i=1}^n |W_{n,i}(X)|\right)\left(\sum_{i=1}^n |W_{n,i}(X)|(m(X_i)-m(X))^2\right)\right\} \\
&\leq D\mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|(m(X_i)-m(X))^2\right\} \\
&= DJ_n'.
\end{aligned}$$

The set of bounded and uniformly continuous functions is dense in L_2, therefore for $\epsilon > 0$ we can choose $\tilde{m}$ bounded and uniformly continuous such that

$$\mathbb{E}\{(m(X)-\tilde{m}(X))^2\} < \epsilon.$$

Then

$$\begin{aligned} J_n' &\leq 3\mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|(m(X_i)-\tilde{m}(X_i))^2\right\} \\ &\quad + 3\mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|(\tilde{m}(X_i)-\tilde{m}(X))^2\right\} \\ &\quad + 3\mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|(\tilde{m}(X)-m(X))^2\right\} \\ &= 3J_{n1}+3J_{n2}+3J_{n3}. \end{aligned}$$

For arbitrary $\delta > 0$,

$$\begin{aligned} J_{n2} &= \mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|\cdot(\tilde{m}(X_i)-\tilde{m}(X))^2 I_{\{\|X_i-X\|>\delta\}}\right\} \\ &\quad + \mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|\cdot(\tilde{m}(X_i)-\tilde{m}(X))^2 I_{\{\|X_i-X\|\leq\delta\}}\right\} \\ &\leq \mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|\cdot(2\tilde{m}(X_i)^2+2\tilde{m}(X)^2) I_{\{\|X_i-X\|>\delta\}}\right\} \\ &\quad + \mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|\cdot(\tilde{m}(X_i)-\tilde{m}(X))^2 I_{\{\|X_i-X\|\leq\delta\}}\right\} \\ &\leq 4\cdot\sup_{u\in\mathbb{R}^d}|\tilde{m}(u)|^2\cdot\mathbb{E}\left\{\sum_{i=1}^n |W_{n,i}(X)|\cdot I_{\{\|X_i-X\|>\delta\}}\right\} \\ &\quad + D\cdot\left(\sup_{u,v\in\mathbb{R}^d\,:\,\|u-v\|\leq\delta}|\tilde{m}(u)-\tilde{m}(v)|\right)^2. \end{aligned}$$

By (iii),

$$\limsup_{n\to\infty} J_{n2} \leq D\cdot\left(\sup_{u,v\in\mathbb{R}^d\,:\,\|u-v\|\leq\delta}|\tilde{m}(u)-\tilde{m}(v)|\right)^2.$$

Using $\tilde{m}$ uniformly continuous we get, with $\delta\to 0$,

$$J_{n2}\to 0.$$

By (ii),

$$J_{n3} \leq D\mathbb{E}\{(\tilde{m}(X) - m(X))^2\} < D\epsilon,$$

moreover, by (i),

$$\limsup_{n\to\infty} J_{n1} \leq c\mathbb{E}\{(\tilde{m}(X) - m(X))^2\} \leq c\epsilon,$$

so

$$\limsup_{n\to\infty} J'_n \leq 3c\epsilon + 3D\epsilon.$$

Put

$$\sigma^2(x) = \mathbb{E}\{(Y - m(X))^2 | X = x\},$$

then $\mathbb{E}Y^2 < \infty$ implies that $\mathbb{E}\sigma^2(X) < \infty$, and

$$\begin{aligned} I_n &= \mathbb{E}\left\{\left(\sum_{i=1}^n W_{n,i}(X)(Y_i - m(X_i))\right)^2\right\} \\ &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\{W_{n,i}(X)W_{n,j}(X)(Y_i - m(X_i))(Y_j - m(X_j))\}. \end{aligned}$$

For $i \neq j$,

$$\begin{aligned} & \mathbb{E}\left\{W_{n,i}(X)W_{n,j}(X)(Y_i - m(X_i))(Y_j - m(X_j))\right\} \\ = \; & \mathbb{E}\left\{\mathbb{E}\left\{W_{n,i}(X)W_{n,j}(X)(Y_i - m(X_i))(Y_j - m(X_j))|X_1, \ldots, X_n, Y_i\right\}\right\} \\ = \; & \mathbb{E}\left\{W_{n,i}(X)W_{n,j}(X)(Y_i - m(X_i))\mathbb{E}\left\{(Y_j - m(X_j))|X_1, \ldots, X_n, Y_i\right\}\right\} \\ = \; & \mathbb{E}\left\{W_{n,i}(X)W_{n,j}(X)(Y_i - m(X_i))(m(X_j) - m(X_j))\right\} \\ = \; & 0, \end{aligned}$$

hence,

$$\begin{aligned} I_n &= \mathbb{E}\left\{\sum_{i=1}^n W_{n,i}(X)^2(Y_i - m(X_i))^2\right\} \\ &= \mathbb{E}\left\{\sum_{i=1}^n W_{n,i}(X)^2\sigma^2(X_i)\right\}. \end{aligned}$$

If $\sigma^2(x)$ is bounded then (v) implies that $I_n \to 0$. Again, for general $\sigma^2(x)$ and $\epsilon > 0$, there exists bounded $\tilde{\sigma}^2(x) \leq L$ such that

$$\mathbb{E}\{|\tilde{\sigma}^2(X) - \sigma^2(X)|\} < \epsilon.$$

Then, by (ii),

$$\begin{aligned} I_n &\leq \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(X)^2 \tilde{\sigma}^2(X_i)\right\} + \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(X)^2 |\sigma^2(X_i) - \tilde{\sigma}^2(X_i)|\right\} \\ &\leq L\mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(X)^2\right\} + D\mathbb{E}\left\{\sum_{i=1}^{n} |W_{n,i}(X)||\sigma^2(X_i) - \tilde{\sigma}^2(X_i)|\right\}, \end{aligned}$$

therefore, by (i) and (v),

$$\limsup_{n\to\infty} I_n \leq cD\mathbb{E}\{|\tilde{\sigma}^2(X) - \sigma^2(X)|\} < cD\epsilon.$$

Concerning the third term

$$L_n = \mathbb{E}\left\{\left(\left(\sum_{i=1}^{n} W_{n,i}(X) - 1\right) m(X)\right)^2\right\} \to 0$$

by conditions (ii), (iv), and by the dominated convergence theorem. □

2.3 Consistency

The purpose of this section is to prove the *weak* universal consistency of the partitioning estimates. This is the first such result that we mention. Later we will prove the same property for other estimates, too. The next theorem provides sufficient conditions for the weak universal consistency of the partitioning estimate. The first condition ensures that the cells of the underlying partition shrink to zero inside a bounded set, so the estimate is local in this sense. The second condition means that the number of cells inside a bounded set is small with respect to n, which implies that with large probability each cell contains many data points.

Theorem 2.2. *If for each sphere S centered at the origin*

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0 \tag{2.1}$$

and

$$\lim_{n\to\infty} \frac{|\{j : A_{n,j} \cap S \neq \emptyset\}|}{n} = 0 \tag{2.2}$$

then the partitioning regression function estimate is weakly universally consistent.

For cubic partitions,

$$\lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n^d = \infty$$

imply (2.1) and (2.2).

In order to prove Theorem 2.2 we will verify the conditions of Stone's theorem. For this we need the following technical lemma. An integer-valued random variable $B(n,p)$ is said to be binomially distributed with parameters n and $0 \le p \le 1$ if

$$\mathbb{P}\{B(n,p) = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Lemma 2.1. *Let the random variable $B(n,p)$ be binomially distributed with parameters n and p. Then:*

(i)

$$\mathbb{E}\left\{\frac{1}{1+B(n,p)}\right\} \le \frac{1}{(n+1)p},$$

(ii)

$$\mathbb{E}\left\{\frac{1}{B(n,p)}\mathbb{I}_{\{B(n,p)>0\}}\right\} \le \frac{2}{(n+1)p}.$$

Proof. Part (i) follows from the following simple calculation:

$$\begin{aligned}
\mathbb{E}\left\{\frac{1}{1+B(n,p)}\right\} &= \sum_{k=0}^{n} \frac{1}{k+1}\binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{1}{(n+1)p} \sum_{k=0}^{n} \binom{n+1}{k+1} p^{k+1} (1-p)^{n-k} \\
&\le \frac{1}{(n+1)p} \sum_{k=0}^{n+1} \binom{n+1}{k} p^k (1-p)^{n-k+1} \\
&= \frac{1}{(n+1)p} \left(p + (1-p)\right)^{n+1} \\
&= \frac{1}{(n+1)p}.
\end{aligned}$$

For (ii) we have

$$\mathbb{E}\left\{\frac{1}{B(n,p)}\mathbb{I}_{\{B(n,p)>0\}}\right\} \leq \mathbb{E}\left\{\frac{2}{1+B(n,p)}\right\} \leq \frac{2}{(n+1)p}$$

by (i). □

Proof of Theorem 2.2. The proof proceeds by checking the conditions of Stone's theorem (Theorem 2.1). Note that if $0/0 = 0$ by definition, then

$$W_{n,i}(\mathbf{x}) = \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}} / \sum_{l=1}^{n} \mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{x})\}}.$$

To verify (i), it suffices to show that there is a constant $c > 0$, such that for any nonnegative function f with $\mathbb{E}f(\mathbf{X}) < \infty$,

$$\mathbb{E}\left\{\sum_{i=1}^{n} f(\mathbf{X}_i)\frac{\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{X})\}}}{\sum_{l=1}^{n}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\right\} \leq c\mathbb{E}f(\mathbf{X}).$$

Observe that

$$\begin{aligned}
& \mathbb{E}\left\{\sum_{i=1}^{n} f(\mathbf{X}_i)\frac{\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{X})\}}}{\sum_{l=1}^{n}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\right\} \\
= & \sum_{i=1}^{n} \mathbb{E}\left\{f(\mathbf{X}_i)\frac{\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{X})\}}}{1+\sum_{l\neq i}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\right\} \\
= & n\mathbb{E}\left\{f(\mathbf{X}_1)\mathbb{I}_{\{\mathbf{X}_1 \in A_n(\mathbf{X})\}}\frac{1}{1+\sum_{l\neq 1}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\right\} \\
= & n\mathbb{E}\left\{\mathbb{E}\left\{f(\mathbf{X}_1)\mathbb{I}_{\{\mathbf{X}_1 \in A_n(\mathbf{X})\}}\frac{1}{1+\sum_{l=2}^{n}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\middle| \mathbf{X}, \mathbf{X}_1\right\}\right\} \\
= & n\mathbb{E}\left\{f(\mathbf{X}_1)\mathbb{I}_{\{\mathbf{X}_1 \in A_n(\mathbf{X})\}}\mathbb{E}\left\{\frac{1}{1+\sum_{l=2}^{n}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\middle| \mathbf{X}, \mathbf{X}_1\right\}\right\} \\
= & n\mathbb{E}\left\{f(\mathbf{X}_1)\mathbb{I}_{\{\mathbf{X}_1 \in A_n(\mathbf{X})\}}\mathbb{E}\left\{\frac{1}{1+\sum_{l=2}^{n}\mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{X})\}}}\middle| \mathbf{X}\right\}\right\}
\end{aligned}$$

by the independence of the random variables $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$. Using Lemma 2.1, the

expected value above can be bounded by

$$\begin{aligned} & n\mathbb{E}\left\{ f(\mathbf{X}_1)\mathbb{I}_{\{\mathbf{X}_1\in A_n(\mathbf{X})\}}\frac{1}{n\mu(A_n(\mathbf{X}))}\right\} \\ = & \sum_j \mathbb{P}\{\mathbf{X}\in A_{nj}\}\int_{A_{nj}} f(u)\mu(du)\frac{1}{\mu(A_{nj})} \\ = & \int_{\mathbb{R}^d} f(u)\mu(du) = \mathbb{E}f(\mathbf{X}). \end{aligned}$$

Therefore, the condition is satisfied with $c = 1$. The weights are sub-probability weights, so (ii) is satisfied. To see that condition (iii) is satisfied first choose a ball S centered at the origin, and then by condition (2.1) a large n such that for $A_{n,j} \cap S \neq \emptyset$ we have $\mathrm{diam}(A_{n,j}) < a$. Thus $\mathbf{X} \in S$ and $\|\mathbf{X}_i - \mathbf{X}\| > a$ imply $\mathbf{X}_i \notin A_n(\mathbf{X})$, therefore

$$\begin{aligned} & \mathbb{I}_{\{\mathbf{X}\in S\}}\sum_{i=1}^n W_{n,i}(\mathbf{X})\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{X}\|>a\}} \\ = & \mathbb{I}_{\{\mathbf{X}\in S\}}\frac{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i\in A_n(\mathbf{X}),\|\mathbf{X}-\mathbf{X}_i\|>a\}}}{n\mu_n(A_n(\mathbf{X}))} \\ = & \mathbb{I}_{\{\mathbf{X}\in S\}}\frac{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i\in A_n(\mathbf{X}),\mathbf{X}_i\notin A_n(\mathbf{X}),\|\mathbf{X}-\mathbf{X}_i\|>a\}}}{n\mu_n(A_n(\mathbf{X}))} \\ = & 0. \end{aligned}$$

Thus

$$\limsup_n \mathbb{E}\sum_{i=1}^n W_{n,i}(\mathbf{X})\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{X}\|>a\}} \leq \mu(S^c).$$

Concerning (iv) note that

$$\begin{aligned} & \mathbb{P}\left\{\sum_{i=1}^n W_{n,i}(\mathbf{X}) \neq 1\right\} \\ = & \mathbb{P}\left\{\mu_n(A_n(\mathbf{X})) = 0\right\} \\ = & \sum_j \mathbb{P}\left\{\mathbf{X}\in A_{n,j}, \mu_n(A_{n,j}) = 0\right\} \\ = & \sum_j \mu(A_{n,j})(1-\mu(A_{n,j}))^n \\ \leq & \sum_{j:A_{n,j}\cap S=\emptyset} \mu(A_{n,j}) + \sum_{j:A_{n,j}\cap S\neq\emptyset} \mu(A_{n,j})(1-\mu(A_{n,j}))^n. \end{aligned}$$

Elementary inequalities

$$x(1-x)^n \le xe^{-nx} \le \frac{1}{en} \qquad (0 \le x \le 1)$$

yield

$$\mathbb{P}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{X}) \neq 1\right\} \le \mu(S^c) + \frac{1}{en}\left|\{j : A_{n,j} \cap S \neq \emptyset\}\right|.$$

The first term on the right-hand side can be made arbitrarily small by the choice of S, while the second term goes to zero by (2.2). To prove that condition (v) holds, observe that

$$\sum_{i=1}^{n} W_{n,i}(\mathbf{x})^2 = \begin{cases} \frac{1}{\sum_{l=1}^{n} \mathbb{I}_{\{\mathbf{X}_l \in A_n(\mathbf{x})\}}} & \text{if } \mu_n(A_n(\mathbf{x})) > 0, \\ 0 & \text{if } \mu_n(A_n(\mathbf{x})) = 0. \end{cases}$$

Then we have

$$\begin{aligned} & \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{X})^2\right\} \\ \le\ & \mathbb{P}\{\mathbf{X} \in S^c\} + \sum_{j : A_{n,j} \cap S \neq \emptyset} \mathbb{E}\left\{\mathbb{I}_{\{\mathbf{X} \in A_{n,j}\}} \frac{1}{n\mu_n(A_{n,j})} \mathbb{I}_{\{\mu_n(A_{n,j})>0\}}\right\} \\ \le\ & \mu(S^c) + \sum_{j : A_{n,j} \cap S \neq \emptyset} \mu(A_{n,j}) \frac{2}{n\mu(A_{n,j})} \\ & \qquad\qquad \text{(by Lemma 2.1)} \\ =\ & \mu(S^c) + \frac{2}{n}\left|\{j : A_{n,j} \cap S \neq \emptyset\}\right|. \end{aligned}$$

A similar argument to the previous one concludes the proof. □

2.4 Rate of convergence

In this section we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for cubic partitions and regression functions which are Lipschitz continuous.

Theorem 2.3. (Györfi et al. (2002)). *For a cubic partition with side length h_n assume that*

$$\mathbb{V}ar(Y|\mathbf{X} = \mathbf{x}) \le \sigma^2,\ \mathbf{x} \in \mathbb{R}^d,$$

$$|m(\mathbf{x}) - m(\mathbf{z})| \leq C\|\mathbf{x} - \mathbf{z}\|,\ \mathbf{x}, \mathbf{z} \in \mathbb{R}^d, \tag{2.3}$$

and that $\mathbf{X}$ *has a compact support* S*. Then*

$$\mathbb{E}\|m_n - m\|^2 \leq \hat{c}\frac{\sigma^2 + \sup_{z\in S}|m(z)|^2}{n \cdot h_n^d} + d \cdot C^2 \cdot h_n^2,$$

where $\hat{c}$ *depends only on* d *and on the diameter of* S*, thus for*

$$h_n = c'\left(\frac{\sigma^2 + \sup_{\mathbf{z}\in S}|m(\mathbf{z})|^2}{C^2}\right)^{1/(d+2)} n^{-1/(d+2)}$$

we get

$$\mathbb{E}\|m_n - m\|^2 \leq c''\left(\sigma^2 + \sup_{\mathbf{z}\in S}|m(\mathbf{z})|^2\right)^{2/(d+2)} C^{2d/(d+2)} n^{-2/(d+2)}.$$

Proof. Set

$$\hat{m}_n(\mathbf{x}) = \mathbb{E}\{m_n(\mathbf{x})|\mathbf{X}_1, \ldots, \mathbf{X}_n\} = \frac{\sum_{i=1}^n m(\mathbf{X}_i)\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{n\mu_n(A_n(\mathbf{x}))}.$$

Then

$$\begin{aligned} & \mathbb{E}\{(m_n(\mathbf{x}) - m(\mathbf{x}))^2|\mathbf{X}_1, \ldots, \mathbf{X}_n\} \\ = \ & \mathbb{E}\{(m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2|\mathbf{X}_1, \ldots, \mathbf{X}_n\} + (\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2. \end{aligned} \tag{2.4}$$

We have

$$\begin{aligned} & \mathbb{E}\{(m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2|\mathbf{X}_1, \ldots, \mathbf{X}_n\} \\ = \ & \mathbb{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(\mathbf{X}_i))\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{n\mu_n(A_n(\mathbf{x}))}\right)^2 \middle| \mathbf{X}_1, \ldots, \mathbf{X}_n\right\} \\ = \ & \frac{\sum_{i=1}^n \mathbb{V}ar(Y_i|\mathbf{X}_i)\mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{(n\mu_n(A_n(\mathbf{x})))^2} \\ \leq \ & \frac{\sigma^2}{n\mu_n(A_n(\mathbf{x}))}\mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))>0\}}. \end{aligned}$$

By Jensen's inequality

$$
\begin{aligned}
(\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 &= \left(\frac{\sum_{i=1}^n (m(\mathbf{X}_i) - m(\mathbf{x})) \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{n\mu_n(A_n(\mathbf{x}))} \right)^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))>0\}} \\
&\quad + m(\mathbf{x})^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))=0\}} \\
&\leq \frac{\sum_{i=1}^n (m(\mathbf{X}_i) - m(\mathbf{x}))^2 \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{n\mu_n(A_n(\mathbf{x}))} \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))>0\}} \\
&\quad + m(\mathbf{x})^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))=0\}} \\
&\leq d \cdot C^2 h_n^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))>0\}} + m(\mathbf{x})^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))=0\}} \\
&\qquad \text{(by (2.3) and } \max_{\mathbf{z} \in A_n(\mathbf{x})} \|\mathbf{x} - \mathbf{z}\|^2 \leq d \cdot h_n^2) \\
&\leq d \cdot C^2 h_n^2 + m(\mathbf{x})^2 \mathbb{I}_{\{n\mu_n(A_n(\mathbf{x}))=0\}}.
\end{aligned}
$$

Without loss of generality assume that S is a cube and the union of $A_{n,1}, \ldots, A_{n,l_n}$ is S. Then

$$
l_n \leq \frac{\tilde{c}}{h_n^d}
$$

for some constant $\tilde{c}$ proportional to the volume of S and, by Lemma 2.1 and (2.4),

$$
\begin{aligned}
&\mathbb{E}\left\{ \int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} \\
= \; &\mathbb{E}\left\{ \int (m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} + \mathbb{E}\left\{ \int (\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} \\
= \; &\sum_{j=1}^{l_n} \mathbb{E}\left\{ \int_{A_{n,j}} (m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} \\
&+ \sum_{j=1}^{l_n} \mathbb{E}\left\{ \int_{A_{n,j}} (\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \mathbb{E}\left\{\int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x})\right\} \\
\leq\ & \sum_{j=1}^{l_n} \mathbb{E}\left\{\frac{\sigma^2 \mu(A_{n,j})}{n\mu_n(A_{n,j})} \mathbb{I}_{\{\mu_n(A_{n,j})>0\}}\right\} + dC^2 h_n^2 \\
& + \sum_{j=1}^{l_n} \mathbb{E}\left\{\int_{A_{n,j}} m(\mathbf{x})^2 \mu(d\mathbf{x}) \mathbb{I}_{\{\mu_n(A_{n,j})=0\}}\right\} \\
\leq\ & \sum_{j=1}^{l_n} \frac{2\sigma^2 \mu(A_{n,j})}{n\mu(A_{n,j})} + dC^2 h_n^2 + \sum_{j=1}^{l_n} \int_{A_{n,j}} m(\mathbf{x})^2 \mu(d\mathbf{x}) \mathbb{P}\{\mu_n(A_{n,j}) = 0\} \\
\leq\ & l_n \frac{2\sigma^2}{n} + dC^2 h_n^2 + \sup_{\mathbf{z}\in S}\left\{m(\mathbf{z})^2\right\} \sum_{j=1}^{l_n} \mu(A_{n,j})(1-\mu(A_{n,j}))^n \\
\leq\ & l_n \frac{2\sigma^2}{n} + dC^2 h_n^2 + l_n \frac{\sup_{\mathbf{z}\in S} m(\mathbf{z})^2}{n} \sup_j n\mu(A_{n,j}) e^{-n\mu(A_{n,j})} \\
\leq\ & l_n \frac{2\sigma^2}{n} + dC^2 h_n^2 + l_n \frac{\sup_{\mathbf{z}\in S} m(\mathbf{z})^2 e^{-1}}{n} \\
& \qquad (\text{since } \sup_z z e^{-z} = e^{-1}) \\
\leq\ & \frac{(2\sigma^2 + \sup_{\mathbf{z}\in S} m(\mathbf{z})^2 e^{-1})\tilde{c}}{nh_n^d} + dC^2 h_n^2.
\end{aligned}
$$

□

Chapter 3

Kernel estimates

3.1 Introduction

Kernel-based rules are derived from the kernel estimate in density estimation originally studied by Parzen (1962), Rosenblatt (1956), Akaike (1954), and Cacoullos (1965); and in regression estimation, introduced by Nadaraya (1964; 1970), and Watson (1964). For particular choices of K, rules of this sort have been proposed by Fix and Hodges (1951; 1952), Sebestyen (1962), Van Ryzin (1966), and Meisel (1969). Statistical analysis of these rules and/or the corresponding regression function estimate can be found in Nadaraya (1964; 1970), Rejtő and Révész (1973), Devroye and Wagner (1976; 1980a; 1980b), Greblicki (1974; 1978b; 1978a), Krzyżak and Pawlak (1984), and Devroye and Krzyżak (1989). Usage of Cauchy kernels in discrimination is investigated by Arkadjew and Braverman (1966), Hand (1981), and Coomans and Broeckaert (1986).

Several authors studied the pointwise properties of the kernel estimates, i.e., the pointwise optimality of the locally polynomial kernel estimates under some regularity conditions on m and μ: Stone (1977; 1980), Katkovnik (1979; 1983; 1985), Korostelev and Tsybakov (1993), Cleveland (1979), Härdle (1990), Fan and Gijbels (1992; 1995), Fan (1993), Tsybakov (1986), and Fan, Hu, and Truong (1994). Kernel regression estimate without bandwidth, called Hilbert kernel estimate, was investigated by Devroye, Györfi, and Krzyżak (1998).

The kernel estimate of a regression function takes the form

$$m_n(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)},$$

if the denominator is nonzero, and 0 otherwise. Here the bandwidth $h_n > 0$ depends

only on the sample size n, and the function $K : \mathbb{R}^d \to [0, \infty)$ is called a kernel. (See Figure 3.1 for some examples.) Usually $K(\mathbf{x})$ is "large" if $\|\mathbf{x}\|$ is "small," therefore the kernel estimate again is a local averaging estimate.

Figures 3.2–3.5 show the kernel estimate for the naive kernel ($K(\mathbf{x}) = \mathbb{I}_{\{\|\mathbf{x}\| \le 1\}}$) and for the Epanechnikov kernel ($K(\mathbf{x}) = (1 - \|\mathbf{x}\|^2)_+$) using various choices for h_n for our simulated data introduced in Chapter 1.

Figure 3.6 shows the L_2 error as a function of h.

3.2 Consistency

In this section we use Stone's theorem (Theorem 2.1) in order to prove the weak universal consistency of kernel estimates under general conditions on h and K.

Theorem 3.1. *Assume that there are balls $S_{0,r}$ of radius r and balls $S_{0,R}$ of radius R centered at the origin ($0 < r \le R$), and constant $b > 0$ such that*

$$\mathbb{I}_{\{\mathbf{x} \in S_{0,R}\}} \ge K(\mathbf{x}) \ge b \mathbb{I}_{\{\mathbf{x} \in S_{0,r}\}}$$

(boxed kernel), and consider the kernel estimate m_n. If $h_n \to 0$ and $n h_n^d \to \infty$, then the kernel estimate is weakly universally consistent.

As one can see in Figure 3.7, the weak consistency holds for a bounded kernel with compact support such that it is bounded away from zero at the origin. The bandwidth must converge to zero but not too fast.

PROOF. Put

$$K_h(\mathbf{x}) = K(\mathbf{x}/h).$$

We check the conditions of Theorem 2.1 for the weights

$$W_{n,i}(\mathbf{x}) = \frac{K_h(\mathbf{x} - \mathbf{X}_i)}{\sum_{j=1}^n K_h(\mathbf{x} - \mathbf{X}_j)}.$$

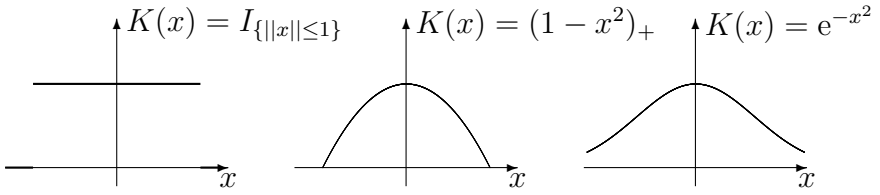

Figure 3.1: Examples for univariate kernels.

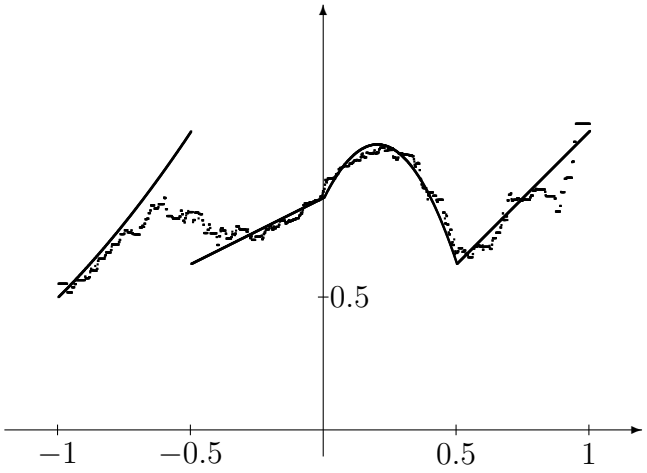

Figure 3.2: Kernel estimate for the naive kernel: $h = 0.1$, L_2 error $= 0.004$.

Condition (i) means that

$$\mathbb{E}\left\{\frac{\sum_{i=1}^{n} K_h(\mathbf{X}-\mathbf{X}_i)f(\mathbf{X}_i)}{\sum_{j=1}^{n} K_h(\mathbf{X}-\mathbf{X}_j)}\right\} \leq c\mathbb{E}\{f(\mathbf{X})\}$$

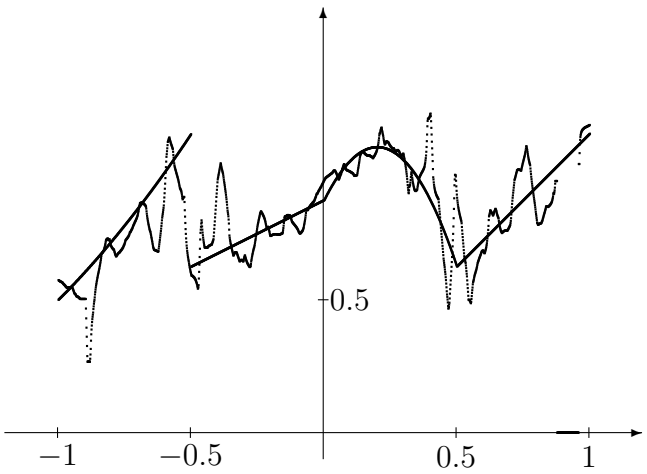

Figure 3.3: Undersmoothing for the Epanechnikov kernel: $h = 0.03$, L_2 error $= 0.032$.

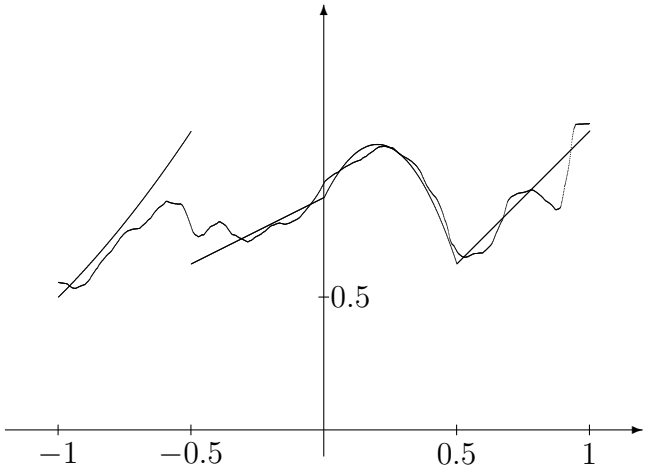

Figure 3.4: Kernel estimate for the Epanechnikov kernel: $h = 0.1$, L_2 error $= 0.004$.

with $c > 0$. Because of

$$\begin{aligned}
& \mathbb{E}\left\{\frac{\sum_{i=1}^n K_h(\mathbf{X}-\mathbf{X}_i)f(\mathbf{X}_i)}{\sum_{j=1}^n K_h(\mathbf{X}-\mathbf{X}_j)}\right\} \\
= \;& n\mathbb{E}\left\{\frac{K_h(\mathbf{X}-\mathbf{X}_1)f(\mathbf{X}_1)}{\sum_{j=1}^n K_h(\mathbf{X}-\mathbf{X}_j)}\right\} \\
= \;& n\mathbb{E}\left\{\frac{K_h(\mathbf{X}-\mathbf{X}_1)f(\mathbf{X}_1)}{K_h(\mathbf{X}-\mathbf{X}_1)+\sum_{j=2}^n K_h(\mathbf{X}-\mathbf{X}_j)}\right\} \\
= \;& n\int f(\mathbf{u})\left[\mathbb{E}\left\{\int \frac{K_h(\mathbf{x}-\mathbf{u})}{K_h(\mathbf{x}-\mathbf{u})+\sum_{j=2}^n K_h(\mathbf{x}-\mathbf{X}_j)}\mu(d\mathbf{x})\right\}\right]\mu(d\mathbf{u})
\end{aligned}$$

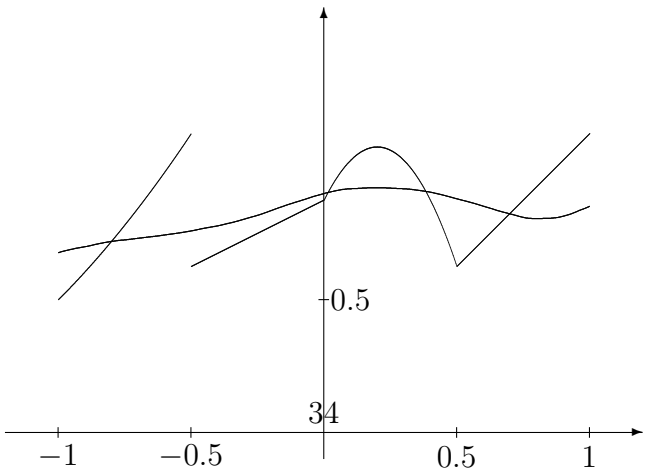

Figure 3.5: Oversmoothing for the Epanechnikov kernel: $h = 0.5$, L_2 error $= 0.013$.

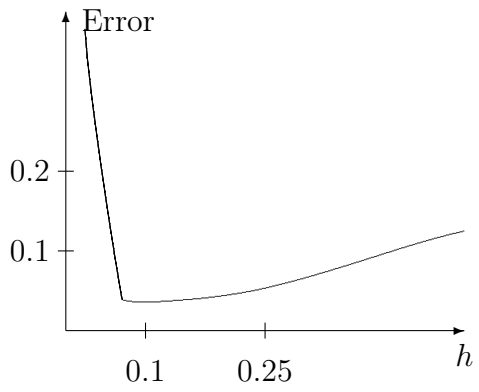

Figure 3.6: The L_2 error for the Epanechnikov kernel as a function of h.

it suffices to show that, for all $\mathbf{u}$ and n,

$$\mathbb{E}\left\{\int \frac{K_h(\mathbf{x}-\mathbf{u})}{K_h(\mathbf{x}-\mathbf{u})+\sum_{j=2}^{n} K_h(\mathbf{x}-\mathbf{X}_j)}\mu(d\mathbf{x})\right\} \leq \frac{c}{n}.$$

The compact support of K can be covered by finitely many balls, with translates of $S_{0,r/2}$, where $r>0$ is the constant appearing in the condition on the kernel K, and with centers $\mathbf{x}_i$, $i=1,2,\ldots,M$. Then, for all $\mathbf{x}$ and $\mathbf{u}$,

$$K_h(\mathbf{x}-\mathbf{u}) \leq \sum_{k=1}^{M} \mathbb{I}_{\{\mathbf{x}\in\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}\}}.$$

Furthermore, $\mathbf{x}\in\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}$ implies that

$$\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2} \subset \mathbf{x}+S_{0,rh}.$$

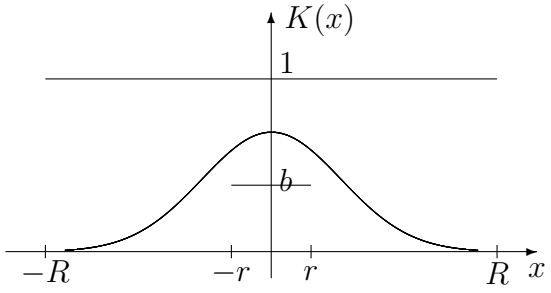

Figure 3.7: Boxed kernel.

Now, by these two inequalities,

$$
\begin{aligned}
& \mathbb{E}\left\{\int \frac{K_h(\mathbf{x}-\mathbf{u})}{K_h(\mathbf{x}-\mathbf{u})+\sum_{j=2}^n K_h(\mathbf{x}-\mathbf{X}_j)}\mu(d\mathbf{x})\right\} \\
\leq & \sum_{k=1}^M \mathbb{E}\left\{\int_{\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}} \frac{K_h(\mathbf{x}-\mathbf{u})}{K_h(\mathbf{x}-\mathbf{u})+\sum_{j=2}^n K_h(\mathbf{x}-\mathbf{X}_j)}\mu(d\mathbf{x})\right\} \\
\leq & \sum_{k=1}^M \mathbb{E}\left\{\int_{\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}} \frac{1}{1+\sum_{j=2}^n K_h(\mathbf{x}-\mathbf{X}_j)}\mu(d\mathbf{x})\right\} \\
\leq & \frac{1}{b}\sum_{k=1}^M \mathbb{E}\left\{\int_{\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}} \frac{1}{1+\sum_{j=2}^n \mathbb{I}_{\{\mathbf{X}_j\in\mathbf{x}+S_{0,rh}\}}}\mu(d\mathbf{x})\right\} \\
\leq & \frac{1}{b}\sum_{k=1}^M \mathbb{E}\left\{\int_{\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}} \frac{1}{1+\sum_{j=2}^n \mathbb{I}_{\{\mathbf{X}_j\in\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}\}}}\mu(d\mathbf{x})\right\} \\
= & \frac{1}{b}\sum_{k=1}^M \mathbb{E}\left\{\frac{\mu(\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2})}{1+\sum_{j=2}^n \mathbb{I}_{\{\mathbf{X}_j\in\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2}\}}}\right\} \\
\leq & \frac{1}{b}\sum_{k=1}^M \frac{\mu(\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2})}{n\mu(\mathbf{u}+h\mathbf{x}_k+S_{0,rh/2})} \\
& \text{(by Lemma 2.1)} \\
\leq & \frac{M}{nb}.
\end{aligned}
$$

The condition (ii) holds since the weights are subprobability weights.

