Performance limits of nonparametric estimators

András Antos

Supervisor: Dr. László Györfi

Budapest, April 22, 1999

Contents

1 Introduction 2

2 Regression estimation 14

2.1 Regression problem 14

2.2 Slow rate of convergence 15

2.3 Minimax lower bounds 16

2.4 Individual lower bounds 22

3 Pattern recognition 26

3.1 The pattern recognition problem 26

3.2 Slow rate of convergence 28

3.3 Lower bounds for smoothness classes 28

3.3.1 Minimax lower bounds 29

3.3.2 Individual lower bounds 37

3.4 Lower bounds for VC classes 39

3.4.1 Minimax lower bounds 40

3.4.2 Counter-examples 42

3.4.3 Individual lower bounds 45

3.4.4 Cumulative error bounds 51

3.4.5 Bounds for the tail probabilities 53

4 Functional estimation 64

4.1 Functional estimation problems 64

4.2 Consistency 65

4.3 Slow rate of convergence 72

4.4 Rate of convergence 83

Bibliography 87

Chapter 1

Introduction

In this work we survey some fields of nonparametric statistics, namely regression function estimation, pattern recognition and functional estimation. In all cases we have a random variable with unknown distribution. Our aim is to estimate a characteristics determined by the distribution. The estimation is based on independent samples drawn from the distribution. In order to measure the goodness of an estimate, we have a loss function defined in the particular models. We are mostly interested in the asymptotic behaviour of the expected loss.

For all of these problems we have universal consistency, that is, we have estimates for which the loss tends to zero in some sense for all distributions. But it turns out that there are no such estimates, for which the expected loss tends to zero with a *guaranteed rate of convergence* for all distributions. Hence the rate of convergence of any estimate can be arbitrary slow for a "bad" distribution.

In order to get rate-of-convergence results, we have to restrict ourselves to a class of distributions instead of allowing every distribution. For certain classes such rates of convergence are known. In many cases they are also proven to be optimal by corresponding lower bounds for the expected loss. Such an optimal rate of convergence determines the rate of convergence of the worst-case expected loss of the best estimate.

However, theoretically, the rate of convergence of the worst-case expected loss can be different from that of the expected loss for every single distribution. For example, it is possible that while the worst-case expected loss tends to zero only at a polynomial rate, the rate is exponential for every single distribution. The former will be called *minimax* approach, the latter as *individual* approach. We are concerned with recent results of the individual approach, especially extending the earlier minimax-rate results to individual rates.

More formally, let X be a random vector variable taking values from a set $\mathcal{X} \subset \mathcal{R}^d$. The distribution of X is unknown. We would like to guess or estimate a characteristics P of the distribution of X. Its value can be real, a function or even a distribution. As an input we are given independent samples $X_1, \dots, X_n$. The estimate is a function P_n on $\mathcal{X}^n$ assigning a possible value of P to every particular sample sequence $(X_1, \dots, X_n)$. We have a nonnegative loss function $l(\cdot, \cdot)$ (a metric or not). The loss $l(P_n, P)$ measures the goodness of an estimate. We are interested in the convergence of $l(P_n, P)$ to zero. (For example, in the case of regression analysis, the random variable is the pair (X, Y), the characteristics is the regression function $m(x) = \mathbf{E}\{Y|X = x\}$, and the loss function is the mean square error $l(m', m) = \mathbf{E}\{(m'(X) - m(X))^2\}$, see Section 2.1.) Henceforth we assume measurability where needed.

DEFINITION 1.1. *An estimate P_n is called* **(strongly) consistent for a distribution** *of X if*

$$\mathbf{E}l(P_n, P) \to 0 \qquad (l(P_n, P) \to 0 \quad a.s.)$$

(where a.s. means almost surely). An estimate P_n is called **(strongly) universally consistent** *if it is (strongly) consistent for all possible distributions of X.* □

Both consistency and strong consistency imply *weak consistency*, that is,

$$l(P_n, P) \to 0 \quad \text{in probability.}$$

As we will see in our problems, usually we have universally consistent estimates (see, e.g., (2.2) for the regression analysis).

The next question is whether there are estimates with the expected loss tending to zero at a specified rate for all distributions. Disappointingly, in many cases, such estimates do not exist. The first kind of negative results shows that for any estimate and for any (or infinitely many) *fixed* n, there exists a distribution such that the expected loss is larger than some ϵ.

Another question is whether a certain universal rate of convergence is achievable for some estimate. For example, the previous kind of results do not exclude the existence of an estimate such that for all n, $\mathbf{E}l(P_n, P) \leq c/n$ for all distributions, for some constant c depending upon the actual distribution. The next kind of negative results is that this cannot be the case. These slow-rate-of-convergence results state that the expected loss of any estimate is larger than (say) $b_n = c/\log\log\log n$ *for every (or infinitely many)* n for some distribution, even if c may depend on the distribution. If we allow all distributions, the rate of convergence of any sequence of estimates can be arbitrary slow, that is, given a sequence $\{b_n\}$ tending to zero, for every sequence of estimates there is a distribution such that $\mathbf{E}l(P_n, P) > b_n$ for every (or infinitely many) n (see, e.g., Theorem 2.1 for the regression analysis).

Thus, in practice, no estimate ensures small loss, unless the actual distribution is known to be a member of a restricted class. Rate-of-convergence studies for particular estimates must necessarily be accompanied by conditions on X. Under certain regularity conditions it is usually possible to obtain upper bounds for the rates of convergence of $\mathbf{E}l(P_n, P)$ for some estimates. Then it is natural to ask what the fastest achievable rate is for the given class of distributions. A minimax theory was worked out by Stone (1982) for regression function estimation. For related results on the general minimax theory of statistical estimates see, for example, Ibragimov and Khasminskii (1980), (1981), (1982), Bretangole and Huber (1979), Birgé (1983), and Korostelev and Tsybakov (1993).

To formulate these concepts introduce the following notations:

DEFINITION 1.2. *For a class $\mathcal{D}$ of distributions and a sequence $\{a_n\}$ of pos-*

itive numbers let

$$l_{\mathrm{mm}}(\mathcal{D},\{a_n\}) \stackrel{\mathrm{def}}{=} \limsup_{n\to\infty} \inf_{P_n} \sup_{X\in\mathcal{D}} \frac{\mathbf{E} l(P_n,P)}{a_n}$$

(for the minimax rates), and

$$l_{\mathrm{ind}}(\mathcal{D},\{a_n\}) \stackrel{\mathrm{def}}{=} \inf_{\{P_n\}} \sup_{X\in\mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E} l(P_n,P)}{a_n}$$

(for the individual rates), where the infimum is taken over all estimates (or sequences of estimates), while the supremum is taken over all distributions in $\mathcal{D}$. □

The obvious relation among these quantities is:

$$l_{\mathrm{mm}}(\mathcal{D},\{a_n\}) \ge l_{\mathrm{ind}}(\mathcal{D},\{a_n\}) \ .$$

If the worst case expected loss of the best estimate has order of magnitude a_n, then its constant coefficient is $l_{\mathrm{mm}}(\mathcal{D},\{a_n\})$. (Moreover, $l_{\mathrm{mm}}(\mathcal{D},\{a_n\}) = 0$ or $l_{\mathrm{mm}}(\mathcal{D},\{a_n\}) = \infty$ mean that the order of magnitude is less or greater than $\{a_n\}$, respectively.)

$l_{\mathrm{ind}}(\mathcal{D},\{a_n\})$ is the constant coefficient of a_n in the order of magnitude of the expected loss in case of the worst fixed distribution and the best sequence of estimates. (Now $l_{\mathrm{ind}}(\mathcal{D},\{a_n\}) = 0$ means that for the best sequence of estimates for every distribution the order of magnitude is less than $\{a_n\}$, while $l_{\mathrm{ind}}(\mathcal{D},\{a_n\}) = \infty$ means that for every sequence of estimates for the worst distribution the order of magnitude is greater than $\{a_n\}$.)

Assume first that $\mathcal{D}$ contains all possible distributions. We realize that universal consistency means that

$$l_{\mathrm{ind}}(\mathcal{D},\{1\}) = 0 \ ,$$

where $\{1\}$ stands for the sequence of constant 1. The weaker kind of negative result above means that

$$l_{\mathrm{mm}}(\mathcal{D},\{1\}) > 0 \ .$$

The slow-rate result means that for every positive sequence $\{b_n\}$ tending to zero

$$l_{\text{ind}}(\mathcal{D}, \{b_n\}) = \infty \ .$$

DEFINITION 1.3. *A positive sequence* $\{a_n\}$ *is called* **minimax lower rate of convergence for the class** $\mathcal{D}$ *if*

$$l_{\text{mm}}(\mathcal{D}, \{a_n\}) > 0 \ . \quad \square$$

In many cases the limes superior in the definition of $l_{\text{mm}}(\mathcal{D}, \{a_n\})$ could be replaced by limes inferior or infimum, because the lower bound for the minimax loss holds for all (sufficiently large) n.

Call $\{a_n\}$ a **minimax upper rate of convergence for the class** $\mathcal{D}$ if for some P_n

$$\limsup_{n\to\infty} \sup_{X\in\mathcal{D}} \frac{\mathbf{E}l(P_n, P)}{a_n} < \infty \ ,$$

that is, if $\sup_{X\in\mathcal{D}} \mathbf{E}l(P_n, P) = O(a_n)$, and call it a **minimax optimal rate of convergence for the class** $\mathcal{D}$ if it is both minimax upper and lower rate of convergence (see, e.g., (2.3) and Theorem 2.2 for the regression analysis).

In some sense, lower bounds in minimax form are not satisfactory. They do not tell us anything about the way the loss decreases as the sample size is increasing for a given distribution. These bounds, for each n, give information about the maximal error within the class, but not about the behaviour of the loss for a single fixed distribution as the sample size n increases. In other words, the "bad" distribution, causing the largest error for an estimate, may be different for each n. For example, the existence of a polynomial minimax lower rate does not exclude the possibility that there exists a sequence of estimates $\{P_n\}$ such that for *every* distribution of X the expected error $\mathbf{E}l(P_n, P)$ decreases at an exponential rate in n. Indeed, there are such examples in Chapter 3. We are interested in "individual" minimax lower bounds that describe the behavior of the loss for a fixed distribution of X as the sample size n grows.

DEFINITION 1.4. *A positive sequence* $\{a_n\}$ *is called* **individual lower rate of convergence for the class** $\mathcal{D}$ *if*

$$l_{\mathrm{ind}}(\mathcal{D}, \{a_n\}) > 0 \ . \quad \square$$

A slightly different, but essentially equivalent, definition requires that for some $c > 0$ for all sequences $\{P_n\}$, there exists a fixed distribution of X such that $\mathbf{E}l(P_n, P) \geq ca_n$ for *infinitely many* n. In some cases the limes superior in the definition of $l_{\mathrm{ind}}(\mathcal{D}, \{a_n\})$ could be replaced by limes inferior or infimum, because the lower bound for the loss of the worst distribution holds for all (sufficiantly large) n.

Call $\{a_n\}$ an **individual upper rate of convergence for the class** $\mathcal{D}$ if for proper $\{P_n\}$

$$\sup_{X \in \mathcal{D}} \limsup_{n\to\infty} \frac{\mathbf{E}l(P_n, P)}{a_n} < \infty \ ,$$

which implies only that for every distribution in $\mathcal{D}$, $\mathbf{E}l(P_n, P) = O(a_n)$, possibly with different constants. Call $\{a_n\}$ an **individual optimal rate of convergence for the class** $\mathcal{D}$ if for every sequence $\{b_n\}$ tending to zero $\{a_n/b_n\}$ is an individual upper and $\{b_n a_n\}$ is an individual lower rate of convergence (see, e.g., Theorem 2.3 for the regression analysis).

To be minimax upper rate is stronger than to be individual upper rate, and to be individual lower rate is stronger than to be minimax lower rate. It is a reasonable aim to extend minimax lower bounds to individual lower bounds. As we will see, the extension is possible for many important classes. Certainly, all lower rate results still hold if we increase the class by leaving some condition on the distributions.

Now the first kind of negative result means that the sequence $\{1\}$ is a minimax lower rate of convergence for the class of all distributions, while the slow-rate result means that every sequence tending to zero is an individual lower rate of convergence for this class.

Let us mention some earlier individual lower bounds for the sake of illustration:

The first example in this respect could be the problem estimating a distribution function $F(x)$ on $\mathcal{X} = \mathcal{R}^d$ ($d \geq 1$), that is, the characteristics of the distribution is F and the loss is the supremum norm

$$l(F_n, F) = \sup_{x \in \mathcal{R}^d} |F_n(x) - F(x)| \ .$$

Here the uniform convergence of the empirical distribution function $F_n(x)$, that is, the Glivenko-Cantelli Theorem holds for all $F(x)$:

$$\lim_{n \to \infty} \sup_{x \in \mathcal{R}^d} |F_n(x) - F(x)| = 0 \quad \text{a.s.}, \tag{1.1}$$

and also there is a c, such that for all $F(x)$

$$\mathbf{E} l(F_n, F) \leq c \sqrt{\frac{d}{n}} \ ,$$

that is, we have strong universal consistency, and $\{1/\sqrt{n}\}$ is a minimax upper rate of convergence for the class of all distributions. 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 classification problem (see Chapter 3) one wants to use empirical distribution functions for two unknown continuous distribution functions for creating a kind of likelihood test, 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 as a loss function: if μ and ν are probability measures on $\mathcal{R}^d$ then the total variation of μ and ν is defined by

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

where the supremum is taken over all Borel sets A.

However, if μ stands for the common distribution of $\{X_i\}$ and μ_n denotes the empirical distribution then for nonatomic μ

$$V(\mu, \mu_n) = 1 \quad \text{a.s.},$$

so the empirical distribution is a bad estimate in total varition.

One may expect to find a more sophisticated sequence $\{\mu_n^*\}$ of distribution estimates of μ which is (strongly) universally consistent in total variation:

$$\lim_{n\to\infty} V(\mu, \mu_n^*) = 0 \quad \text{a.s.}$$

Theorem 1.1. (DEVROYE AND GYÖRFI (1990)) *Given any sequence of distribution estimators $\{\mu_n^*\}$ there exists a probability measure μ for which*

$$V(\mu, \mu_n^*) > 0.5 \qquad \textit{for all } n \textit{ a.s.} \quad \square$$

This implies $l_{\text{ind}}(\mathcal{D}, \{1\}) \geq 0.5$. This negative finding means that the total variation is a much stronger error criterion than the Kolmogorov-Smirnov distance and it is impossible to construct a distribution estimate with universal consistency in total variation. For meaningful results either the class of sets, over which the supremum is taken in the definition of $l(\cdot,\cdot)$, or the class of permitted distributions must be restricted.

If the class of sets is between the two extreme cases mentioned above (only the octants and all Borel sets in $\mathcal{R}^d$, respectively), then the celebrated theory of Vapnik and Chervonenkis yields that for Vapnik-Chervonenkis classes (see Section 3.4) the situation is the same as for Kolmogorov-Smirnov distance. We have strong universal consistency and $\{1/\sqrt{n}\}$ is a minimax upper rate.

Now let the class $\mathcal{D}$ of distributions be the class of absolutely continuous distributions with respect to a σ-finite measure λ. If μ and ν are in $\mathcal{D}$ with densities f and g respectively, then

$$\|f - g\| := \int_{\mathcal{R}^d} |f(x) - g(x)|\lambda(dx) = 2\,V(\mu, \nu)\ .$$

This relation results in a way of distribution estimation consistent in total variation via L_1-consistent density estimation: assume that f_n is (strongly) universally L_1-consistent, that is,

$$\lim_{n\to\infty} \|f_n - f\| = 0 \quad \text{a.s.}$$

Introduce the distribution estimate induced by the density estimate f_n:

$$\mu_n^*(A) = \int_A f_n(x)\lambda(dx),$$

then

$$\lim_{n\to\infty} V(\mu, \mu_n^*) = 0 \quad \text{a.s.}$$

Standard examples of universally L_1-consistent density estimates are the histogram and the kernel estimates when λ is the Lebesgue measure (see Devroye (1983a) and Devroye, Györfi (1985)).

Theorem 1.2. *Assume that μ has a density f. There are estimates (histogram, kernel) that*

$$\lim_{n\to\infty} \|f - f_n\| = 0 \quad a.s. \quad \Box$$

This implies consistency over $\mathcal{D}$ in expectation, that is, $l_{\text{ind}}(\mathcal{D}, \{1\}) = 0$. The beauty of Theorem 1.2 is that L_1-consistency holds without any condition on the density f, thus we have distribution estimates consistent in total variation if μ is absolutely continuous with respect to the Lebesgue measure.

Obviously one can ask for rate-of-convergence results. However, it is again impossible, since for any sequence of density estimators $\{f_n\}$ the rate of convergence of the expected L_1 error $\mathbf{E}\|f - f_n\|$ can be arbitrary slow.

Even an infinite discrete distribution with known support cannot be estimated with a guaranteed rate of convergence in total variation:

Theorem 1.3. (DEVROYE, GYÖRFI AND LUGOSI (1996)) *Assume that μ is a probability measure on the set of positive integers. For any sequence of positive numbers $b_n \le 1/16$ tending to zero, and any sequence of distribution estimators $\{\mu_n^*\}$, there always exists a probability measure μ for which*

$$\mathbf{E}V(\mu_n^*, \mu) \ge b_n \quad for\ all n. \quad \Box$$

This implies that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of distributions on the positive integers. The proof is based on Theorem 3.1, a similar slow-rate result of pattern recognition.

This theorem implies the following negative result for density estimation for a dominated class:

Theorem 1.4. (BIRGÉ (1986), DEVROYE (1983B), (1995)) *Given any sequence of density estimators* $\{f_n\}$ *and any sequence of positive numbers* $b_n \leq 1/32$ *tending to* 0 *there exists a density* f *on* $[0,1]$ *bounded by* 2 *for which*

$$\mathbf{E}\|f - f_n\| \geq b_n \quad \textit{for all} n. \quad \square$$

This means that every sequence tending to zero is an individual lower rate for the class of densities on $[0,1]$ bounded by 2. It is easy to see that the slow-rate result also holds over the class $\mathcal{D}$ of absolute continuous distributions (with the total variation loss), if we allow any distribution estimate (not only absolute continuous ones), that is, every sequence tending to zero is an individual lower rate for $\mathcal{D}$.

According to these facts, if P is the density function and the loss is the L_1-distance one can have an universally L_1-consistent density estimator, but its rate of convergence can be slow unless we have some conditions on the unknown f. Such conditions can be formulated, for example, in term of metric entropy: for the class $\mathcal{D}$ of densities, let N be the cardinality of the smallest subclass $\mathcal{A}$ of $\mathcal{D}$ with the property that for every $f \in \mathcal{D}$ there is a $g \in \mathcal{A}$ such that $\sup_{x \in \mathcal{R}^d} |f(x) - g(x)| < \epsilon$. Let $\log N$ be the ϵ-metric entropy of $\mathcal{D}$.

Theorem 1.5. (BIRGÉ (1986)) *If* $\mathcal{D}^\delta$ *is a class with densities such that its* ϵ*-metric entropy is* $O(\epsilon^{-\delta})$ *as* $\epsilon \to 0$, *then*

$$n^{-\frac{1}{2+\delta}}$$

is a minimax upper rate of convergence for $\mathcal{D}^\delta$. $\square$

In particular, for given $k \in \mathcal{N}_0$ (where $\mathcal{N}_0 = \{0, 1, 2, \ldots\}$), $0 < \beta \le 1$, $p = k + \beta$ and $M > 0$, let $\mathcal{F}^{(p,M)}$ be a class of smooth densities on a convex compact subset of $\mathcal{R}^d$, that is, for $f \in \mathcal{F}^{(p,M)}$ for every $\alpha = (\alpha_1, \ldots, \alpha_d)$ with $\alpha_i \in \mathcal{N}_0$, $\sum_{j=1}^d \alpha_j = k$

$$|D^\alpha f(x) - D^\alpha f(z)| < M\|x - z\|^\beta,$$

where D^α denotes the partial derivative belonging to α. Then $\{n^{-\frac{p}{2p+d}}\}$ is a minimax upper rate of convergence for $\mathcal{F}^{(p,M)}$ (see, e.g., Birgé (1986)). This rate is also a minimax optimal rate, but Birgé (1986) also proves the following stronger statement

Theorem 1.6. (Birgé (1986)) *For the class $\mathcal{F}^{(p,M)}$ of smooth densities on a convex compact subset of $\mathcal{R}^d$ for every sequence $\{b_n\}$*

$$b_n n^{-\frac{p}{2p+d}}$$

is an individual lower rate of convergence. □

(See Chapter 2 and 3 for corresponding results on regression function estimation and pattern recognition.)

For distribution estimation we may consider other error criteria. Such error criteria can be derived from dissimilarity measures of probability measures, like f-divergences introduced by Csiszár (1967) (see also Liese, Vajda (1987) and Vajda (1989)). The three most important f-divergences in mathematical statistics are the total variation, the information divergence and the χ^2-divergence.

If μ and ν are probability measures on $\mathcal{R}^d$ then the information divergence (or I-divergence, relative entropy, Kullback-Leibler number) of μ and ν is defined by

$$l(\mu, \nu) = I(\mu, \nu) = \sup_{\{A_j\}} \sum_j \mu(A_j) \log \frac{\mu(A_j)}{\nu(A_j)} ,$$

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

The following inequality, also called Pinsker's inequality, bounds the total variation in terms of I-divergence (cf. Csiszár (1967), Kemperman (1969) and Kullback (1967)):

$$2\{V(\mu,\nu)\}^2 \le I(\mu,\nu)\ . \tag{1.2}$$

By Pinsker's inequality (1.2), the information divergence dominates the total variation, so it follows from Theorem 1.1 that given any sequence of distribution estimators $\{\mu_n^*\}$ there exists a probability measure μ for which the sequence $\{\mu_n^*\}$ is not consistent in information divergence. The situation is even worse, a discrete distribution with known support cannot be estimated consistently in information divergence:

Theorem 1.7. (GYÖRFI, PÁLI AND VAN DER MEULEN (1994)) *Assume that μ is a probability measure on the set of positive integers. Given any sequence of distribution estimators $\{\mu_n^*\}$ there exists a probability measure μ with finite Shannon entropy $H(\mu) = -\sum_{j=1}^{\infty} \mu(\{i\}) \log \mu(\{i\})$ for which*

$$I(\mu,\mu_n^*) = \infty \quad \textit{for all } n \textit{ a.s.} \quad \Box$$

This implies $l_{\text{ind}}(\mathcal{D}, \{a_n\}) = \infty$ for any sequence $\{a_n\}$ of positive numbers, where $\mathcal{D}$ is the class of distributions on the positive integers with finite entropy. This remains true, if $\mathcal{D}$ is the class of those absolute continuous distributions on $\mathcal{R}$, whose densities have finite differential entropies and arbitrary many derivatives.

In Chapters 2, 3, and 4 we show results concerning regression function estimation, pattern recognition and functional estimation, respectively.

Chapter 2

Regression estimation

2.1 Regression problem

Let (X,Y), (X_1,Y_1), (X_2,Y_2), ... be independent identically distributed $\mathcal{R}^d\times\mathcal{R}$-valued random variables with $\mathbf{E}\{Y^2\}<\infty$. In regression analysis one wishes to estimate Y given X, that is, one wants to find a function f defined on the range of X so that $f(X)$ is "close" to Y. Assume that the main aim of the analysis is to minimize the mean squared error:

$$\inf_f \mathbf{E}\{(f(X)-Y)^2\}. \tag{2.1}$$

Let

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

be the regression function, and denote the distribution of X by μ. For $q\geq 1$ introduce

$$||f||_q \stackrel{\text{def}}{=} \left(\int |f|^q\, d\mu\right)^{1/q} = (\mathbf{E}\{|f(X)|^q\})^{1/q}$$

for the $L_q(\mu)$ norm of a real function on $\mathcal{R}^d$. $||f||$ stands for the $||f||_2$ norm.

It is well-known, that for each measurable function $f:\mathcal{R}^d\to\mathcal{R}$ the relation

$$\mathbf{E}\{(f(X)-Y)^2\}=\int |\ f(x)-m(x)\ |^2\ \mu(dx)+\mathbf{E}\{(m(X)-Y)^2\}$$

holds. Therefore the regression function m achieves the minimum in (2.1), and the mean squared error of an arbitrary function f is close to its minimum (2.1) if and only if

$$\|f-m\|^2=\int |f(x)-m(x)|^2\mu(dx)$$

is close to zero.

