Nonparametric Statistics

László Györfi

February 3, 2014

Contents

Chapter 1

Introduction

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:

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

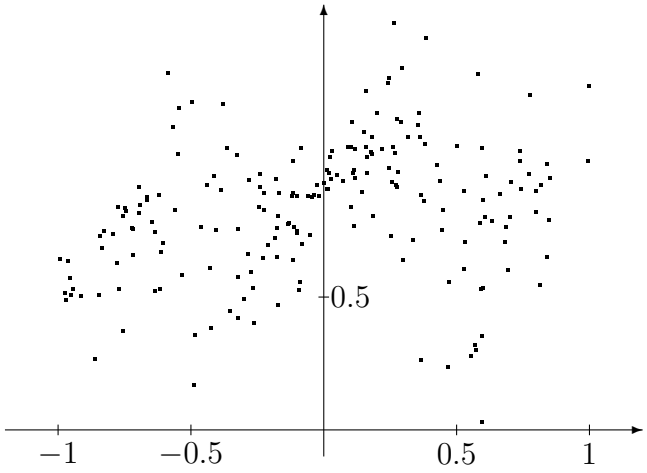

Figure 1.1: Simulated data points.

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

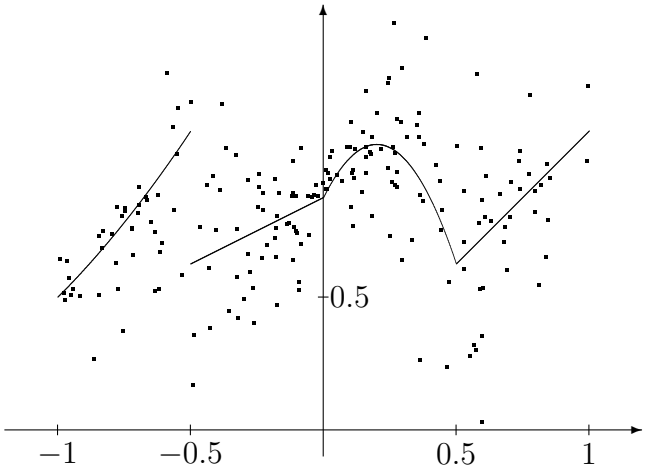

Figure 1.2: Data points and regression function.

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.

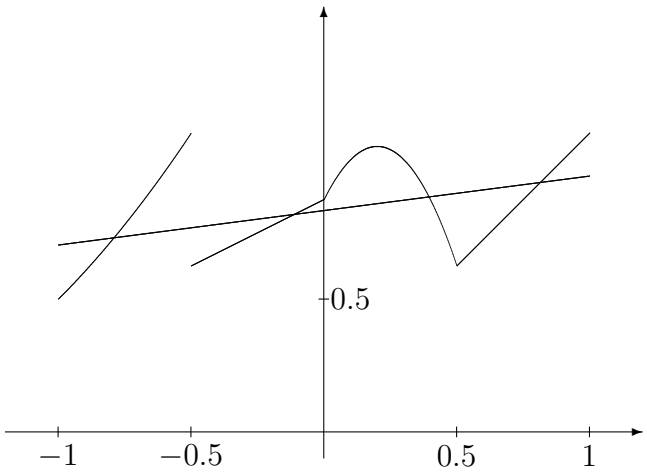

Figure 1.3: Linear regression estimate.

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 [**?**]). 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

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

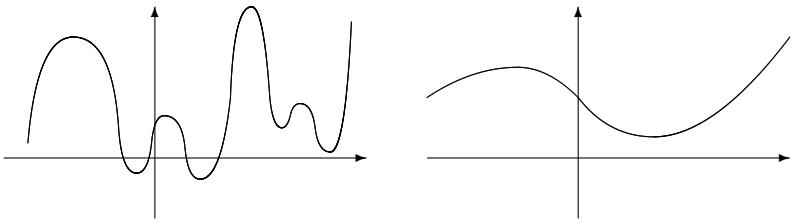

Figure 1.4: The estimate on the right seems to be more reasonable than the estimate on the left, which interpolates the data.

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,

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

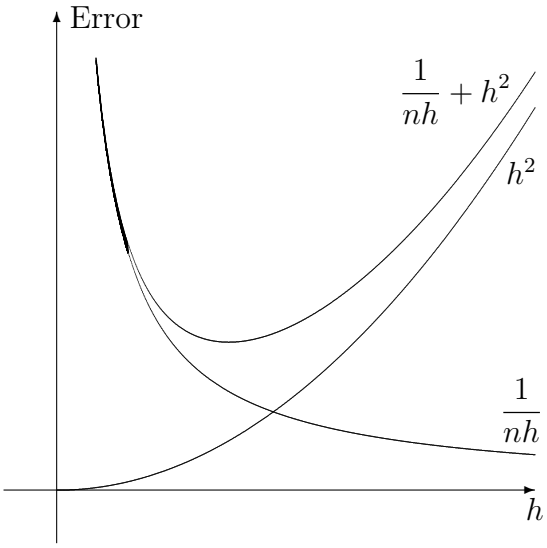

Figure 1.5: Bias–variance tradeoff.

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.* [**?**].

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 $h_{n1}, \ldots, h_{nd}$. 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,

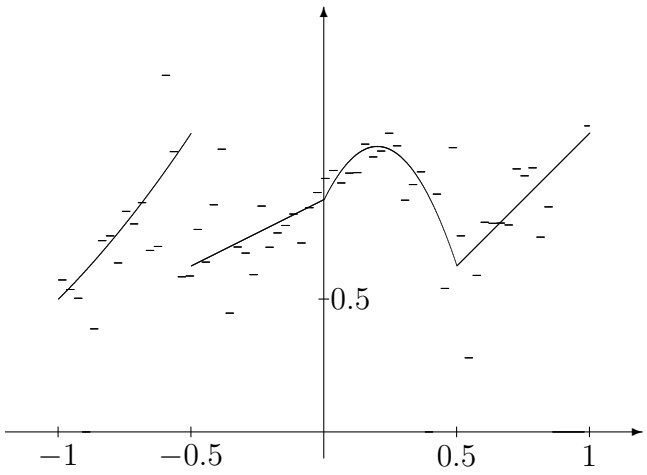

Figure 2.1: Undersmoothing: $h = 0.03$, L_2 error $= 0.062433$.

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 [**?**]).

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.

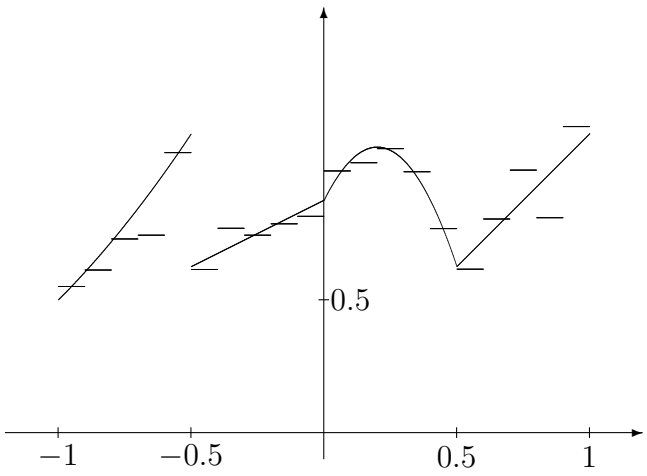

Figure 2.2: Good choice: $h = 0.1$, L_2 error $= 0.003642$.

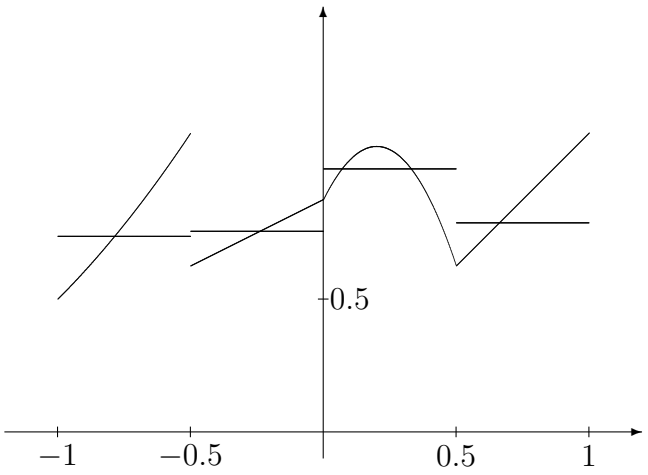

Figure 2.3: Oversmoothing: $h = 0.5$, L_2 error $= 0.013208$.

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.

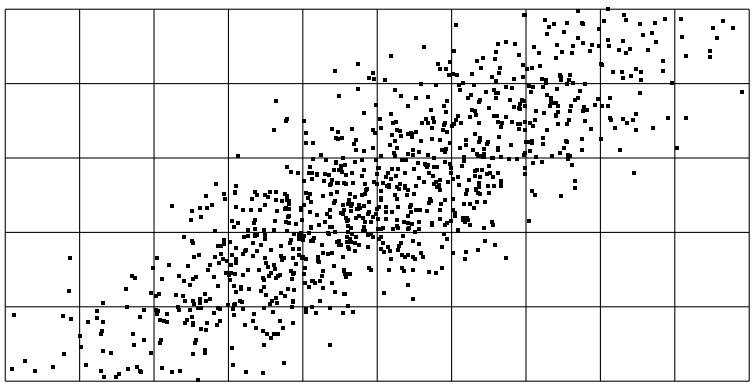

Figure 2.4: Cubic partition.

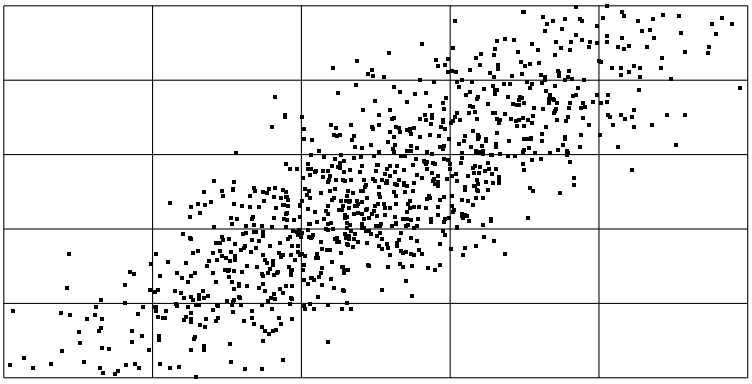

Figure 2.5: Rectangle partition.

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.

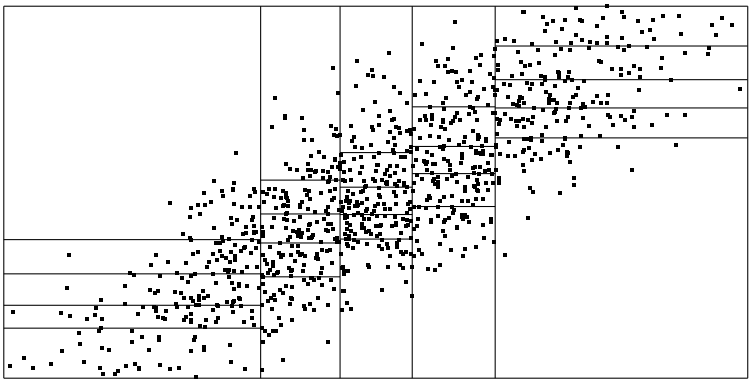

Figure 2.6: Statistically equivalent blocks.

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). *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,

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

(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.

(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.

(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.

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 $0 \leq p \leq 1$ 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)

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

(ii)

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

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

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

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

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

yield

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

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

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

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. *For a cubic partition with side length h_n assume that*

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

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

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

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

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

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

By Jensen's inequality

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

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),

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

□

Chapter 3

Kernel Estimates

3.1 Introduction

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 ($K(\mathbf{x}) = \mathbb{I}_{\{\|\mathbf{x}\|\leq 1\}}$) 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.

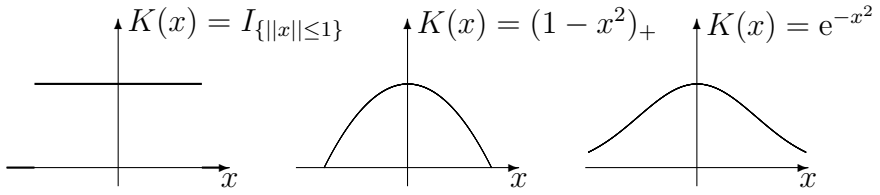

Figure 3.1: Examples for univariate kernels.

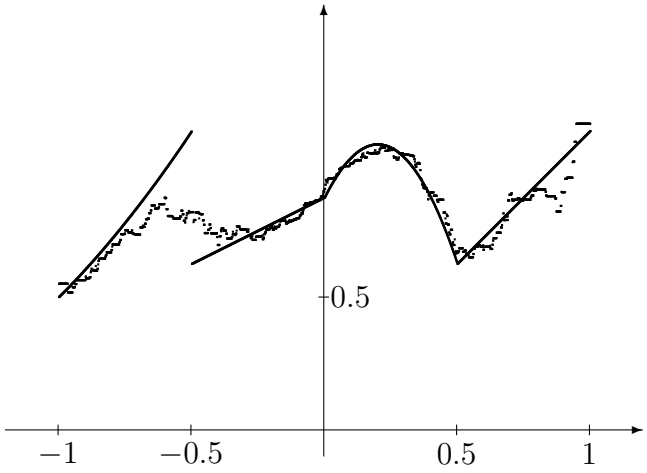

Figure 3.2: Kernel estimate for the naive kernel: $h = 0.1$, L_2 error $= 0.004066$.

