Consistency results in nonparametric regression estimation and classification

Márta Pintér

Supervisor: Dr. László Györfi

Budapest University of Technology and Economics
Computer and Automation Research Institute, Hungarian Academy of Sciences

Budapest, 2001

Contents

Acknowledgement

I am deeply grateful to my supervisor László Györfi for the kind guidance, support and encouragement he gave me, and to Gábor Lugosi for initiating me into the subject of pattern recognition. It was an honour to work with them. I am grateful to Michael Kohler and Kinga Máthé, who are coauthors of some results treated in this dissertation. I would also like to thank my colleagues at the Department of Computer Science and Information Theory of the Budapest University of Technology and Economics and at the Computer and Automation Research Institute of the Hungarian Academy of Sciences for providing such a pleasant working atmosphere. I particularly want to thank Lajos Rónyai for the many valuable suggestions. The financial support of the Soros Foundation (PhD Dissertation Grant No. 230/2/856) is gratefully acknowledged. Finally, I would like to express my debt of gratitude to my family for their warm, loving support and understanding.

Chapter 1

Introduction

Nonparametric estimation has grown to a central topic in statistics. The ambition to avoid misspecification that may occur when parametric techniques are used has joined forces with increased computer power and availability of very large data sets to increase the interest in nonparametric functional estimation. Nonparametric methods are important tools in various fields of applications including mathematics, engineering, medicine, natural and social sciences. In mathematical statistics we intend to learn a probability law or its characteristics from the available data. We wish to derive statements and decisions regarding the nature of the phenomenon examined. Classical mathematical statistics consists mainly of parametric rules, when the statistician has some a priori information about the actual problem and has reasons to believe that the unknown probability law belongs to a parametric class of distributions. A major drawback of this approach is that the observed data may fail to follow a specific parametric model and the incorrect modelling assumption may lead to seriously flawed statistical conclusions. In case of nonparametric statistics the statistician does not have a simple, parametrized family of distributions to select from. As a consequence, the construction of the rules cannot use any specific feature of the law, and the basic properties of the rule should be distribution-free.

From the rich variety of research directions in nonparametric statistics, we studied problems of regression function estimation and pattern recognition (or classification). In both cases we want to guess or predict an unknown quantity (covariate) from an observation. In regression estimation this quantity takes its value on the real line. One can think, as examples for the unknown covariate, the lifetime of a machine or a patient, wheat crop, the profit on the stock market. In pattern recognition the observation is to be classified into one of a finite number of classes such as one or zero, sick or healthy, real or fake. The observation is a collection of measurements, e.g. image, weather data, medical file, digitalized signature, electrocardiogram. The estimation or prediction is based on previously collected data, independent sample pairs of observation and covariate, and a loss function is defined to measure the quality of the estimate.

In regression estimation this loss function is the mean squared error, which is known to be minimized by the regression function. Therefore our aim is to estimate the regression function for which several consistent methods exist. However, often it is not possible to collect data about the unknown quantity of interest. Think of, for example, medical or other experimental studies when we wish to predict the lifetime of an individual or subject taking part in the study. Often (and luckily) some of the patients are still alive at termination of the study, or withdraw alive during the study or die from other causes than those under study. Hence, instead of an observation of the survival time of a patient one observes only the minimum of the survival time and a so called censoring time. We consider regression estimation from censored data in Chapter 2.

In pattern recognition the loss function we wish to minimize is the error probabilty, the probability that our decision is wrong in the sense that we classify the observation into an incorrect class. This happens if we say that a patient is healthy although he has a disease or if we declare a handwritten character “c” when it is really an “o”. Again, we can say that consistent classification rules exist: we have classification rules for which the error probabilty tends to the optimum, but a universal rate of convergence cannot be guaranteed. In Chapter 3 we study the rate of convergence of a particular classification rule and of two error estimates. Very often, in addition to the collected data, some supplementary information is available about the relationship between the observation and the covariate we wish to guess. In Chapter 4 we introduce and analyze methods which effectively exploit this prior knowledge.

1.1 Regression function estimation

Let X be a random vector taking values in $\mathcal{R}^d$ and let Y be a real random variable with $\mathbf{E}Y^2 < \infty$. The coordinates of X may have various types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. Hence, we do not assume anything on the distribution of X. In regression analysis one wants to estimate Y after having observed X, i.e. one wishes to find a function f with $f(X)$ "close" to Y. If the main goal of the analysis is the minimization of the mean squared error, then one wants to find a function f^* with

$$\mathbf{E}\,|f^*(X) - Y|^2 = \min_f \mathbf{E}\,|f(X) - Y|^2 . \tag{1.1}$$

Let $m(x) := \mathbf{E}\{Y|X = x\}$ be the regression function and denote the distribution of X by μ, that is,

$$\mu(A) = \mathbf{P}\{X \in A\} \quad \text{for each measurable set } A.$$

It is well known that for each measurable function f the relation

$$\mathbf{E}|f(X) - Y|^2 = \mathbf{E}|m(X) - Y|^2 + \int_{\mathcal{R}^d} |\, f(x) - m(x) \,|^2 \, \mu(dx) \tag{1.2}$$

holds. Therefore m is the solution of the minimization problem (1.1) and the mean squared error $\mathbf{E}|f(X) - Y|^2$ is close to its optimal value $\mathbf{E}|m(X) - Y|^2$ if and only if

$$\int_{\mathcal{R}^d} |\, f(x) - m(x) \,|^2 \, \mu(dx)$$

(in the following denoted as L_2 error) is close to zero.

For the regression estimation problem the distribution of (X, Y) (and therefore also m) is unknown. Given only a sample (X_1, Y_1),...,(X_n, Y_n) of the distribution of (X, Y) one has to construct an estimate $m_n(x)$=$m_n(x, (X_1, Y_1), ..., (X_n, Y_n))$ of $m(x)$ such that $\int_{\mathcal{R}^d} |m_n(x) - m(x)|^2 \mu(dx)$ is small.

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

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

and it is called strongly consistent for a certain distribution of (X,Y)*, if*

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

A regression function estimate can be consistent for a certain class of distributions of (X,Y), but may not be consistent for others. Since in many situations we do not have any prior information about the distribution, it is essential to have estimates that give good performance for *all* distributions.

DEFINITION 1.2. *A sequence of regression function estimates* $\{m_n\}$ *is called weakly (strongly) universally consistent if it is (strongly) consistent for all distributions of* (X,Y) *with* $\mathbf{E}\{Y^2\}<\infty$.

The existence of weakly universally consistent estimates was shown for the first time in Stone (1977). He considered estimates of form

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

where $W_{ni}(x)=W_{ni}(x,X_1,\ldots,X_n)$ are weights which depend on x and on X_1, ..., X_n. These estimators are of local averaging type, they average those Y_i's where X_i is close to x.

Well-known examples of this type of estimates are k_n–*nearest neighbor estimates*, where $k_n\in\mathbb{N}$ is a parameter of the estimate and one uses weights defined (modulo some suitable tie–breaking strategy – see Devroye et al. (1994) for details) by

$$W_{ni}(x)=\begin{cases}\frac{1}{k_n} & \text{if } X_i \text{ is among the } k_n \text{ nearest neighbors of } x \\ & \ \text{in } \{X_1,\ldots,X_n\}, \\ 0 & \text{else,}\end{cases} \tag{1.3}$$

kernel estimates, which use as parameters a positive number h_n (so–called bandwidth) and a function $K:\mathcal{R}^d\to\mathcal{R}$ (so–called kernel function) and where one uses weights

$$W_{ni}(x)=\frac{K\left(\frac{x-X_i}{h_n}\right)}{\sum_{j=1}^n K\left(\frac{x-X_j}{h_n}\right)} \quad \left(\frac{0}{0}:=0\right), \tag{1.4}$$

and *partitioning estimates*, which use a partition $\mathcal{P}_n = \{A_{nj} \; : \; j\}$ as a parameter and weights defined by

$$W_{ni}(x) = \sum_j \frac{I_{\{X_i \in A_{nj}\}}}{\sum_{k=1}^n I_{\{X_k \in A_{nj}\}}} I_{\{x \in A_{nj}\}} \quad \left(\frac{0}{0} := 0\right). \tag{1.5}$$

Here I_A denotes the indicator function of the set A.

LEMMA 1.1. *(Devroye, Györfi, Krzyżak and Lugosi (1994))* *If* $\lim_{n\to\infty} k_n/\log(n) = \infty$ *and* $\lim_{n\to\infty} k_n/n = 0$ *then the* k_n*-nearest neighbor estimate is strongly universally consistent.*

For the kernel estimate Devroye and Wagner (1980) and Spiegelman and Sacks (1980) proved weak universal consistency.

LEMMA 1.2. *(Devroye and Krzyżak (1989)) The kernel estimate is strongly consistent if* Y *is uniformly bounded,* $h_n \to 0$ *and* $nh_n^d \to \infty$, K *is a regular kernel (nonnegative, Riemann integrable kernel and* $K \geq aI_S$, *where* $a > 0$ *is a constant, and* S *is a ball centered at the origin that has a positive radius).*

Results on weak and strong universal consistencies of the partitioning estimate are in Devroye, Györfi (1983) and Györfi (1991).

LEMMA 1.3. *(Devroye, Györfi (1983)) Let the partition be such that for each sphere* S *centered at the origin* $\lim_{n\to\infty} \max_{j: A_{nj}\cap S \neq \emptyset} diam(A_{nj}) = 0$ *and* $\lim_{n\to\infty} \frac{|\{j: A_{nj}\cap S\neq\emptyset\}|}{n} = 0$ *then the partitioning estimate is strongly universally consistent.*

Another approach to regression estimation that can also lead to consistent estimates is to minimize the empirical error.

Least squares estimates are based on empirical error minimization. Recall that the regression function minimizes the so–called L_2 error $\mathbf{E}|f(X) - Y|^2$ over all (measurable) functions $f : \mathcal{R}^d \to \mathcal{R}$. This motivates to construct an estimate of m by minimizing an estimate of the L_2 error, the so–called *empirical* L_2 *error*

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

If one minimizes (1.6) over all (measurable) functions $f : \mathcal{R}^d \to \mathcal{R}$, then one gets (at least if the $X_1, \dots, X_n$ are all distinct) a function which interpolates the data $(X_1, Y_1), \dots, (X_n, Y_n)$. Obviously, this is not a reasonable estimate of the regression function. Therefore one chooses sets $\mathcal{F}_n$ of functions $f : \mathcal{R}^d \to \mathcal{R}$, which depend on the sample size n and which get more and more complex for n tending to infinity, and minimizes (1.6) only over $\mathcal{F}_n$, i.e., one defines least squares estimates m_n by

$$m_n(\cdot) = \arg \min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2. \tag{1.7}$$

Here we assume for simplicity that the minima in (1.7) exists, but we do not assume that it is unique.

Instead of restricting the class of function over which one minimizes the empirical L_2 error, one can also add a penalty term to the empirical L_2 error. More precisely, let $k \in \mathbb{N}$ with $2k > d$ and denote by $W^k(\mathcal{R}^d)$ the Sobolev space containing all functions $f : \mathcal{R}^d \to \mathcal{R}$ whose derivatives of total order k are in $L_2(\mathcal{R}^d)$. The condition $2k > d$ implies that the functions in $W^k(\mathcal{R}^d)$ are continuous and hence the evaluation of a function at a point is well defined. Set

$$J_k^2(f) = \sum_{\alpha_1,\dots,\alpha_d \in \mathbb{N},\, \alpha_1+\dots+\alpha_d=k} \frac{k!}{\alpha_1! \cdot \ldots \cdot \alpha_d!} \int_{\mathcal{R}^d} \left| \frac{\partial^k f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}(x) \right|^2 dx. \tag{1.8}$$

Then the *penalized least squares* (or *smoothing spline*) *estimate* m_n is defined by

$$m_n(\cdot) = \arg \min_{f \in W^k(\mathcal{R}^d)} \left(\frac{1}{n} \sum_{i=1}^{n} |f(X_i) - Y_i|^2 + \lambda_n J_k^2(f) \right), \tag{1.9}$$

where $\lambda_n > 0$ is a parameter of the estimate. Again we do not require for the minimum to be unique. A function which achieves the minima in (1.9) can be computed by solving a linear equation system (see, e.g., Duchon (1976), Section V).

Strong universal consistency of various least squares estimates was shown in Lugosi and Zeger (1995), Kohler (1997; 1999), and of penalized least squares estimates in Kohler and Krzyżak (2001).

In Chapter 2 we consider regression estimation from censored data, i.e. when it is not possible to observe a sample of (X, Y), but Y is subject to right censoring. Our data consists of (X, Z, δ) triples, where Z is the minimum of Y and the censoring time C, and δ shows whether what we observed Y or C. In Section 2.3 we assume that (X, Y) and C are independent and present versions of kernel, partitioning, nearest neighbor, least squares and penalized least squares estimates that can be used for prediction from randomly right censored data. We prove strong consistency of the estimates for every distribution of (X, Y, C) with Y nonnegative and almost surely bounded. In Section 2.4 we extend the well-known Kaplan-Meier method (Section 2.1) to conditional survival functions and present strongly consistent survival and regression function estimates based on local averaging assuming only that Y and C are conditionally independent given X.