In the regression estimation problem, the distribution of (X, Y) (and therefore also m) is unknown. Given only the independent sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, one wants to construct an estimate

$$m_n(x) = m_n(x, D_n) : \mathcal{R}^d \times (\mathcal{R}^d \times \mathcal{R})^n \mapsto \mathcal{R}$$

of $m(x)$ such that $\|m_n - m\|^2$ is small.

Hence now the pair (X, Y) and m play the roles of X and P of Chapter 1, respectively. The loss function is the mean squared error

$$l(m_n, m) \stackrel{\text{def}}{=} \|m_n - m\|^2 .$$

The class $\mathcal{D}$ of distributions is often determined by giving a class $\mathcal{D}_1$ of allowed distributions of X, a class $\mathcal{D}_2$ of allowed regression functions and the type of the conditional distributions of Y given X. For example, $\mathcal{D}_1$ may consist of all distributions (distribution free approach), or one particular distribution (distribution sensitive approach), or all absolute continuous distributions.

2.2 Slow rate of convergence

It is well-known, that there exist universally consistent estimates, that is, which satisfy

$$\mathbf{E}\{\|m_n - m\|^2\} \to 0 \quad (n \to \infty) \tag{2.2}$$

for all distributions of (X, Y) with $\mathbf{E}\{Y^2\} < \infty$. This was first shown in Stone (1977) for nearest neighbor estimates.

A slow-rate-of-convergence result is also known and follows from a similar result of pattern recognition (see Theorem 3.1):

Theorem 2.1. *Let $\{b_n\}$ be a sequence of positive numbers converging to zero with $1/64 \geq b_1 \geq b_2 \geq \ldots$. For every sequence of regression estimates, there exists a distribution of (X, Y), such that X is uniformly distributed on $[0, 1]$, $Y = m(X)$ and*

$$\mathbf{E}\{\|m_n - m\|^2\} \geq b_n$$

for all n. □

This implies that for every sequence $\{b_n\}$ tending to zero, $l_{\text{ind}}(\mathcal{D}, \{b_n\}) = \infty$ for the class $\mathcal{D}$ of distributions with $X \sim \text{Uniform}[0,1]$ and $Y = m(X)$, hence every sequence tending to zero is an individual lower rate of convergence.

2.3 Minimax lower bounds

In Stone (1982) we find minimax optimal rates of convergence for the class $\mathcal{D}^{(p,M)}$ defined below:

DEFINITION 2.1. *For given $k \in \mathcal{N}_0$, $0 < \beta \leq 1$, $p = k + \beta$ and $M > 0$, let $\mathcal{F}^{(p,M)}$ be the set of functions $f : \mathcal{R}^d \to \mathcal{R}$ such that for every $\alpha = (\alpha_1, \dots, \alpha_d)$ with $\alpha_i \in \mathcal{N}_0$, $\sum_{j=1}^d \alpha_j = k$*

$$|D^\alpha f(x) - D^\alpha f(z)| < M\|x - z\|^\beta,$$

where D^α denotes the partial derivative belonging to α.

$\mathcal{F}^{(p,M)}$ is a class of functions "smooth" enough. In absence of other information, smoothness is a quite natural assumption. We show minimax and individual lower rates on the analogy of the one in Theorem 1.6 for the class of smooth densities.

DEFINITION 2.2. *Let $\mathcal{D}^{(p,M)}$ be the class of distributions of (X, Y) such that*
(i) X is uniformly distributed on $[0,1]^d$,
(ii) $Y = m(X) + N$, where X and N are independent and N is a standard normal random variable,
(iii) $m \in \mathcal{F}^{(p,M)}$.

It is known, that there exist estimates m_n, which satisfy

$$\limsup_{n\to\infty} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{n^{-\frac{2p}{2p+d}}} < \infty \tag{2.3}$$

(see, e.g., Barron, Birgé and Massart (1995)), hence $\{n^{-\frac{2p}{2p+d}}\}$ is a minimax upper rate of convergence for a class $\mathcal{D}^{(p,M)}$. It is also minimax optimal rate by

Theorem 2.2. *(*Stone (1982)*) The sequence*

$$\left\{a_n = n^{-\frac{2p}{2p+d}}\right\}$$

is a minimax lower rate of convergence for the class $\mathcal{D}^{(p,M)}$.

Remark. Stone (1982) also considered the estimation of the derivatives of a regression function, as well as rates for tail probabilities. □

We give a different proof from Stone's. The reason is that the new proof may be easily modified to prove the individual lower bound in Theorem 2.3. Our proof applies the following lemma:

Lemma 2.1. (Antos, Györfi and Kohler (1999)) *Let u be an l-dimensional real vector, let C be a zero mean random variable taking values in $\{-1,+1\}$, and let N be an l-dimensional standard normal random variable, independent of C, and*

$$Z = Cu + N\ .$$

Let $g^* : \mathcal{R}^l \mapsto \{-1,+1\}$ be the Bayes-decision for C based on Z, that is,

$$g^*(z) = \begin{cases} 1 & \text{if } \mathbf{P}\{C=1|Z=z\} > \mathbf{P}\{C=-1|Z=z\}\ , \\ -1 & \text{otherwise} \end{cases}$$

(see also Chapter 3). Then the error probability of g^*

$$\mathbf{P}\{g^*(Z) \neq C\} = \Phi(-\|u\|)\ ,$$

where Φ is the standard normal distribution function.

Proof. The case $u = 0$ is obvious. Assume that $u \neq 0$. Let φ be the density of an l-dimensional standard normal random variable. The conditional density of Z given $C = 1$ is $\varphi(z-u)$, and given $C = -1$ is $\varphi(z+u)$, therefore the equivalent version of the Bayes-decision is

$$g^*(z) = \begin{cases} 1 & \text{if } \varphi(z-u) > \varphi(z+u)\ , \\ -1 & \text{otherwise.} \end{cases}$$

This, together with

$$\varphi(z-u) > \varphi(z+u) \Leftrightarrow z \text{ is closer to } u \text{ than to } -u \Leftrightarrow (u,z) > 0$$

implies that $g^*(z)$ is the sign of the inner product (u,z).

Set $\tilde{N} = (u,N)/\|u\|$. Then $\tilde{N}$ is a one-dimensional standard normal variable independent of C, and $Z = Cu + N$ implies

$$\frac{(u,Z)}{\|u\|} = C \cdot \|u\| + \frac{(u,N)}{\|u\|} = C \cdot \|u\| + \tilde{N}.$$

Therefore the error of the Bayes-decision is given by

$$\begin{aligned}
\mathbf{P}\{g^*(Z) \neq C\} &= \mathbf{P}\{C=-1,(u,Z)>0\} + \mathbf{P}\{C=+1,(u,Z)<0\} \\
&= \mathbf{P}\{C=-1\} \cdot \mathbf{P}\{(u,Z)>0 | C=-1\} \\
&\quad + \mathbf{P}\{C=+1\} \cdot \mathbf{P}\{(u,Z)<0 | C=+1\} \\
&= \frac{1}{2}\mathbf{P}\{C \cdot \|u\| + \tilde{N} > 0 | C=-1\} + \frac{1}{2}\mathbf{P}\{C \cdot \|u\| + \tilde{N} < 0 | C=+1\} \\
&= \frac{1}{2}\mathbf{P}\{\tilde{N} > \|u\|\} + \frac{1}{2}\mathbf{P}\{\tilde{N} < -\|u\|\} = \Phi(-\|u\|) \ . \quad \Box
\end{aligned}$$

Proof of Theorem 2.2 First we define a subclass of distributions (X,Y) contained in $\mathcal{D}^{(p,M)}$. We pack infinitely many disjoint cubes into $[0,1]^d$ in the following way: For a given probability distribution $\{p_j\}$, let $\{B_j\}$ be a partition of $[0,1]$ such that B_j is an interval of length p_j. We pack disjoint cubes of volume p_j^d into the rectangle

$$B_j \times [0,1]^{d-1}.$$

Denote these cubes by

$$A_{j,1}, \ldots, A_{j,S_j} \ ,$$

where

$$S_j = \left\lfloor \frac{1}{p_j} \right\rfloor^{d-1} .$$

Let $a_{j,k}$ be center of $A_{j,k}$. Choose a function $g : \mathcal{R}^d \to \mathcal{R}$ such that

(I) the support of g is a subset of $[-\frac{1}{2}, \frac{1}{2}]^d$,

(II) $\int g^2 \, d\mu > 0$,

(III) $g \in \mathcal{F}^{(p, M2^{\beta-1})}$.

The subclass of regression functions is indexed by a vector

$$c = (c_{1,1}, c_{1,2}, \ldots, c_{1,S_1}, c_{2,1}, c_{2,2}, \ldots, c_{2,S_2}, \ldots)$$

of $+1$ or -1 components. Denote the set of all such vectors by $\mathcal{C}$. For $c \in \mathcal{C}$ define the function

$$m^{(c)}(x) = \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} g_{j,k}(x),$$

where

$$g_{j,k}(x) = p_j^p g(p_j^{-1}(x - a_{j,k})).$$

Then it is easy to check (cf. Stone (1982), p. 1045) that because of (III)

$$m^{(c)} \in \mathcal{F}^{(p,M)}.$$

Hence, each distribution (X, Y) with $Y = m^{(c)}(X) + N$ for some $c \in \mathcal{C}$ is contained in $\mathcal{D}^{(p,M)}$, which implies

$$\inf_{\{m_n\}} \sup_{(X,Y) \in \mathcal{D}^{(p,M)}} \limsup_{n \to \infty} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{b_n a_n}$$

$$\geq \inf_{\{m_n\}} \sup_{(X,Y): X \sim \mathrm{Unif}[0,1]^d, Y = m^{(c)}(X) + N, c \in \mathcal{C}} \limsup_{n \to \infty} \frac{\mathbf{E}\{\|m_n - m^{(c)}\|^2\}}{b_n a_n}. \quad (2.4)$$

Let m_n be an arbitrary estimate. By definition, $\{g_{j,k} : j, k\}$ is an orthogonal system in $L_2(\mu)$, therefore the projection $\hat{m}_n$ of m_n to $\{m^{(c)} : c \in \mathcal{C}\}$ is given by

$$\hat{m}_n(x) = \sum_{j,k} \hat{c}_{n,j,k} g_{j,k}(x),$$

where

$$\hat{c}_{n,j,k} = \frac{\int_{A_{j,k}} m_n(x) g_{j,k}(x) \, dx}{\int_{A_{j,k}} g_{j,k}^2(x) \, dx}.$$

Let $c \in \mathcal{C}$ be arbitrary. Then

$$\|m_n - m^{(c)}\|^2 \geq \|\hat{m}_n - m^{(c)}\|^2 = \sum_{j,k} \int_{A_{j,k}} (\hat{c}_{n,j,k} g_{j,k}(x) - c_{j,k} g_{j,k}(x))^2 \, dx$$

$$= \sum_{j,k} \int_{A_{j,k}} (\hat{c}_{n,j,k} - c_{j,k})^2 g_{j,k}^2(x) \, dx = ||g||^2 \sum_{j,k} (\hat{c}_{n,j,k} - c_{j,k})^2 p_j^{2p+d} .$$

Let $\tilde{c}_{n,j,k}$ be 1 if $\hat{c}_{n,j,k} \geq 0$ and -1 otherwise. Because of

$$|\hat{c}_{n,j,k} - c_{j,k}| \geq |\tilde{c}_{n,j,k} - c_{j,k}|/2 = I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} \ ,$$

where I_A denotes the indicator function of the event A, we get

$$\|m_n - m^{(c)}\|^2 \geq ||g||^2 \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{2p+d}.$$

This proves

$$\mathbf{E}\{\|m_n - m^{(c)}\|^2\} \geq ||g||^2 \cdot R_n(c) \ , \tag{2.5}$$

where

$$R_n(c) = \sum_{j:np_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{2p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq c_{j,k}\}. \tag{2.6}$$

(2.4) and (2.5) imply

$$\begin{aligned} &\inf_{\{m_n\}} \sup_{(X,Y) \in \mathcal{D}^{(p,M)}} \limsup_{n \to \infty} \frac{\mathbf{E}\{\|m_n - m\|^2\}}{b_n a_n} \\ \geq \ & ||g||^2 \inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \end{aligned} \tag{2.7}$$

To bound the last term, we fix a sequence $\{m_n\}$ of estimates and choose $c \in \mathcal{C}$ randomly. Let $(C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots)$ be a sequence of independent identically distributed random variables independent of (X_1, N_1), (X_2, N_2), ..., which satisfy

$$\mathbf{P}\{C_{1,1} = 1\} = \mathbf{P}\{C_{1,1} = -1\} = \frac{1}{2}.$$

Set

$$C = (C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots).$$

Next we derive a lower bound for

$$\mathbf{E}R_n(C) = \sum_{j:np_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{2p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\} \ .$$

$\tilde{c}_{n,j,k}$ can be interpreted as a decision on $C_{j,k}$ using D_n. Its error probability is minimal for the Bayes decision $\bar{C}_{n,j,k}$, which is 1 if $\mathbf{P}\{C_{j,k} = 1|D_n\} > 1/2$ and -1 otherwise (see Chapter 3), therefore

$$\mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\} \geq \mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} \ .$$

Let $X_{i_1}, \dots, X_{i_l}$ be those $X_i \in A_{j,k}$. Then

$$(Y_{i_1}, \dots, Y_{i_l}) = C_{j,k} \cdot (g_{j,k}(X_{i_1}), \dots, g_{j,k}(X_{i_l})) + (N_{i_1}, \dots, N_{i_l}),$$

while

$$(Y_1, \dots, Y_n) \setminus (Y_{i_1}, \dots, Y_{i_l})$$

depends only on $C \setminus \{C_{j,k}\}$ and on X_r's and N_r's with $r \notin \{i_1, \dots, i_l\}$, therefore is independent of $C_{j,k}$ given $X_1, \dots, X_n$. Now conditioning on $X_1, \dots, X_n$, the error of the conditional Bayes decision for $C_{j,k}$ based on $(Y_1, \dots, Y_n)$ depends only on $(Y_{i_1}, \dots, Y_{i_l})$, hence Lemma 2.1 implies

$$\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k} | X_1, \dots, X_n\} = \Phi\left(-\sqrt{\sum_{r=1}^{l} g_{j,k}^2(X_{i_r})}\right) = \Phi\left(-\sqrt{\sum_{i=1}^{n} g_{j,k}^2(X_i)}\right).$$

Since $\Phi(-\sqrt{x})$ is convex, by Jensen-inequality

$$\begin{aligned}
\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} &= \mathbf{E}\{\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k} | X_1, \dots, X_n\}\} \\
&= \mathbf{E}\left\{\Phi\left(-\sqrt{\sum_{i=1}^{n} g_{j,k}^2(X_i)}\right)\right\} \\
&\geq \Phi\left(-\sqrt{\mathbf{E}\{\sum_{i=1}^{n} g_{j,k}^2(X_i)\}}\right) \\
&= \Phi\left(-\sqrt{n\mathbf{E}\{g_{j,k}^2(X_1)\}}\right) \\
&= \Phi\left(-\sqrt{np_j^{2p+d}||g||^2}\right)
\end{aligned}$$

independently of k. Thus

$$\begin{aligned}
\mathbf{E}R_n(C) &\geq \sum_{j:np_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{2p+d} \Phi\left(-\sqrt{np_j^{2p+d}||g||^2}\right) \\
&\geq \Phi\left(-||g||\right) \sum_{j:np_j^{2p+d} \leq 1} S_j p_j^{2p+d} \\
&\geq K_1 \cdot \sum_{j:np_j^{2p+d} \leq 1} p_j^{2p+1}, \qquad (2.8)
\end{aligned}$$

where

$$K_1 = \Phi\left(-||g||\right) \left(\frac{1}{2}\right)^{d-1}.$$

Setting

$$p_j = p_{j,n} = \left(\frac{1}{n}\right)^{\frac{1}{2p+d}} \text{ for } j \le n^{\frac{1}{2p+d}} ,$$

$$\mathbf{E}R_n(C) \ge K_1 \lfloor n^{\frac{1}{2p+d}} \rfloor n^{-\frac{2p+1}{2p+d}} = K_1 a_n (1 - o(1)),$$

so

$$\limsup_{n\to\infty} \inf_{m_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \ge \limsup_{n\to\infty} \inf_{m_n} \frac{\mathbf{E}R_n(C)}{a_n} \ge K_1 > 0 \ . \tag{2.9}$$

This together with (2.7) implies the assertion. □

The vast majority of the rate-of-convergence results for nonparametric regression estimation is for bounded $|Y|$ (see, e.g., Devroye, Györfi, Krzyżak and Lugosi (1994)), so it is reasonable to consider the minimax lower bounds for such classes.

Definition 2.3. *Let $\mathcal{D}^{*(p,M)}$ be the class of distributions of (X,Y) such that*
(i') X is uniformly distributed on $[0,1]^d$,
(ii') $Y \in \{0,1\}$ a.s. (thus $m(x) = \mathbf{P}\{Y = 1|X = x\}$),
(iii') $m \in \mathcal{F}^{(p,M)}$.

It turns out that the minimax lower (and upper) rates are the same for $\mathcal{D}^{(p,M)}$ and $\mathcal{D}^{*(p,M)}$. (The lower bound follows from results on pattern recognition, Theorem 3.2).

2.4 Individual lower bounds

We will show that for every sequence $\{b_n\}$ tending to zero, $\{b_n n^{-\frac{2p}{2p+d}}\}$ is an individual lower rate of convergence for the class $\mathcal{D}^{(p,M)}$. Hence there exist individual lower rates of these classes, which are arbitrarily close to the minimax optimal rates.

Theorem 2.3. (Antos, Györfi and Kohler (1999)) *Let $\{b_n\}$ be an arbitrary positive sequence tending to zero. Then the sequence*

$$\left\{ b_n a_n = b_n n^{-\frac{2p}{2p+d}} \right\}$$

is an individual lower rate of convergence for the class $\mathcal{D}^{(p,M)}$.

REMARK 1. Applying for the sequence $\{\sqrt{b_n}\}$, Theorem 2.3 implies

$$l_{\mathrm{ind}}(\mathcal{D}^{(p,M)}, \{b_n a_n\}) = \infty \ . \ \Box$$

REMARK 2. Since $\{n^{-\frac{2p}{2p+d}}\}$ is also an individual upper rate of convergence, $\{n^{-\frac{2p}{2p+d}}\}$ is an individual optimal rate of convergence. $\Box$

The following lemma provides a simple way of proving individual lower bounds. A somewhat weaker version is implicitely used by Schuurmans (1996). A stronger form and more discussion can be found in Antos and Lugosi (1998).

Lemma 2.2. (ANTOS AND LUGOSI (1998)) *Let $\{R_n(c)\}$ be a sequence of nonnegative numbers parametrized by an abstract parameter c from a set $\mathcal{C}$. If there exists a random variable C taking its values from $\mathcal{C}$ such that the sequence*

$$\left\{\frac{\sup_{c\in\mathcal{C}} R_n(c)}{\mathbf{E}R_n(C)}\right\} \tag{2.10}$$

is bounded, then

$$\sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{\mathbf{E}R_n(C)} \geq 1,$$

that is, there exists a $c \in \mathcal{C}$ such that for every $0 < \epsilon < 1$,

$$R_n(c) > (1-\epsilon)\mathbf{E}R_n(C) \quad \textit{for infinitely many } n.$$

PROOF. By condition (2.10) Fatou's lemma may be applied to the sequence of random variables $\{R_n(C)/\mathbf{E}R_n(C)\}$:

$$\sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{\mathbf{E}R_n(C)} \geq \mathbf{E}\left\{\limsup_{n\to\infty} \frac{R_n(C)}{\mathbf{E}R_n(C)}\right\} \geq \limsup_{n\to\infty} \mathbf{E}\left\{\frac{R_n(C)}{\mathbf{E}R_n(C)}\right\} = 1. \ \Box$$

REMARK. Lemma 2.2 states that minimax lower bounds obtained by using the simplest form

$$\sup_{c\in\mathcal{C}} R_n(c) \geq \mathbf{E}R_n(C)$$

of the probabilistic method can be extended to their individual form if the randomization C does not depend on n, and, in addition, the boundedness of

$$\left\{\frac{\sup_{c\in\mathcal{C}} R_n(c)}{\mathbf{E}R_n(c)}\right\}$$

is verified. An example in Chapter 3 demonstrates that this additional condition cannot be dropped, and some kind of stability condition is necessary. $\square$

PROOF OF THEOREM 2.3. We use the notations and results of the proof of Theorem 2.2. Now we have by (2.5)

$$\begin{aligned}
&\inf_{\{m_n\}} \sup_{(X,Y)\in\mathcal{D}^{(p,M)}} \limsup_{n\to\infty} \frac{\mathbf{E}\{||m_n - m^{(c)}||^2\}}{b_n a_n} \\
&\geq \quad ||g||^2 \inf_{\{m_n\}} \sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{b_n a_n} \ . \qquad (2.11)
\end{aligned}$$

In this case we have to choose $\{p_j\}$ independently from n. Since b_n and a_n tend to zero we can take a subsequence $\{n_t\}_{t\in\mathcal{N}}$ of $\{n\}_{n\in\mathcal{N}}$ with

$$b_{n_t} \leq 2^{-t}$$

and

$$a_{n_t}^{1/2p} \leq 2^{-t} \ .$$

Define q_t such that

$$\frac{2^{-t}}{q_t} = \left\lceil \frac{2^{-t}}{a_{n_t}^{1/2p}} \right\rceil \ ,$$

and choose $\{p_j\}$ as

$$q_1, \ldots, q_1, q_2, \ldots, q_2, \ldots, q_t, \ldots, q_t, \ldots \ ,$$

where q_t is repeated $2^{-t}/q_t$ times. So

$$\begin{aligned}
\sum_{j:np_j^{2p+d}\leq 1} p_j^{2p+1} &= \sum_{t:nq_t^{2p+d}\leq 1} \frac{2^{-t}}{q_t} q_t^{2p+1} \\
&\geq \sum_{t:nq_t^{2p+d}\leq 1} b_{n_t} q_t^{2p}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t:\lceil 2^{-t} a_{n_t}^{-1/2p}\rceil \geq 2^{-t} a_n^{-1/2p}} b_{n_t} \left(\frac{2^{-t}}{\left\lceil \frac{2^{-t}}{a_{n_t}^{1/2p}} \right\rceil} \right)^{2p} \\
&\geq \sum_{t:a_{n_t} \leq a_n} b_{n_t} \left(\frac{2^{-t}}{\frac{2^{-t}}{a_{n_t}^{1/2p}} + 1} \right)^{2p} \\
&= \sum_{t:n_t \geq n} b_{n_t} \left(\frac{a_{n_t}^{1/2p}}{1 + 2^t a_{n_t}^{1/2p}} \right)^{2p} \\
&\geq \sum_{t:n_t \geq n} \frac{b_{n_t} a_{n_t}}{2^{2p}}
\end{aligned}
$$

by $a_{n_t}^{1/2p} \leq 2^{-t}$, and, in particular, for $n = n_s$, (2.8) implies

$$
\mathbf{E} R_{n_s}(C) \geq K_1 \sum_{j:n_s p_j^{2p+d} \leq 1} p_j^{2p+1} \geq \frac{K_1}{2^{2p}} \sum_{t \geq s} b_{n_t} a_{n_t} \geq \frac{K_1}{2^{2p}} b_{n_s} a_{n_s} \ . \tag{2.12}
$$

Using (2.12) one gets

$$
\begin{aligned}
\inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} &\geq \inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{b_{n_s} a_{n_s}} \\
&\geq \frac{K_1}{2^{2p}} \inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E} R_{n_s}(C)} \ .
\end{aligned}
$$

Because of (2.8) and the fact that for all $c \in \mathcal{C}$

$$
R_n(c) \leq \sum_{j:np_j^{2p+d} \leq 1} S_j p_j^{2p+d} \leq \sum_{j:np_j^{2p+d} \leq 1} p_j^{2p+1} \ ,
$$

the sequence $\{\sup_c R_n(c)/\mathbf{E} R_n(C)\}$ is bounded, so we can apply Lemma 2.2 for the subsequence $\{n_s\}$ to get

$$
\inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \geq \frac{K_1}{2^{2p}} \inf_{\{m_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E} R_{n_s}(C)} \geq \frac{K_1}{2^{2p}} > 0 \ .
$$

This, together with (2.11) implies the assertion. □

Again, the individual lower (and upper) rates are the same for $\mathcal{D}^{(p,M)}$ and $\mathcal{D}^{*(p,M)}$ (see Theorem 3.3).