Concerning (iii) notice that, for $h_nR < a$,

$$
\sum_{i=1}^n |W_{n,i}(\mathbf{X})|\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{X}\|>a\}} = \frac{\sum_{i=1}^n K_{h_n}(\mathbf{X}-\mathbf{X}_i)\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{X}\|>a\}}}{\sum_{i=1}^n K_{h_n}(\mathbf{X}-\mathbf{X}_i)} = 0.
$$

In order to show (iv), mention that

$$
1-\sum_{i=1}^n W_{n,i}(\mathbf{X}) = \mathbb{I}_{\left\{\sum_{i=1}^n K_{h_n}(\mathbf{X}-\mathbf{X}_i)=0\right\}},
$$

therefore,

$$\begin{aligned}
\mathbb{P}\left\{1 \neq \sum_{i=1}^{n} W_{n,i}(\mathbf{X})\right\} &= \mathbb{P}\left\{\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i) = 0\right\} \\
&\leq \mathbb{P}\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \notin S_{\mathbf{X},rh_n}\}} = 0\right\} \\
&= \mathbb{P}\left\{\mu_n(S_{\mathbf{X},rh_n}) = 0\right\} \\
&= \int (1-\mu(S_{\mathbf{x},rh_n}))^n \mu(d\mathbf{x}).
\end{aligned}$$

Choose a sphere S centered at the origin, then

$$\begin{aligned}
&\mathbb{P}\left\{1 \neq \sum_{i=1}^{n} W_{n,i}(\mathbf{X})\right\} \\
\leq &\int_S e^{-n\mu(S_{\mathbf{x},rh_n})}\mu(d\mathbf{x}) + \mu(S^c) \\
= &\int_S n\mu(S_{\mathbf{x},rh_n}) e^{-n\mu(S_{\mathbf{x},rh_n})} \frac{1}{n\mu(S_{\mathbf{x},rh_n})}\mu(d\mathbf{x}) + \mu(S^c) \\
= &\max_u ue^{-u} \int_S \frac{1}{n\mu(S_{\mathbf{x},rh_n})}\mu(d\mathbf{x}) + \mu(S^c).
\end{aligned}$$

By the choice of S, the second term can be small. For the first term we can find $\mathbf{z}_1, \ldots, \mathbf{z}_{M_n}$ such that the union of $S_{\mathbf{z}_1,rh_n/2}, \ldots, S_{\mathbf{z}_{M_n},rh_n/2}$ covers S, and

$$M_n \leq \frac{\tilde{c}}{h_n^d}.$$

Then

$$\begin{aligned}
\int_S \frac{1}{n\mu(S_{\mathbf{x},rh_n})}\mu(d\mathbf{x}) &\leq \sum_{j=1}^{M_n} \int \frac{\mathbb{I}_{\{\mathbf{x}\in S_{\mathbf{z}_j,rh_n/2}\}}}{n\mu(S_{\mathbf{x},rh_n})}\mu(d\mathbf{x}) \\
&\leq \sum_{j=1}^{M_n} \int \frac{\mathbb{I}_{\{\mathbf{x}\in S_{\mathbf{z}_j,rh_n/2}\}}}{n\mu(S_{\mathbf{z}_j,rh_n/2})}\mu(d\mathbf{x}) \\
&\leq \frac{M_n}{n} \\
&\leq \frac{\tilde{c}}{nh_n^d} \to 0. \qquad (3.1)
\end{aligned}$$

Concerning (v), since $K(\mathbf{x}) \leq 1$ we get that, for any $\delta > 0$,

$$\begin{aligned}
\sum_{i=1}^{n} W_{n,i}(\mathbf{X})^2 &= \frac{\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i)^2}{(\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i))^2} \\
&\leq \frac{\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i)}{(\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i))^2} \\
&\leq \min\left\{\delta, \frac{1}{\sum_{i=1}^{n} K_{h_n}(\mathbf{X}-\mathbf{X}_i)}\right\} \\
&\leq \min\left\{\delta, \frac{1}{\sum_{i=1}^{n} b\mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}}}\right\} \\
&\leq \delta + \frac{1}{\sum_{i=1}^{n} b\mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}}} \mathbb{I}_{\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}} > 0\right\}},
\end{aligned}$$

therefore it is enough to show that

$$\mathbb{E}\left\{\frac{1}{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}}} \mathbb{I}_{\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}} > 0\right\}}\right\} \to 0.$$

Let S be as above, then

$$\begin{aligned}
& \mathbb{E}\left\{\frac{1}{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}}} \mathbb{I}_{\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}} > 0\right\}}\right\} \\
\leq\ & \mathbb{E}\left\{\frac{1}{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}}} \mathbb{I}_{\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{X},rh_n}\}} > 0\right\}} \mathbb{I}_{\{\mathbf{X} \in S\}}\right\} + \mu(S^c) \\
\leq\ & 2\mathbb{E}\left\{\frac{1}{(n+1)\mu(S_{\mathbf{X},h_n})} \mathbb{I}_{\{\mathbf{X} \in S\}}\right\} + \mu(S^c) \\
& \text{(by Lemma 2.1)} \\
\to\ & \mu(S^c)
\end{aligned}$$

as above. □

3.3 Rate of convergence

In this section we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for a naive kernel and a Lipschitz continuous regression function.

Theorem 3.2. (GYÖRFI ET AL. (2002)). *For a kernel estimate with a naive kernel assume that*

$$\mathbb{V}ar(Y|\mathbf{X}=\mathbf{x}) \leq \sigma^2,\ \mathbf{x} \in \mathbb{R}^d,$$

and

$$|m(\mathbf{x}) - m(\mathbf{z})| \leq C\|\mathbf{x}-\mathbf{z}\|,\ \mathbf{x},\mathbf{z} \in \mathbb{R}^d,$$

and $\mathbf{X}$ *has a compact support* S^**. Then*

$$\mathbb{E}\|m_n - m\|^2 \leq \hat{c}\frac{\sigma^2 + \sup_{\mathbf{z}\in S^*}|m(\mathbf{z})|^2}{n \cdot h_n^d} + C^2 h_n^2,$$

where $\hat{c}$ *depends only on the diameter of* S^* *and on* d*, thus for*

$$h_n = c'\left(\frac{\sigma^2 + \sup_{\mathbf{z}\in S^*}|m(\mathbf{z})|^2}{C^2}\right)^{1/(d+2)} n^{-\frac{1}{d+2}}$$

we have

$$\mathbb{E}\|m_n - m\|^2 \leq c''\left(\sigma^2 + \sup_{\mathbf{z}\in S^*}|m(\mathbf{z})|^2\right)^{2/(d+2)} C^{2d/(d+2)} n^{-2/(d+2)}.$$

PROOF. We proceed similarly to Theorem 2.3. Put

$$\hat{m}_n(\mathbf{x}) = \frac{\sum_{i=1}^n m(\mathbf{X}_i)\mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{n\mu_n(S_{\mathbf{x},h_n})},$$

then we have the decomposition (2.4). If $B_n(\mathbf{x}) = \{n\mu_n(S_{\mathbf{x},h_n}) > 0\}$, then

$$\begin{aligned}
& \mathbb{E}\{(m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2 | \mathbf{X}_1, \ldots, \mathbf{X}_n\} \\
= & \mathbb{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(\mathbf{X}_i))\mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{n\mu_n(S_{\mathbf{x},h_n})}\right)^2 | \mathbf{X}_1, \ldots, \mathbf{X}_n\right\} \\
= & \frac{\sum_{i=1}^n \mathbb{V}ar(Y_i|\mathbf{X}_i)\mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{(n\mu_n(S_{\mathbf{x},h_n}))^2} \\
\leq & \frac{\sigma^2}{n\mu_n(S_{\mathbf{x},h_n})}\mathbb{I}_{B_n(\mathbf{x})}.
\end{aligned}$$

By Jensen's inequality and the Lipschitz property of m,

$$\begin{aligned}
& (\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \\
= & \left(\frac{\sum_{i=1}^n (m(\mathbf{X}_i) - m(\mathbf{x})) \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{n\mu_n(S_{\mathbf{x},h_n})} \right)^2 \mathbb{I}_{B_n(\mathbf{x})} + m(\mathbf{x})^2 \mathbb{I}_{B_n(\mathbf{x})^c} \\
\leq & \frac{\sum_{i=1}^n (m(\mathbf{X}_i) - m(\mathbf{x}))^2 \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{n\mu_n(S_{\mathbf{x},h_n})} \mathbb{I}_{B_n(\mathbf{x})} + m(\mathbf{x})^2 \mathbb{I}_{B_n(\mathbf{x})^c} \\
\leq & C^2 h_n^2 \mathbb{I}_{B_n(\mathbf{x})} + m(\mathbf{x})^2 \mathbb{I}_{B_n(\mathbf{x})^c} \\
\leq & C^2 h_n^2 + m(\mathbf{x})^2 \mathbb{I}_{B_n(\mathbf{x})^c}.
\end{aligned}$$

Using this, together with Lemma 2.1,

$$\begin{aligned}
& \mathbb{E}\left\{ \int (m_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} \\
= & \mathbb{E}\left\{ \int (m_n(\mathbf{x}) - \hat{m}_n(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} + \mathbb{E}\left\{ \int (\hat{m}_n(\mathbf{x}) - m(\mathbf{x}))^2 \mu(d\mathbf{x}) \right\} \\
\leq & \int_{S^*} \mathbb{E}\left\{ \frac{\sigma^2}{n\mu_n(S_{\mathbf{x},h_n})} \mathbb{I}_{\{\mu_n(S_{\mathbf{x},h_n})>0\}} \right\} \mu(d\mathbf{x}) + C^2 h_n^2 \\
& + \int_{S^*} \mathbb{E}\left\{ m(\mathbf{x})^2 \mathbb{I}_{\{\mu_n(S_{\mathbf{x},h_n})=0\}} \right\} \mu(d\mathbf{x}) \\
\leq & \int_{S^*} \frac{2\sigma^2}{n\mu(S_{\mathbf{x},h_n})} \mu(d\mathbf{x}) + C^2 h_n^2 + \int_{S^*} m(\mathbf{x})^2 (1 - \mu(S_{\mathbf{x},h_n}))^n \mu(d\mathbf{x}) \\
\leq & \int_{S^*} \frac{2\sigma^2}{n\mu(S_{\mathbf{x},h_n})} \mu(d\mathbf{x}) + C^2 h_n^2 + \sup_{z \in S^*} m(z)^2 \int_{S^*} e^{-n\mu(S_{\mathbf{x},h_n})} \mu(d\mathbf{x}) \\
\leq & 2\sigma^2 \int_{S^*} \frac{1}{n\mu(S_{\mathbf{x},h_n})} \mu(d\mathbf{x}) + C^2 h_n^2 \\
& + \sup_{\mathbf{z} \in S^*} m(\mathbf{z})^2 \max_u u e^{-u} \int_{S^*} \frac{1}{n\mu(S_{\mathbf{x},h_n})} \mu(d\mathbf{x}).
\end{aligned}$$

Now we refer to (3.1) such that there the set S is a sphere containing S^*. Combining these inequalities the proof is complete. □

Chapter 4

k-nearest-neighbor estimates

4.1 Introduction

The k-nearest neighbor rule, since its conception in 1951 and 1952 (Fix and Hodges (1951; 1952; 1991a; 1991b)), has attracted many followers and continues to be studied by many researchers. For surveys of various aspects of the nearest neighbor or related methods, see Beck (1979), Biau and Devroye (2015), Bhattacharya and Mack (1987), Bickel and Breiman (1983), Cheng (1995), Collomb (1979; 1980; 1981), Cover (1968), Cover and Hart (1967), Dasarathy (1991), Devijver (1980), Devroye (1978; 1981a; 1981b; 1982), Devroye and Györfi (1985), Devroye et al. (1994), Devroye and Wagner (1982), Fritz (1974), Guerre (2000) Györfi (1978) Györfi and Györfi (1975; 1978), Kulkarni and Posner (1995), Mack (1981), Stone (1977), Stute (1984), and Zhao (1987).

Storing the n data pairs in an array and searching for the k nearest neighbors may take time proportional to nkd if done in a naive manner—the "d" accounts for the cost of one distance computation. This complexity may be reduced in terms of one or more of the three factors involved. Typically, with k and d fixed, $O(n^{1/d})$ worst-case time (Papadimitriou and Bentley (1980)) and $O(\log n)$ expected time (Friedman, Bentley, and Finkel (1977)) may be achieved. Multidimensional search trees that partition the space and guide the search are invaluable—for this approach, see Fukunaga and Narendra (1975), Friedman, Bentley, and Finkel (1977), Niemann and Goppert (1988), Kim and Park (1986), and Broder (1990). We refer to a survey in Dasarathy (1991) for more references. Other approaches are described by Yunck (1976), Friedman, Baskett, and Shustek (1975), Vidal (1986), Sethi (1981), and Faragó, Linder, and Lugosi (1993). Generally, with preprocessing, one may considerably reduce the overall complexity in terms of n and d.

We fix $\mathbf{x} \in \mathbb{R}^d$, and reorder the data $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ according to increasing values of $\|\mathbf{X}_i - \mathbf{x}\|$. The reordered data sequence is denoted by

$$(\mathbf{X}_{(1,n)}(\mathbf{x}), Y_{(1,n)}(\mathbf{x})), \ldots, (\mathbf{X}_{(n,n)}(\mathbf{x}), Y_{(n,n)}(\mathbf{x}))$$

or by

$$(\mathbf{X}_{(1,n)}, Y_{(1,n)}), \ldots, (\mathbf{X}_{(n,n)}, Y_{(n,n)})$$

if no confusion is possible. $\mathbf{X}_{(k,n)}(\mathbf{x})$ is called the kth nearest neighbor (k-NN) of $\mathbf{x}$.

The k_n-NN regression function estimate is defined by

$$m_n(\mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y_{(i,n)}(\mathbf{x}).$$

If $\mathbf{X}_i$ and $\mathbf{X}_j$ are equidistant from $\mathbf{x}$, i.e., $\|\mathbf{X}_i - \mathbf{x}\| = \|\mathbf{X}_j - \mathbf{x}\|$, then we have a tie. There are several rules for tie breaking. For example, $\mathbf{X}_i$ might be declared "closer" if $i < j$, i.e., the tie breaking is done by indices. For the sake of simplicity we assume that ties occur with probability 0. In principle, this is an assumption on μ, so the statements are formally not universal, but adding a component to the observation vector $\mathbf{X}$ we can automatically satisfy this condition as follows: Let $(\mathbf{X}, Z)$ be a random vector, where Z is independent of $(\mathbf{X}, Y)$ and uniformly distributed on $[0, 1]$. We also artificially enlarge the data set by introducing $Z_1, Z_2, \ldots, Z_n$, where the Z_i's are i.i.d. uniform $[0, 1]$ as well. Thus, each $(\mathbf{X}_i, Z_i)$ is distributed as $(\mathbf{X}, Z)$. Then ties occur with probability 0. In the sequel we shall assume that $\mathbf{X}$ has such a component and, therefore, for each $\mathbf{x}$ the random variable $\|\mathbf{X} - \mathbf{x}\|^2$ is absolutely continuous, since it is a sum of two independent random variables such that one of the two is absolutely continuous.

Figures 4.1 – 4.3 show k_n-NN estimates for various choices of k_n for our simulated data introduced in Chapter 1. Figure 4.4 shows the L_2 error as a function of k_n.

4.2 Consistency

In this section we use Stone's theorem (Theorem 2.1) in order to prove weak universal consistency of the k-NN estimate. The main result is the following theorem:

Theorem 4.1. *If $k_n \to \infty$, $k_n/n \to 0$, then the k_n-NN regression function estimate is weakly consistent for all distributions of $(\mathbf{X}, Y)$ where ties occur with probability zero and $\mathbb{E}Y^2 < \infty$.*

According to Theorem 4.1 the number of nearest neighbors (k_n), over which one averages in order to estimate the regression function, should on the one hand converge to infinity but should, on the other hand, be small with respect to the sample size n. To verify the conditions of Stone's theorem we need several lemmas.

We will use Lemma 4.1 to verify condition (iii) of Stone's theorem. Denote the probability measure for $\mathbf{X}$ by μ, and let $S_{\mathbf{x},\epsilon}$ be the closed ball centered at $\mathbf{x}$ of radius $\epsilon > 0$. The collection of all $\mathbf{x}$ with $\mu(S_{\mathbf{x},\epsilon}) > 0$ for all $\epsilon > 0$ is called the support of $\mathbf{X}$ or μ. This set plays a key role because of the following property:

Lemma 4.1. *If* $\mathbf{x} \in$ *support*(μ) *and* $\lim_{n\to\infty} k_n/n = 0$*, then*

$$\|\mathbf{X}_{(k_n,n)}(\mathbf{x}) - \mathbf{x}\| \to 0$$

with probability one.

PROOF. Take $\epsilon > 0$. By definition, $\mathbf{x} \in \text{support}(\mu)$ implies that $\mu(S_{\mathbf{x},\epsilon}) > 0$. Observe that

$$\{\|\mathbf{X}_{(k_n,n)}(\mathbf{x}) - \mathbf{x}\| > \epsilon\} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},\epsilon}\}} < \frac{k_n}{n} \right\}.$$

By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},\epsilon}\}} \to \mu(S_{\mathbf{x},\epsilon}) > 0$$

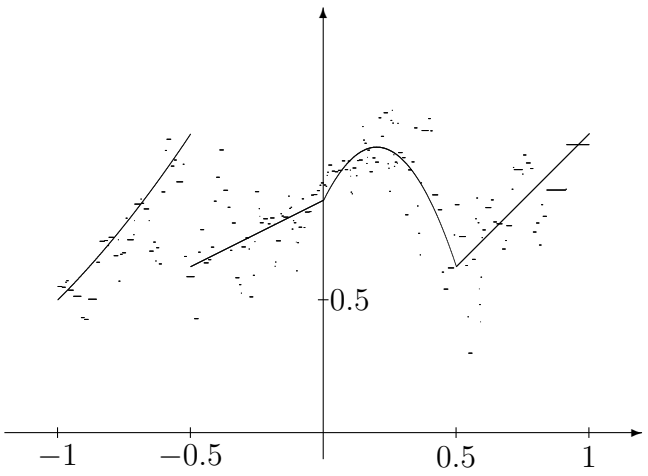

Figure 4.1: Undersmoothing: $k_n = 3$, L_2 error =0.011703.

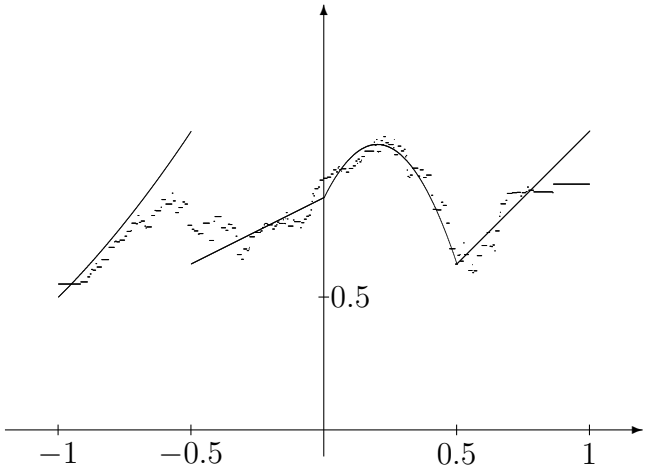

Figure 4.2: Good choice: $k_n = 12$, L_2 error =0.004247.

with probability one, while, by assumption,

$$\frac{k_n}{n} \to 0.$$

Therefore, $\|\mathbf{X}_{(k_n,n)}(\mathbf{x}) - \mathbf{x}\| \to 0$ with probability one. □

The next two lemmas will enable us to establish condition (i) of Stone's theorem.

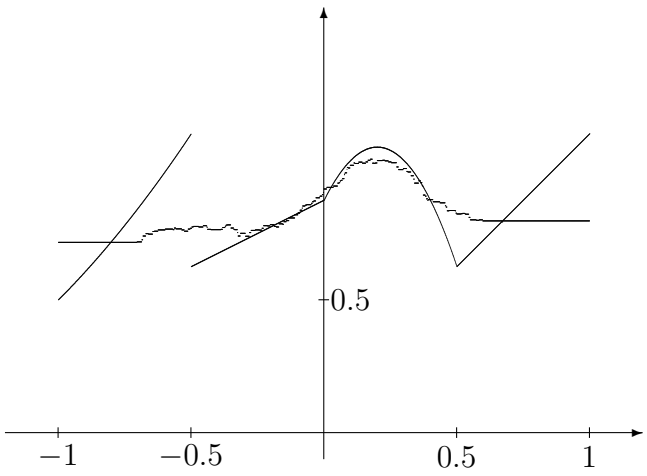

Figure 4.3: Oversmoothing: $k_n = 50$, L_2 error =0.009931.

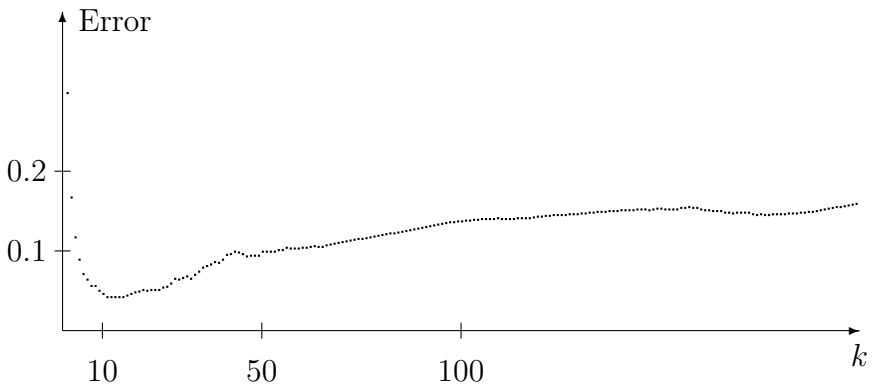

Figure 4.4: L_2 error of the k-NN estimate as a function of k.

Lemma 4.2. *Let*

$$B_a(\mathbf{x}') = \left\{\mathbf{x} : \mu(S_{\mathbf{x},\|\mathbf{x}-\mathbf{x}'\|}) \le a\right\}.$$

Then, for all $\mathbf{x}' \in \mathbb{R}^d$,

$$\mu(B_a(\mathbf{x}')) \le \gamma_d a,$$

where γ_d *depends on the dimension* d *only.*

PROOF. Let $C_j \subset \mathbb{R}^d$ be a cone of angle $\pi/3$ and centered at 0. It is a property of cones that if $\mathbf{u}, \mathbf{u}' \in C_j$ and $\|\mathbf{u}\| < \|\mathbf{u}'\|$, then $\|\mathbf{u} - \mathbf{u}'\| < \|\mathbf{u}'\|$ (cf. Figure 4.5). Let $C_1, \ldots, C_{\gamma_d}$

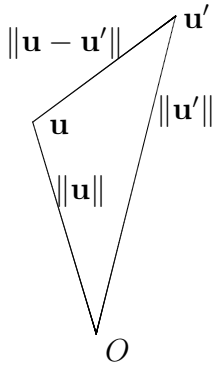

Figure 4.5: The cone property.

be a collection of such cones with different central directions such that their union covers $\mathbb{R}^d$:

$$\bigcup_{j=1}^{\gamma_d} C_j = \mathbb{R}^d.$$

Then

$$\mu(B_a(\mathbf{x}')) \le \sum_{i=1}^{\gamma_d} \mu(\{\mathbf{x}' + C_i\} \cap B_a(\mathbf{x}')).$$

Let $\mathbf{x}^* \in \{\mathbf{x}' + C_i\} \cap B_a(\mathbf{x}')$. Then, by the property of cones mentioned above, we have

$$\mu(\{\mathbf{x}' + C_i\} \cap S_{\mathbf{x}', \|\mathbf{x}' - \mathbf{x}^*\|} \cap B_a(\mathbf{x}')) \le \mu(S_{\mathbf{x}^*, \|\mathbf{x}' - \mathbf{x}^*\|}) \le a,$$

where we use the fact that $\mathbf{x}^* \in B_a(\mathbf{x}')$. Since $\mathbf{x}^*$ is arbitrary,

$$\mu(\{\mathbf{x}' + C_i\} \cap B_a(\mathbf{x}')) \le a,$$

which completes the proof of the lemma. □

An immediate consequence of the lemma is that the number of points among $\mathbf{X}_1, \ldots, \mathbf{X}_n$, such that $\mathbf{X}$ is one of their k nearest neighbors, is not more than a constant times k.

Corollary 4.1. *Assume that ties occur with probability zero. Then*

$$\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X} \textit{ is among the } k \textit{ NNs of } \mathbf{X}_i \textit{ in } \{\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n\}\}} \le k\gamma_d$$

a.s.

PROOF. Apply Lemma 4.2 with $a = k/n$ and let μ be the empirical measure μ_n of $\mathbf{X}_1, \ldots, \mathbf{X}_n$, i.e., for each Borel set $A \subseteq \mathbb{R}^d$, $\mu_n(A) = (1/n) \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in A\}}$. Then

$$B_{k/n}(\mathbf{X}) = \left\{\mathbf{x} \ : \ \mu_n(S_{\mathbf{x}, \|\mathbf{x} - \mathbf{X}\|}) \le k/n\right\}$$

and

$$\begin{aligned} & \mathbf{X}_i \in B_{k/n}(\mathbf{X}) \\ \Leftrightarrow \quad & \mu_n(S_{\mathbf{X}_i, \|\mathbf{X}_i - \mathbf{X}\|}) \le k/n \\ \Leftrightarrow \quad & \mathbf{X} \text{ is among the } k \text{ NNs of } \mathbf{X}_i \text{ in } \{\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_n\} \end{aligned}$$

a.s., where for the second $\Leftrightarrow$ we applied the condition that ties occur with probability zero. This, together with Lemma 4.2, yields

$$\begin{aligned}
& \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X} \text{ is among the } k \text{ NNs of } \mathbf{X}_i \text{ in } \{\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{X},\mathbf{X}_{i+1},\ldots,\mathbf{X}_n\}\}} \\
= & \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in B_{k/n}(\mathbf{X})\}} \\
= & \; n \cdot \mu_n(B_{k/n}(\mathbf{X})) \\
\leq & \; k\gamma_d
\end{aligned}$$

a.s. □

Lemma 4.3. *Assume that ties occur with probability zero. Then for any integrable function f, any n, and any $k \leq n$,*

$$\sum_{i=1}^{k} \mathbb{E}\left\{|f(\mathbf{X}_{(i,n)}(\mathbf{X}))|\right\} \leq k\gamma_d \mathbb{E}\{|f(\mathbf{X})|\},$$

where γ_d depends upon the dimension only.

PROOF. If f is a nonnegative function,

$$\begin{aligned}
& \sum_{i=1}^{k} \mathbb{E}\left\{f(\mathbf{X}_{(i,n)}(\mathbf{X}))\right\} \\
= & \; \mathbb{E}\left\{\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \text{ is among the k NNs of } \mathbf{X} \text{ in } \{\mathbf{X}_1,\ldots,\mathbf{X}_n\}\}} f(\mathbf{X}_i)\right\} \\
= & \; \mathbb{E}\left\{f(\mathbf{X}) \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X} \text{ is among the k NNs of } \mathbf{X}_i \text{ in } \{\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{X},\mathbf{X}_{i+1},\ldots,\mathbf{X}_n\}\}}\right\} \\
& \text{(by exchanging } \mathbf{X} \text{ and } \mathbf{X}_i) \\
\leq & \; \mathbb{E}\{f(\mathbf{X})k\gamma_d\},
\end{aligned}$$

by Corollary 4.1. This concludes the proof of the lemma. □

PROOF OF THEOREM 4.1. We proceed by checking the conditions of Stone's weak convergence theorem (Theorem 2.1) under the condition that ties occur with probability

zero. The weight $W_{n,i}(\mathbf{X})$ in Theorem 2.1 equals $1/k_n$ if $\mathbf{X}_i$ is among the k_n nearest neighbors of $\mathbf{X}$, and equals 0 otherwise, thus the weights are probability weights, and (ii) and (iv) are automatically satisfied. Condition (v) is obvious since $k_n \to \infty$. For condition (iii) observe that, for each $\epsilon > 0$,

$$\begin{aligned}
& \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{X})\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{X}\|>\epsilon\}}\right\} \\
= & \int \mathbb{E}\left\{\sum_{i=1}^{n} W_{n,i}(\mathbf{x})\mathbb{I}_{\{\|\mathbf{X}_i-\mathbf{x}\|>\epsilon\}}\right\}\mu(d\mathbf{x}) \\
= & \int \mathbb{E}\left\{\frac{1}{k_n}\sum_{i=1}^{k_n} \mathbb{I}_{\{\|\mathbf{X}_{(i,n)}(\mathbf{x})-\mathbf{x}\|>\epsilon\}}\right\}\mu(d\mathbf{x}) \to 0
\end{aligned}$$

holds whenever

$$\int \mathbb{P}\left\{\|\mathbf{X}_{(k_n,n)}(\mathbf{x}) - \mathbf{x}\| > \epsilon\right\}\mu(d\mathbf{x}) \to 0, \tag{4.1}$$

where $\mathbf{X}_{(k_n,n)}(\mathbf{x})$ denotes the k_nth nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$. For $\mathbf{x} \in$ *support*(μ), $k_n/n \to 0$, together with Lemma 4.1, implies

$$\mathbb{P}\left\{\|\mathbf{X}_{(k_n,n)}(\mathbf{x}) - \mathbf{x}\| > \epsilon\right\} \to 0 \quad (n \to \infty).$$

This together with the dominated convergence theorem implies (4.1). Finally, we consider condition (i). It suffices to show that for any nonnegative measurable function f with $\mathbb{E}\{f(\mathbf{X})\} < \infty$, and any n,

$$\mathbb{E}\left\{\sum_{i=1}^{n} \frac{1}{k_n}\mathbb{I}_{\{\mathbf{X}_i \text{ is among the } k_n \text{ NNs of } \mathbf{X}\}} f(\mathbf{X}_i)\right\} \le c \cdot \mathbb{E}\left\{f(\mathbf{X})\right\}$$

for some constant c. But we have shown in Lemma 4.3 that this inequality always holds with $c = \gamma_d$. Thus, condition (i) is verified. □

4.3 Rate of convergence

In this section we bound the rate of convergence of $\mathbb{E}\|m_n - m\|^2$ for a k_n-nearest neighbor estimate.

Theorem 4.2. (GYÖRFI ET AL. (2002)). *Assume that* $\mathbf{X}$ *is bounded,*

$$\sigma^2(\mathbf{x}) = \mathbb{V}ar(Y|\mathbf{X}=\mathbf{x}) \le \sigma^2 \quad (\mathbf{x} \in \mathbb{R}^d)$$

and

$$|m(\mathbf{x}) - m(\mathbf{z})| \le C\|\mathbf{x}-\mathbf{z}\| \quad (\mathbf{x},\mathbf{z} \in \mathbb{R}^d).$$

Assume that $d \ge 3$. *Let* m_n *be the* k_n*-NN estimate. Then*

$$\mathbb{E}\|m_n - m\|^2 \le \frac{\sigma^2}{k_n} + c_1 \cdot C^2 \left(\frac{k_n}{n}\right)^{2/d},$$

thus for $k_n = c' \left(\sigma^2/C^2\right)^{d/(2+d)} n^{\frac{2}{d+2}}$,

$$\mathbb{E}\|m_n - m\|^2 \le c'' \sigma^{\frac{4}{d+2}} C^{\frac{2d}{2+d}} n^{-\frac{2}{d+2}}.$$

For the proof of Theorem 4.2 we need the rate of convergence of nearest neighbor distances.

Lemma 4.4. *Assume that* $\mathbf{X}$ *is bounded. If* $d \ge 3$, *then*

$$\mathbb{E}\{\|\mathbf{X}_{(1,n)}(\mathbf{X}) - \mathbf{X}\|^2\} \le \frac{\tilde{c}}{n^{2/d}}.$$

PROOF. For fixed $\epsilon > 0$,

$$\mathbb{P}\{\|\mathbf{X}_{(1,n)}(\mathbf{X}) - \mathbf{X}\| > \epsilon\} = \mathbb{E}\{(1-\mu(S_{\mathbf{X},\epsilon}))^n\}.$$

Let $A_1, \ldots, A_{N(\epsilon)}$ be a cubic partition of the bounded support of μ such that the A_j's have diameter ϵ and

$$N(\epsilon) \le \frac{c}{\epsilon^d}.$$

If $\mathbf{x} \in A_j$, then $A_j \subset S_{\mathbf{x},\epsilon}$, therefore

$$\begin{aligned}
\mathbb{E}\{(1-\mu(S_{\mathbf{X},\epsilon}))^n\} &= \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1-\mu(S_{\mathbf{x},\epsilon}))^n \mu(d\mathbf{x}) \\
&\le \sum_{j=1}^{N(\epsilon)} \int_{A_j} (1-\mu(A_j))^n \mu(d\mathbf{x}) \\
&= \sum_{j=1}^{N(\epsilon)} \mu(A_j)(1-\mu(A_j))^n.
\end{aligned}$$

Obviously,

$$\begin{aligned}\sum_{j=1}^{N(\epsilon)} \mu(A_j)(1-\mu(A_j))^n &\leq \sum_{j=1}^{N(\epsilon)} \max_z z(1-z)^n \\ &\leq \sum_{j=1}^{N(\epsilon)} \max_z z e^{-nz} \\ &= \frac{e^{-1}N(\epsilon)}{n}.\end{aligned}$$

If L stands for the diameter of the support of μ, then

$$\begin{aligned}\mathbb{E}\{\|\mathbf{X}_{(1,n)}(\mathbf{X})-\mathbf{X}\|^2\} &= \int_0^\infty \mathbb{P}\{\|\mathbf{X}_{(1,n)}(\mathbf{X})-\mathbf{X}\|^2 > \epsilon\}\, d\epsilon \\ &= \int_0^{L^2} \mathbb{P}\{\|\mathbf{X}_{(1,n)}(\mathbf{X})-\mathbf{X}\| > \sqrt{\epsilon}\}\, d\epsilon \\ &\leq \int_0^{L^2} \min\left\{1, \frac{e^{-1}N(\sqrt{\epsilon})}{n}\right\} d\epsilon \\ &\leq \int_0^{L^2} \min\left\{1, \frac{c}{en}\epsilon^{-d/2}\right\}\, d\epsilon \\ &= \int_0^{(c/(en))^{2/d}} 1\, d\epsilon + \frac{c}{en}\int_{(c/(en))^{2/d}}^{L^2} \epsilon^{-d/2} d\epsilon \\ &\leq \frac{\tilde{c}}{n^{2/d}}\end{aligned}$$

for $d \geq 3$. □

Proof of Theorem 4.2. We have the decomposition

$$\begin{aligned}\mathbb{E}\{(m_n(\mathbf{x})-m(\mathbf{x}))^2\} &= \mathbb{E}\{(m_n(\mathbf{x})-\mathbb{E}\{m_n(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_n\})^2\} \\ &\quad + \mathbb{E}\{(\mathbb{E}\{m_n(\mathbf{x})|\mathbf{X}_1,\ldots,\mathbf{X}_n\}-m(\mathbf{x}))^2\} \\ &= I_1(\mathbf{x}) + I_2(\mathbf{x}).\end{aligned}$$

The first term is easier:

$$\begin{aligned} I_1(\mathbf{x}) &= \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}\left(Y_{(i,n)}(\mathbf{x}) - m(\mathbf{X}_{(i,n)}(\mathbf{x}))\right)\right)^2\right\} \\ &= \mathbb{E}\left\{\frac{1}{k_n^2}\sum_{i=1}^{k_n}\sigma^2(\mathbf{X}_{(i,n)}(\mathbf{x}))\right\} \\ &\leq \frac{\sigma^2}{k_n}. \end{aligned}$$

For the second term

$$\begin{aligned} I_2(\mathbf{x}) &= \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}(m(\mathbf{X}_{(i,n)}(\mathbf{x})) - m(\mathbf{x}))\right)^2\right\} \\ &\leq \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}|m(\mathbf{X}_{(i,n)}(\mathbf{x})) - m(\mathbf{x})|\right)^2\right\} \\ &\leq \mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{i=1}^{k_n}C\|\mathbf{X}_{(i,n)}(\mathbf{x}) - \mathbf{x}\|\right)^2\right\}. \end{aligned}$$

Put $N = k_n\lfloor\frac{n}{k_n}\rfloor$. Split the data $\mathbf{X}_1, \ldots, \mathbf{X}_n$ into $k_n + 1$ segments such that the first k_n segments have length $\lfloor\frac{n}{k_n}\rfloor$, and let $\tilde{\mathbf{X}}_j^{\mathbf{x}}$ be the first nearest neighbor of $\mathbf{x}$ from the jth segment. Then $\tilde{\mathbf{X}}_1^{\mathbf{x}}$, …, $\tilde{\mathbf{X}}_{k_n}^{\mathbf{x}}$ are k_n different elements of $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, which implies

$$\sum_{i=1}^{k_n}\|\mathbf{X}_{(i,n)}(\mathbf{x}) - \mathbf{x}\| \leq \sum_{j=1}^{k_n}\|\tilde{\mathbf{X}}_j^{\mathbf{x}} - \mathbf{x}\|,$$

therefore, by Jensen's inequality,

$$\begin{aligned} I_2(\mathbf{x}) &\leq C^2\mathbb{E}\left\{\left(\frac{1}{k_n}\sum_{j=1}^{k_n}\|\tilde{\mathbf{X}}_j^{\mathbf{x}}-\mathbf{x}\|\right)^2\right\} \\ &\leq C^2\frac{1}{k_n}\sum_{j=1}^{k_n}\mathbb{E}\left\{\|\tilde{\mathbf{X}}_j^{\mathbf{x}}-\mathbf{x}\|^2\right\} \\ &= C^2\mathbb{E}\left\{\|\tilde{\mathbf{X}}_1^{\mathbf{x}}-\mathbf{x}\|^2\right\} \\ &= C^2\mathbb{E}\left\{\|\mathbf{X}_{(1,\lfloor\frac{n}{k_n}\rfloor)}(\mathbf{x})-\mathbf{x}\|^2\right\}. \end{aligned}$$

Thus, by Lemma 4.4,

$$\begin{aligned} \frac{1}{C^2}\left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\int I_2(\mathbf{x})\mu(d\mathbf{x}) &\leq \left\lfloor\frac{n}{k_n}\right\rfloor^{2/d}\mathbb{E}\left\{\|\mathbf{X}_{(1,\lfloor\frac{n}{k_n}\rfloor)}(\mathbf{X})-\mathbf{X}\|^2\right\} \\ &\leq const. \end{aligned}$$

□

Chapter 5

Splitting the sample

In the previous chapters the parameters of the estimates with the optimal rate of convergence depend on the unknown distribution of (X, Y), especially on the smoothness of the regression function. In this and in the following chapter we present data-dependent choices of the smoothing parameters. We show that for bounded Y the estimates with parameters chosen in such an adaptive way achieve the optimal rate of convergence.