As in the above cited results concerning universally consistent estimation of the regression function in the uncensored case, our results require no restriction on the distribution of X.

1.2 Pattern recognition

In pattern recognition Y can take only two values 0 and 1. Let (X, Y) be a pair of random variables, taking their values from some set $\mathcal{X}$ and $\{0, 1\}$, respectively. The value of the *label* Y is to be predicted upon observing the *feature vector* X. The *prediction rule* or *classifier* g is a function $\mathcal{X} \to \{0, 1\}$, whose performance is measured by the *probability of error*

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

The joint distribution of (X, Y) is determined by the *regression function*

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

(also known as the *a posteriori probability function*) and the measure μ of X on $\mathcal{X}$. The Bayes classifier

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

is well-known to have minimal probability of error among all possible classifiers. Its error probability $L(g^*)$ is called the Bayes risk (Bayes error), and is denoted by L^*.

$$L^* = \mathbf{E}\{\min(\eta^*(X), 1 - \eta^*(X))\}.$$

Assume that n independent copies of (X, Y) form the available data sequence:

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n)).$$

These data may be used to obtain the classification rule $g_n(x)$, whose probability of error is the random variable

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

Like in regression estimation, we can define weak and strong consistency.

DEFINITION 1.3. *A classification rule is consistent (or asymptotically Bayes-risk efficient) for a certain distribution of* (X, Y) *if*

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X, D_n) \neq Y\} \to L^* \quad \textit{as } n \to \infty,$$

and strongly consistent if

$$\lim_{n\to\infty} L_n = L^* \quad \textit{with probability } 1.$$

Consistency is defined as the convergence of the expected value of L_n to L^*. Since L_n is a random variable bounded between L^* and 1, this convergence is equivalent to the convergence of L_n to L^* in probability, which means that for every $\epsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\{L_n - L^* > \epsilon\} = 0.$$

A decision rule can be consistent for a certain class of distributions of (X, Y), but may not be consistent for others. It is clearly desirable to have a rule that is consistent for a large class of distributions. Since in many situations we do not have any prior information about the distribution, it is essential to have a rule that gives good performance for *all* distributions.

DEFINITION 1.4. *A sequence of decision rules is called universally (strongly) consistent if it is (strongly) consistent for any distribution of the pair* (X, Y).

If $\eta : \mathcal{X} \to [0,1]$ is an arbitrary measurable function, and we define the corresponding *plug-in* classifier by

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

then the following elementary property holds:

$$\begin{aligned} L(g) - L^* &= 2\mathbf{E}\left\{ I_{\{g(X)\neq g^*(X)\}} \left| \eta^*(X) - \frac{1}{2} \right| \right\} \\ &\leq 2\mathbf{E}\left\{ I_{\{g(X)\neq g^*(X)\}} \left| \eta^*(X) - \eta(X) \right| \right\}, \end{aligned} \tag{1.10}$$

see, for example, Devroye, Györfi and Lugosi (1996, p.16). Let $L(\eta)$ denote the probability of error $L(g)$ of the corresponding classifier $g(x)$. We will always denote $g'(x) = I_{\{\eta'(x)\geq 1/2\}}$, $g^*(x) = I_{\{\eta^*(x)\geq 1/2\}}$, $g_n(x) = I_{\{\eta_n(x)\geq 1/2\}}$, etc.

Consistency of classification rules can be deduced from consistent regression estimation. If η_n is a regression function estimate and g_n is the corresponding plug-in rule, then from (1.10) and the Cauchy-Schwarz inequality

$$L(g_n) - L^* \leq 2\sqrt{\int_{\mathcal{R}^d} |\eta^*(x) - \eta_n(x)|^2 \mu(dx)}. \tag{1.11}$$

Therefore the mere existence of a regression function estimate $\eta_n(x)$ for which

$$\int |\eta(x) - \eta_n(x)|^2 \mu(dx) \to 0$$

in probability or with probability one implies that the plug-in decision rule g_n is consistent or strongly consistent, respectively.

Classification rules based on local majority voting, such as partitioning, kernel, nearest neighbor rule, can be derived in a natural way as the plug-in classifiers of the local averaging regression estimates, and their consistency follows from the consistency of the regression estimates.

In Chapter 3 partitioning classification rules are studied. We show in Section 3.1 that, although universal rate of convergence doesn't exist, the error probability of the partitioning classification rule converges to the Bayes error faster than $1/\sqrt{n}$ under certain conditions. In Section 3.2 we investigate two error estimates, the resubstitution and the deleted estimate (or cross validation) and prove asymptotic normality of the error estimates.

Another type of classificaton rules select a member of a fixed class of classifiers by minimizing the *empirical error*

$$\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^{n} I_{\{g(X_i)\neq Y_i\}}.$$

For details on the behaviour see Chapter 4.

Very often, apart from the training sequence, some prior information is available about the joint distribution of (X, Y). In Chapter 4 the classical binary classification (pattern recognition) problem is investigated when it is known in advance that the posterior probability function (or regression function) belongs to some class of functions. We introduce and analyze methods which effectively exploit this knowledge. These methods are based on minimizing the empirical risk over a carefully selected "skeleton" of the class of regression functions. The skeletons are coverings of the class based on metrics, especially fitted for classification. A new scale-sensitive dimension is introduced which is more suitable for the studied classification problem than other, previously defined, dimension measures. This fact is demonstrated by performance bounds for the skeleton estimates in terms of the new dimension.

Chapter 2

Prediction from randomly right censored data

This chapter deals with nonparametric regression analysis in the presence of randomly right censored data.

Let Y be a non-negative random variable representing the survival time of an individual or subject taking part in a medical or other experimental study, and let X be a random vector of covariates, e.g., the medical file of a patient, jointly distributed with Y. In many medical studies it is not possible to observe a sample of (X, Y), the survival time Y is subject to right censoring. Often some of the patients are still alive at termination of the study, or withdraw alive during the study or die from other causes than those under study. Hence, instead of an observation of the survival time of a patient one observes only the minimum of the survival time and a censoring time, i.e. the observable random variables are given by X_i, $Z_i = \min(Y_i, C_i)$, and $\delta_i = I_{\{Y_i \le C_i\}}$. Here C_i is a non-negative random variable, representing the time of censoring, which could be the subject's time to withdrawal or the time until the end of the study.

Another example of the same model can be taken from the world of medical insurance. Suppose an insurance company sets its various insurance rates on basis of the lifetimes Y_i of its clients having a specific health condition (heart disease, diabetes, etc.). In addition, for each client a vector X_i of additional measurements is available from medical examinations. However, for several reasons, a patient may terminate the contract with the insurance company at time C_i (due to lack of money or because of more favorable conditions elsewhere), so that the actual file of the patient at the insurance company is incomplete in the above sense: it does not contain Y_i, it contains only C_i with the indication that it is not the real lifetime. Moreover, in this example it can be assumed that the reasons for stopping the contract are unrelated to any characteristics of the individual itself.

The distribution of the observable random vector (X, Z, δ) does not identify the conditional survival distribution of Y given X. The problem of identifiability can be resolved by imposing additional conditions on the possible distributions for (X, Y, C).

Model A: Assume that C and (X, Y) are independent.

Model B: Assume that Y and C are conditionally independent given X.

Obviously Model A is a special case of Model B, therefore the estimation problem for Model A is easier. Model A is plausible in all situations where the censoring is caused circumstances not related to any characteristics of the individual. This can clearly be argued to be the case in the medical insurance example described above, as well as in many classical survival studies. One could think e.g. of volunteers participating in a medical study, but for reasons not related to the data contained in (X, Y), such as lack of enthusiasm they withdraw prematurely from the study. Censoring may also be caused by the (random) termination of the study, which can be assumed to occur independently of the persons participating in the study.

Statistical approaches to the survival analysis can be broadly classified as regression based and hazard risk based methods. In the latter, one studies the contribution of the risk factors by modelling the conditional hazard rate function under specific assumptions on the conditional distribution of Y, see, e.g., Cox (1972; 1975). An overview of the literature related to survival analysis can be found, e.g., in Fan and Gijbels (1995).

An estimation of the slope in a linear regression model via the proportional hazard model was proposed by Cox (1972). Based on the Kaplan-Meier estimator, Buckley and James (1979) intoduced an estimator, which estimates a linear regression function directly from the censored data. The consistency of this estimator was investigated by James and Smith (1984). For a slight modification of the estimator Ritov (1990) and Lai and Ying (1991) established the asymptotic normality.

In practical applications it is normally not known whether or not the model is linear. Important diagnostic tools for uncensored regression as scatterplots or residualplots are not directly applicable in the censored case. Therefore we are interested in nonparametric regression estimates for the randomly right censored case.

Without censoring, i.e. for $C = \infty$ almost surely, it is well–known that there exist estimates which are weakly and strongly universally consistent, i.e. for which the L_2 error tends to 0 in probability or almost surely for every distribution of (X, Y) with $\mathbf{E}|Y|^2 < \infty$ (see Section 1.1 of Introduction).

In Section 2.1 we introduce the Kaplan-Meier estimator of survival functions and give a proof of uniform consistency.

In Section 2.3 we present versions of kernel, partitioning, nearest neighbor, least squares and penalized least squares estimates that can be used for prediction from randomly right censored data. Under some assumptions we prove the strong consistency of the estimates for every distribution of (X, Y, C) with Y nonnegative and almost surely bounded.

In Section 2.4 we extend the Kaplan-Meier method to conditional survival functions and present strongly consistent survival and regression function estimates based on local averaging.

We want to stress that the results of Section 2.3 and 2.4 hold without any assumption on the distribution of X or on the smoothness of the underlying regression function.

2.1 Kaplan-Meier estimate

A crucial point in the definition of regression estimates is estimation of a mean from i.i.d. data. In the definition of (penalized) least squares estimates the L_2 risk is estimated by (the sum of a penalty term and the) empirical L_2 risk (1.6), and in the definition of local averaging estimates the local average

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

is used as an approximation of the conditional expectation of Y given $X = x$. We will show how these estimates can be adapted to the case of censored data. For this we use an idea introduced in Carbonez, Györfi and van der Meulen (1995).

Let

$$T_F = \sup\{t : F(t) > 0\},$$

$$T_G = \sup\{t : G(t) > 0\},$$

$$T_K = \min\{T_F, T_G\},$$

where F and G are the survival functions of Y and C, i.e. $F(t) = \mathbf{P}\{Y > t\}$ and $G(t) = \mathbf{P}\{C > t\}$.

Let us first consider the case when there are no covariates. If the Y_i are i.i.d. and available, then the arithmetic mean

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

is the usual estimate of $\mathbf{E}\{Y\}$. If instead of the Y_i only the (Z_i, δ_i) are available, and the censoring distribution G is known, then an unbiased estimate of $\mathbf{E}\{Y\}$ can be given by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Z_i}{G(Z_i)}, \tag{2.1}$$

Indead, using properties of conditional expectation and the assumed independence of Y and C:

$$\begin{aligned}
\mathbf{E}\left\{\frac{\delta Z}{G(Z)}\right\} &= \mathbf{E}\left\{\frac{I_{\{Y\le C\}}\min(Y,C)}{G(\min(Y,C))}\right\} \\
&= \mathbf{E}\left\{I_{\{Y\le C\}}\frac{Y}{G(Y)}\right\} \\
&= \mathbf{E}\left\{\mathbf{E}\left\{I_{\{Y\le C\}}\frac{Y}{G(Y)} \mid Y\right\}\right\} \\
&= \mathbf{E}\left\{G(Y)\frac{Y}{G(Y)}\right\} \\
&= \mathbf{E}\left\{Y\right\}.
\end{aligned}$$

In an application, G will usually be unknown, hence we have to estimate it.

Let F_n and G_n be the Kaplan-Meier estimator of F and G, respectively, introduced by Kaplan and Meier (1958).

$$F_n(t) = \begin{cases} \prod_{i:Z_{(i)}\le t}\left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}} & \text{if } t \le Z_{(n)} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and

$$G_n(t) = \begin{cases} \prod_{i:Z_{(i)}\le t}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}} & \text{if } t \le Z_{(n)} \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where $(Z_{(i)}, \delta_{(i)}), i = 1, \ldots, n$ are the n pairs of observed (Z_i, δ_i) ordered on the $Z_{(i)}$, i.e.

$$Z_{(1)} \le Z_{(2)} \le \ldots \le Z_{(n)} =: T_{K_n}. \tag{2.4}$$

It is easy to show that $T_{K_n} \to T_K$ as $n \to \infty$. For G arbitrary choosen, some of the Z_i may be identical, so that the ordering is not unique. However, the Kaplan-Meier estimator is unique (see Peterson (1977)).