Chapter 3

Pattern recognition

3.1 The pattern recognition problem

Let (X,Y),(X_1,Y_1), (X_2,Y_2), ... be independent identically distributed $\mathcal{R}^d \times \{0,1\}$-valued random variables. In pattern recognition (or classification) one wishes to decide whether Y (the label) is 0 or 1 given X (the observation), that is, one wants to find a decision function g defined on the range of X taking values 0 or 1 so that $g(X)$ equals Y with high probability. Assume that the main aim of the analysis is to minimize the probability of error:

$$\inf_g L(g) \stackrel{\text{def}}{=} \inf_g \mathbf{P}\{g(X) \neq Y\} \ . \tag{3.1}$$

Let

$$\eta(x) = \mathbf{P}\{Y=1|X=x\} = \mathbf{E}\{Y|X=x\}$$

be the a posteriori probability (or regression) function. Let

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \ , \\ 0 & \text{otherwise} \end{cases}$$

be the Bayes-decision. Let

$$L^* \stackrel{\text{def}}{=} L(g^*) = \mathbf{P}\{g^*(X) \neq Y\}$$

be the Bayes-error and denote the distribution of X by μ.

It is well-known (see Devroye et al. (1996)), that for each measurable function $g : \mathcal{R}^d \mapsto \{0,1\}$ the relation

$$L(g) - L^* = 2\int \left|\eta - \frac{1}{2}\right| I_{\{g \neq g^*\}}\, d\mu \tag{3.2}$$

holds, where I_A denotes the indicator function of the event A. Therefore the function g^* achieves the minimum in (3.1) and the minimum is L^*.

In the classification problem we consider here, the distribution of (X, Y) (and therefore also η and g^*) is unknown. Given only the independent sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, one wants to construct a decision rule

$$g_n(x) = g_n(x, D_n) : \mathcal{R}^d \times (\mathcal{R}^d \times \{0,1\})^n \mapsto \{0,1\}$$

such that

$$L(g_n) = \mathbf{P}\{g_n(X) \neq Y | D_n\}$$

is close to L^*.

Hence now the pair (X, Y) and g^* play the roles of X and P of Chapter 1, respectively. The loss function is the additional error

$$l(g_n, g^*) \stackrel{\text{def}}{=} L(g_n) - L^* \ .$$

Since now μ and η determines the distribution of (X, Y), the class $\mathcal{D}$ of distributions is often given as a product of a class $\mathcal{D}_1$ of allowed distributions of X and a class $\mathcal{D}_2$ of allowed regression functions. In this chapter we are going to examine two kinds of restricted class. First, η has to be smooth as in Chapter 2, second, η has to be $\{0,1\}$-valued and $\mathcal{D}_2$ is a Vapnik-Chervonenkis class (see Section 3.4). For minimax lower rate results on other types of distribution classes see Devroye et al. (1996) and the references therein.

If we have an estimate η_n of the regression function η and we derive a plug-in rule g_n from η_n quite naturally by

$$g_n(x) = \begin{cases} 1 & \text{if } \eta_n(x) > 1/2 \ , \\ 0 & \text{otherwise,} \end{cases}$$

then from (3.2) we get easily (see Devroye et al. (1996))

$$L(g_n) - L^* \leq 2||\eta_n - \eta||_1 \leq 2||\eta_n - \eta|| \qquad (3.3)$$

(as in Chapter 2, $||\cdot||_1$ and $||\cdot||$ denote$L_1(\mu)$ and $L_2(\mu)$ distances, respectively). This shows that if $||\eta_n - \eta|| \to 0$ then $L(g_n) \to L^*$ in the same sense, and the latter has at least the same rate, that is, in a sense, classification is not more complex than regression function estimation.

3.2 Slow rate of convergence

As a consequence of (3.3), there also exist universally consistent rules, that is, rules which satisfy

$$\mathbf{E}L(g_n) \to L^* \quad (n \to \infty)$$

for all distributions of (X, Y).

The slow-rate-of-convergence result is also known. First Cover (1968) showed that for every sequence of classification rules, for sequences $\{b_n\}$ converging to zero at arbitrary slow algebraic rates, there exists a distribution, such that $\mathbf{E}L(g_n) - L^* \geq b_n$ infinitely often. Devroye (1982) strengthened Cover's result allowing sequences tending to zero arbitrary slowly. The next result asserts that $\mathbf{E}L(g_n) - L^* \geq b_n$ for *every* n.

Theorem 3.1. (Devroye et al. (1996)) *Let $\{b_n\}$ be a sequence of positive numbers converging to zero with $1/16 \geq b_1 \geq b_2 \geq \ldots$. For every sequence of classification rules, there exists a distribution of (X, Y), such that X is uniformly distributed on $[0, 1]$, $Y = \eta(X)$ ($L^* = 0$) and*

$$\mathbf{E}L(g_n) \geq b_n$$

for all n. □

This implies that for every sequence $\{b_n\}$ tending to zero, $l_{\text{ind}}(\mathcal{D}, \{b_n\}) = \infty$ for the class $\mathcal{D}$ of distributions with $X \sim \text{Uniform}[0, 1]$ and $Y = \eta(X)$, hence every sequence tending to zero is an individual lower rate of convergence.

3.3 Lower bounds for smoothness classes

Classification is actually easier than regression estimation in the sense that if $\mathbf{E}||\eta_n - \eta||_1 \to 0$, then for the plug-in rule

$$\frac{\mathbf{E}L(g_n) - L^*}{\sqrt{\mathbf{E}\{||\eta_n - \eta||^2\}}} \to 0 \,. \tag{3.4}$$

(see Devroye et al. (1996) Chapter 6), that is, the relative expected error of g_n decreases faster than the expected $L_2(\mu)$ error of η_n. Moreover, if $||\eta_n - \eta||_1 \to 0$ a.s., then for the plug-in rule

$$\frac{L(g_n) - L^*}{||\eta_n - \eta||} \to 0 \quad \text{a.s.}$$

(see Antos (1995)), that is, the relation also holds for strong consistency. However the value of the ratio above can not be universally bounded, the convergence can be arbitrary slow. It depends on the behavior of η near $1/2$ and the rate of convergence of η_n.

3.3.1 Minimax lower bounds

Yang (1999) points out that while (3.4) holds for every fixed *distribution* for which $\{\eta_n\}$ is consistent, the minimax optimal rates of convergence for many usual *classes* are the same in classification and regression estimation. He shows many examples and some counterexamples to this phenomenon with rates of convergence in terms of metric entropy. Classification seems to have the same complexity as regression function estimation for classes which are rich near 1/2. (See also Mammen and Tsybakov (1999).)

For example, it was shown in Yang (1999) that for classes $\mathcal{D}^{*(p,M)}$, the minimax optimal rate of convergence is $n^{-\frac{p}{2p+d}}$, the same as for regression estimation (see Chapter 2). It is obviously a minimax upper rate of convergence as a consequence of regression estimation (see (2.3) in Chapter 2).

Theorem 3.2. (Yang (1999)) *The sequence*

$$\left\{a_n = n^{-\frac{p}{2p+d}}\right\}$$

is a minimax lower rate of convergence for the class $\mathcal{D}^{*(p,M)}$.

Yang's proof is based on some information-theoretic tools such as Fano's inequality. Here we give a new proof which applies the following lemma:

Lemma 3.1. (Antos (1999a)) *Let $u = (u_1, \ldots, u_l)$ be an l-dimensional real vector taking values in $[-1/4, 1/4]^l$, let C be a zero mean random variable* *taking values in* $\{-1,+1\}$*, and let* $Y_1,\ldots,Y_l$ *be indepedent binary variables given* C *with*

$$\mathbf{P}\{Y_i=1|C\}=\frac{1}{2}+Cu_i \quad i=1,\ldots,l\ .$$

Then for the error probability of the Bayes decision for C *based on* $Y=(Y_1,\ldots,Y_l)$*,*

$$L^*\geq q\left(1-\frac{5\sqrt{\sum_i u_i^2+4\sum_{i\neq i'}u_i^2u_{i'}^2}}{\log\frac{1-q}{q}}\right)$$

for any $q>0$*. In particular,*

$$L^*\geq\frac{1}{4}e^{-10\sqrt{\sum_i u_i^2+4\sum_{i\neq i'}u_i^2u_{i'}^2}}\ .$$

Proof. The Bayes decision is 1 if $\mathbf{P}\{C=1|Y\}\geq\frac{1}{2}$ and -1 otherwise. Therefore,

$$L^*=\mathbf{E}\{\min(\mathbf{P}\{C=1|Y\},\mathbf{P}\{C=-1|Y\})\}\ .$$

Denote $\min(\mathbf{P}\{C=1|Y\},\mathbf{P}\{C=-1|Y\})$ by π. One can verify that

$$\mathbf{P}\{C=1|Y\}=\frac{T}{T+1}\ ,$$

where

$$\begin{aligned}T &\stackrel{\mathrm{def}}{=} \prod_{i\leq l}\left(\frac{\frac{1}{2}+u_i}{\frac{1}{2}-u_i}I_{\{Y_i=1\}}+\frac{\frac{1}{2}-u_i}{\frac{1}{2}+u_i}I_{\{Y_i=0\}}\right)=\prod_{i\leq l}\left(\frac{\frac{1}{2}+u_i}{\frac{1}{2}-u_i}\right)^{2Y_i-1}\\ &= e^{\sum_{i\leq l}(2Y_i-1)\log\frac{1+2u_i}{1-2u_i}}=e^{\sum_{i\leq l}Z_i},\end{aligned}$$

where

$$Z_i\stackrel{\mathrm{def}}{=}(2Y_i-1)\log\frac{1+2u_i}{1-2u_i}\ .$$

For arbitrary $0<q<1/2$

$$\pi\geq q$$

if and only if

$$|\log T|\leq\log\frac{1-q}{q}\ ,$$

and therefore

$$L^*=\mathbf{E}\{\pi\}\geq q\mathbf{P}\{\pi\geq q\}=q\mathbf{P}\left\{|\log T|\leq\log\frac{1-q}{q}\right\}\ .$$

By Markov's inequality,

$$\mathbf{P}\left\{|\log T| \le \log\frac{1-q}{q}\right\} \ge 1 - \frac{\mathbf{E}\{|\log T|\}}{\log\frac{1-q}{q}} \ .$$

Moreover because of

$$|\log T| = \left|\sum_{i\le l} Z_i\right| \ ,$$

we get

$$\begin{aligned}\mathbf{E}\{|\log T|\} &= \mathbf{E}\left|\sum_{i\le l} Z_i\right| \le \sqrt{\mathbf{E}\left(\sum_{i\le l} Z_i\right)^2} = \sqrt{\mathbf{E}\left\{\sum_i Z_i^2 + \sum_{i\ne i'} Z_i Z_{i'}\right\}} \\ &= \sqrt{\sum_i \mathbf{E}\{Z_i^2\} + \sum_{i\ne i'} \mathbf{E}\{Z_i Z_{i'}\}} \ .\end{aligned}$$

For $-1/4 \le x \le 1/4$

$$\begin{aligned}\left|\log\frac{1+2x}{1-2x}\right| &= |\log(1+2x) - \log(1-2x)| = \log(1+2|x|) - \log(1-2|x|) \\ &\le 2|x| + \log 4 \cdot 2|x| \le 5|x| \ .\end{aligned}$$

Using the above inequality, we obtain on the one hand

$$\mathbf{E}\{Z_i^2\} = \log^2\frac{1+2u_i}{1-2u_i} \le 25u_i^2 \ ,$$

and on the other hand

$$\begin{aligned}&\mathbf{E}\{(2Y_i-1)(2Y_{i'}-1)\} = 4\mathbf{E}\{Y_iY_{i'}\} - 2\mathbf{E}\{Y_i\} - 2\mathbf{E}\{Y_{i'}\} + 1 \\ &= 4(\mathbf{E}\{Y_iY_{i'}|C=1\}\mathbf{P}\{C=1\} + \mathbf{E}\{Y_iY_{i'}|C=-1\}\mathbf{P}\{C=-1\}) - 1 \\ &= 4\left(\left(\frac{1}{2}+u_i\right)\left(\frac{1}{2}+u_{i'}\right)\frac{1}{2} + \left(\frac{1}{2}-u_i\right)\left(\frac{1}{2}-u_{i'}\right)\frac{1}{2}\right) - 1 \\ &= 4u_iu_{i'} \ ,\end{aligned}$$

so

$$\begin{aligned}\mathbf{E}\{Z_iZ_{i'}\} &= \mathbf{E}\{(2Y_i-1)(2Y_{i'}-1)\}\log\frac{1+2u_i}{1-2u_i}\log\frac{1+2u_{i'}}{1-2u_{i'}} \\ &= 4u_iu_{i'}\log\frac{1+2u_i}{1-2u_i}\log\frac{1+2u_{i'}}{1-2u_{i'}} \\ &\le 4|u_i|\left|\log\frac{1+2u_i}{1-2u_i}\right| |u_{i'}| \left|\log\frac{1+2u_{i'}}{1-2u_{i'}}\right| \\ &\le 100u_i^2u_{i'}^2 \ .\end{aligned}$$

Hence

$$\mathbf{E}|\log T| \;\leq\; \sqrt{25\sum_i u_i^2 + 100\sum_{i\neq i'} u_i^2 u_{i'}^2}\;.$$

Thus,

$$L^* \geq q\left(1 - \frac{\mathbf{E}\{|\log T|\}}{\log\frac{1-q}{q}}\right) \geq q\left(1 - \frac{5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}{\log\frac{1-q}{q}}\right)\;.$$

By choosing

$$q = \frac{1}{1 + e^{1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}}\;,$$

we obtain

$$\begin{aligned}
L^* \;&\geq\; \frac{1}{1+\exp(1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2})}\,\frac{1}{1+5\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}} \\
&\geq\; \frac{1}{(e+1)e^{10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}} \\
&\geq\; \frac{1}{4}e^{-10\sqrt{\sum_i u_i^2 + 4\sum_{i\neq i'} u_i^2 u_{i'}^2}}\;.
\end{aligned}$$

$\square$

PROOF OF THEOREM 3.2. Unlike Yang's proof, our proof can be easily modified to individual lower bound in Theorem 3.3. First we define a subclass of distributions (X, Y) contained in $\mathcal{D}^{*(p,M)}$. We pack infinitely many disjoint cubes into $[0,1]^d$ in the following way: For a given probability distribution $\{p_j\}$, let $\{B_j\}$ be a partition of $[0,1]$ such that B_j is an interval of length p_j. We pack disjoint cubes of volume p_j^d into the rectangle

$$B_j \times [0,1]^{d-1}.$$

Denote these cubes by

$$A_{j,1}, \ldots, A_{j,S_j}\;,$$

where

$$S_j = \left\lfloor \frac{1}{p_j} \right\rfloor^{d-1}.$$

Let $a_{j,k}$ be the center of $A_{j,k}$. Choose a function $m : \mathcal{R}^d \to [0, 1/4]$ such that
(I) the support of m is a subset of $[-\frac{1}{2}, \frac{1}{2}]^d$,
(II) $\int m \, d\mu > 0$,
(III) $m \in \mathcal{F}^{(p, M2^{\beta-1})}$.
The class of a posteriori probability functions is indexed by a vector

$$c = (c_{1,1}, c_{1,2}, \ldots, c_{1,S_1}, c_{2,1}, c_{2,2}, \ldots, c_{2,S_2}, \ldots)$$

of $+1$ or -1 components. Denote the set of all such vectors by $\mathcal{C}$. For $c \in \mathcal{C}$ define the function

$$\eta^{(c)}(x) = \frac{1}{2} + \sum_{j=1}^{\infty} \sum_{k=1}^{S_j} c_{j,k} m_{j,k}(x),$$

where

$$m_{j,k}(x) = p_j^p m(p_j^{-1}(x - a_{j,k})).$$

Then it is easy to check (cf. Stone (1982), p. 1045) that $\eta^{(c)}(x) \in [0, 1]$ for all $x \in [0, 1]^d$ and because of (III)

$$\eta^{(c)} \in \mathcal{F}^{(p,M)}.$$

Hence, each distribution (X, Y) with $Y \in \{0, 1\}$ and $\mathbf{E}\{Y|X = x\} = \mathbf{P}\{Y = 1|X = x\} = \eta^{(c)}(x)$ for all $x \in [0, 1]^d$ for some $c \in \mathcal{C}$ is contained in $\mathcal{D}^{*(p,M)}$, which implies

$$\begin{aligned} &\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}^{*(p,M)}} \frac{\mathbf{E}L(g_n) - L^*}{a_n} \\ &\geq \limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y):X\sim\mathrm{Unif}[0,1]^d, \mathbf{E}\{Y|X=x\}=\eta^{(c)}(x), c\in\mathcal{C}} \frac{\mathbf{E}L(g_n) - L^*}{a_n} \, . \end{aligned} \qquad (3.5)$$

Introduce the measure ν by

$$\nu(A) \stackrel{\text{def}}{=} \int_A \left|\eta - \frac{1}{2}\right| \, d\mu \, .$$

Then by (3.2), for an arbitrary rule g_n

$$L(g_n) - L^* = 2 \int \left|\eta - \frac{1}{2}\right| I_{\{g_n \neq g^*\}} \, d\mu = 2 \int I_{\{g_n \neq g^*\}} \, d\nu = 2 \int (g_n - g^*)^2 \, d\nu \, .$$

Let $c \in \mathcal{C}$ be arbitrary and $\eta = \eta^{(c)}$. Then

$$\nu(A) = \int_A \left|\eta^{(c)} - \frac{1}{2}\right| d\mu = \int_A \sum_{j,k} m_{j,k}\, d\mu \ .$$

By definition, $\{I_{A_{j,k}}/2 : j,k\}$ is an orthogonal system in $L_2(\nu)$, therefore the projection $\hat{g}_n - \frac{1}{2}$ of $g_n - \frac{1}{2}$ is given by

$$\hat{g}_n(x) - \frac{1}{2} = \sum_{j,k} \hat{c}_{n,j,k} \frac{I_{A_{j,k}}(x)}{2} \ ,$$

where

$$\begin{aligned} \hat{c}_{n,j,k} = \frac{\int (g_n - 1/2) I_{A_{j,k}}/2\, d\nu}{\int \left(I_{A_{j,k}}/2\right)^2 d\nu} &= 2\frac{\int_{A_{j,k}} (g_n - 1/2)\, d\nu}{\int_{A_{j,k}} 1\, d\nu} \\ = \quad 2\frac{\int_{A_{j,k}} (g_n - 1/2) m_{j,k}\, d\mu}{\int_{A_{j,k}} m_{j,k}\, d\mu} &= 2\frac{\int_{A_{j,k}} g_n m_{j,k}\, d\mu}{\int_{A_{j,k}} m_{j,k}\, d\mu} - 1 \ . \end{aligned}$$

Note that $\hat{c}_{n,j,k} \in [-1,1]$ and $g^* = \frac{1}{2} + \sum_{j,k} c_{j,k} \frac{I_{A_{j,k}}}{2}$. Thus

$$\begin{aligned} L(g_n) - L^* &= 2\int (g_n - g^*)^2\, d\nu \\ &= 2\int \left(\left(g_n - \frac{1}{2}\right) - \left(g^* - \frac{1}{2}\right)\right)^2 d\nu \\ &\geq 2\int \left(\left(\hat{g}_n - \frac{1}{2}\right) - \left(g^* - \frac{1}{2}\right)\right)^2 d\nu \\ &= 2\int \sum_{j,k} m_{j,k} (\hat{g}_n - g^*)^2\, d\mu \\ &= 2\sum_{j,k} \frac{(\hat{c}_{n,j,k} - c_{j,k})^2}{4} \int_{A_{j,k}} m_{j,k}\, d\mu \\ &= \frac{1}{2}||m||_1 \sum_{j,k} (\hat{c}_{n,j,k} - c_{j,k})^2 p_j^{p+d} \ . \end{aligned}$$

Let $\tilde{c}_{n,j,k}$ be 1 if $\hat{c}_{n,j,k} \geq 0$ and -1 otherwise. Because of

$$|\hat{c}_{n,j,k} - c_{j,k}| \geq |\tilde{c}_{n,j,k} - c_{j,k}|/2 = I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} \ ,$$

we get

$$L(g_n) - L^* \geq \frac{1}{2}||m||_1 \sum_{j,k} I_{\{\tilde{c}_{n,j,k} \neq c_{j,k}\}} p_j^{p+d} \ .$$

This proves

$$\mathbf{E}L(g_n) - L^* \geq \frac{1}{2}||m||_1 R_n(c) \ , \tag{3.6}$$

where

$$R_n(c) = \sum_{j: np_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq c_{j,k}\} \ . \tag{3.7}$$

(3.5) and (3.6) imply

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{(X,Y)\in\mathcal{D}^*(p,M)} \frac{\mathbf{E}L(g_n) - L^*}{a_n} \geq \frac{1}{2}||m||_1 \limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \ . \tag{3.8}$$

To bound the last term, we fix the rule g_n and choose $c \in \mathcal{C}$ randomly. Let $(C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots)$ be a sequence of independent identically distributed random variables independent of X_1, X_2, $\ldots$, which satisfy

$$\mathbf{P}\{C_{1,1} = 1\} = \mathbf{P}\{C_{1,1} = -1\} = \frac{1}{2} \ .$$

Set

$$C = (C_{1,1}, \ldots, C_{1,S_1}, C_{2,1}, \ldots, C_{2,S_2}, \ldots) \ .$$

Next we derive a lower bound for

$$\mathbf{E}R_n(C) = \sum_{j: np_j^{2p+d} \leq 1} \sum_{k=1}^{S_j} p_j^{p+d} \cdot \mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\}.$$

$\tilde{c}_{n,j,k}$ can be interpreted as a decision on $C_{j,k}$ using D_n. Its error probability is minimal for the Bayes decision $\bar{C}_{n,j,k}$, which is 1 if $\mathbf{P}\{C_{j,k} = 1|D_n\} \geq \frac{1}{2}$ and -1 otherwise, therefore

$$\mathbf{P}\{\tilde{c}_{n,j,k} \neq C_{j,k}\} \geq \mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} \ .$$

Let $X_{i_1}, \ldots, X_{i_l}$ be those X_i which fall in $A_{j,k}$. Then given $X_1, \ldots, X_n$

$$(Y_{i_1}, \ldots, Y_{i_l})$$

is distributed as $(Y_1, \ldots, Y_l)$ under the conditions of Lemma 3.1 with $u_r = m_{j,k}(X_{i_r})$, while

$$(Y_1, \ldots, Y_n) \setminus (Y_{i_1}, \ldots, Y_{i_l})$$

depends only on $C\setminus\{C_{j,k}\}$ and on the X_r's with $r \notin \{i_1,\dots,i_l\}$, and therefore it is conditionally independent of $C_{j,k}$ given $X_1,\dots,X_n$. Now conditioning on $X_1,\dots,X_n$, the error of the conditional Bayes decision for $C_{j,k}$ based on $(Y_1,\dots,Y_n)$ depends only on $(Y_{i_1},\dots,Y_{i_l})$, hence Lemma 3.1 implies

$$\begin{aligned}
&\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}|X_1,\dots,X_n\} \\
&\geq \frac{1}{4}e^{-10\sqrt{\sum_r m_{j,k}^2(X_{i_r})+4\sum_{r\neq r'} m_{j,k}^2(X_{i_r})m_{j,k}^2(X_{i_{r'}})}} \\
&= \frac{1}{4}e^{-10\sqrt{\sum_i m_{j,k}^2(X_i)+4\sum_{i\neq i'} m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})}} .
\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}
&\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}\} = \mathbf{E}\{\mathbf{P}\{\bar{C}_{n,j,k} \neq C_{j,k}|X_1,\dots,X_n\}\} \\
&\geq \frac{1}{4}\mathbf{E}\left\{e^{-10\sqrt{\sum_i m_{j,k}^2(X_i)+4\sum_{i\neq i'} m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})}}\right\} \\
&\geq \frac{1}{4}e^{-10\mathbf{E}\left\{\sqrt{\sum_i m_{j,k}^2(X_i)+4\sum_{i\neq i'} m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})}\right\}} \\
&\geq \frac{1}{4}e^{-10\sqrt{\sum_i \mathbf{E}\{m_{j,k}^2(X_i)\}+4\sum_{i\neq i'} \mathbf{E}\{m_{j,k}^2(X_i)m_{j,k}^2(X_{i'})\}}} \\
&= \frac{1}{4}e^{-10\sqrt{||m||^2\cdot np_j^{2p+d}+4||m||^4 n(n-1)p_j^{4p+2d}}}
\end{aligned}$$

independently of k. Thus,

$$\begin{aligned}
\mathbf{E}R_n(C) &\geq \frac{1}{4}\sum_{j:np_j^{2p+d}\leq 1}\sum_{k=1}^{S_j} p_j^{p+d}e^{-10||m||\sqrt{np_j^{2p+d}+4||m||^2 n(n-1)p_j^{4p+2d}}} \\
&\geq \frac{1}{4}e^{-10||m||\sqrt{1+4||m||^2}}\sum_{j:np_j^{2p+d}\leq 1} S_j p_j^{p+d} \\
&\geq K_1 \sum_{j:np_j^{2p+d}\leq 1} p_j^{p+1} , \qquad (3.9)
\end{aligned}$$

where

$$K_1 = \frac{1}{4}e^{-10||m||\sqrt{1+4||m||^2}}\left(\frac{1}{2}\right)^{d-1} .$$

Setting

$$p_j = p_{j,n} = \left(\frac{1}{n}\right)^{\frac{1}{2p+d}} \text{ for } j \leq n^{\frac{1}{2p+d}} ,$$

$$\mathbf{E}R_n(C) \geq K_1 \lfloor n^{\frac{1}{2p+d}} \rfloor n^{-\frac{p+1}{2p+d}} = K_1 a_n (1 - o(1)) ,$$

so

$$\limsup_{n\to\infty} \inf_{g_n} \sup_{c\in\mathcal{C}} \frac{R_n(c)}{a_n} \geq \limsup_{n\to\infty} \inf_{g_n} \frac{\mathbf{E}R_n(C)}{a_n} \geq K_1 > 0 .$$

This, together with (3.8), implies the assertion. □

3.3.2 Individual lower bounds

We show that for every sequence $\{b_n\}$ tending to zero, $\{b_n n^{-\frac{p}{2p+d}}\}$ is an individual lower rate of convergence of the class $\mathcal{D}^{*(p,M)}$. Hence there exist individual lower rates of these classes, which are arbitrarily close to the minimax optimal rates.