5.1 Best random choice of a parameter

Let $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be the sample as before. Assume a finite set $\mathcal{Q}_n$ of parameters such that for every parameter $h \in \mathcal{Q}_n$ there is a regression function estimate $m_n^{(h)}(\cdot) = m_n^{(h)}(\cdot, D_n)$. Let $\hat{h} = \hat{h}(D_n) \in \mathcal{Q}_n$ be such that

$$\int |m_n^{(\hat{h})}(x) - m(x)|^2 \mu(dx) = \min_{h \in \mathcal{Q}_n} \int |m_n^{(h)}(x) - m(x)|^2 \mu(dx),$$

where $\hat{h}$ is called the best random choice of the parameter. Obviously, $\hat{h}$ is not an estimate, it depends on the unknown m and μ.

This best random choice can be approximated by splitting the data. Let $D_{n_l} = \{(X_1, Y_1), \dots, (X_{n_l}, Y_{n_l})\}$ be the learning (training) data of size n_l and $D_n \backslash D_{n_l}$ the testing data of size n_t ($n = n_l + n_t \geq 2$). For every parameter $h \in \mathcal{Q}_n$ let $m_{n_l}^{(h)}(\cdot) = m_{n_l}^{(h)}(\cdot, D_{n_l})$ be an estimate of m depending only on the learning data D_{n_l} of the sample D_n. Use the testing data to choose a parameter $H = H(D_n) \in \mathcal{Q}_n$:

$$\frac{1}{n_t} \sum_{i=n_l+1}^{n_l+n_t} |m_{n_l}^{(H)}(X_i) - Y_i|^2 = \min_{h \in \mathcal{Q}_n} \frac{1}{n_t} \sum_{i=n_l+1}^{n_l+n_t} |m_{n_l}^{(h)}(X_i) - Y_i|^2. \tag{5.1}$$

Define the estimate by

$$m_n(x) = m_n(x, D_n) = m_{n_l}^{(H)}(x, D_{n_l}). \tag{5.2}$$

We show that H approximates the best random choice $\hat{h}$ in the sense that $\mathbb{E}\int |m_n(x) - m(x)|^2 \mu(dx)$ approximates $\mathbb{E}\int |m_{n_l}^{(\hat{h})}(x) - m(x)|^2 \mu(dx)$.

Theorem 5.1. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Let $0 < L < \infty$. Assume*

$$|Y| \le L \quad a.s. \tag{5.3}$$

and

$$\max_{h \in \mathcal{Q}_n} \|m_{n_l}^{(h)}\|_\infty \le L \quad a.s. \tag{5.4}$$

Then, for any $\delta > 0$,

$$\begin{aligned} &\mathbb{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\ &\le (1+\delta)\mathbb{E}\int |m_{n_l}^{(\hat{h})}(x) - m(x)|^2 \mu(dx) + c\frac{1 + \log(|\mathcal{Q}_n|)}{n_t}, \end{aligned} \tag{5.5}$$

where $\hat{h} = \hat{h}(D_{n_l})$ and $c = L^2(16/\delta + 35 + 19\delta)$.

The only assumption on the underlying distribution in Theorem 5.1 is the boundedness of $|Y|$ (cf. (5.3)). It can be applied to any estimate which is bounded in supremum norm by the same bound as the data (cf. (5.4)). We can always truncate an estimate at $\pm L$, which implies that (5.4) holds. If (5.3) holds, then the regression function will be bounded in absoulte value by L, too, and hence the L_2 error of the truncated estimate will be less than or equal to the L_2 error of the original estimate, so the truncation has no negative consequence in view of the error of the estimate.

In the next section we will apply this theorem to partitioning, kernel, and nearest neighbor estimates. We will choose $\mathcal{Q}_n$ and n_t such that the second term on the right-hand side of (5.5) is less than the first term. This implies that the expected L_2 error of the estimate is bounded by some constant times the expected L_2 error of an estimate, which is applied to a data set of size n_l (rather than n) and where the parameter is chosen in an optimal way for this data set. Observe that this is not only true asymptotically, but true for each finite sample size.

5.2 Partitioning, kernel, and nearest neighbor estimates

In Theorems 2.3, 3.2, and 4.2 we showed that partitioning, kernel, and nearest neighbor estimates are able to achieve the minimax lower bound for the estimation of (p, C)-smooth regression functions if $p = 1$ and if the parameters are chosen depending on C (the Lipschitz constant of the regression function). Obviously, the value of C will be unknown in an application, therefore, one cannot use estimates where the parameters depend on C in applications. In the sequel we show that, in the case of bounded data, one can also derive similar bounds for estimates where the parameters are chosen by splitting the sample.

We start with the kernel estimate. Let $m_n^{(h)}$ be the kernel estimate with naive kernel and bandwidth h. We choose the finite set $\mathcal{Q}_n$ of bandwidths such that we can approach the choice of the bandwidth in Theorem 3.2 up to some factor less than some constant, e.g., up to factor 2. This can be done, e.g., by setting

$$\mathcal{Q}_n = \left\{2^k \ : \ k \in \{-n, -(n-1), \ldots, 0, \ldots, n-1, n\}\right\}.$$

Theorems 5.1 and 3.2 imply

Corollary 5.1. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that X is bounded,*

$$|m(x) - m(z)| \le C \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d)$$

and $|Y| \le L$ a.s. Set

$$n_l = \left\lceil \frac{n}{2} \right\rceil \text{ and } n_t = n - n_l.$$

Let m_n be the kernel estimate with naive kernel and bandwidth $h \in \mathcal{Q}_n$ chosen as in Theorem 5.1, where $\mathcal{Q}_n$ is defined as above. Then $(\log n)^{(d+2)/(2d)} n^{-1/2} \le C$ implies, for $n \ge 2$,

$$\mathbb{E} \int |m_n(x) - m(x)|^2 \mu(dx) \le c_1 C^{2d/(d+2)} n^{-2/(d+2)}$$

for some constant c_1 which depends only on L, d, and the diameter of the support of X.

Proof. Without loss of generality we can assume $C \le n^{1/d}$ (otherwise, the assertion is

trivial because of boundedness of Y). Theorems 5.1 and 3.2 imply

$$\begin{aligned}
&\mathbb{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq 2 \min_{h \in \mathcal{Q}_n} \mathbb{E}\int |m_{n_l}^{(h)}(x) - m(x)|^2 \mu(dx) + c \cdot \frac{1 + \log(|\mathcal{Q}_n|)}{n_t} \\
&\leq 2 \min_{h \in \mathcal{Q}_n} \left(\hat{c} \cdot \frac{2L^2}{n_l h^d} + C^2 h^2 \right) + c \cdot \frac{1 + \log(2n+1)}{n_t} \\
&\leq 2 \left(\hat{c} \cdot \frac{2L^2}{n_l h_n^d} + C^2 h_n^2 \right) + c \cdot \frac{1 + \log(2n+1)}{n_t},
\end{aligned}$$

where $h_n \in \mathcal{Q}_n$ is chosen such that

$$C^{-2/(d+2)} n^{-1/(d+2)} \leq h_n \leq 2C^{-2/(d+2)} n^{-1/(d+2)}.$$

The choices of h_n, n_l, and n_t together with $C \geq (\log n)^{(d+2)/(2d)} n^{-1/2}$ imply

$$\begin{aligned}
&\mathbb{E}\int |m_n(x) - m(x)|^2 \mu(dx) \\
&\leq \tilde{c} \cdot C^{2d/(d+2)} n^{-2/(d+2)} + 4c \cdot \frac{1 + \log(2n+1)}{n} \\
&\leq c_1 \cdot C^{2d/(d+2)} n^{-2/(d+2)}.
\end{aligned}$$

□

Similarly, one can show the following result concerning the partitioning estimate:

Corollary 5.2. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that X is bounded,*

$$|m(x) - m(z)| \leq C \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d)$$

and $|Y| \leq L$ a.s. Set

$$n_l = \left\lceil \frac{n}{2} \right\rceil \text{ and } n_t = n - n_l.$$

Let m_n be the partitioning estimate with cubic partition and grid size $h \in \mathcal{Q}_n$ chosen as in Theorem 5.1, where $\mathcal{Q}_n$ is defined as above. Then $(\log n)^{(d+2)/(2d)} n^{-1/2} \leq C$ implies, for $n \geq 2$,

$$\mathbb{E}\int |m_n(x) - m(x)|^2 \mu(dx) \leq c_2 C^{2d/(d+2)} n^{-2/(d+2)}$$

for some constant c_2 which depends only on L, d, and the diameter of the support of X.

Finally we consider the k-nearest neighbor estimates. Here we can set $\mathcal{Q}_n = \{1, \dots, n\}$, so the optimal value from Theorem 4.2 is contained in $\mathcal{Q}_n$. Immediately from Theorems 5.1 and 4.2 we can conclude

Corollary 5.3. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that X is bounded,*

$$|m(x) - m(z)| \leq C \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d)$$

and $|Y| \leq L$ a.s. Set

$$n_l = \left\lceil \frac{n}{2} \right\rceil \text{ and } n_t = n - n_l.$$

Let m_n be the k-nearest neighbor estimate with $k \in \mathcal{Q}_n = \{1, \dots, n_l\}$ chosen as in Theorem 5.1. Then $(\log n)^{(d+2)/(2d)} n^{-1/2} \leq C$ together with $d \geq 3$ implies, for $n \geq 2$,

$$\mathbb{E} \int |m_n(x) - m(x)|^2 \mu(dx) \leq c_3 C^{2d/(d+2)} n^{-2/(d+2)}$$

for some constant c_3 which depends only on L, d, and the diameter of the support of X.

Here we use for each component of X the same smoothing parameter. But the results can be extended to optimal scaling, where one uses for each component a different smoothing parameter. Splitting of the data can be used to approximate the optimal scaling parameters, which depend on the underlying distribution.

In Corollaries 5.1–5.3 the expected L_2 error of the estimates is bounded from above up to a constant by the corresponding minimax lower bound for (p, C)-smooth regression functions, if $p = 1$. We would like to mention two important aspects of these results: First, the definition of the estimates does not depend on C, therefore they adapt automatically to the unknown smoothness of the regression function measured by the Lipschitz constant C. Second, the bounds are valid for finite sample size. So we are able to approach the minimax lower bound not only asymptotically but even for finite sample sizes.

Approaching the minimax lower bound for fixed sample size by some constant does not imply that one can get asymptotically the minimax rate of convergence with the optimal constant in front of $n^{-2p/(2p+d)}$. But as we show in the next theorem, this goal can also be reached by splitting the sample:

Theorem 5.2. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Under the conditions of Theorem 5.1 assume that*

$$\log |\mathcal{Q}_n| \leq \tilde{c} \log n$$

and

$$\mathbb{E}\left\{\min_{h\in\mathcal{Q}_n}\int|m_n^{(h)}(x)-m(x)|^2\mu(dx)\right\}\le C_{opt}(1+o(1))n^{-\gamma}$$

for some $0<\gamma<1$. *Choose* $\gamma<\gamma'<1$ *and set*

$$n_t=\left\lceil n^{\gamma'}\right\rceil \quad \textit{and} \quad n_l=n-n_t.$$

Then

$$\mathbb{E}\int|m_n(x)-m(x)|^2\mu(dx)\le C_{opt}(1+o(1))n^{-\gamma}.$$

PROOF. Theorem 5.1 implies that

$$\begin{aligned}
&\mathbb{E}\int|m_n(x)-m(x)|^2\mu(dx)\\
\le\;& (1+\delta)\mathbb{E}\left\{\min_{h\in\mathcal{Q}_n}\int|m_{n_l}^{(h)}(x)-m(x)|^2\mu(dx)\right\}+c\frac{1+\log(|\mathcal{Q}_n|)}{n_t}\\
\le\;& (1+\delta)C_{opt}(1+o(1))n_l^{-\gamma}+c\frac{1+\tilde{c}\log n}{n_t}\\
\le\;& (1+\delta)C_{opt}(1+o(1))(1-o(1))^{-\gamma}n^{-\gamma}+c\frac{1+\tilde{c}\log n}{n^{\gamma'}}\\
&\qquad(\text{since } n_l=n-\lceil n^{\gamma'}\rceil \text{ and } n-n^{\gamma'}=(1-n^{-(1-\gamma')})\cdot n\ )\\
=\;& (1+\delta)C_{opt}(1+o(1))n^{-\gamma}.
\end{aligned}$$

Since $\delta>0$ is arbitrary we get that

$$\mathbb{E}\int|m_n(x)-m(x)|^2\mu(dx)\le C_{opt}(1+o(1))n^{-\gamma}.$$

□

Chapter 6

Cross-validation

6.1 Best deterministic choice of the parameter

Let $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be the sample as before. Assume a finite set $\mathcal{Q}_n$ of parameters such that for every parameter $h \in \mathcal{Q}_n$ there is a regression function estimate $m_n^{(h)}(\cdot) = m_n^{(h)}(\cdot, D_n)$. Let $\bar{h}_n \in \mathcal{Q}_n$ be such that

$$\mathbb{E}\left\{\int |m_n^{(\bar{h}_n)}(x) - m(x)|^2 \mu(dx)\right\} = \min_{h \in \mathcal{Q}_n} \mathbb{E}\left\{\int |m_n^{(h)}(x) - m(x)|^2 \mu(dx)\right\},$$

where $\bar{h}_n$ is called the best deterministic choice of the parameter. Obviously, $\bar{h}_n$ is not an estimate, it depends on the unknown distribution of (X, Y), in particular on m and μ.

This best deterministic choice can be approximated by cross-validation. For every parameter $h \in \mathcal{Q}_n$ let $m_n^{(h)}$ and $m_{n,i}^{(h)}$ be the regression estimates from D_n and $D_n \backslash (X_i, Y_i)$, respectively, where

$$D_n \backslash (X_i, Y_i) = \{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)\}.$$

The cross-validation selection of h is

$$H = H_n = \arg\min_{h \in \mathcal{Q}_n} \frac{1}{n} \sum_{i=1}^{n} (m_{n,i}^{(h)}(X_i) - Y_i)^2.$$

Define the cross-validation regression estimate by

$$m_n(x) = m_n^{(H)}(x). \tag{6.1}$$

Throughout this chapter we use the notation

$$\Delta_n^{(h)} = \mathbb{E} \int |m_n^{(h)}(x) - m(x)|^2 \mu(dx).$$

In the sequel we show that H_n approximates the best deterministic choice $\bar{h} = \bar{h}_{n-1}$ for sample size $n-1$ in the sense that $\mathbb{E}\left\{\Delta_{n-1}^{(H_n)}\right\}$ approximates $\Delta_{n-1}^{(\bar{h}_{n-1})}$ with an asymptotically small correction term.

6.2 Partitioning and kernel estimates

Theorem 6.1 yields relations between $\mathbb{E}\left\{\Delta_{n-1}^{(H_n)}\right\}$ and $\Delta_{n-1}^{(\bar{h}_{n-1})}$.

Theorem 6.1. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Let $|Y| \leq L < \infty$. Choose $m_n^{(h)}$ of the form*

$$m_n^{(h)}(x) = \frac{\sum_{j=1}^n Y_j K_h(x, X_j)}{\sum_{j=1}^n K_h(x, X_j)}$$

where the binary valued function $K_h : \mathbb{R}^d \times \mathbb{R}^d \to \{0,1\}$ with $K_h(x,x) = 1$ fulfills the covering assumption (C) that a constant $\rho > 0$ depending only on $\{K_h; h \in \cup_n \mathcal{Q}_n\}$ exists with

$$\int \frac{K_h(x,z)}{\int K_h(x,t)\mu(dt)} \mu(dx) \leq \rho$$

for all $z \in \mathbb{R}^d$, all $h \in \cup_n \mathcal{Q}_n$, and all probability measures μ.
(a)

$$\mathbb{E}\left\{\Delta_{n-1}^{(H_n)}\right\} \leq \Delta_{n-1}^{(\bar{h}_{n-1})} + c\sqrt{\frac{\log(|\mathcal{Q}_n|)}{n}}$$

for some constant c depending only on L and ρ.
(b) *For any $\delta > 0$*

$$\mathbb{E}\left\{\triangle_{n-1}^{(H_n)}\right\} \leq (1+\delta)\, \triangle_{n-1}^{(\bar{h}_{n-1})} + c\frac{|\mathcal{Q}_n|}{n} \log n,$$

where c depends only on δ, L, and ρ.

The covering assumption (C) in Theorem 6.1 is fulfilled for kernel estimates using naive kernel and partitioning estimates (see below). Before we consider the application of

Theorem 6.1 to these estimates in detail, we give some comments concerning convergence order.

Neglecting $\log n$, the correction terms in parts (a) and (b) are both of the order $n^{-1/2}$ if $|\mathcal{Q}_n| = O(n^{1/2})$. One is interested that the correction term is less than $\triangle_{n-1}^{(\overline{h}_{n-1})}$. For Lipschitz-continuous m one has

$$\triangle_{n-1}^{(\overline{h}_{n-1})} = O\left(n^{-2/(d+2)}\right)$$

in naive kernel estimation and cubic partitioning estimation according to Theorems 3.2 and 2.3, respectively. In this case, for $d \geq 3$ and $\log(|\mathcal{Q}_n|) = O(\log n)$, i.e., $|\mathcal{Q}_n| \leq n^s$ for some $s > 0$, or for $\log(|\mathcal{Q}_n|) = O(n^t)$ for some $0 < t < (d-2)/(d+2)$, part (a) yields the desired result, and for $d \geq 1$ and $\log(|\mathcal{Q}_n|) \leq c^* \log n$ with $c^* < d/(d+2)$, part (b) yields the desired result. The latter also holds if

$$\triangle_{n-1}^{(\overline{h}_{n-1})} = O(n^{-\gamma})$$

with $\gamma < 1$ near to 1, if c^* is chosen sufficiently small.

Now we give more detailed applications of Theorem 6.1 to kernel and partitioning estimates. Let $\mathcal{P}_h$ be a partition of $\mathbb{R}^d$, and denote by $m_n^{(h)}$ the partitioning estimate for this partition and sample size n. Because of the proof of Theorem 2.2 the covering assumption (C) is satisfied with $\rho = 1$:

$$\begin{aligned}\int \frac{I_{\{z \in A_n(x)\}}}{\mu(A_n(x))} \mu(dx) &= \int \frac{I_{\{z \in A_n(x)\}}}{\mu(A_n(z))} \mu(dx) \\ &= \frac{\mu(A_n(z))}{\mu(A_n(z))} \\ &= 1.\end{aligned}$$

Or, let

$$m_n^{(h)}(x) = \frac{\sum_{j=1}^n Y_j K\left(\frac{x - X_j}{h}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)}.$$

be the kernel estimate with bandwidth h and naive kernel K. The covering assumption is satisfied with a ρ depending on d only.

For these estimates Theorem 6.1, together with Theorems 3.2 and 2.3, implies

Corollary 6.1. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that X is bounded,*

$$|m(x) - m(z)| \leq C \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d)$$

and $|Y| \le L$ *a.s.*

Let m_n *be the partitioning estimate with cubic partitioning and grid size* $h \in \mathcal{Q}_n$ *chosen as in Theorem 6.1, or let* m_n *be the kernel estimate with naive kernel and bandwidth* $h \in \mathcal{Q}_n$ *chosen as in Theorem 6.1. Let* $d \ge 3$*,*

$$\mathcal{Q}_n = \left\{2^k \ : \ k \in \{-n, -(n-1), \dots, 0, \dots, n-1, n\}\right\}$$

and

$$(\log n)^{(d+2)/(4d)} n^{-(d-2)/(4d)} \le C,$$

or, let $d \ge 1$*,*

$$\mathcal{Q}_n = \left\{\lceil 2^{-n^{1/4}+k} \rceil \ : \ k \in \{1, 2, \dots, 2\lceil n^{1/4} \rceil\}\right\}$$

and

$$(\log n)^{(d+2)/(2d)} n^{-(3d-2)/(8d)} \le C.$$

Then, in each of the four cases,

$$\mathbb{E}\left\{\Delta_{n-1}^{(H_n)}\right\} \le c_1 C^{2d/(d+2)} n^{-2/(d+2)}$$

for some constant c_1 *which depends only on* L*,* d*, and the diameter of the support of* X*.*

As in the previous chapter the results can be extended to optimal scaling and adapting to the optimal constant in front of $n^{-2/(d+2)}$.

6.3 Nearest neighbor estimates

Theorem 6.1 cannot be applied for a nearest neighbor estimate. Let $m_n^{(k)}$ be the k-NN estimate for sample size $n \ge 2$. Then $h = k$ can be considered as a parameter, and we choose $\mathcal{Q}_n = \{1, \dots, n\}$. Let m_n denote the cross-validation nearest neighbor estimate, i.e., put

$$H = H_n = \arg\min_h \frac{1}{n} \sum_{i=1}^{n} (m_{n,i}^{(h)}(X_i) - Y_i)^2$$

and

$$m_n = m_n^{(H)}.$$

For the nearest neighbor estimate again we have covering (Corollary 4.1) with $\rho = \gamma_d$.

Theorem 6.2. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that* $|Y| \leq L$. *Then, for the cross-validation nearest neighbor estimate* m_n,

$$\mathbb{E}\{\Delta_{n-1}^{(H_n)}\} \leq \Delta_{n-1}^{(\bar{h}_{n-1})} + c\sqrt{\frac{\log n}{n}}$$

for some constant c *depending only on* L *and* γ_d.

Theorems 6.2 and 4.2 imply

Corollary 6.2. *(*Györfi, Kohler, Krzyzak, Walk (2002) *) Assume that* X *is bounded,*

$$|m(x) - m(z)| \leq C \cdot \|x - z\| \quad (x, z \in \mathbb{R}^d)$$

and $|Y| \leq L$ *a.s. Let* m_n *be the* k*-nearest neighbor estimate with* k *chosen as in Theorem 6.2. Then for* $d \geq 3$ *and*

$$(\log n)^{(d+2)/(4d)} n^{-(d-2)/(4d)} \leq C,$$

one has

$$\mathbb{E}\{\Delta_{n-1}^{(H_n)}\} \leq c_1 C^{2d/(d+2)} n^{-2/(d+2)}$$

for some constant c_1 *which depends only on* L, d, *and the diameter of the support of* X.

Chapter 7

Estimating the residual variance

7.1 Introduction

In this chapter we study the problem of estimating the smallest achievable mean-squared error in regression function estimation in multivariate problems. We introduce and analyze a nearest neighbor-based estimate of the second moment of the regression function. The second moment of the regression function is closely tied to the best possible achievable mean squared error. It is shown that the estimate is asymptotically normally distributed. It is remarkable that the asymptotic variance only depends on conditional moments of the regression function but not on its smoothness. Moreover, the non-asymptotic variance is bounded by a constant that is independent of the dimension. We also establish a non-asymptotic exponential concentration inequality. We illustrate these results studying variable selection. In particular, we construct and analyze a test for deciding whether a component of the observational vector has predictive power.

The formal setup is as follows. Let $(\mathbf{X}, Y)$ be a pair of random variables such that $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ takes values in $\mathbb{R}^d$ and Y is a real-valued random variable with $\mathbb{E}[Y^2] < \infty$. We denote by μ the distribution of the observation vector $\mathbf{X}$, that is, for all measurable sets $A \subset \mathbb{R}^d$, $\mu(A) = \mathbb{P}\{\mathbf{X} \in A\}$. Then the *regression function*

$$m(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] \tag{7.1}$$

is well defined for μ-almost all $\mathbf{x}$. The center of our investigations is the functional

$$L^* = \mathbb{E}\left[(m(\mathbf{X}) - Y)^2\right] \ .$$

The importance of this functional stems from the fact that for each measurable function $g : \mathbb{R}^d \to \mathbb{R}$ one has

$$\mathbb{E}\left[(g(\mathbf{X}) - Y)^2\right] = L^* + \mathbb{E}\left[(m(\mathbf{X}) - g(\mathbf{X}))^2\right]$$

and, in particular,

$$L^* = \min_g \mathbb{E}\left[(g(\mathbf{X}) - Y)^2\right] \ ,$$

where the minimum is taken over all measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. In other words, L^* is the minimal mean squared error of any "predictor" of Y based on observing $\mathbf{X}$. L^* is often referred to as the *residual variance*.

In regression analysis the residual variance L^* is of obvious interest as it provides a lower bound for the performance of any regression function estimator. In this chapter we study the problem of estimating L^* based on data consisting of independent, identically distributed (i.i.d.) copies of the pair $(\mathbf{X}, Y)$. It is convenient to assume that the number of samples is even and the $2n$ samples are split into two halves as

$$D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\} \qquad \text{and} \qquad D'_n = \{(\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)\}$$

such that the $2n+1$ pairs $(\mathbf{X}, Y), (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n), (\mathbf{X}'_1, Y'_1), \ldots, (\mathbf{X}'_n, Y'_n)$ are independent and identically distributed.

An estimator $\widehat{L}_n$ of L^* is simply a function of the data D_n, D'_n. We are interested in "nonparametric" estimators of L^* that work under minimal assumptions on the underlying distribution. In particular, a desirable feature of any estimate is that it is strongly universally consistent, that is, $\widehat{L}_n \to L^*$ with probability one, for all possible distributions of $(\mathbf{X}, Y)$ with $\mathbb{E}Y^2 < \infty$. Such estimators may be constructed, for example, by constructing a strongly universally consistent regression function estimator m_n based on the data D_n (i.e., a function m_n is such that $\mathbb{E}[(m_n(\mathbf{X}) - Y)^2 | D_n] \to L^*$ with probability one for all distributions) and estimating its mean squared error by $(1/n)\sum_{i=1}^n (m_n(\mathbf{X}'_i) - Y'_i)^2$. (For a detailed theory of universally consistent regression function estimation see Györfi et al. (2002).) However, the rate of convergence of such estimators is determined by the rate of convergence of the mean squared error of m_n which can be quite slow even under regularity assumptions on the underlying distribution. Estimating the entire regression function $m(\mathbf{x})$ is, intuitively, "harder" than estimating the value of L^*. Indeed, nearest-neighbor-based estimators of L^* have been constructed and analyzed by Devroye, Ferrario, Györfi, and Walk (2013), Devroye, Schäfer, Györfi, and Walk (2003), Evans and Jones (2008), Liitiäinen, Corona, and Lendasse (2008), (2010), Liitiäinen, Verleysen, Corona, and Lendasse (2009), and Ferrario and Walk (2012). These estimates have been shown to have a faster rate of convergence—under some natural assumptions– than estimates based on estimating the error of consistent regression function estimators. Moreover, the estimate in Devroye, Ferrario, Györfi, and Walk (2013) is strongly universally consistent.

In this chapter we introduce yet another universally consistent nearest-neighbor-based estimator of L^*. The advantage of this estimator, apart from sharing the fast rates of

convergence of previously defined estimators, is that its random fluctuations may be bounded by dimension-, and distribution-independent quantities. In particular, we prove a central limit theorem and a distribution-free upper bound for the variance for the new estimator that show that it is concentrated around its expected value in an interval of width $O(1/\sqrt{n})$, independently of the dimension. The established concentration property is crucial in a variable-selection procedure that we discuss as an application. In particular, we design a test for deciding whether exclusion of a certain component of $\mathbf{X}$ increases L^* or not.

The chapter is organized as follows. In Section 7.2 we introduce a novel estimate of L^* and establish some of its properties such as asymptotic normality and a non-asymptotic concentration inequality. The central limit theorem holds without any smoothness condition on the regression function, and the asymptotic variance depends only on the conditional moments of Y (Theorem 7.1). We prove a non-asymptotic bound on the variance that does not depend on the dimension of $\mathbf{X}$ (Theorem 7.2), and show an exponential concentration inequality for the centered estimate (Theorem 7.3). All these results are universal in the sense that we only assume that $\mathbf{X}$ has a density and Y is bounded.

In Section 7.3 we briefly describe how the results method based on the results of Section 7.2 may be relevant for variable selection. Finally, the proofs are presented in Section 7.4.

7.2 A nearest-neighbor based estimate and its asymptotic normality

Denoting the second moment of the regression function by

$$S^* = \mathbb{E}\left[m(\mathbf{X})^2\right] \ ,$$

we have

$$L^* = \mathbb{E}\left[Y^2\right] - S^* \ ,$$

and therefore estimating L^* is essentially equivalent to estimating S^* (as the "easy" part $\mathbb{E}\left[Y^2\right]$ may be estimated by, e.g., $(1/n)\sum_{i=1}^n Y_i^2$ whose behavior is well understood).

Next we introduce a nearest neighbor-based estimator of S^*. Based on the data D_n, we start by constructing a nearest-neighbor (1-NN) regression function estimator as follows. Let $\mathbf{X}_{1,n}(\mathbf{x})$ be the first nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$ (with respect to the Euclidean distance in $\mathbb{R}^d$) and let $Y_{1,n}(\mathbf{x})$ be its label. (In order to rigorously define the nearest neighbor, we assume that ties are broken in order to favor points with

smaller index. Since we assume the distribution of $\mathbf{X}$ to be absolutely continuous, this issue is immaterial since ties occur with probability zero.) The 1-NN estimator of the regression function m is defined as

$$m_n(\mathbf{x}) = Y_{1,n}(\mathbf{x}) \ .$$

The proposed estimate of S^* is

$$S_n = \frac{1}{n} \sum_{i=1}^{n} Y_i' m_n(\mathbf{X}_i') \ .$$

By a straightforward adjustment of the arguments of Devroye, Ferrario, Györfi, and Walk (2013), one may show that S_n is a strongly universal consistent estimate of S^*, that is,

$$\lim_n S_n = S^*$$

with probability one for any distribution of $(\mathbf{X}, Y)$ with $\mathbb{E}[Y^2] < \infty$. Note that the consistent functional estimate S_n is based on a non-consistent regression function estimate m_n.

Next we establish asymptotic normality of S_n under the condition that the response variable Y is bounded. In order to describe the asymptotic variance, we introduce the dimension-dependent constant $\alpha(d)$ as follows.

Let $B_{\mathbf{x},r}$ denote the closed ball of radius $r > 0$ centered at $\mathbf{x}$ in $\mathbb{R}^d$ and let λ denote the Lebesgue measure on $\mathbb{R}^d$. Let V be a random vector uniformly distributed in $B_{0,1}$. Define $\overline{1} = (1, 0, 0, \ldots, 0) \in \mathbb{R}^d$ and let $\overline{B} = B_{\overline{1},1} \bigcup B_{V,\|V\|}$. Introduce the random variable

$$W = \frac{\lambda(\overline{B})}{\lambda(B_{0,1})}$$

and define

$$\alpha(d) = \mathbb{E}\left[\frac{2}{W^2}\right] \ . \tag{7.2}$$

Theorem 7.1. (Devroye, Györfi, Lugosi, and Walk (2018)) *Assume that μ has a density and that there exists a constant $L > 0$ such that*

$$\mathbb{P}\{|Y| < L\} = 1 \ . \tag{7.3}$$

Denote

$$M_2(\mathbf{X}) = \mathbb{E}[Y^2 \mid \mathbf{X}]$$

and define

$$\sigma_1^2 = \int M_2(\mathbf{x})^2 \mu(d\mathbf{x}) - \left(\int m(\mathbf{x})^2 \mu(d\mathbf{x}) \right)^2$$

and

$$\sigma_2^2 = \alpha(d) \left(\int M_2(\mathbf{x}) m(\mathbf{x})^2 \mu(d\mathbf{x}) - \int m(\mathbf{x})^4 \mu(d\mathbf{x}) \right) .$$

If $\sigma_1 > 0$*, then*

$$\sqrt{n} \left(S_n - \mathbb{E}\{S_n\} \right) / \sigma \xrightarrow{\mathcal{D}} N(0,1) ,$$

where

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 .$$

The dependence of the asymptotic variance on the dimension d is weak, merely via the constant $\alpha(d)$. Given $\mathbf{X}_1, \ldots, \mathbf{X}_n$, Devroye, Györfi, Lugosi, and Walk (2017) considered the probability measures of the Voronoi cells. They proved that the asymptotic variance of n-times the probability measure of the Voronoi cell is equal to $\alpha(d) - 1$. Thus, this asymptotic variance is universal in the sense that it does not depend on the underlying density. A few values are $\alpha(1) = 1.5$, $\alpha(2) \approx 1.28$, $\alpha(3) \approx 1.18$. In general, $1 \leq \alpha(d) \leq 2$ and $\alpha(d) \to 1$ exponentially fast as $d \to \infty$. Thus, by (7.3) we have $\sigma^2 \leq 3L^4$, and therefore Theorem 7.1 implies that

$$\limsup_{n \to \infty} n \mathbb{V}ar(S_n) \leq 3L^4.$$

The next theorem shows that, up to a constant factor, this bound holds non-asymptotically.

Theorem 7.2. (Devroye, Györfi, Lugosi, and Walk (2018)) *Assume that* μ *has a density and that* $|Y| < L$*. Then for all* $n \geq 1$*,*

$$\mathbb{V}ar(S_n) \leq \frac{9 \cdot L^4}{n}.$$

The next result is a non-asymptotic exponential inequality that extends Theorem 7.2. It implies that for all $t > 0$,

$$\mathbb{P}\left\{ \sqrt{n} |S_n - \mathbb{E}S_n| > t \right\} \leq c e^{-\left(t/(cL^2)\right)^{2/3}}$$

for a universal constant $c > 0$. It is an interesting open question whether the right-hand side can be improved to $e^{-\left(t/(cL^2)\right)^2}$. This would give a non-asymptotic analog of the central limit theorem of Theorem 7.1.

Theorem 7.3. (Devroye, Györfi, Lugosi, and Walk (2018)) *Assume that μ has a density and that $|Y| < L$. Write*

$$S_n - \mathbb{E}[S_n] = U_n + V_n$$

with

$$U_n = S_n - \mathbb{E}[S_n \mid D_n] \quad \text{and} \quad V_n := \mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n] \ .$$

Then for every $n \geq 1$ and $\epsilon > 0$, we have

$$\mathbb{P}\{|U_n| > \epsilon\} \leq 2e^{-n\epsilon^2/(2L^4)}$$

and

$$\mathbb{P}\{|V_n| \geq \epsilon\} \leq 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1} \ . \tag{7.4}$$

The proofs of Theorems 7.1, 7.2 and 7.3 are presented in Section 7.4.