The estimates $F_n(t)$ and $G_n(t)$ have been invented by Kaplan and Meier (1958). These estimates are often called product-limit estimates. It is not at all obvious why they should be consistent. The first interpretation in this respect is due to Kaplan and Meier (1958) showing that the estimates are of maximum likelihood type.

Peterson (1977) proved that if F and G don't have common jumps, then the Kaplan-Meier estimates are consistent, i.e. $F_n(t) \longrightarrow F(t)$ and $G_n(t) \longrightarrow G(t)$ almost surely. Stute and Wang (1993) showed that if in addition either F is continuous in T_K or F has a jump in T_K but $G(T_K^-) > 0$ then F_n is uniformly consistent on $(-\infty, T_K]$. We present a much simpler proof for the uniform convergence of the Kaplan-Meier estimates on $(-\infty, T_K)$ without the additional conditions.

Theorem 2.1. *(Stute and Wang (1993)) If F and G don't have common jumps, then*

(i) $\sup_{t<T_K} |G_n(t) - G(t)| \longrightarrow 0$

(ii) $\sup_{t<T_K} |F_n(t) - F(t)| \longrightarrow 0$

a.s. as $n \to \infty$.

Proof. (i): We prove pointwise convergence of $G_n(t)$ to $G(t)$ for all $t < T_K$. Since G_n and G are bounded and nonincreasing functions uniform convergence follows from pointwise convergence (see e.g. the proof of Glivenko-Cantelli theorem in Rényi (1968), p. 333).

For a function S let $D(S)$ denote the set of the jump points of S. Then any survival function can be expressed as

$$S(t) = \exp\left(\int_{[0,t]\setminus D(S)} \frac{dS(s)}{S(s)}\right) \cdot \prod_{\substack{s\in D(S)\\ s\le t}} \frac{S(s)}{S(s^-)}$$

(see Cox (1972))).
Thus

$$G(t) = \exp\left(\int_{[0,t]\setminus D(G)} \frac{dG(s)}{G(s)}\right) \cdot \prod_{\substack{s\in D(G)\\ s\le t}} \frac{G(s)}{G(s^-)} \tag{2.5}$$

and

$$G_n(t) = \prod_{\substack{s\in D(G_n)\\ s\le t}} \frac{G_n(s)}{G_n(s^-)}$$

since $G_n(\cdot)$ is a discrete survival function.

Define the subsurvival functions

$$G^*(t) = \mathbf{P}(Z > t, \delta = 0),$$

$$F^*(t) = \mathbf{P}(Z > t, \delta = 1).$$

$G^*(\cdot)$ can be expressed in terms of the survival functions $G(\cdot)$ and $F(\cdot)$

$$G^*(t) = \int_t^\infty F(s)[-dG(s)], \tag{2.6}$$

because the random variables Y and C are independent. The integral is well-defined since by assumption the integrator and the integrand have no jumps in common.
Note that

$$\begin{aligned} G^*(t) + F^*(t) &= \mathbf{P}(Z > t) = K(t) \\ &= \mathbf{P}(\min(Y, C) > t) = G(t) \cdot F(t). \end{aligned} \tag{2.7}$$

For real valued functions H_1 and H_2 define a function ϕ by

$$\phi(H_1, H_2, t) = \exp\left(\int_{[0,t]\setminus D(H_1)} \frac{dH_1(s)}{H_1(s) + H_2(s)}\right) \cdot \prod_{\substack{s\in D(H_1)\\ s\le t}} \frac{H_1(s) + H_2(s)}{H_1(s^-) + H_2(s^-)}.$$

We will show that

$$G(t) = \phi(G^*(\cdot), F^*(\cdot), t). \tag{2.8}$$

Clearly, G^* has the same jump points as G. Thus (2.8) follows from (2.5), since

$$\begin{aligned}
\frac{dG^*(s)}{G^*(s)+F^*(s)} &= \frac{dG^*(s)}{G(s)\cdot F(s)} && \text{(from (2.7))} \\
&= \frac{F(s)\cdot dG(s)}{F(s)\cdot G(s)} && \text{(from (2.6))} \\
&= \frac{dG(s)}{G(s)}
\end{aligned}$$

and

$$\begin{aligned}
\frac{G^*(s)+F^*(s)}{G^*(s^-)+F^*(s^-)} &= \frac{G(s)\cdot F(s)}{G(s^-)\cdot F(s^-)} && \text{(from (2.7))} \\
&= \frac{G(s)}{G(s^-)}
\end{aligned}$$

if s is a jump point of $G(\cdot)$, because by assumption $G(\cdot)$ and $F(\cdot)$ don't have common jumps, thus s cannot be a jump point of $F(\cdot)$, so $F(s) = F(s^-)$.

Now we express the Kaplan-Meier estimator in terms of the empirical sub-survival functions

$$G_n^*(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i>t,\delta_i=0\}}$$

and

$$F_n^*(t) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i>t,\delta_i=1\}}.$$

First note that (2.7) is true also for the estimates, since

$$\begin{aligned}
G_n^*(t)+F_n^*(t) &= \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i>t,\delta_i=0\}} + \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i>t,\delta_i=1\}} \\
&= \frac{1}{n}\sum_{i=1}^{n} I_{\{Z_i>t\}} \\
&=: K_n(t)
\end{aligned}$$

and

$$G_n(t)\cdot F_n(t) = \prod_{Z_{(i)}\le t}\left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}} \cdot \prod_{Z_{(i)}\le t}\left(\frac{n-i}{n-i+1}\right)^{\delta_{(i)}}$$

$$
\begin{aligned}
&= \prod_{Z_{(i)} \le t} \left(\frac{n-i}{n-i+1} \right) \\
&= \frac{n-1}{n} \cdot \frac{n-2}{n-1} \cdots \frac{n - \sum_{i=1}^{n} I_{\{Z_i \le t\}}}{n - \sum_{i=1}^{n} I_{\{Z_i \le t\}} + 1} \\
&= \frac{1}{n} \sum_{i=1}^{n} I_{\{Z_i > t\}} \\
&= K_n(t).
\end{aligned}
$$

Therefore

$$\frac{G_n^*(s) + F_n^*(s)}{G_n^*(s^-) + F_n^*(s^-)} = \frac{G_n(s) \cdot F_n(s)}{G_n(s^-) \cdot F_n(s^-)} = \frac{G_n(s)}{G_n(s^-)}$$

if s is a jump point of $G_n(\cdot)$, because by assumption $G_n(\cdot)$ and $F_n(\cdot)$ don't have common jumps, thus $F_n(s) = F_n(s^-)$. Hence

$$G_n(t) = \phi(G_n^*(\cdot), F_n^*(\cdot), t)$$

for $t \le Z_{(n)}$.

The empirical subsurvival functions $G_n^*(t)$ and $F_n^*(t)$ converge almost surely uniformly in t (Glivenko-Cantelli theorem) and $Z_{(n)} = T_{K_n}$ converge almost surely to T_K. The almost sure convergence of $G_n(\cdot)$ to $G(\cdot)$ follows then from the fact that $\phi(H_1(\cdot), H_2(\cdot), t)$ is a continuous function of the argument $H_1(\cdot)$ and $H_2(\cdot)$, where the metric on the space of bounded functions is the supremum norm.

(ii) can be proved similarly.

□

Remark.

In order to get convergence also at T_K, the estimates need to be slightly modified, since if the last observation is censored, then G_n jumps to 0 and thus if $G(T_K) > 0$ then G_n cannot converge to G at T_K. But if we modify the estimate such that instead of jumping to 0 in the last observation it keeps the previous value, i.e.

$$G_n(t) = \begin{cases} \prod_{i: Z_{(i)} \le t} \left(\frac{n-i}{n-i+1}\right)^{1-\delta_{(i)}} & \text{if} \quad t < T_{K_n}, \\ \lim\limits_{s \nearrow T_{K_n}} G_n(s) & \text{if} \quad t \ge T_{K_n}, \end{cases} \tag{2.9}$$

then it becomes consistent in T_K if, e.g., G is continuous in T_K, and thus uniform convergence can be extended to the closed interval $(-\infty, T_K]$.

$\mathbf{E}\{Y\}$ can be estimated by the so-called (untruncated) Kaplan-Meier mean:

$$\begin{aligned} M_n := \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Z_i}{G_n(Z_i)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i Y_i}{G_n(Y_i)} \\ &\stackrel{(a)}{=} \int_0^{+\infty} F_n(y)dy \\ &= \int_0^{T_F} F_n(y)dy, \end{aligned} \tag{2.10}$$

where (a) follows from an identity in Susarla, Tsai and Van Ryzin (1984), which is based on the observation that $\frac{\delta_i}{nG_n(Z_i)}$ is the jump of F_n in Z_i.

Theorem 2.1 implies that $M_n \to \mathbf{E}\{Y\}$ a.s.

2.2 Covering numbers and dimensions

Now we make a digression and define some covering numbers and dimensions of function classes that will be useful later. First recall the notion of shattering and Vapnik-Chervonenkis dimension.

Definition 2.1. *$S_{\mathcal{F}}(x_1^n)$ is the* shatter coefficient *of the class $\mathcal{F}$ restricted to $x_1^n = (x_1, \dots, x_n)$, that is, the number of different ways the members of $\mathcal{F}$ can classify the n points $x_1, \dots, x_n$. x_1^n is shattered by $\mathcal{F}$ if for all subset $E \subseteq x_1^n$ there exists some function $\eta_E \in \mathcal{F}$ such that $\eta_E(x) \geq 1/2$ if $x \in E$, and $\eta_E(x) < 1/2$ if $x \in x_1^n \setminus E$. The* Vapnik-Chervonenkis or VC dimension *of $\mathcal{F}$, $\mathcal{V}_{\mathcal{F}}$, is the largest positive integer n for which there exists $x_1^n = (x_1, \dots, x_n)$ that is shattered by $\mathcal{F}$. If for every n there is a set x_1^n that is shattered by $\mathcal{F}$ then we say that $P_\gamma = \infty$.*

Let $\mathcal{A}$ be the class of sets of the form $\{x : \eta(x) \geq 1/2\}$, $\eta \in \mathcal{F}$. Then we have the following important result by Vapnik and Chervonenkis.

Lemma 2.1. *(Vapnik and Chervonenkis (1971)) For any probability measure ν and class of sets $\mathcal{A}$, and for any n and $\epsilon > 0$,*

$$\mathbf{P}\left\{\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon\right\} \leq 8 \sup_{x_1,\dots,x_n} S_{\mathcal{F}}(x_1^n) e^{-n\epsilon^2/32}.$$

Several generalizations of the VC dimension have been proposed, the following one is due to Kearns and Shapire (1994).

DEFINITION 2.2. *The* scale-sensitive dimension P_γ *(also known as the* fat shattering function*) for $0 < \gamma \leq 1/2$ is defined as follows: we say that $\mathcal{F}$* γ-fat-shatters x_1^n *if there exists some function $s : x_1^n \to [0,1]$ such that for every subset $E \subset x_1^n$ there is a function $\eta_E \in \mathcal{F}$ such that $\eta_E(x) \geq s(x) + \gamma$ if $x \in E$ and $\eta_E(x) \leq s(x) - \gamma$ if $x \in x_1^n \setminus E$. P_γ of $\mathcal{F}$ is the largest positive integer n for which there exists x_1^n that is γ-fat-shattered by $\mathcal{F}$. If for every n there is a set x_1^n that is γ-fat-shattered by $\mathcal{F}$ then we say that $P_\gamma = \infty$.*

Besides dimensions, one can measure the richness of a class of functions with covering numbers. Let $d_{H,n}(\eta, \eta') = \frac{1}{n}\sum_{i=1}^n I_{\{g(x_i) \neq g'(x_i)\}}$ denote the *empirical (normalized) Hamming distance* between two classifiers, and $d_{1,n}(\eta, \eta') = \frac{1}{n}\sum_{i=1}^n |\eta(x_i) - \eta'(x_i)|$ the *empirical L_1 distance.* Let $d_H(\eta, \eta') = \mathbf{P}\{g(X) \neq g'(X)\}$, and $d_1(\eta, \eta') = \mathbf{E}\{|\eta(X) - \eta'(X)|\}$, denote the corresponding *theoretical distances.*

DEFINITION 2.3. *$N_H(\epsilon, x_1^n, \mathcal{F})$ is the* empirical Hamming covering number *of $\mathcal{F}$ restricted to the set $\{x_1, \ldots, x_n\}$, that is, the cardinality of the smallest set of functions satisfying the property that for every $\eta \in \mathcal{F}$ there is an η' in the set such that $d_{H,n}(\eta, \eta') < \epsilon$,*

The empirical L_1 covering number *$N_1(\epsilon, x_1^n, \mathcal{F})$ is defined similarly, replacing the distance $d_{H,n}$ above by the empirical L_1 distance $d_{1,n}$.*

The covering numbers of $\mathcal{F}$ with respect to the theoretical distances d_H and d_1 are denoted by $N_H(\epsilon, \mu, \mathcal{F})$ and $N_1(\epsilon, \mu, \mathcal{F})$, respectively.