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*

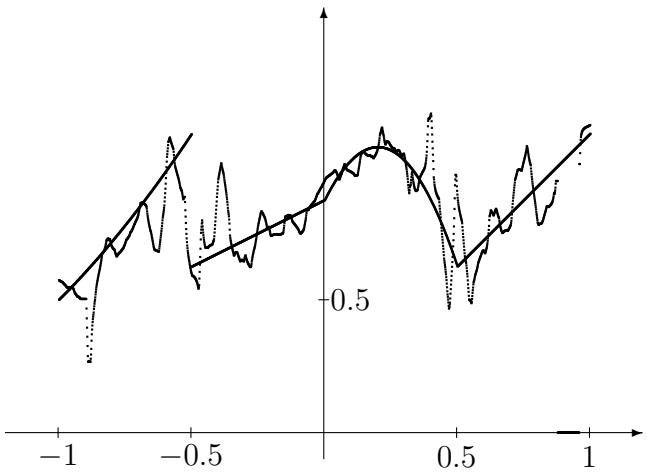

Figure 3.3: Undersmoothing for the Epanechnikov kernel: $h = 0.03$, L_2 error $= 0.031560$.

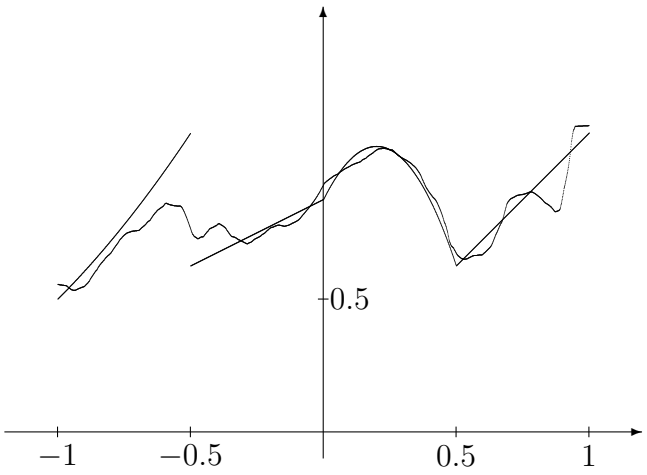

Figure 3.4: Kernel estimate for the Epanechnikov kernel: $h = 0.1$, L_2 error $= 0.003608$.

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 $nh_n^d \to \infty$, 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

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

Condition (i) means that

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

with $c > 0$. Because of

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

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

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

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

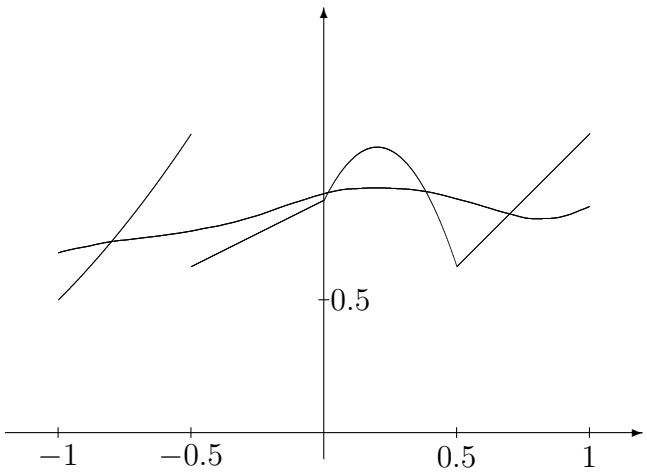

Figure 3.5: Oversmoothing for the Epanechnikov kernel: $h = 0.5$, L_2 error $= 0.012551$.

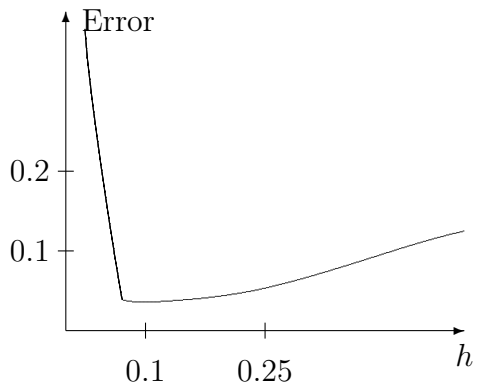

Figure 3.6: The L_2 error for the Epanechnikov kernel as a function of h.

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

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

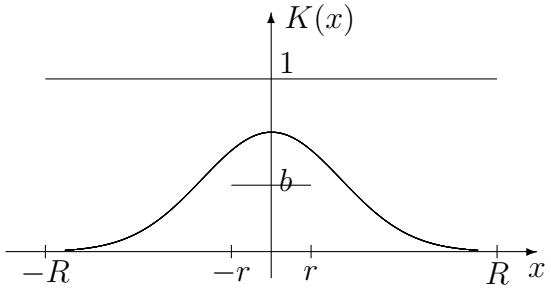

Figure 3.7: Boxed kernel.

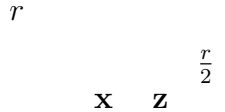

Figure 3.8: If $\mathbf{x} \in S_{\mathbf{z},r/2}$, then $S_{\mathbf{z},r/2} \subseteq S_{\mathbf{x},r}$.

(cf. Figure 3.8). 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 $h_n R < a$,

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

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

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

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

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

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. *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*

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

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-NN Estimates

4.1 Introduction

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

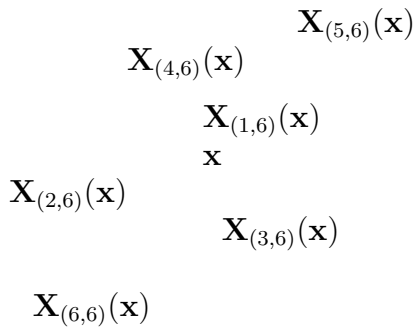

Figure 4.1: Illustration of nearest neighbors.

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 4.2 – 4.4 show k_n-NN estimates for various choices of k_n for our simulated data introduced in Chapter 1. Figure 4.5 shows the L_2 error as a function of k_n.

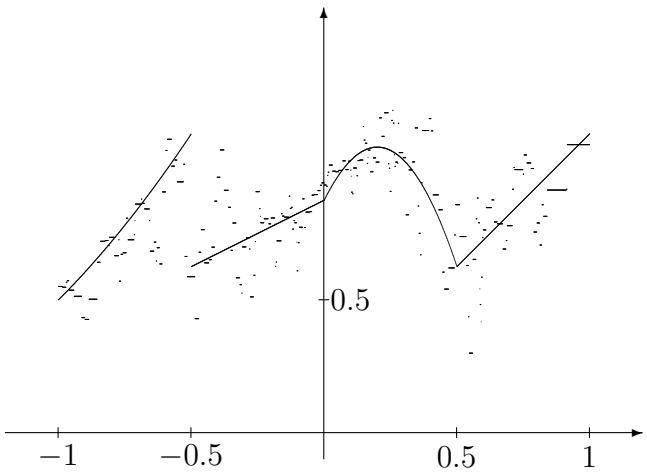

Figure 4.2: Undersmoothing: $k_n = 3$, L_2 error =0.011703.

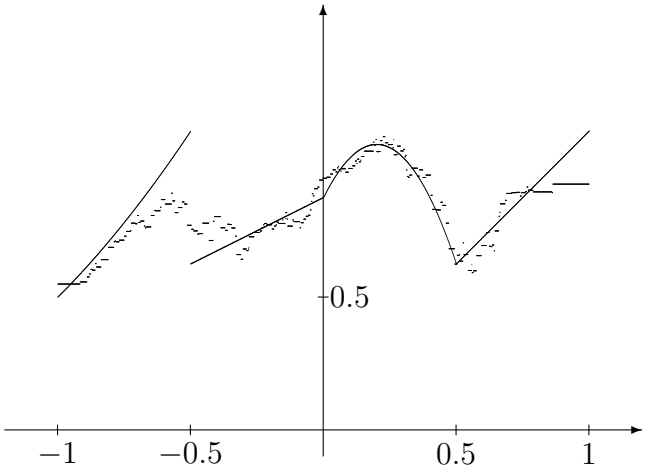

Figure 4.3: Good choice: $k_n = 12$, L_2 error =0.004247.

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*

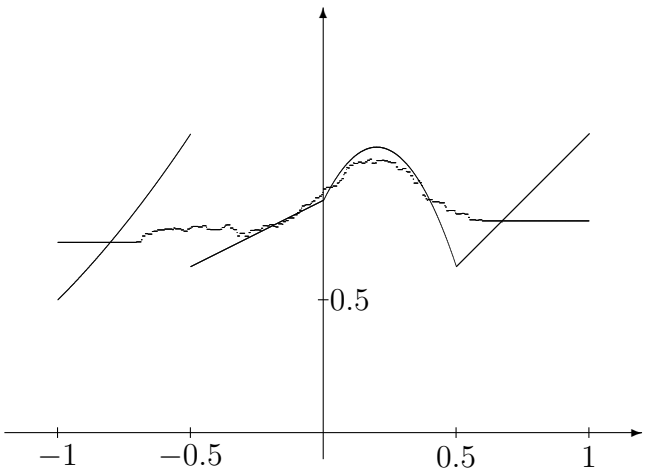

Figure 4.4: Oversmoothing: $k_n = 50$, L_2 error =0.009931.

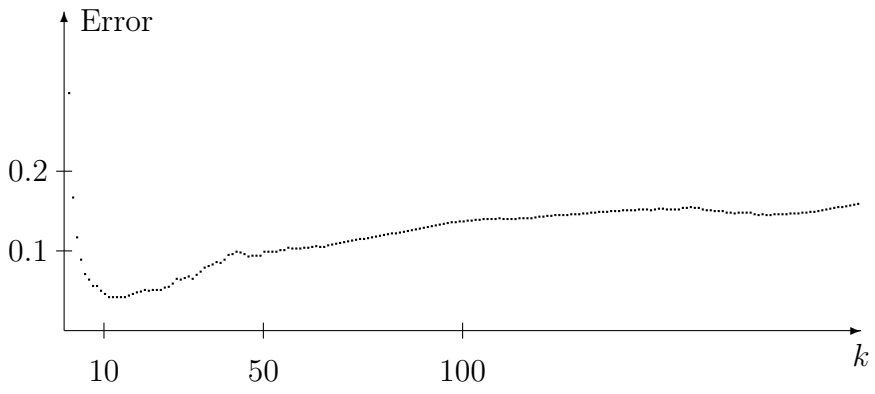

Figure 4.5: L_2 error of the k-NN estimate as a function of k.

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,

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

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.

Lemma 4.2. *Let*

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

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

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

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

PROOF. Let $C_j \subset \mathcal{R}^d$ 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 4.6). Let $C_1, \ldots, C_{\gamma_d}$ be a collection of such cones with different central directions such that their union covers $\mathbb{R}^d$:

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

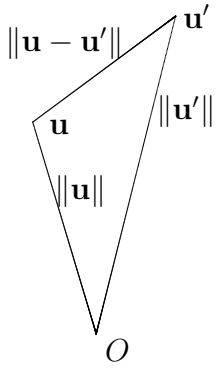

Figure 4.6: The cone property.

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

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

a.s. □

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

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

where γ_d depends upon the dimension only.

PROOF. If f is a nonnegative function,

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

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,

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

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. *Assume that* $\mathbf{X}$ *is bounded,*

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

and

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

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

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

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

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

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* $d \geq 3$*, then*

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

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

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

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

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

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

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

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

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

therefore, by Jensen's inequality,

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

Thus, by Lemma 4.4,

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

□

Chapter 5

Prediction of time series

5.1 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 [**?**], who showed that for any prediction strategy g and stationary ergodic process $\{(\mathbf{X}_n, Y_n)\}_{-\infty}^{\infty}$,

$$\liminf_{n\to\infty} L_n(g) \geq L^* \quad \text{almost surely,} \tag{5.1}$$

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 5.1. *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 [**?**] for a survey.

5.2 Universally consistent predictions: bounded Y

5.2.1 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 [**?**]). 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

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/c} \tag{5.2}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} \ , \tag{5.3}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{5.4}$$

The prediction strategy g is defined by

$$g_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(\mathbf{x}_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \ldots \tag{5.5}$$

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 5.1. (Györfi and Lugosi [**?**]) *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 [**?**].

Lemma 5.1. *Let $\tilde{h}_1, \tilde{h}_2, \dots$ 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

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

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

$$\left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(\mathbf{x}_1^t, y_1^{t-1})\right)\right]^2 \leq -c \ln \frac{W_{t+1}}{W_t} \tag{5.6}$$

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 5.1 is known as Breiman's generalized ergodic theorem, see also Algoet [**?**] and Györfi et al. [**?**].

Lemma 5.2. (Breiman [**?**]). *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 5.1. Because of (5.1), 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 5.2, 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 [**?**]) 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 5.1,

$$L_n(g) \le \inf_{k,\ell}\left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n}\right), \tag{5.7}$$

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. □

5.2.2 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

$$\lim_{\ell\to\infty} r_{k,\ell} = 0, \tag{5.8}$$

and

$$\lim_{\ell\to\infty} r'_{k,\ell} = 0. \tag{5.9}$$

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

$$h_n^{(k,\ell)}(\mathbf{x}_1^n, y_1^{n-1}) = \frac{\sum_{\{t\in J_n^{(k,\ell)}\}} y_t}{|J_n^{(k,\ell)}|}, \qquad n > k+1, \tag{5.10}$$

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. (5.2), (5.3), (5.4) and (5.5)).

Theorem 5.2. *Suppose that (5.8) and (5.9) 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* $|Y_0| \leq B$.

5.2.3 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

$$\lim_{\ell \to \infty} p_\ell = 0\,, \tag{5.11}$$

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. (5.2), (5.3), (5.4) and (5.5)).

Theorem 5.3. *Suppose that (5.11) 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* $|Y_0| \le B$.

5.2.4 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 [**?**]. 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

$$\sum_{t=k+1}^{n-1} \left(\sum_{j=1}^{\ell} c_j \phi_j^{(k)}(\mathbf{x}_{t-k}^t, y_{t-k}^{t-1}) - y_t \right)^2 \tag{5.12}$$

if $n > k+1$, and the all-zero vector otherwise. It can be shown using a recursive technique (see e.g., Tsypkin [**?**], Györfi [**?**] and Györfi and Lugosi [**?**]) 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. (5.2), (5.3), (5.4) and (5.5)).