7.3 Illustration: testing for dimension reduction

In standard nonparametric regression design, one considers a finite number of real-valued features $X^{(i)}$, $i \in I \subset \{1, \ldots, d\}$ for predicting the value of a response variable Y. A first question one may try to answer is whether these features suffice to explain Y. In case they do, an estimation method can be applied on the basis of the features already under consideration. Otherwise more or different features need to be considered. The quality of a subvector $\{X^{(i)}, i \in I\}$ of $\mathbf{X}$ is measured by the minimum mean squared error

$$L^*(I) := \mathbb{E}\left[Y - \mathbb{E}[Y \mid X^{(i)} : i \in I]\right]^2$$

that can be achieved using the features as explanatory variables. $L^*(I)$ depends upon the unknown distribution of $(Y, X^{(i)} : i \in I)$.

Thus, even before a regression function estimate is chosen, one may be interested in estimating L^*. For possible dimensionality reduction, one needs, in general, to test the hypothesis

$$L^* = L^*(I) \tag{7.5}$$

for a particular (proper) subset I of $\{1, \ldots, d\}$. A natural way of approaching this testing problem is by estimating both L^* and $L^*(I)$, and accept the hypothesis if the two estimates are close to each other (De Brabanter, Ferrario and Györfi (2014)).

Introduce the notation

$$S^*(I) := \mathbb{E}\left[\mathbb{E}[Y \mid X^{(i)}, i \in I]^2\right] \ .$$

Then the hypothesis (7.5) is equivalent to

$$S^* = S^*(I) \ .$$

Without loss of generality, consider the case $I = \{1, \ldots, d-1\}$, that is, the case when one tests whether the last component $X^{(d)}$ of the observation vector $(X^{(1)}, \ldots, X^{(d)})$ is ineffective. Let the transformation T be defined by

$$T((x^{(1)}, \ldots, x^{(d)})) = (x^{(1)}, \ldots, x^{(d-1)}) \ .$$

Thus, dropping the component $X^{(d)}$ from the observation vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$ leads to the observation vector

$$\widehat{\mathbf{X}} = T(\mathbf{X}) = (X^{(1)}, \ldots, X^{(d-1)})$$

of dimension $d-1$.

Using the notation

$$m(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}] \text{ and } \widetilde{m}(T(\mathbf{X})) = \mathbb{E}[Y \mid T(\mathbf{X})]$$

and

$$S^* = \mathbb{E}[m(\mathbf{X})^2] \text{ and } \widehat{S}^* = \mathbb{E}[\widetilde{m}(T(\mathbf{X}))^2] \ ,$$

the null-hypothesis $\widehat{S}^* = S^*$ is equivalent to

$$m(\mathbf{X}) = \widetilde{m}(T(\mathbf{X})) \quad \text{with probability one.} \tag{7.6}$$

We propose to approach this testing problem by considering the nearest-neighbor estimate defined in Section 7.2. Let S_n be the estimate of S^* using the sample

$$\mathcal{D}_{2n} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{2n}, Y_{2n})\} \ .$$

Assume that an independent sample of size $2n$ is available:

$$\overline{\mathcal{D}}_{2n} = \{(\overline{\mathbf{X}}_1, \overline{Y}_1), \ldots, (\overline{\mathbf{X}}_{2n}, \overline{Y}_{2n})\} \ .$$

We use $\overline{\mathcal{D}}_{2n}$ to construct an estimate $\widetilde{S}_n$ of $\widehat{S}^*$. $\widetilde{S}_n$ is defined as the nearest-neighbor estimate computed from the sample

$$\{(T(\overline{\mathbf{X}}_1), \overline{Y}_1), \ldots, (T(\overline{\mathbf{X}}_{2n}), \overline{Y}_{2n})\} \ .$$

The proposed test is based of the test statistic

$$T_n = S_n - \widetilde{S}_n$$

and accepts the null hypothesis (7.6) if and only if

$$T_n \le a_n := \omega_n \left(n^{-1/2} + n^{-2/d} \right)$$

where ω_n is an increasing unbounded sequence such that $a_n \to 0$. Under the alternative hypothesis, according to the consistency result of Devroye, Ferrario, Györfi, and Walk (2013), for bounded Y,

$$T_n \to S^* - \widehat{S}^* > 0 \quad \text{with probability one,} \tag{7.7}$$

and this convergence is universal, that is, it holds without any conditions. Thus, since $a_n \to 0$, if $\widehat{S}^* \neq S^*$, then, with probability one, the test does not make any mistake for a sufficiently large n.

Theorem 7.1 implies that

$$\sqrt{n} \left(S_n - \mathbb{E}S_n \right) / \sigma \xrightarrow{\mathcal{D}} N(0,1)$$

and

$$\sqrt{n} \left(\widetilde{S}_n - \mathbb{E}\widetilde{S}_n \right) / \widetilde{\sigma} \xrightarrow{\mathcal{D}} N(0,1)$$

with $\sigma^2, \widetilde{\sigma}^2 < 3L^4$. Since S_n and $\widetilde{S}_n$ are independent, we have

$$\sqrt{n}(T_n - \mathbb{E}T_n)/(\sqrt{\sigma^2 + \widetilde{\sigma}^2}) \xrightarrow{\mathcal{D}} N(0,1) \ . \tag{7.8}$$

In order to understand the behavior of the test, one needs to study the difference of the biases of the estimates

$$\mathbb{E}T_n = \mathbb{E}S_n - \mathbb{E}\widetilde{S}_n$$

under the null hypothesis (7.6). In this case we have

$$\mathbb{E}S_n - \mathbb{E}\widetilde{S}_n = (\mathbb{E}S_n - \mathbb{E}\{m(\mathbf{X})^2\}) - (\mathbb{E}\widetilde{S}_n - \mathbb{E}\{\widetilde{m}(T(\mathbf{X}))^2\}) \ .$$

If $\widetilde{m}$ and f are Lipschitz continuous and f is bounded away from 0, then, by Devroye, Ferrario, Györfi, and Walk (2013),

$$n^{2/d}(\mathbb{E}S_n - \mathbb{E}\{m(\mathbf{X})^2\}) = O(1)$$

when $d \geq 2$ and

$$n^{2/(d-1)}(\mathbb{E}\widetilde{S}_n - \mathbb{E}\{\widetilde{m}(T(\mathbf{X}))^2\}) = O(1)$$

when $d \geq 3$.

Thus, under the null hypothesis (7.6),

$$\mathbb{E}T_n = O(n^{-2/d}) \ , \tag{7.9}$$

for $d \geq 2$. Note that for $d \leq 4$, the bias is at most of the order of the random fluctuations of the test statistic. However, for $d > 4$ the bias may dominate. Such a dependence on the dimension is inevitable under fully nonparametric conditions like the ones assumed here.

Under the null hypothesis, (7.8) and (7.9) imply that the probability of error may be bounded as

$$\mathbb{P}\{T_n > a_n\} \leq \mathbb{P}\{T_n - \mathbb{E}T_n > \omega_n \cdot n^{-1/2}\} + \mathbb{I}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}} \to 0 \ .$$

Thus, the test is consistent.

The condition that the density f is bounded away from zero may be avoided at the price of a worse rate of convergence. In particular, if m is C-Lipschitz and $\mathbf{X}$ is bounded, then

$$\begin{aligned}
& n^{1/d}|\mathbb{E}S_n - \mathbb{E}[m(\mathbf{X})^2]| \\
&= n^{1/d}|\mathbb{E}[m(\mathbf{X})m_n(\mathbf{X})] - \mathbb{E}[m(\mathbf{X})^2]| \\
&= n^{1/d}|\mathbb{E}[m(\mathbf{X})m(\mathbf{X}_{1,n}(\mathbf{X}))] - \mathbb{E}[m(\mathbf{X})^2]| \\
&\leq n^{1/d}LC\mathbb{E}\|\mathbf{X}_{1,n}(\mathbf{X}) - \mathbf{X}\| \\
&= O(1) \quad \text{(by a packing argument of Liitiäinen et al. (2010, Theorem 3.2)} \\
&\qquad\qquad \text{and by Biau and Devroye (2015, Theorem 2.1)).}
\end{aligned}$$

In this case the threshold should be larger:

$$a_n := \omega_n \left(n^{-1/2} + n^{-1/d}\right)$$

One may prove that the test is not only consistent in the sense that $\mathbb{P}\{T_n > a_n\} \to 0$ under the null hypothesis but also in the sense that $\limsup_{n\to\infty} \mathbb{I}_{\{T_n > a_n\}} = 0$ with probability one. For a discussion and references on the notion of strong consistency we refer the reader to Devroye and Lugosi (2002), Biau and Györfi (2005), Gretton and Györfi (2010).

The proof of strong consistency under the alternative hypothesis follows simply from (7.7). Under the null hypothesis it follows from Theorem 7.3. Indeed, Theorem 7.3 implies that

$$\mathbb{P}\left\{|T_n - \mathbb{E}T_n| > \epsilon\right\} \leq 2e^{-n\epsilon^2/(2L^4)} + 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1} \ .$$

For $\delta > 3/2$, choose

$$a_n := (\ln n)^\delta n^{-1/2} + \omega_n \cdot n^{-2/d}$$

with increasing unbounded $\omega_n = o(n^{2/d})$. Then, under the null hypothesis

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}\{T_n > a_n\} &\leq \sum_{n=1}^{\infty} \left(\mathbb{P}\{T_n - \mathbb{E}T_n > (\ln n)^\delta n^{-1/2}\} + \mathbb{I}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}}\right) \\
&\leq \sum_{n=1}^{\infty} \left(2e^{-(\ln n)^{2\delta}/(2L^4)} + 2e^{-(\ln n)^{2\delta/3}/(42eL^4)^{1/3}+1} + \mathbb{I}_{\{\mathbb{E}T_n > \omega_n \cdot n^{-2/d}\}}\right) \\
&< \infty \ ,
\end{aligned}$$

and so the Borel-Cantelli Lemma implies that the test makes error only finitely many times almost surely.

Remark. In applications, one would like to test not only if a given component of $\mathbf{X}$ carries predictive information but rather test the same for each of the d variables. In such cases, one faces a *multiple testing* problem with d dependent tests. In order to analyze such multiple testing procedures, say, by the Bonferroni approach, one needs a uniform control over the fluctuations of the test statistic. In such cases a non-asymptotic concentration inequality of Theorem 7.3 is particularly useful.

7.4 Proofs

In the proofs below we use two lemmas on the measure of Voronoi cells. Let

$$A_n(\mathbf{X}_j) = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{X}_j \text{ is the nearest neighbor of } \mathbf{x} \text{ among } \mathbf{X}_1, \ldots, \mathbf{X}_n\}$$

$(j = 1, \ldots, n)$, be the cells of the Voronoi partition of $\mathbb{R}^d$.

Lemma 7.1. (Devroye, Györfi, Lugosi, and Walk (2018)) *If μ has a density, then*

$$n^k \mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^k\right] \leq k! \ ,$$

$k = 1, 2, \ldots$

PROOF. Devroye, Györfi, Lugosi, and Walk (2017) proved that there exists a positive constant c_k such that

$$n^k \mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^k\right] \leq c_k \ ,$$

and $n\mu(A_n(\mathbf{X}_1))$ converges in distribution to a random variable Z such that

$$\mathbb{E}\left[Z^k\right] \leq k! \ ,$$

$k = 1, 2, \ldots$ This lemma is on the same non-asymptotic bound. We show that

$$\begin{aligned} &\mathbb{E}\left\{\mu(A_n(\mathbf{X}_1))^k\right\} \\ &\leq \mathbb{P}\left\{\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+k} \text{ are the nearest neighbors of } \mathbf{X}_1 \text{ among } \mathbf{X}_2, \ldots, \mathbf{X}_{n+k}\right\} \ , \end{aligned} \tag{7.10}$$

which implies that

$$\mathbb{E}\left\{(n\mu(A_n(\mathbf{X}_1)))^k\right] \leq \frac{n^k}{\binom{n+k-1}{k}} \leq k! \ .$$

Recall that $B_{\mathbf{x},r}$ denotes the closed ball of radius $r > 0$ centered at $\mathbf{x}$ and note that

$$\begin{aligned} \mathbb{E}\left\{\mu(A_n(\mathbf{X}_1))^k\right\} &= \mathbb{P}\left\{\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+k} \in A_n(\mathbf{X}_1)\right\} \\ &= \mathbb{E}\left[(1 - \mu(B_{\mathbf{X}_{n+1}, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|} \cup \cdots \cup B_{\mathbf{X}_{n+k}, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|}))^{n-1}\right] \\ &\leq \mathbb{E}\left[(1 - \max\{\mu(B_{\mathbf{X}_{n+1}, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}), \ldots, \mu(B_{\mathbf{X}_{n+k}, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|})\})^{n-1}\right] \ , \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}\left\{\mathbf{X}_{n+1} \ldots, \mathbf{X}_{n+k} \text{ are the nearest neighbors of } \mathbf{X}_1 \text{ among } \mathbf{X}_2, \ldots, \mathbf{X}_{n+k}\right\} \\ &= \mathbb{E}\left[(1 - \max\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}), \ldots, \mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|})\})^{n-1}\right] \ . \end{aligned}$$

(7.10) follows from comparing the right-hand sides of the two equations above. On the one hand,

$$\begin{aligned} &\mathbb{P}\left\{\max\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}), \ldots, \mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|})\} \leq z\right\} \\ &= \mathbb{P}\left\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}) \leq z, \ldots, \mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|}) \leq z\right\} \\ &= \mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}) \leq z, \ldots, \mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|}) \leq z \mid \mathbf{X}_1\right\}\right] \\ &= \mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}) \leq z \mid \mathbf{X}_1\right\} \cdot \ldots \cdot \mathbb{P}\left\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+k} - \mathbf{X}_1\|}) \leq z \mid \mathbf{X}_1\right\}\right] \\ &= \mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_1, \|\mathbf{X}_{n+1} - \mathbf{X}_1\|}) \leq z \mid \mathbf{X}_1\right\}^k\right] \\ &= z^k \ , \end{aligned}$$

while on the other hand,

$$\begin{aligned}
&\mathbb{P}\left\{\max\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|}),\ldots,\mu(B_{\mathbf{X}_{n+k},\|\mathbf{X}_{n+k}-\mathbf{X}_1\|})\}\le z\right\}\\
&=\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z,\ldots,\mu(B_{\mathbf{X}_{n+k},\|\mathbf{X}_{n+k}-\mathbf{X}_1\|})\le z\right\}\\
&=\mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z,\ldots,\mu(B_{\mathbf{X}_{n+k},\|\mathbf{X}_{n+k}-\mathbf{X}_1\|})\le z\mid\mathbf{X}_1\right\}\right]\\
&=\mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z\mid\mathbf{X}_1\right\}\cdot\ldots\cdot\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+k},\|\mathbf{X}_{n+k}-\mathbf{X}_1\|})\le z\mid\mathbf{X}_1\right\}\right]\\
&=\mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z\mid\mathbf{X}_1\right\}^k\right]\\
&\ge\mathbb{E}\left[\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z\mid\mathbf{X}_1\right\}\right]^k\\
&=\mathbb{P}\left\{\mu(B_{\mathbf{X}_{n+1},\|\mathbf{X}_{n+1}-\mathbf{X}_1\|})\le z\right\}^k\\
&=z^k\ .
\end{aligned}$$

□

Lemma 7.2. (Devroye, Györfi, Lugosi, and Walk (2017)) *Assume that μ has a density. Then*

$$n^2\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^2\mid\mathbf{X}_1=\mathbf{x}\right]\to\alpha(d)$$

for μ-almost all $\mathbf{x}$, where α_d is defined in (7.2).

Proof of Theorem 7.2

We prove the variance bound of Theorem 7.2 first. The proof relies of the following version of the Efron-Stein inequality, see, for example, Boucheron et al. (2010, Theorem 3.1).

Lemma 7.3. *(Efron-Stein inequality) Let $Z=(Z_1,\ldots,Z_n)$ be a collection of independent random variables taking values in some measurable set A and denote by $Z^{(i)}=(Z_1,\ldots,Z_{i-1},Z_{i+1},\ldots,Z_n)$ the collection with the i-th random variable dropped. Let $f:A^n\to\mathbb{R}$ and $g:A^{n-1}\to\mathbb{R}$ be measurable real-valued functions. Then*

$$\mathbb{V}ar(f(Z))\le\mathbb{E}\left[\sum_{i=1}^n\left(f(Z)-g(Z^{(i)})\right)^2\right]\ .$$

By the decomposition

$$S_n=S_n-\mathbb{E}\left[S_n\mid D_n\right]+\mathbb{E}\left[S_n\mid D_n\right],$$

we have that

$$\mathbb{V}ar(S_n) = \mathbb{E}\left[(S_n - \mathbb{E}\left[S_n \mid D_n\right])^2\right] + \mathbb{V}ar(\mathbb{E}\left[S_n \mid D_n\right]) \ .$$

Conditionally on D_n, S_n is an average of independent, identically distributed (i.i.d.) random variables bounded by L^2, and therefore

$$\mathbb{E}\left[(S_n - \mathbb{E}\left[S_n \mid D_n\right])^2\right] \leq \frac{L^4}{n} \ .$$

Notice that we may write

$$m_n(\mathbf{x}) = \sum_{j=1}^{n} Y_j \mathbb{I}_{\{\mathbf{x} \in A_n(\mathbf{X}_j)\}} \ .$$

Then

$$\mathbb{E}\left[S_n \mid D_n\right] = \int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x}) = \sum_{j=1}^{n} Y_j \int_{A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \ .$$

Putting $L_n = \mathbb{E}\left[S_n \mid D_n\right]$, this implies

$$L_n = \sum_{i=1}^{n} Y_i \mathbb{E}\{\mathbb{I}_{\mathbf{X} \in A_n(\mathbf{X}_i)} m(\mathbf{X}) \mid D_n\} \ .$$

Considering L_n as a function of the n i.i.d. pairs $(\mathbf{X}_i, Y_i)_{i=1}^n$, we may use the Efron-Stein inequality to bound the variance of L_n. Define $L_n^{(j)}$ as L_n when $(\mathbf{X}_j, Y_j)$ is omitted from the sample. By Lemma 7.3,

$$\mathbb{V}ar(L_n) \leq \mathbb{E}\left[\sum_{j=1}^{n} \left(L_n - L_n^{(j)}\right)^2\right] = n\mathbb{E}\left[\left(L_n - L_n^{(1)}\right)^2\right] \ .$$

Let $\{A_n'(\mathbf{X}_2), \ldots, A_n'(\mathbf{X}_n)\}$ be the Voronoi partition, when $\mathbf{X}_1$ is omitted from the sample. Then

$$\begin{aligned}
|L_n - L_n^{(1)}| &= \left| Y_1 \int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x}) - \sum_{i=2}^{n} Y_i \int_{A_n'(\mathbf{X}_i) \setminus A_n(\mathbf{X}_i)} m(\mathbf{x}) \mu(d\mathbf{x}) \right| \\
&\leq L^2 \left(\mu(A_n(\mathbf{X}_1)) + \sum_{i=2}^{n} \mu(A_n'(\mathbf{X}_i) \setminus A_n(\mathbf{X}_i)) \right) \\
&= 2L^2 \mu(A_n(\mathbf{X}_1)) \ .
\end{aligned}$$

Thus, Lemma 7.1 implies

$$\mathbb{V}ar(L_n) \leq 4nL^4\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^2\right] \leq 8L^4/n$$

leading to

$$\mathbb{V}ar\left(\mathbb{E}\left[S_n \mid D_n\right]\right) \leq \frac{8L^4}{n} \ ,$$

and therefore to the desired bound

$$\mathbb{V}ar(S_n) \leq \frac{9L^4}{n} \ .$$

Proof of Theorem 7.1

Introduce the notation

$$\sqrt{n}\left(S_n - \mathbb{E}S_n\right) = U_n + V_n + W_n \ ,$$

where

$$U_n = \sqrt{n}\left(S_n - \mathbb{E}[S_n \mid D_n]\right)$$

and

$$V_n = \sqrt{n}\left(\mathbb{E}[S_n \mid D_n] - \mathbb{E}[S_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n]\right)$$

and

$$W_n = \sqrt{n}\left(\mathbb{E}[S_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n] - \mathbb{E}S_n\right) \ .$$

We prove Theorem 7.1 by showing that, for any $u, v \in \mathbb{R}$,

$$\mathbb{P}\{U_n \leq u, V_n \leq v\} \to \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right) \ , \tag{7.11}$$

where Φ denotes the standard normal distribution function, and that

$$\mathbb{V}ar(W_n) \to 0. \tag{7.12}$$

Györfi and Walk (2015) proved that

$$\begin{aligned}
&\left|\mathbb{P}\{U_n \leq u, V_n \leq v\} - \Phi\left(\frac{u}{\sigma_1}\right)\Phi\left(\frac{v}{\sigma_2}\right)\right| \\
&\leq \mathbb{E}\left|\mathbb{P}\{U_n \leq u \mid D_n\} - \Phi\left(\frac{u}{\sigma_1}\right)\right| + \left|\mathbb{P}\{V_n \leq v\} - \Phi\left(\frac{v}{\sigma_2}\right)\right| \ .
\end{aligned}$$

Thus, (7.11) holds if

$$\mathbb{P}\{U_n \leq u \mid D_n\} \to \Phi\left(\frac{u}{\sigma_1}\right) \quad \text{in probability} \tag{7.13}$$

and

$$\mathbb{P}\{V_n \leq v\} \to \Phi\left(\frac{v}{\sigma_2}\right) \ . \tag{7.14}$$

Proof of (7.13).
Let's start with the decomposition

$$\begin{aligned} U_n &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i' m_n(\mathbf{X}_i') - \mathbb{E}[Y_i' m_n(\mathbf{X}_i') \mid D_n])\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i' m_n(\mathbf{X}_i') - \mathbb{E}[Y_i' m_n(\mathbf{X}_i') \mid D_n]) \ . \end{aligned}$$

Next we apply a Berry-Esseen type central limit theorem (see Theorem 14 in Petrov (1975)). For a universal constant $c > 0$, we have

$$\left|\mathbb{P}\{U_n \leq u \mid D_n\} - \Phi\left(\frac{u}{\sqrt{\mathbb{V}ar(Y_1' m_n(\mathbf{X}_1') \mid D_n)}}\right)\right| \leq \frac{c}{\sqrt{n}} \frac{\mathbb{E}[|Y_1' m_n(\mathbf{X}_1')|^3 \mid D_n]}{\sqrt{\mathbb{V}ar(Y_1' m_n(\mathbf{X}_1') \mid D_n)}^3} \ .$$

Since

$$\mathbb{E}[Y_1' m_n(\mathbf{X}_1') \mid D_n] = \int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x}) \ , \tag{7.15}$$

we have

$$\begin{aligned} \mathbb{V}ar(Y_1' m_n(\mathbf{X}_1') \mid D_n) &= \mathbb{E}[{Y_1'}^2 m_n(\mathbf{X}_1')^2 \mid D_n] - \mathbb{E}[Y_1' m_n(\mathbf{X}_1') \mid D_n]^2 \\ &= \int M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) - \left(\int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x})\right)^2 . \end{aligned}$$

We need to show that

$$\int M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) \to \int M_2(\mathbf{x})^2 \mu(d\mathbf{x}) \tag{7.16}$$

in probability and

$$\int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x}) \to \int m(\mathbf{x})^2 \mu(d\mathbf{x}) \tag{7.17}$$

in probability. Since $m_n(\mathbf{x}) = Y_j$ if $\mathbf{x} \in A_n(\mathbf{X}_j)$, we get that

$$\begin{aligned}\int M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) &= \sum_{j=1}^{n} \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) \\ &= \sum_{j=1}^{n} Y_j^2 \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) \mu(d\mathbf{x}) \ .\end{aligned}$$

We use this to prove (7.16). Indeed,

$$\begin{aligned}&\int M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) - \int M_2(\mathbf{x})^2 \mu(d\mathbf{x}) \\ &= \sum_{j=1}^{n} Y_j^2 \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) \mu(d\mathbf{x}) - \sum_{j=1}^{n} \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x})^2 \mu(d\mathbf{x}) \\ &= \sum_{j=1}^{n} (Y_j^2 - M_2(\mathbf{X}_j)) \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) \mu(d\mathbf{x}) \\ &\quad + \sum_{j=1}^{n} \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) (M_2(\mathbf{X}_j) - M_2(\mathbf{x})) \mu(d\mathbf{x}) \ .\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left[\left|\int M_2(\mathbf{x}) m_n(\mathbf{x})^2 \mu(d\mathbf{x}) - \int M_2(\mathbf{x})^2 \mu(d\mathbf{x})\right|\right] \\ &\leq \mathbb{E}\left[\left|\sum_{j=1}^{n} (Y_j^2 - M_2(\mathbf{X}_j)) \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) \mu(d\mathbf{x})\right|\right] \\ &\quad + \mathbb{E}\left[\left|\sum_{j=1}^{n} \int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x}) (M_2(\mathbf{X}_j) - M_2(\mathbf{x})) \mu(d\mathbf{x})\right|\right] \ ,\end{aligned}$$

and so

$$\begin{aligned}
&\mathbb{E}\left[\left|\int M_2(\mathbf{x})m_n(\mathbf{x})^2\mu(d\mathbf{x}) - \int M_2(\mathbf{x})^2\mu(d\mathbf{x})\right|\right] \\
&\leq \sqrt{\mathbb{V}ar\left(\sum_{j=1}^{n}(Y_j^2 - M_2(\mathbf{X}_j))\int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x})\mu(d\mathbf{x})\right)} \\
&\quad + \mathbb{E}\left[\sum_{j=1}^{n}\int_{A_n(\mathbf{X}_j)} M_2(\mathbf{x})|M_2(\mathbf{X}_j) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&\leq \sqrt{n\mathbb{E}\left[(Y_1^2 - M_2(\mathbf{X}_1))^2\left(\int_{A_n(\mathbf{X}_1)} M_2(\mathbf{x})\mu(d\mathbf{x})\right)^2\right]} \\
&\quad + n\mathbb{E}\left[\int_{A_n(\mathbf{X}_1)} M_2(\mathbf{x})|M_2(\mathbf{X}_1) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&\leq L^4\sqrt{n\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^2\right]} + L^2 n\mathbb{E}\left[\int_{A_n(\mathbf{X}_1)} |M_2(\mathbf{X}_1) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right]
\end{aligned}$$

To complete the proof of (7.16), it suffices to show that the sum above converges to zero as $n \to \infty$. To this end, note that Lemma 7.1 implies that

$$n\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^2\right] \leq c_2/n \to 0 \ ,$$

and furthermore

$$\begin{aligned}
&n\mathbb{E}\left[\int_{A_n(\mathbf{X}_1)} |M_2(\mathbf{X}_1) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&= n\mathbb{E}\left[\int_{A_n(\mathbf{X}_1)} |M_2(\mathbf{X}_{1,n}(\mathbf{x})) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&= \mathbb{E}\left[\int |M_2(\mathbf{X}_{1,n}(\mathbf{x})) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \ .
\end{aligned}$$

It remains to show that

$$\mathbb{E}\left[\int |M_2(\mathbf{X}_{1,n}(\mathbf{x})) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \to 0 \ . \tag{7.18}$$

Fix any $\epsilon > 0$ and choose a bounded continuous function $\widetilde{M}_2$ such that

$$\int |M_2(\mathbf{x}) - \widetilde{M}_2(\mathbf{x})|\mu(d\mathbf{x}) < \epsilon \ .$$

Then, with $M_2^* = M_2 - \widetilde{M}_2$, one has

$$\begin{aligned}
&\mathbb{E}\left[\int |M_2(\mathbf{X}_{1,n}(\mathbf{x})) - M_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&\leq \mathbb{E}\left[\int |\widetilde{M}_2(\mathbf{X}_{1,n}(\mathbf{x})) - \widetilde{M}_2(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&+ \mathbb{E}\left[\int |M_2^*(\mathbf{X}_{1,n}(\mathbf{x}))|\mu(d\mathbf{x})\right] + \int |M_2^*(\mathbf{x})|\mu(d\mathbf{x}) \ . \qquad (7.19)
\end{aligned}$$

The first term on the right-hand side converges to 0 by the dominated convergence theorem, since, by Lemma 6.1 in Györfi et al. (2002),

$$\mathbf{X}_{1,n}(\mathbf{x}) \to \mathbf{x} \qquad \text{a.s. for } \mu\text{-almost all } \mathbf{x} \ .$$

To bound the second term, we introduce some notation. A set $C \subset \mathbb{R}^d$ is a cone of angle $\pi/3$ centered at 0 if there exists an $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| = 1$ such that

$$C = \left\{\mathbf{y} \in \mathbb{R}^d : \frac{(\mathbf{x}, \mathbf{y})}{\|\mathbf{y}\|} \geq \cos(\pi/6)\right\} \ .$$

Let γ_d be the minimal number of cones $C_1, \ldots, C_{\gamma_d}$ of angle $\pi/3$ centered at 0 such that their union covers $\mathbb{R}^d$. The second term on the right-hand side of (7.19) is bounded by

$$\gamma_d \int |M_2^*(\mathbf{x})|\mu(d\mathbf{x}) \leq \gamma_d \epsilon$$

by Lemma 6.3 in Györfi et al. (2002). Thus, (7.18) is proved and hence so is (7.16). For the proof of (7.17), we have that

$$\begin{aligned}
\int m(\mathbf{x})m_n(\mathbf{x})\mu(d\mathbf{x}) &= \sum_{j=1}^n \int_{A_n(\mathbf{X}_j))} m(\mathbf{x})m_n(\mathbf{x})\mu(d\mathbf{x}) \\
&= \sum_{j=1}^n Y_j \int_{A_n(\mathbf{X}_j))} m(\mathbf{x})\mu(d\mathbf{x}). \qquad (7.20)
\end{aligned}$$

Similarly, the derivation for (7.16) implies that

$$\begin{aligned}
&\mathbb{E}\left[\left|\int m(\mathbf{x})m_n(\mathbf{x})\mu(d\mathbf{x}) - \int m(\mathbf{x})^2\mu(d\mathbf{x})\right|\right] \\
&\leq L^2\sqrt{n\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^2\right]} + Ln\mathbb{E}\left[\int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_1) - m(\mathbf{x})|\mu(d\mathbf{x})\right] \\
&\to 0,
\end{aligned}$$

and so (7.17) is proved, too. Thus,

$$\mathbb{V}ar(Y_1' m_n(\mathbf{X}_1') \mid D_n) \to \sigma_1^2$$

in probability. Moreover,

$$\mathbb{E}[|Y_1' m_n(\mathbf{X}_1')|^3 \mid D_n] \le L^6.$$

These relations imply (7.13).

Proof of (7.12).
(7.15) and (7.20) imply that

$$\mathbb{E}[S_n \mid D_n] = \mathbb{E}[Y_1' m_n(\mathbf{X}_1') \mid D_n] = \int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x}) = \sum_{j=1}^{n} Y_j \int_{A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \ .$$

Hence

$$\mathbb{E}[S_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n] = \sum_{j=1}^{n} m(\mathbf{X}_j) \int_{A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) = \int m(\mathbf{x}) m(\mathbf{X}_{1,n}(\mathbf{x})) \mu(d\mathbf{x}).$$

We prove (7.12) by a slight extension of the proof of Theorem 7.2. Set

$$L_n := \sqrt{n} \int m(\mathbf{x}) m(\mathbf{X}_{1,n}(\mathbf{x})) \mu(d\mathbf{x}) = \sqrt{n} \sum_{j=1}^{n} m(\mathbf{X}_j) \int_{A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \ .$$

Define $L_n^{(j)}$ as L_n when $\mathbf{X}_j$ is dropped. As in the proof of Theorem 7.2,

$$\mathbb{V}ar(W_n) = \mathbb{V}ar(L_n) \le \mathbb{E}\left[\sum_{j=1}^{n} \left(L_n - L_n^{(j)} \right)^2 \right] = n \mathbb{E}\left[\left(L_n - L_n^{(1)} \right)^2 \right] \ .$$

Then

$$L_n^{(1)} = \sqrt{n} \sum_{j=2}^{n} m(\mathbf{X}_j) \int_{A_n'(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \ ,$$

and so

$$\begin{aligned} L_n - L_n^{(1)} &= \sqrt{n} m(\mathbf{X}_1) \int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x}) - \sqrt{n} \sum_{j=2}^{n} m(\mathbf{X}_j) \int_{A_n'(\mathbf{X}_j) \setminus A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \\ &= \sqrt{n} \left(\int_{A_n(\mathbf{X}_1)} m(\mathbf{X}_{1,n}(\mathbf{x})) m(\mathbf{x}) \mu(d\mathbf{x}) - \int_{A_n(\mathbf{X}_1)} m(\mathbf{X}_{2,n}(\mathbf{x})) m(\mathbf{x}) \mu(d\mathbf{x}) \right) \ , \end{aligned}$$

where $\mathbf{X}_{2,n}(\mathbf{x})$ denotes the second nearest neighbor of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Therefore

$$|L_n - L_n^{(1)}| \leq \sqrt{n} L \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{X}_{2,n}(\mathbf{x}))| \mu(d\mathbf{x})$$

by (7.3). Hence,

$$\mathbb{V}ar(W_n) \leq L^2 \mathbb{E}\left[\left(n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{X}_{2,n}(\mathbf{x}))| \mu(d\mathbf{x}) \right)^2 \right] . \tag{7.21}$$

As it is well known, for a real-valued random variable Z, by Hölder's inequality,

$$\mathbb{E}\left[Z^2\right] = \mathbb{E}\left[|Z|^{2/3} |Z|^{4/3}\right] \leq \mathbb{E}\left[|Z|\right]^{2/3} \mathbb{E}\left[Z^4\right]^{1/3} . \tag{7.22}$$

One has

$$\begin{aligned}
&\mathbb{E}\left[n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{X}_{2,n}(\mathbf{x}))| \mu(d\mathbf{x}) \right] \\
&\leq \mathbb{E}\left[n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{x})| \mu(d\mathbf{x}) \right] + \mathbb{E}\left[n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{2,n}(\mathbf{x})) - m(\mathbf{x})| \mu(d\mathbf{x}) \right] \\
&= \mathbb{E}\left[\int |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{x})| \mu(d\mathbf{x}) \right] + \mathbb{E}\left[\int |m(\mathbf{X}_{2,n}(\mathbf{x})) - m(\mathbf{x})| \mu(d\mathbf{x}) \right] \\
&\to 0
\end{aligned} \tag{7.23}$$

as $n \to \infty$, where the latter can be shown as the limit relation (7.18). Furthermore

$$\begin{aligned}
\mathbb{E}\left[\left(n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{X}_{2,n}(\mathbf{x}))| \mu(d\mathbf{x}) \right)^4 \right] &\leq 16 L^4 \mathbb{E}\left[n^4 \mu(A_n(\mathbf{X}_1))^4 \right] \\
&\leq 16 L^4 c_4
\end{aligned} \tag{7.24}$$

by (7.3) and Lemma 7.1. With the notation

$$Z = n \int_{A_n(\mathbf{X}_1)} |m(\mathbf{X}_{1,n}(\mathbf{x})) - m(\mathbf{X}_{2,n}(\mathbf{x}))| \mu(d\mathbf{x})$$

(7.21), (7.22), (7.23) and (7.24) imply (7.12).

Proof of (7.14).
For

$$V_n = \frac{\sum_{j=1}^n V_{n,j}}{\sqrt{n}}$$

with

$$V_{n,j} = n(Y_j - m(\mathbf{X}_j)) \int_{A_n(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x}) \ ,$$

notice that the triangular array $V_{n,j}$, $n = 1, 2, \ldots$, $j = 1, \ldots, n$ is (row-wise) exchangeable, for which there is a classical central limit theorem:

Theorem 7.4. (Blum et al. (1958), Weber (1980)) *Let $\{V_{n,j}\}$ be a triangular array of exchangeable random variables with zero mean and finite variance. Assume that*

(i)

$$\mathbb{E}[V_{n,1}V_{n,2}] = o(1/n) \ ,$$

(ii)

$$\lim_{n\to\infty} \max\{|V_{n,j}|; j = 1, \ldots, n\}/\sqrt{n} = 0$$

in probability,

(iii)

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} V_{n,j}^2 = \sigma^2$$

in probability.

Then

$$\frac{\sum_{j=1}^{n} V_{n,j}}{\sqrt{n}}$$

is asymptotically normal with mean zero and variance σ^2.