Note that for $\epsilon > 1/n$, $N_H(\epsilon, x_1^n, \mathcal{F}) \leq S_{\mathcal{F}}(x_1^n)$.

If $Z_1^n = (Z_1, ..., Z_n)$ is a sequence of $\mathcal{R}^d$–valued random variables, then $N_1(\epsilon, Z_1^n, \mathcal{F})$ is a random variable with expected value $\mathbf{E}N_1(\epsilon, Z_1^n, \mathcal{F})$. The next result due to Pollard will be very useful.

LEMMA 2.2. *(Pollard (1984), section II.5, Th. 24) Let $\mathcal{F}$ be a class of functions $f : \mathcal{R}^d \to [0, B]$, and let $Z_1^n = (Z_1, ..., Z_n)$ be $\mathcal{R}^d$–valued i.i.d.*

random variables. Then for any $\epsilon > 0$

$\mathbf{P}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbf{E}f(Z_1)\right| > \epsilon\right] \leq 8\mathbf{E}\left(N_1(\frac{\epsilon}{8}, Z_1^n, \mathcal{F})\right)\exp\left(-\frac{n\,\epsilon^2}{128\,B^2}\right).$

A connection between covering numbers and VC dimension is given by the following lemma, which uses the notation $\mathcal{V}_{\mathcal{F}^+}$ for the VC dimension of the set

$$\mathcal{F}^+ := \left\{\{(x,t)\in\mathcal{R}^d\times\mathcal{R}\ :\ t\leq f(x)\}\ :\ f\in\mathcal{F}\right\}$$

of all graphs of functions of $\mathcal{F}$.

Lemma 2.3. *(Haussler (1992), Th. 6) Let* $\mathcal{F}$ *be a class of functions* $f : \mathcal{R}^d \to [-B, B]$. *Then one has for any* $z_1^n \in \mathcal{R}^{d\cdot n}$ *and any* $0 < \epsilon < B$

$$N_1(\epsilon, z_1^n, \mathcal{F}) \leq 2\left(\frac{4eB}{\epsilon}\log\left(\frac{4eB}{\epsilon}\right)\right)^{\mathcal{V}_{\mathcal{F}^+}}.$$

In this chapter we use only a part of the notions introduced here, the rest will be useful in Chapter 4.

2.3 Regression estimation for Model A

The regression function $m(x)$ is to be estimated from the censored data, which is an i.i.d. sequence of random variables: $\{(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)\}$, where $\delta_i = I_{\{Y_i\leq C_i\}}$, and $Z_i = \min(Y_i, C_i)$.

In Carbonez, Györfi and van der Meulen (1995) the same model of censored data is considered, partitioning estimates are defined and it is shown that they are strongly consistent for bounded Y. Our results are based on a simplification of their proof; furthermore we show the strong consistency in the randomly right censored case not only of partitioning, but also of kernel, nearest neighbor, least squares and penalized least squares estimates.

Let h be a function $h : \mathcal{R}^d \times \mathcal{R} \to \mathcal{R}$ and consider the problem of estimating the mean $\mathbf{E}\{h(X,Y)\}$ from the censored data $\mathcal{D}_n$. Then an unbiased estimate of $\mathbf{E}\{h(X,Y)\}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i h(X_i, Z_i)}{G(Z_i)}. \tag{2.11}$$

Unbiasedness can be verified the same way as in case of (2.1), using the independence of (X, Y) and C and properties of conditional expectations.

We use this idea to adapt the definiton of local averaging, least squares and penalized least squares estimates to the case of censored data.

Let G_n be the modified Kaplan-Meier estimate of G defined by (2.9). Throughout this section we rely on the following assumption:

ASSUMPTION 2.1. *(X, Y) and C are independent, $T_F < \infty$, G is continuous in T_K and $G(T_F) > 0$.*

For such a G the Kaplan-Meier estimate is uniformly consistent on $(-\infty, T_K]$. The last condition implies that $T_F < T_G$ and hence $T_K = T_F$ by definition of T_K.

We use the following notation: For any $0 \le t < \infty$ and $x \in \mathcal{R}$ define

$$T_{[0,t]}(x) = \begin{cases} t & \text{if} \quad x > t, \\ x & \text{if} \quad 0 \le x \le t, \\ 0 & \text{if} \quad x < 0. \end{cases} \tag{2.12}$$

For $f : \mathcal{R}^d \to \mathcal{R}$ define $T_{[0,t]}f : \mathcal{R}^d \to \mathcal{R}$ by $(T_{[0,t]}f)(x) := T_{[0,t]}(f(x))$ $(x \in \mathcal{R}^d)$.

2.3.1 Local averaging estimates

We define local averaging estimates for censored data by

$$\tilde{m}_{n,1}(x) = \tilde{m}_{n,1}(x, \mathcal{D}_n) = \sum_{i=1}^{n} W_{ni}(x) \cdot \frac{\delta_i \cdot Z_i}{G_n(Z_i)}. \tag{2.13}$$

The fact $0 \le Y \le T_F < \infty$ *a.s.* implies that $0 \le m(x) \le T_F$ $(x \in \mathcal{R}^d)$. In order to ensure that the estimates are bounded in the same way, set

$$m_{n,1}(x) = m_{n,1}(x, \mathcal{D}_n) = T_{[0,T_{K_n}]}\tilde{m}_{n,1}(x) \tag{2.14}$$

with T_{K_n} as in (2.4). Observe $0 \le T_{K_n} \le T_K = T_F < \infty$.

Our first result concerns the strong consistency of local averaging estimates for censored data.

THEOREM 2.2. *(Kohler, Máthé and Pintér (2000)) Assume that the weights $\{W_{ni}\}$ are nonnegative, that the sum of the weights is at most one and that for all distributions of (X, Y) with $|Y|$ bounded almost surely*

$$\int_{\mathcal{R}^d} \left(\sum_{i=1}^{n} W_{ni}(x) Y_i - m(x) \right)^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s. \tag{2.15}$$

Then the estimate $m_{n,1}$ defined by (2.13) and (2.14) satisfies

$$\int_{\mathcal{R}^d} |m_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for all distributions of (X, Y, C) satisfying Assumption 2.1.

In analogy to (2.14) set

$$m^{\star}_{n,1}(x) = T_{[0,T_F]} \tilde{m}_{n,1}(x) \tag{2.16}$$

In order to prove the theorem we need the following lemma:

LEMMA 2.4.

$$\int_{\mathcal{R}^d} |m_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

if and only if

$$\int_{\mathcal{R}^d} |m^{\star}_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

PROOF. It suffices to show

$$\int_{\mathcal{R}^d} |m^{\star}_{n,1}(x) - m_{n,1}(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s. \tag{2.17}$$

Since $T_{K_n} \le T_K = T_F$, it is easy to see that

$$|m^{\star}_{n,1}(x) - m_{n,1}(x)| = |T_{[0,T_F]} \tilde{m}_{n,1}(x) - T_{[0,T_{K_n}]} \tilde{m}_{n,1}(x)| \le T_F - T_{K_n} \quad (x \in \mathcal{R}^d).$$

Thus

$$\int_{\mathcal{R}^d} |m^{\star}_{n,1}(x) - m_{n,1}(x)|^2 \mu(dx) \le \int_{\mathcal{R}^d} (T_F - T_{K_n})^2 \mu(dx) = (T_F - T_{K_n})^2.$$

Since $T_{K_n} \to T_K$ $(n \to \infty)$ $a.s.$ and $T_K = T_F$, (2.17) follows. □

PROOF OF THEOREM 2.2 Because of Lemma 2.4 it suffices to show

$$\int_{\mathcal{R}^d} |m^\star_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

Set

$$\bar{m}_n(x) = T_{[0,T_F]}\left(\sum_{i=1}^n W_{ni}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)}\right).$$

Then

$$\int_{\mathcal{R}^d} |m^\star_{n,1}(x) - m(x)|^2 \mu(dx)$$
$$\leq 2\int_{\mathcal{R}^d} |m^\star_{n,1}(x) - \bar{m}_n(x)|^2 \mu(dx) + 2\int_{\mathcal{R}^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx). \quad (2.18)$$

Because of $0 \leq \frac{\delta_1 Z_1}{G(Z_1)} \leq \frac{T_F}{G(T_F)}$ $a.s.$ and

$$\begin{aligned}
\mathbf{E}\left\{\frac{\delta_1 Z_1}{G(Z_1)}\Big| X_1\right\} &= \mathbf{E}\left\{\frac{I_{[Y_1 \leq C_1]} Y_1}{G(Y_1)}\Big| X_1\right\} \\
&= \mathbf{E}\left\{\frac{Y_1}{G(Y_1)}\mathbf{E}\{I_{[Y_1 \leq C_1]}|X_1, Y_1\}\Big| X_1\right\} = m(X_1),
\end{aligned}$$

where the last equation follows from the independence of (X, Y) and C, (2.15) and $0 \leq m(x) \leq T_F$ $(x \in \mathcal{R}^d)$ imply

$$\begin{aligned}
&\int_{\mathcal{R}^d} |\bar{m}_n(x) - m(x)|^2 \mu(dx) \\
&\leq \int_{\mathcal{R}^d} \left|\sum_{i=1}^n W_{ni}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)} - m(x)\right|^2 \mu(dx) \to 0 \qquad (2.19)
\end{aligned}$$

$a.s.$ as $n \to \infty$. Furthermore, the assumptions that the weights are nonnegative and that their sum is at most one imply

$$\begin{aligned}
&\int_{\mathcal{R}^d} |m^\star_{n,1}(x) - \bar{m}_n(x)|^2 \mu(dx) \\
&\leq T_F \int_{\mathcal{R}^d} \left|\sum_{i=1}^n W_{ni}(x) \cdot \frac{\delta_i Z_i}{G_n(Z_i)} - \sum_{i=1}^n W_{ni}(x) \cdot \frac{\delta_i Z_i}{G(Z_i)}\right| \mu(dx) \\
&\leq T_F \int_{\mathcal{R}^d} \sum_{i=1}^n W_{ni}(x) \cdot T_F \cdot \left|\frac{1}{G_n(Z_i)} - \frac{1}{G(Z_i)}\right| \mu(dx) \\
&\leq T_F^2 \frac{1}{G(T_F) \cdot G_n(T_F)} \sup_{t \leq T_F} |G(t) - G_n(t)| \to 0 \quad (n \to \infty) \quad a.s.,
\end{aligned}$$

where we have used $G(T_F) > 0$ and the uniform consistency of G_n. This together with (2.18) and (2.19) implies the assertion. □

By Theorem 2.2 and Lemma 1.3 we get the following result concerning the consistency of partitioning estimates.

Corollary 2.1. *(Carbonez, Györfi and van der Meulen (1995), Kohler, Máthé and Pintér (2000)). Let* $\mathcal{P}_n = \{A_{nj} : j = 1, 2, \ldots\}$ *be partitions of* $\mathcal{R}^d$ *such that for each sphere* S *centered at the origin*

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

and

$$\lim_{n\to\infty} \frac{|\{j : A_{nj} \cap S \neq 0\}|}{n} = 0.$$

Define the partitioning estimate $m_{n,1}$ *by (2.13), (2.14) and (1.5). Then*

$$\int_{\mathcal{R}^d} |m_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for all distributions of (X, Y, C) *satisfying Assumption 2.1.*

By Theorem 2.2 and Lemma 1.2 we get the following result concerning the consistency of kernel estimates.

Corollary 2.2. *(Kohler, Máthé and Pintér (2000)) Let* K *be a regular kernel, i.e. it is nonnegative, and there is a ball* $S_{0,r}$ *of radius* $r > 0$ *centered at the origin, and constant* $b > 0$ *such that*

$$K(x) \geq bI_{\{x\in S_{0,r}\}}$$

and

$$\int_{\mathcal{R}^d} \sup_{u\in x+S_{0,r}} K(u)dx < \infty.$$

If $h_n \to 0$ *and* $nh_n^d \to \infty$, *then for the kernel regression estimate* $m_{n,1}$ *defined by (2.13), (2.14) and (1.4),*

$$\int_{\mathcal{R}^d} |m_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for all distributions of (X, Y, C) *satisfying Assumption 2.1.*

By Theorem 2.2 and Lemma 1.1 we get the following result concerning the consistency of nearest neighbor estimates.

Corollary 2.3. *(Kohler, Máthé and Pintér (2000)) If $k_n \to \infty$ and $k_n/n \to 0$, then for the nearest neighbor regression estimate $m_{n,1}$ defined by (2.13), (2.14) and (1.3),*

$$\int_{\mathcal{R}^d} |m_{n,1}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for all distributions of (X, Y, C) satisfying Assumption 2.1, and for which for each $x \in \mathcal{R}^d$ the random variable $\|X - x\|$ is absolutely continuous.