Theorem 5.4. (Györfi and Lugosi [**?**]) *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 $|Y_0| \le B$.

5.3 Universally consistent predictions: unbounded Y

5.3.1 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 5.2.1. 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 5.2.1.

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

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/\sqrt{t}} \tag{5.13}$$

and their normalized values

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{W_t} \ , \tag{5.14}$$

where

$$W_t = \sum_{i,j=1}^{\infty} w_{t,i,j} \ . \tag{5.15}$$

The prediction strategy g is defined by

$$g_t(\mathbf{x}_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(\mathbf{x}_1^t, y_1^{t-1}) \ , \qquad t = 1, 2, \ldots \tag{5.16}$$

Theorem 5.5. (GYÖRFI AND OTTUCSÁK [?]) *Assume that the conditions (a), (b), (c) and (d) of Theorem 5.1 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 5.3. (GYÖRFI AND OTTUCSÁK [?]) *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,

$$\begin{aligned}
&\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t} \\
&\leq \frac{1}{\eta_t}\ln\left(1-\eta_t\sum_{k=1}^{\infty} p_{t,k}\lambda_t(h^{(k)})+\frac{\eta_t^2}{2}\sum_{k=1}^{\infty} p_{t,k}\lambda_t^2(h^{(k)})\right) \\
&\leq -\sum_{k=1}^{\infty} p_{t,k}\lambda_t(h^{(k)})+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\lambda_t^2(h^{(k)}) \qquad (5.17) \\
&= -\sum_{k=1}^{\infty} p_{t,k}\lambda(h_t^{(k)},Y_t)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\lambda_t^2(h^{(k)}) \\
&\leq -\lambda\left(\sum_{k=1}^{\infty} p_{t,k}h_t^{(k)},Y_t\right)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\lambda_t^2(h^{(k)}) \qquad (5.18) \\
&= -\lambda_t(g)+\frac{\eta_t}{2}\sum_{k=1}^{\infty} p_{t,k}\lambda_t^2(h^{(k)}) \qquad (5.19)
\end{aligned}$$

where (5.17) follows from the fact that $\ln(1+x) \leq x$ for all $x > -1$ and in (5.18) we used the convexity of the loss $\lambda(h, y)$ in its first argument h. From (5.19) 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 5.5. Because of (5.1), 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 5.1, 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 5.2, 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 5.1, 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 5.3 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

$$\begin{aligned}
L_n(g) \le \inf_{k,\ell}&\left(L_n(h^{(k,\ell)}) - \frac{2\ln q_{k,\ell}}{\sqrt{n}}\right) \\
&+\frac{1}{2n}\sum_{t=1}^{n}\frac{1}{\sqrt{t}}\sum_{k,\ell=1}^{\infty}p_{t,k,\ell}\big(h^{(k,\ell)}(\mathbf{X}_1^t, Y_1^{t-1}) - Y_t\big)^4. \qquad (5.20)
\end{aligned}$$

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. □

5.3.2 Kernel-based prediction strategies

Apply the notations of Section 5.2.2. 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. (5.13), (5.14), (5.15) and (5.16)).

Theorem 5.6. (Biau et al [?]) *Suppose that (5.8) and (5.9) 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*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

5.3.3 Nearest neighbor-based prediction strategy

Apply the notations of Section 5.2.3. 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. (5.13), (5.14), (5.15) and (5.16)).

Theorem 5.7. (Biau et al [?]) *Suppose that (5.11) 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*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

5.3.4 Generalized linear estimates

Apply the notations of Section 5.2.4. 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. (5.13), (5.14), (5.15) and (5.16)).

Theorem 5.8. (BIAU ET AL [?]) *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*

$$\mathbb{E}\{Y_0^4\} < \infty.$$

5.3.5 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.

It is well known for gaussian time series that the best predictor is a linear function of the past:

$$\mathbb{E}\{Y_n \mid Y_{n-1}, Y_{n-2}, \ldots\} = \sum_{j=1}^{\infty} c_j^* Y_{n-j},$$

where the c_j^* minimize the criterion

$$\mathbb{E}\left\{ \left(\sum_{j=1}^{\infty} c_j Y_{n-j} - Y_n \right)^2 \right\}.$$

We apply the principle of generalized linear estimates to the prediction of gaussian time series by considering the special case

$$\phi_j^{(k)}(y_{n-k}^{n-1}) = y_{n-j} \mathbb{I}_{\{1 \leq j \leq k\}},$$

i.e.,

$$\tilde{h}_n^{(k)}(y_1^{n-1}) = \sum_{j=1}^{k} c_{n,j} y_{n-j}.$$

Once again, the coefficients $c_{n,j}$ are calculated according to the past observations y_1^{n-1} by minimizing the criterion:

$$\sum_{t=k+1}^{n-1} \left(\sum_{j=1}^{k} c_j y_{t-j} - y_t \right)^2$$

if $n > k$, and the all-zero vector otherwise.

We set

$$h_n^{(k)}(y_1^{n-1}) = T_{\min\{n^\delta, k\}} \left(\tilde{h}_n^{(k)}(y_1^{n-1}) \right),$$

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 5.9. (Biau et al [?]) *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 5.1. (Biau et al [?]) *Under the conditions of Theorem 5.9,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \left(\mathbb{E}\{Y_t \mid Y_1^{t-1}\} - g(Y_1^{t-1}) \right)^2 = 0 \quad \textit{almost surely.}$$

Corollary 5.1 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

$$\lim_{n\to\infty} \left(\mathbb{E}\{Y_n | Y_1^{n-1}\} - g(Y_1^{n-1}) \right) = 0 \quad \text{almost surely} \tag{5.21}$$

for all stationary and ergodic gaussian processes.

Chapter 6

Pattern Recognition

6.1 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

$$R(g^*) = \min_{g:\mathbb{R}^d \to \{1,2,\ldots,M\}} R(g), \tag{6.1}$$

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,

$$\sum_{y=1}^{m} C(y, g^*(\mathbf{X})) P_y(\mathbf{X}) \leq \sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X}). \tag{6.2}$$

Theorem 6.1. *For any decision function g, we have that*

$$R(g^*) \leq R(g).$$

PROOF. For a decision function g, let's calculate the risk.

$$\begin{aligned}
R(g) &= \mathbb{E}\{C(Y, g(\mathbf{X}))\} \\
&= \mathbb{E}\{\mathbb{E}\{C(Y, g(\mathbf{X})) \mid \mathbf{X}\}\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{m}\sum_{y'=1}^{M} C(y, y') \mathbb{P}\{Y = y, g(\mathbf{X}) = y' \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{m}\sum_{y'=1}^{M} C(y, y') \mathbb{I}_{\{g(\mathbf{X})=y'\}} \mathbb{P}\{Y = y \mid \mathbf{X}\}\right\} \\
&= \mathbb{E}\left\{\sum_{y=1}^{M} C(y, g(\mathbf{X})) P_y(\mathbf{X})\right\}.
\end{aligned}$$

(6.2) 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 6.1 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 6.1 illustrates the Bayes decision, while the red area in Figure 6.2 is equal to the Bayes error probability.

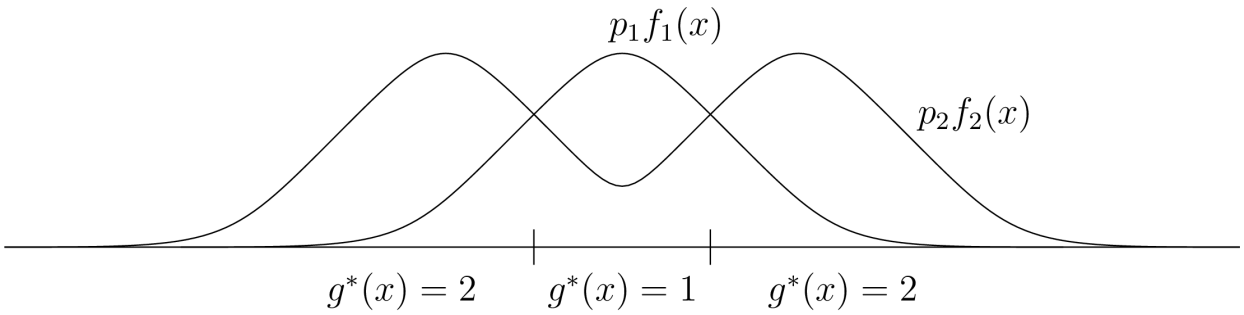

Figure 6.1: Bayes decision.

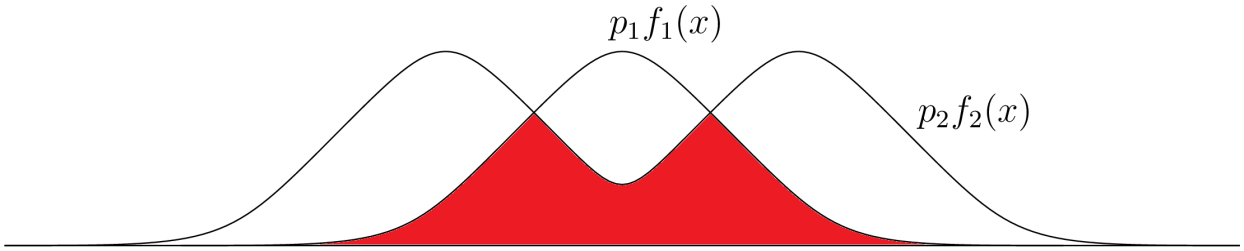

Figure 6.2: Bayes error probability.

6.2 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 6.1. *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 6.1 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

$$g_n(\mathbf{x}) = \arg \max_{1 \le y \le M} m_n^{(y)}(\mathbf{x}) \tag{6.3}$$

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 [**?**].)

6.3 Pattern recognition for time series

In this section we apply the ideas of Chapter 5 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 [**?**] proved that for any classification strategy f and stationary ergodic process $\{(\mathbf{X}_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n \to \infty} R_n(f) \ge R^* \quad \text{a.s.}, \tag{6.4}$$

where

$$R^* = \mathbb{E}\Big\{\min\big(\mathbb{P}\{Y_0 = 1|\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\}, \mathbb{P}\{Y_0 = 0|\mathbf{X}_{-\infty}^0, Y_{-\infty}^{-1}\}\big)\Big\},$$

therefore the following definition is meaningful:

Definition 6.1. *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 5.2.1 or 5.2.2 or 5.2.3 or 5.2.4, 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 6.2. (Györfi and Ottucsák [?]) *Assume that the conditions of Theorems 5.1 or 5.2 or 5.3 or 5.4 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 6.2 we derive a corollary of Theorem 5.1, which shows that asymptotically, the predictor g_t defined by (5.5) 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 6.1. *Under the conditions of Theorem 5.1*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \big(\mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\big)^2 = 0 \qquad \textit{almost surely.}$$

PROOF. By Theorem 5.1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right)^2 = L^* \qquad \text{almost surely.}$$

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

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|Y_{-\infty}^{i-1}\} - g_i(\mathbf{X}_1^i, Y_1^{i-1})\right) = 0. \tag{6.5}$$

almost surely. But this is a straightforward consequence of Kolmogorov's classical strong law of large numbers for martingale differences due to Chow [**?**] (see also Stout [**?**, Theorem 3.3.1]). It states that if $\{Z_i\}$ is a martingale difference sequence with

$$\sum_{n=1}^{\infty}\frac{\mathbb{E}Z_n^2}{n^2} < \infty, \tag{6.6}$$

then

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}Z_i = 0 \qquad \text{almost surely.}$$

Thus, (6.5) is implied by Chow's theorem since the martingale differences $Z_i = \left(Y_i - \mathbb{E}\{Y_i|\mathbf{X}_{-\infty}^i, Y_{-\infty}^{i-1}\}\right)\left(\mathbb{E}\{Y_i|\right.$ are bounded by $4B^2$. □

PROOF OF THEOREM 6.2 Because of (6.4) we have to show that

$$\limsup_{n\to\infty} R_n(f) \le R^* \quad \text{a.s.}$$

By Corollary 6.1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{E}\{Y_t \mid \mathbf{X}_{-\infty}^{t}, Y_{-\infty}^{t-1}\} - g_t(\mathbf{X}_1^t, Y_1^{t-1})\right)^2 = 0 \qquad \text{a.s.} \tag{6.7}$$

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

$$\limsup_{n\to\infty}(\bar{R}_n(f) - \bar{R}_n(f^*)) \le 0 \qquad \text{a.s.}$$

Lemma 6.1 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 (6.7). □

Chapter 7

Density Estimation

7.1 Why and how 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. The Scheffé Theorem below shows that the total variation is the half of the L_1 distance of the corresponding densities.

Theorem 7.1. *(*Scheffé [?]*) 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

$$L_1(f, g) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x} \tag{7.1}$$

*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 7.1 is equal to the L_1 distance between the densities f and g.

The Scheffé Theorem implies an equivalent definition of the total variation:

$$V(\mu, \nu) = \frac{1}{2} \sup_{\{A_j\}} \sum_j |\mu(A_j) - \nu(A_j)|, \tag{7.2}$$

where the supremum is taken over all finite Borel measurable partitions $\{A_j\}$.

For the distribution of $\mathbf{X}$, introduce the notation

$$\mu(A) = \mathbb{P}\{\mathbf{X} \in A\}.$$

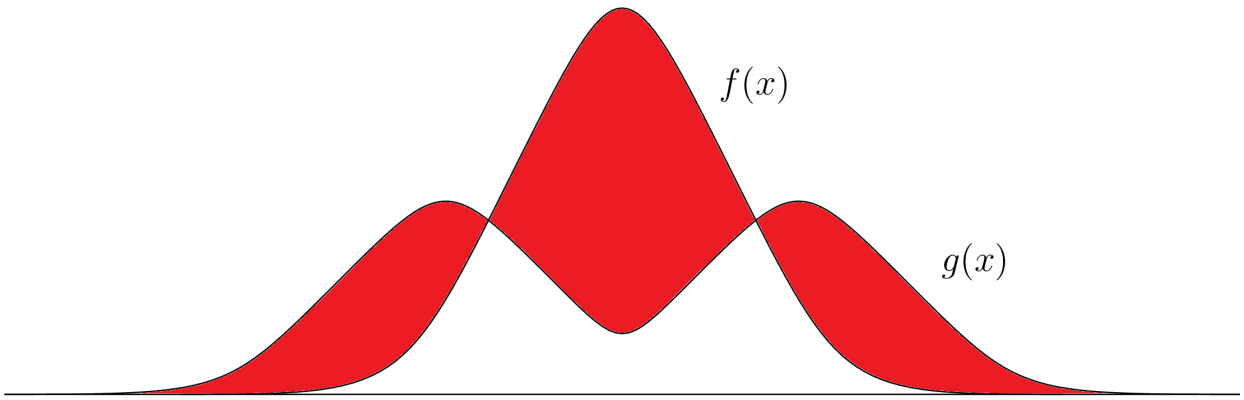

Figure 7.1: L_1 error.