Condition (i) of Theorem 7.4 is satisfied since

$$\mathbb{E}[V_{n,1}V_{n,2}] = 0.$$

Condition (ii) of Theorem 7.4 follows from (7.3), Lemma 7.1 and Jensen's inequality:

$$\begin{aligned} n\mathbb{E}\left[\max_j \mu(A_n(\mathbf{X}_j)\right] &\leq n\mathbb{E}\left[\left(\sum_j \mu(A_n(\mathbf{X}_j)^3\right)^{1/3}\right] \\ &\leq n\left(\mathbb{E}\left[\sum_j \mu(A_n(\mathbf{X}_j)^3\right]\right)^{1/3} \\ &\leq n\left(n\frac{c_3}{n^3}\right)^{1/3} \\ &= o(\sqrt{n}) \ . \end{aligned}$$

Condition (iii) in Theorem 7.4 is fulfilled if

$$\lim_{n\to\infty} \mathbb{E}[V_{n,1}^2] = \sigma_2^2 \tag{7.25}$$

and

$$\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2\right) \to 0. \tag{7.26}$$

We have that

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}[V_{n,1}^2] &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(\mathbf{X}_1))^2\left(\int_{A_n(\mathbf{X}_1)} m(\mathbf{x})\mu(d\mathbf{x})\right)^2\right] \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(Y_1 - m(\mathbf{X}_1))^2 m(\mathbf{X}_1)^2\mu(A_n(\mathbf{X}_1))^2\right] \\ &= \lim_{n\to\infty} n^2\mathbb{E}\left[(M_2(\mathbf{X}_1)m(\mathbf{X}_1)^2 - m(\mathbf{X}_1)^4)\mu(A_n(\mathbf{X}_1))^2\right] \ . \end{aligned} \tag{7.27}$$

(7.27) follows from

$$
\begin{aligned}
& n^2 \Big| \mathbb{E}\left[(Y_1 - m(\mathbf{X}_1))^2 \left(\int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x}) \right)^2 \right] \\
& - \mathbb{E}\left[(Y_1 - m(\mathbf{X}_1))^2 m(\mathbf{X}_1)^2 \mu(A_n(\mathbf{X}_1))^2 \right] \Big| \\
& \leq n^2 4L^2 \mathbb{E}\left[\left| \left(\int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x}) \right)^2 - m(\mathbf{X}_1)^2 \mu(A_n(\mathbf{X}_1))^2 \right| \right] \\
& \leq n^2 8L^3 \mathbb{E}\left[\left| \int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x}) - m(\mathbf{X}_1) \mu(A_n(\mathbf{X}_1)) \right| \mu(A_n(\mathbf{X}_1)) \right] \\
& = n^2 8L^3 \mathbb{E}\left[\left| \frac{\int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x})}{\mu(A_n(\mathbf{X}_1))} - m(\mathbf{X}_1) \right| \mu(A_n(\mathbf{X}_1))^2 \right] \\
& \leq n^2 8L^3 \sqrt{ \mathbb{E}\left[\left| \frac{\int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x})}{\mu(A_n(\mathbf{X}_1))} - m(\mathbf{X}_1) \right|^2 \right] } \sqrt{ \mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^4 \right] } \\
& \leq 8L^3 \sqrt{c_4} \sqrt{ \mathbb{E}\left[\left| \frac{\int_{A_n(\mathbf{X}_1)} m(\mathbf{x}) \mu(d\mathbf{x})}{\mu(A_n(\mathbf{X}_1))} - m(\mathbf{X}_1) \right|^2 \right] } \; .
\end{aligned}
$$

The expression on the right-hand side converges to zero. To show this, fix an arbitrary $\epsilon > 0$ and choose a decomposition $m = m^* + m^{**}$ such that m^* is Lipschitz continuous with bounded support and $\mathbb{E}[m^{**}(\mathbf{X})^2] < \epsilon$. Then it suffices to show the limit relation for m^*. But this follows from the fact that $\operatorname{diam}(A_n(\mathbf{X}_1)) \to 0$ in probability (Devroye, Györfi, Lugosi, and Walk (2017, Section 5)). Lemma 7.2 implies that

$$
\mathbb{E}\left[n^2 \mu(A_n(\mathbf{X}_1))^2 \mid \mathbf{X}_1 \right] \to \alpha(d) \quad \text{with probability one.} \tag{7.28}
$$

Set

$$
Z_n = (M_2(\mathbf{X}_1) m(\mathbf{X}_1)^2 - m(\mathbf{X}_1)^4) \mathbb{E}\left[n^2 \mu(A_n(\mathbf{X}_1))^2 \mid \mathbf{X}_1 \right] \; .
$$

By (7.3) and Lemma 7.1 for $k = 4$ together with Jensen's inequality for conditional expectations we obtain

$$
\mathbb{E}[Z_n^2] \leq L^8 c_4
$$

and thus uniform integrability of $\{Z_n\}$, i.e.,

$$
\lim_{K \to \infty} \sup_n \mathbb{E}[Z_n \mathbb{I}_{\{Z_n > K\}}] = 0.
$$

Then (7.28) yields

$$\begin{aligned}
&n^2\mathbb{E}\left[(M_2(\mathbf{X}_1)m(\mathbf{X}_1)^2 - m(\mathbf{X}_1)^4)\mu(A_n(\mathbf{X}_1))^2\right] \\
&= \mathbb{E}\left[(M_2(\mathbf{X}_1)m(\mathbf{X}_1)^2 - m(\mathbf{X}_1)^4)\mathbb{E}\left[n^2\mu(A_n(\mathbf{X}_1))^2 \mid \mathbf{X}_1\right]\right] \\
&\to \alpha(d)\mathbb{E}\left[M_2(\mathbf{X}_1)m(\mathbf{X}_1)^2 - m(\mathbf{X}_1)^4\right] \\
&= \sigma_2^2 \ ,
\end{aligned}$$

verifying (7.25).

One may check (7.26) similarly to (7.12). Indeed, put

$$L_n := \frac{1}{n}\sum_{j=1}^n V_{n,j}^2 = n\sum_{j=1}^n (Y_j - m(\mathbf{X}_j))^2 \left(\int_{A_n(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x})\right)^2 .$$

Thus,

$$\begin{aligned}
&|L_n - L_n^{(1)}| \\
&\le n(Y_1 - m(\mathbf{X}_1))^2 \left(\int_{A_n(\mathbf{X}_1)} m(\mathbf{x})\mu(d\mathbf{x})\right)^2 \\
&+ n\sum_{j=2}^n (Y_j - m(\mathbf{X}_j))^2 \left|\left(\int_{A_n(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x})\right)^2 - \left(\int_{A_n'(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x})\right)^2\right| .
\end{aligned}$$

Therefore

$$\begin{aligned}
&|L_n - L_n^{(1)}| \\
&\le 4L^4 n\mu(A_n(\mathbf{X}_1))^2 \\
&+ 4L^2 n\sum_{j=2}^n (Y_j - m(\mathbf{X}_j))^2 \left|\int_{A_n(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x}) + \int_{A_n'(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x})\right| \\
&\cdot \left|\int_{A_n'(\mathbf{X}_j)\setminus A_n(\mathbf{X}_j)} m(\mathbf{x})\mu(d\mathbf{x})\right| \\
&\le 4L^4 n\mu(A_n(\mathbf{X}_1))^2 + 8L^4 n\sum_{j=2}^n \mu(A_n'(\mathbf{X}_j))\mu(A_n'(\mathbf{X}_j)\setminus A_n(\mathbf{X}_j)) \\
&\le 4L^4 n\mu(A_n(\mathbf{X}_1))^2 + 8L^4 n\left(\max_{j=2,\dots n}\mu(A_n'(\mathbf{X}_j))\right)\mu(A_n(\mathbf{X}_1)) \ ,
\end{aligned}$$

which implies that

$$
\begin{aligned}
&\mathbb{V}ar\left(\frac{1}{n}\sum_{j=1}^{n} V_{n,j}^2\right) \\
&\leq n\mathbb{E}\left[\left(L_n - L_n^{(1)}\right)^2\right] \\
&\leq 32L^8 n^3 \mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^4\right] \\
&+ 128L^8 n^3 \sqrt{\mathbb{E}\left[\max_{j=2,\ldots n} \mu(A_n'(\mathbf{X}_j))^4\right]} \sqrt{\mathbb{E}\left[\mu(A_n(\mathbf{X}_1))^4\right]} \\
&\leq 32L^8 c_4/n + 128L^8 n \sqrt{\mathbb{E}\left[\sum_{j=2}^{n} \mu(A_n'(\mathbf{X}_j))^4\right]} \sqrt{c_4}
\end{aligned}
$$

by Lemma 7.1. Noticing that

$$
\mathbb{E}\left[\sum_{j=2}^{n} \mu(A_n'(\mathbf{X}_j))^4\right] = (n-1)\mathbb{E}\left[\mu(A_n'(\mathbf{X}_2))^4\right] = O(n^{-3})
$$

by Lemma 7.1, we obtain (7.26).

Proof of Theorem 7.3

As we mentioned in the proof (7.13), for given D_n, S_n is an average of i.i.d. random variables bounded by L^2. Therefore, by the Hoeffding inequality, one has

$$
\mathbb{P}\left\{|U_n| > \epsilon \mid D_n\right\} \leq 2e^{-n\epsilon^2/(2L^4)} .
$$

For the term V_n, apply the extension of the Efron-Stein inequality for the centered higher moments, which is a slight modification of Theorem 15.5 in Boucheron et al. (2010):

Lemma 7.4. (Devroye, Györfi, Lugosi, and Walk (2018)) *Let $Z = (Z_1, \ldots, Z_n)$ be a collection of independent random variables taking values in some measurable set A and denote by $Z^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$ the collection with the i-th random variable dropped. Let $f : A^n \to \mathbb{R}$ be a measurable real-valued function and the function*

$g_i : A^{n-1} \to \mathbb{R}$ is obtained from f by dropping the i-th argument, $i = 1, \ldots, n$. Then for any integer $q \geq 1$,

$$\begin{aligned}
&\mathbb{E}\left[(f(Z) - \mathbb{E}f(Z))^{2q}\right] \\
&\leq (cq)^q \Big(\mathbb{E}\left[\left(\sum_{i=1}^n \left(f(Z) - g_i(Z^{(i)}) \right)^2 \right)^q \right] \\
&\quad + \mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{E}\left[\left(f(Z) - g_i(Z^{(i)}) \right)^2 \mid Z_1, \ldots, Z_{i-1}, Z_{i-1}, \ldots, Z_n \right] \right)^q \right] \Big) \,, \qquad (7.29)
\end{aligned}$$

with a universal constant $c < 5.1$.

Proof. If $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ are i.i.d. and

$$Z'^{(i)} = (Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n)$$

then from Theorem 15.5 in (2010) one gets

$$\mathbb{E}\left[(f(Z) - \mathbb{E}f(Z))_+^{2q}\right] \leq (2\kappa q)^q \mathbb{E}\left[\left(V^+\right)^q\right] \,,$$

and

$$\mathbb{E}\left[(f(Z) - \mathbb{E}f(Z))_-^{2q}\right] \leq (2\kappa q)^q \mathbb{E}\left[\left(V^-\right)^q\right] \,,$$

with $\kappa = \sqrt{e}/(2(\sqrt{e} - 1)) < 1.271$ and with

$$\begin{aligned}
V^+ &\leq \sum_{i=1}^n \mathbb{E}\left\{ (f(Z) - f(Z'^{(i)}))^2 \mid Z_1, \ldots, Z_n \right\} \\
&\leq 2 \sum_{i=1}^n \left((f(Z) - g_i(Z^{(i)}))^2 + \mathbb{E}\left[(g_i(Z^{(i)}) - f(Z'^{(i)}))^2 \mid Z_1, \ldots, Z_n \right] \right)
\end{aligned}$$

and

$$V^- \leq 2 \sum_{i=1}^n \left((f(Z) - g_i(Z^{(i)}))^2 + \mathbb{E}\left[(g_i(Z^{(i)}) - f(Z'^{(i)}))^2 \mid Z_1, \ldots, Z_n \right] \right) \,.$$

Therefore, c_r-inequality implies

$$\begin{aligned}
&\mathbb{E}\left[(f(Z) - \mathbb{E}f(Z))^{2q}\right] \\
&\leq 2(2\kappa q)^q 2^{q-1} \mathbb{E}\left[\left(\sum_{i=1}^n (f(Z) - g_i(Z^{(i)}))^2 \right)^q + \left(\sum_{i=1}^n \mathbb{E}\left[(g_i(Z^{(i)}) - f(Z'^{(i)}))^2 \mid Z_1, \ldots, Z_n \right] \right)^q \right] \,.
\end{aligned}$$

By the equality

$$\mathbb{E}\left[(g_i(Z^{(i)}) - f(Z'^{(i)}))^2 \mid Z_1, \ldots, Z_n\right] = \mathbb{E}\left[(g_i(Z^{(i)}) - f(Z))^2 \mid Z_1, \ldots, Z_{i-1}, Z_{i-1}, \ldots, Z_n\right] \ ,$$

the lemma is proved. □

Notice that

$$m_n(\mathbf{x}) = \sum_{j=1}^{n} Y_j I_{\{\mathbf{x} \in A_n(\mathbf{X}_j)\}} \ .$$

Then

$$L_n := \mathbb{E}\left[S_n \mid D_n\right] = \int m(\mathbf{x}) m_n(\mathbf{x}) \mu(d\mathbf{x}) = \sum_{j=1}^{n} Y_j \int_{A_n(\mathbf{X}_j)} m(\mathbf{x}) \mu(d\mathbf{x}) \ .$$

Consider now L_n as a function of n i.i.d. vectors $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. Define $L_n^{(i)}$ as L_n when the pair $(\mathbf{X}_i, Y_i)$ is dropped. As in the proof of Theorem 7.1

$$L_n - L_n^{(i)} = \int_{A_n(\mathbf{X}_i)} (Y_{1,n}(\mathbf{x}) - Y_{2,n}(\mathbf{x})) m(\mathbf{x}) \mu(d\mathbf{x}) \ ,$$

where $Y_{2,n}(\mathbf{x})$ denotes the label of the second nearest neighbor $\mathbf{X}_{2,n}(\mathbf{x})$ of $\mathbf{x}$ among $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Thus,

$$\begin{aligned} \left(L_n - L_n^{(i)}\right)^2 &= \left(\int_{A_n(\mathbf{X}_i)} (Y_{1,n}(\mathbf{x}) - Y_{2,n}(\mathbf{x})) m(\mathbf{x}) \mu(d\mathbf{x})\right)^2 \\ &\leq (2L^2)^2 \left(\mu(A_n(\mathbf{X}_i))\right)^2 \ . \end{aligned}$$

(7.29) implies that

$$\begin{aligned} \mathbb{E}[|L_n - \mathbb{E}[L_n]|^{2q}] \leq (cq)^q (2L^2)^{2q} \Bigg(& \mathbb{E}\left[\left(\sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i))^2\right)^q\right] \\ & + \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}[\mu(A_n(\mathbf{X}_i))^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n]\right)^q\right] \Bigg) \ . \end{aligned} \tag{7.30}$$

Because of

$$\sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i)) = 1,$$

the Jensen inequality implies that

$$\left(\sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i))^2\right)^q \le \sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i))^{q+1} ,$$

and so from Lemma 7.1 we get

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i))^2\right)^q\right] \le \mathbb{E}\left[\sum_{i=1}^{n} \mu(A_n(\mathbf{X}_i))^{q+1}\right] \le n^{-q}(q+1)! . \tag{7.31}$$

Apply the Jensen inequality twice and Lemma 7.1:

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{E}[\mu(A_n(\mathbf{X}_i))^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n]\right)^q\right] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} n\mathbb{E}[\mu(A_n(\mathbf{X}_i))^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n]\right)^q\right] \\
&\le \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \left(n\mathbb{E}[\mu(A_n(\mathbf{X}_i))^2 \mid \mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i-1}, \ldots, \mathbf{X}_n]\right)^q\right] \\
&= \mathbb{E}\left[\left(n\mathbb{E}[\mu(A_n(\mathbf{X}_1))^2 \mid \mathbf{X}_2, \ldots, \mathbf{X}_n]\right)^q\right] \\
&\le n^{-q}\mathbb{E}\left[(n\mu(A_n(\mathbf{X}_1)))^{2q}\right] \\
&\le n^{-q}(2q)! .
\end{aligned} \tag{7.32}$$

(7.30), (7.31) and (7.32) imply that

$$\begin{aligned}
\mathbb{P}\{|V_n| \ge \epsilon\} &= \mathbb{P}\{|L_n - \mathbb{E}[L_n]| \ge \epsilon\} \\
&\le \frac{\mathbb{E}[|L_n - \mathbb{E}[L_n]|^{2q}]}{\epsilon^{2q}} \\
&\le 2\epsilon^{-2q}(cq)^q(2L^2)^{2q}n^{-q}(2q)! \\
&\le 2\epsilon^{-2q}(cq)^q(2L^2)^{2q}(2q)^{2q}e^{-2q/3}n^{-q} \\
&\le 2\left(\frac{q^3}{n\epsilon^2/(42L^4)}\right)^q ,
\end{aligned}$$

because $c \cdot 4 \cdot 4 \cdot e^{-2/3} < 42$. We assume that $n\epsilon^2/(42eL^4) \ge 1$, otherwise the bound (7.4) is trivial. Put

$$q = \lfloor [n\epsilon^2/(42eL^4)]^{1/3} \rfloor \ge 1 .$$

Thus,

$$\mathbb{P}\{|V_n| \geq \epsilon\} \leq 2\left(\frac{\lfloor [n\epsilon^2/(42eL^4)]^{1/3} \rfloor^3}{n\epsilon^2/(42L^4)}\right)^{\lfloor [n\epsilon^2/(42eL^4)]^{1/3} \rfloor} \leq 2e^{-n^{1/3}\epsilon^{2/3}/(42eL^4)^{1/3}+1} .$$

Chapter 8

Prediction of time series for squared loss

8.1 The prediction problem

We study the problem of sequential prediction of a real valued sequence. At each time instant $t = 1, 2, \ldots$, the predictor is asked to guess the value of the next outcome y_t of a sequence of real numbers $y_1, y_2, \ldots$ with knowledge of the pasts $y_1^{t-1} = (y_1, \ldots, y_{t-1})$ (where y_1^0 denotes the empty string) and the side information vectors $\mathbf{x}_1^t = (\mathbf{x}_1, \ldots, \mathbf{x}_t)$, where $\mathbf{x}_t \in \mathbb{R}^d$. Thus, the predictor's estimate, at time t, is based on the value of $\mathbf{x}_1^t$ and y_1^{t-1}. A prediction strategy is a sequence $g = \{g_t\}_{t=1}^{\infty}$ of functions

$$g_t : \left(\mathbb{R}^d\right)^t \times \mathbb{R}^{t-1} \to \mathbb{R}$$

so that the prediction formed at time t is $g_t(\mathbf{x}_1^t, y_1^{t-1})$.

In this study we assume that $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots$ are realizations of the random variables $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots$ such that $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ is a stationary and ergodic process.

After n time instants, the *normalized cumulative prediction error* is

$$L_n(g) = \frac{1}{n} \sum_{t=1}^{n} (g_t(\mathbf{X}_1^t, Y_1^{t-1}) - Y_t)^2.$$

Our aim to achieve small $L_n(g)$ when n is large.

For this prediction problem, an example can be the forecasting daily relative prices y_t of an asset, while the side information vector $\mathbf{x}_t$ may contain some information on other assets in the past days or the trading volume in the previous day or some news related

to the actual assets, etc. This is a widely investigated research problem. However, in the vast majority of the corresponding literature the side information is not included in the model, moreover, a parametric model (AR, MA, ARMA, ARIMA, ARCH, GARCH, etc.) is fitted to the stochastic process $\{Y_t\}$, its parameters are estimated, and a prediction is derived from the parameter estimates. Formally, this approach means that there is a parameter θ such that the best predictor has the form

$$\mathbb{E}\{Y_t \mid Y_1^{t-1}\} = g_t(\theta, Y_1^{t-1}),$$

for a function g_t. The parameter θ is estimated from the past data Y_1^{t-1}, and the estimate is denoted by $\hat{\theta}$. Then the data-driven predictor is

$$g_t(\hat{\theta}, Y_1^{t-1}).$$

Here we don't assume any parametric model, so our results are fully nonparametric. This modelling is important for financial data when the process is only approximately governed by stochastic differential equations, so the parametric modelling can be weak, moreover the error criterion of the parameter estimate (usually the maximum likelihood estimate) has no relation to the mean square error of the prediction derived. The main aim of this research is to construct predictors, called universally consistent predictors, which are consistent for all stationary time series. Such universal feature can be proven using the recent principles of nonparametric statistics and machine learning algorithms.

The results below are given in an autoregressive framework, that is, the value Y_t is predicted based on $\mathbf{X}_1^t$ and Y_1^{t-1}. The fundamental limit for the predictability of the sequence can be determined based on a result of Algoet (1994), who showed that for any prediction strategy g and stationary ergodic process $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$,

$$\liminf_{n\to\infty} L_n(g) \geq L^* \quad \text{almost surely,} \tag{8.1}$$

where

$$L^* = \mathbf{E}\big(Y_0 - \mathbf{E}Y_0 \big| \mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\big)^2$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}$.

This lower bound gives sense to the following definition:

Definition 8.1. *A prediction strategy g is called* universally consistent with respect to a class $\mathcal{C}$ of stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$, *if for each process in the class,*

$$\lim_{n\to\infty} L_n(g) = L^* \quad \textit{almost surely.}$$

Next we introduce several simple prediction strategies which build on a methodology worked out in recent years for prediction of individual sequences, see Cesa-Bianchi and Lugosi (2006) for a survey.

8.2 Universally consistent predictions: bounded Y

8.2.1 Partition-based prediction strategies

In this section we introduce our first prediction strategy for bounded ergodic processes. We assume throughout the section that $|Y_0|$ is bounded by a constant $B > 0$, with probability one, and the bound B is known.

The prediction strategy is defined, at each time instant, as a convex combination of *elementary predictors*, where the weighting coefficients depend on the past performance of each elementary predictor.

We define an infinite array of elementary predictors $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\mathcal{P}_\ell = \{A_{\ell,j}, j = 1, 2, \ldots, m_\ell\}$ be a sequence of finite partitions of $\mathbb{R}$, and let $\mathcal{Q}_\ell = \{B_{\ell,j}, j = 1, 2, \ldots, m'_\ell\}$ be a sequence of finite partitions of $\mathbb{R}^d$. Introduce the corresponding quantizers:

$$F_\ell(y) = j, \text{ if } y \in A_{\ell,j}$$

and

$$G_\ell(\mathbf{x}) = j, \text{ if } \mathbf{x} \in B_{\ell,j} \ .$$

With some abuse of notation, for any n and $y_1^n \in \mathbb{R}^n$, we write $F_\ell(y_1^n)$ for the sequence $F_\ell(y_1), \ldots, F_\ell(y_n)$, and similarly, for $\mathbf{x}_1^n \in (\mathbb{R}^d)^n$, we write $G_\ell(\mathbf{x}_1^n)$ for the sequence $G_\ell(\mathbf{x}_1), \ldots, G_\ell(\mathbf{x}_n)$.

Fix positive integers k, ℓ, and for each $k+1$-long string z of positive integers, and for each k-long string s of positive integers, define the partitioning regression function estimate

$$\widehat{E}_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}, z, s) = \frac{\sum_{\{k<t<n: G_\ell(\mathbf{x}_{t-k}^t)=z,\, F_\ell(y_{t-k}^{t-1})=s\}} y_t}{\left|\{k < t < n : G_\ell(\mathbf{x}_{t-k}^t) = z,\, F_\ell(y_{t-k}^{t-1}) = s\}\right|},$$

for all $n > k+1$ where $0/0$ is defined to be 0.

Define the elementary predictor $h^{(k,\ell)}$ by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \widehat{E}_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}, G_\ell(\mathbf{x}_{n-k}^n), F_\ell(y_{n-k}^{n-1})),$$

for $n = 1, 2, \ldots$. That is, $h_n^{(k,\ell)}$ quantizes the sequence $\mathbf{x}_1^n, y_1^{n-1}$ according to the partitions $\mathcal{Q}_\ell$ and $\mathcal{P}_\ell$, and looks for all appearances of the last seen quantized strings $G_\ell(\mathbf{x}_{n-k}^n)$ of length $k+1$ and $F_\ell(y_{n-k}^{n-1})$ of length k in the past. Then it predicts according to the average of the y_t's following the string.

In contrast to the nonparametric regression estimation problem from i.i.d. data, for ergodic observations, it is impossible to choose $k = k_n$ and $\ell = \ell_n$ such that the corresponding predictor is universally consistent for the class of bounded ergodic processes.

The very important new principle is the combination or aggregation of elementary predictors (cf. Cesa-Bianchi and Lugosi (2006)). The proposed prediction algorithm proceeds as follows: let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) of positive integers such that for all k,ℓ, $q_{k,\ell} > 0$. Put $c = 8B^2$, and define the weights

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/c} \tag{8.2}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} \ , \tag{8.3}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{8.4}$$

The prediction strategy g is defined by

$$g_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(\mathbf{x}_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \ldots \tag{8.5}$$

i.e., the prediction g_t is the convex linear combination of the elementary predictors such that an elementary predictor has non-negligible weight in the combination if it has good performance until time $t-1$.

Theorem 8.1. (Györfi and Lugosi (2002)) *Assume that*
(a) the sequences of partition $\mathcal{P}_\ell$ is nested, that is, any cell of $\mathcal{P}_{\ell+1}$ is a subset of a cell of $\mathcal{P}_\ell$, $\ell = 1, 2, \ldots$;
(b) the sequences of partition $\mathcal{Q}_\ell$ is nested;
(c) the sequences of partition $\mathcal{P}_\ell$ is asymptotically fine, that is, for each sphere S centered at the origin

$$\lim_{\ell \to \infty} \max_{A \in \mathcal{P}_\ell, \, A \cap S \neq \emptyset} \mathrm{diam}(A) = 0;$$

(d) the sequences of partition $\mathcal{Q}_\ell$ is asymptotically fine;
Then the prediction scheme g defined above is universal with respect to the class of all stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ such that $|Y_0| \leq B$.

One of the main ingredients of the proof is the following lemma, whose proof is a straightforward extension of standard arguments in the prediction theory of individual sequences, see, for example, Kivinen and Warmuth (1999).

Lemma 8.1. *Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $\tilde{h}_i(\mathbf{x}_1^n, y_1^{n-1}) \in [-B, B]$ and $y_1^n \in [-B, B]^n$. Define*

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1}) \qquad t = 1, 2, \ldots$$

then for every $n \geq 1$,

$$L_n(\tilde{g}) \leq \inf_k \left(L_n(\tilde{h}_k) - \frac{c \ln q_k}{n} \right).$$

Here $-\ln 0$ is treated as ∞.

PROOF. Introduce

$$W_1 = 1$$

and

$$W_t = \sum_{k=1}^{\infty} w_{t,k}$$

for $t > 1$. Note that

$$W_{t+1} = \sum_{k=1}^{\infty} w_{t,k} e^{-\left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)^2/c} = W_t \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)^2/c},$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)^2/c} \right).$$

Introduce the function

$$F_t(z) = e^{-(y_t - z)^2/c}$$

Because of $c \geq 8B^2$, the function F_t is concave on $[-B, B]$, therefore Jensen's inequality implies that

$$\left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)\right]^2 \leq -c \ln \frac{W_{t+1}}{W_t} \tag{8.6}$$

Thus,

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(\mathbf{x}_1^t, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)\right]^2 \\ &\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\ &= -c \ln W_{n+1} \end{aligned}$$

and therefore

$$\begin{aligned} nL_n(\tilde{g}) &\leq -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\ &= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\ &\leq -c \ln \left(\sup_k q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\ &= \inf_k \left(-c \ln q_k + nL_n(\tilde{h}_k)\right) \ , \end{aligned}$$

which concludes the proof. □

Another main ingredient of the proof of Theorem 8.1 is known as Breiman's generalized ergodic theorem, see also Algoet (1994) and Györfi et al. (2002).

Lemma 8.2. (Breiman (1957)). *Let $Z = \{Z_i\}_{-\infty}^{\infty}$ be a stationary and ergodic process. Let T denote the left shift operator. Let f_i be a sequence of real-valued functions such that for some function f, $f_i(Z) \to f(Z)$ almost surely. Assume that $\mathbb{E}\{\sup_i |f_i(Z)|\} < \infty$. Then*

$$\lim_{t \to \infty} \frac{1}{n} \sum_{i=1}^{n} f_i(T^i Z) = \mathbb{E}\{f(Z)\} \qquad \textit{almost surely.}$$

PROOF OF THEOREM 8.1. Because of (8.1), it is enough to show that

$$\limsup_{n\to\infty} L_n(g) \le L^* \qquad \text{a.s.}$$

By a double application of the ergodic theorem, as $n \to \infty$, almost surely,

$$\begin{aligned}
\widehat{E}_n^{(k,\ell)}(\mathbf{X}_1^n, Y_1^{n-1}, z, s) &= \frac{\frac{1}{n}\sum_{\{k<i<n: G_\ell(\mathbf{X}_{t-k}^t)=z,\, F_\ell(Y_{t-k}^{t-1})=s\}} Y_i}{\frac{1}{n}\left|\{k<i<n : G_\ell(\mathbf{X}_{t-k}^t)=z,\, F_\ell(Y_{t-k}^{t-1})=s\}\right|} \\
&\to \frac{\mathbb{E}\{Y_0 I_{\{G_\ell(\mathbf{X}_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}}\}}{\mathbb{P}\{G_\ell(\mathbf{X}_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}} \\
&= \mathbb{E}\{Y_0 | G_\ell(\mathbf{X}_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\},
\end{aligned}$$

and therefore

$$\lim_{n\to\infty} \sup_z \sup_s |\widehat{E}_n^{(k,\ell)}(\mathbf{X}_1^n, Y_1^{n-1}, z, s) - \mathbb{E}\{Y_0 | G_\ell(\mathbf{X}_{-k}^0)=z,\, F_\ell(Y_{-k}^{-1})=s\}| = 0$$

almost surely. Thus, by Lemma 8.2, as $n \to \infty$, almost surely,

$$\begin{aligned}
L_n(h^{(k,\ell)}) &= \frac{1}{n}\sum_{i=1}^n (h^{(k,\ell)}(\mathbf{X}_1^i, Y_1^{i-1}) - Y_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (\widehat{E}_n^{(k,\ell)}(\mathbf{X}_1^i, Y_1^{i-1}, G_\ell(\mathbf{X}_{i-k}^i), F_\ell(Y_{i-k}^{i-1})) - Y_i)^2 \\
&\to \mathbb{E}\{(Y_0 - \mathbb{E}\{Y_0 | G_\ell(\mathbf{X}_{-k}^0),\, F_\ell(Y_{-k}^{-1})\})^2\} \\
&\stackrel{\text{def}}{=} \epsilon_{k,\ell}.
\end{aligned}$$

Since the partitions $\mathcal{P}_\ell$ and $\mathcal{Q}_\ell$ are nested, $\mathbb{E}\left\{Y_0 | G_\ell(\mathbf{X}_{-k}^0),\, F_\ell(Y_{-k}^{-1})\right\}$ is a martingale indexed by the pair (k,ℓ). Thus, the martingale convergence theorem (see, e.g., Stout (1974)) and assumption (c) and (d) for the sequence of partitions implies that

$$\inf \epsilon_{k,\ell} = \lim_{k,\ell\to\infty} \epsilon_{k,\ell} = \mathbb{E}\left\{\left(Y_0 - \mathbb{E}\{Y_0 | \mathbf{X}_{-\infty}^0,\, Y_{-\infty}^{-1}\}\right)^2\right\} = L^*.$$

Now by Lemma 8.1,

$$L_n(g) \le \inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n}\right), \tag{8.7}$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\leq \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\leq \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\leq \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&= \inf_{k,\ell} \epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty} \epsilon_{k,\ell} \\
&= L^*
\end{aligned}$$

and the proof of the theorem is finished. □

8.2.2 Kernel-based prediction strategies

We introduce in this section a class of *kernel-based* prediction strategies for stationary and ergodic sequences. The main advantage of this approach in contrast to the partition-based strategy is that it replaces the rigid discretization of the past appearances by more flexible rules. This also often leads to faster algorithms in practical applications.

To simplify the notation, we start with the simple "moving-window" scheme, corresponding to a naiv kernel function. Just like before, we define an array of experts $h^{(k,\ell)}$, where k and ℓ are positive integers. We associate to each pair (k,ℓ) two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$ such that, for any fixed k

$$\lim_{\ell\to\infty} r_{k,\ell} = 0, \tag{8.8}$$

and

$$\lim_{\ell\to\infty} r'_{k,\ell} = 0. \tag{8.9}$$

Finally, let the location of the matches be

$$J_n^{(k,\ell)} = \left\{ k < t < n : \|\mathbf{x}_{t-k}^t - \mathbf{x}_{n-k}^n\| \leq r_{k,\ell},\ \|y_{t-k}^{t-1} - y_{n-k}^{n-1}\| \leq r'_{k,\ell} \right\}$$

Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}, \qquad n > k+1, \tag{8.10}$$

where 0/0 is defined to be 0. The pool of experts is mixed the same way as in the case of the partition-based strategy (cf. (8.2), (8.3), (8.4) and (8.5)).

Theorem 8.2. *Suppose that (8.8) and (8.9) are verified. Then the kernel-based strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes* $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ *such that* $|Y_0| \leq B$.

8.2.3 Nearest neighbor-based prediction strategy

This strategy is yet more robust with respect to the kernel strategy and thus also with respect to the partition strategy. Since it does not suffer from scaling problem as partition and kernel-based strategies where the quantizer and the radius has to be carefully chosen to obtain "good" performance. As well as this, in practical applications it runs extremely fast compared with the kernel and partition schemes as it is much less likely to get bogged down in calculations for certain experts.

To introduce the strategy, we start again by defining an infinite array of experts $h^{(k,\ell)}$, where k and ℓ are positive integers. Just like before, k is the length of the past observation vectors being scanned by the elementary expert and, for each ℓ, choose $p_\ell \in (0,1)$ such that

$$\lim_{\ell \to \infty} p_\ell = 0\,, \tag{8.11}$$

and set

$$\bar{\ell} = \lfloor p_\ell n \rfloor$$

(where $\lfloor . \rfloor$ is the floor function). At time n, for fixed k and ℓ ($n > k + \bar{\ell} + 1$), the expert searches for the $\bar{\ell}$ nearest neighbors (NN) of the last seen observation $\mathbf{x}_{n-k}^n$ and y_{n-k}^{n-1} in the past and predicts accordingly. More precisely, let

$$\begin{aligned} J_n^{(k,\ell)} = \big\{\, & k < t < n : (\mathbf{x}_{t-k}^t, y_{t-k}^{t-1}) \text{ is among the } \bar{\ell} \text{ NN of } (\mathbf{x}_{n-k}^n, y_{n-k}^{n-1}) \text{ in} \\ & (\mathbf{x}_1^{k+1}, y_1^k), \ldots, (\mathbf{x}_{n-k-1}^{n-1}, y_{n-k-1}^{n-2}) \big\} \end{aligned}$$

and introduce the elementary predictor

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \frac{\sum_{\{t \in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}$$

if the sum is nonvoid, and 0 otherwise. Finally, the experts are mixed as before (cf. (8.2), (8.3), (8.4) and (8.5)).

Theorem 8.3. *Suppose that (8.11) is verified and that for each vector* $\mathbf{s}$ *the random variable*

$$\|(\mathbf{X}_1^{k+1}, Y_1^k) - \mathbf{s}\|$$

has a continuous distribution function. Then the nearest neighbor strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ *such that* $|Y_0| \le B$.

8.2.4 Generalized linear estimates

This section is devoted to an alternative way of defining a universal predictor for stationary and ergodic processes. It is in effect an extension of the approach presented in Györfi and Lugosi (2002). Once again, we apply the method described in the previous sections to combine elementary predictors, but now we use elementary predictors which are generalized linear predictors. More precisely, we define an infinite array of elementary experts $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$ as follows. Let $\{\phi_j^{(k)}\}_{j=1}^{\ell}$ be real-valued functions defined on $(\mathbb{R}^d)^{(k+1)} \times \mathbb{R}^k$. The elementary predictor $h_n^{(k,\ell)}$ generates a prediction of form

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \sum_{j=1}^{\ell} c_{n,j} \phi_j^{(k)}(\mathbf{x}_{n-k}^n, y_{n-k}^{n-1}) \,,$$

where the coefficients $c_{n,j}$ are calculated according to the past observations $\mathbf{x}_1^n$, y_1^{n-1}. More precisely, the coefficients $c_{n,j}$ are defined as the real numbers which minimize the criterion

$$\sum_{t=k+1}^{n-1} \left(\sum_{j=1}^{\ell} c_j \phi_j^{(k)}(\mathbf{x}_{t-k}^t, y_{t-k}^{t-1}) - y_t \right)^2 \tag{8.12}$$

if $n > k+1$, and the all-zero vector otherwise. It can be shown using a recursive technique (see e.g., Tsypkin (1971), Györfi (1984) and Györfi and Lugosi (2002)) that the $c_{n,j}$ can be calculated with small computational complexity.