2.3.2 Least squares estimates

We define least squares estimates for censored data by

$$\tilde{m}_{n,2}(\cdot) = \tilde{m}_{n,2}(\cdot, \mathcal{D}_n) = \arg\min_{f \in \mathcal{F}_n} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \cdot |f(X_i) - Z_i|^2}{G_n(Z_i)} \tag{2.20}$$

In order to ensure that the estimates are bounded in the same way, set

$$m_{n,2}(x) = m_{n,2}(x, \mathcal{D}_n) = T_{[0,T_{K_n}]} \tilde{m}_{n,2}(x) \tag{2.21}$$

In the next theorem we use the notion $\mathcal{V}_{\mathcal{F}_n^+}$ for the VC dimension of the set of graphs of functions in $\mathcal{F}_n$.

Theorem 2.3. *(Kohler, Máthé and Pintér (2000)) Assume that the sets $\mathcal{F}_n$ of functions $f : \mathcal{R}^d \to \mathcal{R}$ satisfy*

$$\frac{\mathcal{V}_{\mathcal{F}_n^+}}{n} \to 0 \quad (n \to \infty) \tag{2.22}$$

and

$$\inf_{\substack{f \in \mathcal{F}_n, \\ 0 \le f(x) \le T_F \ (x \in \mathcal{R}^d)}} \sup_{x \in [-A,A]^d} |f(x) - g(x)| \to 0 \quad (n \to \infty) \tag{2.23}$$

for every $A \in \mathcal{R}_+$ and for every continuous function g with $0 \le g(x) \le T_F$ $(x \in \mathcal{R}^d)$ and $g(x) = 0$ for all $x \in \mathcal{R}^d \setminus [-A, A]^d$. Let $m_{n,2}$ be defined by (2.20) and (2.21). Then

$$\int_{\mathcal{R}^d} |m_{n,2}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

for all distributions of (X, Y, C) satisfying Assumption 2.1.

We now give an example for sets of functions which satisfy the conditions of Theorem 2.3.

EXAMPLE.

Let $M \in \mathbb{N}_0$, $A_n \in \mathcal{R}_+$, $H_n \in \mathbb{N}$, such that

$$\frac{H_n^d}{n} \to 0 \quad (n \to \infty), \tag{2.24}$$

$$A_n \to \infty \quad (n \to \infty) \quad \text{and} \quad \frac{A_n}{K_n} \to 0 \quad (n \to \infty). \tag{2.25}$$

Let $\mathcal{P}_n = \{B_{n,1}, B_{n,2}, \dots B_{n,H_n^d}\}$ be the partition of $[-A_n, A_n]^d$ into H_n^d equidistant cubes. Let $\mathcal{F}_n$ be the set of all (multivariate) piecewise polynomials of degree M (or less, in each coordinate) with respect to $\mathcal{P}_n$. We use the following lemma to upper bound the VC dimension of $\mathcal{F}_n$.

LEMMA 2.5. *(Steele (1975), Dudley (1978)) Let $\mathcal{F}$ be a k–dimensional vector space of functions $f : \mathcal{R}^d \to \mathcal{R}$. Then the class of sets of the form $\{x \in \mathcal{R}^d : f(x) \geq 0\}$, $f \in \mathcal{F}$, has* VC *dimension less than or equal to k.*

$\mathcal{F}_n$ is a $H_n^d \cdot (M+1)^d$-dimensional vector space, thus it follows that $V_{\mathcal{F}_n^+} \leq H_n^d \cdot (M+1)^d + 1$. Condition (2.22) is then satisfied by (2.24). In order to show (2.23) let g be a continuous function with $0 \leq g(x) \leq T_F$ $(x \in \mathcal{R}^d)$ and $g(x) = 0$ for all $x \in \mathcal{R}^d \setminus [-A_n, A_n]^d$. Choose fixed $x_i \in B_{n,i}$ $(i = 1, \dots, H_n^d)$ and define $f_g \in \mathcal{F}_n$ by

$$f_g(x) = \sum_i g(x_i) I_{B_{n,i}}(x) \quad (x \in \mathcal{R}^d).$$

For arbitrary points $u_1, u_2 \in B_{n,i}$ $(i = 1, \dots, H_n^d)$ it holds that $||u_1 - u_2|| \leq \sqrt{d} \cdot \frac{2A_n}{H_n}$. Furthermore for given x there exists a cube $B_{n,i}$ by the definition of f_g, so that we have $f_g(x) = g(x_i)$. It follows

$$\begin{aligned}
\inf_{\substack{f \in \mathcal{F}_n, \\ 0 \leq f(x) \leq T_F \ (x \in \mathcal{R}^d)}} \sup_{x \in [-A_n, A_n]^d} |f(x) - g(x)| &\leq \sup_{x \in [-A_n, A_n]^d} |f_g(x) - g(x)| \\
&\leq \sup_{u_1, u_2 \in B_{n,i}} |g(u_1) - g(u_2)| \\
&\leq \sup_{||u_1 - u_2|| \leq \sqrt{d} \cdot \frac{2A_n}{H_n}} |g(u_1) - g(u_2)| \\
&\longrightarrow 0
\end{aligned}$$

by (2.25) and by uniform continuity of g, so that condition (2.23) is satisfied.

PROOF OF THEOREM 2.3

Define

$$\mathcal{B}_n^\star \mathcal{F}_n := \{f \in \mathcal{F}_n : \ 0 \le f(x) \le T_F \ (x \in \mathcal{R}^d)\} \tag{2.26}$$

and

$$\mathcal{T}_n^\star \mathcal{F}_n := \{g : \mathcal{R}^d \to \mathcal{R}_+, \ g = T_{[0,T_F]} f \text{ for some } f \in \mathcal{F}_n\}. \tag{2.27}$$

Let $m_{n,2}^\star = T_{[0,T_F]} \tilde{m}_{n,2}(x)$. Then the statement of Lemma 2.4 is true also for $m_{n,2}$, thus, in order to prove the theorem, it suffices to show that

$$\int_{\mathcal{R}^d} |m_{n,2}^\star(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

The proof of this relation is divided into three steps.

In the first step we show

$$\begin{aligned} &\int_{\mathcal{R}^d} |m_{n,2}^\star(x) - m(x)|^2 \mu(dx) \\ &\le \ 2 \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Y_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \\ &\quad + \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \int_{\mathcal{R}^d} |f(x) - m(x)|^2 \mu(dx). \end{aligned} \tag{2.28}$$

Because of (1.2) we have the following decomposition of the L_2 risk:

$$\begin{aligned} &\int_{\mathcal{R}^d} |m_{n,2}^\star(x) - m(x)|^2 \mu(dx) \\ &= \ \left\{ \mathbf{E}\left(|m_{n,2}^\star(X) - Y|^2 \,|\mathcal{D}_n \right) - \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 \right\} \\ &\quad + \left\{ \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 \right\}. \end{aligned}$$

Again by (1.2)

$$\inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 = \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \int_{\mathcal{R}^d} |f(x) - m(x)|^2 \mu(dx).$$

Thus we just need an upper bound for the first term. It holds that

$$\mathbf{E}\left(|m_{n,2}^\star(X) - Y|^2 \,|\mathcal{D}_n \right) - \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \mathbf{E}|f(X) - Y|^2$$

$$
\begin{aligned}
= & \sup_{f\in\mathcal{B}_n^\star\mathcal{F}_n}\Bigg[\mathbf{E}\left(|m_{n,2}^\star(X)-Y|^2\,|\mathcal{D}_n\right)-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,2}^\star(X_i)-Z_i|^2}{G_n(Z_i)}\\
&+\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,2}^\star(X_i)-Z_i|^2}{G_n(Z_i)}-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|\tilde m_{n,2}(X_i)-Z_i|^2}{G_n(Z_i)}\\
&+\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|\tilde m_{n,2}(X_i)-Z_i|^2}{G_n(Z_i)}-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G_n(Z_i)}\\
&+\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G_n(Z_i)}-\mathbf{E}|f(X)-Y|^2\Bigg].
\end{aligned}
$$

Now we give upper bounds for the terms on the right-hand side of the last equality.

The third term is by (2.20) and $f\in\mathcal{B}_n^\star\mathcal{F}_n\subset\mathcal{F}_n$ less than or equal to 0.

The same is true for the second term on the right-hand side, because of (2.16) and $0\le Z_i\le T_{K,n}\le T_K=T_F$ $(i=1,\dots n)$ $a.s.$ we have for all $i\in\{1,\dots,n\}$

$$|\tilde m_{n,2}(X_i)-Z_i|=|m_{n,2}^\star(X_i)-Z_i|+|m_{n,2}^\star(X_i)-\tilde m_{n,2}(X_i)|\ge|m_{n,2}^\star(X_i)-Z_i|.$$

Since $m_{n,2}^\star\in\mathcal{T}_n^\star\mathcal{F}_n$ and $f\in\mathcal{B}_n^\star\mathcal{F}_n\subset\mathcal{T}_n^\star\mathcal{F}_n$ we can bound the remaining two terms by

$$\sup_{f\in\mathcal{T}_n^\star\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Y_i|^2}{G_n(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|,$$

which proves (2.28).

In the second step of the proof we show

$$\inf_{f\in\mathcal{B}_n^\star\mathcal{F}_n}\int_{\mathcal{R}^d}|f(x)-m(x)|^2\mu(dx)\longrightarrow 0\ (n\to\infty)\ a.s. \tag{2.29}$$

Let $\epsilon>0$ be arbitrary. Denote the set of all continuous functions with compact support by $C_0(\mathcal{R}^d)$. Since $m\in L_2(\mu)$ and $C_0(\mathcal{R}^d)$ dense in $L_2(\mu)$ for any distribution μ of X, there exists a function $h\in C_0(\mathcal{R}^d)$ with

$$\int_{\mathcal{R}^d}|h(x)-m(x)|^2\mu(dx)<\epsilon.$$

Choose $A>0$ with

$$h(x)=0\ (x\in\mathcal{R}^d\setminus[-A,A]^d)\ \text{ and }\ \mu(\mathcal{R}^d\setminus[-A,A]^d)\le\frac{\epsilon}{T_F^2}. \tag{2.30}$$

Set

$$\bar h(x):=T_{[0,T_F]}h(x)\ \ (x\in\mathcal{R}^d).$$

Then $\bar{h} \in C_0(\mathcal{R}^d)$, $0 \le \bar{h}(x) \le T_F$ $(x \in \mathcal{R}^d)$, and $0 \le m(x) \le T_F$ $(x \in \mathcal{R}^d)$ implies

$$\int_{\mathcal{R}^d} |\bar{h}(x) - m(x)|^2 \mu(dx) \quad \le \quad \int_{\mathcal{R}^d} |h(x) - m(x)|^2 \mu(dx) \le \epsilon.$$

From this, $|f(x) - m(x)|^2 \le T_F^2$ $(x \in \mathcal{R}^d)$ for all $f \in \mathcal{B}_n^\star \mathcal{F}_n$ and (2.30) we conclude

$$\begin{aligned}
&\inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \int_{\mathcal{R}^d} |f(x) - m(x)|^2 \mu(dx) \\
&\le \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \int_{[-A,A]^d} |f(x) - m(x)|^2 \mu(dx) + T_F^2 \cdot \mu\left(\mathcal{R}^d \setminus [-A,A]^d\right) \\
&\le \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} 2 \int_{[-A,A]^d} |f(x) - \bar{h}(x)|^2 \mu(dx) + 2 \int_{[-A,A]^d} |\bar{h}(x) - m(x)|^2 \mu(dx) + \epsilon \\
&\le 2 \inf_{f \in \mathcal{B}_n^\star \mathcal{F}_n} \sup_{x \in [-A,A]^d} |f(x) - \bar{h}(x)|^2 + 3\epsilon.
\end{aligned}$$

With (2.23) and $\epsilon \to 0$ we have proved (2.29).