In the sequel assume that the distribution μ has a density, which is denoted by f:

$$\mu(A) = \int_A f(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

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.

7.2 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 7.2. *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

$$\int |f_n(\mathbf{x}) - f(\mathbf{x})|\, \mathrm{d}\mathbf{x} \leq \underbrace{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})|\, \mathrm{d}\mathbf{x}}_{\text{variation term}} + \underbrace{\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\, \mathrm{d}\mathbf{x}}_{\text{bias}}.$$

The histogram is constant on a cell, therefore

$$\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} = \sum_j \int\limits_{A_{nj}} |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} = \sum_j |\mu_n(A_{nj}) - \mu(A_{nj})|.$$

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

$$\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \leq \sum_{j=1}^{M_n} |\mu_n(A_{nj}) - \mu(A_{nj})| + \mu_n(S_n^c) + \mu(S_n^c),$$

therefore the Cauchy-Schwarz and the Jensen inequalities imply that

$$\begin{aligned}
\mathbb{E}\left\{\int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x}\right\} &\leq \sum_{j=1}^{M_n} \mathbb{E}\{|\mu_n(A_{nj}) - \mu(A_{nj})|\} + 2\mu(S_n^c) \\
&\leq \sum_{j=1}^{M_n} \sqrt{\mathbb{E}\{|\mu_n(A_{nj}) - \mu(A_{nj})|^2\}} + 2\mu(S^c) \\
&\leq \sum_{j=1}^{M_n} \sqrt{\frac{\mu(A_{nj})}{n}} + 2\mu(S^c) \\
&\leq \sqrt{\frac{M_n}{n}} + 2\mu(S^c) \qquad (7.3) \\
&\to 2\mu(S^c).
\end{aligned}$$

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

$$\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} \leq \int\int |f(\mathbf{z}) - f(\mathbf{x})| K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \tag{7.4}$$

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

$$\begin{aligned}
& \int |f(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \\
= & \int \left| f(\mathbf{x}) - \int f(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
\leq & \int |f(\mathbf{x}) - \tilde{f}(\mathbf{x})| \, \mathrm{d}\mathbf{x} + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
& + \int \left| \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} - \int f(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
\leq & \varepsilon + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} \\
& + \int \left(\int |\tilde{f}(\mathbf{z}) - f(\mathbf{z})| K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{x} \right) \, \mathrm{d}\mathbf{z} \\
= & \varepsilon + \int \left| \tilde{f}(\mathbf{x}) - \int \tilde{f}(\mathbf{z}) K_n(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{z} \right| \, \mathrm{d}\mathbf{x} + \int |\tilde{f}(\mathbf{z}) - f(\mathbf{z})| \, \mathrm{d}\mathbf{z} \\
\to & \ 2\varepsilon.
\end{aligned}$$

□

Theorem 7.3. *Assume that 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

$$\mathbb{E} \int |f - f_n| \leq \frac{c_1}{\sqrt{n h_n^d}} + c_2 h_n,$$

so for the choice

$$h_n = c_3 n^{-\frac{1}{d+2}}$$

we have that

$$\mathbb{E} \int |f - f_n| \leq c_4 n^{-\frac{1}{d+2}}.$$

Proof. For the variation term, (7.3) implies that

$$\mathbb{E} \left\{ \int |f_n(\mathbf{x}) - \mathbb{E}f_n(\mathbf{x})| \, \mathrm{d}\mathbf{x} \right\} \leq \sqrt{\frac{M_n}{n}} \leq \sqrt{\frac{\lambda(S)}{n h_n^d}}.$$

Concerning the bias term, (7.4) implies that

$$\begin{aligned}\int |\mathbb{E}f_n(\mathbf{x}) - f(\mathbf{x})|\,\mathrm{d}\mathbf{x} &\leq \int\int |f(\mathbf{z}) - f(\mathbf{x})| K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\ &\leq \int\int C\|\mathbf{z}-\mathbf{x}\| K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\ &\leq \int\int Ch_n\sqrt{d} K_n(\mathbf{x},\mathbf{z})\,\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} \\ &\leq Ch_n\sqrt{d}\lambda(S).\end{aligned}$$

□

7.3 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

$$f_n(\mathbf{x}) = \frac{1}{nh_n^d}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right).$$

Theorem 7.4. *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

$$\lim_{n\to\infty} \mathbb{E}\int |f(\mathbf{x}) - f_n(\mathbf{x})|\,\mathrm{d}\mathbf{x} = 0.$$

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

$$K(\mathbf{x}) = c(1 - \|\mathbf{x}\|^2)\mathbb{I}_{\{\|\mathbf{x}\| \leq 1\}}.$$

Theorem 7.5. *Assume that f is zero outside a sphere S and it is differentiable with Lipschitz continuous gradient, i.e.,*

$$\|f'(\mathbf{x}) - f'(\mathbf{z})\| \leq C\|\mathbf{x} - \mathbf{z}\|.$$

Then for the kernel estimate f_n, one has that

$$\mathbb{E}\int |f - f_n| \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2,$$

so for the choice

$$h_n = c_3 n^{-\frac{1}{d+4}}$$

we have that

$$\mathbb{E}\int |f - f_n| \leq c_4 n^{-\frac{2}{d+4}}.$$

For further reading on L_1 density estimation, the books Devroye, Györfi [**?**], Devroye [**?**] and Devroye, Lugosi [**?**] are suggested.

Chapter 8

Testing Simple Hypotheses

8.1 α-level tests

In this section we consider decision problems, where the consequences of the various errors are very much different. For example, if in a diagnostic problem $Y = 0$ means that the patient is OK, while $Y = 1$ means that the patient is ill, then for $Y = 0$ the false decision is that the patient is ill, which implies some superfluous medical treatment, while for $Y = 1$ the false decision is that the illness is not detected, and the patient's state may become worse. A similar situation happens for radar detection.

The event $Y = 0$ is called null hypothesis and is denoted by $\mathcal{H}_0$, and the event $Y = 1$ is called alternative hypothesis and is denoted by $\mathcal{H}_1$. The decision, the test is formulated by a set $A \subset \mathbb{R}^d$, called acceptance region such that accept $\mathcal{H}_0$ if $\mathbf{X} \in A$, otherwise reject $\mathcal{H}_0$, i.e., accept $\mathcal{H}_1$. The set A^c is called critical region.

Let P_0 and P_1 be the probability distributions of $\mathbf{X}$ under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. There are two types of errors:

- Error of the first kind, if under the null hypothesis $\mathcal{H}_0$ we reject $\mathcal{H}_0$. This error is $P_0(A^c)$.
- Error of the second kind, if under the alternative hypothesis $\mathcal{H}_1$ we reject $\mathcal{H}_1$. This error is $P_1(A)$.

Obviously, one decreases the error of the first kind $P_0(A^c)$ if the error of the second kind $P_1(A)$ increases. We can formulate the optimization problem such that minimize the error of the second kind under the condition that the error of the first kind is at most $0 < \alpha < 1$:

$$\min_{A:\, P_0(A^c) \le \alpha} P_1(A). \tag{8.1}$$

In order to solve this problem the Neyman-Pearson Lemma plays an important role.

Theorem 8.1. (NEYMAN, PEARSON [?]) *Assume that the distributions P_0 and P_1 have densities f_0 and f_1:*

$$P_0(B) = \int_B f_0(\mathbf{x})d\mathbf{x} \quad \textit{and} \quad P_1(B) = \int_B f_1(\mathbf{x})d\mathbf{x}.$$

For a $\gamma > 0$, put

$$A_\gamma = \{\mathbf{x} : f_0(\mathbf{x}) \geq \gamma f_1(\mathbf{x})\}.$$

If for any set A

$$P_0(A^c) \leq P_0(A_\gamma^c)$$

then

$$P_1(A) \geq P_1(A_\gamma).$$

PROOF. Because of the condition of the theorem, we have the following chain of inequalities:

$$\begin{aligned} P_0(A^c) &\leq P_0(A_\gamma^c) \\ P_0(A^c \cap A_\gamma) + P_0(A^c \cap A_\gamma^c) &\leq P_0(A \cap A_\gamma^c) + P_0(A^c \cap A_\gamma^c) \\ \int_{A^c \cap A_\gamma} f_0(x)dx &\leq \int_{A \cap A_\gamma^c} f_0(x)dx. \end{aligned}$$

The definition of A_γ implies that

$$\gamma \int_{A^c \cap A_\gamma} f_1(\mathbf{x})d\mathbf{x} \leq \int_{A^c \cap A_\gamma} f_0(\mathbf{x})d\mathbf{x} \leq \int_{A \cap A_\gamma^c} f_0(\mathbf{x})d\mathbf{x} \leq \gamma \int_{A \cap A_\gamma^c} f_1(\mathbf{x})d\mathbf{x},$$

therefore using the previous chain of derivations in a reverse order we get that

$$P_1(A^c) \leq P_1(A_\gamma^c).$$

□

In Figure 8.1 the blue area illustrates the error of the first kind, while the red area is the error of the second kind.

If for an $0 < \alpha < 1$ there is a $\gamma = \gamma(\alpha)$, which solves the equation

$$P_0(A_\gamma^c) = \alpha,$$

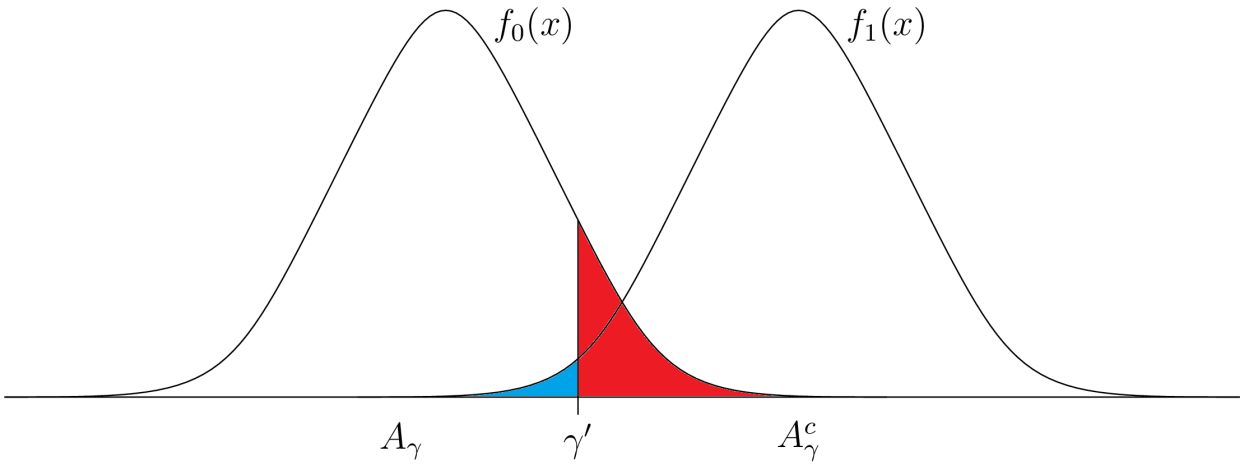

Figure 8.1: Error of the first and second kind.

then the Neyman-Pearson Lemma implies that in order to solve the problem (8.1), it is enough to search for set of form A_γ, i.e.,

$$\min_{A:\, P_0(A^c)\le\alpha} P_1(A) = \min_{A_\gamma:\, P_0(A_\gamma^c)\le\alpha} P_1(A_\gamma).$$

Then A_γ is called the *most powerful α-level test.*

Because of the Neyman-Pearson Lemma, we introduce the likelihood ratio statistic

$$T(\mathbf{X}) = \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})},$$

and so the null hypothesis $\mathcal{H}_0$ is accepted if $T(\mathbf{X}) \ge \gamma$.