The experts are mixed via an exponential weighting, which is defined the same way as earlier (cf. (8.2), (8.3), (8.4) and (8.5)).

Theorem 8.4. (Györfi and Lugosi (2002)) *Suppose that* $|\phi_j^{(k)}| \le 1$ *and, for any fixed* k*, suppose that the set*

$$\left\{ \sum_{j=1}^{\ell} c_j \phi_j^{(k)}; \quad (c_1, \ldots, c_\ell), \ \ell = 1, 2, \ldots \right\}$$

is dense in the set of continuous functions of $d(k+1)+k$ variables. Then the generalized linear strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ such that $|Y_0| \le B$.

8.3 Universally consistent predictions: unbounded Y

8.3.1 Partition-based prediction strategies

Let $\widehat{E}_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}, z, s)$ be defined as in Section 8.2.1. Introduce the truncation function

$$T_m(z) = \begin{cases} m & \text{if } z > m \\ z & \text{if } |z| < m \\ -m & \text{if } z < -m, \end{cases}$$

Define the elementary predictor $h^{(k,\ell)}$ by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = T_{n^\delta}\left(\widehat{E}_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}, G_\ell(\mathbf{x}_{n-k}^n), F_\ell(y_{n-k}^{n-1}))\right),$$

where

$$0 < \delta < 1/8,$$

for $n = 1, 2, \ldots$. That is, $h_n^{(k,\ell)}$ is the truncation of the elementary predictor introduced in Section 8.2.1.

The proposed prediction algorithm proceeds as follows: let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k, ℓ) of positive integers such that for all k, ℓ, $q_{k,\ell} > 0$. For a time dependent learning parameter $\eta_t > 0$, define the weights

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/\sqrt{t}} \tag{8.13}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} \ , \tag{8.14}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{8.15}$$

The prediction strategy g is defined by

$$g_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(\mathbf{x}_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \ldots \tag{8.16}$$

Theorem 8.5. (GYÖRFI AND OTTUCSÁK (2007)) *Assume that the conditions (a), (b), (c) and (d) of Theorem 8.1 are satisfied. Then the prediction scheme g defined above is universally consistent with respect to the class of all stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ such that*

$$\mathbb{E}\{Y_1^4\} < \infty.$$

Here we describe a result, which is used in the analysis.

Lemma 8.3. (GYÖRFI AND OTTUCSÁK (2007)) *Let $h^{(1)}, h^{(2)}, \dots$ be a sequence of prediction strategies (experts). Let $\{q_k\}$ be a probability distribution on the set of positive integers. Denote the normalized loss of the expert $h = (h_1, h_2, \dots)$ by*

$$L_n(h) = \frac{1}{n} \sum_{t=1}^{n} \lambda_t(h),$$

where

$$\lambda_t(h) = \lambda(h_t, Y_t)$$

and the loss function λ is convex in its first argument h. Define

$$w_{t,k} = q_k e^{-\eta_t (t-1) L_{t-1}(h^{(k)})}$$

where $\eta_t > 0$ is monotonically decreasing, and

$$p_{t,k} = \frac{w_{t,k}}{W_t}$$

where

$$W_t = \sum_{k=1}^{\infty} w_{t,k} \ .$$

If the prediction strategy $g = (g_1, g_2, \dots)$ is defined by

$$g_t = \sum_{k=1}^{\infty} p_{t,k} h_t^{(k)} \qquad t = 1, 2, \dots$$

then for every $n \geq 1$,

$$L_n(g) \leq \inf_k \left(L_n(h^{(k)}) - \frac{\ln q_k}{n \eta_{n+1}} \right) + \frac{1}{2n} \sum_{t=1}^{n} \eta_t \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}).$$

PROOF. Introduce some notations:

$$w'_{t,k} = q_k e^{-\eta_{t-1}(t-1)L_{t-1}(h^{(k)})},$$

which is the weight $w_{t,k}$, where η_t is replaced by η_{t-1} and the sum of these are

$$W'_t = \sum_{k=1}^{\infty} w'_{t,k}.$$

We start the proof with the following chain of bounds:

$$\begin{aligned}
\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} &= \frac{1}{\eta_t} \ln \frac{\sum_{k=1}^{\infty} w_{t,k} e^{-\eta_t \lambda_t(h^{(k)})}}{W_t} \\
&= \frac{1}{\eta_t} \ln \sum_{k=1}^{\infty} p_{t,k} e^{-\eta_t \lambda_t(h^{(k)})} \\
&\leq \frac{1}{\eta_t} \ln \sum_{k=1}^{\infty} p_{t,k} \left(1 - \eta_t \lambda_t(h^{(k)}) + \frac{\eta_t^2}{2} \lambda_t^2(h^{(k)})\right)
\end{aligned}$$

because of $e^{-x} \leq 1 - x + x^2/2$ for $x \geq 0$. Moreover,

$$\begin{aligned}
&\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} \\
&\leq \frac{1}{\eta_t} \ln \left(1 - \eta_t \sum_{k=1}^{\infty} p_{t,k} \lambda_t(h^{(k)}) + \frac{\eta_t^2}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)})\right) \\
&\leq -\sum_{k=1}^{\infty} p_{t,k} \lambda_t(h^{(k)}) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \qquad (8.17) \\
&= -\sum_{k=1}^{\infty} p_{t,k} \lambda(h_t^{(k)}, Y_t) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \\
&\leq -\lambda \left(\sum_{k=1}^{\infty} p_{t,k} h_t^{(k)}, Y_t\right) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \qquad (8.18) \\
&= -\lambda_t(g) + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \qquad (8.19)
\end{aligned}$$

where (8.17) follows from the fact that $\ln(1+x) \leq x$ for all $x > -1$ and in (8.18) we used the convexity of the loss $\lambda(h, y)$ in its first argument h. From (8.19) after rearranging we obtain

$$\lambda_t(g) \leq -\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t} + \frac{\eta_t}{2} \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \ .$$

Then write a telescope formula:

$$\begin{aligned} \frac{1}{\eta_t} \ln W_t - \frac{1}{\eta_t} \ln W'_{t+1} &= \left(\frac{1}{\eta_t} \ln W_t - \frac{1}{\eta_{t+1}} \ln W_{t+1} \right) \\ &\quad + \left(\frac{1}{\eta_{t+1}} \ln W_{t+1} - \frac{1}{\eta_t} \ln W'_{t+1} \right) \\ &= (A_t) + (B_t). \end{aligned}$$

We have that

$$\begin{aligned} \sum_{t=1}^{n} A_t &= \sum_{t=1}^{n} \left(\frac{1}{\eta_t} \ln W_t - \frac{1}{\eta_{t+1}} \ln W_{t+1} \right) \\ &= \frac{1}{\eta_1} \ln W_1 - \frac{1}{\eta_{n+1}} \ln W_{n+1} \\ &= -\frac{1}{\eta_{n+1}} \ln \sum_{k=1}^{\infty} q_k e^{-\eta_{n+1} n L_n(h^{(k)})} \\ &\leq -\frac{1}{\eta_{n+1}} \ln \sup_k q_k e^{-\eta_{n+1} n L_n(h^{(k)})} \\ &= -\frac{1}{\eta_{n+1}} \sup_k \left(\ln q_k - \eta_{n+1} n L_n(h^{(k)}) \right) \\ &= \inf_k \left(n L_n(h^{(k)}) - \frac{\ln q_k}{\eta_{n+1}} \right). \end{aligned}$$

$\frac{\eta_{t+1}}{\eta_t} \leq 1$, therefore applying Jensen's inequality for concave function, we get that

$$\begin{aligned} W_{t+1} &= \sum_{i=1}^{\infty} q_i e^{-\eta_{t+1} t L_t(h^{(i)})} \\ &= \sum_{i=1}^{\infty} q_i \left(e^{-\eta_t t L_t(h^{(i)})}\right)^{\frac{\eta_{t+1}}{\eta_t}} \\ &\leq \left(\sum_{i=1}^{\infty} q_i e^{-\eta_t t L_t(h^{(i)})}\right)^{\frac{\eta_{t+1}}{\eta_t}} \\ &= \left(W'_{t+1}\right)^{\frac{\eta_{t+1}}{\eta_t}} . \end{aligned}$$

Thus,

$$\begin{aligned} B_t &= \frac{1}{\eta_{t+1}} \ln W_{t+1} - \frac{1}{\eta_t} \ln W'_{t+1} \\ &\leq \frac{1}{\eta_{t+1}} \frac{\eta_{t+1}}{\eta_t} \ln W'_{t+1} - \frac{1}{\eta_t} \ln W'_{t+1} \\ &= 0. \end{aligned}$$

We can summarize the bounds:

$$L_n(g) \leq \inf_k \left(L_n(h^{(k)}) - \frac{\ln q_k}{n\eta_{n+1}} \right) + \frac{1}{2n} \sum_{t=1}^{n} \eta_t \sum_{k=1}^{\infty} p_{t,k} \lambda_t^2(h^{(k)}) \ .$$

□

Proof of Theorem 8.5. Because of (8.1), it is enough to show that

$$\limsup_{n\to\infty} L_n(g) \leq L^* \qquad \text{a.s.}$$

Because of the proof of Theorem 8.1, as $n \to \infty$, a.s.,

$$\widehat{E}_n^{(k,\ell)}(\mathbf{X}_1^n, Y_1^{n-1}, z, s) \to \mathbb{E}\{Y_0 \mid G_\ell(\mathbf{X}_{-k}^0) = z,\ F_\ell(Y_{-k}^{-1}) = s\},$$

and therefore for all z and s

$$T_{n^\delta}\left(\widehat{E}_n^{(k,\ell)}(\mathbf{X}_1^n, Y_1^{n-1}, z, s)\right) \to \mathbb{E}\{Y_0 \mid G_\ell(\mathbf{X}_{-k}^0) = z,\ F_\ell(Y_{-k}^{-1}) = s\}.$$

By Lemma 8.2, as $n \to \infty$, almost surely,

$$\begin{aligned}
&L_n(h^{(k,\ell)}) \\
&= \frac{1}{n}\sum_{t=1}^{n}(h^{(k,\ell)}(\mathbf{X}_1^t, Y_1^{t-1}) - Y_t)^2 \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(T_{t^\delta}\left(\widehat{E}_t^{(k,\ell)}(\mathbf{X}_1^t, Y_1^{t-1}, G_\ell(\mathbf{X}_{t-k}^t), F_\ell(Y_{t-k}^{t-1}))\right) - Y_t\right)^2 \\
&\to \mathbb{E}\{(Y_0 - \mathbb{E}\{Y_0 \mid G_\ell(\mathbf{X}_{-k}^0),\, F_\ell(Y_{-k}^{-1})\})^2\} \\
&\stackrel{\text{def}}{=} \epsilon_{k,\ell}.
\end{aligned}$$

In the same way as in the proof of Theorem 8.1, we get that

$$\inf_{k,l}\epsilon_{k,l} = \lim_{k,\ell\to\infty}\epsilon_{k,\ell} = \mathbb{E}\left\{\left(Y_0 - \mathbb{E}\{Y_0|\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\}\right)^2\right\} = L^*.$$

Apply Lemma 8.3 with choice $\eta_t = \frac{1}{\sqrt{t}}$ and for the squared loss $\lambda_t(h) = (h_t - Y_t)^2$, then the square loss is convex in its first argument h, so

$$\begin{aligned}
L_n(g) \le \inf_{k,\ell}&\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right) \\
&+\frac{1}{2n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{t,k,\ell}\big(h^{(k,\ell)}(\mathbf{X}_1^t, Y_1^{t-1}) - Y_t\big)^4. \qquad (8.20)
\end{aligned}$$

On the one hand, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right) \\
&\le \inf_{k,\ell}\limsup_{n\to\infty}\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right) \\
&= \inf_{k,\ell}\limsup_{n\to\infty}L_n(h^{(k,\ell)}) \\
&= \inf_{k,\ell}\epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty}\epsilon_{k,\ell} \\
&= L^*.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell}p_{t,k,\ell}(h^{(k,\ell)}(\mathbf{X}_1^t,Y_1^{t-1})-Y_t)^4\\
&\quad\leq\ \frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell}p_{t,k,\ell}\left(h^{(k,\ell)}(\mathbf{X}_1^t,Y_1^{t-1})^4+Y_t^4\right)\\
&\quad\leq\ \frac{8}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell}p_{t,k,\ell}\left(t^{4\delta}+Y_t^4\right)\\
&\quad=\ \frac{8}{n}\sum_{t=1}^{n}\frac{t^{4\delta}+Y_t^4}{\sqrt{t}},
\end{aligned}$$

therefore, almost surely,

$$\begin{aligned}
&\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell}p_{t,k,\ell}(h^{(k,\ell)}(\mathbf{X}_1^t,Y_1^{t-1})-Y_t)^4\\
\leq\ &\limsup_{n\to\infty}\frac{8}{n}\sum_{t=1}^{n}\frac{Y_t^4}{\sqrt{t}}\\
=\ &0,
\end{aligned}$$

where we applied that $\mathbb{E}\{Y_1^4\}<\infty$ and $0<\delta<\frac{1}{8}$. Summarizing these bounds, we get that, almost surely,

$$\limsup_{n\to\infty}L_n(g)\leq L^*$$

and the proof of the theorem is finished. □

8.3.2 Kernel-based prediction strategies

Apply the notations of Section 8.2.2. Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n,y_1^{n-1})=T_{\min\{n^\delta,\ell\}}\left(\frac{\sum_{\{t\in J_n^{(k,\ell)}\}}y_t}{|J_n^{(k,\ell)}|}\right),\qquad n>k+1,$$

where $0/0$ is defined to be 0 and $0<\delta<1/8$. The pool of experts is mixed the same way as in the case of the partition-based strategy (cf. (8.13), (8.14), (8.15) and (8.16)).

Theorem 8.6. (Biau et al (2010)) *Suppose that (8.8) and (8.9) are verified. Then the kernel-based strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes* $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ *such that*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

8.3.3 Nearest neighbor-based prediction strategy

Apply the notations of Section 8.2.3. Then the elementary expert $h_n^{(k,\ell)}$ at time n is defined by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = T_{\min\{n^\delta,\ell\}}\left(\frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}\right), \qquad n > k+1,$$

if the sum is nonvoid, and 0 otherwise and $0 < \delta < 1/8$. The pool of experts is mixed the same way as in the case of the histogram-based strategy (cf. (8.13), (8.14), (8.15) and (8.16)).

Theorem 8.7. (Biau et al (2010)) *Suppose that (8.11) is verified, and that for each vector* $\mathbf{s}$ *the random variable*

$$\|(\mathbf{X}_1^{k+1}, Y_1^k) - \mathbf{s}\|$$

has a continuous distribution function. Then the nearest neighbor strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ *such that*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

8.3.4 Generalized linear estimates

Apply the notations of Section 8.2.4. The elementary predictor $h_n^{(k,\ell)}$ generates a prediction of form

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = T_{\min\{n^\delta,\ell\}}\left(\sum_{j=1}^{\ell} c_{n,j}\phi_j^{(k)}(\mathbf{x}_{n-k}^n, y_{n-k}^{n-1})\right),$$

with $0 < \delta < 1/8$. The pool of experts is mixed the same way as in the case of the histogram-based strategy (cf. (8.13), (8.14), (8.15) and (8.16)).

Theorem 8.8. (BIAU ET AL (2010)) *Suppose that* $|\phi_j^{(k)}| \leq 1$ *and, for any fixed* k, *suppose that the set*

$$\left\{ \sum_{j=1}^{\ell} c_j \phi_j^{(k)}; \quad (c_1, \ldots, c_\ell), \ \ell = 1, 2, \ldots \right\}$$

is dense in the set of continuous functions of $d(k+1)+k$ *variables. Then the generalized linear strategy defined above is universally consistent with respect to the class of all stationary and ergodic processes* $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ *such that*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

8.3.5 Prediction of gaussian processes

We consider in this section the classical problem of gaussian time series prediction. In this context, parametric models based on distribution assumptions and structural conditions such as AR(p), MA(q), ARMA(p,q) and ARIMA(p,d,q) are usually fitted to the data. However, in the spirit of modern nonparametric inference, we try to avoid such restrictions on the process structure. Thus, we only assume that we observe a string realization y_1^{n-1} of a zero mean, stationary and ergodic, gaussian process $\{Y_n\}_{-\infty}^{\infty}$, and try to predict y_n, the value of the process at time n. Note that there is no side information vectors $\mathbf{x}_1^n$ in this purely time series prediction framework.

For Gaussian time series and for any integer $k > 0$, $\mathbb{E}\{Y_n \mid Y_{n-k}^{n-1}\}$ is a linear function of Y_{n-k}^{n-1}:

$$\mathbb{E}\{Y_n \mid Y_{n-k}^{n-1}\} = \sum_{j=1}^{k} c_j^{(k)} Y_{n-j}, \tag{8.21}$$

where the coefficients $c_j^{(k)}$ minimize the risk

$$\mathbb{E}\left\{ \left(\sum_{j=1}^{k} c_j Y_{-j} - Y_0 \right)^2 \right\},$$

therefore the main ingredient is the estimate of the coefficients $c_1^{(k)}, \ldots, c_k^{(k)}$ from the data Y_1^{n-1}. Such an estimate is called elementary predictor, it is denoted by $\tilde{h}^{(k)}$ generating a prediction of form

$$\tilde{h}^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{k} C_{n,j}^{(k)} Y_{n-j}$$

such that the coefficients $C_{n,j}^{(k)}$ minimize the empirical risk

$$\sum_{i=k+1}^{n-1}\left(\sum_{j=1}^{k} c_j Y_{i-j} - Y_i\right)^2$$

if $n > k$, and the all-zero vector otherwise. Even though the minimum always exists, it is not unique in general, and therefore the minimum is not well-defined. It is shown by Györfi (1984) that there is a unique vector $C_n^{(k)} = (C_{n,1}^{(k)}, \ldots, C_{n,k}^{(k)})$ such that

$$\sum_{i=k+1}^{n-1}\left(\sum_{j=1}^{k} C_{n,j}^{(k)} Y_{i-j} - Y_i\right)^2 = \min_{(c_1,\ldots,c_k)} \sum_{i=k+1}^{n-1}\left(\sum_{j=1}^{k} c_j Y_{i-j} - Y_i\right)^2,$$

and it has the smallest Euclidean norm among the minimizer vectors.

We set

$$T_a(z) = \begin{cases} a & \text{if } z > a; \\ z & \text{if } |z| < a; \\ -a & \text{if } z < -a. \end{cases}$$

and

$$h_n^{(k)}(y_1^{n-1}) = T_{\min\{n^\delta, k\}}\left(\tilde{h}_n^{(k)}(y_1^{n-1})\right),$$

where $0 < \delta < \frac{1}{8}$, and combine these experts as before. Precisely, let $\{q_k\}$ be an arbitrarily probability distribution over the positive integers such that for all k, $q_k > 0$, define the weights

$$w_{k,n} = q_k e^{-(n-1)L_{n-1}(h_n^{(k)})/\sqrt{n}}$$

and their normalized values

$$p_{k,n} = \frac{w_{k,n}}{\sum_{i=1}^{\infty} w_{i,n}}.$$

The prediction strategy g at time n is defined by

$$g_n(y_1^{n-1}) = \sum_{k=1}^{\infty} p_{k,n} h_n^{(k)}(y_1^{n-1}), \qquad n = 1, 2, \ldots$$

Theorem 8.9. (Biau et al (2010)) *The prediction strategy g defined above is universally consistent with respect to the class of all stationary and ergodic zero-mean gaussian processes $\{Y_n\}_{-\infty}^{\infty}$.*

The following corollary shows that the strategy g provides asymptotically a good estimate of the regression function in the following sense:

Corollary 8.1. (Biau et al (2010)) *Under the conditions of Theorem 8.9,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\{Y_t \mid Y_1^{t-1}\} - g(Y_1^{t-1}) \right)^2 = 0 \quad \textit{almost surely.}$$

Corollary 8.1 is expressed in terms of an almost sure Cesáro consistency. It is an *open problem* to know whether there exists a prediction rule g such that

$$\lim_{n\to\infty} \left(\mathbb{E}\{Y_n | Y_1^{n-1}\} - g(Y_1^{n-1}) \right) = 0 \quad \text{almost surely} \tag{8.22}$$

for all stationary and ergodic gaussian processes.

Schäfer (2002) investigated the following predictor: choose $L_n \uparrow \infty$, then his predictor is

$$\bar{g}_n(Y_1^{n-1}) = \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} T_{L_n}(Y_{n-j}).$$

Schäfer (2002) proved that, under some conditions on the Gaussian process, we have that

$$\lim_{n\to\infty} \left(\mathbb{E}\{Y_n \mid Y_{n-k_n}^{n-1}\} - \bar{g}_n(Y_1^{n-1}) \right) = 0 \quad \text{a.s.}$$

His conditions include that the process has the MA(∞) representation

$$\sum_{j=0}^{\infty} a_j^* Z_{n-j}, \tag{8.23}$$

with i.i.d. Gaussian innovations $\{Z_n\}$ and with

$$\sum_{i=1}^{\infty} |a_i^*|^2 < \infty, \tag{8.24}$$

such that

$$\sum_{i=1}^{\infty} |a_i^*| < \infty, \tag{8.25}$$

and therefore it is purely nondeterministic and the spectral density exists. Moreover, he assumed that

$$\mathbb{E}\{Y_n \mid Y_{-\infty}^{n-1}\} - \mathbb{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} \to 0$$

a.s. For example, he proved the strong consistency with $k_n = n^{1/4}$ if the spectral density is bounded away from zero. The question left is how to avoid these conditions such that we pose conditions only on the covariances.

Györfi, Sanchetta (2014) studied the open problem (8.22). For fixed k, an elementary predictor

$$\tilde{h}^{(k)}(Y_1^{n-1}) = \sum_{j=1}^{k} C_{n,j}^{(k)} Y_{n-j}$$

cannot be consistent. In order to get consistent predictions there are three main principles:

- k is a deterministic function of n,
- k depends on the data Y_1^{n-1},
- aggregate the elementary predictors $\{\tilde{h}^{(k)}(Y_1^{n-1}), k = 1, 2, \ldots, n-2\}$.

For a deterministic sequence $k_n, n = 1, 2, \ldots$, consider the predictor

$$\tilde{g}_n(Y_1^{n-1}) = \tilde{h}^{(k_n)}(Y_1^{n-1}) = \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} Y_{n-j}.$$

We guess that the following is true:

Conjecture 8.1. *For any deterministic sequence k_n, there is a stationary, ergodic Gaussian process such that the prediction error*

$$\mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \sum_{j=1}^{k_n} C_{n,j}^{(k_n)} Y_{n-j}$$

does not converge to 0 *a.s.*

For the prediction error $\mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \tilde{g}_n(Y_1^{n-1})$ we have the decomposition

$$\mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \tilde{g}_n(Y_1^{n-1}) = I_n + J_n,$$

where

$$I_n = \mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \mathbb{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\}$$

is the approximation error, and

$$J_n = \mathbb{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} - \tilde{g}_n(Y_1^{n-1}) = \sum_{j=1}^{k_n} (c_j^{(k_n)} - C_{n,j}^{(k_n)}) Y_{n-j}$$

is the estimation error. In order to have small approximation error, we need $k_n \to \infty$, while the control of the estimation error is possible if this convergence to ∞ is slow.

The approximation error tends to zero in L_2 without any condition:

Proposition 8.1. (Györfi, Sanchetta (2014)) *For any sequence $k_n \to \infty$ and for any stationary process $\{Y_n\}_{-\infty}^{\infty}$,*

$$\lim_{n\to\infty} \mathbb{E}\{(I_n)^2\} = 0.$$

However, concerning the strong convergence of the approximation error, we have a negative finding:

Proposition 8.2. (Györfi, Sanchetta (2014)) *Put $k_n = (\ln n)^{1-\delta}$ with $0 < \delta < 1$. Then for the MA(1) process defined by*

$$Y_n = Z_n - Z_{n-1}, \tag{8.26}$$

where the innovations $\{Z_n\}$ are i.i.d. standard Gaussian, the approximation error does not converge to zero a.s.

Under some condition on the covariances $r(j), j = 1, 2, \ldots$, one may get positive result on the approximation error.

Proposition 8.3. (Györfi, Sanchetta (2014)) *Assume that for all $n > k$,*

$$\sum_{j=k+1}^{n-1} c_j^{(n-1)} r(j) \le C_1 k^{-\gamma},$$

and

$$\sum_{j=1}^{k} (c_j^{(n-1)} - c_j^{(k)}) r(j) \le C_2 k^{-\gamma}, \tag{8.27}$$

with $\gamma > 0$. If

$$k_n = (\ln n)^{(1+\delta)/\gamma} \tag{8.28}$$

($\delta > 0$), then for the approximation error, we have that

$$I_n = \mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \mathbb{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} \to 0$$

a.s.

The partial autocorrelation function of Y_n is $\alpha(j) := c_j^{(j)}$ where $c_j^{(j)}$ is as defined before, i.e. the j^{th} coefficient from the AR(j) approximation of Y_n. It is possible to explicitly bound the approximation error I_n using $\alpha(j)$.

Proposition 8.4. (Györfi, Sanchetta (2014)) *Suppose that*

$$\sum_{j=k+1}^{\infty} \alpha^2(j) \leq ck^{-\gamma}$$

with $\gamma > 0$, $c > 0$. For the choice (8.28), we have that

$$I_n = \mathbb{E}\{Y_n \mid Y_1^{n-1}\} - \mathbb{E}\left\{Y_n | Y_{n-k_n}^{n-1}\right\} \to 0$$

a.s.

The estimation error is even more interesting. Under some conditions on the covariances, Györfi and Sanchetta (2014) had some positive results on the strong convergence of the estimation error. However, they guessed the following:

Conjecture 8.2. *There is a stationary, ergodic Gaussian process and a fixed k such that the estimation error*

$$J_n = \sum_{j=1}^{k} (c_j^{(k)} - C_{n,j}^{(k)}) Y_{n-j}$$

does not converge to 0 a.s.

Chapter 9

Estimation and prediction for pinball loss

9.1 The absolute loss

In the previous chapters we studied the squared loss

$$\ell(\hat{y}, y) = (y - \hat{y})^2. \tag{9.1}$$

In some applications the squared loss is too sensitive for large error values, therefore we introduce the absolute loss:

$$\ell(\hat{y}, y) = |y - \hat{y}|. \tag{9.2}$$

The absolute loss or the related l_1 norm became an important quantity for high-dimensional statistics and for compressed sensing, see Bühlmann and van de Geer (2011), Candés, Romberg and Tao (2006), Donoho (2006), Elad (2010), Eldar and Kutyniok (2012). Another application of the l_1 norm is in the photogrammetry, see Förstner, Wrobel (2016), Kraus (2007), Luhmann, et al. (2013).

Similarly to the previous chapters, we are given a random vector $(\mathbf{X}, Y)$, where the observation vector $\mathbf{X}$ takes values in $\mathbb{R}^d$, and the label Y is real valued. For absolute loss, we search for the function $f : \mathbb{R}^d \to \mathbb{R}$ such that $|f(\mathbf{X}) - Y|$ is "small", i.e., let the L_1 *error* or the *mean absolute error*

$$\mathbb{E}\{|f(\mathbf{X}) - Y|\}$$

take the smallest value. It means that we want to construct the function $r^* : \mathbb{R}^d \to \mathbb{R}$, for which

$$\mathbb{E}\{|r^*(\mathbf{X}) - Y|\} = \min_{f:\mathbb{R}^d \to \mathbb{R}} \mathbb{E}\{|f(\mathbf{X}) - Y|\}. \tag{9.3}$$

For the squared loss, we applied the Steiner theorem for conditional distributions and so got the optimal function, called regression function. For absolute loss, the problem is a bit more involved.

Consider the following optimization task:

$$\mathbb{E}\{|y_{opt} - Y|\} = \min_{y \in \mathbb{R}} \mathbb{E}\{|y - Y|\}!$$

Because of the absolute value, the expression

$$\mathbb{E}\{|y - Y|\}$$

is not differentiable at each point y. However, it is convex, therefore the right and left derivatives exist. At the point y_{opt} the right derivative is non-negative, while the left derivative is non-positive.

If the function is differentiable at y, then calculating formally the derivative, one can guess y_{opt}.

$$\begin{aligned}
\frac{d}{dy}\mathbb{E}\{|y - Y|\} &= \mathbb{E}\left\{\frac{d}{dy}|y - Y|\right\} \\
&= \mathbb{E}\left\{sign(y - Y)\right\} \\
&= \mathbb{P}\left\{Y \leq y\right\} - \mathbb{P}\left\{Y > y\right\} \\
&= 2(\mathbb{P}\left\{Y \leq y\right\} - 1/2).
\end{aligned}$$

Thus y_{opt} is the solution of the equation

$$\mathbb{P}\left\{Y \leq y\right\} = 1/2,$$

i.e., y_{opt} is the median of the random variable Y. Let

$$F(y) = \mathbb{P}\left\{Y \leq y\right\}$$

be the distribution function of the random variable Y. Then

$$y_{opt} = F^{-1}(1/2),$$

which is the median of the random variable Y. A formal proof of the optimality can be found in Stroock (2011).

Here we have two problems: one the one hand the inverse does not exist, in general, on the other hand it may not be unique. However, one can define the inverse uniquely:

$$F^{-1}(u) = \max\{y; F(y) \leq u\}. \tag{9.4}$$

Now, consider the task (9.3). Introduce the conditional distribution function

$$F(y \mid \mathbf{x}) = \mathbb{P}\left\{Y \leq y \mid \mathbf{X} = \mathbf{x}\right\}. \tag{9.5}$$

Because of

$$\mathbb{E}\{|f(\mathbf{X}) - Y|\} = \mathbb{E}\{\mathbb{E}\{|f(\mathbf{X}) - Y| \mid \mathbf{X}\}\} = \int \mathbb{E}\{|f(\mathbf{x}) - Y| \mid \mathbf{X} = \mathbf{x}\}\mu(d\mathbf{x}),$$

we have that

$$r^*(\mathbf{x}) = F^{-1}(1/2 \mid \mathbf{x}), \tag{9.6}$$

which is the conditional median of the random variable Y, given $\mathbf{X} = \mathbf{x}$.

9.2 The pinball loss

Before studying the estimation of the function $r^*(\mathbf{x})$, we generalize the concept of absolute loss. In many real life applications the loss is different depending, whether or not $f(\mathbf{X})$ overestimates Y. For example, if Y is a future price of an asset, then overestimating Y the loss is much larger than for underestimating. This type of costs shows up for inventory problems, see Toomey (2000), Prékopa (2006). The other important practical example is the prediction of electricity consumption, cf. Abu-Shikhah, Elkarmi and Aloquili (2011), Alfares and Nazeeruddin (2002), Almeshaiei and Soltan (2011), Aung et al. (2012), Ba et al. (2012), Bozic, Stojanovic and Stajic (2010), Bruhns, Deurveilher and Roy (2005), Cancelo, Espasa and Grafe (2008), Devaine et al. (2013), Dordonnat et al. (2008), Elattar, Goulermas and Wu (2010), Feinberg and Genethliou (2005), Gaillard and Goude (2011), Misiti et al. (2010), Nagi et al. (2008), Pierrot and Goude (2011), Sevlian and Rajagopal (2018), Taylor and McSharry (2008).

Put $0 < \tau < 1$. If x^+ denotes the positive part of x, then the generalization of the absolute loss is defined by

$$\ell_\tau(\hat{y}, y) = \tau(y - \hat{y})^+ + (1 - \tau)(\hat{y} - y)^+. \tag{9.7}$$

This loss is called τ-pinball loss, see Steinwart and Christmann (2011). Our optimization task is as follows:

$$\begin{aligned} &\tau\mathbb{E}\{(Y - r^*(\mathbf{X}))^+\} + (1 - \tau)\mathbb{E}\{(r^*(\mathbf{X}) - Y)^+\} \\ &= \min_{f:\mathbb{R}^d \to \mathbb{R}} [\tau\mathbb{E}\{(Y - f(\mathbf{X}))^+\} + (1 - \tau)\mathbb{E}\{(f(\mathbf{X}) - Y)^+\}]. \end{aligned} \tag{9.8}$$

For the notation

$$C_+ = \frac{\tau}{(1-\tau)},$$

(9.8) is equivalent to

$$\begin{aligned} & C_+\mathbb{E}\{(Y - r^*(\mathbf{X}))^+\} + \mathbb{E}\{(r^*(\mathbf{X}) - Y)^+\} \\ & = \min_{f:\mathbb{R}^d \to \mathbb{R}} [C_+\mathbb{E}\{(Y - f(\mathbf{X}))^+\} + \mathbb{E}\{(f(\mathbf{X}) - Y)^+\}]. \end{aligned} \tag{9.9}$$

Similarly to the previous section, start with a special optimization problem:

$$C_+\mathbb{E}\{(Y - y_{opt})^+\} + \mathbb{E}\{(y_{opt} - Y)^+\} = \min_{y \in \mathbb{R}} [C_+\mathbb{E}\{(Y - y)^+\} + \mathbb{E}\{(y - Y)^+\}].$$

Again search for y_{opt} calculating the derivative:

$$\begin{aligned} \frac{d}{dy}[C_+\mathbb{E}\{(Y - y)^+\} + \mathbb{E}\{(y - Y)^+\}] &= C_+\mathbb{E}\left\{\frac{d}{dy}(Y - y)^+\right\} + \mathbb{E}\left\{\frac{d}{dy}(y - Y)^+\right\} \\ &= -C_+\mathbb{E}\left\{\mathbb{I}_{Y>y}\right\} + \mathbb{E}\left\{\mathbb{I}_{Y \le y}\right\} \\ &= -C_+\mathbb{P}\left\{Y > y\right\} + \mathbb{P}\left\{Y \le y\right\} \\ &= (1 + C_+)\mathbb{P}\left\{Y \le y\right\} - C_+. \end{aligned}$$

Therefore

$$y_{opt} = F^{-1}\left(\frac{C_+}{1 + C_+}\right) = F^{-1}\left(\tau\right),$$

which is the quantile of the random variable Y at level τ. Concerning a formal proof of optimality, see Biau and Patra (2011).

It implies the solution of (9.8):

$$r^*(\mathbf{x}) = F^{-1}\left(\tau \mid \mathbf{x}\right) \tag{9.10}$$

and so $r^*(\mathbf{x})$ is the conditional quantile of the random variable Y at level τ, given $\mathbf{X} = \mathbf{x}$. The function $r^*(\mathbf{x})$ is called quantile regression function.

9.3 Estimates of quantile regression function

In an application the distribution of $(\mathbf{X}, Y)$ is unknown and so we cannot calculate the quantile regression function. Assume we are given data

$$D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n), \}$$

which is a sequence of i.i.d. pairs of random variables.

From the definition (9.5) it is obvious, that for any fixed y, the function $F(y \mid \mathbf{x})$ is a regression function, which can be estimated from the data D_n. Let $F_n(y \mid \mathbf{x})$ an arbitrary regression estimate depending on $\mathbf{x}$ and D_n. From $F_n(y \mid \mathbf{x})$ we derive a quantile regression estimate:

$$r_n(\mathbf{x}) = F_n^{-1}(\tau \mid \mathbf{x}).$$

(See Bhattacharya and Gangopadhyay (1990), Caner (2002), Chaudhuri (1991), Dette and Volgushev (2008), Hall and Müller (2003), Koenker (2005), Lejeune and Sarda (1988), Stone (1977).)

For a sequence of real numbers $z_1, ..., z_N$ let

$$Q_\tau(z_1, ..., z_N)$$

be the quantile of the sequence $z_1, ..., z_N$ at level τ.