Both these individual lower rates and the ones in Chapter 2 are optimal (see (2.3) and also Remark 2 below), hence we extended Yang's observation for the individual rates for these classes.

Our results also imply that the ratio $(\mathbf{E}L(g_n) - L^*)/\sqrt{\mathbf{E}\{||\eta_n - \eta||^2\}}$ can tend to zero arbitrary slowly (even for a fixed sequence $\{\eta_n\}$).

Theorem 3.3. (Antos (1999a)) *Let $\{b_n\}$ be an arbitrary positive sequence tending to zero. Then the sequence*

$$\left\{ b_n a_n = b_n n^{-\frac{p}{2p+d}} \right\}$$

is an individual lower rate of convergence for the class $\mathcal{D}^{*(p,M)}$.

Remark 1. Applying for the sequence $\{\sqrt{b_n}\}$, Theorem 3.3 implies

$$l_{\text{ind}}(\mathcal{D}^{*(p,M)}, \{b_n a_n\}) = \infty . \quad \square$$

Remark 2. Since $\{n^{-\frac{p}{2p+d}}\}$ is also an individual upper rate of convergence, $\{n^{-\frac{p}{2p+d}}\}$ is an individual optimal rate of convergence. Moreover (2.3) in Chapter 2 (more exactly, its version for $\mathcal{D}^{*(p,M)}$) and (3.4) imply that for classification

$$l_{\text{ind}}(\mathcal{D}^{*(p,M)}, \{n^{-\frac{p}{2p+d}}\}) = 0,$$

which shows that Theorem 3.3 cannot be improved by dropping b_n. This shows the strange nature of individual lower bounds, that while every sequence tending to zero faster than $\{n^{-\frac{p}{2p+d}}\}$ is an individual lower rate for $\mathcal{D}^{(p,M)}$, $\{n^{-\frac{p}{2p+d}}\}$ itself is not that. □

Proof. We use the notation and results of the proof of Theorem 3.2. Now we have by (3.6)

$$\begin{aligned}
&\inf_{\{g_n\}} \sup_{(X,Y)\in\mathcal{D}^{*(p,M)}} \limsup_{n\to\infty} \frac{\mathbf{E}L(g_n) - L^*}{b_n a_n} \\
&\geq \frac{1}{2}||m||_1 \inf_{\{g_n\}} \sup_{c\in\mathcal{C}} \limsup_{n\to\infty} \frac{R_n(c)}{b_n a_n} . \qquad (3.10)
\end{aligned}$$

In this case we have to choose $\{p_j\}$ independently from n. Since b_n and a_n tend to zero we can take a subsequence $\{n_t\}_{t\in\mathcal{N}}$ of $\{n\}_{n\in\mathcal{N}}$ with

$$b_{n_t} \leq 2^{-t}$$

and

$$a_{n_t}^{1/p} \leq 2^{-t} .$$

Define q_t such that

$$\frac{2^{-t}}{q_t} = \left\lceil \frac{2^{-t}}{a_{n_t}^{1/p}} \right\rceil ,$$

and choose $\{p_j\}$ as

$$q_1, \ldots, q_1, q_2, \ldots, q_2, \ldots, q_t, \ldots, q_t, \ldots ,$$

where q_t is repeated $2^{-t}/q_t$ times. So

$$\begin{aligned}
\sum_{j:np_j^{2p+d}\leq 1} p_j^{p+1} &= \sum_{t:nq_t^{2p+d}\leq 1} \frac{2^{-t}}{q_t} q_t^{p+1} \\
&\geq \sum_{t:nq_t^{2p+d}\leq 1} b_{n_t} q_t^p \\
&= \sum_{t:\lceil 2^{-t}a_{n_t}^{-1/p}\rceil \geq 2^{-t}a_n^{-1/p}} b_{n_t} \left(\frac{2^{-t}}{\left\lceil \frac{2^{-t}}{a_{n_t}^{1/p}} \right\rceil} \right)^p
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{t: a_{n_t} \leq a_n} b_{n_t} \left(\frac{2^{-t}}{\frac{2^{-t}}{a_{n_t}^{1/p}} + 1} \right)^p \\
&= \sum_{t: n_t \geq n} b_{n_t} \left(\frac{a_{n_t}^{1/p}}{1 + 2^t a_{n_t}^{1/p}} \right)^p \\
&\geq \sum_{t: n_t \geq n} \frac{b_{n_t} a_{n_t}}{2^p}
\end{aligned}
$$

by $a_{n_t}^{1/p} \leq 2^{-t}$, and, in particular, for $n = n_s$, (3.9) implies

$$
\mathbf{E} R_{n_s}(C) \geq K_1 \sum_{j: n_s p_j^{2p+d} \leq 1} p_j^{p+1} \geq \frac{K_1}{2^p} \sum_{t \geq s} b_{n_t} a_{n_t} \geq \frac{K_1}{2^p} b_{n_s} a_{n_s} . \tag{3.11}
$$

Using (3.11) one gets

$$
\begin{aligned}
\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} &\geq \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{b_{n_s} a_{n_s}} \\
&\geq \frac{K_1}{2^p} \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E} R_{n_s}(C)} .
\end{aligned}
$$

Because of (3.9) and the fact that for all $c \in \mathcal{C}$

$$
R_n(c) \leq \sum_{j: n p_j^{2p+d} \leq 1} S_j p_j^{p+d} \leq \sum_{j: n p_j^{2p+d} \leq 1} p_j^{p+1} ,
$$

the sequence $\{\sup_c R_n(c)/\mathbf{E} R_n(C)\}$ is bounded, so we can apply Lemma 2.2 for the subsequence $\{n_s\}$ to get

$$
\inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{n \to \infty} \frac{R_n(c)}{b_n a_n} \geq \frac{K_1}{2^p} \inf_{\{g_n\}} \sup_{c \in \mathcal{C}} \limsup_{s \to \infty} \frac{R_{n_s}(c)}{\mathbf{E} R_{n_s}(C)} \geq \frac{K_1}{2^p} > 0 .
$$

This, together with (3.10) implies the assertion. □

3.4 Lower bounds for VC classes

In this section we are concerned with the special case of concept learning (see Devroye et al. (1996) and the references there). Let $\mathcal{C}$ be a class of subsets of $\mathcal{R}^d$. Members of $\mathcal{C}$ are called concepts, and $\mathcal{C}$ is a concept class. Let $\mathcal{D}$ be the class of distributions of (X, Y) such that $Y = I_{\{X \in C\}}$, $C \in \mathcal{C}$. Hence in this case an unknown concept (or target) $C \in \mathcal{C}$ is to be learnt based on the data

$$D_n = ((X_1, I_{\{X_1 \in C\}}), \ldots, (X_n, I_{\{X_n \in C\}})).$$

For these distributions $L^* = 0$, thus $l(g, g^*) = L(g)$. The joint distribution of (X, Y) is determined by the pair (μ, C), which will be referred to as a *distribution-target pair*.

The expected probability of error is a useful quantity in describing the behavior of $L(g_n)$. However, it is rather the tail probabilities

$$\mathbf{P}\{L(g_n) \geq \epsilon\} \ , \quad \epsilon \in [0, 1]$$

that completely describe the distribution of the probability of error. In this section we are also concerned with tail probabilities.

3.4.1 Minimax lower bounds

The minimax behavior of the expected probability of error has been thoroughly studied. It is a beautiful fact that for a given n, the minimax expected probability of error is basically determined by the VC *dimension* V of the class $\mathcal{C}$, and it is insensitive to other properties of $\mathcal{C}$. V is defined as the largest integer $k \geq 1$ with $s(k) = 2^k$, where the k-th *shatter coefficient* $s(k)$ of the class $\mathcal{C}$ is defined as the maximal number of different sets in

$$\{\{x_1, \ldots, x_k\} \cap C; C \in \mathcal{C}\},$$

where the maximum is taken over all $x_1, \ldots, x_k \in \mathcal{R}^d$. If $s(k) = 2^k$ for all k, then, by definition, $V = \infty$. If $V < \infty$, then $\mathcal{C}$ is said to be a Vapnik-Chervonenkis class.

Haussler, Littlestone, and Warmuth (1994) showed that there exists a learning rule such that for all distribution-target pairs,

$$\mathbf{E}L(g_n) \leq \frac{V}{n} \ , \tag{3.12}$$

that is, $\{V/n\}$ is a minimax upper rate of convergence for the class $\mathcal{D}$. It is also minimax optimal rate by

Theorem 3.4. (Vapnik and Chervonenkis (1974)) *For every $n \geq V - 1$, and every classifier g_n, there exists a distribution-target pair such that*

$$\mathbf{E}L(g_n) \geq \frac{V-1}{2en}\left(1 - \frac{1}{n}\right) ,$$

that is, the sequence $\{V/n\}$ is a minimax lower rate of convergence for the class $\mathcal{D}$. □

(3.12) and Theorem 3.4 together state that for $V \geq 2$

$$\frac{1}{4e} \leq l_{\mathrm{mm}}(\mathcal{D}, \{V/n\}) \leq 1 .$$

The minimax problem for the tail probabilities is a more interesting (and harder) problem. Here one is interested in the quantity

$$\inf_{g_n} \sup_{(\mu,C)} \mathbf{P}\{L(g_n) \geq \epsilon\} ,$$

if n, ϵ and $\mathcal{C}$ are given. Now the loss function is the indicator

$$l(\{g_n\}, g^*) \stackrel{\text{def}}{=} I_{\{2\|g_n - g^*\|_{L_1(\nu)} \geq \epsilon\}} ,$$

which depends on n if $\epsilon = \epsilon_n$. The VC dimension also appears in minimax upper and lower bounds for the tail probabilities. For example, results of Anthony et al. (1993) and Lugosi (1995) (see also Vapnik and Chervonenkis (1974), Blumer et al. (1989)) state that if g_n is any classifier such that $g_n(X_i) = I_{\{X_i \in C\}}$ for all $i = 1, \ldots, n$, and $\{x : g_n(x) = 1\} \in \mathcal{C}$, then for $n\epsilon > V + 2$,

$$\mathbf{P}\{L(g_n) \geq \epsilon\} \leq 2\left(\left\lceil \frac{n^2\epsilon}{V} \right\rceil e^2\right)^V e^{-n\epsilon} , \tag{3.13}$$

that is, $\{(n^2\epsilon/V)^V e^{-n\epsilon}\}$ is a minimax upper rate of convergence for the class $\mathcal{D}$. Corresponding minimax lower bounds were first proved by Blumer, Ehrenfeucht, Haussler, and Warmuth (1989), and Ehrenfeucht, Haussler, Kearns, and Valiant (1989).

Theorem 3.5. (DEVROYE AND LUGOSI (1995), SEE ALSO DEVROYE ET AL. (1996)) *For any classifier g_n, there exists a distribution-target pair such that*

$$\mathbf{P}\{L(g_n) \geq \epsilon\} \geq \frac{1}{2e\sqrt{\pi(V-1)}} \left(\frac{2ne\epsilon}{V-1}\right)^{(V-1)/2} e^{\frac{-4n\epsilon}{1-4\epsilon}},$$

whenever $V \geq 2$, $n \geq V-1$ and $\epsilon < 1/4$, that is, the sequence $\left\{\left(\frac{n\epsilon}{V}\right)^{\frac{V-1}{2}} e^{\frac{-4n\epsilon}{1-4\epsilon}}\right\}$ is a minimax lower rate of convergence for the class $\mathcal{D}$. □

The combination of (3.13) and Theorem 3.5 yield that for any concept class $\mathcal{C}$, with $V \geq 2$, $n\epsilon > V+2$ and $\epsilon < 1/4$,

$$\begin{aligned} &\frac{1}{2e\sqrt{\pi(V-1)}} \left(\frac{2ne\epsilon}{V-1}\right)^{(V-1)/2} e^{\frac{-4n\epsilon}{1-4\epsilon}} \leq \\ &\leq \inf_{g_n} \sup_{(\mu,C)} \mathbf{P}\{L(g_n) \geq \epsilon\} \leq 2\left(\left\lceil \frac{n^2\epsilon}{V} \right\rceil e^2\right)^V e^{-n\epsilon}. \end{aligned}$$

3.4.2 Counter-examples

As mentioned in Chapter 1, the lower bound of Theorem 3.4 does not exclude the possibility that there exists a sequence of classifiers $\{g_n\}$ such that for *every* μ and C the expected error $\mathbf{E}L(g_n)$ decreases at an exponential rate in n. Indeed, it is easy to see that such classes exist with arbitrarily large, and even with infinite, VC dimension (see Theorem 3.6 and 3.7 below). Schuurmans (1996) studied the question when such exponential decrease occurs, and characterized it among certain "one-dimensional" problems. We illustrate through a simple example why it is impossible to give an individual extension of the lower bound of Theorem 3.4 for all VC classes. It can be seen similarly that no individual extension of Theorem 3.5 can be given for all VC classes either.

Let $\mathcal{C}$ be any class containing finitely many concepts. Then consider a learning rule that selects a concept C_n from $\mathcal{C}$ which is consistent with the data D_n, that is, $g_n(x) = I_{\{x \in C_n\}}$ for some $C_n \in \mathcal{C}$, and

$$g_n(X_i) = I_{\{X_i \in C\}} \quad \text{for all } i = 1, \ldots, n,$$

where $C \in \mathcal{C}$ is the true concept. Then (3.13) implies that

$$\mathbf{E}L(g_n) \le \frac{\log|\mathcal{C}|+1}{n} \ .$$

The beauty of this bound is that it is independent of the distribution-target pair, and that it is essentially the best such bound by Theorem 3.4. However, for *all* distribution-target pairs, the error decreases at a much faster rate. This can be seen from the simple fact that g_n can only make an error if there is at least one concept $C' \in \mathcal{C}$ with $\mu(C' \triangle C) > 0$ such that no one of $X_1, \ldots, X_n$ falls in the symmetric difference of C' and C. The probability of this event is at most

$$\sum_{C' \in \mathcal{C}:\mu(C' \triangle C)>0} (1-\mu(C' \triangle C))^n \le |\mathcal{C}| \max_{C' \in \mathcal{C}:\mu(C' \triangle C)>0} (1-\mu(C' \triangle C))^n \ ,$$

which converges to zero exponentially rapidly. Since a finite concept class can have an arbitrary VC dimension, this proves the following:

Theorem 3.6. (ANTOS AND LUGOSI (1998)) *Let V be an arbitrary positive integer. There exists a class $\mathcal{C}$ with* VC *dimension V and a corresponding sequence of learning rules $\{g_n\}$ such that for all distribution-target pairs (μ, C) with $C \in \mathcal{C}$ and for all n,*

$$\mathbf{E}L(g_n) \le a \cdot b^n \ ,$$

where $b < 1$. The positive constants a and b depend on the distribution-target pair. □

In other words, while $\{V/n\}$ is a minimax optimal rate for this class, for every subexponential sequence $\{a_n\}$ (e.g., $a_n = e^{-\sqrt{n}}$) $l_{\text{ind}}(\mathcal{D}, \{a_n\}) = 0$, that is, $\{a_n\}$ is an individual upper rate and cannot be an individual lower rate of convergence.

If a concept class $\mathcal{C}$ is finite, its n-th shatter coefficient $s(n)$ is bounded above by $|\mathcal{C}|$ for all n, that is, the shatter coefficients do not increase with n for large n. In such cases it is not surprising that the error can decrease at an exponential rate for all distribution-target pairs. It is natural to ask whether the growth of $s(n)$ determines the rate of convergence of the error. This conjecture is false, and in fact, we may have an exponential rate of convergence for all distribution-target pairs even for classes with infinite VC dimension (for which $s(n) = 2^n$ for all n):

Theorem 3.7. (Antos and Lugosi (1998)) *There exists a class $\mathcal{C}$ with $V = \infty$ and a corresponding sequence of learning rules $\{g_n\}$, such that for all distribution-target pairs and for all n,*

$$\mathbf{E}L(g_n) \leq a \cdot b^n ,$$

where $b < 1$. The positive constants a and b depend on the distribution-target pair.

Proof. Let $d = 1$ and let $\mathcal{C}$ contain all finite subsets of $\mathcal{R}$. Let $g_n(x) = 1$ if and only if there exists an X_i such that $x = X_i$ and $I_{\{X_i \in C\}} = 1$. $\square$

Now while $\{1\}$ is a minimax optimal rate for this class, every subexponential sequence is an individual upper (and not lower) rate of convergence.

The following example demonstrates the fact that the stability condition in Lemma 2.2 cannot be dropped. Let $\mathcal{C} = \{\{i\} : i \in \mathcal{N}\}$ be the class of one-point concepts on the domain $\mathcal{N}$ of positive integers. Let $\{g_n\}$ be an arbitrary sequence of learning rules, and for $z \in \mathcal{N}$, define $R_n(z) = \mathbf{P}\{g_n(X, D_n(z)) \neq Y(z)\}$, where $Y(z) = I_{\{X=z\}}$ and $D_n(z) = ((X_1, I_{\{X_1=z\}}), \ldots, (X_n, I_{\{X_n=z\}}))$.

Let $\mathbf{P}\{X = i\} = c/(i \log^2 i)$ for an appropriate normalizing constant c, and introduce the random variable Z distributed as X, and independent of $X, X_1, \ldots, X_n$. Using a similar argument as in the proof of Theorem 3.8 below, one sees, for example, that for every n, $\mathbf{E}R_n(Z) \geq \text{const.}/n \log^4(n+1)$. However, similarly to the proof of Theorem 3.7, one can show easily that there exists a sequence $\{g_n\}$ such that for every z, $R_n(z)$ converges to zero exponentially rapidly. This demonstrates the fact that a minimax lower bound for $R_n(z)$ cannot necessarily be converted into an individual lower bound, even if the randomization Z is independent of n. An additional condition, such as the boundedness of

$$\frac{\sup_z R_n(z)}{\mathbf{E}R_n(Z)}$$

needs to be satisfied.

Finding a fixed random variable Z such that $\mathbf{E}R_n(Z) \geq a_n$ for all n, is useful in a different situation, even if the additional stability property above cannot be verified. It allows us to derive lower bounds for the cumulative error (see Section 3.4.4). In particular, in such a case we have, for every n, that

$$\sup_{z \in \mathcal{Z}} \left(\sum_{i=0}^{n} R_i(z) \right) \geq \sum_{i=0}^{n} \mathbf{E}R_i(Z) \ . \tag{3.14}$$

3.4.3 Individual lower bounds

Because of the reason mentioned above, extending the lower bounds of Theorem 3.4 and 3.5 to their individual forms is clearly not possible for all VC classes. However, the extension is possible for many important geometric concept classes, and the role of the VC dimension is played by the number of parameters of the class, which, in all of our examples, is closely related to the VC dimension of the class. Thus, the situation here significantly differs from that of the usual minimax theory, where a single combinatorial parameter—the VC dimension—completely determines the behavior of the concept class.

In this subsection we provide examples of concept classes for which the minimax lower bound of Theorem 3.4 for the expected probability of error $\mathbf{E}L(g_n) = \mathbf{P}\{g_n(X) \neq I_{\{X \in C\}}\}$ can be extended to its individual version. All examples shown here are based on lower bounds obtained for a very simple concept class introduced in Haussler et al. (1994): The class $\mathcal{C}_k$ of unions of k initial segments is defined as follows: let $\mathcal{X} = [0,1] \times \{1, 2, \ldots, k\}$ be the support of X, and

$$\mathcal{C}_k = \left\{ \bigcup_{j=1}^{k} ([0, z_j] \times \{j\}) : z \in [0,1]^k \right\} \ . \tag{3.15}$$

The class $\mathcal{C}_k$ is therefore parametrized by a vector of k parameters: $z = (z_1, \ldots, z_k) \in [0,1]^k$. Clearly, the VC dimension of $\mathcal{C}_k$ is also k, thus the following result states that there always exists a distribution-target pair such that the error is essentially within a factor of two of the upper bound of (3.12) infinitely many times.

Theorem 3.8. (SCHUURMANS (1996), ANTOS AND LUGOSI (1998)) *Let μ be the uniform distribution on $\mathcal{X}$. For every sequence of learning rules $\{g_n\}$, there exist a $C \in \mathcal{C}_k$ such that if C is the "true" concept, then for all $0 < \epsilon < 1$,*

$$\mathbf{E}L(g_n) > (1-\epsilon)\frac{k}{2n} \quad \text{for infinitely many } n,$$

that is, $l_{\text{ind}}(\mathcal{D}, \{k/2n\}) \geq 1$ and $\{k/n\}$ is an individual lower rate for $\mathcal{D}$.

REMARK. Haussler et al. (1994, Theorem 3.2) showed for the class $\mathcal{C}_k$ of unions of k initial segments that for every learning rule, and for every n, there exists a $C \in \mathcal{C}_k$ such that

$$\mathbf{E}L(g_n) \geq \frac{k}{2n} - O\left(n^{-2}\right) .$$

Furthermore, in their proof of this lower bound, the randomization is independent of n. To make the proof of Theorem 3.8 short, we use many elements of the proof of the above inequality. □

REMARK. Note that the lower bound $k/(2n) - O\left(n^{-2}\right)$ for the minimax expected error is better in the constant factor than the bound of Theorem 3.4. However, it is less general, since it does not apply to any VC class. □

REMARK. It is clear from the proof of the theorem that the uniform distribution may be replaced by any nonatomic distribution on $\mathcal{X}$. □

PROOF. Let $z \in [0,1]^k$ be the parameter that determines $C \in \mathcal{C}_k$. First we introduce some notation. Let $Y(z) = I_{\{X \in C\}}$, $Y_i(z) = I_{\{X_i \in C\}}$, and

$D_n(z) = ((X_1, Y_1(z)), \ldots, (X_n, Y_n(z)))$. Denote $X = \langle U, M \rangle$ so that U is uniformly distributed on $[0,1]$, M is uniform on $\{1, \ldots, k\}$, and U and M are independent. Introduce

$$l = \max\{u \in [0,1] : u \leq U \text{ and } \langle u, M \rangle \in \{X_1, \ldots, X_n\} \cup \langle 0, M \rangle\}$$

and

$$r = \min\{u \in [0,1] : u \geq U \text{ and } \langle u, M \rangle \in \{X_1, \ldots, X_n\} \cup \langle 1, M \rangle\} ,$$

that is, l and r are the left and right neighbors of U among the data points falling on the M-th segment. Finally, define the following (random) sets of parameters:

$$L_n = \left\{z \in [0,1]^k : z_M \in [l, U)\right\} \quad \text{and} \quad R_n = \left\{z \in [0,1]^k : z_M \in [U, r)\right\}.$$

Clearly,

$$\mathbf{E}L(g_n) \geq R_n(z) \stackrel{\text{def}}{=} \mathbf{P}\{g_n(X, D_n(z)) \neq Y(z), z \in L_n \cup R_n\} .$$

We apply Lemma 2.2 for $R_n(z)$. (The reason why we do not define $R_n(z)$ as the expected probability of error $\mathbf{E}L(g_n)$ itself is that this is the only way we can ensure that the additional stability property (2.10) required by Lemma 2.2 holds.) We will show that if the random vector $Z = (Z_1, \ldots, Z_k)$ is uniformly distributed on $[0,1]^k$ and independent of $X, X_1, \ldots, X_n$, then

$$\mathbf{E}\{R_n(Z)\} \geq \frac{k}{2(n+1)} - \frac{k^2}{2(n+1)(n+2)}\left(1 - \left(1 - \frac{1}{k}\right)^{n+2}\right) , \tag{3.16}$$

and $\{\sup_z R_n(z)/\mathbf{E}\{R_n(Z)\}\}$ is bounded, from which the theorem follows.