EXAMPLE 1. As an illustration of the Neyman-Pearson Lemma, consider the example of an experiment, where the null hypothesis is that the components of $\mathbf{X}$ are i.i.d. normal with mean $m = m_0 > 0$ and with variance σ^2, while under the alternative hypothesis the components of $\mathbf{X}$ are i.i.d. normal with mean $m_1 = 0$ and with the same variance σ^2. Then

$$f_0(\mathbf{x}) = f_0(x_1, \dots, x_d) = \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - m)^2}{2\sigma^2}} \right)$$

and

$$f_1(\mathbf{x}) = f_1(x_1, \dots, x_d) = \prod_{i=1}^{d} \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x_i^2}{2\sigma^2}} \right)$$

and

$$\frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} \geq \gamma$$

means that

$$-\sum_{i=1}^{d} \frac{(X_i - m)^2}{2\sigma^2} + \sum_{i=1}^{d} \frac{X_i^2}{2\sigma^2} \geq \ln \gamma,$$

or equivalently,

$$\sum_{i=1}^{d} (2X_i m - m^2) \geq 2\sigma^2 \ln \gamma.$$

This test accepts the null hypothesis if

$$\frac{1}{d} \sum_{i=1}^{d} X_i \geq \frac{2\sigma^2 \ln \gamma / d + m^2}{2m} = \frac{\sigma^2 \ln \gamma}{dm} + \frac{m}{2} =: \gamma'.$$

The test is based on the linear statistic $\sum_{i=1}^{d} X_i/d$, and the question left is how to choose the critical value γ', for which it is an α-level test, i.e., the error of the first kind is α:

$$\mathbb{P}_0 \left\{ \frac{1}{d} \sum_{i=1}^{d} X_i \leq \gamma' \right\} = \alpha.$$

Under the null hypothesis, the distribution of $\frac{1}{d} \sum_{i=1}^{d} X_i$ is normal with mean m and with variance σ^2/d, therefore

$$\mathbb{P}_0 \left\{ \frac{1}{d} \sum_{i=1}^{d} X_i \leq \gamma' \right\} = \Phi \left(\frac{\gamma' - m}{\sigma/\sqrt{d}} \right),$$

where Φ denotes the standard normal distribution function, and so the critical value γ' of an α-level test solves the equation

$$\Phi \left(-\frac{m - \gamma'}{\sigma/\sqrt{d}} \right) = \alpha,$$

i.e.,

$$\gamma' = m - \Phi^{-1}(1 - \alpha)\sigma/\sqrt{d}.$$

REMARK 1. In many situations, when d is large enough, one can refer to the central limit theorem such that the log-likelihood ratio

$$\ln \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}$$

is asymptotically normal. The argument of Example 1 can be extended if under $\mathcal{H}_0$, the log-likelihood ratio is approximately normal with mean m_0 and with variance σ_0^2. Let the test be defined such that it accepts $\mathcal{H}_0$ if

$$\ln \frac{f_0(\mathbf{X})}{f_1(\mathbf{X})} \geq \gamma',$$

where

$$\gamma' = m_0 - \Phi^{-1}(1-\alpha)\sigma_0.$$

Then this test is approximately an α-level test.

8.2 ϕ-divergences

In the analysis of repeated observations the divergences between distribution play an important role. Imre Csiszár [**?**] introduced the concept of ϕ-divergences. Let $\phi : (0, \infty) \to \mathbb{R}$ be a convex function, extended on $[0, \infty)$ by continuity such that $\phi(1) = 0$. For the probability distributions μ and ν, let λ be a σ-finite dominating measure of μ and ν, for example, $\lambda = \mu + \nu$. Introduce the notations

$$f = \frac{d\mu}{d\lambda}$$

and

$$g = \frac{d\nu}{d\lambda}.$$

Then the *ϕ-divergence* of μ and ν is defined by

$$D_\phi(\mu, \nu) = \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}). \tag{8.2}$$

The Jensen inequality implies the most important property of the ϕ-divergences:

$$D_\phi(\mu, \nu) = \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \geq \phi\left(\int_{\mathbb{R}^d} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\lambda(d\mathbf{x})\right) = \phi(1) = 0.$$

It means that $D_\phi(\mu, \nu) \geq 0$ and if $\mu = \nu$ then $D_\phi(\mu, \nu) = 0$. If, in addition, ϕ is strictly convex at 1 then $D_\phi(\mu, \nu) = 0$ iff $\mu = \nu$.

Next we show some examples.

- For
$$\phi_1(t) = |t-1|,$$
we get the L_1 *distance*
$$D_{\phi_1}(\mu,\nu) = \int_{\mathbb{R}^d} |f(\mathbf{x}) - g(\mathbf{x})| \lambda(d\mathbf{x}).$$

- For
$$\phi_2(t) = (\sqrt{t}-1)^2,$$
we get the *squared Hellinger distance*
$$\begin{aligned} D_{\phi_2}(\mu,\nu) &= \int_{\mathbb{R}^d} \left(\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})}\right)^2 \lambda(d\mathbf{x}) \\ &= 2\left(1 - \int_{\mathbb{R}^d} \sqrt{f(\mathbf{x})g(\mathbf{x})}\lambda(d\mathbf{x})\right). \end{aligned}$$

- For
$$\phi_3(t) = -\ln t,$$
we get the *I-divergence* (called also relative entropy or Kullback-Leibler divergence)
$$I(\mu,\nu) = D_{\phi_3}(\mu,\nu) = \int_{\mathbb{R}^d} \ln\left(\frac{g(\mathbf{x})}{f(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

- For
$$\phi_4(t) = (t-1)^2,$$
we get the χ^2-*divergence*
$$\chi^2(\mu,\nu) = D_{\phi_4}(\mu,\nu) = \int_{\mathbb{R}^d} \frac{(f(\mathbf{x}) - g(\mathbf{x}))^2}{g(\mathbf{x})}\lambda(d\mathbf{x}).$$

An equivalent definition of the ϕ-divergence is
$$D_\phi(\mu,\nu) = \sup_{\mathcal{P}} \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right)\nu(A_j), \tag{8.3}$$
where the supremum is taken over all finite Borel measurable partitions $\mathcal{P} = \{A_j\}$ of $\mathbb{R}^d$.

The main reasoning of this equivalence is that for any partition $\mathcal{P} = \{A_j\}$, the Jensen inequality implies that

$$\begin{aligned} D_\phi(\mu,\nu) &= \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\ &= \sum_j \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}) \\ &= \sum_j \frac{1}{\nu(A_j)} \int_{A_j} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x})\nu(A_j) \\ &\geq \sum_j \phi\left(\frac{1}{\nu(A_j)} \int_{A_j} \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x})\lambda(d\mathbf{x})\right) \nu(A_j) \\ &= \sum_j \phi\left(\frac{\mu(A_j)}{\nu(A_j)}\right) \nu(A_j). \end{aligned} \tag{8.4}$$

The sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is called nested if any cell $A \in \mathcal{P}_{n+1}$ is a subset of a cell $A' \in \mathcal{P}_n$. Next we show that for nested sequence of partitions

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right) \nu(A) \uparrow .$$

Again, this property is the consequence of the Jensen inequality:

$$\begin{aligned} \sum_{A'\in\mathcal{P}_{n+1}} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \nu(A') &= \sum_{A\in\mathcal{P}_n} \left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \nu(A') \right) \\ &= \sum_{A\in\mathcal{P}_n} \left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \phi\left(\frac{\mu(A')}{\nu(A')}\right) \frac{\nu(A')}{\nu(A)} \right) \nu(A) \\ &\geq \sum_{A\in\mathcal{P}_n} \phi\left(\sum_{A'\in\mathcal{P}_{n+1}, A'\subset A} \frac{\mu(A')}{\nu(A')} \frac{\nu(A')}{\nu(A)} \right) \nu(A) \\ &= \sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right) \nu(A). \end{aligned}$$

It implies that there is a nested sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ such that

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) \uparrow \sup_{\mathcal{P}_n} \sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A).$$

The sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is called asymptotically fine if for any sphere S centered at the origin

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n, A\cap S\neq 0} diam(A) = 0. \tag{8.5}$$

One can show that if the nested sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine then

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) \uparrow \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

This final step will be verified in the particular case of L_1 distance, cf. (9.7). In general, we may introduce a cell wise constant approximation of $\frac{f(\mathbf{x})}{g(\mathbf{x})}$:

$$F_n(\mathbf{x}) := \frac{\mu(A)}{\nu(A)} \text{ if } \mathbf{x}\in A.$$

Thus,

$$\sum_{A\in\mathcal{P}_n} \phi\left(\frac{\mu(A)}{\nu(A)}\right)\nu(A) = \int_{\mathbb{R}^d} \phi\left(F_n(\mathbf{x})\right) g(\mathbf{x})\lambda(d\mathbf{x})$$

and

$$F_n(\mathbf{x}) \to \frac{f(\mathbf{x})}{g(\mathbf{x})}$$

for almost all $\mathbf{x} \mod \lambda$ with $g(\mathbf{x}) > 0$ such that

$$\int_{\mathbb{R}^d} \phi\left(F_n(\mathbf{x})\right) g(\mathbf{x})\lambda(d\mathbf{x}) \to \int_{\mathbb{R}^d} \phi\left(\frac{f(\mathbf{x})}{g(\mathbf{x})}\right) g(\mathbf{x})\lambda(d\mathbf{x}).$$

8.3 Repeated observations

The error probabilities can be decreased if instead of an observation vector $\mathbf{X}$, we are given n vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ such that under $\mathcal{H}_0$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed (i.i.d.) with distribution P_0, while under $\mathcal{H}_1$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. with distribution P_1. In this case the likelihood ratio statistic is of form

$$T(\mathbf{X}) = \frac{f_0(\mathbf{X}_1)\cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1)\cdot \ldots \cdot f_1(\mathbf{X}_n)}.$$

The Stein Lemma below says that there are tests, for which both the error of the first kind α_n and the error of the second kind β_n tend to 0, if $n \to \infty$.

In order to formulate the Stein Lemma, we remember the *I-divergence*

$$I(P_0, P_1) = D(f_0, f_1) = \int_{\mathbb{R}^d} f_0(\mathbf{x}) \ln \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} d\mathbf{x}. \tag{8.6}$$

.

Theorem 8.2. (CF. CHERNOFF [**?**]) *For any $0 < \delta < D(f_0, f_1)$, there is a test such that the error of the first kind*

$$\alpha_n \to 0,$$

and for the error of the second kind

$$\beta_n \leq e^{-n(D(f_0,f_1)-\delta)} \to 0.$$

PROOF. Construct a test such that accept the null hypothesis $\mathcal{H}_0$ if

$$\frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq e^{n(D(f_0,f_1)-\delta)},$$

or equivalently

$$\frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \geq D(f_0, f_1) - \delta.$$

Under $\mathcal{H}_0$, the strong law of large numbers implies that

$$\frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \to D(f_0, f_1)$$

almost surely (a.s.), therefore for the error of the first kind α_n, we get that

$$\alpha_n = \mathbb{P}_0 \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} < D(f_0, f_1) - \delta \right\} \to 0.$$

Concerning the error of the second kind β_n we have the following simple bound:

$$\begin{aligned}
& \beta_n \\
= \; & \mathbb{P}_1 \left\{ \frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\} \\
= \; & \int_{\left\{ \frac{f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n)}{f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\}} f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n) d\mathbf{x}_1, \ldots, d\mathbf{x}_n \\
\leq \; & e^{-n(D(f_0,f_1)-\delta)} \int_{\left\{ \frac{f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n)}{f_1(\mathbf{x}_1) \cdot \ldots \cdot f_1(\mathbf{x}_n)} \geq e^{n(D(f_0,f_1)-\delta)} \right\}} f_0(\mathbf{x}_1) \cdot \ldots \cdot f_0(\mathbf{x}_n) d\mathbf{x}_1, \ldots, d\mathbf{x}_n \\
\leq \; & e^{-n(D(f_0,f_1)-\delta)}.
\end{aligned}$$

□

The critical value of the test in the proof of the Stein Lemma used the I-divergence $D(f_0, f_1)$. Without knowing $D(f_0, f_1)$, the Chernoff Lemma below results in exponential rate of convergence of the errors.

Theorem 8.3. (CHERNOFF [?]). *Construct a test such that accept the null hypothesis $\mathcal{H}_0$ if*

$$\frac{f_0(\mathbf{X}_1) \cdot \ldots \cdot f_0(\mathbf{X}_n)}{f_1(\mathbf{X}_1) \cdot \ldots \cdot f_1(\mathbf{X}_n)} \geq 1,$$

or equivalently

$$\sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} \geq 0.$$

(This test is called maximum likelihood test.) Then

$$\alpha_n \leq \left(\inf_{s>0} \int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s} d\mathbf{x} \right)^n$$

and

$$\beta_n \leq \left(\inf_{s>0} \int_{\mathbb{R}^d} f_0(\mathbf{x})^s f_1(\mathbf{x})^{1-s} d\mathbf{x} \right)^n.$$

PROOF. Apply the Chernoff bounding technique such that for any $s > 0$ the Markov

inequality implies that

$$\begin{aligned}
\alpha_n &= \mathbb{P}_0\left\{\sum_{i=1}^{n} \ln \frac{f_0(\mathbf{X}_i)}{f_1(\mathbf{X}_i)} < 0\right\} \\
&= \mathbb{P}_0\left\{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)} > 0\right\} \\
&= \mathbb{P}_0\left\{e^{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}} > 1\right\} \\
&\leq \mathbb{E}_0\left\{e^{s\sum_{i=1}^{n} \ln \frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}}\right\} \\
&= \mathbb{E}_0\left\{\prod_{i=1}^{n}\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\}.
\end{aligned}$$

Under $\mathcal{H}_0$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., therefore

$$\begin{aligned}
\alpha_n &\leq \mathbb{E}_0\left\{\prod_{i=1}^{n}\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\} \\
&= \prod_{i=1}^{n}\mathbb{E}_0\left\{\left(\frac{f_1(\mathbf{X}_i)}{f_0(\mathbf{X}_i)}\right)^s\right\} \\
&= \mathbb{E}_0\left\{\left(\frac{f_1(\mathbf{X}_1)}{f_0(\mathbf{X}_1)}\right)^s\right\}^n \\
&= \left(\int_{\mathbb{R}^d}\left(\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}\right)^s f_0(\mathbf{x})d\mathbf{x}\right)^n.
\end{aligned}$$

Since $s > 0$ is arbitrary, the first half of the lemma is proved, and the proof of the second half is similar. □

REMARK 2. The Chernoff Lemma results in exponential rate of convergence if

$$\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} < 1$$

and

$$\inf_{s>0}\int_{\mathbb{R}^d} f_0(\mathbf{x})^s f_1(\mathbf{x})^{1-s}d\mathbf{x} < 1.$$

The Cauchy-Schwartz inequality implies that

$$\begin{aligned}\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} &\leq \int_{\mathbb{R}^d} f_1(\mathbf{x})^{1/2} f_0(\mathbf{x})^{1/2}d\mathbf{x} \\ &\leq \sqrt{\int_{\mathbb{R}^d} f_1(\mathbf{x})d\mathbf{x}\int_{\mathbb{R}^d} f_0(\mathbf{x})d\mathbf{x}} \\ &= 1,\end{aligned}$$

with equality in the second inequality if and only if $f_0 = f_1$. Morover, one can check that the function

$$g(s) := \int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x}$$

is convex such that $g(0) = 1$ and $g(1) = 1$, therefore

$$\inf_{s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x} = \inf_{1>s>0}\int_{\mathbb{R}^d} f_1(\mathbf{x})^s f_0(\mathbf{x})^{1-s}d\mathbf{x}.$$

The quantity

$$He(f_0, f_1) = \int_{\mathbb{R}^d} f_1(\mathbf{x})^{1/2} f_0(\mathbf{x})^{1/2}d\mathbf{x} \tag{8.7}$$

is called *Hellinger integral*. The previous derivations imply that

$$\alpha_n \leq He(f_0, f_1)^n$$

and

$$\beta_n \leq He(f_0, f_1)^n.$$

The squared Hellinger distance $D_{\phi_2}(\mu, \nu)$ was introduced in previous section. One can check that

$$D_{\phi_2}(\mu, \nu) = 2\left(1 - He(f_0, f_1)\right).$$

REMARK 3. Besides the concept of α-level consistency, there is a new kind of consistency, called *strong consistency*, meaning that both on $\mathcal{H}_0$ and on its complement the tests make a.s. no error after a random sample size. In other words, denoting by $\mathbb{P}_0$ (*resp.* $\mathbb{P}_1$) the probability distributions under the null hypothesis (*resp.* under the alternative), we have

$$\mathbb{P}_0\{\text{rejecting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1 \tag{8.8}$$

and

$$\mathbb{P}_1\{\text{accepting } \mathcal{H}_0 \text{ for only finitely many } n\} = 1. \tag{8.9}$$

Because of the Chernoff bound, both errors tend to 0 exponentially fast, so the Borel-Cantelli Lemma implies that the maximum likelihood test is strongly consistent. In a real life problem, for example, when we get the data sequentially, one gets data just once, and should make good inference for these data. Strong consistency means that the single sequence of inference is a.s. perfect if the sample size is large enough. This concept is close to the definition of discernability introduced by Dembo and Peres [**?**]. For a discussion and references, we refer the reader to Devroye and Lugosi [**?**].