In the third step of the proof we show

$$\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Y_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \longrightarrow 0 \ (n \to \infty) \ a.s. \tag{2.31}$$

We use the decomposition

$$\begin{aligned}
&\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right| \\
&\le \quad \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} \right| \\
&\quad + \sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|f(X) - Y|^2 \right|.
\end{aligned}$$

Using $0 \le f(x) \le T_F$ $(x \in \mathcal{R}^d)$ for all $f \in \mathcal{T}_n^\star \mathcal{F}_n$, $0 \le Z_i \le T_{K,n} \le T_K = T_F$ $(i = 1, \dots, n)$ $a.s.$ and $G(T_F) > 0$, we have

$$\begin{aligned}
&\sup_{f \in \mathcal{T}_n^\star \mathcal{F}_n} \left| \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n} \sum_{i=1}^n \frac{\delta_i |f(X_i) - Z_i|^2}{G(Z_i)} \right| \\
&= \quad \frac{1}{n} \sum_{i=1}^n \delta_i |f(X_i) - Z_i|^2 \left| \frac{1}{G(Z_i)} - \frac{1}{G_n(Z_i)} \right| \\
&\le \quad \frac{T_F^2}{G(T_F) \cdot G_n(T_F)} \sup_{t \le T_F} |G(t) - G_n(t)|,
\end{aligned}$$

which together with the uniform consistency of G_n implies

$$\sup_{f\in\mathcal{T}_n^\star\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G_n(Z_i)}-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G(Z_i)}\right|\longrightarrow 0$$

almost surely as $n\longrightarrow\infty$. It remains to prove that

$$\sup_{f\in\mathcal{T}_n^\star\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|\longrightarrow 0\quad(n\longrightarrow\infty)\quad a.s. \tag{2.32}$$

holds. For the proof of (2.32) we will use Lemma 2.2 and Lemma 2.3.
For $V:=(X,Z,\delta)$, $V_1:=(X_1,Z_1,\delta_1),\dots,V_n:=(X_n,Z_n,\delta)$ i.i.d, define

$$\begin{aligned}\mathcal{H}_n\ :=\ &\Big\{h:\mathcal{R}^d\times[0,T_F]\times\{0,1\}\longrightarrow\mathcal{R}_+:\exists f\in\mathcal{T}_n^\star\mathcal{F}_n\text{ with}\\ &\quad h(x,z,\delta):=\frac{\delta|f(x)-z|^2}{G(z)}\ \text{ for all }(x,z,\delta)\in\mathcal{R}^d\times[0,T_F]\times\{0,1\}\Big\}.\end{aligned}$$

Since $0\le Y\le T_F\le\infty$ $a.s.$ and $0\le f(x)\le T_F$ for all $x\in\mathcal{R}^d$ and all $f\in\mathcal{T}_n^\star\mathcal{F}_n$ the functions in $\mathcal{H}_n$ are nonnegative and bounded by $T_F^2/G(T_F)$. Furthermore we have

$$\mathbf{E}h(V)=\mathbf{E}\frac{\delta|f(X)-Z|^2}{G(Z)}=\mathbf{E}|f(X)-Z|^2$$

and therefore

$$\sup_{f\in\mathcal{T}_n^\star\mathcal{F}_n}\left|\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|f(X_i)-Z_i|^2}{G(Z_i)}-\mathbf{E}|f(X)-Y|^2\right|=\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n h(V_i)-\mathbf{E}h(V)\right|.$$

If $h_1,h_2\in\mathcal{H}_n$ and f_1,f_2 are the corresponding functions in $\mathcal{T}_n^\star\mathcal{F}_n$, then

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n|h_1(V_i)-h_2(V_i)|\\ &\quad=\ \frac{1}{n}\sum_{i=1}^n\left|\frac{\delta_i|f_1(X_i)-Z_i|^2}{G(Z_i)}-\frac{\delta_i|f_2(X_i)-Z_i|^2}{G(Z_i)}\right|\\ &\quad\le\ \frac{1}{G(T_F)}\frac{1}{n}\sum_{i=1}^n|(f_1(X_i)+f_2(X_i)-2Z_i)(f_1(X_i)-f_2(X_i))|\\ &\quad\le\ \frac{2T_F}{G(T_F)}\frac{1}{n}\sum_{i=1}^n|f_1(X_i)-f_2(X_i)|,\end{aligned}$$

which implies that

$$N_1(\epsilon,V_1^n,\mathcal{H}_n)\le N_1\left(\frac{\epsilon\cdot G(T_F)}{2T_F},X_1^n,\mathcal{T}_n^\star\mathcal{F}_n\right). \tag{2.33}$$

From this, Lemma 2.2 and Lemma 2.3 we get for arbitrary $t > 0$ that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^{n} h(V_i) - \mathbf{E}h(V)\right| > t\right\} \\
&\leq \ 8\mathbf{E}\left\{N_1\left(\frac{t\cdot G(T_F)}{16T_F}, X_1^n, \mathcal{T}_n^\star\mathcal{F}_n\right)\right\}\cdot\exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right) \\
&\leq \ 8\left(\frac{64eT_F^3}{t\cdot G(T_F)}\right)^{2\mathcal{V}_{\mathcal{T}_n^\star\mathcal{F}_n^+}}\cdot\exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right).
\end{aligned}$$

Because of $\mathcal{V}_{\mathcal{T}_n^\star\mathcal{F}_n^+} \leq \mathcal{V}_{\mathcal{F}_n^+}$ and of condition (2.22) of the theorem the above probability is summable. Hence by the Borel-Cantelli lemma we get (2.31).

The assertion follows from steps 1 to 3. □

2.3.3 Penalized least squares estimates

We define penalized least squares estimates for censored data by

$$\tilde{m}_{n,3}(\cdot) = \tilde{m}_{n,3}(\cdot,\mathcal{D}_n) = \arg\min_{f\in W^k(\mathcal{R}^d)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\cdot|f(X_i)-Z_i|^2}{G_n(Z_i)} + \lambda_n J_k^2(f)\right), \tag{2.34}$$

where $J_k^2(f)$ is as defined in (1.8). In order to ensure that the estimates are bounded in the same way, set

$$m_{n,3}(x) = m_{n,3}(x,\mathcal{D}_n) = T_{[0,T_{K_n}]}\tilde{m}_{n,3}(x) \tag{2.35}$$

Theorem 2.4. *(Kohler, Máthé and Pintér (2000)) Let $k \in \mathbb{N}$ with $2k > d$, and for $n \in \mathbb{N}$ choose $\lambda_n > 0$ such that*

$$\lambda_n \to 0 \quad (n\to\infty) \tag{2.36}$$

and

$$n\cdot\lambda_n \to \infty \quad (n\to\infty). \tag{2.37}$$

Let $m_{n,3}$ be defined by (2.34) and (2.35). Then

$$\int_{\mathcal{R}^d} |m_{n,3}(x) - m(x)|^2\mu(dx) \to 0 \quad a.s. \quad (n\to\infty)$$

for all distributions of (X,Y,C) satisfying Assumption 2.1, and $|X|$ bounded a.s.

REMARK.

The assumption $|X|$ bounded *a.s.* may be dropped if one modifies the estimate as described in Remark 3 in Kohler and Krzyżak (2001).

Let $m_{n,3}^{\star} = T_{[0,T_F]}\tilde{m}_{n,3}(x)$. Then the statement of Lemma 2.4 is true also for $m_{n,3}$, thus, in order to prove the theorem, it suffices to show that

$$\int_{\mathcal{R}^d} |m_{n,3}^{\star}(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

We will need the following lemma.

LEMMA 2.6. *Assume that $2k > d$ and that conditions (2.36) and (2.37) hold. Then*

$$\mathbf{E}\left(|m_{n,3}^{\star}(X) - Y|^2 \,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i |m_{n,3}^{\star}(X_i) - Z_i|^2}{G(Z_i)} \to 0$$

a.s. as $n \to \infty$ for all distributions of (X, Y, C) satisfying Assumption 2.1. with $X \in [0,1]^d$ a.s.

PROOF. By the definition of the estimate

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \cdot |\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} + \lambda_n J_k^2(\tilde{m}_{n,3}) \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \cdot |0 - Z_i|^2}{G_n(Z_i)} + \lambda_n \cdot 0 \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \cdot Z_i^2}{G(Z_i)} + \frac{1}{n}\sum_{i=1}^{n} \left|\frac{\delta_i \cdot Z_i^2}{G_n(Z_i)} - \frac{\delta_i \cdot Z_i^2}{G(Z_i)}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \cdot Z_i^2}{G(Z_i)} + \frac{T_F^2}{G(T_F)G_n(T_F)} \sup_{t \leq T_F} |G(t) - G_n(t)| \\
&\to \mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\} \quad a.s.
\end{aligned}$$

as $n \to \infty$ because of the strong law of large numbers and the uniform consistency of G_n. This together with (2.37) implies that for sufficiently large n

$$J_k^2(\tilde{m}_{n,3}) \leq \frac{2}{\lambda_n}\mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\} = n \cdot \frac{2\mathbf{E}\left\{\frac{\delta \cdot Z^2}{G(Z)}\right\}}{n \cdot \lambda_n} \leq n,$$

hence for n sufficiently large

$$\tilde{m}_{n,3} \in \mathcal{F}_n := \left\{ f : f \in W^k(\mathcal{R}^d) \text{ and } J_k^2(f) \leq n \right\},$$

thus

$$m^\star_{n,3} \in \mathcal{T}_n^\star \mathcal{F}_n.$$

Similarly as in the proof of Theorem 2.3, let

$$\mathcal{H}_n := \left\{ \begin{array}{l} h : \mathcal{R}^d \times [0, T_F] \times \{0,1\} \longrightarrow \mathcal{R}^+ : \\ \qquad h(x,z,\delta) := \dfrac{\delta |f(x) - z|^2}{G(z)} \text{ for some } f \in \mathcal{T}_n^\star \mathcal{F}_n \end{array} \right\}.$$

Hence using the notations $V := (X, Z, \delta)$ and $V_i := (X_i, Z_i, \delta_i)$ $(i = 1, \ldots, n)$, and the fact that

$$\mathbf{E}h(V) = \mathbf{E}\left\{ \frac{\delta |f(X) - Z|^2}{G(Z)} \right\} = \mathbf{E}|f(X) - Z|^2,$$

it suffices to show that

$$\sup_{h \in \mathcal{H}_n} \left| \frac{1}{n} \sum_{i=1}^n h(V_i) - \mathbf{E}h(V) \right| \to 0 \quad (n \to \infty) \quad a.s.$$

From (2.33) we know that

$$N_1(\epsilon, V_1^n, \mathcal{H}_n) \leq N_1\left(\frac{\epsilon \cdot G(T_F)}{2T_F}, X_1^n, \mathcal{T}_n^\star \mathcal{F}_n \right). \tag{2.38}$$

The following lemma gives an upper bound for the covering number.

LEMMA 2.7. *(Kohler, Krzyżak (2001)) Let $L, c > 0$ and set*

$$\mathcal{F} = \left\{ T_{[0,L]} f : f \in W^k(\mathcal{R}^d) \text{ and } J_k^2(f) \leq c \right\},$$

where the definition of $J_k^2(f)$ is the same as in (1.8). Then for any $\epsilon > 0$ and $x_1, ..., x_n \in [0,1]^d$

$$N_1(\epsilon, x_1^n, \mathcal{F}) \leq \left(c_1 \frac{Ln}{\epsilon} \right)^{c_2 \left(\frac{\sqrt{c}}{\epsilon} \right)^{\frac{d}{k}} + c_3},$$

where $c_1, c_2, c_3 \in \mathcal{R}_+$ are constants which only depend on k and d.

Thus from this, (2.38) and Lemma 2.2 we get for any $t > 0$ that

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{h\in\mathcal{H}_n}\left|\frac{1}{n}\sum_{i=1}^n h(V_i) - \mathbf{E}h(V)\right| > t\right\} \\
&\leq \; 8\mathbf{E}\left\{N_1\left(\frac{t\cdot G(T_F)}{16T_F}, X_1^n, \mathcal{T}_n^\star\mathcal{F}_n\right)\right\}\cdot\exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right) \\
&\leq \; 8\left(c_1\frac{T_F\cdot n}{\frac{t\cdot G(T_F)}{16T_F}}\right)^{c_2\left(\frac{\sqrt{n}\cdot 16T_F}{t\cdot G(T_F)}\right)^{\frac{d}{k}}+c_3} \exp\left(-\frac{nt^2G(T_F)^2}{128T_F^4}\right),
\end{aligned}$$

which is summable since $2k > d$. Hence by the Borel-Cantelli lemma we get the assertion. $\square$

PROOF OF THEOREM 2.4. We may assume w.l.o.g. that $X \in [0,1]^d$ a.s. Let $\epsilon > 0$ be arbitrary, and choose a bounded infinitely often differentiable function $g_\epsilon : \mathcal{R}^d \to \mathcal{R}$ such that $\int_{\mathcal{R}^d} |m(x) - g_\epsilon(x)|^2\mu(dx) < \epsilon$ and $J_k^2(g_\epsilon) < \infty$. We use the following error decomposition

$$\begin{aligned}
&\int_{\mathcal{R}^d} |m_{n,3}^\star(x) - m(x)|^2\mu(dx) \\
&= \; \mathbf{E}\left(|m_{n,3}^\star(X) - Y|^2\,|\mathcal{D}_n\right) - \mathbf{E}|m(X) - Y|^2 \\
&= \; \mathbf{E}\left(|m_{n,3}^\star(X) - Y|^2\,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,3}^\star(X_i) - Z_i|^2}{G(Z_i)} \\
&\quad +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,3}^\star(X_i) - Z_i|^2}{G(Z_i)} - \frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,3}^\star(X_i) - Z_i|^2}{G_n(Z_i)} \\
&\quad +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|m_{n,3}^\star(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n\frac{\delta_i|\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} \\
&\quad +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n\frac{\delta_i|g_\epsilon(X_i) - Z_i|^2}{G_n(Z_i)} \\
&\quad +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|g_\epsilon(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n\frac{\delta_i|g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} \\
&\quad +\frac{1}{n}\sum_{i=1}^n\frac{\delta_i|g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|g_\epsilon(X) - Y|^2 \\
&\quad +\mathbf{E}|g_\epsilon(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2.
\end{aligned}$$