First we prove the lower bound for the expected value of $R_n(Z)$. By the independence of Z and $X, X_1, \ldots, X_n$,

$$R_n(Z) = \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n | Z\} , \tag{3.17}$$

and

$$\begin{aligned} \mathbf{E}\{R_n(Z)\} &= \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n\} \\ &= \mathbf{E}\left\{\mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z), Z \in L_n \cup R_n | X, X_1, \ldots, X_n\}\right\} \end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\Big\{\mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z) | Z \in L_n, X, X_1, \ldots, X_n\} \\
&\qquad \mathbf{P}\{Z \in L_n | X, X_1, \ldots, X_n\} \\
&\quad + \mathbf{P}\{g_n(X, D_n(Z)) \neq Y(Z) | Z \in R_n, X, X_1, \ldots, X_n\} \\
&\qquad \mathbf{P}\{Z \in R_n | X, X_1, \ldots, X_n\}\Big\} \\
&\geq \mathbf{E}\Big\{\min(\mathbf{P}\{Z \in L_n | X, X_1, \ldots, X_n\}, \mathbf{P}\{Z \in R_n | X, X_1, \ldots, X_n\})\Big\} \\
&= \mathbf{E}\left\{\min(U - l, r - U)\right\} \\
&= \frac{k}{2(n+1)} - \frac{k^2}{2(n+1)(n+2)}\left(1 - \left(1 - \frac{1}{k}\right)^{n+2}\right),
\end{aligned}
$$

where the last equality follows from direct calculation, which is detailed in the proof of Theorem 3.2 in Haussler et al. (1994).

On the other hand, we observe that for each fixed $z \in [0,1]^k$,

$$
R_n(z) \leq \mathbf{P}\{z \in L_n \cup R_n\} \leq \frac{2k}{n+1}. \tag{3.18}
$$

This may be seen by conditioning on the *set* $\{X, X_1, \ldots, X_n\}$, and observing that since $X, X_1, \ldots, X_n$ are i.i.d., the probability remains the same by permuting them. Of the $(n+1)!$ permutations, there are at most $2kn!$ such that X is a neighbor of one of the z_i's. Therefore,

$$
\frac{\sup_z R_n(z)}{\mathbf{E} R_n(Z)} \leq 4 + o(1),
$$

so the condition of Lemma 2.2 is satisfied, and the proof of the theorem is complete. □

We extend Theorem 3.8 to more general classes of geometric concepts by embedding, showing that the above lower bound remains true for many other important classes of "dimension" k. (Here by dimension we mean the number of parameters of the class, which, in most of our cases, essentially coincides with the VC dimension of the class.) These examples include the class of halfspaces, the class of d-dimensional intervals, the class of euclidean balls, the class of all ellipsoids, certain classes of neural networks, etc.

For example, we have the following straightforward corollary of Theorem 3.8.

COROLLARY 3.1. (ANTOS AND LUGOSI (1998)) *Let $\mathcal{C}$ be a class of concepts. If there exist invertible measurable mappings $f_1, \ldots, f_k : [0,1] \to \mathcal{R}^d$ such that the sets $f_i([0,1])$ are disjoint and for all $z = (z_1, \ldots, z_k) \in [0,1]^k$ there exists a $C \in \mathcal{C}$ with*

$$C \cap (f_1([0,1]) \cup \cdots \cup f_k([0,1])) = f_1([0,z_1]) \cup \cdots \cup f_k([0,z_k]),$$

then for any sequence of classifiers $\{g_n\}$, there exists a distribution-target pair (μ, C) with μ concentrated on $f_1([0,1]) \cup \cdots \cup f_k([0,1])$ and $C \in \mathcal{C}$ such that for all $0 < \epsilon < 1$,

$$\mathbf{E}L(g_n) > (1-\epsilon)\frac{k}{2n} \quad \textit{for infinitely many } n,$$

that is, $l_{\mathrm{ind}}(\mathcal{D}, \{k/2n\}) \geq 1$ and $\{k/n\}$ is an individual lower rate for $\mathcal{D}$. $\square$

The above corollary may be applied to many important geometric concept classes. Below we give a short list of examples. The proofs are quite straightforward, most of them can be found in Haussler et al. (1994, p.279).

1. If $\mathcal{C}$ is the class of subsets of $\mathcal{R}$ that can be written as a union of m intervals, then $k = 2m$.

2. $k = d$ for the class of d-dimensional octants:

$$\left\{x \in \mathcal{R}^d : x_i \leq a_i, i = 1, \ldots, d\right\}, \quad a_1, \ldots, a_d \in \mathcal{R}\ ,$$

where $x_1, \ldots, x_d$ are the components of the vector x.

3. $k = 2d$ for the class of d-dimensional intervals:

$$\left\{x \in \mathcal{R}^d : a_i \leq x_i \leq b_i, i = 1, \ldots, d\right\}\ ,$$

where $a_1, b_1, \ldots, a_d, b_d \in \mathcal{R}$.

4. $k = d$ if $\mathcal{C}$ is the class of halfspaces of $\mathcal{R}^d$, that is, sets of the form

$$\left\{x : \sum_{i=1}^{d} a_i x_i + a_0 \geq 0\right\}, \quad a_0, a_1, \ldots, a_d \in \mathcal{R}\ .$$

5. $k = d$ if

$$\mathcal{C} = \left\{ \left\{ x \in \mathcal{R}^d : \prod_{i=1}^{d} (x_i - a_i) \geq 0 \right\} : \quad a_1, \ldots, a_d \in \mathcal{R} \right\} .$$

6. $k = d+1$ for the class of balls in $\mathcal{R}^d$:

$$\left\{ x \in \mathcal{R}^d : \sum_{i=1}^{d} (x_i - a_i)^2 \leq r \right\} ,$$

where $a_1, \ldots, a_d, r \in \mathcal{R}, r \geq 0$.

7. $k = 2d$ for the class of all d-dimensional ellipsoids:

$$\left\{ x \in \mathcal{R}^d : \sum_{i=1}^{d} \frac{(x_i - a_i)^2}{b_i} \leq 1 \right\} ,$$

where $a_1, b_1, \ldots, a_d, b_d \in \mathcal{R}$.

8. $k = md$ for the class of convex polyhedra of m faces in $\mathcal{R}^d$.

9. $k = md$ for the class $\mathcal{C}$ of all neural network classifiers on $\mathcal{R}^d$ with m hidden nodes in their single hidden layer, that is, each $C \in \mathcal{C}$ is of the form

$$\left\{ x : \sum_{i=1}^{m} a_i \sigma(b_i x^T + c_i) + a_0 \geq 0 \right\} ,$$

where $a_0, \ldots, a_m, c_1, \ldots, c_m \in \mathcal{R}$, $b_1, \ldots, b_m \in \mathcal{R}^d$, and σ is the threshold sigmoid $\sigma(x) = I_{\{x>0\}}$. (x^T denotes the transpose of a vector x.)

Remark. Based on Corollary 3.1, we may define a new "dimension" Δ for a concept class $\mathcal{C}$ as follows: let Δ be the largest integer k such that there exist k invertible measurable mappings $f_1, \ldots, f_k : [0,1] \to \mathcal{R}^d$ such that the sets $f_i([0,1])$ are disjoint and for all $z = (z_1, \ldots, z_k) \in [0,1]^k$ there exists a $C \in \mathcal{C}$ with

$$C \cap (f_1([0,1]) \cup \cdots \cup f_k([0,1])) = f_1([0,z_1]) \cup \cdots \cup f_k([0,z_k]).$$

If no such mapping exists then $\Delta = 0$, and if for each k there are k mappings with the above property, then $\Delta = \infty$.

Corollary 3.1 shows the relation of Δ to individual lower bounds for the expected error. Lower bounds for Δ may be obtained in specific cases by construction. For upper bounds, note that it is easy to see that $\Delta \le V$, since a set $\{x_1, \ldots, x_\Delta\}$ is shattered by $\mathcal{C}$ if for each $i \le \Delta$, $x_i \in f_i([0,1])$. Further, it is easy to see that for each n,

$$s(n) \ge \left\lfloor \frac{n}{\Delta} \right\rfloor^{\Delta}$$

(just put at least $\lfloor n/\Delta \rfloor$ of the n points on each image $f_i([0,1])$ of the segment $[0,1]$, $i = 1, \ldots, \Delta$), which means that also $\Delta \le D$, where D is the *Assouad density* of $\mathcal{C}$, defined as

$$D = \inf \left\{ r > 0 : \sup_n \frac{s(n)}{n^r} < \infty \right\} ,$$

see Assouad (1983). (It is well-known that $D \le V$, $D < \infty$ if and only if $V < \infty$, and that for each k there exists a class $\mathcal{C}$ with $V = k$ and $D = 0$.) Thus, we have

$$\Delta \le D \le V .$$

On the other hand, it follows from Theorem 3.7 and Corollary 3.1 that there exists a class $\mathcal{C}$ such that $D = V = \infty$, but $\Delta = 0$. $\square$

3.4.4 Cumulative error bounds

Let $\{g_n\}$ be a sequence of learning rules. The *cumulative error* is defined as

$$\frac{1}{n} \sum_{i=1}^{n} I_{\{g_i(X_{i+1}, D_i) \neq I_{\{X_{i+1} \in C\}}\}} ,$$

that is, the relative frequency of errors committed by the sequence in the first n steps, if the first i labelled examples are always used to predict the label of the $i+1$-th example. Now the loss function is the cumulative error probability

$$l(\{g_i\}_{i=1}^n, g^*) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} L(g_i) .$$

Based on the results of the previous subsection, it is easy to obtain individual lower bounds for the expected value of the cumulative error.

It is a direct consequence of (3.12) that there exists a sequence of learning rules such that for all distribution-target pairs

$$\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} \le \frac{V\log(n+1)+1}{n} ,$$

which means that for the expected cumulative error $\{V\frac{\log n}{n}\}$ is a minimax upper rate of convergence for the class $\mathcal{D}$ (see Haussler et al. (1994)).

Haussler et al. (1994) considered minimax lower bounds for the expected cumulative error. The observation (3.14) is at the basis of the proof of their result:

Theorem 3.9. (Haussler et al. (1994)) *For the class $\mathcal{C}_k$ introduced in Section 3.4.3, for every n, and for every sequence of learning rules, there exists a distribution-target pair such that the expected cumulative error satisfies*

$$\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} \ge \frac{k}{2n}\left(\log\frac{n+1}{k}-1\right)-\frac{1}{4n} ,$$

that is, for the expected cumulative error the sequence $\{k\frac{\log n}{n}\}$ is a minimax lower rate of convergence for the class $\mathcal{D}$. □

We have the following individual extension of the above minimax lower bound:

Theorem 3.10. (Antos and Lugosi (1998)) *Let μ be the uniform distribution on $[0,1]\times\{1,2,\dots,k\}$. For every sequence of learning rules $\{g_n\}$, there exist a $C\in\mathcal{C}_k$ such that for all $0<\epsilon<1$,*

$$\mathbf{E}\left\{\frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}}\right\} > (1-\epsilon)\frac{k}{2n}\log n \quad \textit{for infinitely many } n,$$

that is, for the expected cumulative error $l_{\mathrm{ind}}(\mathcal{D},\{\frac{k}{2n}\log n\})\ge 1$ and $\{k\frac{\log n}{n}\}$ is an individual lower rate for $\mathcal{D}$.

Proof. We apply Lemma 2.2 with

$$R_n(z) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}\{g_i(X_{i+1},D_i(z))\neq Y_{i+1}(z), z\in L_i\cup R_i\} .$$

(Recall the definition of $Y_i(z)$, $D_i(z)$, L_i and R_i from Section reflower.) Clearly,

$$\begin{aligned} & \mathbf{E}\left\{ \frac{1}{n}\sum_{i=1}^{n} I_{\{g_i(X_{i+1},D_i)\neq I_{\{X_{i+1}\in C\}}\}} \right\} \\ & = \frac{1}{n}\sum_{i=1}^{n} \mathbf{P}\{g_i(X_{i+1},D_i(z)) \neq Y_{i+1}(z)\} \geq R_n(z) \ . \end{aligned}$$

Then it follows from (3.16) that if Z is uniform on $[0,1]\times\{1,2,\ldots,k\}$, and independent of $X, X_1, X_2 \ldots$, then

$$\mathbf{E}R_n(Z) \geq \frac{k}{2n}\left(\log\frac{n+1}{k} - 1\right) - \frac{1}{4n} \ .$$

(For the details see Haussler et al. (1994, p.278).) On the other hand, (3.18) implies that for each z,

$$R_n(z) \leq \frac{1}{n}\sum_{i=1}^{n}\frac{2k}{i+1} \leq \frac{2k}{n}\log(n+1) \ ,$$

so condition (2.10) is satisfied, which completes the proof. □

3.4.5 Bounds for the tail probabilities

The purpose of this subsection is to give individual lower bounds of the following type: let $\mathcal{C}$ be a class of concepts, and let $\{\epsilon_n\}$ be a sequence of positive numbers. Then for any sequence of learning rules $\{g_n\}$, there exists a distribution-target pair (μ, C) with $C \in \mathcal{C}$ such that

$$\mathbf{P}\{L(g_n) \geq \epsilon_n\} \geq a_n \quad \text{for infinitely many } n.$$

Here we would like to have

$$a_n \approx (c_0 n\epsilon_n)^{c_1 k} e^{-c_2 n\epsilon_n}$$

for some constants c_0, c_1, c_2, where k is the "dimension" of $\mathcal{C}$ so that the result is indeed an extension of Theorem 3.5. Clearly, these bounds are much more informative than bounds for the expected value of $L(g_n)$. For some sequences of $\{\epsilon_n\}$'s (which we believe to be the most interesting ones) we will be able to prove such results if $\mathcal{C}$ is one of the geometric concept classes discussed in Section 3.4.3.

Clearly, the most interesting values of ϵ_n are constant multiples of $1/n$, since this is the range where the probability of error $L(g_n)$ of a good learning rule g_n is expected to be with high probability. Our result extends Theorem 3.5 to such values of ϵ_n:

Theorem 3.11. (ANTOS AND LUGOSI (1998)) *Let $\mathcal{C}_k$ be the class of unions of k initial segments as defined in (3.15), and let μ be the uniform distribution on $\mathcal{X} = [0,1] \times \{1,\dots,k\}$. Let $\epsilon_1, \epsilon_2, \dots$ be nonnegativ numbers such that $\{\gamma_n = n\epsilon_n\}$ does not tend to ∞ as $n \to \infty$. For any sequence $\{g_n\}$ there exists a $C \in \mathcal{C}_k$ such that for each $\delta \in (0,1)$,*

$$\mathbf{P}\{L(g_n) \geq \epsilon_n\} \geq (1-\delta)\frac{1}{2}\sum_{i=0}^{k-1}\frac{(c\gamma_n)^i}{i!}e^{-c\gamma_n} \quad \textit{for infinitely many } n, \tag{3.19}$$

where $c = \log 256 \approx 5.545$, so $l_{\mathrm{ind}}(\mathcal{D}, \{\frac{1}{2}(cn\epsilon_n/(k-1))^{k-1}e^{-cn\epsilon_n}\}) \geq 1$, that is, $\{(cn\epsilon_n/k)^{k-1}e^{-cn\epsilon_n}\}$ and so $\{1\}$ is an individual lower rate for $\mathcal{D}$. □

REMARK. At the price of more complicated arguments, the value of the constant c may be improved to something slightly larger than 2. □

REMARK. A possible choice is $\epsilon_n = \gamma/n$ for any $\gamma \geq 0$. □

REMARK. Note that since

$$\sum_{i=0}^{k-1}\frac{(c\gamma_n)^i}{i!}e^{-c\gamma_n} \geq \left(\frac{c\gamma_n}{k-1}\right)^{k-1} e^{-c\gamma_n} ,$$

apart from constants, the lower bound of Theorem 3.11 has the same form as that of Theorem 3.5. □

REMARK. By the same embedding argument as the one used in Corollary 3.1, Theorem 3.11 can be extended to the concept classes listed in Section 3.4.3. □

The intuitive idea behind the proof of Theorem 3.11 is that in each of the k segments, inside the interval between the rightmost data point labelled by 1 and the leftmost data point labelled by 0, no learning rule can do better than mere guessing. Thus, the sum of the lengths of these intervals determines the size of the minimal probability of error. In the proof we exploit the fact that the length of these intervals have approximately exponential distribution, and they are almost independent, therefore we may approximate the minimax tail distribution of $L(g_n)$ by the tail of an appropriate gamma distribution. The proof is quite technical, so parts of it are given in lemmas after the main line of the proof.

Proof of Theorem 3.11. We assume that $\gamma_n \equiv \gamma \geq 0$, the proof of the general case is identical. First we introduce some notation:

$$U_{nj}^- = \max\{u \in [0,1] : \langle u, j\rangle \in \{X_i : Y_i = 1\} \cup \langle 0, j\rangle\} \ ,$$

$$U_{nj}^+ = \min\{u \in [0,1] : \langle u, j\rangle \in \{X_i : Y_i = 0\} \cup \langle 1, j\rangle\} \ ,$$

$$A'_{nj} = (U_{nj}^-, U_{nj}^+) \ ,$$

$$A_{nj} = A'_{nj} \times \{j\} \ ,$$

$$A_n = \bigcup_{j=1}^{k} A_{nj} \ .$$

Step 1. We apply Lemma 2.2 for $R_n(z) = R_{n,\epsilon_n}(z)$, where for each $\epsilon > 0$,

$$R_{n,\epsilon}(z) \stackrel{\text{def}}{=} \mathbf{P}\{L(g_n) > \epsilon\} = \mathbf{P}\left\{\int_{\mathcal{X}} I_{\{g_n(x,D_n(z)) \neq Y(x,z)\}}\, d\mu(x) > \epsilon\right\} \ .$$

(Here $Y(x,z) = I_{\{x \in C_z\}}$, where C_z is the concept associated with the parameter $z \in [0,1]^k$.) Just like in the proof of Theorem 3.8, let $Z = (Z_1, \ldots, Z_k)$ be uniformly distributed on $[0,1]^k$, and independent of $X, X_1, \ldots, X_n$. Since $R_{n,\epsilon_n}(z)$ is always bounded above by 1, and since the desired lower bound of (3.19) does not tend to 0 as $n \to \infty$, it suffices to prove a suitable lower bound for $\mathbf{E}R_{n,\epsilon}(Z)$, as the stability property (2.10) is automatically satisfied, that is, $\{\sup_z R_n(z)/\mathbf{E}R_n(Z)\}$ (or at least one of its subsequences) is bounded.

We will show that for each n and ϵ,

$$\mathbf{E}R_{n,\epsilon}(Z) \geq \frac{1}{2}\sum_{i=0}^{k-1}\frac{(c\gamma)^i}{i!}e^{-c\gamma} - \frac{1}{2}\left[\left(\sum_{i=0}^{k-1}\frac{(c\gamma)^i}{i!}e^{-c\gamma}\right)ke^{-n/k} + ke^{-\frac{n}{k}\left(1-\frac{\sqrt{\epsilon}}{2}\right)}\right] \tag{3.20}$$

(where $\gamma = n\epsilon$), which proves the theorem, since the term inside the brackets converges to zero rapidly as $n \to \infty$.

Clearly, by the independence of Z and $X, X_1, \ldots, X_n$,

$$\begin{aligned} R_{n,\epsilon}(Z) &= \mathbf{P}\left\{\int_{\mathcal{X}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) > \epsilon \,\middle|\, Z\right\} \\ &\geq \mathbf{P}\left\{\int_{A_n} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) > \epsilon \,\middle|\, Z\right\} . \end{aligned}$$

Thus, we have

$$\begin{aligned} \mathbf{E}R_{n,\epsilon}(Z) &\geq \mathbf{P}\left\{\int_{A_n} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) > \epsilon\right\} \\ &= \mathbf{E}\left\{\mathbf{P}\left\{\int_{A_n} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) > \epsilon \,\middle|\, D_n(Z)\right\}\right\} \\ &= \mathbf{E}\left\{\mathbf{P}\left\{\sum_{j=1}^{k}\int_{A_{nj}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) > \epsilon \,\middle|\, D_n(Z)\right\}\right\} . \end{aligned}$$

Step 2. In this step we obtain a lower bound for $\mathbf{P}\{L(g_n) > \epsilon\}$ in terms of the spacings containing the Z_i's. Let $\xi_{nj} = U_{nj}^+ - U_{nj}^-$. For all n and $\epsilon > 0$,

$$\mathbf{E}R_{n,\epsilon}(Z) \geq \frac{1}{2}\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon\right\} .$$

PROOF OF STEP 2. Clearly,

$$\int_{A_{nj}} I_{\{g_n(x,D_n(Z))\neq Y(x,Z)\}}\,d\mu(x) = \frac{1}{k}\lambda(A_{nj}\cap(B_{nj}\triangle C)) = \frac{1}{k}\lambda(B_{nj}\triangle C_{nj}) ,$$

where λ is the one-dimensional Lebesgue measure, and

$$B_{nj} = \{x \in A_{nj} : g_n(x, D_n(Z)) = 1\} ,$$

and

$$C_{nj} = C \cap A_{nj} \ .$$

Then it follows by Lemma 3.2 below that

$$\int_{A_{nj}} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}} \, d\mu(x) \geq \frac{1}{k} \left| U_{nj}^- + \lambda(B_{nj}) - Z_j \right| \ ,$$

and therefore

$$\begin{aligned} & \mathbf{P} \left\{ \sum_{j=1}^{k} \int_{A_{nj}} I_{\{g_n(x,D_n(Z)) \neq Y(x,Z)\}} \, d\mu(x) > \epsilon \,\middle|\, D_n(Z) \right\} \\ \geq \ & \mathbf{P} \left\{ \frac{1}{k} \sum_{j=1}^{k} |U_{nj}^- + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z) \right\} \\ \geq \ & \frac{1}{2} I_{\left\{ \frac{1}{k} \sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon \right\}} \ , \end{aligned}$$

where the last inequality is proved in Lemma 3.3 below. Taking expected values of both sides, we obtain

$$\mathbf{E} R_{n,\epsilon}(Z) \geq \frac{1}{2} \mathbf{P} \left\{ \frac{1}{k} \sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon \right\}$$

as desired. □

Step 3. Obviously,

$$\mathbf{P} \left\{ \frac{1}{k} \sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon \right\} \geq \mathbf{P} \left\{ \frac{1}{k} \sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon, \forall N_j > 0 \right\} \ ,$$

where N_j denotes the number of X_i's falling on the j'th segment $[0,1] \times \{j\}$. If $N_j > 0$, given N_j, the conditional distribution of ξ_{nj} is the same as the distribution of the sum of two spacings defined by N_j+1 i.i.d. uniform random variables on $[0,1]$, that is, for all $\epsilon \in [0,1]$,

$$\mathbf{P}\{\xi_{nj} \geq \epsilon | N_1, \ldots, N_k\} = \mathbf{P}\{\xi_{nj} \geq \epsilon | N_j\} = (1-\epsilon)^{N_j}(1 + N_j \epsilon)$$

(see, e.g., Reiss (1989)). A crucial step of the proof is approximating the conditional distributions of the ξ_{nj}'s by appropriate exponential distributions. For $N_j > 0$, define $\lambda_j = N_j \log 4 - 2\log(1 + N_j/2)$, and define the random variables $\xi'_{n1}, \ldots, \xi'_{nk}$ such that given $N_1, \ldots, N_k$, they are conditionally independent, and the conditional distribution of ξ'_{nj} is exponential with parameter λ_j, that is,

$$\mathbf{P}\{\xi'_{nj} \geq \epsilon | N_1, \ldots, N_k\} = \mathbf{P}\{\xi'_{nj} \geq \epsilon | N_j\} = e^{-\lambda_j \epsilon} .$$

(If $N_j = 0$ for some j, then $\mathbf{P}\{\xi'_{nj} \geq \epsilon | N_1, \ldots, N_k\}$ is defined arbitrarily.) Then, by Lemma 3.4 below, for all ϵ,

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi_{nj} \geq 4\epsilon, \forall N_j > 0\right\} \geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} - ke^{-\frac{n}{k}\left(1-\frac{\sqrt{\epsilon}}{2}\right)} .$$

Step 4. To finish the proof of (3.20), it remains to show that

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} \geq \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma}\left(1 - ke^{-n/k}\right) .$$

To do this, we may proceed as follows:

$$\begin{aligned}
\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon, \forall N_j > 0\right\} &= \mathbf{E}\left\{I_{\{\forall N_j>0\}}\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi'_{nj} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\geq \mathbf{E}\left\{I_{\{\forall N_j>0\}}\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k} \xi''_{nj} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(where the } \xi''_{nj} \text{ are defined exactly as the} \\
&\quad \xi'_{nj} \text{ but with } \lambda_j \text{ replaced by } \lambda'_j = N_j \log 4) \\
&\geq \mathbf{E}\left\{I_{\{\forall N_j>0\}}\mathbf{P}\left\{\frac{\sum_{j=1}^{k} \lambda'_j \xi''_{nj}}{\sum_{i=1}^{k} \lambda'_j} \geq 4\epsilon \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(by Lemma 3.5 below)} \\
&= \mathbf{E}\left\{I_{\{\forall N_j>0\}}\mathbf{P}\left\{\Phi_k \geq 4\epsilon \sum_{i=1}^{k} \lambda'_j \,\middle|\, N_1, \ldots, N_k\right\}\right\} \\
&\quad \text{(where given } N_1, \ldots, N_k \text{, the random variable } \Phi_k \text{ has} \\
&\quad k\text{-th order gamma distribution with parameter 1,} \\
&\quad \text{since given } N_j \text{, each } \lambda'_j \xi''_{nj} \text{ has exponential} \\
&\quad \text{distribution with parameter 1.)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left\{ I_{\{\forall N_j>0\}} \mathbf{P}\left\{ \Phi_k \geq 4n\epsilon \log 4 | \, N_1, \ldots, N_k \right\} \right\} \\
&\quad (\text{since } \textstyle\sum_{i=1}^{k} \lambda'_j = n \log 4) \\
&= \mathbf{E}\left\{ I_{\{\forall N_j>0\}} \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \right\} \\
&= \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \mathbf{P}\{\forall N_j > 0\} \\
&\geq \sum_{i=0}^{k-1} \frac{(c\gamma)^i}{i!} e^{-c\gamma} \left(1 - k e^{-n/k}\right) ,
\end{aligned}
$$

since

$$
\mathbf{P}\{\forall N_j > 0\} \geq 1 - k\mathbf{P}\{N_1 = 0\} = 1 - k\left(1 - \frac{1}{k}\right)^n \geq 1 - ke^{-n/k} ,
$$

and the proof of (3.20) is finished, so the proof of the theorem is complete.
□

The following lemmas are used in the proof above. We use the notation introduced in the text.