Chapter 9

Testing Simple versus Composite Hypotheses

9.1 Total variation and I-divergence

If μ and ν are probability distributions on $\mathbb{R}^d$ ($d \geq 1$), then the *total variation distance* between μ and ν was defined by

$$V(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|,$$

where the supremum is taken over all Borel sets A. According to the Scheffé Theorem (Theorem 7.1), the total variation is the half of the L_1 distance of the corresponding densities.

The following inequality, called Pinsker's inequality, gives an upper bound to the total variation in terms of I-divergence:

Theorem 9.1. (Csiszár [?], Kullback [?] and Kemperman [?])

$$2\{V(\mu, \nu)\}^2 \leq I(\mu, \nu). \tag{9.1}$$

Proof. Applying the notations of the proof of the Scheffé Theorem, put

$$A^* = \{f > g\},$$

then the Scheffé Theorem implies that

$$V(\mu, \nu) = \mu(A^*) - \nu(A^*).$$

Moreover, from (8.4) we get that

$$I(\mu,\nu) \geq \mu(A^*) \ln \frac{\mu(A^*)}{\nu(A^*)} + (1-\mu(A^*)) \ln \frac{1-\mu(A^*)}{1-\nu(A^*)}$$

Introduce the notations

$$q = \nu(A^*) \text{ and } p = \mu(A^*) > q,$$

and

$$h_p(q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

then we have to prove that

$$2(p-q)^2 \leq h_p(q),$$

which follows from the facts on the derivative:

$$\begin{aligned} \frac{d}{dq}(h_p(q) - 2(p-q)^2) &= -\frac{p}{q} + \frac{1-p}{1-q} + 4(p-q) \\ &= -\frac{p-q}{q(1-q)} + 4(p-q) \\ &\leq 0. \end{aligned}$$

□

9.2 Large deviation of L_1 distance

Consider the sample of $\mathbb{R}^d$-valued random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ with *i.i.d.* components such that the common distribution is denoted by ν. For a fixed distribution μ, we consider the problem of testing hypotheses

$$\mathcal{H}_0 : \nu = \mu \text{ versus } \mathcal{H}_1 : \nu \neq \mu$$

by means of test statistics $T_n = T_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$.

For testing a simple hypothesis $\mathcal{H}_0$ that the distribution of the sample is μ, versus a composite alternative, Györfi and van der Meulen [**?**] introduced a related goodness of fit test statistic L_n defined as

$$L_n = \sum_{j=1}^{m_n} |\mu_n(A_{n,j}) - \mu(A_{n,j})|,$$

where μ_n denotes the empirical measure associated with the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, so that

$$\mu_n(A) = \frac{\#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}}{n}$$

for any Borel subset A, and $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ is a finite partition of $\mathbb{R}^d$.

Next we characterize the large deviation properties of L_n:

Theorem 9.2. (Beirlant, Devroye, Györfi and Vajda [?]). *Assume that*

$$\lim_{n\to\infty} \max_j \mu(A_{n,j}) = 0 \tag{9.2}$$

and

$$\lim_{n\to\infty} \frac{m_n \ln n}{n} = 0. \tag{9.3}$$

Then for all $0 < \epsilon < 2$

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{L_n > \epsilon\} = -g_L(\epsilon), \tag{9.4}$$

where

$$g_L(\epsilon) = \inf_{0<p<1-\epsilon/2} \left(p \ln \frac{p}{p+\epsilon/2} + (1-p) \ln \frac{1-p}{1-p-\epsilon/2} \right). \tag{9.5}$$

Biau and Györfi [?] provided an alternative derivation of $g_L(\varepsilon)$ and non-asymptotic upper bound.

Theorem 9.3. (Biau and Györfi [?]). *For any* $\epsilon > 0$,

$$\mathbb{P}\{L_n > \epsilon\} \le 2^{m_n} e^{-n\epsilon^2/2}.$$

Proof. By Scheffé's theorem for partitions

$$L_n = \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu(A)| = 2 \max_{A\in\sigma(\mathcal{P}_n)} (\mu_n(A) - \mu(A)),$$

where the class of sets $\sigma(\mathcal{P}_n)$ contains all sets obtained by unions of cells of $\mathcal{P}_n$. Therefore, for any $s > 0$, by the Markov inequality

$$\mathbb{P}\{L_n > \epsilon\} = \mathbb{P}\{L_n/2 > \epsilon/2\} = \mathbb{P}\{e^{nsL_n/2} > e^{ns\epsilon/2}\} \le \frac{\mathbb{E}\{e^{nsL_n/2}\}}{e^{ns\epsilon/2}}.$$

Moreover,

$$\begin{aligned}
\mathbb{E}\{e^{snL_n/2}\} &= \mathbb{E}\{\max_{A\in\sigma(\mathcal{P}_n)} e^{sn(\mu_n(A)-\mu(A))}\} \\
&\leq \sum_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn(\mu_n(A)-\mu(A))}\} \\
&\leq 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn(\mu_n(A)-\mu(A))}\} \\
&= 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{sn\mu_n(A)}\} e^{-sn\mu(A)}.
\end{aligned}$$

For any fixed Borel set A,

$$\mathbb{E}\{e^{sn\mu_n(A)}\} = \mathbb{E}\{e^{s\sum_{i=1}^n \mathbb{I}_{\mathbf{X}_i\in A}}\} = \prod_{i=1}^n \mathbb{E}\{e^{s\mathbb{I}_{\mathbf{X}_i\in A}}\} = \left(e^s\mu(A) + 1 - \mu(A)\right)^n.$$

Thus, for any $s > 0$, we have that

$$\mathbb{P}\{L_n > \epsilon\} \leq 2^{m_n} \left[\max_{A\in\sigma(\mathcal{P}_n)} e^{-s(\mu(A)+\epsilon/2)} \left(e^s\mu(A) + 1 - \mu(A)\right)\right]^n.$$

For fixed set A, choose

$$e^s = \frac{\mu(A)+\epsilon/2}{1-(\mu(A)+\epsilon/2)} \frac{1-\mu(A)}{\mu(A)},$$

then for this s,

$$\begin{aligned}
e^{-s(\mu(A)+\epsilon/2)} \left(e^s\mu(A) + 1 - \mu(A)\right) &= e^{-I((\mu(A)+\epsilon/2,1-\mu(A)-\epsilon/2),(\mu(A),1-\mu(A)))} \\
&\leq e^{-\epsilon^2/2},
\end{aligned}$$

where the last step follows from the Pinsker inequality. Thus,

$$\mathbb{P}\{L_n > \epsilon\} \leq 2^{m_n} e^{-n\epsilon^2/2}.$$

□

REMARK 5. As a special case of relative frequencies, in the previous proof the Chernoff inequality

$$\mathbb{P}\{\mu_n(A) - \mu(A) \geq \epsilon\} \leq e^{-nI((\mu(A)+\epsilon/2,1-\mu(A)-\epsilon/2),(\mu(A),1-\mu(A)))}$$

and the Hoeffding inequality is contained:

$$\mathbb{P}\{\mu_n(A) - \mu(A) \geq \epsilon\} \leq e^{-2n\epsilon^2}. \tag{9.6}$$

The Hoeffding inequality can be extended as follows: Let $X_1, \ldots, X_n$ be independent real-valued random variables, let $a, b \in \mathbb{R}$ with $a < b$, and assume that $X_i \in [a, b]$ with probability one ($i = 1, \ldots, n$). Then, for all $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}\{X_i\})\right| > \epsilon\right\} \leq 2e^{-\frac{2n\epsilon^2}{|b-a|^2}}.$$

(Cf. Hoeffding [**?**].) A further refinement is the Berstein inequality such that it takes into account the variances, too: let $X_1, \ldots, X_n$ be independent real-valued random variables, let $a, b \in \mathbb{R}$ with $a < b$, and assume that $X_i \in [a, b]$ with probability one ($i = 1, \ldots, n$). Let

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{V}ar\{X_i\} > 0.$$

Then, for all $\epsilon > 0$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}\{X_i\})\right| > \epsilon\right\} \leq 2e^{-\frac{n\epsilon^2}{2\sigma^2 + 2\epsilon(b-a)/3}}.$$

(Cf. Berstein [**?**].)

9.3 L_1-distance-based strongly consistent test

Theorem 9.3 results in a strongly consistent test such that reject the null-hypothesis $\mathcal{H}_0$ if

$$L_n > c_1\sqrt{\frac{m_n}{n}},$$

where

$$c_1 > \sqrt{2\ln 2} \approx 1.177.$$

Moreover, assume that the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine. (Cf. (8.5)). Then, under the null hypothesis $\mathcal{H}_0 = \{\nu = \mu\}$, the inequality in Theorem 9.3 implies an upper bound on the error of the first kind

$$\mathbb{P}\left\{L_n > c_1\sqrt{\frac{m_n}{n}}\right\} \leq 2^{m_n}e^{-nc_1^2 m_n/(2n)} = e^{-m_n(c_1^2/2 - \ln 2)} \to 0$$

If $m_n/\ln n \to \infty$ then

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ L_n > c_1 \sqrt{\frac{m_n}{n}} \right\} < \infty,$$

therefore the Borel-Cantelli lemma implies that the goodness of fit test based on the statistic L_n is strongly consistent under the null hypothesis $\mathcal{H}_0$, independently of the underlying distribution μ.

Under the alternative hypothesis $\mathcal{H}_1 = \{\nu \neq \mu\}$, the triangle inequality implies that

$$\begin{aligned} L_n &= \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu(A_{nj})| \\ &\geq \sum_{j=1}^{m_n} |\mu(A_{nj}) - \nu(A_{nj})| - \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \nu(A_{nj})|. \end{aligned}$$

Because of the argument above,

$$\sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \nu(A_{nj})| \to 0,$$

a.s., while the condition (8.5) and $\{\nu \neq \mu\}$ imply that

$$\sum_{j=1}^{m_n} |\mu(A_{nj}) - \nu(A_{nj})| \to 2 \sup_B |\mu(B) - \nu(B)| = 2V(\mu, \nu) > 0. \tag{9.7}$$

therefore

$$\liminf_{n\to\infty} L_n \geq 2V(\mu, \nu) > 0 \tag{9.8}$$

a.s., therefore $L_n > c_1\sqrt{m_n/n}$ a.s. for n large enough, and so the goodness of fit test based on L_n is strongly consistent under the alternative hypothesis $\mathcal{H}_1$, too.

In order to show (9.7) we apply the technique from Barron, Györfi and van der Meulen [**?**]. Choose a measure λ which dominates μ and ν, for example, $\lambda = \mu + \nu$, and denote

by f the Radon-Nikodym derivative of $\mu - \nu$ with respect to λ. Then, on the one hand,

$$\begin{aligned}
\sum_{A\in\mathcal{P}_n} |\mu(A) - \nu(A)| &= \sum_{A\in\mathcal{P}_n} \left| \int_A f \,\mathrm{d}\lambda \right| \\
&\leq \sum_{A\in\mathcal{P}_n} \int_A |f| \,\mathrm{d}\lambda \\
&= \int |f| \,\mathrm{d}\lambda \\
&= 2 \sup_B |\mu(B) - \nu(B)|.
\end{aligned}$$

On the other hand, for uniformly continuous f, using (8.5),

$$\sum_{A\in\mathcal{P}_n} \left| \int_A f \,\mathrm{d}\lambda \right| \to \int |f| \,\mathrm{d}\lambda.$$

If f is arbitrary then, for a given $\delta > 0$, choose a uniformly continuous $\tilde{f}$ such that

$$\int |f - \tilde{f}| \,\mathrm{d}\lambda < \delta.$$

Thus

$$\begin{aligned}
\sum_{A\in\mathcal{P}_n} \left| \int_A f \,\mathrm{d}\lambda \right| &\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \,\mathrm{d}\lambda \right| - \sum_{A\in\mathcal{P}_n} \left| \int_A (f - \tilde{f}) \,\mathrm{d}\lambda \right| \\
&\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \,\mathrm{d}\lambda \right| - \int |f - \tilde{f}| \,\mathrm{d}\lambda \\
&\geq \sum_{A\in\mathcal{P}_n} \left| \int_A \tilde{f} \,\mathrm{d}\lambda \right| - \delta \\
&\to \int |\tilde{f}| \,\mathrm{d}\lambda - \delta \\
&\geq \int |f| \,\mathrm{d}\lambda - 2\delta \\
&= 2 \sup_B |\mu(B) - \nu(B)| - 2\delta.
\end{aligned}$$

The result follows since δ was arbitrary.