For the partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots\}$, we slightly modify the partitioning estimate defined in Chapter 2 such that

$$F_n(y \mid \mathbf{x}) = \begin{cases} \frac{\sum_{i=1}^n \mathbb{I}_{\{Y_i \le y\}} \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}}, & \text{if} \quad \sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}} > 0, \\ \mathbb{I}_{\{\frac{1}{n}\sum_{i=1}^n Y_i \le y\}} & \text{otherwise,} \end{cases}$$

where $A_n(\mathbf{x})$ denotes the $A_{n,j}$ of the partition $\mathcal{P}_n$ into which $\mathbf{x}$ falls. If $\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}} > 0$, then let $Y_{n,1}(\mathbf{x}), \ldots, Y_{n,N}(\mathbf{x})$ be the subsequence of $Y_1, \ldots, Y_n$, for which $\mathbf{X}_i \in A_n(\mathbf{x})$. (Here $N = \sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}}$.) Then the partitioning based quantile estimate is given by

$$r_n(\mathbf{x}) = \begin{cases} Q_\tau(Y_{n,1}(\mathbf{x}), \ldots, Y_{n,N}(\mathbf{x})), & \text{if} \quad \sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in A_n(\mathbf{x})\}} > 0, \\ Q_\tau(Y_1, \ldots, Y_n) & \text{otherwise.} \end{cases} \tag{9.11}$$

Because of computational complexity, we introduce the kernel based quantile regression estimate only in the special case of naive kernel. Put

$$F_n(y \mid \mathbf{x}) = \begin{cases} \frac{\sum_{i=1}^n \mathbb{I}_{\{Y_i \le y\}} \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}{\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}}, & \text{if} \quad \sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}} > 0, \\ \mathbb{I}_{\{\frac{1}{n}\sum_{i=1}^n Y_i \le y\}} & \text{otherwise.} \end{cases}$$

If $\sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}} > 0$, then let $Y_{n,1}(\mathbf{x}), \ldots, Y_{n,N}(\mathbf{x})$ be the subsequence of $Y_1, \ldots, Y_n$, for which $\mathbf{X}_i \in S_{\mathbf{x},h_n}$. (Here $N = \sum_{i=1}^n \mathbb{I}_{\{\mathbf{X}_i \in S_{\mathbf{x},h_n}\}}$.) Then with these notations, the kernel based quantile regression estimate $r_n(\mathbf{x})$ is defined by (9.11).

Using the notations of Chapter 4 we define k_n nearest neighbor quantile regression estimate. Put

$$F_n(y \mid \mathbf{x}) = \frac{1}{k_n} \sum_{i=1}^{k_n} \mathbb{I}_{\{Y_{(i,n)}(\mathbf{x}) \le y\}}.$$

Then nearest neighbor quantile regression estimate is

$$r_n(\mathbf{x}) = Q_\tau(Y_{(1,n)}(\mathbf{x}), \dots, Y_{(k_n,n)}(\mathbf{x})).$$

9.4 Aggregation of finitely many elementary predictors

Introduce the notations

$$\ell_\tau(\hat{y}_t, y_t) = \tau \mathbb{E}\{(y_t - \hat{y}_t))^+\} + (1-\tau)\mathbb{E}\{(\hat{y}_t - y_t)^+\} \tag{9.12}$$

and

$$L_n(g) = \frac{1}{n} \sum_{t=1}^{n} \ell_\tau(g_t(\mathbf{x}_1^t, y_1^{t-1}), y_t).$$

Theorem 9.1. (Cesa-Bianchi, Lugosi (2006)) *Let $\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_K$ be a sequence of predictions. Assume that the cost function ℓ is convex in the first argument, and*

$$0 \le \ell_\tau(\tilde{h}_k(\mathbf{x}_1^n, y_1^{n-1}), y_n) \le B.$$

Introduce the weights

$$w_{t,k} = \frac{1}{K} e^{-\eta(t-1)L_{t-1}(\tilde{h}_k)}$$

and their normalizations

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^K w_{t,i}}.$$

If the aggregated prediction $\tilde{g}$ is defined by

$$\tilde{g}_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k=1}^{K} v_{t,k} \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1}) \qquad t = 1, 2, \dots,$$

then for each $n \ge 1$,

$$L_n(\tilde{g}) \le \min_{1 \le k \le K} L_n(\tilde{h}_k) + \frac{\ln K}{n\eta} + \frac{\eta B^2}{8}.$$

Thus, for the choice

$$\eta = \frac{\sqrt{8}}{B}\sqrt{\frac{\ln K}{n}},$$

we get

$$L_n(\tilde{g}) \le \min_{1\le k\le K} L_n(\tilde{h}_k) + \sqrt{\frac{2\ln K}{n}}B.$$

PROOF. For the proof of this theorem we apply the Hoeffding (1963) lemma. If for a random variable Z, $a \le Z \le b$, then for any real number s, one gets

$$\mathbb{E}\left\{\exp\left(s\cdot(Z-\mathbb{E}Z)\right)\right\} \le \exp\left(\frac{s^2(b-a)^2}{8}\right). \tag{9.13}$$

Put

$$Y = Z - \mathbb{E}Z.$$

Then $Y \in [a-\mathbb{E}Z, b-\mathbb{E}Z] =: [a', b']$, $a'-b' = a-b$, and $\mathbb{E}Y = 0$. We show, that

$$\mathbb{E}\left\{\exp(sY)\right\} \le \exp\left(\frac{s^2(b-a)^2}{8}\right). \tag{9.14}$$

Because of the convexity of e^{sx},

$$e^{sx} \le \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{ha } a \le x \le b,$$

therefore

$$\begin{aligned}\mathbb{E}\{\exp(sY)\} &\le \frac{\mathbb{E}\{Y\}-a}{b-a}e^{sb} + \frac{b-\mathbb{E}\{Y\}}{b-a}e^{sa}\\ &= e^{sa}\left(1+\frac{a}{b-a}-\frac{a}{b-a}e^{s(b-a)}\right)\\ &\quad (\text{ since } \mathbb{E}\{Y\}=0).\end{aligned}$$

Put

$$p = -\frac{a}{b-a},$$

then

$$\mathbb{E}\{\exp(sY)\} \le (1-p+p\cdot e^{s(b-a)})e^{-s\,p\,(b-a)} = e^{\Phi(s(b-a))},$$

where

$$\Phi(u) = \ln\left((1-p+pe^u)e^{-pu}\right) = \ln\left(1-p+pe^u\right) - pu.$$

Calculate the Taylor series of Φ! Because of

$$\Phi(0) = 0,$$

$$\Phi'(u) = \frac{pe^u}{1-p+pe^u} - p, \text{ therefore } \Phi'(0) = 0$$

and

$$\Phi''(u) = \frac{(1-p+pe^u)pe^u - pe^u pe^u}{(1-p+pe^u)^2} = \frac{(1-p)pe^u}{(1-p+pe^u)^2}$$
$$\leq \frac{(1-p)pe^u}{4(1-p)pe^u} = \frac{1}{4}.$$

Thus, for any $u > 0$,

$$\Phi(u) = \Phi(0) + \Phi'(0)u + \frac{1}{2}\Phi''(\eta)u^2 \leq \frac{1}{8}u^2$$

with some $\eta \in [0, u]$. We get, that

$$\mathbb{E}\{\exp(sY)\} \leq e^{\Phi(s(b-a))} \leq \exp\left(\frac{1}{8}s^2(b-a)^2\right),$$

which proves (9.14). The Hoeffding lemma implies, that

$$\ln \mathbb{E}\left\{\exp\left(s \cdot Z\right)\right\} \leq s\mathbb{E}\left\{Z\right\} + \frac{s^2(b-a)^2}{8}. \tag{9.15}$$

Put

$$W_1 = 1$$

and

$$W_t = \sum_{k=1}^{K} w_{t,k}$$

for $t > 1$. Then on the one hand

$$\ln \frac{W_n}{W_1} = \ln\left(\sum_{k=1}^{K} e^{-\eta(n-1)L_{n-1}(\tilde{h}_k)}\right) - \ln K$$
$$\geq \ln\left(\max_{1\leq k\leq K} e^{-\eta(n-1)L_{n-1}(\tilde{h}_k)}\right) - \ln K$$
$$= -\eta(n-1)\min_{1\leq k\leq K} L_{n-1}(\tilde{h}_k) - \ln K,$$

and on the other hand (9.15) implies that

$$\begin{aligned}
\ln \frac{W_{t+1}}{W_t} &= \ln \frac{\sum_{k=1}^K e^{-\eta t L_t(\tilde h_k)}}{\sum_{k=1}^K e^{-\eta (t-1) L_{t-1}(\tilde h_k)}} \\
&= \ln \frac{\sum_{k=1}^K e^{-\eta (t-1) L_{t-1}(\tilde h_k)} e^{-\eta \ell_\tau(\tilde h_{k,t}(\mathbf{x}_1^t, y_1^{t-1}), y_t)}}{\sum_{k=1}^K e^{-\eta (t-1) L_{t-1}(\tilde h_k)}} \\
&= \ln \frac{\sum_{k=1}^K w_{t,k} e^{-\eta \ell_\tau(\tilde h_{k,t}(\mathbf{x}_1^t, y_1^{t-1}), y_t)}}{\sum_{k=1}^K w_{t,k}} \\
&\le -\eta \frac{\sum_{k=1}^K w_{t,k} \ell_\tau(\tilde h_{k,t}(\mathbf{x}_1^t, y_1^{t-1}), y_t)}{\sum_{k=1}^K w_{t,k}} + \frac{\eta^2 B^2}{8} \\
&\le -\eta \ell_\tau \left(\frac{\sum_{k=1}^K w_{t,k} \tilde h_{k,t}(\mathbf{x}_1^t, y_1^{t-1})}{\sum_{k=1}^K w_{t,k}}, y_t \right) + \frac{\eta^2 B^2}{8} \\
&= -\eta \ell_\tau \left(\tilde g_t(\mathbf{x}_1^t, y_1^{t-1}), y_t \right) + \frac{\eta^2 B^2}{8},
\end{aligned}$$

where the second inequality follows from the Jensen inequality, because we assume, that the cost function ℓ_τ is convex in the first argument. Thus,

$$\begin{aligned}
\ln \frac{W_n}{W_1} &= \sum_{t=1}^{n-1} \ln \frac{W_{t+1}}{W_t} \\
&\le -\eta \sum_{t=1}^{n-1} \ell_\tau \left(\tilde g_t(\mathbf{x}_1^t, y_1^{t-1}), y_t \right) + (n-1) \frac{\eta^2 B^2}{8} \\
&= -\eta (n-1) L_{n-1}(\tilde g) + (n-1) \frac{\eta^2 B^2}{8}.
\end{aligned}$$

Combining the two inequalities for $\ln \frac{W_n}{W_1}$, we get that

$$-\eta (n-1) \min_{1 \le k \le K} L_{n-1}(\tilde h_k) - \ln K \le -\eta (n-1) L_{n-1}(\tilde g) + (n-1) \frac{\eta^2 B^2}{8},$$

and the theorem is proved. □

The Theorem 9.1 holds for any sequence $\{\mathbf{x}_n, y_n\}$. In the theory of machine learning one says, that for any individual sequence we have a worst case inequality. For particular cases this inequality can be improved, while there are results for the good choice of η.

9.5 Prediction of time series for pinball loss

With the notation of Section 8.2.1, put

$$J_n^{(k,\ell)} = \left\{ k < t < n : G_\ell(\mathbf{x}_{t-k}^t) = G_\ell(\mathbf{x}_{n-k}^n), F_\ell(y_{t-k}^{t-1}) = F_\ell(y_{n-k}^{n-1}) \right\}.$$

Define the elementary predictor $h^{(k,\ell)}$ by

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \begin{cases} Q_\tau(y_i, i \in J_n^{(k,\ell)}), & \text{if} \quad J_n^{(k,\ell)} \neq \emptyset, \\ Q_\tau(y_1, \ldots, y_{n-1}) & \text{otherwise.} \end{cases}$$

According to (8.13), (8.14), (8.15) and (8.16), aggregate the elementary predictors, which is called as partitioning based predictor for pinball loss. Then the expert lemma for unbounded y_is (Lemma 8.3) implies, that the partitioning based predictor universally consistent for the class of stationary and ergodic time series with $\mathbb{E}\{Y^2\} < \infty$ and for pinball loss.

The kernel based predictor is defined as before such that $J_n^{(k,\ell)}$ was introduced in Section 3, and the same consistency result can be formulated as for partitioning based predictor.

For the nearest neighbor predictor, $J_n^{(k,\ell)}$ is defined in Section 8.2.3, and its consistency was proved by Biau and Patra (2011).

Chapter 10

Prediction of time series for $0-1$ loss

10.1 Bayes decision

For the statistical inference, a d-dimensional observation vector $\mathbf{X}$ is given, and based on $\mathbf{X}$, the statistician has to make an inference on a random variable Y, which takes finitely many values, i.e., it takes values from the set $\{1, 2, \ldots, M\}$. In fact, the inference is a decision formulated by a decision function

$$g : \mathbb{R}^d \to \{1, 2, \ldots, M\}.$$

If $g(\mathbf{X}) \neq Y$ then the decision makes error.

In the formulation of the Bayes decision problem, introduce a cost function $C(y, y') \geq 0$, which is the cost if the label $Y = y$ and the decision $g(\mathbf{X}) = y'$. For a decision function g, the risk is the expectation of the cost:

$$R(g) = \mathbb{E}\{C(Y, g(\mathbf{X}))\}.$$

In Bayes decision problem, the aim is to minimize the risk, i.e., the goal is to find a function $g^* : \mathbb{R}^d \to \{1, 2, \ldots, M\}$ such that

$$R(g^*) = \min_{g:\mathbb{R}^d \to \{1,2,\ldots,M\}} R(g), \tag{10.1}$$

where g^* is called the Bayes decision function, and $R^* = R(g^*)$ is the Bayes risk.

For the posteriori probabilities, introduce the notations:

$$P_y(\mathbf{X}) = \mathbb{P}\{Y = y \mid \mathbf{X}\}.$$

Let the decision function g^* be defined by

$$g^*(\mathbf{X}) = \arg\min_{y'} \sum_{y=1}^{M} C(y, y') P_y(\mathbf{X}).$$

If arg min is not unique then choose the smallest y', which minimizes $\sum_{y=1}^{m} C(y, y') P_y(\mathbf{X})$. This definition implies that for any decision function g,

$$\sum_{y=1}^{m} C(y, g^*(\mathbf{X})) P_y(\mathbf{X}) \leq \sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X}). \tag{10.2}$$

Theorem 10.1. *For any decision function g, we have that*

$$R(g^*) \leq R(g).$$

PROOF. For a decision function g, let's calculate the risk.

$$\begin{aligned}
R(g) &= \mathbb{E}\{C(Y, g(\mathbf{X}))\} \\
&= \mathbb{E}\{\mathbb{E}\{C(Y, g(\mathbf{X})) \mid \mathbf{X}\}\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{m}\sum_{y'=1}^{M} C(y, y') \mathbb{P}\{Y = y, g(\mathbf{X}) = y' \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{m}\sum_{y'=1}^{M} C(y, y') \mathbb{I}_{\{g(\mathbf{X})=y'\}} \mathbb{P}\{Y = y \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X})\right\}.
\end{aligned}$$

(10.2) implies that

$$\begin{aligned}
R(g) &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X})\right\} \\
&\geq \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X})) P_y(\mathbf{X})\right\} \\
&= R(g^*).
\end{aligned}$$

□

Concerning the cost function, the most frequently studied example is the so called $0-1$ loss:

$$C(y,y') = \begin{cases} 1 & \text{if} \quad y \neq y', \\ 0 & \text{if} \quad y = y'. \end{cases}$$

For the $0-1$ loss, the corresponding risk is the error probability:

$$R(g) = \mathbb{E}\{C(Y, g(\mathbf{X}))\} = \mathbb{E}\{\mathbb{I}_{\{Y \neq g(\mathbf{X})\}}\} = \mathbb{P}\{Y \neq g(\mathbf{X})\},$$

and the Bayes decision is of form

$$g^*(\mathbf{X}) = \arg\min_{y'} \sum_{y=1}^{M} C(y,y') P_y(\mathbf{X}) = \arg\min_{y'} \sum_{y \neq y'} P_y(\mathbf{X}) = \arg\max_{y'} P_{y'}(\mathbf{X}),$$

which is called maximum posteriori decision, too.

If the distribution of the observation vector $\mathbf{X}$ has density, then the Bayes decision has an equivalent formulation. Introduce the notations for density of $\mathbf{X}$ by

$$\mathbb{P}\{\mathbf{X} \in B\} = \int_B f(\mathbf{x}) d\mathbf{x}$$

and for the conditional densities by

$$\mathbb{P}\{\mathbf{X} \in B \mid Y = y\} = \int_B f_y(\mathbf{x}) d\mathbf{x}$$

and for a priori probabilities

$$q_y = \mathbb{P}\{Y = y\},$$

then it is easy to check that

$$P_y(\mathbf{X}) = \mathbb{P}\{Y = y \mid \mathbf{X} = \mathbf{x}\} = \frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})}$$

and therefore

$$\begin{aligned} g^*(\mathbf{x}) &= \arg\min_{y'} \sum_{y=1}^{M} C(y,y') P_y(\mathbf{x}) \\ &= \arg\min_{y'} \sum_{y=1}^{M} C(y,y') \frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})} \\ &= \arg\min_{y'} \sum_{y=1}^{M} C(y,y') q_y f_y(\mathbf{x}). \end{aligned}$$

From the proof of Theorem 10.1 we may derive a formula for the optimal risk:

$$R(g^*) = \mathbb{E}\left\{\min_{y'} \sum_{y=1}^{M} C(y, y') P_y(\mathbf{X})\right\}.$$

If $\mathbf{X}$ has density then

$$\begin{aligned} R(g^*) &= \mathbb{E}\left\{\min_{y'} \sum_{y=1}^{M} C(y, y') \frac{q_y f_y(\mathbf{X})}{f(\mathbf{X})}\right\} \\ &= \int_{\mathbb{R}^d} \min_{y'} \sum_{y=1}^{M} C(y, y') \frac{q_y f_y(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \min_{y'} \sum_{y=1}^{M} C(y, y') q_y f_y(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

For the $0-1$ loss, we get that

$$R(g^*) = \mathbb{E}\left\{\min_{y'} (1 - P_{y'}(\mathbf{X}))\right\},$$

which has the form, for densities,

$$R(g^*) = \int_{\mathbb{R}^d} \min_{y'} (f(\mathbf{x}) - q_{y'} f_{y'}(\mathbf{x})) d\mathbf{x} = 1 - \int_{\mathbb{R}^d} \max_{y'} q_{y'} f_{y'}(\mathbf{x}) d\mathbf{x}.$$

For $M = 2$, we have that

$$R(g^*) = \mathbb{E}\left\{\min(P_1(\mathbf{X}), P_2(\mathbf{X}))\right\},$$

and, for densities,

$$R(g^*) = \int_{\mathbb{R}^d} \min(q_1 f_1(\mathbf{x}), q_2 f_2(\mathbf{x})) d\mathbf{x}.$$

Figure 10.1 illustrates the Bayes decision, while the red area in Figure 10.2 is equal to the Bayes error probability.

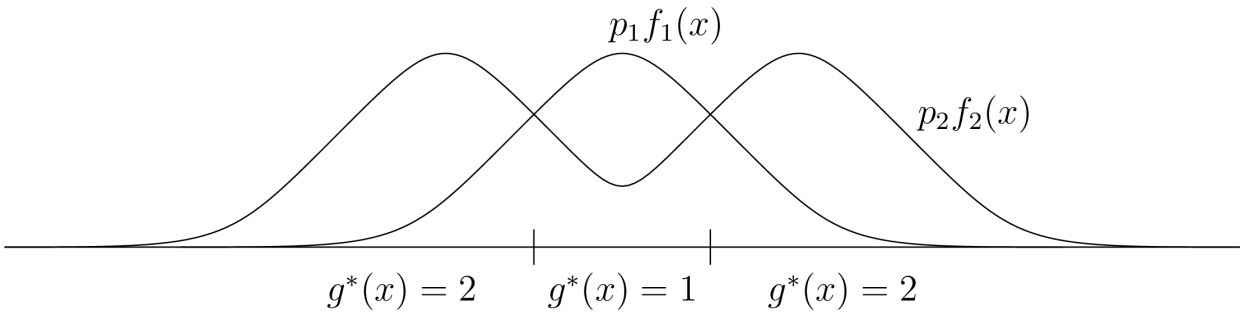

Figure 10.1: Bayes decision.

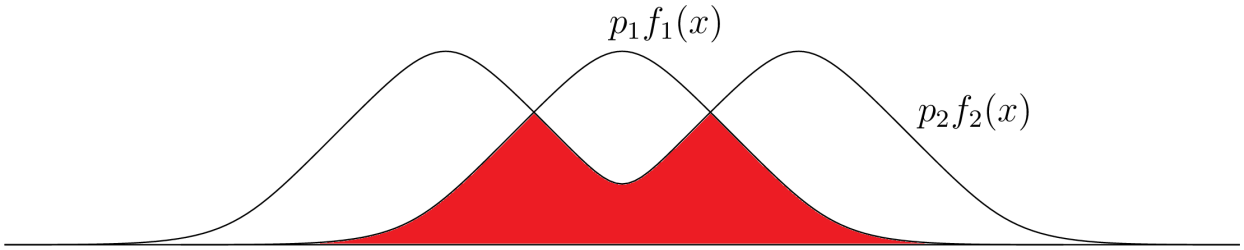

Figure 10.2: Bayes error probability.

10.2 Approximation of Bayes decision

In practice, the posteriori probabilities $\{P_y(\mathbf{X})\}$ are unknown. If we are given some approximations $\{\hat{P}_y(\mathbf{X})\}$, from which one may derive some approximate decision

$$\hat{g}(\mathbf{X}) = \arg\min_{y'} \sum_{y=1}^{M} C(y, y') \hat{P}_y(\mathbf{X})$$

then the question is how well $R(\hat{g})$ approximates R^*.

Lemma 10.1. *Put* $C_{max} = \max_{y,y'} C(y, y')$, *then*

$$0 \leq R(\hat{g}) - R(g^*) \leq 2C_{max} \sum_{y=1}^{M} \mathbb{E}\left\{ |P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})| \right\}.$$

PROOF. We have that

$$\begin{aligned} R(\hat{g}) - R(g^*) &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))P_y(\mathbf{X})\right\} - \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X}))P_y(\mathbf{X})\right\} \\ &= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))P_y(\mathbf{X}) - \sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X})\right\} \\ &\quad +\mathbb{E}\left\{\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y, g^*(\mathbf{X}))\hat{P}_y(\mathbf{X})\right\} \\ &\quad +\mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y, g^*(\mathbf{X}))P_y(\mathbf{X})\right\}. \end{aligned}$$

The definition of $\hat{g}$ implies that

$$\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))\hat{P}_y(\mathbf{X}) - \sum_{y=1}^{M} C(y, g^*(\mathbf{X}))\hat{P}_y(\mathbf{X}) \leq 0,$$

therefore

$$\begin{aligned} R(\hat{g}) - R(g^*) &\leq \mathbb{E}\left\{\sum_{y=1}^{M} C(y, \hat{g}(\mathbf{X}))|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\} \\ &\quad +\mathbb{E}\left\{\sum_{y=1}^{M} C(y, g^*(\mathbf{X}))|\hat{P}_y(\mathbf{X}) - P_y(\mathbf{X})|\right\} \\ &\leq 2C_{max}\sum_{y=1}^{M} \mathbb{E}\left\{|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\}. \end{aligned}$$

□

In the special case of the approximate maximum posteriori decision the inequality in Lemma 10.1 can be slightly improved:

$$0 \leq R(\hat{g}) - R(g^*) \leq \sum_{y=1}^{M} \mathbb{E}\left\{|P_y(\mathbf{X}) - \hat{P}_y(\mathbf{X})|\right\}.$$

Based on this relation, one can introduce efficient pattern recognition rules. The a posteriori probabilities are the regression functions

$$\mathbb{P}\{Y = y|\mathbf{X} = \mathbf{x}\} = \mathbb{E}\{\mathbb{I}_{\{Y=y\}}|\mathbf{X} = \mathbf{x}\} = m^{(y)}(\mathbf{x}).$$

Given data $D_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, estimates $m_n^{(y)}$ of $m^{(y)}$ can be constructed from the data set

$$D_n^{(y)} = \{(\mathbf{X}_1, \mathbb{I}_{\{Y_1=y\}}), \ldots, (\mathbf{X}_n, \mathbb{I}_{\{Y_n=y\}})\},$$

and one can use a plug-in estimate

$$g_n(\mathbf{x}) = \arg \max_{1 \le y \le M} m_n^{(y)}(\mathbf{x}) \tag{10.3}$$

to estimate g^*. If the estimates $m_n^{(y)}$ are close to the a posteriori probabilities, then again the error of the plug-in estimate is close to the optimal error. (For the details, see Devroye, Györfi, and Lugosi (1996).)

10.3 Pattern recognition for time series

In this section we apply the ideas of Chapter 8 to the seemingly more difficult pattern recognition problem for time series. The setup is the following: let $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$ be a stationary and ergodic sequence of pairs taking values in $\mathbb{R}^d \times \{0, 1\}$. The problem is to predict the value of Y_n given the data $(\mathbf{X}_1^n, Y_1^{n-1})$.

We may formalize the prediction (classification) problem as follows. The strategy of the classifier is a sequence $f = \{f_t\}_{t=1}^{\infty}$ of decision functions

$$f_t : \left(\mathbb{R}^d\right)^t \times \{0, 1\}^{t-1} \to \{0, 1\}$$

so that the classification formed at time t is $f_t(\mathbf{X}_1^t, Y_1^{t-1})$. The *normalized cumulative* $0 - 1$ *loss* for any fixed pair of sequences $\mathbf{X}_1^n,\ Y_1^n$ is now

$$R_n(f) = \frac{1}{n} \sum_{t=1}^{n} I_{\{f_t(\mathbf{X}_1^t, Y_1^{t-1}) \neq Y_t\}}.$$

In this case there is a fundamental limit for the predictability of the sequence, i.e., Algoet (1994) proved that for any classification strategy f and stationary ergodic process $\{(\mathbf{X}_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n \to \infty} R_n(f) \ge R^* \quad \text{a.s.}, \tag{10.4}$$

where

$$R^* = \mathbb{E}\left\{\min\left(\mathbb{P}\{Y_0 = 1|\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\}, \mathbb{P}\{Y_0 = 0|\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\}\right)\right\},$$

therefore the following definition is meaningful:

Definition 10.1. *A classification strategy f is called* universally consistent *if for all stationary and ergodic processes $\{\mathbf{X}_n, Y_n\}_{-\infty}^{\infty}$,*

$$\lim_{n\to\infty} R_n(f) = R^* \quad \textit{almost surely.}$$

Therefore, universally consistent strategies asymptotically achieve the best possible loss for all ergodic processes. We present a simple (non-randomized) on-line classification strategy, and prove its universal consistency. Consider the prediction scheme $g_t(\mathbf{X}_1^t, Y_1^{t-1})$ introduced in Sections 8.2.1 or 8.2.2 or 8.2.3 or 8.2.4, and then introduce the corresponding classification scheme:

$$f_t(\mathbf{X}_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } g_t(\mathbf{X}_1^t, Y_1^{t-1}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

The main result of this section is the universal consistency of this simple classification scheme:

Theorem 10.2. (Györfi and Ottucsák (2007)) *Assume that the conditions of Theorems 8.1 or 8.2 or 8.3 or 8.4 are satisfied. Then the classification scheme f defined above satisfies*

$$\lim_{n\to\infty} R_n(f) = R^* \quad \textit{almost surely}$$

for any stationary and ergodic process $\{(\mathbf{X}_n, Y_n)\}_{n=-\infty}^{\infty}$.

In order to prove Theorem 10.2 we derive a corollary of Theorem 8.1, which shows that asymptotically, the predictor g_t defined by (8.5) predicts as well as the optimal predictor given by the regression function $\mathbb{E}\{Y_t|Y_{-\infty}^{t-1}\}$. In fact, g_t gives a good estimate of the regression function in the following (Cesáro) sense:

Corollary 10.1. *Under the conditions of Theorem 8.1*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)^2 = 0 \qquad \textit{almost surely.}$$

PROOF. By Theorem 8.1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)^2 = L^* \qquad \text{almost surely.}$$

Consider the following decomposition:

$$\begin{aligned} & \left(Y_i - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)^2 \\ = \; & \left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)^2 \\ & +2\left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right) \\ & + \left(\mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)^2. \end{aligned}$$

Then the ergodic theorem implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)^2 = L^* \qquad \text{almost surely.}$$

It remains to show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right) = 0. \tag{10.5}$$

almost surely. But this is a straightforward consequence of Kolmogorov's classical strong law of large numbers for martingale differences due to Chow (1965) (see also Stout (1974, Theorem 3.3.1)). It states that if $\{Z_i\}$ is a martingale difference sequence with

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}Z_n^2}{n^2} < \infty, \tag{10.6}$$

then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = 0 \qquad \text{almost surely.}$$

Thus, (10.5) is implied by Chow's theorem since the martingale differences $Z_i = \left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)$ are bounded by $4B^2$. □

PROOF OF THEOREM 10.2 Because of (10.4) we have to show that

$$\limsup_{n\to\infty} R_n(f) \le R^* \quad \text{a.s.}$$

By Corollary 10.1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\{Y_t \mid \mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}\} - g_t(\mathbf{X}_1^t, Y_1^{t-1})\right)^2 = 0 \qquad \text{a.s.} \tag{10.7}$$

Introduce the Bayes classification scheme using the infinite past:

$$f_t^*(\mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}) = \begin{cases} 1 & \text{if } \mathbb{P}\{Y_t = 1 \mid \mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}\} > 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

and its normalized cumulative $0-1$ loss:

$$R_n(f^*) = \frac{1}{n}\sum_{t=1}^{n} I_{\{f_t^*(\mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1})\neq Y_t\}}.$$

Put

$$\bar{R}_n(f) = \frac{1}{n}\sum_{t=1}^{n}\mathbb{P}\{f_t(\mathbf{X}_1^t, Y_1^{t-1}) \neq Y_t \mid \mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}\}$$

and

$$\bar{R}_n(f^*) = \frac{1}{n}\sum_{t=1}^{n}\mathbb{P}\{f_t^*(\mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}) \neq Y_t \mid \mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}\}.$$

Then

$$R_n(f) - \bar{R}_n(f) \to 0 \qquad \text{a.s.}$$

and

$$R_n(f^*) - \bar{R}_n(f^*) \to 0 \qquad \text{a.s.},$$

since they are the averages of bounded martingale differences. Moreover, by the ergodic theorem

$$\bar{R}_n(f^*) \to R^* \qquad \text{a.s.},$$

so we have to show that

$$\limsup_{n\to\infty}(\bar{R}_n(f) - \bar{R}_n(f^*)) \leq 0 \qquad \text{a.s.}$$

Lemma 10.1 implies that

$$\begin{aligned}
\bar{R}_n(f) - \bar{R}_n(f^*) &= \frac{1}{n}\sum_{t=1}^{n}\Big(\mathbb{P}\{f_t(\mathbf{X}_1^t, Y_1^{t-1}) \neq Y_t \mid \mathbf{X}_{-\infty}^t, Y_{-\infty}^{t-1}\} \\
&\quad -\mathbb{P}\{f_t^*(\mathbf{X}_{-\infty}^t, Y_{-\infty}^{t-1}) \neq Y_t \mid \mathbf{X}_{-\infty}^t, Y_{-\infty}^{t-1}\}\Big) \\
&\leq 2\frac{1}{n}\sum_{t=1}^{n}\left|\mathbb{E}\{Y_t \mid \mathbf{X}_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(\mathbf{X}_1^t, Y_1^{t-1})\right| \\
&\leq 2\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left|\mathbb{E}\{Y_t \mid \mathbf{X}_{-\infty}^t, Y_{-\infty}^{t-1}\} - g_t(\mathbf{X}_1^t, Y_1^{t-1})\right|^2} \\
&\to 0 \qquad \text{a.s.},
\end{aligned}$$

where in the last step we applied (10.7). □

Chapter 11

Density estimation

11.1 Why density estimation: the L_1 error

The classical nonparametric example is the problem estimating a distribution function

$$F(\mathbf{x}) = \mathbb{P}\{\mathbf{X} < \mathbf{x}\}.$$

from i.i.d. samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ taking values in $\mathbb{R}^d$. Here on the one hand the construction of the empirical distribution function

$$F_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i < \mathbf{x}\}}.$$

is distribution-free, and on the other hand its uniform convergence, the Glivenko-Cantelli Theorem holds for all F

$$\lim_{n \to \infty} \sup_{\mathbf{x} \in \mathbb{R}^d} |F_n(\mathbf{x}) - F(\mathbf{x})| = 0$$

a.s.

The Glivenko-Cantelli Theorem is really distribution-free, and the convergence in Kolmogorov- Smirnov distance means uniform convergence, so virtually it seems that there is no need to go further. However, if, for example, in a decision problem one wants to use empirical distribution functions for two unknown continuous distribution functions for creating a kind of likelihood then these estimates are useless. It turns out that we should look for stronger error criteria. For this purpose it is obvious to consider the total variation: if μ and ν are probability distributions on $\mathbb{R}^d$ ($d \geq 1$), then the *total variation distance* between μ and ν is defined by

$$V(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Borel sets A.

However, if μ stands for the common distribution of $\{X_i\}$ and μ_n denotes the empirical distribution then for nonatomic μ

$$V(\mu, \mu_n) = 1$$

a.s., so the empirical distribution is a bad estimate in total variation.

One may expect to find a more sophisticated sequence $\{\mu_n^*\}$ of distribution estimates of μ which is consistent in total variation:

$$\lim_{n\to\infty} V(\mu, \mu_n^*) = 0 \text{ a.s.}$$

Theorem 11.1. (DEVROYE AND GYÖRFI (1992)) *Given any sequence of distribution estimators $\{\mu_n^*\}$ there always exists a probability measure μ for which*

$$V(\mu, \mu_n^*) > 1/3 \text{ for all } n \text{ a.s.}$$

PROOF. This negative finding means that the total variation is a much stronger error criterion than the Kolmogorov-Smirnov distance such that it is impossible to construct a distribution estimate with distribution-free consistency in total variation. The proof borrows some arguments from Devroye (1983) and Rényi (1959). First, we need a rich family of singular continuous probability measures. The family of probability measures considered here is parametrized by a number $b \in [0, 1]$ with binary expansion $b = 0.b_{(1)}b_{(2)}b_{(3)}\ldots$, $b_{(i)} \in \{0, 1\}$. Let the random variables $Y_{(1)}, Y_{(2)}, \ldots$ be i.i.d. and uniformly distributed on $\{0, 1, 2\}$. We define the random variable $X = X(Y, b)$ by setting $X = 0.X_{(1)}X_{(2)}X_{(3)}\ldots$ in the ternary radix system used for $Y = 0.Y_{(1)}Y_{(2)}Y_{(3)}\ldots$, where

$$X_{(k)} = \begin{cases} 0, & \text{if } b_{(k)} = 0, \\ Y_{(k)}, & \text{if } b_{(k)} = 1. \end{cases}$$

Let μ_b denote the probability measure of $X = X(Y, b)$. If in the binary expansion of b there are finitely many (L) zeros, then μ_b is absolutely continuous and distributes its mass uniformly on a set of Lebesgue measure 3^{-L}.If in the binary expansion of b there are finitely many (L) ones, then μ_b is discrete and puts its mass uniformly on a set of cardinality 3^L. In other cases, μ_b is singular.

We write $X(Y_1, b), \ldots, X(Y_n, b)$ to denote a sample drawn from the distribution of $X(Y, b)$. We will replace b at a crucial step in the argument by a uniform $[0, 1]$ random variable B, which is independent of $Y_1, \ldots Y_n$. Put

$$A_k = \{0.x_{(1)}x_{(2)}\cdots : x_{(i)} \in \{0, 1, 2\} \text{ for all } i;\ x_{(k)} = 0\}.$$

Then

$$\mu_b(A_k) = \begin{cases} 1, & \text{if } b_{(k)} = 0, \\ 1/3, & \text{if } b_{(k)} = 1. \end{cases}$$

Let μ_n^* be an arbitrary distribution estimate based upon $X(Y_1, b), \ldots, X(Y_n, b)$. Let us now define the parameter estimate $b_n = 0.b_{n1}b_{n2}\ldots$ by its binary expansion with bits

$$b_{nk} = \begin{cases} 0, & \text{if } \mu_n^*(A_k) > 2/3, \\ 1, & \text{otherwise.} \end{cases}$$

Then

$$|\mu_n^*(A_k) - \mu_b(A_k)| \geq 1/3 I_{\{b_{nk} \neq b_{(k)}\}}.$$

Therefore

$$\begin{aligned} \sup_b \inf_n V(\mu_n^*, \mu_b) &= \sup_b \inf_n \sup_A |\mu_n^*(A) - \mu_b(A)| \\ &\geq \sup_b \inf_n \sup_k |\mu_n^*(A_k) - \mu_b(A_k)| \\ &\geq \sup_b \inf_n \sup_k 1/3 I_{\{b_{nk} \neq b_{(k)}\}}. \end{aligned}$$

Replace b by B and resulting b_{nk} by B_{nk}. Then

$$\begin{aligned} \sup_b \inf_n V(\mu_n^*, \mu_b) &\geq \inf_n \sup_k 1/3 I_{\{B_{nk} \neq B_{(k)}\}} \\ &= 1/3 \inf_n Z_n. \end{aligned}$$

Our theorem is proved if we can show that $Z_n = 1$ almost surely for all n. Put $Z_{Nn} = I_{\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\}}$. Then $Z_{Nn} \uparrow Z_n = I_{\{\cup_{k=1}^\infty [B_{nk} \neq B_{(k)}]\}}$. Therefore it suffices to show that

$$\lim_{N \to \infty} \mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\} = 1.$$

But $\mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\}$ is the error probability of the decision $(B_{n1}, \ldots, B_{nN})$ on $(B_{(1)}, \ldots B_{(N)})$ for the observations $X_1, \ldots, X_n$. For this decision problem the Bayes decision is

$$\tilde{B}_{nk} = \begin{cases} 0, & \text{if } X_{i(k)} = 0 \text{ for all } i = 1, \ldots, n, \\ 1, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned} \mathbb{P}\{Z_{Nn} = 1\} &= \mathbb{P}\{\cup_{k=1}^N [B_{nk} \neq B_{(k)}]\} \\ &\geq \mathbb{P}\{\cup_{k=1}^N [\tilde{B}_{nk} \neq B_{(k)}]\} \\ &= 1 - \left(1 - \frac{1}{2\,3^n}\right)^N \\ &\uparrow 1. \end{aligned}$$

□

In the sequel assume that the distribution μ has a density, which is denoted by f:

$$\mu(A) = \int_A f(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Then we show a way how to estimate a distribution consistently in total variation. The Scheffé Theorem below shows that the total variation is the half of the L_1 distance of the corresponding densities.