By Lemma 2.6

$$\mathbf{E}\left(|m^\star_{n,3}(X) - Y|^2\,|\mathcal{D}_n\right) - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |m^\star_{n,3}(X_i) - Z_i|^2}{G(Z_i)} \to 0 \quad (n\to\infty) \quad a.s.$$

Since $0 \le m^\star_{n,3}(X_i) \le T_{K,n}$ and $0 \le Z_i \le T_{K,n} \le T_K = T_F$ $(i = 1,\dots,n)$ $a.s.$, $G(T_F) > 0$ one can conclude from the uniform consistency of G_n

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\delta_i |m^\star_{n,3}(X_i) - Z_i|^2}{G(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |m^\star_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} \\
\le\ & \frac{1}{n}\sum_{i=1}^n \delta_i |m^\star_{n,3}(X_i) - Z_i|^2 \left|\frac{1}{G(Z_i)} - \frac{1}{G_n(Z_i)}\right| \\
\le\ & \frac{T_F^2}{G(T_F)G_n(T_F)} \sup_{t\le T_F} |G(t) - G_n(t)| \\
\to\ & 0 \quad a.s.
\end{aligned}$$

If $a, b \in \mathcal{R}$ with $0 \le b \le T_F$ and $c = T_{[0,T_F]}a$ then $|c - b| \le |a - b|$, which implies that

$$\frac{1}{n}\sum_{i=1}^n \frac{\delta_i |m^\star_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} \le 0.$$

It follows from the definition of the estimate and (2.36) that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\delta_i |\tilde{m}_{n,3}(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G_n(Z_i)} \\
\le\ & \lambda_n J_k^2(g_\epsilon) - \lambda_n J_k^2(\tilde{m}_{n,3}) \\
\le\ & \lambda_n J_k^2(g_\epsilon) \to 0 \quad (n\to\infty).
\end{aligned}$$

Since $G(T_F) > 0$ one can conclude from the uniform consistency of G_n

$$\frac{1}{G(T_F)G_n(T_F)} \sup_{t\le T_F} |G(t) - G_n(t)| \to 0 \quad (n\to\infty),$$

thus

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G_n(Z_i)} - \frac{1}{n}\sum_{i=1}^n \frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} \\
\le\ & \frac{1}{G(T_F)G_n(T_F)} \sup_{t\le T_F} |G(t) - G_n(t)| \, \frac{1}{n}\sum_{i=1}^n \delta_i |g_\epsilon(X_i) - Z_i|^2 \\
\le\ & \frac{1}{G(T_F)G_n(T_F)} \sup_{t\le T_F} |G(t) - G_n(t)| \cdot (2T_F^2 + 2\sup_{x\in\mathcal{R}^d} |g_\epsilon(x)|^2) \to 0
\end{aligned}$$

almost surely as $n \to \infty$.

By the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i |g_\epsilon(X_i) - Z_i|^2}{G(Z_i)} - \mathbf{E}|g_\epsilon(X) - Y|^2 \to 0 \quad (n \to \infty) \quad a.s.$$

Finally,

$$\mathbf{E}|g_\epsilon(X) - Y|^2 - \mathbf{E}|m(X) - Y|^2 = \int_{\mathcal{R}^d} |m(x) - g_\epsilon(x)|^2 \mu(dx) < \epsilon.$$

With $\epsilon \to 0$ the result follows. □

2.4 Survival analysis and regression estimation for Model B

Under the assumption that Y and C are conditionally independent given X, regression estimation was considered by many authors. The first fully nonparametric approach was given by Beran (1981) who introduced a class of nonparametric regression estimates to estimate conditional survival functions (regression functions) in the presence of right censoring. Dabrowska (1987), (1989) proved some consistency results for this estimate. An alternative approach was taken by Horváth (1981) who proposed to estimate the conditional survival function by integrating an estimate of the conditional density.

Let

$$F(t|x) = P[Y > t|X = x]$$

$$G(t|x) = P[C > t|X = x]$$

$$K(t|x) = P[Z > t|X = x] = F(t|x)G(t|x)$$

and

$$T_F(x) = \sup\{t : F(t|x) > 0\},$$

$$T_G(x) = \sup\{t : G(t|x) > 0\},$$

$$T_K(x) = \min\{T_F(x), T_G(x)\},$$

.

We observe

$$(X_1, Z_1, \delta_1), \ldots, (X_n, Z_n, \delta_n)$$

where $\delta_i = I_{\{Y_i \leq C_i\}}$, and $Z_i = \min(Y_i, C_i)$. Two related problems are studied:
(i) estimating the conditional survival functions $F(t|x)$ and $G(t|x)$
(ii) estimating the regression function $m(x) = \mathbf{E}(Y|X = x)$
from the censored data.

2.4.1 Product-limit estimator

Define the local averaging estimates of $F(t|x)$ and $G(t|x)$ in the following way:

$$F_n(t|x) = \begin{cases} \prod_{\substack{i:Z_{(i)} \leq t \\ \sum_{j=i}^n W_{n(j)}(x) > 0}} \left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)} \right)^{\delta_{(i)}} & \text{if } t \leq Z_{(n)} \\ 0 & \text{otherwise} \end{cases} \tag{2.39}$$

and

$$G_n(t|x) = \begin{cases} \prod_{\substack{i:Z_{(i)} \leq t \\ \sum_{j=i}^n W_{n(j)}(x) > 0}} \left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)} \right)^{1-\delta_{(i)}} & \text{if } t \leq Z_{(n)} \\ 0 & \text{otherwise,} \end{cases} \tag{2.40}$$

where $(Z_{(i)}, \delta_{(i)}), i = 1, \ldots, n$ are the n pairs of observed (Z_i, δ_i) ordered on the $Z_{(i)}$, i.e.

$$Z_{(1)} \leq Z_{(2)} \leq \ldots \leq Z_{(n)},$$

and W_{ni} are subprobability weights depending on $X_1, \ldots, X_n$, i.e. $W_{ni}(x) \geq 0$ and $\sum_{j=1}^n W_{ni}(x) \leq 1$, and $W_{n(i)}$ are these weights ordered on the $Z_{(i)}$.

Examples of local averaging type estimates are the k_n*–nearest neighbor estimates* with weights defined in (1.3), the *kernel estimates* with weights defined in (1.4) and the *partitioning estimates* with weights defined in (1.5).

The estimates defined in (2.39) and (2.40) extend the Kaplan-Meier method to conditional survival functions. If $W_{ni}(x) = \frac{1}{n}$ for all i then

$$F_n(t|x) = \prod_{i:Z_{(i)} \leq t} \left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)} \right)^{\delta_{(i)}} = \prod_{i:Z_{(i)} \leq t} \left(\frac{n-i}{n-i+1} \right)^{\delta_{(i)}}$$

for all x, which is exactly the Kaplan-Meier estimator.

Now we will show, that these estimates are consistent if the weights are the consistent k-NN, kernel or partitioning weights and $F(t|x)$ and $G(t|x)$ have no common jumps. Beran (1981) proved uniform consistency of these estimates, here we present a much simpler and shorter proof.

THEOREM 2.5. *(Beran (1981)) If $F(\cdot|x)$ and $G(\cdot|x)$ don't have common jumps and the $W_{ni}(x)$ weights are such that the local averaging regression estimate using these weights is consistent, then*

(i) $\sup_{t<T_K(x)} |F_n(t|x) - F(t|x)| \longrightarrow 0$

(ii) $\sup_{t<T_K(x)} |G_n(t|x) - G(t|x)| \longrightarrow 0$

a.s. as $n \to \infty$ for all x.

PROOF. The argument follows the reasoning of the proof of Theorem 2.1, here we prove the statement for F. Again, it is enough to show pointwise convergence.

(i): $F(\cdot|x)$ can be expressed as

$$F(t|x) = \exp\left(\int_{[0,t]\setminus D(F(\cdot|x))} \frac{dF(s|x)}{F(s|x)}\right) \cdot \prod_{\substack{s\in D(F(\cdot|x))\\ s\le t}} \frac{F(s|x)}{F(s^-|x)} \tag{2.41}$$

and

$$F_n(t|x) = \prod_{\substack{s\in D(F_n(\cdot|x))\\ s\le t}} \frac{F_n(s|x)}{F_n(s^-|x)},$$

where $D(S)$ is, as earlier, the set of jump points of function S.

Define the conditional subsurvival functions

$$F^*(t|x) = \mathbf{P}(Z > t, \delta = 1|X = x),$$

$$G^*(t|x) = \mathbf{P}(Z > t, \delta = 0|X = x).$$

$F^*(\cdot|x)$ can be expressed in terms of the survival functions $F(\cdot|x)$ and $G(\cdot|x)$

$$F^*(t|x) = \int_t^\infty G(s|x)[-dF(s|x)], \tag{2.42}$$

because the random variables Y and C are conditionally independent given X. The integral is well-defined since by assumption the integrator and the integrand have no jumps in common.

Note that

$$\begin{aligned} F^*(t|x) + G^*(t|x) &= \mathbf{P}(Z > t|X = x) = K(t|x) \\ &= \mathbf{P}(\min(Y, C) > t|X = x) = F(t|x) \cdot G(t|x) \quad (2.43) \end{aligned}$$

For real valued functions H_1 and H_2 define a function ϕ by

$$\phi(H_1, H_2, t) = \exp\left(\int_{[0,t]\setminus D(H_1)} \frac{dH_1(s)}{H_1(s) + H_2(s)}\right) \cdot \prod_{\substack{s\in D(H_1)\\ s\le t}} \frac{H_1(s) + H_2(s)}{H_1(s^-) + H_2(s^-)}.$$

We will show that

$$F(t|x) = \phi(F^*(\cdot|x), G^*(\cdot|x), t). \tag{2.44}$$

Clearly, F^* has the same jump points as F. Thus (2.44) follows from (2.41), since

$$\begin{aligned} \frac{dF^*(s|x)}{F^*(s|x) + G^*(s|x)} &= \frac{dF^*(s|x)}{F(s|x) \cdot G(s|x)} && \text{(from (2.43))} \\ &= \frac{-G(s|x) \cdot dF(s|x)}{G(s|x) \cdot F(s|x)} && \text{(from (2.42))} \\ &= \frac{-dF(s|x)}{F(s|x)} \end{aligned}$$

and

$$\begin{aligned} \frac{F^*(s|x) + G^*(s|x)}{F^*(s^-|x) + G^*(s^-|x)} &= \frac{F(s|x) \cdot G(s|x)}{F(s^-|x) \cdot G(s^-|x)} && \text{(from (2.43))} \\ &= \frac{F(s|x)}{F(s^-|x)} \end{aligned}$$

if s is a jump point of $F(\cdot|x)$, because by assumption $F(\cdot|x)$ and $G(\cdot|x)$ don't have common jumps, thus s cannot be a jump point of $G(\cdot|x)$, so $G(s|x) = G(s^-|x)$.

Now we express $F_n(\cdot|x)$ in terms of the local averaging estimates

$$F_n^*(t|x) = \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i > t, \delta_i = 1\}}$$

and

$$G_n^*(t|x) = \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i>t,\delta_i=0\}}$$

of the conditional subsurvival functions.

First note that (2.43) is true also for the estimates, since

$$\begin{aligned} F_n^*(t|x) + G_n^*(t|x) &= \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i>t,\delta_i=1\}} + \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i>t,\delta_i=0\}} \\ &= \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i>t\}} \end{aligned}$$

and from definition

$$\begin{aligned} F_n(t|x) \cdot G_n(t|x) &= \prod_{\substack{Z_{(i)}\le t \\ \sum_{j=i}^n W_{n(j)}(x)>0}} \left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)} \right) \\ &= \frac{\sum_{i=1}^n W_{ni}(x) I_{\{Z_i>t\}}}{\sum_{j=1}^n W_{nj}(x)} \\ &= \sum_{i=1}^{n} W_{ni}(x) I_{\{Z_i>t\}}. \end{aligned}$$

Therefore

$$\frac{F_n^*(s|x) + G_n^*(s|x)}{F_n^*(s^-|x) + G_n^*(s^-|x)} = \frac{F_n(s|x) \cdot G_n(s|x)}{F_n(s^-|x) \cdot G_n(s^-|x)} = \frac{F_n(s|x)}{F_n(s^-|x)}$$

if s is a jump point of $F_n(\cdot|x)$, because by assumption $F_n(\cdot|x)$ and $G_n(\cdot|x)$ don't have common jumps, thus $G_n(s|x) = G_n(s^-|x)$. Hence

$$F_n(t|x) = \phi(F_n^*(\cdot|x), G_n^*(\cdot|x), t)$$

for $t \le Z_{(n)}$.