9.4 L_1-distance-based α-level test

Beirlant, Györfi and Lugosi [**?**] proved, under conditions

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0,$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0,$$

that

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

where $\xrightarrow{\mathcal{D}}$ indicates convergence in distribution and $\sigma^2 = 1 - 2/\pi$.

Let $\alpha \in (0,1)$. Consider the test which rejects $\mathcal{H}_0$ when

$$L_n \;>\; c_2\sqrt{\frac{m_n}{n}} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) \approx c_2\sqrt{\frac{m_n}{n}},$$

where

$$c_2 = \sqrt{2/\pi} \approx 0.798.$$

Then the test is asymptotically an α-level test.

Comparing c_2 above with c_1 in the strong consistent test, both tests behave identically with respect to $\sqrt{m_n/n}$ for large enough n, but c_2 is smaller.

Under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(L_n - \mathbb{E}\{L_n\})/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the asymptotically α-level test rejects the null hypothesis if

$$L_n > \mathbb{E}\{L_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha).$$

Beirlant, Györfi and Lugosi [**?**] proved an upper bound

$$\mathbb{E}\{L_n\} \le \sqrt{2/\pi}\sqrt{\frac{m_n}{n}}.$$

Chapter 10

Testing Homogeneity

10.1 The testing problem

Consider two mutually independent samples of $\mathbb{R}^d$-valued random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$ with *i.i.d.* components distributed according to unknown probability measures μ and μ'. We are interested in testing the null hypothesis that the two samples are homogeneous, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

Such tests have been extensively studied in the statistical literature for special parametrized models, *e.g.* for linear or loglinear models. For example, the analysis of variance provides standard tests of homogeneity when μ and μ' belong to a normal family on the line. For multinomial models these tests are discussed in common statistical textbooks, together with the related problem of testing independence in contingency tables. For testing homogeneity in more general parametric models, we refer the reader to the monograph of Greenwood and Nikulin [**?**] and further references therein.

However, in many real life applications, the parametrized models are either unknown or too complicated for obtaining asymptotically α-level homogeneity tests by the classical methods. For $d = 1$, there are nonparametric procedures for testing homogeneity, for example, the Cramer-Mises, Kolmogorov-Smirnov, Wilcoxon tests. The problem of $d > 1$ is much more complicated, but nonparametric tests based on finite partitions of $\mathbb{R}^d$ may provide a welcome alternative. Such results are the extensions of Read and Cressie [**?**].

In the present chapter, we discuss a simple approach based on a L_1 distance test statistic. The advantage of our test procedure is that, besides being explicit and relatively easy to carry out, it requires very few assumptions on the partition sequence, and

it is consistent. Let us now describe our test statistic.

Denote by μ_n and μ'_n the empirical measures associated with the samples $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{X}'_1, \ldots, \mathbf{X}'_n$, respectively, so that

$$\mu_n(A) = \frac{\#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}}{n},$$

and, similarly,

$$\mu'_n(A) = \frac{\#\{i : \mathbf{X}'_i \in A, i = 1, \ldots, n\}}{n}.$$

Based on a finite partition $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ of $\mathbb{R}^d$ ($m_n \in \mathbb{N}^*$), we let the test statistic comparing μ_n and μ'_n be defined as

$$T_n = \sum_{j=1}^{m_n} |\mu_n(A_{n,j}) - \mu'_n(A_{n,j})|.$$

10.2 L_1-distance-based strongly consistent test

The following theorem extends the results of Beirlant, Devroye, Györfi and Vajda [**?**], and Devroye and Györfi [**?**] to the statistic T_n.

Theorem 10.1. (Biau, Györfi [**?**].) *Assume that conditions*

$$\lim_{n\to\infty} m_n = \infty, \qquad \lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{10.1}$$

and

$$\lim_{n\to\infty} \max_{j=1,\ldots,m_n} \mu(A_{nj}) = 0, \tag{10.2}$$

are satisfied. Then, under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,

$$\lim_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\{T_n > \varepsilon\} = -g_T(\varepsilon),$$

where

$$g_T(\varepsilon) = (1 + \varepsilon/2) \ln(1 + \varepsilon/2) + (1 - \varepsilon/2) \ln(1 - \varepsilon/2).$$

Proof. We prove only the upper bound

$$\mathbb{P}\{T_n > \epsilon\} \leq 2^{m_n} e^{-n g_T(\epsilon)} \leq 2^{m_n} e^{-n\epsilon^2/4}. \tag{10.3}$$

For any $s > 0$, the Markov inequality implies that

$$\mathbb{P}\{T_n > \epsilon\} \;=\; \mathbb{P}\{e^{snT_n} > e^{sn\epsilon}\} \leq \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\epsilon}}.$$

By Scheffé's theorem for partitions

$$T_n = \sum_{A\in\mathcal{P}_n} |\mu_n(A) - \mu'_n(A)| = 2 \max_{A\in\sigma(\mathcal{P}_n)} (\mu_n(A) - \mu'_n(A)),$$

where the class of sets $\sigma(\mathcal{P}_n)$ contains all sets obtained by unions of cells of $\mathcal{P}_n$. Therefore

$$\begin{aligned}
\mathbb{E}\{e^{snT_n}\} &= \mathbb{E}\{\max_{A\in\sigma(\mathcal{P}_n)} e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&\leq \sum_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&\leq 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn(\mu_n(A)-\mu'_n(A))}\} \\
&= 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \mathbb{E}\{e^{2sn\mu_n(A)}\}\mathbb{E}\{e^{-2sn\mu'_n(A)}\}.
\end{aligned}$$

Clearly,

$$\begin{aligned}
\mathbb{E}\{e^{2sn\mu_n(A)}\} &= \sum_{k=0}^{n} e^{2sk} \binom{n}{k} \mu(A)^k (1-\mu(A))^{n-k} \\
&= \left(e^{2s}\mu(A) + 1 - \mu(A)\right)^n,
\end{aligned}$$

and, similarly, under $\mathcal{H}_0$,

$$\begin{aligned}
\mathbb{E}\{e^{-2sn\mu'_n(A)}\} &= \sum_{k=0}^{n} e^{-2sk} \binom{n}{k} \mu(A)^k (1-\mu(A))^{n-k} \\
&= \left(e^{-2s}\mu(A) + 1 - \mu(A)\right)^n.
\end{aligned}$$

The remainder of the proof is under the null hypothesis $\mathcal{H}_0$. From above, we deduce that

$$\begin{aligned}
&\mathbb{E}\{e^{snT_n}\} \\
&\quad \le 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left(e^{2s}\mu(A)+1-\mu(A)\right)^n \left(e^{-2s}\mu(A)+1-\mu(A)\right)^n \\
&\quad = 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left[\left(e^{2s}\mu(A)+1-\mu(A)\right)\left(e^{-2s}\mu(A)+1-\mu(A)\right)\right]^n \\
&\quad = 2^{m_n} \max_{A\in\sigma(\mathcal{P}_n)} \left[1+\mu(A)\left(1-\mu(A)\right)\left(e^{2s}+e^{-2s}-2\right)\right]^n \\
&\quad \le 2^{m_n} \left[1+(e^{2s}+e^{-2s}-2)/4\right]^n \\
&\quad = 2^{m_n} \left[1/2+(e^{2s}+e^{-2s})/4\right]^n .
\end{aligned}$$

It implies that

$$\mathbb{P}\{T_n>\epsilon\} \;\le\; \inf_{s>0} \frac{\mathbb{E}\{e^{snT_n}\}}{e^{sn\epsilon}} \le 2^{m_n} \left[\inf_{s>0} \frac{1/2+(e^{2s}+e^{-2s})/4}{e^{s\epsilon}}\right]^n$$

One can verify that the infimum is achieved at

$$e^{2s} = \frac{1+\varepsilon/2}{1-\varepsilon/2},$$

and then

$$\mathbb{P}\{T_n>\epsilon\} \le 2^{m_n} e^{-ng_T(\epsilon)}.$$

The Pinsker inequality implies that

$$g_T(\epsilon) \ge \epsilon^2/4$$

therefore

$$\mathbb{P}\{T_n>\epsilon\} \le 2^{m_n} e^{-n\epsilon^2/4}.$$

□

The technique of Theorem 10.1 yields a distribution-free strong consistent test of homogeneity, which rejects the null hypothesis if T_n becomes large. We insist on the fact that the test presented in Corollary 10.1 is entirely distribution-free, i.e., the measures μ and μ' are completely arbitrary.

Corollary 10.1. (Biau, Györfi [?].) *Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_1\sqrt{\frac{m_n}{n}},$$

where

$$c_1 > 2\sqrt{\ln 2} \approx 1.6651.$$

Assume that condition (10.1) is satisfied and

$$\lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

Then, under $\mathcal{H}_0$, after a random sample size the test makes a.s. no error. Moreover, if

$$\mu \neq \mu',$$

and the sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \ldots$ is asymptotically fine, (cf. (8.5)), then after a random sample size the test makes a.s. no error.

PROOF. Under $\mathcal{H}_0$, by (10.3),

$$\begin{aligned}
\mathbb{P}\left\{T_n > c_1\sqrt{\frac{m_n}{n}}\right\} &\leq 2^{m_n} e^{-ng_T\left(c_1\sqrt{m_n/n}\right)} \\
&= 2^{m_n} e^{-nc_1^2(m_n/n)/4 + n\cdot o(m_n/n)} \\
&= e^{-\left(c_1^2/4 - \ln 2 + o(1)\right)m_n},
\end{aligned}$$

as $n \to \infty$. Therefore the condition $m_n/\ln n \to \infty$ implies that

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{T_n > c_1\sqrt{\frac{m_n}{n}}\right\} < \infty,$$

and by the Borel-Cantelli lemma we are ready with the first half of the corollary. Concerning the second half, in the same way as for (9.7) we can show that by the additional condition (8.5),

$$\liminf_{n\to\infty} T_n \geq 2 \sup_B |\mu(B) - \mu'(B)| > 0 \tag{10.4}$$

a.s. □

10.3 L_1-distance-based α-level test

Again, one can prove the following asymptotic normality:

Theorem 10.2. (Biau, Györfi [?].) *Assume that conditions (10.1) and (10.2) are satisfied. Then, under $\mathcal{H}_0$, there exists a centering sequence $C_n = \mathbb{E}\{T_n\}$ such that*

$$\sqrt{n}\,(T_n - C_n)\,/\sigma \xrightarrow{\mathcal{D}} \mathcal{N}(0,1),$$

where $\sigma^2 = 2(1-2/\pi)$.

Theorem 10.2 yields the asymptotic null distribution of a consistent homogeneity test, which rejects the null hypothesis if T_n becomes large. In contrast to Corollary 10.1, and because of condition (10.2), this new test is *not* distribution-free. In particular, the measures μ and μ' have to be nonatomic.

Corollary 10.2. (Biau, Györfi [?].) *Put $\alpha \in (0,1)$, and let $C^* \approx 0.7655$ denote a universal constant. Consider the test which rejects $\mathcal{H}_0$ when*

$$T_n > c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha) \approx c_2\sqrt{\frac{m_n}{n}},$$

where

$$\sigma^2 = 2(1-2/\pi) \quad \textit{and} \quad c_2 = \frac{2}{\sqrt{\pi}} \approx 1.1284.$$

Then, under the conditions of Theorem 10.2, the test is an asymptotically α-level test. Moreover, under the additional condition (8.5), the test is consistent.

Proof. According to Theorem 10.2, under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(T_n - \mathbb{E}\{T_n\})/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the α-level test rejects the null hypothesis if

$$T_n > \mathbb{E}\{T_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha).$$

However, $\mathbb{E}\{T_n\}$ depends on the unknown distribution, thus we apply an upper bound on $\mathbb{E}\{T_n\}$, and so decrease the error probability. The following inequality is valid:

$$\mathbb{E}\{T_n\} \;\le\; c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n},$$

(cf. Biau, Györfi [**?**]). Thus

$$\begin{aligned}\alpha &\approx \mathbf{P}\left\{T_n > \mathbb{E}\{T_n\} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\} \\ &\geq \mathbf{P}\left\{T_n > c_2\sqrt{\frac{m_n}{n}} + C^*\,\frac{m_n}{n} + \frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\right\}.\end{aligned}$$

This proves that the test has asymptotic error probability at most α.
Under $\mu \neq \mu'$, the consistency of the test follows from (10.4). □

Chapter 11

Testing Independence

11.1 The testing problem

Consider a sample of $\mathbb{R}^d \times \mathbb{R}^{d'}$-valued random vectors $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$ with independent and identically distributed (i.i.d.) pairs. The distribution of $(\mathbf{X}, \mathbf{Y})$ is denoted by ν, while μ_1 and μ_2 stand for the distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively. We are interested in testing the null hypothesis that $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$\mathcal{H}_0 : \nu = \mu_1 \times \mu_2, \tag{11.1}$$

while making minimal assumptions regarding the distribution.

We obtain two kinds of tests for each statistic: first, we derive *strong consistent* tests — meaning that both on $\mathcal{H}_0$ and on its complement the tests make a.s. no error after a random sample size — based on large deviation bounds. While such tests are not common in the classical statistics literature, they are well suited to data analysis from streams, where we receive a sequence of observations rather than a sample of fixed size, and must return the best possible decision at each time using only current and past observations. Our strong consistent tests are *distribution-free*, meaning they require no conditions on the distribution being tested; and *universal*, meaning the test threshold holds independent of the distribution. Second, we obtain tests based on the asymptotic distribution of the L_1, which assume only that ν is nonatomic. Subject to this assumption, the tests are *consistent*: for a given asymptotic error rate on $\mathcal{H}_0$, the probability of error on $\mathcal{H}_1$ drops to zero as the sample size increases. Moreover, the thresholds for the asymptotic tests are distribution-independent. We emphasize that our tests are explicit, easy to carry out, and require very few assumptions on the partition sequences.