Theorem 11.2. *(*Scheffé (1947)*) If μ and ν are absolutely continuous with densities f and g, respectively, then*

$$\int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} = 2\, V(\mu, \nu).$$

(The quantity

$$L_1(f, g) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} \tag{11.1}$$

is called L_1-distance.*)*

Proof. Note that

$$\begin{aligned} V(\mu, \nu) &= \sup_A |\mu(A) - \nu(A)| \\ &= \sup_A \left| \int_A f - \int_A g \right| \\ &= \sup_A \left| \int_A (f - g) \right| \\ &= \int_{f>g} (f - g) \\ &= \int_{g>f} (g - f) \\ &= \frac{1}{2} \int |f - g|. \end{aligned}$$

□

The red area in Figure 11.1 is equal to the L_1 distance between the densities f and g.

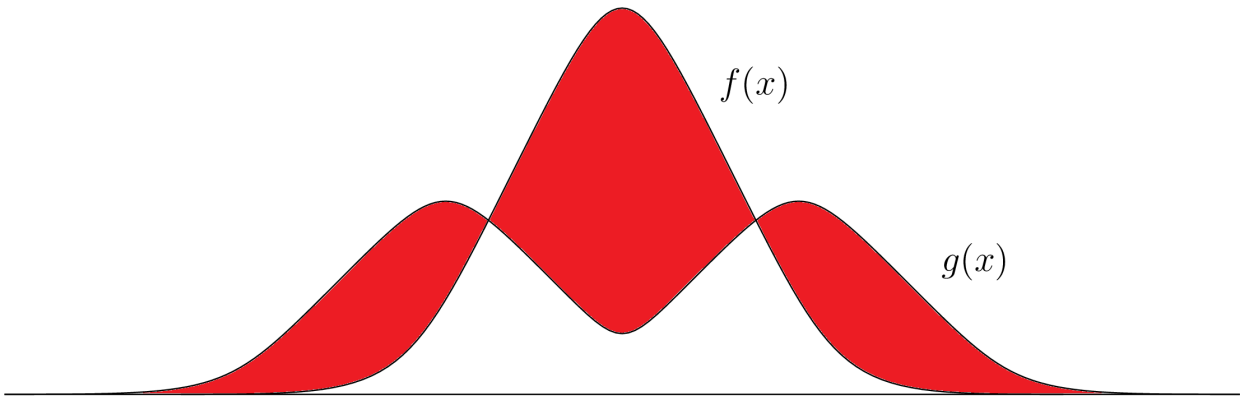

Figure 11.1: L_1 error.

The Scheffé Theorem implies an equivalent definition of the total variation:

$$V(\mu,\nu) = \frac{1}{2}\sup_{\{A_j\}}\sum_j |\mu(A_j) - \nu(A_j)|, \tag{11.2}$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$.

From i.i.d. samples $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ we may estimate the density function f, and such an estimate is denoted by

$$f_n(\mathbf{x}) = f_n(\mathbf{x}, \mathbf{X}_1, \ldots, \mathbf{X}_n).$$

In an obvious manner one can derive a distribution estimate μ_n^* as follows:

$$\mu_n^*(A) = \int_A f_n(\mathbf{x})\mathrm{d}\mathbf{x}.$$

Then the Scheffé theorem implies that

$$V(\mu,\mu_n^*) = \frac{1}{2}\int_{\mathbb{R}^d} |f(\mathbf{x}) - f_n(\mathbf{x})|\mathrm{d}\mathbf{x},$$

therefore if the density estimate f_n is consistent in L_1, i.e.,

$$\lim_{n\to\infty}\int |f(\mathbf{x}) - f_n(\mathbf{x})|\ \mathrm{d}\mathbf{x} = 0$$

a.s. then the corresponding distribution estimate μ_n^* is consistent in total variation:

$$\lim_{n\to\infty} V(\mu, \mu_n^*) = 0$$

a.s.

11.2 The histogram

Let μ_n denote the empirical distribution

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\mathbf{X}_i \in A\}}.$$

Let $\mathcal{P}_n = \{A_{n1}, A_{n2}, \dots\}$ be a partition of $\mathbb{R}^d$ such that the cells A_{nj} have positive and finite volume (Lebesgue measure λ). Then the histogram is defined by

$$f_n(\mathbf{x}) = \frac{\mu_n(A_n(\mathbf{x}))}{\lambda(A_n(\mathbf{x}))},$$

where

$$A_n(\mathbf{x}) = A_{nj}, \text{ if } \mathbf{x} \in A_{nj}.$$

For the partition $\mathcal{P}_n$, an example can be the cubic partition, when the cells are cubes of side length h_n. In this special case

$$f_n(\mathbf{x}) = \frac{\mu_n(A_n(\mathbf{x}))}{h_n^d}$$

Theorem 11.3. (Devroye and Györfi (1985)) *Assume that for each sphere S centered at the origin we have that*

$$\lim_{n\to\infty} \sup_{j: A_{nj}\cap S \neq \emptyset} \operatorname{diam}(A_{nj}) = 0$$

and

$$\lim_{n\to\infty} \frac{|\{j : A_{nj} \cap S \neq \emptyset\}|}{n} = 0,$$

then

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |f(\mathbf{x}) - f_n(\mathbf{x})|\, \mathrm{d}\mathbf{x}\right\} = 0.$$

PROOF. The triangle inequality implies that

$$\int |f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \le \underbrace{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x}}_{\text{variation term}} + \underbrace{\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x}}_{\text{bias}}.$$

The histogram is constant on a cell, therefore

$$\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \sum_j \int_{A_{nj}} |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \sum_j |\mu_n(A_{nj}) - \mu(A_{nj})|.$$

Put $M_n = |\{j : \ A_{nj} \cap S \neq \emptyset\}|$, and choose the numbering of the cells such that $A_{nj} \cap S \neq \emptyset$, $j = 1, \ldots, M_n$. Because of the condition of the theorem,

$$\frac{M_n}{n} \to 0.$$

Denote

$$S_n = \bigcup_{j=1}^{M_n} A_{nj}.$$

Then

$$\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x} \le \sum_{j=1}^{M_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + \mu_n(S_n^c) + \mu(S_n^c),$$

therefore the Cauchy-Schwarz and the Jensen inequalities imply that

$$\begin{aligned}
\mathbb{E}\left\{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x}\right\} &\le \sum_{j=1}^{M_n} \mathbb{E}\{|\mu_n(A_{nj}) - \mu(A_{nj})|\} + 2\mu(S_n^c) \\
&\le \sum_{j=1}^{M_n} \sqrt{\mathbb{E}\{|\mu_n(A_{nj}) - \mu(A_{nj})|^2\}} + 2\mu(S^c) \\
&\le \sum_{j=1}^{M_n} \sqrt{\frac{\mu(A_{nj})}{n}} + 2\mu(S^c) \\
&\le \sqrt{\frac{M_n}{n}} + 2\mu(S^c) \qquad (11.3) \\
&\to 2\mu(S^c).
\end{aligned}$$

The sphere S is arbitrary therefore

$$\mathbb{E}\left\{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x}\right\} \to 0.$$

Concerning the bias term, we have that

$$\mathbb{E}f_n(\mathbf{x}) = \frac{\mu(A_n(\mathbf{x}))}{\lambda(A_n(\mathbf{x}))} = \frac{1}{\lambda(A_n(\mathbf{x}))}\int\limits_{A_n(\mathbf{x})} f(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int f(\mathbf{z})K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z},$$

where

$$K_n(\mathbf{x},\mathbf{z}) = \frac{\mathbb{I}_{\{\mathbf{z}\in A_n(\mathbf{x})\}}}{\lambda(A_n(\mathbf{x}))}.$$

Then

$$\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} = \int \left|\int f(\mathbf{z})K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z} - f(\mathbf{x})\right|\,\mathrm{d}\mathbf{x}.$$

If f is continuous and is zero outside of a compact set then it is uniformly continuous, and the inequality

$$\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \le \int\int |f(\mathbf{z}) - f(\mathbf{x})|K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \tag{11.4}$$

implies that

$$\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \to 0.$$

If the density f is arbitrary then for any $\varepsilon > 0$ there is a density $\tilde{f}$ such that it is continuous and is zero outside of a compact set, and

$$\int |f(\mathbf{x}) - \tilde{f}(\mathbf{x})|\,\mathrm{d}\mathbf{x} < \varepsilon.$$

Then

$$\begin{aligned}
& \int |f(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \\
= & \int \left| f(\mathbf{x}) - \int f(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
\leq & \int |f(\mathbf{x}) - \tilde{f}(\mathbf{x})| \, \mathrm{d}\mathbf{x} + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
& + \int \left| \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} - \int f(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
\leq & \varepsilon + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
& + \int \left(\int |\tilde{f}(\mathbf{z}) - f(\mathbf{z})| K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{x} \right) \, \mathrm{d}\mathbf{z} \\
= & \varepsilon + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} + \int |\tilde{f}(\mathbf{z}) - f(\mathbf{z})| \, \mathrm{d}\mathbf{z} \\
\to & \ 2\varepsilon.
\end{aligned}$$

□

Without any tail and smoothness conditions on the density f, any slow rate of convergence can happen, for any density estimate. (Cf. Devroye (1983).)

Theorem 11.4. (Devroye and Györfi (1985)) *Assume that the density f is zero outside a sphere S and it is Lipschitz continuous, i.e.,*

$$|f(\mathbf{x}) - f(\mathbf{z})| \leq C \|\mathbf{x} - \mathbf{z}\|.$$

If the partition $\mathcal{P}_n$ is a cubic partition with side length h_n then for the histogram f_n, one has that

$$\mathbb{E}\left\{ \int |f(\mathbf{x}) - f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \right\} \leq \frac{c_1}{\sqrt{n h_n^d}} + c_2 h_n,$$

so for the choice

$$h_n = c_3 n^{-\frac{1}{d+2}}$$

we have that

$$\mathbb{E}\left\{ \int |f(\mathbf{x}) - f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \right\} \leq c_4 n^{-\frac{1}{d+2}}.$$

PROOF. For the variation term, (11.3) implies that

$$\mathbb{E}\left\{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x}\right\} \leq \sqrt{\frac{M_n}{n}} \leq \sqrt{\frac{\lambda(S)}{nh_n^d}}.$$

Concerning the bias term, (11.4) implies that

$$\begin{aligned}
\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} &\leq \int\int |f(\mathbf{z}) - f(\mathbf{x})| K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\
&\leq \int\int C\|\mathbf{z}-\mathbf{x}\| K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\
&\leq \int\int Ch_n\sqrt{d} K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\
&\leq Ch_n\sqrt{d}\lambda(S).
\end{aligned}$$

□

11.3 Kernel density estimate

Introduce the kernel density estimate such that choose a density $K(\mathbf{x})$, called kernel function. For a positive bandwidth h_n, the kernel estimate is defined by

$$f_n(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right).$$

Examples for kernels:

- Naive or window kernel
$$K(\mathbf{x}) = c\mathbb{I}_{\{\mathbf{x}\in S_{0,r}\}},$$
where $S_{0,r}$ is a sphere centered at the origin and with radius r.
- Gauss kernel
$$K(\mathbf{x}) = c\mathrm{e}^{-\|\mathbf{x}\|^2}.$$
- Cauchy kernel
$$K(\mathbf{x}) = \frac{c}{1+\|\mathbf{x}\|^{d+1}}.$$

- Epanechnikov kernel

$$K(\mathbf{x}) = c(1 - \|\mathbf{x}\|^2)\mathbb{I}_{\{\|\mathbf{x}\| \le 1\}}.$$

Theorem 11.5. (Devroye and Györfi (1985)) *If*

$$\lim_{n\to\infty} h_n = 0 \qquad \textit{and} \qquad \lim_{n\to\infty} nh_n^d = \infty.$$

then for the kernel density estimate f_n, one has

$$\lim_{n\to\infty} \mathbb{E}\left\{\int |f(\mathbf{x}) - f_n(\mathbf{x})|\, \mathrm{d}\mathbf{x}\right\} = 0.$$

Proof. With the notation

$$K_n(\mathbf{x}, \mathbf{z}) = \frac{1}{h_n^d} K\left(\frac{\mathbf{x} - \mathbf{z}}{h_n}\right),$$

one can prove Theorem 11.5 similarly to the proof of Theorem 11.3. □

Theorem 11.6. (Devroye and Györfi (1985)) *Assume that f is zero outside a sphere S and it is differentiable with Lipschitz continuous gradient, i.e.,*

$$\|f'(\mathbf{x}) - f'(\mathbf{z})\| \le C\|\mathbf{x} - \mathbf{z}\|.$$

If the kernel K has bounded support and

$$\int \mathbf{x} K(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbf{0}, \tag{11.5}$$

then for the kernel estimate f_n, one has that

$$\mathbb{E}\left\{\int |f(\mathbf{x}) - f_n(\mathbf{x})|\, \mathrm{d}\mathbf{x}\right\} \le \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2,$$

so for the choice

$$h_n = c_3 n^{-\frac{1}{d+4}}$$

we have that

$$\mathbb{E}\left\{\int |f(\mathbf{x}) - f_n(\mathbf{x})|\, \mathrm{d}\mathbf{x}\right\} \le c_4 n^{-\frac{2}{d+4}}.$$

PROOF. One can manage the variation term in Theorem 11.6 similarly to the variation term in Theorem 11.4. Concerning the bias, we have that

$$\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x}) = \int f(\mathbf{z}) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} - f(\mathbf{x}).$$

Then calculate the second order Taylor expansion of $f(\mathbf{z})$ at $\mathbf{x}$:

$$f(\mathbf{z}) = f(\mathbf{x}) + (f'(\mathbf{z}_\mathbf{x}), \mathbf{z}-\mathbf{x}),$$

where

$$\|\mathbf{z}_\mathbf{x} - \mathbf{x}\| \leq \|\mathbf{z}-\mathbf{x}\|.$$

Then

$$\begin{aligned}
&\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x}) \\
&= \int (f(\mathbf{x}) + (f'(\mathbf{z}_\mathbf{x}), \mathbf{z}-\mathbf{x})) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} - f(\mathbf{x}) \\
&= \int (f'(\mathbf{z}_\mathbf{x}) - f'(\mathbf{x}), \mathbf{z}-\mathbf{x}) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} + \int (f'(\mathbf{x}), \mathbf{z}-\mathbf{x}) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z}.
\end{aligned}$$

Because of (11.5), we have that

$$\int (f'(\mathbf{x}), \mathbf{z}-\mathbf{x}) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} = 0.$$

Furthermore, the Lipschitz condition implies that

$$\begin{aligned}
\left| \int (f'(\mathbf{z}_\mathbf{x}) - f'(\mathbf{x}), \mathbf{z}-\mathbf{x}) \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} \right| &\leq C \int \|\mathbf{x}-\mathbf{z}\|^2 \frac{1}{h_n^d} K\left(\frac{\mathbf{x}-\mathbf{z}}{h_n}\right) \mathrm{d}\mathbf{z} \\
&= O(h_n^2).
\end{aligned}$$

□

For further reading on L_1 density estimation, the books Devroye, Györfi (1985), Devroye (1987) and Devroye, Lugosi (2001) are suggested.

Bibliography

Abu-Shikhah, N., Elkarmi, F., and Aloquili, O. M. (2011). Medium-term electric load forecasting using multivariable linear and non-linear regression. *Smart Grid and Renewable Energy*, 2:126–135.

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6:127–132.

Alfares, H. K. and Nazeeruddin, M. (2002). Electric load forecasting: literature survey and classiÂŽcation of methods. *International Journal of Systems Science*, 33:23–34.

Algoet, P. (1994). The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, 40:609–633.

Almeshaiei, E. and Soltan, H. (2011). A methodology for electric power load forecasting. *Alexandria Engineering Journal*, 50:137–144.

Arkadjew, A. G. and Braverman, E. M. (1966). *Zeichenerkennung und Maschinelles Lernen*. Oldenburg Verlag, München, Wien.

Aung, Z., Toukhy, M., Williams, J. R., Sanchez, A., and Herrero, S. (2012). Towards accurate electricity load forecasting in smart grids. In *DBKDA 2012 : The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*, pages 51–57.

Ba, A., Sinn, M., Goude, Y., and Pompey, P. (2012). Adaptive learning of smoothing functions: application to electricity load forecasting. In *Advances in Neural Information Processing Systems*, pages 2510–2518.

Beck, J. (1979). The exponential rate of convergence of error for k_n-NN nonparametric regression and decision. *Problems of Control and Information Theory*, 8:303–311.

Beirlant, J. and Györfi, L. (1998). On the asymptotic L_2-error in partitioning regression estimation. *Journal of Statistical Planning and Inference*, 71:93–107.

Bhattacharya, P. K. and Gangopadhyay, A. K. (1990). Kernel and nearest neighbor estimation of a conditional quantile. *Annals of Statistics*, 18:1400–1415.

Bhattacharya, P. K. and Mack, Y. P. (1987). Weak convergence of k–NN density and regression estimators with varying k and applications. *Annals of Statistics*, 15:976–994.

Biau, G., Bleakley, K., Györfi, L., and Ottucsák, G. (2010). Nonparametric sequential prediction of time series. *Journal of Nonparametric Statistics*, 22:297–317.

Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer-Verlag, New York.

Biau, G. and Györfi, L. (2005). On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51:3965–3973.

Biau, G. and Patra, B. (2011). Sequential quantile prediction of time series. *IEEE Trans. Information Theory*, 57:1664–1674.

Bickel, P. J. and Breiman, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Annals of Probability*, 11:185–214.

Blum, J. R., Chernoff, H., Rosenblatt, M., and Teicher, H. (1958). Central limit theorems for interexchangeable processes. *Canadian J. of Mathematics*, 10:222–229.

Bosq, D. and Lecoutre, J. P. (1987). *Théorie de l' Estimation Fonctionnelle*. Economica, Paris.

Boucheron, S., Lugosi, G., and Massart, P. (2010). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford.

Bozic, M., Stojanovic, M., and Stajic, Z. (2010). Short-term electric load forecasting using least square support vector machines. *Facta Universitatis*, 9:141–150.

Breiman, L. (1957). The individual ergodic theorem of information theory. *Annals of Mathematical Statistics*, 28:809–811.

Broder, A. J. (1990). Strategies for efficient incremental nearest neighbor search. *Pattern Recognition*, 23:171–178.

Bruhns, A., Deurveilher, G., and Roy., J.-S. (2005). A non-linear regression model for mid-term load forecasting and improvements in seasonnality. In *Proceedings of the Fifteenth Power Systems Computation Conference (PSCC)*.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York.

Cacoullos, T. (1965). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 18:179–190.

Cancelo, J. R., Espasa, A., and Grafe, R. (2008). Forecasting the electricity load from one day to one week ahead for the spanish system operator. *International Journal of Forecasting*, 24:588–602.

Candés, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete fourier information. *IEEE Trans. Information Theory*, 52:489–509.

Caner, M. (2002). A note on least absolute deviation estimation of a threshold model. *Econometric Theory*, 18:800–814.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press, New York.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics*, 19:760–777.

Cheng, P. E. (1995). A note on strong convergence rates in nonparametric regression. *Statistics and Probability Letters*, 24:357–364.

Chow, C. K. (1965). Statistical independence and threshold functions. *IEEE Transactions on Computers*, E-14:66–68.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Collomb, G. (1977). Quelques proprietes de la méthode du noyau pour l'estimation nonparametrique de la regression en un point fixe. *Comptes Rendus de l'Académie des Sciences de Paris*, 285:28–292.

Collomb, G. (1979). Estimation de la regression par la méthode des k points les plus proches: propriétés de convergence ponctuelle. *Comptes Rendus de l'Académie des Sciences de Paris*, 289:245–247.

Collomb, G. (1980). *Estimation de la regression par la méthode des k points les plus proches avec noyau*. Lecture Notes in Mathematics #821, Springer-Verlag, Berlin. 159–175.

Collomb, G. (1981). Estimation non parametrique de la regression: revue bibliographique. *International Statistical Review*, 49:75–93.

Coomans, D. and Broeckaert, I. (1986). *Potential Pattern Recognition in Chemical and Medical Decision Making*. Research Studies Press, Letchworth, Hertfordshire, England.

Cover, T. M. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55.

Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

Dasarathy, B. V. (1991). *Nearest Neighbor Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA.

De Brabanter, K., Ferrario, P. G., and Györfi, L. (2014). Detecting ineffective features for nonparametric regression. In *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 177–194. Chapman & Hall/CRC Machine Learning and Pattern Recognition Series.

Dette, H. and Volgushev, S. (2008). Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B*, 70:609–627.

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, 90:231–260.

Devijver, P. A. (1980). An overview of asymptotic properties of nearest neighbor rules. In *Pattern Recognition in Practice*, Gelsema, E. S. and Kanal, L. N., editors, pages 343–350. Elsevier Science Publishers, Amsterdam.

Devroye, L. (1978). The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24:142–151.

Devroye, L. (1981a). On the almost everywhere convergence of nonparametric regression function estimates. *Annals of Statistics*, 9:1310–1319.

Devroye, L. (1981b). On the asymptotic probability of error in nonparametric discrimination. *Annals of Statistics*, 9:1320–1327.

Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61:467–481.

Devroye, L. (1983). On arbitrarily slow rates of global convergence in density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:475–483.

Devroye, L. (1987). *A Course in Density Estimation*. Birkhäuser, Boston.

Devroye, L., Ferrario, P., Györfi, L., and Walk, H. (2013). Strong universal consistent estimate of the minimum mean squared error. In *Empirical Inference - Festschrift in Honor of Vladimir N. Vapnik*, pages 143–160. Springer.

Devroye, L. and Györfi, L. (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, Konecny, F., Mogyoródi, J., and Wertz, W., editors, pages 67–76. Akadémiai Kiadó, Budapest, Hungary.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Devroye, L. and Györfi, L. (1992). No empirical probability measure can converge in the total variation sense for all distributions. *Annals of Statistics*, 18:1496–1499.

Devroye, L., Györfi, L., and Krzyżak, A. (1998). The Hilbert kernel regression estimate. *Journal of Multivariate Analysis*, 65:209–227.

Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Devroye, L., Györfi, L., Lugosi, G., and Walk, H. (2017). On the measure of voronoi cells. *Journal of Applied Probability*, 54:394–408.

Devroye, L., Györfi, L., Lugosi, G., and Walk, H. (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12:1752–1778.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23:71–82.

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer-Verlag, New York.

Devroye, L. and Lugosi, G. (2002). Almost sure classification of densities. *J. Nonparametr. Stat.*, 14:675–698.

Devroye, L., Schäfer, D., Györfi, L., and Walk, H. (2003). The estimation problem of minimum mean squared error. *Statistics and Decisions*, 21:15–28.

Devroye, L. and Wagner, T. J. (1976). Nonparametric discrimination and density estimation. Technical Report 183, Electronics Research Center, University of Texas.

Devroye, L. and Wagner, T. J. (1980a). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8:231–239.

Devroye, L. and Wagner, T. J. (1980b). On the L_1 convergence of kernel estimators of regression functions with applications in discrimination. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 51:15–21.

Devroye, L. and Wagner, T. J. (1982). Nearest neighbor methods in discrimination. In *Handbook of Statistics*, Krishnaiah, P. R. and Kanal, L., editors, volume 2, pages 193–197. North Holland, Amsterdam.

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Information Theory*, 52:1289–1306.

Dordonnat, V., Koopman, S., Ooms, M., Dessertaine, A., and Collet, J. (2008). An hourly periodic state space model for modelling french national electricity load. *International Journal of Forecasting*, 24:566–587.

Elad, M. (2010). *Sparse and Redundant Representations*. Springer, New York.

Elattar, E. E., Goulermas, J., and Wu, Q. H. (2010). Electric load forecasting based on locally weighted support vector regression. *IEEE Trans. Systems, Man, and Cybernetics*, 40:438–447.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge.

Evans, D. and Jones, A. J. (2008). Non-parametric estimation of residual moments and covariance. *Proceedings of the Royal Society*, A 464:2831–2846.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21:196–216.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, 20:2008–2036.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

Fan, J., Hu, T. C., and Truong, Y. K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, 21:433–446.

Faragó, A., Linder, T., and Lugosi, G. (1993). Fast nearest neighbor search in dissimilarity spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:957–962.

Feinberg, E. A. and Genethliou, D. (2005). Load forecasting. In *Applied Mathematics for Restructured Electric Power Systems*, pages 269–285. Springer.

Ferrario, P. G. and Walk, H. (2012). Nonparametric partitioning estimation of residual and local variance based on first and second nearest neighbors. *Journal of Nonparametric Statistics*, 24:1019–1039.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical Report 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Fix, E. and Hodges, J. L. (1952). Discriminatory analysis: small sample performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX.

Fix, E. and Hodges, J. L. (1991a). Discriminatory analysis, nonparametric discrimination, consistency properties. In *Nearest Neighbor Pattern Classification Techniques*, Dasarathy, B., editor, pages 32–39. IEEE Computer Society Press, Los Alamitos, CA.

Fix, E. and Hodges, J. L. (1991b). Discriminatory analysis: small sample performance. In *Nearest Neighbor Pattern Classification Techniques*, Dasarathy, B. V., editor, pages 40–56. IEEE Computer Society Press, Los Alamitos, CA.

Förstner, W. and Wrobel, B. P. (2016). *Photogrammetric Computer Vision: Statistics, Geometry, Orientation and Reconstraction*. Springer, Cham.

Friedman, J. H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers*, 26:404–408.

Friedman, J. H., Baskett, F., and Shustek, L. J. (1975). An algorithm for finding nearest neighbor. *IEEE Transactions on Computers*, 24:1000–1006.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3:209–226.

Fritz, J. (1974). Learning from ergodic training sequence. In *Limit Theorems of Probability Theory*, Révész, P., editor, pages 79–91. North-Holland, Amsterdam.

Fukunaga, K. and Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. *IEEE Transactions on Computers*, 24:750–753.

Gaillard, P. and Goude, Y. (2011). Forecasting electricity consumption by aggregating experts; how to design a good set of experts. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pages 95–115. Springer.

Greblicki, W. (1974). Asymptotically optimal probabilistic algorithms for pattern recognition and identification. Technical Report, Monografie No. 3, Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawskiej No. 18, Wroclaw, Poland.

Greblicki, W. (1978a). Asymptotically optimal pattern recognition procedures with density estimates. *IEEE Transactions on Information Theory*, 24:250–251.

Greblicki, W. (1978b). Pattern recognition procedures with nonparametric density estimates. *IEEE Transactions on Systems, Man and Cybernetics*, 8:809–812.

Gretton, A. and Györfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Multivariate Analysis*, 11:1391–1423.

Guerre, E. (2000). Design adaptive nearest neighbor regression estimation. *Journal of Multivariate Analysis*, 75:219–255.

Györfi, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Transactions on Information Theory*, 29:509–512.

Györfi, L. (1984). Adaptive linear procedures under general conditions. *IEEE Transactions on Information Theory*, 30:262–267.

Györfi, L. (1991). Universal consistencies of a regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, pages 329–338. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Györfi, L. and Györfi, Z. (1975). On the nonparametric estimate of a posteriori probabilities of simple statistical hypotheses. In *Colloquia Mathematica Societatis János Bolyai: Topics in Information Theory*, pages 299–308. Keszthely, Hungary.

Györfi, L. and Györfi, Z. (1978). An upper bound on the asymptotic error probability of the k-nearest neighbor rule for multiple classes. *IEEE Transactions on Information Theory*, 24:512–514.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

Györfi, L. and Lugosi, G. (2002). Strategies for sequential prediction of stationary time series. In *Modeling Uncertainity: An Examination of its Theory, Methods and Applications*, Dror, M., L'Ecuyer, P., and Szidarovszky, F., editors, pages 225–248. Kluwer Academic Publishers, Dordrecht.

Györfi, L. and Ottucsák, G. (2007). Sequential prediction of unbounded time series. *IEEE Transactions on Information Theory*, 53:1866–1872.

Györfi, L. and Sanchetta, A. (2014). An open problem on strongly consistent learning of the best prediction for gaussian processes. In *Topics in Nonparametric Statistics. Proceedings of the First Conference of the International Society of Nonparametric Statistics*, pages 115–136. Springer.

Györfi, L., Schäfer, D., and Walk, H. (2002). Relative stability of global errors in nonparametric function estimations. *IEEE Transactions on Information Theory*, 48:2230–2242.

Györfi, L. and Walk, H. (2015). On the asymptotic normality of an estimate of a regression functional. *Journal of Machine Learning Research*, 16:1863–1877.

Hall, P. and Müller, H.-G. (2003). Order-preserving nonparametric regression, with applications to conditional distribution and quantile function estimation. *J. Am. Stat. Assoc.,*, 98:598–608.

Hand, D. J. (1981). *Discrimination and Classification.* John Wiley, Chichester.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, Cambridge, UK.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30.

Katkovnik, V. Y. (1979). Linear and nonlinear methods for nonparametric regression analysis. *Avtomatika*, 5:35–46.

Katkovnik, V. Y. (1983). Convergence of linear and nonlinear nonparametric estimates of "kernel" type. *Automation and Remote Control*, 44:495–506.

Katkovnik, V. Y. (1985). *Nonparametric Identification and Data Smoothing: Local Approximation Approach (in Russian).* Nauka, Moscow.

Kim, B. S. and Park, S. B. (1986). A fast k-nearest neighbor finding algorithm based on the ordered partition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:761–766.

Kivinen, J. and Warmuth, M. K. (1999). Averaging expert predictions. In *Computational Learning Theory: Proceedings of the Fourth European Conference, Eurocolt'99*, Simon, H. U. and Fischer, P., editors, pages 153–167. Springer.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, Cambridge.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction*. Springer-Verlag, Berlin.

Kraus, K. (2007). *Photogrammetry: Geometry from Images and Laser Scans*. De Gruyter, Berlin, New York.

Krzyżak, A. and Pawlak, M. (1984). Distribution-free consistency of a nonparametric kernel regression estimate and classification. *IEEE Transactions on Information Theory*, 30:78–81.

Kulkarni, S. R. and Posner, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41:1028–1039.

Lecoutre, J. P. (1980). Estimation d'une fonction de régression pour par la méthode du regressogramme á blocks équilibrés. *Comptes Rendus de l'Académie des Sciences de Paris*, 291:355–358.

Lejeune, M. G. and Sarda, P. (1988). Quantile regression: A nonparametric approach. *Computational Statistics and Data Analysis*, 6:229–239.

Liitiäinen, E., Corona, F., and Lendasse, A. (2008). On nonparametric residual variance estimation. *Neural Processing Letters*, 28:155–167.

Liitiäinen, E., Corona, F., and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101:811–823.

Liitiäinen, E., Verleysen, M., Corona, F., and Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72:3692–3703.

Lugosi, G. and Nobel, A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics*, 24:687–706.

Luhmann, T., Robson, S., Kyle, S., and Boehm, J. (2013). *Close-Range Photogrammetry and 3D Imaging*. De Gruyter, Berlin, Boston.

Mack, Y. P. (1981). Local properties of k–nearest neighbor regression estimates. *SIAM Journal on Algebraic and Discrete Methods*, 2:311–323.

Meisel, W. (1969). Potential functions in mathematical pattern recognition. *IEEE Transactions on Computers*, 18:911–918.

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J. (2010). Optimized clusters for dissaggregated electricity load forecasting. *Revstat – Statistical Journal*, 8:105–124.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9:141–142.

Nadaraya, E. A. (1970). Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, 15:134–137.

Nagi, J., Yap, K. S., Tiong, S. K., and Ahmed, S. K. (2008). Electrical power load forecasting using hybrid self-organizing maps and support vector machines. In *The 2nd International Power Engineering and Optimization Conference (PEOCO 2008), Shah Alam, Selangor, MALAYSIA. 4-5 June 2008.*, pages 51–56.

Niemann, H. and Goppert, R. (1988). An efficient branch-and-bound nearest neighbour classifier. *Pattern Recognition Letters*, 7:67–72.

Papadimitriou, C. H. and Bentley, J. L. (1980). A worst-case analysis of nearest neighbor searching by projection. In *Automata, Languages and Programming 1980*, pages 470–482. Lecture Notes in Computer Science #85, Springer-Verlag, Berlin.

Parzen, E. (1962). On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, 33:1065–1076.

Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer-Verlag, Berlin.

Pierrot, A. and Goude, Y. (2011). Short-term electricity load forecasting with generalized additive models. In *Proceedings of the Sixteenth International Conference on Intelligent System Application to Power Systems (ISAP).*

Prékopa, A. (2006). On the Hungarian inventory control model. *European Journal of Operational Research*, 171:894–914.

Rejtő, L. and Révész, P. (1973). Density estimation and pattern classification. *Problems of Control and Information Theory*, 2:67–80.

Rényi, A. (1959). On the dimension and entropy of probability distributions. *Acta Mathematica Acad. Sci. Hungar.*, 10:193–215.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837.

Ryzin, J. V. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankhya Series A*, 28:161–170.

Schäfer, D. (2002). Strongly consistent online forecasting of centered gaussian processes. *IEEE Trans. Information Theory*, 48:791–799.

Scheffé, H. (1947). A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–458.

Sebestyen, G. (1962). *Decision Making Processes in Pattern Recognition.* Macmillan, New York.

Sethi, I. K. (1981). A fast algorithm for recognizing nearest neighbors. *IEEE Transactions on Systems, Man and Cybernetics*, 11:245–248.

Sevlian, R. and Rajagopal, R. (2018). A scaling law for short term electricity load forecasting on varying levels of aggregation. *International Journal of Electrical Power and Energy Systems*, 98:350–361.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17:211–225.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5:595–645.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8:1348–1360.

Stout, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.

Stroock, D. W. (2011). *Probability Theory: An Analytic View*. Cambridge University Press, Cambridge.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Annals of Statistics*, 12:917–926.

Taylor, J. W. and McSharry, P. E. (2008). Short-term load forecasting methods: an evaluation based on european data. *IEEE Trans. Power Systems*, 22:2213–2219.

Toomey, J. W. (2000). *Inventory Management: Principles, Concepts and Techniques*. Kluwer, Boston, Dordrecht, London.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, 22:133–146.

Tsypkin, Y. Z. (1971). *Adaptation and Learning in Automatic Systems*. Academic Press, New York.

Tukey, J. W. (1947). Nonparametric estimation II. Statistically equivalent blocks and tolerance regions. *Annals of Mathematical Statistics*, 18:529–539.

Tukey, J. W. (1961). Curves as parameters and touch estimation. *Proceedings of the Fourth Berkeley Symposium*, pages 681–694.

Vidal, E. (1986). An algorithm for finding nearest neighbors in (approximately) constant average time. *Pattern Recognition Letters*, 4:145–157.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, 26:359–372.

Weber, N. C. (1980). A martingale approach to central limit theorems for exchangeable random variables. *Journal of Applied Probability*, 17:662–673.

Yunck, T. P. (1976). A technique to identify nearest neighbors. *IEEE Transactions on Systems, Man, and Cybernetics*, 6:678–683.

Zhao, L. C. (1987). Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *Journal of Multivariate Analysis*, 21:168–178.

NEMZETI KÖZSZOLGÁLATI EGYETEM