Lemma 3.2. *Let*

$$
E_{nj} = (U_{nj}^-, U_{nj}^- + \lambda(B_{nj})) \times \{j\} .
$$

For all n, $j \in \{1, \ldots, k\}$, $z \in [0,1]^k$, and data points $X_1, \ldots, X_n$,

$$
\lambda(B_{nj} \triangle C_{nj}) \geq \lambda(E_{nj} \triangle C_{nj}) .
$$

Proof. Clearly, $\lambda(E_{nj}) = \lambda(B_{nj})$. Assume, on the contrary, that

$$
\lambda(E_{nj} \triangle C_{nj}) > \lambda(B_{nj} \triangle C_{nj}) .
$$

Then

$$
\text{either} \quad \lambda(E_{nj} \cap \overline{C_{nj}}) > \lambda(B_{nj} \cap \overline{C_{nj}}) \quad \text{or} \quad \lambda(\overline{E_{nj}} \cap C_{nj}) > \lambda(\overline{B_{nj}} \cap C_{nj}) ,
$$

where $\overline{A} = [0,1] \times \{j\} - A$ is the complement of a set $A \subset [0,1] \times \{j\}$. In the first case $C_{nj} \subseteq E_{nj}$, so we have

$$
\lambda(E_{nj}) = \lambda(E_{nj} \cap \overline{C_{nj}}) + \lambda(E_{nj} \cap C_{nj}) > \lambda(B_{nj} \cap \overline{C_{nj}}) + \lambda(C_{nj}) \geq \lambda(B_{nj}) ,
$$

a contradiction. In the second case $E_{nj} \subseteq C_{nj}$. Then similarly to the first case,

$$\lambda(\overline{E_{nj}}) = \lambda(\overline{E_{nj}} \cap \overline{C_{nj}}) + \lambda(\overline{E_{nj}} \cap C_{nj}) > \lambda(\overline{C_{nj}}) + \lambda(\overline{B_{nj}} \cap C_{nj}) \geq \lambda(\overline{B_{nj}}) \ ,$$

again a contradiction. □

Lemma 3.3.

$$\mathbf{P}\left\{ \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z) \right\} \geq \frac{1}{2} I_{\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon\}} \ .$$

Proof. Since given D_n, the Z_j's are independent and uniform on the sets A'_{nj},

$$\begin{aligned} &\mathbf{P}\left\{ \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - Z_j| > \epsilon \,\middle|\, D_n(Z) \right\} \\ &= \frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - z_j| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})} \end{aligned}$$

where λ is the one-dimensional, and λ_k is the k-dimensional Lebesgue measure. Define $M_{nj} = \frac{1}{2}(U_{nj}^{-} + U_{nj}^{+})$, and

$$T_n \stackrel{\text{def}}{=} \bigotimes_{j=1}^{k} (M_{nj}, U_{nj}^{+}) \ .$$

Then

$$\begin{aligned} &\frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |U_{nj}^{-} + \lambda(B_{nj}) - z_j| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})} \geq \\ \geq\ & \frac{\lambda_k(\{z \in \bigotimes_{j=1}^{k} A'_{nj} : \frac{1}{k}\sum_{j=1}^{k} |z_j - M_{nj}| > \epsilon\})}{\lambda_k(\bigotimes_{j=1}^{k} A'_{nj})} \\ =\ & \frac{2^k \lambda_k(\{z \in T_n : \frac{1}{k}\sum_{j=1}^{k} (z_j - M_{nj}) > \epsilon\})}{2^k \lambda_k(T_n)} \\ \geq\ & \frac{\lambda_k(\{z \in T_n : \frac{1}{k}\sum_{j=1}^{k} (z_j - M_{nj}) > \frac{1}{k}\sum_{j=1}^{k} \frac{\xi_{nj}}{4}\})}{\lambda_k(T_n)} \\ =\ & \frac{\lambda_k(T_n \cap \{z : \sum_{j=1}^{k} (z_j - (M_{nj} + \frac{\xi_{nj}}{4})) > 0\})}{\lambda_k(T_n)} \ , \end{aligned}$$

whenever $\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq 4\epsilon$. Observe that the last expression equals 1/2, since the numerator is the volume of the intersection of the rectangle T_n with a halfspace defined by a hyperplane containing the center of the rectangle. The proof is complete. □

Lemma 3.4. *Let the random variables ξ_{nj} and ξ'_{nj} be as defined in the proof of Theorem 3.11. Then for all ϵ,*

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon, \forall N_j > 0\right\} \geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \forall N_j > 0\right\} - ke^{-\frac{n}{k}\left(1-\frac{\sqrt{\epsilon}}{2}\right)} .$$

Proof. It is easy to see that for all $N_j > 0$,

$$\mathbf{P}\{\xi_{nj} \geq \epsilon | N_j\} \geq \mathbf{P}\{\xi'_{nj} \geq \epsilon | N_j\} ,$$

whenever $\epsilon \leq 1/2$, and we have equality for $\epsilon = 1/2$. Thus, if $\forall N_j > 0$,

$$\begin{aligned}
&\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_{nj} \geq \epsilon \,\middle|\, \max_{j\leq k}\xi_{nj} < \frac{1}{2}, N_1,\ldots,N_k\right\} \\
&\qquad \mathbf{P}\left\{\max_{j\leq k}\xi_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, \max_{j\leq k}\xi'_{nj} < \frac{1}{2}, N_1,\ldots,N_k\right\} \\
&\qquad \mathbf{P}\left\{\max_{j\leq k}\xi'_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&= \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \max_{j\leq k}\xi'_{nj} < \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&= \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} \\
&\qquad - \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon, \max_{j\leq k}\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\} \\
&\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \,\middle|\, N_1,\ldots,N_k\right\} - \mathbf{P}\left\{\max_{j\leq k}\xi'_{nj} \geq \frac{1}{2} \,\middle|\, N_1,\ldots,N_k\right\}
\end{aligned}$$

$$\geq \mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi'_{nj} \geq \epsilon \middle| N_1,\ldots,N_k\right\} - \sum_{j=1}^{k}\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}\middle| N_j\right\} .$$

Clearly, for each j with $N_j > 0$,

$$\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}\middle| N_j\right\} = e^{-\lambda_j/2} \leq e^{-N_j \log 2 + N_j/2} .$$

Using the fact that N_j is a binomial random variable with parameters n and $1/k$, we get, by straightforward calculation, that

$$\begin{aligned}
\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}, N_j > 0\right\} &= \mathbf{E}\left\{I_{\{N_j>0\}}\mathbf{P}\left\{\xi'_{nj} \geq \frac{1}{2}\middle| N_j\right\}\right\} \\
&\leq \mathbf{E}\left\{e^{-N_j(\log 2 - 1/2)}\right\} \\
&= \left[1 - \frac{1}{k}\left(1 - \frac{\sqrt{e}}{2}\right)\right]^n \\
&\leq e^{-\frac{n}{k}\left(1-\frac{\sqrt{e}}{2}\right)} .
\end{aligned}$$

Taking expected values, we get the desired inequality. □

Lemma 3.5. *Let $\xi_1, \ldots, \xi_k$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_k > 0$, respectively. Then for each ϵ,*

$$\mathbf{P}\left\{\frac{1}{k}\sum_{j=1}^{k}\xi_j > \epsilon\right\} \geq \mathbf{P}\left\{\frac{\sum_{j=1}^{k}\lambda_j\xi_j}{\sum_{j=1}^{k}\lambda_j} > \epsilon\right\} = \mathbf{P}\{\Phi_k > k\epsilon\lambda\} ,$$

where $\lambda = \frac{1}{k}\sum_{j=1}^{k}\lambda_j$, and the random variable Φ_k has k-th order gamma distribution with parameter 1.

PROOF. We prove the lemma by induction for k. We will use two simple facts:

FACT 1. Let η, ξ, ξ' be real-valued random variables such that η is independent of (ξ, ξ') and $\mathbf{P}\{\xi > x\} \geq \mathbf{P}\{\xi' > x\}$ for all $x \in \mathcal{R}$. Then $\mathbf{P}\{\xi + \eta > x\} \geq \mathbf{P}\{\xi' + \eta > x\}$ for all x.

FACT 2. Let ξ_1 and ξ_2 be independent, exponential random variables with parameters λ_1 and λ_2, respectively. Assume $0 < \lambda_1 \leq \lambda_2$, and let $\delta =$

$(\lambda_2 - \lambda_1)/2$. Then for all $\epsilon > 0$, the probability $\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\}$ is monotone increasing in δ (while holding $\lambda_1 + \lambda_2$ fixed). In particular, $\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} \geq \mathbf{P}\{\Phi_2 > \lambda'\epsilon\}$ for $\lambda' = (\lambda_2 + \lambda_1)/2$.

Proof of Fact 2. Straightforward calculation shows that for $\delta > 0$,

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} = \frac{\lambda_2 e^{-\lambda_1 \epsilon} - \lambda_1 e^{-\lambda_2 \epsilon}}{\lambda_2 - \lambda_1} = e^{-\lambda' \epsilon}\left(\lambda'\epsilon \frac{\sinh(\delta\epsilon)}{\delta\epsilon} + \cosh(\delta\epsilon)\right) ,$$

and for $\delta = 0$,

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} = \mathbf{P}\{\Phi_2 > \lambda'\epsilon\} = (1 + \lambda'\epsilon)e^{-\lambda'\epsilon}.$$

Since $\sinh(x)/x$ and $\cosh(x)$ are monotone increasing on $[0, \infty)$, Fact 2 follows.

Now we are ready to prove the lemma. The statement is trivially true for $k = 1$, and by Fact 2 for $k = 2$. Let $k \geq 3$, and assume that the statement is true for $k - 1$. There exist two indices $j, j' \leq k$ such that $\lambda_j \leq \lambda$ and $\lambda_{j'} \geq \lambda$. Without loss of generality, we assume that $\lambda_1 \leq \lambda$ and $\lambda_2 \geq \lambda$. Let ξ_1' and ξ_2' be independent exponential random variables with parameter λ and $\lambda_1 + \lambda_2 - \lambda$, also independent of all ξ_j. Since

$$\frac{|\lambda_1 - \lambda_2|}{2} \geq \left|\lambda - \frac{\lambda_1 + \lambda_2}{2}\right| ,$$

Fact 2 implies that

$$\mathbf{P}\{\xi_1 + \xi_2 > \epsilon\} \geq \mathbf{P}\{\xi_1' + \xi_2' > \epsilon\} .$$

Also, by the inductive assumption,

$$\mathbf{P}\left\{\xi_2' + \sum_{j=3}^{k} \xi_j > \epsilon\right\} \geq \mathbf{P}\{\Phi_{k-1}/\lambda > \epsilon\} .$$

Using these and Fact 1 twice, we obtain

$$\begin{aligned} \mathbf{P}\left\{\sum_{j=1}^{k} \xi_j > \epsilon\right\} &\geq \mathbf{P}\left\{\xi_1' + \xi_2' + \sum_{j=3}^{k} \xi_j > \epsilon\right\} \\ &\geq \mathbf{P}\{\xi_1' + \Phi_{k-1}/\lambda > \epsilon\} \\ &= \mathbf{P}\{\Phi_k > \lambda\epsilon\} . \end{aligned}$$

□

Chapter 4

Functional estimation

4.1 Functional estimation problems

We are given n i.i.d. observations $D_n = \{X_1, \ldots, X_n\}$, drawn from the common unknown distribution of X on the discrete set $\mathcal{X}$. Let F be a functional assigning a real (possibly infinite) value to each possible distribution of X. In many cases it is of great importance to be able to estimate F accurately if it is finite, that is, find an estimate $F_n = F_n(X_1, \ldots, X_n)$ of F such that $|F_n - F|$ is small. Hence the loss function is the obvious distance

$$l(F_n, F) \stackrel{\text{def}}{=} |F_n - F| \ .$$

Important special cases are, when F is the expectation, the entropy, the mutual information and the Bayes-error in pattern recognition.

Expectation. In the first case X has a distribution $p(i)$ on $\mathcal{X} \subset \mathcal{R}$, and the expectation of X is

$$F = m \stackrel{\text{def}}{=} \mathbf{E}X = \sum_{i \in \mathcal{X}} ip(i) \ .$$

Also one would like to estimate the moments (and so the variance) of X,

$$m^{(k)} \stackrel{\text{def}}{=} \mathbf{E}\{X^k\} = \sum_{i \in \mathcal{X}} i^k p(i) \ ,$$

which leads again to the estimation of the expectation.

Entropy. In the second case assume that X has a distribution $p(i)$ on the set of the natural numbers $\mathcal{N}$, and consider the entropy of X, given by

$$F = H \stackrel{\text{def}}{=} -\sum_{i=1}^{\infty} p(i) \log_2 p(i) \ .$$

This quantity has crucial importance in information theory (see, e.g., Cover and Thomas (1991)).

Mutual information. In the third case we consider, we have a two-dimensional variable (V, W) with distribution $p(i,j)$ on $\mathcal{N}^2$ and marginal distributions $p_V(i)$ and $p_W(j)$. The mutual information of V and W is defined as

$$F = I \stackrel{\text{def}}{=} \sum_{i,j} p(i,j) \log_2 \frac{p(i,j)}{p_V(i) p_W(j)} .$$

Bayes error. In the fourth case, we consider a two-dimensional variable (X, Y) with distribution $\nu(i,j)$ on $\mathcal{N} \times \{0,1\}$. The distribution of X is $\mu(i)$ and the conditional distribution of Y given X is given by $\mathbf{P}\{Y = 1 | X = i\} = \eta(i)$. The Bayes-error

$$F = L^* \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} \mu(i) \min(\eta(i), 1 - \eta(i)) = \sum_{i=1}^{\infty} \min(\nu(i,1), \nu(i,0))$$

is the best possible probability of error we can achieve in classification, see Chapter 3. Thus it is important to estimate L^* accurately, even before pattern recognition is attempted. Also, a comparison of estimate of the error probability $L(g_n)$ of a decision rule and that of L^* gives us an idea how much room is left for improvement.

4.2 Consistency

First we prove a consistency result under general assumptions. If $p(i)$ determines the distribution of X, let the functional F be given in the form:

$$F \stackrel{\text{def}}{=} g\left(\sum_i f(i, p(i))\right) ,$$

where $f : \mathcal{N} \times [0,1] \mapsto \mathcal{R}$ and $g : \mathcal{R} \mapsto \mathcal{R}$ are real-valued functions. The plug-in estimate of F is defined by

$$F_n \stackrel{\text{def}}{=} g\left(\sum_i f(i, p_n(i))\right) ,$$

where

$$p_n(i) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j=i\}}$$

denotes the relative frequency of the occurence of i in the sample of size n.

Theorem 4.1. *(*ANTOS (1999B)*) Assume that for each $i \in \mathcal{N}$, $f(i,\cdot) : [0,1] \mapsto \mathcal{R}$ is nonnegative, concave, continuous and for each integers n and $0 \le j < n$ it satisfies*

$$\left| f\left(i, \frac{j+1}{n}\right) - f\left(i, \frac{j}{n}\right) \right| \le \frac{K}{n^p}$$

for some positive constants p and K. Also suppose that $g : [0,\infty) \mapsto \mathcal{R}$ is monotone increasing, concave, and Lipschitz(q, K'), that is,

$$|g(x) - g(y)| \le K'|x-y|^q \ , \ x, y \in [0,\infty) \ .$$

If $pq > 1/2$, then the plug-in estimate of F is strongly universally consistent, that is,

$$\lim_{n\to\infty} F_n = F \quad a.s.$$

If $F < \infty$, the estimate is also consistent in L_2, that is,

$$\lim_{n\to\infty} \mathbf{E}\{(F_n - F)^2\} = 0 \ .$$

The proof of Theorem 4.1 uses the following lemmas:

Lemma 4.1. *(*MCDIARMID (1989)*) Let $X_1, \dots, X_n$ be independent random variables on $\mathcal{X}$, and assume that $f : \mathcal{X}^n \mapsto \mathcal{R}$ satisfies*

$$\sup_{x_1,\dots,x_n,x_i' \in \mathcal{X}} |f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \le c_i \, , \ 1 \le i \le n \ .$$

Then for all $\epsilon > 0$

$$\mathbf{P}\{f(X_1,\dots,X_n) - \mathbf{E}f(X_1,\dots,X_n) \ge \epsilon\} \le e^{-2\epsilon^2/\sum_{i=1}^n c_i^2} \ . \ \square$$

Lemma 4.2. *(*DEVROYE (1991)*) If the conditions of Lemma 4.1 hold, then*

$$\mathbf{Var}\{f(X_1,\dots,X_n)\} \le \frac{1}{4}\sum_{i=1}^{n} c_i^2 \ . \ \square$$

PROOF OF THEOREM 4.1. We prove the following properties of F_n:

(a) For every n

$$g(0) \le F_n \ .$$

(b) For every n

$$\mathbf{E}F_n \le F \ .$$

(c) For every $\epsilon > 0$

$$\mathbf{P}\left\{|F_n - \mathbf{E}F_n| > \epsilon\right\} \le 2e^{-K_1 n^{2pq-1}\epsilon^2} \ ,$$

where $K_1 = \frac{2}{K'^2(2K)^{2q}}$.

(d)

$$\mathbf{Var}\, F_n \le \frac{1}{2K_1 n^{2pq-1}}.$$

(a) is obvious, because g is monotone.

(b) follows directly from Jensen's inequality and the concavity of the functions $f(i,\cdot)$ and g.

Lemma 4.1 implies (c). To see this, note that by changing the value of one sample point X_i, there can be two values, j and k such that $p_n(j)$ increases by $1/n$ and $p_n(k)$ decreases by the same amount. Then by the properties of the functions $f(i,\cdot)$, the value of $\sum_i f(i, p_n(i))$ cannot change by more than

$$\frac{2K}{n^p} \ ,$$

hence, by the properties of the function g, the value of F_n cannot change by more than

$$\frac{K'(2K)^q}{n^{pq}} \ .$$

Lemma 4.2 implies (d), similarly.

Noting that, by the strong law of large numbers, for each i, $\lim_{n\to\infty} p_n(i) = p_i$ a.s., continuity and Fatou's lemma imply

$$\liminf_{n\to\infty} \sum_i f(i, p_n(i)) \ge \sum_i f(i, p(i)) \quad \text{a.s.}$$

Since g is increasing and continuous on $(0,\infty)$

$$\begin{aligned}\liminf_{n\to\infty} F_n &= \liminf_{n\to\infty} g\left(\sum_i f(i,p_n(i))\right) \ge g\left(\liminf_{n\to\infty}\sum_i f(i,p_n(i))\right)\\ &\ge g\left(\sum_i f(i,p(i))\right) = F \quad \text{a.s.} \end{aligned} \tag{4.1}$$

For $pq > 1/2$ it follows from (c) and the Borel-Cantelli lemma that

$$\limsup_{n\to\infty} F_n = \limsup_{n\to\infty} \mathbf{E}F_n \quad \text{a.s. },$$

which is at most F by (b). This and (4.1) together imply

$$\lim_{n\to\infty} F_n = F \quad \text{a.s.}$$

To get the L_2 consistency, again by Fatou's lemma, we obtain

$$F \ge \limsup_{n\to\infty} \mathbf{E}F_n \ge \liminf_{n\to\infty} \mathbf{E}F_n \ge \mathbf{E}\{\liminf_{n\to\infty} F_n\} = F \ .$$

Thus, $\lim_{n\to\infty} \mathbf{E}F_n = F$, and since

$$\mathbf{E}\{(F_n - F)^2\} = \mathbf{Var}\, F_n + (\mathbf{E}F_n - F)^2 \ , \tag{4.2}$$

using (d), we conclude that for $F < \infty$, $\lim_{n\to\infty} \mathbf{E}\{(F_n - F)^2\} = 0$. □

We show that the plug-in estimates of the four functionals mentioned above are consistent, that is, $l_{\text{ind}}(\mathcal{D}, \{1\}) = 0$, where $\mathcal{D}$ is the class of all allowed distributions.

Expectation. The plug-in estimate of the expectation m is just the obvious average estimate

$$m_n = \frac{\sum_{j=1}^n X_j}{n} = \sum_{i\in\mathcal{X}} i p_n(i) \ .$$

Theorem 4.2. (LAW OF LARGE NUMBERS) *The plug-in estimate of m is strongly universally consistent, that is,*

$$\lim_{n\to\infty} m_n = m \quad \textit{a.s.}$$

For $|m| < \infty$ it is also consistent in L_1, that is,

$$\lim_{n\to\infty} \mathbf{E}\{|m_n - m|\} = 0 \ . \quad \square$$

REMARK. Obviously an analogous statement holds for the moments of X.
□

Entropy. The plug-in estimate of the entropy H is

$$H_n = -\sum_{i=1}^{\infty} p_n(i) \log_2 p_n(i) \ .$$

Theorem 4.3. *(*ANTOS (1999B)*) The plug-in estimate of H is strongly universally consistent, that is,*

$$\lim_{n\to\infty} H_n = H \quad a.s.$$

For $H < \infty$ it is also consistent in L_2, that is,

$$\lim_{n\to\infty} \mathbf{E}\{(H_n - H)^2\} = 0 \ .$$

PROOF. We apply Theorem 4.1. Now

$$f(i, p(i)) = -p(i) \log_2 p(i) \quad \text{and} \quad g(x) = x \ .$$

The conditions are satisfied with $K' = 1$, $q = 1$, with any $1/2 < p < 1$ and $K = \frac{1}{e(1-p)\ln 2}$ (e.g., $p = 3/4$ and $K = 3$ are suitable). □

REMARK. Directly we could prove the following properties of H_n:

(a) For every n

$$0 \le H_n \le \log_2 n \ ,$$

because H_n is just the entropy of a distribution concentrating on at most n different points.