Additional independence testing approaches also exist in the statistics literature. For $d = d' = 1$, an early nonparametric test for independence, due to Hoeffding [**?**], Blum et

al. [?], De Wet [?] is based on the notion of differences between the joint distribution function and the product of the marginals. The associated independence test is consistent under appropriate assumptions. Two difficulties arise when using this statistic in a test, however. First, quantiles of the null distribution are difficult to estimate. Second, and more importantly, the quality of the empirical distribution function estimates becomes poor as the dimensionality of the spaces $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ increases, which limits the utility of the statistic in a multivariate setting.

Rosenblatt [?] defined the statistic as the L_2 distance between the joint density estimate and the product of marginal density estimates. Let K and K' be density functions (called kernels) defined on $\mathbb{R}^d$ and on $\mathbb{R}^{d'}$, respectively. For the bandwidth $h > 0$, define

$$K_h(\mathbf{x}) = \frac{1}{h^d} K\left(\frac{\mathbf{x}}{h}\right) \quad \text{and} \quad K'_h(\mathbf{y}) = \frac{1}{h^{d'}} K'\left(\frac{\mathbf{y}}{h}\right).$$

The Rosenblatt-Parzen kernel density estimates of the density of $(\mathbf{X}, \mathbf{Y})$ and $\mathbf{X}$ are respectively

$$f_n(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i) K'_h(\mathbf{y} - \mathbf{Y}_i) \text{ and } f_{n,1}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i), \tag{11.2}$$

with $f_{n,2}(\mathbf{y})$ defined by analogy. Rosenblatt [?] introduced the kernel-based independence statistic

$$T_n = \int_{\mathbb{R}^d \times \mathbb{R}^{d'}} (f_n(\mathbf{x}, \mathbf{y}) - f_{n,1}(\mathbf{x}) f_{n,2}(\mathbf{y}))^2 d\mathbf{x}\, d\mathbf{y}. \tag{11.3}$$

Further approaches to independence testing can be employed when particular assumptions are made on the form of the distributions, for instance that they should exhibit symmetry. We do not address these approaches in the present study.

11.2 L_1-distance-based strongly consistent test

Denote by ν_n, $\mu_{n,1}$ and $\mu_{n,2}$ the empirical measures associated with the samples $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$, $\mathbf{X}_1, \ldots, \mathbf{X}_n$, and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, respectively, so that

$$\begin{aligned} \nu_n(A \times B) &= n^{-1} \#\{i : (\mathbf{X}_i, \mathbf{Y}_i) \in A \times B, i = 1, \ldots, n\}, \\ \mu_{n,1}(A) &= n^{-1} \#\{i : \mathbf{X}_i \in A, i = 1, \ldots, n\}, \quad \text{and} \\ \mu_{n,2}(B) &= n^{-1} \#\{i : \mathbf{Y}_i \in B, i = 1, \ldots, n\}. \end{aligned}$$

Given the finite partitions $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ of $\mathbb{R}^d$ and $\mathcal{Q}_n = \{B_{n,1}, \ldots, B_{n,m'_n}\}$ of $\mathbb{R}^{d'}$, we define the L_1 test statistic comparing ν_n and $\mu_{n,1} \times \mu_{n,2}$ as

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) = \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|.$$

In the following two sections, we derive the large deviation and limit distribution properties of this L_1 statistic, and the associated independence tests.

For testing a simple hypothesis versus a composite alternative, Györfi and van der Meulen [**?**] introduced a related goodness of fit test statistic L_n defined as

$$L_n(\mu_{n,1}, \mu_1) = \sum_{A \in \mathcal{P}_n} |\mu_{n,1}(A) - \mu_1(A)|.$$

Biau and Györfi [**?**] proved that, for all $0 < \varepsilon$,

$$\mathbb{P}\{L_n(\mu_{n,1}, \mu_1) > \varepsilon\} \leq 2^{m_n} e^{-n\varepsilon^2/2}, \tag{11.4}$$

(cf. Theorem 9.3). We now describe a similar result for our L_1 independence statistic.

Theorem 11.1. *(*Gretton, Györfi [**?**].) *Under $\mathcal{H}_0$, for all $0 < \varepsilon_1$, $0 < \varepsilon_2$ and $0 < \varepsilon_3$,*

$$\mathbb{P}\{L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > \varepsilon_1 + \varepsilon_2 + \varepsilon_3\} \leq 2^{m_n \cdot m'_n} e^{-n\varepsilon_1^2/2} + 2^{m_n} e^{-n\varepsilon_2^2/2} + 2^{m'_n} e^{-n\varepsilon_3^2/2}.$$

Proof. We bound $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$ according to

$$\begin{aligned}
L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) &= \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)| \\
&\leq \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu_n(A \times B) - \nu(A \times B)| \\
&\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu(A \times B) - \mu_1(A) \cdot \mu_2(B)| \\
&\quad + \sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\mu_1(A) \cdot \mu_2(B) - \mu_{n,1}(A) \cdot \mu_{n,2}(B)|.
\end{aligned}$$

Under the null hypothesis $\mathcal{H}_0$, we have that

$$\sum_{A \in \mathcal{P}_n} \sum_{B \in \mathcal{Q}_n} |\nu(A \times B) - \mu_1(A) \cdot \mu_2(B)| = 0.$$

Moreover

$$\begin{aligned}
& \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_2(B) - \mu_{n,1}(A)\cdot\mu_{n,2}(B)| \\
\le\ & \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_2(B) - \mu_1(A)\cdot\mu_{n,2}(B)| \\
& + \sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n} |\mu_1(A)\cdot\mu_{n,2}(B) - \mu_{n,1}(A)\cdot\mu_{n,2}(B)| \\
=\ & \sum_{B\in\mathcal{Q}_n} |\mu_2(B) - \mu_{n,2}(B)| + \sum_{A\in\mathcal{P}_n} |\mu_1(A) - \mu_{n,1}(A)| \\
=\ & L_n(\mu_{n,1},\mu_1) + L_n(\mu_{n,2},\mu_2).
\end{aligned}$$

Thus, (11.4) implies

$$\begin{aligned}
& \mathbb{P}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) > \varepsilon_1+\varepsilon_2+\varepsilon_3\} \\
\le\ & \mathbb{P}\left\{L_n(\nu_n,\nu) > \varepsilon_1\right\} + \mathbb{P}\left\{L_n(\mu_{n,1},\mu_1) > \varepsilon_2\right\} + \mathbb{P}\left\{L_n(\mu_{n,2},\mu_2) > \varepsilon_3\right\} \\
\le\ & 2^{m_n\cdot m'_n} e^{-n\varepsilon_1^2/2} + 2^{m_n} e^{-n\varepsilon_2^2/2} + 2^{m'_n} e^{-n\varepsilon_3^2/2}.
\end{aligned}$$

□

Theorem 11.1 yields a strong consistent test of independence, which rejects the null hypothesis if $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ becomes large. The test is distribution-free, i.e., the probability distributions ν, μ_1 and μ_2 are completely arbitrary; and the threshold is universal, i.e., it does not depend on the distribution.

Corollary 11.1. *(*Gretton, Györfi [**?**].*) Consider the test which rejects $\mathcal{H}_0$ when*

$$L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) > c_1\left(\sqrt{\frac{m_n m'_n}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m'_n}{n}}\right) \approx c_1\sqrt{\frac{m_n m'_n}{n}},$$

where

$$c_1 > \sqrt{2\ln 2} \approx 1.177. \tag{11.5}$$

Assume that conditions

$$\lim_{n\to\infty}\frac{m_n m'_n}{n} = 0, \tag{11.6}$$

and

$$\lim_{n\to\infty}\frac{m_n}{\ln n} = \infty, \qquad \lim_{n\to\infty}\frac{m'_n}{\ln n} = \infty, \tag{11.7}$$

are satisfied. Then under $\mathcal{H}_0$, the test makes a.s. no error after a random sample size. Moreover, if

$$\nu \neq \mu_1 \times \mu_2,$$

and for any sphere S centered at the origin,

$$\lim_{n\to\infty} \max_{A\in\mathcal{P}_n,\, A\cap S\neq 0} \operatorname{diam}(A) = 0 \tag{11.8}$$

and

$$\lim_{n\to\infty} \max_{B\in\mathcal{Q}_n,\, B\cap S\neq 0} \operatorname{diam}(B) = 0, \tag{11.9}$$

then after a random sample size the test makes a.s. no error.

PROOF. Under $\mathcal{H}_0$, we obtain from Theorem 11.1 a non-asymptotic bound for the tail of the distribution of $L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})$, namely

$$\begin{aligned}
& \mathbb{P}\left\{ L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} \\
\leq\ & 2^{m_n m_n'} e^{-c_1^2 m_n m_n'/2} + 2^{m_n} e^{-c_1^2 m_n/2} + 2^{m_n'} e^{-c_1^2 m_n'/2} \\
\leq\ & e^{-(c_1^2/2 - \ln 2) m_n m_n'} + e^{-(c_1^2/2-\ln 2) m_n} + e^{-(c_1^2/2-\ln 2) m_n'}
\end{aligned}$$

as $n \to \infty$. Therefore the condition (11.7) implies

$$\sum_{n=1}^{\infty} \mathbb{P}\left\{ L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > c_1 \left(\sqrt{\frac{m_n m_n'}{n}} + \sqrt{\frac{m_n}{n}} + \sqrt{\frac{m_n'}{n}} \right) \right\} < \infty,$$

and the proof under the null hypothesis is completed by the Borel-Cantelli lemma. For the result under the alternative hypothesis, we first apply the triangle inequality

$$\begin{aligned}
L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) \ \geq\ & \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu(A\times B) - \mu_1(A)\cdot\mu_2(B)| \\
& - \sum_{A\in\mathcal{P}_n} \sum_{B\in\mathcal{Q}_n} |\nu_n(A\times B) - \nu(A\times B)| \\
& - \sum_{B\in\mathcal{Q}_n} |\mu_2(B) - \mu_{n,2}(B)| \\
& - \sum_{A\in\mathcal{P}_n} |\mu_1(A) - \mu_{n,1}(A)|.
\end{aligned}$$

The condition in (11.6) implies the three last terms of the right hand side tend to 0 a.s. Moreover, using the technique for (9.7) we can prove that by conditions (11.8) and (11.9),

$$\sum_{A\in\mathcal{P}_n}\sum_{B\in\mathcal{Q}_n}|\nu(A\times B)-\mu_1(A)\cdot\mu_2(B)|\to 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0$$

as $n\to\infty$, where the last supremum is taken over all Borel subsets C of $\mathbb{R}^d\times\mathbb{R}^{d'}$, and therefore

$$\liminf_{n\to\infty} L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\geq 2\sup_C|\nu(C)-\mu_1\times\mu_2(C)|>0 \tag{11.10}$$

a.s. □

11.3 L_1-distance-based α-level test

Again, one can prove the following asymptotic normality:

Theorem 11.2. *(*Gretton, Györfi [?].*) Assume that conditions (11.6) and*

$$\lim_{n\to\infty}\max_{A\in\mathcal{P}_n}\mu_1(A)=0,\quad \lim_{n\to\infty}\max_{B\in\mathcal{Q}_n}\mu_2(B)=0, \tag{11.11}$$

are satisfied. Then, under $\mathcal{H}_0$, there exists a centering sequence $C_n=\mathbb{E}\{L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})\}$ depending on ν such that

$$\sqrt{n}\left(L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})-C_n\right)/\sigma\xrightarrow{\mathcal{D}}\mathcal{N}(0,1),$$

where $\sigma^2=1-2/\pi$.

Theorem 11.2 yields the asymptotic null distribution of a consistent independence test, which rejects the null hypothesis if $L_n(\nu_n,\mu_{n,1}\times\mu_{n,2})$ becomes large. In contrast to Corollary 11.1, and because of condition (11.11), this new test is *not* distribution-free: the measures μ_1 and μ_2 have to be nonatomic.

Corollary 11.2. *(*Gretton, Györfi [?].*) Let $\alpha\in(0,1)$. Consider the test which rejects $\mathcal{H}_0$ when*

$$\begin{aligned} L_n(\nu_n,\mu_{n,1}\times\mu_{n,2}) &> c_2\sqrt{\frac{m_nm_n'}{n}}+\frac{\sigma}{\sqrt{n}}\,\Phi^{-1}(1-\alpha)\\ &\approx c_2\sqrt{\frac{m_nm_n'}{n}}, \end{aligned}$$

where

$$\sigma^2 = 1 - 2/\pi \quad \textit{and} \quad c_2 = \sqrt{2/\pi} \approx 0.798.$$

Then, under the conditions of Theorem 11.2, the test is an asymptotically α-level test. Moreover, under the additional conditions (11.8) and (11.9), the test is consistent.

Before proceeding to the proof, we examine how the above test differs from that in Corollary 11.1. In particular, comparing c_2 above with c_1 in (11.5), both tests behave identically with respect to $\sqrt{m_n m'_n/n}$ for large enough n, but c_2 is smaller.

PROOF. According to Theorem 11.2, under $\mathcal{H}_0$,

$$\mathbb{P}\{\sqrt{n}(L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) - C_n)/\sigma \le x\} \approx \Phi(x),$$

therefore the error probability with threshold x is

$$\alpha = 1 - \Phi(x).$$

Thus the α-level test rejects the null hypothesis if

$$L_n(\nu_n, \mu_{n,1} \times \mu_{n,2}) > C_n + \frac{\sigma}{\sqrt{n}}\, \Phi^{-1}(1-\alpha).$$

As C_n depends on the unknown distribution, we apply an upper bound

$$C_n = \mathbb{E}\{L_n(\nu_n, \mu_{n,1} \times \mu_{n,2})\} \le \sqrt{2/\pi}\sqrt{\frac{m_n m'_n}{n}}$$

(cf. Gretton, Györfi [**?**]). □