(b) For every n

$$\mathbf{E}H_n \le H \ .$$

(c) For every $\epsilon > 0$

$$\mathbf{P}\left\{|H_n - \mathbf{E}H_n| > \epsilon\right\} \leq 2e^{-n\epsilon^2/2\log_2^2 n} ,$$

because for each integers n and $0 \leq j < n$

$$\left|\frac{j+1}{n}\log_2\frac{j+1}{n} - \frac{j}{n}\log_2\frac{j}{n}\right| \leq \frac{\log_2 n}{n} .$$

(d)

$$\mathbf{Var}\, H_n \leq \frac{\log_2^2 n}{n} .$$

□

Mutual information. For the mutual information let

$$p_n(i,j) = \frac{1}{n}\sum_{k=1}^{n} I_{\{V_k=i, W_k=j\}}$$

denote the relative frequency of the occurence of (i,j) in the sample of size n, with marginal distributions $\{p_{V,n}(i)\}$ and $\{p_{W,n}(j)\}$. Using the identity $I = H(V) + H(W) - H(V,W)$, the results for entropy estimation show the consistency of the plug-in estimate

$$I_n = \sum_{i,j} p_n(i,j)\log_2\frac{p_n(i,j)}{p_{V,n}(i)p_{W,n}(j)} .$$

COROLLARY 4.1. *(*ANTOS (1999B)*) If $H(V,W)$ is finite then the plug-in estimate of I is strongly universally consistent and consistent in L_2, that is,*

$$\lim_{n\to\infty} I_n = I \quad a.s.$$

and

$$\lim_{n\to\infty} \mathbf{E}\{(I_n - I)^2\} = 0 . \quad \square$$

REMARK. Note that it is possible that $H(V,W) = \infty$ while $I = 0$. □

Bayes-error. In a first group of methods, the Bayes-error L^* is estimated by an estimate $\widehat{L}_n$ of the error probability $L(g_n)$ of some consistent classification rule g_n. As such, this problem has been attempted, for example, by Fukunaga and Kessel (1971), Chen and Fu (1973), Fukunaga and Hummels (1987) and Garnett and Yau (1977). Concerning the error estimation of specific classification rules, see Chapter 10 in McLachlan (1992). Clearly, if the estimate $\widehat{L}_n$ we use is strongly consistent in the sense that $\widehat{L}_n - L(g_n) \to 0$ with probability one as $n \to \infty$, and the rule is strongly consistent, then $\widehat{L}_n \to L^*$ with probability one. In other words, we have a strongly consistent estimate of the Bayes error probability.

For example, the plug-in estimate of L^* is

$$\widehat{L}_n = \sum_{i=1}^{\infty} \min(\nu_n(i,1), \nu_n(i,0)) \ ,$$

where

$$\nu_n(i,j) = \frac{1}{n} \sum_{k=1}^{n} I_{\{X_k=i, Y_k=j\}} \ , \ j = 0,1$$

denotes the relative frequency of the occurence of (i,j) in the sample of size n. This is just the resubstitution estimate of the error probability of the histogram rule, if the cells of the partition are the points $\{i\}$ (see Devroye et al. (1996)). The consistency of this estimate can be proved the same way as Theorem 4.1. But it is also a consequence of the consistency of the histogram rule and that of the resubstitution estimate.

Theorem 4.4. (Devroye et al. (1996)) *The plug-in estimate of L^* is strongly universally consistent and consistent in L_p, that is,*

$$\lim_{n\to\infty} \widehat{L}_n = L^* \quad a.s.$$

and

$$\lim_{n\to\infty} \mathbf{E}\{|\widehat{L}_n - L^*|^p\} = 0 \ . \ \square$$

Remark. The theorem also holds for any (not discrete) distribution, if the plug-in estimate is defined by $\widehat{L}_n = \sum_i \min(\nu_n(A_{ni},1), \nu_n(A_{ni},0))$, where

$\nu_n(A_{ni}, j) = \frac{1}{n}\sum_{k=1}^{n} I_{\{X_k \in A_{ni}, Y_k = j\}}$ $(j = 0, 1)$, and $\{A_{ni}\}$ is an appropriate sequence of partitions (see, e.g., Devroye et al. (1996)). □

4.3 Slow rate of convergence

In this section we show that without assuming some conditions on F, there is no method that guarantees a certain rate of convergence for all distributions with finite value of F, that is, the convergence of the error of any estimates can be arbitrary slow. Let F be a general functional and let $\mathcal{D}$ be a class of distributions. The next result establishes a general result for the convergence in probability under some technical conditions.

Theorem 4.5. *(*ANTOS (1999B)*) Let $f : \mathcal{R}^+ \mapsto \mathcal{R}^+$ be continuous, stricktly decreasing function with $f(1) \geq 1/2$ and $\sum_{k=1}^{\infty} f(k) \leq 1$. Assume that for any discrete weight vector $\{q_0, q_1, \ldots\}$ with $\sum_{i=0}^{\infty} q_i = 1$ and $0 < q_i \leq f(i)$ $(i \geq 1)$, there is a subclass of distributions $\{\mu_u : u \in \{0,1\}^{\mathcal{N}}\} \subseteq \mathcal{D}$ parametrized by a binary sequence $u = (u_1, u_2, \ldots)$ with the following properties: There are disjoint sets $B_0, B_1, B_2, \ldots$ on $\mathcal{X}$ with $\mu_u(B_i) = q_i$ independently of u. For $i \geq 1$, the restriction of μ_u to B_i is chosen from two possibilities according to the value of u_i. Let $F(u) = F(\mu_u)$. If u and u' are two bit vectors coinciding in all but the k-th bit ($k \geq 1$), then $|F(u) - F(u')| \geq f(k)$. Assume moreover that for all u, $F(u)$ is finite. Then for any sequence of estimates $\{F_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a distribution on $\mathcal{X}$ may be found in $\mathcal{D}$ to be the distribution of X, such that $F < \infty$ and for any $\epsilon > 0$*

$$\mathbf{P}\{|F_n - F| > b_n\} > \frac{1}{2} - \epsilon \quad \textit{infinitely often.}$$

REMARK. Applying for the sequence $\{\sqrt{b_n}\}$ instead of $\{b_n\}$, Theorem 4.5 implies that for any $\{F_n\}$ there is a distribution in $\mathcal{D}$ such that for any $K > 0$

$$\limsup_{n\to\infty} \mathbf{P}\{|F_n - F| > Kb_n\} \geq \frac{1}{2}\,. \quad \Box$$

Denote the distribution of D_n by $\mu^{(n)}$. The proof of Theorem 4.5 uses the following lemma:

Lemma 4.3. *(*ANTOS (1999B)*) Consider a class of distributions $\mu_{u,z}$ parametrized by a binary sequence $u = (u_1, u_2, \ldots)$ and z with the following properties: The value of $F(u) = F(\mu_{u,z})$ does not depend on z. Let Z be a random variable taking the possible values of z and let $A_{n,k} \subset \mathcal{X}^n$ such that, on one hand, $c_{n,k} = \mu_{u,z}^{(n)}(A_{n,k})$ does not depend on (u, z), and on the other hand, if u and u' are two bit vectors coinciding in all but the k-th bit ($k \geq 1$), then for every $x_1^n = (x_1, \ldots, x_n) \in A_{n,k}$*

$$\mathbf{E}\mu_{u,Z}^{(n)}(x_1^n) = \mathbf{E}\mu_{u',Z}^{(n)}(x_1^n) \ . \tag{4.3}$$

For such coinciding u and u' assume moreover that $|F(u) - F(u')| \geq f(k)$. Then for any sequence of estimates $\{F_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero,

$$\sup_{u,z} \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u,z)) - F(u)| > b_n\} \geq \frac{1}{2} \limsup_{n\to\infty} c_{n,k_n} \ ,$$

for any $\{k_n\}$ such that $k_n \geq 1$ and $f(k_n) > 2b_n$ for all sufficiantly large n.

PROOF. We use randomization such that (u, z) is replaced by the independent random variables (U, Z), where $U = (U_1, U_2, \ldots)$ is an i.i.d. Bernoulli($\frac{1}{2}$) sequence.

Let U_+^k and U_-^k denote the vector U, with the difference that U_+^k forces the k-th bit to be 1 and U_-^k forces the k-th bit to be 0.

Introduce the notation

$$R_n(u,z) = \mathbf{P}\{|F_n(D_n(u,z)) - F(u)| > b_n\} \ .$$

Now $R_n(U,Z) \leq 1$, thus by Fatou's lemma

$$\begin{aligned}\sup_{u,z} \limsup_{n\to\infty} R_n(u,z) &\geq \mathbf{E}\{\limsup_{n\to\infty} R_n(U,Z)\} \\ &\geq \limsup_{n\to\infty} \mathbf{E}\{R_n(U,Z)\}\end{aligned}$$

$$
\begin{aligned}
&= \limsup_{n\to\infty} \mathbf{E}\{\mathbf{P}\{|F_n(D_n(U,Z)) - F(U)| > b_n | U, Z\}\} \\
&= \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(U,Z)) - F(U)| > b_n\} \\
&\geq \limsup_{n\to\infty} \mathbf{E}\{\mathbf{P}\{|F_n(D_n(U,Z)) - F(U)| > b_n | D_n(U,Z)\} I_{\{D_n(U,Z)\in A_{n,k}\}}\} .
\end{aligned}
$$

Conditioning on $D_n(U,Z)$ in $A_{n,k}$, for $k = k_n \geq 1$,

$$
\begin{aligned}
&\mathbf{P}\{|F_n(D_n(U,Z)) - F(U)| > b_n | D_n(U,Z)\} \\
&\quad = \mathbf{P}\{|F_n(D_n(U,Z)) - F(U_-^k)| > b_n, U_k = 0 | D_n(U,Z)\} \\
&\qquad + \mathbf{P}\{|F_n(D_n(U,Z)) - F(U_+^k)| > b_n, U_k = 1 | D_n(U,Z)\} \\
&\quad = \mathbf{P}\{|F_n(D_n(U,Z)) - F(U_-^k)| > b_n | D_n(U,Z)\} \mathbf{P}\{U_k = 0 | D_n(U,Z)\} \\
&\qquad + \mathbf{P}\{|F_n(D_n(U,Z)) - F(U_+^k)| > b_n | D_n(U,Z)\} \mathbf{P}\{U_k = 1 | D_n(U,Z)\} ,
\end{aligned}
$$

because (4.3) implies that $\mathbf{P}\{U_k = 1 | U_-^k, D_n(U,Z)\} = \frac{1}{2}$ for every possible U_-^k and $D_n(U,Z) \in A_{n,k}$, and thus U_-^k and U_k (or U_+^k and U_k) are conditionally independent given $D_n(U,Z) \in A_{n,k}$. Moreover $\mathbf{P}\{U_k = 1 | D_n(U,Z)\} = \frac{1}{2}$ for $D_n(U,Z) \in A_{n,k}$, so

$$
\begin{aligned}
&\mathbf{P}\{|F_n(D_n(U,Z)) - F(U_-^k)| > b_n | D_n(U,Z)\} \mathbf{P}\{U_k = 0 | D_n(U,Z)\} \\
&\qquad + \mathbf{P}\{|F_n(D_n(U,Z)) - F(U_+^k)| > b_n | D_n(U,Z)\} \mathbf{P}\{U_k = 1 | D_n(U,Z)\} \\
&\quad = \frac{1}{2}\mathbf{P}\{|F_n(D_n(U,Z)) - F(U_-^k)| > b_n | D_n(U,Z)\} \\
&\qquad + \frac{1}{2}\mathbf{P}\{|F_n(D_n(U,Z)) - F(U_+^k)| > b_n | D_n(U,Z)\} \\
&\quad \geq \frac{1}{2}\mathbf{P}\{|F_n(D_n(U,Z)) - F(U_-^k)| + |F_n(D_n(U,Z)) - F(U_+^k)| > \\
&\qquad > 2b_n | D_n(U,Z)\} \\
&\quad \geq \frac{1}{2}\mathbf{P}\{|F(U_-^k) - F(U_+^k)| > 2b_n | D_n(U,Z)\} \\
&\quad \geq \frac{1}{2}\mathbf{P}\{f(k) > 2b_n | D_n(U,Z)\} \\
&\quad = \frac{I_{\{f(k)>2b_n\}}}{2} ,
\end{aligned}
$$

Taking expectation on $A_{n,k}$

$$
\mathbf{E}\{\mathbf{P}\{|F_n(D_n(U,Z)) - F(U)| > b_n | D_n(U,Z)\} I_{\{D_n(U,Z)\in A_{n,k}\}}\}
$$

$$\geq \frac{I_{\{f(k)>2b_n\}}}{2} c_{n,k} \ .$$

In summary,

$$\sup_{u,z} \limsup_{n\to\infty} R_n(u,z) \geq \limsup_{n\to\infty} \frac{I_{\{f(k_n)>2b_n\}}}{2} c_{n,k_n} \geq \frac{1}{2} \limsup_{n\to\infty} c_{n,k_n}$$

by the choice of k_n. □

PROOF OF THEOREM 4.5. Since $b_n \to 0$, we can assume that $b_i < 1/4$, $i = 1, 2, \ldots$. For a given $\{q_i\}$ apply Lemma 4.3 in the case when z is fixed to, for example, 0. Let $A_{n,k} = \{x_1^n : \nexists i \ \ x_i \in B_k\}$. Then

$$c_{n,k} = \mu_u^{(n)}(A_{n,k}) = (1 - q_k)^n$$

independently of u, and if u and u' differ only in the k-th bit and $x_1^n \in A_{n,k}$, then

$$\mu_u^{(n)}(x_1^n) = \mu_{u'}^{(n)}(x_1^n) \ .$$

Hence by Lemma 4.3 we get that for any sequence of estimates $\{F_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero,

$$\sup_{u,z} \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u,z)) - F(u)| > b_n\} \geq \frac{1}{2} \limsup_{n\to\infty} (1 - q_{k_n})^n \geq \frac{1}{2} \ ,$$

if $f(k_n) > 2b_n$ and $q_{k_n} \leq 1/n^2$ for all sufficiantly large n.

This is satisfied if, for example, $k_n = \lceil f^{-1}(2b_n) \rceil - 1 (\to \infty)$, $q_k = \min\left(\frac{1}{\max\{n^2 : k_n = k\}}, f(k)\right)$ and $q_0 = 1 - \sum_{k=1}^{\infty} q_k$.

So for all $\{b_n\}$, there is $\{q_k\}$ and u that $\limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u)) - F(u)| > b_n\} \geq 1/2$, which implies that for all $\{b_n\}$, there is $\{q_k\}$ and u that for any $\epsilon > 0$

$$\mathbf{P}\{|F_n(D_n(u)) - F(u)| > b_n\} > \frac{1}{2} - \epsilon \quad \text{infinitely often.} \ \square$$

COROLLARY 4.2. (ANTOS (1999B)) *Assume the conditions of Theorem 4.5. Then for any sequence of estimates $\{F_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a distribution on $\mathcal{X}$ may be found in $\mathcal{D}$ to be the distribution of X, such that $F < \infty$ and*

$$\mathbf{E}\{|F_n - F|\} \geq b_n \quad \textit{infinitely often.} \ \square$$

REMARK. This means that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of distributions. $\square$

REMARK. Applying for the sequence $\{\sqrt{b_n}\}$ instead of $\{b_n\}$, Corollary 4.2 implies

$$l_{\text{ind}}(\mathcal{D}, \{b_n\}) = \infty\ . \quad \square$$

REMARK. The phrase "infinitely often" cannot be dropped from Theorem 4.5 and Corollary 4.2 (that is, the limes superior cannot be replaced by limes inferior here). Indeed, there exist deterministic sequences $\{f_n\}$ with $|f_n - F| \leq c/\sqrt{n}$ infinitely often for some constant c for every F. Just consider the dyadic sequence

$$\{f'_n\} = \left\{\frac{1}{2^0}, \frac{1}{2^1}, \frac{2}{2^1}, \frac{3}{2^1}, \frac{4}{2^1}, \frac{1}{2^2}, \frac{2}{2^2}, \ldots, \frac{16}{2^2}, \frac{1}{2^3}, \frac{2}{2^3}, \ldots, \frac{64}{2^3}, \ldots\right\}$$

and let $f_{2n-1} = f'_n$ and $f_{2n} = -f'_n$. Now for every F, for every i large enough, there is an element of the sequence in the first $2(1 + 4 + \cdots + 4^i) < 4 \cdot 4^i$ elements, whose distance from F is at most 2^{-i}. Thus $c = 2$ is a suitable choice. With $F_n \equiv f_n$, we thus obtain a very good estimate along an (unknown) subsequence for every real functional. (This can be generalized from $\mathcal{R}$ to certain finite dimensional spaces, see also Birgé (1986).) $\square$

Now we can apply Lemma 4.3, Theorem 4.5 and Corollary 4.2 for the four particular functionals.

Expectation. As an application of Corollary 4.2, we obtain

Theorem 4.6. *(*ANTOS (1999B)*) For any sequence of estimates $\{m_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a discrete distribution of X on $\mathcal{R}$ may be found such that $|m| < \infty$ and*

$$\mathbf{E}\{|m_n - m|\} \geq b_n \quad \textit{infinitely often.}$$

PROOF. Take, for example, $f(x) = 2^{-x}$. Given $\{q_0, q_1, \ldots\}$ with $0 < q_i \leq 2^{-i}$ $(i \geq 1)$, we find a sequence $\{l_1, l_2, \ldots\}$ of positive integers to be specified later. Then we consider the sets $B_0 = \{0\}$ and $B_i = \{2^{-i}, 2^{-i} + l_i\}$, each with two points. Let $\mu_u(B_i) = q_i$ independently of u, and for a bit vector u the distribution μ_u of X on B_i $(i \geq 1)$ is described constructively as follows: if $u_i = 1$ then $X = 2^{-i} + l_i$, while if $u_i = 0$, then $X = 2^{-i}$. For this distribution, it is easy to verify that

$$m = m(u) = \sum_{i=1}^{\infty} q_i(2^{-i} + u_i l_i) = \sum_{i=1}^{\infty} q_i 2^{-i} + \sum_{i=1}^{\infty} u_i q_i l_i \leq K + \sum_{i=1}^{\infty} q_i l_i,$$

where $K = \sum_{i=1}^{\infty} q_i 2^{-i}$. If u and u' differ only in the k-th bit, then

$$m(u) - m(u') = \sum_{i=1}^{\infty} (u_i - u_i') q_i l_i = (u_k - u_k') q_k l_k,$$

and thus, $|m(u) - m(u')| = q_k l_k$. For $k \geq 1$ the inequality $2^{-k} \leq q_k l_k \leq 2 \cdot 2^{-k}$ is satisfied if, for example, $l_k = \lceil \frac{1}{q_k 2^k} \rceil$, and thus $m(u) \leq K + 2$. □

This means that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of all discrete distributions.

REMARK. Examine the parameter $l^{(p)} = \mathbf{E}\{|X| \log_2^p(1 + |X|)\}$ of the distributions in the class above. For a distribution corresponding to u, $l^{(p)}(u) = \sum_k q_k(2^{-k} + u_k l_k) \log_2^p(1 + 2^{-k} + u_k l_k)$, so

$$\sum_k q_k u_k l_k \log_2^p(1 + u_k l_k) \leq l^{(p)}(u) \leq K_1 \sum_k q_k u_k l_k \log_2^p(1 + u_k l_k) + K_2 \ .$$

By the choice of q_k, $l^{(p)}(u) < \infty$ if and only if $\sum_k q_k u_k l_k \log_2^p(1 + u_k l_k) < \infty$, thus $l^{(p)}$ is finite over the class if and only if $\sum_k q_k l_k \log_2^p(1 + l_k) < \infty$, that is, $\sum_k q_k l_k \log_2^p l_k < \infty$. For monotone decreasing b_n, a good choice is $q_k \sim \min\left(\frac{1}{\max\{n : b_n \geq 2^{-k}\}}, 2^{-k}\right)$ and

$$\sum_k q_k l_k \log_2^p l_k = \sum_k 2^{-k} \left(\log_2 \frac{1}{q_k} - \log_2 \frac{1}{q_k l_k}\right)^p \sim \sum_k 2^{-k} \left(\log_2 \frac{1}{q_k} - k\right)^p$$

is finite if $\sum_k 2^{-k} \log_2^p \frac{1}{q_k}$ is finite. For example, if $b_n = \log_2^{-r} n, r > 0$, then $q_k \sim \min(2^{-2^{k/r}}, 2^{-k}) \sim 2^{-2^{k/r}}$ and

$$\sum_k q_k l_k \log_2^p \frac{1}{q_k} \sim \sum_k 2^{-k(1-p/r)},$$

which is infinite for $r \le p$ and finite for $r > p$. This means that for $r > p$ the rate $\{\frac{1}{\log_2^r n}\}$ is an individual lower rate of convergence for the class of distributions with finite $l^{(p)}$, and suggests that the rate $\{\frac{1}{\log_2^p n}\}$ can be achived for this class. □

REMARK. Examine the moment parameter $m^{(p)} = \mathbf{E}\{|X|^p\}$ $(p > 1)$ of the distributions in the class above. For a distribution corresponding to u, $m^{(p)}(u) = \sum_k q_k(2^{-k} + u_k l_k)^p$, so

$$\sum_k q_k u_k l_k^p \le m^{(p)}(u) \le 2^{p-1}\left(\sum_k q_k 2^{-kp} + \sum_k q_k u_k l_k^p\right).$$

By the choice of q_k, $m^{(p)}(u) < \infty$ if and only if $\sum_k q_k u_k l_k^p < \infty$, thus $m^{(p)}$ is finite over the class if and only if $\sum_k q_k l_k^p < \infty$. For monotone decreasing b_n, a good choice is $q_k \sim \min\left(\frac{1}{\max\{n: b_n \ge 2^{-k}\}}, 2^{-k}\right)$ and $\sum_k q_k l_k^p \sim \sum_k \frac{1}{q_k^{p-1} 2^{kp}}$. For example, if $b_n = n^{-r}, r > 0$, then $q_k \sim \min(2^{-k/r}, 2^{-k})$ and $\sum_k q_k l_k^p \sim \sum_k 2^{-k(p-(p-1)\max(\frac{1}{r},1))}$, which is infinite for $r \le 1 - 1/p$ and finite for $r > 1 - 1/p$. This means that for $r > 1 - 1/p$ the rate $\{\frac{1}{n^r}\}$ is an individual lower rate of convergence for the class of distributions with finite $m^{(p)}$, and suggests that the rate $\{\frac{1}{n^{1-1/p}}\}$ can be achived for this class (see Theorem 4.11). □

Entropy. As an application of Corollary 4.2, we obtain

Theorem 4.7. *(*ANTOS (1999B)*) For any sequence of estimates $\{H_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a distribution of X on $\mathcal{N}$ may be found such that $H < \infty$ and*

$$\mathbf{E}\{|H_n - H|\} \ge b_n \quad \textit{infinitely often.}$$

PROOF. Take $f(x) = 2^{-x}$. Given $\{q_0, q_1, \ldots\}$ with $0 < q_i \le 2^{-i}$ $i \ge 1$, we find a sequence $\{l_0, l_1, \ldots\}$ of positive integers to be specified later. Let $l_i = \log_2 l'_i$. Then we partition the positive integers into consecutive blocks B_i of cardinality $l'_0 = 1$, l'_1, l'_2, l'_3, ... Let $\mu_u(B_i) = q_i$ independently of u, and for a bit vector u the distribution μ_u of X on B_i is described constructively as follows: if $u_i = 1$ then X is drawn uniformly over the l'_i integers in that block, while if $u_i = 0$, then X takes the first point in the block. For this distribution, it is easy to verify that

$$H = H(u) = \sum_{i=1}^{\infty} u_i q_i \log_2 l'_i - \sum_{i=0}^{\infty} q_i \log_2 q_i = \sum_{i=1}^{\infty} u_i q_i l_i - \sum_{i=0}^{\infty} q_i \log_2 q_i \ .$$

If u and u' differ only in the k-th bit, then

$$H(u) - H(u') = \sum_{i=1}^{\infty} (u_i - u'_i) q_i l_i = (u_k - u'_k) q_k l_k,$$

and thus, $|H(u) - H(u')| = q_k l_k$. For $k \geq 1$ the inequality $2^{-k} \leq q_k l_k \leq 2 \cdot 2^{-k}$ is satisfied if, for example, $l'_k = \lceil 2^{1/(q_k 2^k)} \rceil$, and thus $H(u) \leq 4 + \sum_{i=2}^{\infty} i 2^{-i}$.
$\square$

This means that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of all distributions on $\mathcal{N}$.

REMARK. Examine the moment parameter $H^{(p)} = \sum_i p(i) \log_2^p(1/p(i))$ of the distributions in the class above. For a distribution corresponding to u, $H^{(p)}(u) = \sum_k q_k(-\log_2 q_k + u_k l_k)^p$, so

$$\sum_k q_k u_k l_k^p \leq H^{(p)}(u) \leq 2^{p-1} \left(\sum_k q_k \log_2^p \frac{1}{q_k} + \sum_k q_k u_k l_k^p \right) \ .$$

By the choice of q_k, $H^{(p)}(u) < \infty$ if and only if $\sum_k q_k u_k l_k^p < \infty$, thus $H^{(p)}$ is finite over the class if and only if $\sum_k q_k l_k^p < \infty$. For monotone decreasing b_n, a good choice is $q_k \sim \min\left(\frac{1}{\max\{n : b_n \geq 2^{-k}\}}, 2^{-k}\right)$ and $\sum_k q_k l_k^p \sim \sum_k \frac{1}{q_k^{p-1} 2^{kp}}$. For example, if $b_n = n^{-r}, r > 0$, then $q_k \sim \min(2^{-k/r}, 2^{-k})$ and $\sum_k q_k l_k^p \sim \sum_k 2^{-k(p-(p-1)\max(1,\frac{1}{r}))}$, which is infinite for $r \leq 1 - 1/p$ and finite for $r > 1 - 1/p$. This means that for $r > 1 - 1/p$ the rate $\{\frac{1}{n^r}\}$ is an individual lower rate of convergence for the class of distributions with finite $H^{(p)}$, and suggests that the rate $\{\frac{1}{n^{1-1/p}}\}$ can be achived for this class. $\square$

Mutual information. As an application of Corollary 4.2, we obtain

Theorem 4.8. (ANTOS (1999B)) *For any sequence of estimates* $\{I_n\}$ *and for any sequence* $\{b_n\}$ *of positive numbers converging to zero, a distribution of* (V, W) *on* $\mathcal{N} \times \mathcal{N}$ *may be found such that* $I < \infty$ *and*

$$\mathbf{E}\{|I_n - I|\} \geq b_n \quad \textit{infinitely often.}$$

PROOF. Take $f(x) = 2^{-x}$. Given $\{q_0, q_1, \ldots\}$ with $0 < q_i \leq 2^{-i}$ $(i \geq 1)$, we find a sequence $\{l_0, l_1, \ldots\}$ of positive integers to be specified later. Let $l_i = \log_2 l'_i$. Then we partition the positive integers into consecutive blocks B'_i of cardinality $l'_0 = 1$, l'_1, l'_2, l'_3, ... Let $B_i = B'_i \times B'_i$. Let $\mu_u(B_i) = q_i$ independently of u, and for a bit vector u the distribution μ_u of (V, W) on B_i is described constructively as follows: if $u_i = 1$ then V is drawn uniformly over the l'_i integers in that block and $W = V$, while if $u_i = 0$, then V and W are drawn uniformly and independenly over the l'_i integers in that block. For this distribution, it is easy to verify that

$$I = I(u) = \sum_{i=1}^{\infty} u_i q_i \log_2 l'_i - \sum_{i=0}^{\infty} q_i \log_2 q_i = \sum_{i=1}^{\infty} u_i q_i l_i - \sum_{i=0}^{\infty} q_i \log_2 q_i.$$

If u and u' differ only in the k-th bit, then

$$I(u) - I(u') = \sum_{i=1}^{\infty} (u_i - u'_i) q_i \log_2 l_i = (u_k - u'_k) q_k l_k,$$

and thus, $|I(u) - I(u')| = q_k l_k$. For $k \geq 1$ the inequality $2^{-k} \leq q_k l_k \leq 2 \cdot 2^{-k}$ is satisfied if, for example, $l'_k = \lceil 2^{1/(q_k 2^k)} \rceil$ and thus $I(u) \leq 4 + \sum_{i=2}^{\infty} i 2^{-i}$. $\square$

This means that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of the symmetric distributions of (V, W) on $\mathcal{N}^2$ (that is, $p(i, j) = p(j, i)$).

Bayes-error. If $\widehat{L}_n$ is an estimate of $L(g_n)$, though for many classifiers, $\widehat{L}_n - L(g_n)$ can be guaranteed to converge to zero rapidly, regardless what the distribution of (X, Y) is (see Chapters 8, 23, 24 and 31 of Devroye et al. (1996)), in view of Theorem 3.1, the rate of convergence of $L(g_n)$ to L^* using such a method may be arbitrarily slow. Thus, we cannot expect a good performance for all distributions from such a method.

The question thus is whether it is possible to come up with another method of estimating L^* such that the difference $\widehat{L}_n - L^*$ converges to zero rapidly for all distributions.

A weaker kind of negative result is known for this general case:

Theorem 4.9. *(*DEVROYE ET AL. (1996)*) For every n, for any estimate $\widehat{L}_n$ of the Bayes error probability L^*, and for every $\epsilon > 0$, there exists a distribution of (X, Y), such that*

$$\mathbf{E}\left\{|\widehat{L}_n - L^*|\right\} \geq 1/4 - \epsilon . \quad \square$$

Using Lemma 4.3, we can show the following slow-rate result:

Theorem 4.10. *For any sequence of estimates $\{\widehat{L}_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a distribution of (X, Y) on $\mathcal{N} \times \{0, 1\}$ may be found such that for any $\epsilon > 0$*

$$\mathbf{P}\{|\widehat{L}_n - L^*| > b_n\} > \frac{1}{2} - \epsilon \quad \textit{infinitely often.}$$

PROOF. Given $\{b_n\}$, we find a sequence $\{l_1, l_2, \ldots\}$ of positive integers to be specified later. Then we partition the positive integers into consecutive blocks of cardinality $l_1, l_2, l_3, \ldots$. Let $z = (z_1, z_2, \ldots)$ be a vector assigning a bit to each integer, and let $u = (u_1, u_2, \ldots)$ be a vector assigning a bit to each block. Then the distribution $\nu_{u,z}$ of (X, Y) is described constructively as follows: first a block B is drawn from the geometric distribution:

$$\mathbf{P}\{B = i\} = q_i = \frac{1}{2^i} , \ i \geq 1 .$$

Then X is drawn uniformly over the l_B integers in that block. If $u_B = 0$, then $Y = z_X$, while if $u_B = 1$, Y is Bernoulli(1/2), independent of X. For this distribution, it is easy to verify that

$$L^* = L^*(u) = \sum_{i=1}^{\infty} \frac{u_i}{2^{i+1}} .$$

Observe that L^* depends upon u only.

Apply Lemma 4.3 by replacing z by the i.i.d. Bernoulli(1/2) sequence $Z = (Z_1, Z_2, \ldots)$. Let $A_{n,k} = \{(x_1, y_1), \ldots, (x_n, y_n) : \forall i, j; x_i, x_j \in B_k \Rightarrow x_i \neq x_j\}$. Then

$$c_{n,k} = \nu_{u,z}^{(n)}(A_{n,k}) \geq 1 - \frac{n^2 q_k^2}{2l_k}$$

independently of (u, z), and if u and u' differ only in the k-th bit and $(x_1, y_1), \ldots, (x_n, y_n) \in A_{n,k}$, then

$$\mathbf{E}\nu_{u,Z}^{(n)}((x_1, y_1), \ldots, (x_n, y_n)) = \mathbf{E}\nu_{u',Z}^{(n)}((x_1, y_1), \ldots, (x_n, y_n)) \ .$$

Moreover, for such coinciding u and u', $|L^*(u) - L^*(u')| = 2^{-k-1}$. Hence by Lemma 4.3 we get that for any sequence of estimates $\{\widehat{L}_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero,

$$\begin{aligned} &\sup_{u,z} \limsup_{n\to\infty} \mathbf{P}\{|\widehat{L}_n(D_n(u, z)) - L^*(u)| > b_n\} \\ &\geq \frac{1}{2} \limsup_{n\to\infty} \left(1 - \frac{n^2 q_{k_n}^2}{2l_{k_n}}\right) \geq \frac{1}{2} - \frac{1}{2} \liminf_{n\to\infty} \frac{1}{2^{2k_n+1}} = \frac{1}{2} \ , \end{aligned}$$

if $2^{-k_n-1} > 2b_n$, $l_{k_n} \geq n^2$ and $\lim_{n\to\infty} k_n = \infty$. This is satisfied if, for example, $k_n = \lceil \log_2(1/4b_n) \rceil - 1 (\to \infty)$ and $l_i = \max^2\{n : k_n \leq i\}$.

So for all $\{b_n\}$, there is $\{l_k\}$,u and z that $\limsup_{n\to\infty} \mathbf{P}\{|\widehat{L}_n(D_n(u, z)) - L^*(u)| > b_n\} \geq 1/2$, which implies that for all $\{b_n\}$, there is $\{l_k\}$,u and z that for any $\epsilon > 0$

$$\mathbf{P}\{|\widehat{L}_n(D_n(u, z)) - L^*(u)| > b_n\} > \frac{1}{2} - \epsilon \quad \text{infinitely often.} \ \Box$$

Corollary 4.3. *(*Antos, Devroye and Györfi (1999)*) For any sequence of estimates $\{\widehat{L}_n\}$ and for any sequence $\{b_n\}$ of positive numbers converging to zero, a distribution of (X, Y) on $\mathcal{N} \times \{0, 1\}$ may be found such that*

$$\mathbf{E}\{|\widehat{L}_n - L^*|\} \geq b_n \quad \textit{infinitely often.} \ \Box$$

This means that every sequence tending to zero is an individual lower rate of convergence for the class $\mathcal{D}$ of all distributions on $\mathcal{N} \times \{0, 1\}$.

Remark. In pattern recognition, another important quantity is the asymptotic error probability of the nearest neighbor rule, $L_{\mathrm{NN}} = 2\sum_{i=1}^{\infty}\mu(i)\eta(i)(1-\eta(i))$ (see Devroye et al. (1996), Chapter 5). We can also estimate L_{NN} consistently by a plug-in estimate, the proof is similar to that of Therorem 4.1. Observing that for the subclass of distributions used in the proof of Theorem 4.10, $L_{\mathrm{NN}} = L^*$, we get that the slow-rate result also holds for the estimation of L_{NN}. The same is true for the asymptotic error probability of the k-nearest neighbor rule, $L_{k\mathrm{NN}} = \sum_{i=1}^{\infty}\mu(i)\alpha_k(\eta(i))$, where α_k is a suitable polinom (see Devroye et al. (1996), Chapter 5). □

Remark. Noting that if $\eta \equiv p$ (on the support of X) then $L^* = \min(p, 1-p)$ (even, $L^* = 1/2$ if and only if $\eta \equiv 1/2$), one can interpret deciding whether $\eta \equiv p$ or not, as a first small step toward estimating L^*. Interestingly, in spite of the slow-rate result above, for given p, testing whether $\eta \equiv p$ (and so whether $L^* = 1/2$) or not can be performed at an exponential rate, that is, there is a sequence of tests $\{T_n : (\mathcal{N} \times \{0,1\})^n \mapsto \{0,1\}\}$, such that $\mathbf{P}\{T_n(D_n) \neq I_{\{\eta\equiv p\}}\} \leq a \cdot b^n$, $b < 1$ for all distributions of (X, Y). □

4.4 Rate of convergence

According to the results of the previous section, in order to get rate-of-convergence results, we have to assume conditions to the distribution. There are such results for the expectation, and we show a rate-of-convergence result for the plug-in estimation of entropy and mutual information.

Expectation. As it was predicted in the second Remark after Theorem 4.6, assuming $m^{(p)} = \mathbf{E}\{|X|^p\} < \infty$, the theorem below implies the upper bound of

$$O\left(\frac{1}{n^{1-1/p}}\right)$$

for the L_1 error of the plug-in estimates of the expectation.

Theorem 4.11. *(*BAHR AND ESSEEN (1965), SEE THEOREM 2.6.20 IN PETROV (1995)*) For $1 \le p \le 2$, if $\mathbf{E}\{|X|^p\} < \infty$ then for*

$$m_n = \frac{\sum_{i=1}^n X_i}{n} \ ,$$

$$\mathbf{E}\{|m_n - m|\} \le (\mathbf{E}\{|m_n - m|^p\})^{1/p} \le \frac{(2\mathbf{E}\{|X|^p\})^{1/p}}{n^{1-1/p}} \ . \quad \square$$

This means that for $1 \le p \le 2$ for any sequence $\{b_n\}$ tending to zero, $l_{\text{ind}}(\mathcal{D}^p, \{\frac{1}{b_n n^{1-1/p}}\}) = 0$, that is, $\{\frac{1}{b_n n^{1-1/p}}\}$ is an individual upper rate of convergence for the class $\mathcal{D}^p$ of distributions with finite $m^{(p)}$.

Entropy. Assuming appropriate tail condition, the theorem below implies the upper bound of

$$O\left(\frac{\log n}{n^{\alpha/(1+\alpha)}}\right)$$

for the L_2 error of the plug-in estimates of the entropy.

Theorem 4.12. *(*ANTOS (1999B)*) Assume that for some $0 < \alpha \le 1$ there exist positive constants c_1, $c_2 > 0$ and a suitable indexing of the weight vector $\{p_i\}$ such that $c_1/i^{1+\alpha} \le p_i \le c_2/i^{1+\alpha}$. Then for the L_2 error of the plug-in estimate*

$$\mathbf{E}\{(H_n - H)^2\} = O\left(\frac{\log^2 n}{n^{2\alpha/(1+\alpha)}}\right) .$$

REMARK. If $\alpha > 1$ the proof gives $O(\log^2 n/n)$. $\square$

PROOF. By (4.2) and (b) and (d) following the proof of Theorem 4.3, it is enough to prove that

$$H - \mathbf{E}H_n = O\left(\frac{\log n}{n^{\alpha/(1+\alpha)}}\right) \ .$$

First we prove that

$$H - \mathbf{E}H_n \le \sum_{i=1}^{\infty} p_i \log_2 \left(1 + \frac{1-p_i}{np_i}\right) \ .$$

Observe that for any i,

$$
\begin{aligned}
&\mathbf{E}\left[-p_{n,i}\log_2 p_{n,i}\right] \\
&= -\sum_{k=0}^{n}\frac{k}{n}\log_2\frac{k}{n}\binom{n}{k}p_i^k(1-p_i)^{n-k} \\
&= -p_i\sum_{k=1}^{n}\log_2\frac{k}{n}\binom{n-1}{k-1}p_i^{k-1}(1-p_i)^{n-1-(k-1)} \\
&= -p_i\sum_{k=0}^{n-1}\log_2\frac{k+1}{n}\binom{n-1}{k}p_i^{k}(1-p_i)^{n-1-k} \\
&= -p_i\mathbf{E}\left\{\log_2\left(\frac{(n-1)p_{n-1,i}+1}{n}\right)\right\} \\
&\geq -p_i\log_2\left(\frac{(n-1)p_i+1}{n}\right),
\end{aligned}
$$

where in the last step we applied Jensen's inequality for the concave function $\log_2 x$. Summing for i

$$
\mathbf{E}H_n \geq -\sum_{i=1}^{\infty}p_i\log_2\left(\frac{(n-1)p_i+1}{n}\right),
$$

and so

$$
H-\mathbf{E}H_n \leq \sum_{i=1}^{\infty}p_i\log_2\left(1+\frac{1-p_i}{np_i}\right) \leq \sum_{i=1}^{\infty}p_i\log_2\left(1+\frac{1}{np_i}\right).
$$

Splitting this sum into two terms

$$
\begin{aligned}
\sum_{i=1}^{\infty}p_i\log_2\left(1+\frac{1}{np_i}\right) &\leq \sum_{i:np_i\leq 1}^{\infty}p_i\log_2\frac{2}{np_i}+\sum_{i:np_i>1}^{\infty}p_i\frac{1}{np_i} \\
&\leq \sum_{i:np_i\leq 1}^{\infty}p_i\log_2\frac{1}{p_i}+\sum_{i:np_i>1}^{\infty}\frac{1}{n}
\end{aligned}
$$

for $n \geq 2$. Now taking the bounds on p_i, $p_i > 1/n$ implies $i < (c_2 n)^{\frac{1}{1+\alpha}}$ and $p_i \leq 1/n$ implies $i \geq (c_1 n)^{\frac{1}{1+\alpha}}$. So for the second term

$$
\sum_{i:np_i>1}^{\infty}\frac{1}{n} \leq \frac{\lfloor (c_2 n)^{\frac{1}{1+\alpha}}\rfloor}{n} \leq c_2^{\frac{1}{1+\alpha}}n^{-\frac{\alpha}{1+\alpha}}.
$$

For n large enough, the first term

$$
\sum_{i:np_i\leq 1}^{\infty}p_i\log_2\frac{1}{p_i} \leq \sum_{i\geq (c_1 n)^{1/(1+\alpha)}}^{\infty}\frac{K_1}{i^{1+\alpha}}\log_2 i
$$

$$
\begin{aligned}
&\leq \frac{K_1}{\log 2} \int\limits_{(c_1 n)^{1/(1+\alpha)}-1}^{\infty} \frac{\log x}{x^{1+\alpha}}\, dx \\
&= \frac{K_1}{\log 2} \left[\frac{1}{\alpha} \frac{\log x}{x^\alpha} + \frac{1}{\alpha^2} \frac{1}{x^\alpha} \right]_{\infty}^{(c_1 n)^{1/(1+\alpha)}-1} \\
&\leq K_2 \frac{\log n}{n^{\alpha/(1+\alpha)}}
\end{aligned}
$$

concludes the theorem. □

Mutual information. Again, using the identity $I = H(V) + H(W) - H(V,W)$, the results for entropy estimation show that assuming appropriate tail condition, the theorem below implies the upper bound of

$$O\left(\frac{\log n}{n^{\alpha/(1+\alpha)}}\right)$$

for the L_2 error of the plug-in estimates of the mutual information.

COROLLARY 4.4. *(*ANTOS (1999B)*) Assume that the tail condition of Theorem 4.12 holds for the distributions of V, W and also (V,W). Then for the L_2 error of the plug-in estimate*

$$\mathbf{E}\{(I_n - I)^2\} = O\left(\frac{\log^2 n}{n^{2\alpha/(1+\alpha)}}\right). \quad \square$$

REMARK. If $\alpha > 1$ the proof gives $O(\log^2 n/n)$. □

Bibliography

Anthony, M., Biggs, N. L., and Shawe-Taylor, J. (1993). Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 42:65–73.

Antos, A. (1995). Függvényosztályok tulajdonságai és szerepe az alakfelismerésben (Properties of classes of functions and their roles in pattern recognition). Master's Thesis, Technical University of Budapest, H-1521 Stoczek u. 2, Budapest, Hungary.

Antos, A. (1999a). Lower bounds on the rate of convergence of nonparametric pattern recognition. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99, Proceedings*, Fischer, P. and Simon, H., editors, volume 1572 of *LNAI/LNCS*, pages 241–252. DFG,IFIP, Springer, Berlin. (IFIP WG 1.4: Student Author Award).

Antos, A. (1999b). On nonparametric estimates of the expectation. In *Colloquium on Limit Theorems in Probability and Statistics, Abstracts of the Talks.* BJMT. Balatonlelle, Hungary.

Antos, A., Devroye, L., and Györfi, L. (1999). Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* In print.

Antos, A., Györfi, L., and Kohler, M. (1999). Lower bounds on the rate of convergence of nonparametric regression estimates. *Journal of Statistical Planning and Inference.* In print. Preprint 98-11, Universität Stuttgart, Math. Inst. A, D-70511 Stuttgart, 1998.

Antos, A. and Lugosi, G. (1998). Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56. Economics Working Paper 197, Universitat Pompeu Fabra, 1997.

Assouad, P. (1983). Densité et dimension. *Annales de l'Institut Fourier*, 33:233–282.

Bahr, B. V. and Esseen, C. (1965). Inequalities for the rth absolute moment of a sum of independent random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.*, 36:299–303.

Barron, A. R., Birgé, L., and Massart, P. (1995). Risk bounds for model selection via penalization. Technical Report 95.54, Université Paris–Sud. Published in Probability Theory and Related Fields.

Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 65:181–237.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71:271–291.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965.

Bretagnolle, J. and Huber, C. (1979). Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47:119–137.

Chen, Z. and Fu, K. S. (1973). Nonparametric Bayes risk estimation for pattern classification. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*. Boston, MA.

Cover, T. M. (1968). Rates of convergence for nearest neighbor procedures. In *Proceedings of the Hawaii International Conference on Systems Sciences*, pages 413–415. Honolulu, HI.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, 2:299–318.

Devroye, L. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:154–157.

Devroye, L. (1983a). The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. *Annals of Statistics*, 11:896–904.

Devroye, L. (1983b). On arbitrarily slow rates of global convergence in density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:475–483.

Devroye, L. (1991). Exponential inequalities in nonparametric estimation. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, pages 31–44. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Devroye, L. (1995). Another proof of a slow convergence result of Birgé. *Statistics and Probability Letters*, 23:63–67.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley, New York.

Devroye, L. and Györfi, L. (1990). No empirical probability measure can converge in the total variation sense for all distributions. *Annals of Statistics*, 18:1496–1499.

Devroye, L., Györfi, L., Krzyżak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics, Stochastic Modelling and Applied Probability*. Springer-Verlag, New York.

Devroye, L. and Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28:1011–1018.

Ehrenfeucht, A., Haussler, D., Kearns, M., and Valiant, L. (1989). A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261.

Fukunaga, K. and Hummels, D. M. (1987). Bayes error estimation using Parzen and k-NN procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:634–643.

Fukunaga, K. and Kessel, D. L. (1971). Estimation of classification error. *IEEE Transactions Computers*, 20:1521–1527.

Garnett, J. M. and Yau, S. S. (1977). Nonparametric estimation of the Bayes error of feature extractors using ordered nearest neighbour sets. *IEEE Transactions on Computers*, 26:46–54.

Györfi, L., Páli, I., and van der Meulen, E. C. (1994). There is no universal source code for infinite alphabet. *IEEE Transactions on Information Theory*, 40:267–271.

Haussler, D., Littlestone, N., and Warmuth, M. (1994). Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115:248–292.

Ibragimov, I. A. and Khasminkii, R. Z. (1980). On nonparametric estimation of regression. *Doklady Acad. Nauk SSSR*, 252:780 – 784.

Ibragimov, I. A. and Khasminskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

Ibragimov, I. A. and Khasminskii, R. Z. (1982). On the bounds for quality of nonparametric regression function estimation. *Theory of Probability and its applications*, 27:81–94.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*, pages 126–169. Springer Lecture Notes in Mathematics, Springer-Verlag, Berlin.

Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction*. Springer, Berlin.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127.

Liese, F. and Vajda, I. (1987). *Convex statistical distances*. Teubner, Leipzig.

Lugosi, G. (1995). Improved upper bounds for probabilities of uniform deviations. *Statistics and Probability Letters*, 25:71–77.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. Revised.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley, New York.

Petrov, V. V. (1995). *Limit theorems of probability theory.* Clarendon Press, Oxford.

Reiss, R. D. (1989). *Approximate distributions of order statistics.* Springer-Verlag, New York.

Schuurmans, D. (1996). *Effective classification learning.* PhD Thesis, University of Toronto, Toronto, CA.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 8:1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053.

Vajda, I. (1989). *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht.

Vapnik, V. N. and Chervonenkis, A. Y. (1974). *Theory of Pattern Recognition.* Nauka, Moscow. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

Yang, Y. (1999). Minimax nonparametric classification — part I: Rates of convergence. *IEEE Transactions on Information Theory.* Accepted.