Since $\phi(H_1(\cdot), H_2(\cdot), t)$ is a continuous function of the argument $H_1(\cdot)$ and $H_2(\cdot)$, $F_n(\cdot|x)$ converges to $F(\cdot|x)$ almost surely if the empirical subsurvival functions $F_n^*(\cdot|x)$ and $G_n^*(\cdot|x)$ converge to $F^*(\cdot|x)$ and $G^*(\cdot|x)$, respectively. Since $F^*(\cdot|x)$ is the regression function of an indicator random variable, and similarly $G^*(\cdot|x)$, $F_n(\cdot|x)$ and $G_n(\cdot|x)$ are regression estimates and they are consistent because of the condition on the weights W_{ni}.

(ii) can be proved similarly.

□

Thus using the consistency results for the nearest neighbor (Devroye et al. (1994)), the kernel (Devroye and Krzyżak (1989)) and the partitioning (Devroye and Györfi (1983)) regression estimates, we get the following results.

COROLLARY 2.4. *If $k_n \to \infty$ and $k_n/n \to 0$, and for which for each $x \in \mathcal{R}^d$ the random variable $\|X - x\|$ is absolutely continuous, then for the nearest neighbor estimates defined by (2.39), (2.40) and (1.3),*

(i) $\sup_{t<T_K(x)} |F_n(t|x) - F(t|x)| \longrightarrow 0$

(ii) $\sup_{t<T_K(x)} |G_n(t|x) - G(t|x)| \longrightarrow 0$

a.s. as $n \to \infty$ for all x.

COROLLARY 2.5. *Let K be a regular kernel, i.e. it is nonnegative, and there is a ball $S_{0,r}$ of radius $r > 0$ centered at the origin, and constant $b > 0$ such that*

$$K(x) \geq bI_{\{x \in S_{0,r}\}}$$

and

$$\int_{\mathcal{R}^d} \sup_{u \in x + S_{0,r}} K(u)dx < \infty.$$

If $h_n \to 0$ and $nh_n^d \to \infty$, then for the kernel regression estimate defined by (2.39), (2.40) and (1.4),

(i) $\sup_{t<T_K(x)} |F_n(t|x) - F(t|x)| \longrightarrow 0$

(ii) $\sup_{t<T_K(x)} |G_n(t|x) - G(t|x)| \longrightarrow 0$

a.s. as $n \to \infty$ for all x.

COROLLARY 2.6. *Let $\mathcal{P}_n = \{A_{nj} : j\}$ be partitions of $\mathcal{R}^d$ such that for each sphere S centered at the origin*

$$\lim_{n\to\infty} \sup_{j: A_{nj} \cap S \neq 0} diam(A_{nj}) = 0$$

and

$$\lim_{n\to\infty} \frac{|\{j : A_{nj} \cap S \neq 0\}|}{n} = 0.$$

Then for the partitioning estimate defined by (2.39), (2.40) and (1.5),

(i) $\sup_{t<T_K(x)} |F_n(t|x) - F(t|x)| \longrightarrow 0$

(ii) $\sup_{t<T_K(x)} |G_n(t|x) - G(t|x)| \longrightarrow 0$

a.s. as $n \to \infty$ *for all* x.

2.4.2 Regression estimation

From a consistent estimator $F_n(\cdot|x)$ of the conditional survival function $F(\cdot|x)$ we can obtain a consistent nonparametric estimator $m_n(x)$ of the regression function $m(x)$. In addition to the assumption of Model B, i.e. that Y and C are conditionally independent given X, we need:

Assumption 2.2. *$T_F(x) < T_G(x)$ for all x.*

Let

$$m_n(x) = \int_0^\infty F_n(t|x)dt = \int_0^{T_F(x)} F_n(t|x)dt. \tag{2.45}$$

Then

$$\begin{aligned}\int |m(x) - m_n(x)|\mu(dx) &= \int \left| \int_0^\infty (F(t|x) - F_n(t|x))dt \right| \mu(dx) \\ &\leq \int \int_0^{T_F(x)} |F(t|x) - F_n(t|x)| dt \mu(dx),\end{aligned}$$

which, under Assumption 2.2, tends to zero a.s. by the dominated convergence theorem.

Since Y is bounded by T_F, we have $m(x) \leq T_F$ and hence also

$$|m(x) - m_n(x)| \leq T_F.$$

Thus

$$\int |m(x) - m_n(x)|^2 \mu(dx) \leq T_F \int |m(x) - m_n(x)|\mu(dx),$$

so consistency in L_2-error also follows.

To see that the integration in (2.45) can be carried out easily, we calculate the jump of the estimator $F_n(\cdot|x)$ defined by (2.39) in Z_i.

Lemma 2.8. *The jump of the estimator $F_n(\cdot|x)$ in an observation Z_i is given by*

$$dF_n(Z_i|x) = \frac{W_{ni}(x)\delta_i}{G_n(Z_i|x)},$$

where $G_n(\cdot|x)$ is the estimator of $G(\cdot|x)$ defined by (2.40).

Proof.

Since we know that the jump of $F_n(\cdot|x)$ in a censored observation is zero, we consider only the jump in an uncensored observation, i.e. let $\delta_{(i)} = 1$. Then by definition the difference $dF_n(Z_{(i)}|x) = F_n(Z_{(i-1)}|x) - F_n(Z_{(i)}|x)$ is

$$\begin{aligned}
&\prod_{\substack{k\le i-1\\ \sum_{j=k}^n W_{n(j)}(x)>0}} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} - \prod_{\substack{l:Z_{(l)}\le Z_{(i)}\\ \sum_{j=l}^n W_{n(j)}(x)>0}} \left(\frac{\sum_{j=l+1}^n W_{n(j)}(x)}{\sum_{j=l}^n W_{n(j)}(x)}\right)^{\delta_{(l)}}\\
&= \prod_{\substack{k\le i-1\\ \sum_{j=k}^n W_{n(j)}(x)>0}} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} \left(1-\left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)}\right)^{\delta_{(i)}}\right)\\
&= \prod_{\substack{k\le i-1\\ \sum_{j=k}^n W_{n(j)}(x)>0}} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} \frac{W_{n(i)}(x)}{\sum_{j=i}^n W_{n(j)}(x)}
\end{aligned}$$

If $W_{n(i)}(x) = 0$, then $dF_n(Z_{(i)}|x) = 0$, otherwise $\sum_{j=k}^n W_{n(j)}(x) > 0$ for all $k \le i$, thus

$$\begin{aligned}
&dF_n(Z_{(i)}|x)\\
&= \prod_{k\le i-1} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} \frac{W_{n(i)}(x)}{\sum_{j=i}^n W_{n(j)}(x)}\\
&= \prod_{k\le i} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} \left(\frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=i}^n W_{n(j)}(x)}\right)^{-1} \frac{W_{n(i)}(x)}{\sum_{j=i}^n W_{n(j)}(x)}\\
&\qquad \text{(since } \delta_{(i)} = 1\text{)}\\
&= \prod_{k\le i} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}} W_{n(i)}(x) \frac{1}{\sum_{j=i+1}^n W_{n(j)}(x)}.
\end{aligned}$$

Since $W_{n(j)}(x)$ are probability weights

$$\sum_{j=i+1}^n W_{n(j)}(x) = \frac{\sum_{j=i+1}^n W_{n(j)}(x)}{\sum_{j=1}^n W_{n(j)}(x)} = \prod_{k\le i} \left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right).$$

Thus by the definition of $G_n(\cdot|x)$ and since $\delta_{(i)} = 1$

$$\begin{aligned} dF_n(Z_{(i)}|x) &= \prod_{k\leq i}\left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{\delta_{(k)}-1} W_{n(i)}(x) \\ &= \frac{1}{\prod_{k\leq i}\left(\frac{\sum_{j=k+1}^n W_{n(j)}(x)}{\sum_{j=k}^n W_{n(j)}(x)}\right)^{1-\delta_{(k)}}} W_{n(i)}(x) \\ &= \frac{W_{n(i)}(x)\delta_{(i)}}{G_n(Z_{(i)}|x)} \end{aligned}$$

□

The integral of $F_n(t|x)$ can be expressed as the sum of Z_i times the jump of F_n in Z_i, what is apperent once we calculate the area in question by decomposing it into horizontal stripes, rather than into vertical ones, as it is customary in the calculation of integrals. Thus

$$\int_0^\infty F_n(t|x)dt = \sum_{i=1}^n Z_i \frac{W_{ni}(x)\delta_i}{G_n(Z_i|x)}.$$

Hence the regression estimate defined by (2.45) can be written in the following form:

$$m_n(x) = \sum_{i=1}^n W_{ni}(x)\frac{Z_i\delta_i}{G_n(Z_i|x)}. \tag{2.46}$$

COROLLARY 2.7. *If $k_n \to \infty$, $k_n/n \to 0$, Assumption 2.2 holds and for each $x \in \mathcal{R}^d$ the random variable $\|X - x\|$ is absolutely continuous, then for the nearest neighbor estimate defined by (2.46) and (1.3),*

$$\int_{\mathcal{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

COROLLARY 2.8. *Let K be a regular kernel, i.e. it is nonnegative, and there is a ball $S_{0,r}$ of radius $r > 0$ centered at the origin, and constant $b > 0$ such that*

$$K(x) \geq bI_{\{x\in S_{0,r}\}}$$

and

$$\int_{\mathcal{R}^d} \sup_{u\in x+S_{0,r}} K(u)dx < \infty.$$

If $h_n \to 0$, $nh_n^d \to \infty$ and Assumption 2.2 holds, then for the kernel regression estimate defined by (2.46) and (1.4),

$$\int_{\mathcal{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

COROLLARY 2.9. *Let $\mathcal{P}_n = \{A_{n,j} : j\}$ be partitions of $\mathcal{R}^d$ such that for each sphere S centered at the origin*

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

and

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

If Assumption 2.2 holds, then for the partitioning estimate defined by (2.46) and (1.5),

$$\int_{\mathcal{R}^d} |m_n(x) - m(x)|^2 \mu(dx) \to 0 \quad (n \to \infty) \quad a.s.$$

Chapter 3

Error Estimates for the Partitioning Classification Rule

Many important classification rules partition $\mathcal{R}^d$ into disjoint cells and classify in each cell according to the majority vote among the labels of the sample points X_i falling in the same cell. Let $\mathcal{P}_n = \{A_{nj}, j = 1, 2, \ldots\}$ be a partition of $\mathcal{R}^d$ and let $A_n(x)$ denote the cell in the partition that includes x. Then the partitioning classification rule:

$$g_n(x) = \begin{cases} 0, & \text{if } \sum_{i=1}^n I_{\{Y_i=1\}} I_{\{X_i \in A_n(x)\}} \\ & \quad \leq \sum_{i=1}^n I_{\{Y_i=0\}} I_{\{X_i \in A_n(x)\}} \\ 1, & \text{otherwise.} \end{cases}$$

For cubic partitions consistency was established under some additional conditions by Glick (1973). Universal consistency follows from the result of Gordon and Olshen (1978), (1980). The usual conditions needed for the consistency are that the diameter of the cells tends to zero and the number of cells in a bounded region is small compared to the sample size. Strong universal consistency of the partitioning classification rule was proved by Devroye and Györfi (1983). We will show in Section 3.1 that, although universal rate of convergence doesn't exist, the error probability of the partitioning classification rule converges to the Bayes error faster than $1/\sqrt{n}$ under certain conditions. In Section 3.2 we investigate two error estimates, the resubstitution and the deleted estimate (or cross validation) and prove asymptotic normality of the estimates.

We will need in all results the assumption, that the labels in a cell are not uniformly chosen, more precisely we will need

Assumption 3.1. *There is a constant $c > 0$ such that*

$$|\tau(A_{nj})| := |\mu(A_{nj}) - 2\nu(A_{nj})| > c\mu(A_{nj})$$

for all n and j, where $\nu(A) = \mathbf{E}\left\{YI_{\{X\in A\}}\right\}$.

The necessity of such an assumption is demonstrated in Proposition 3.2, where we show that if the label is uniform in every cell then the difference between the resubstitution estimate and the true error probability is too big and asymptotic normality doesn't hold.

Partitioning estimates, which were introduced by Tukey (1947) for regression estimation, are present also in density estimation, their behaviour has been studied in several papers, and they are known to be consistent. Asymptotic normality of the L_1- and L_2-errors in partitioning density estimation (histogram) was proved by Beirlant, Györfi and Lugosi (1994) and Berlinet, Devroye and Györfi (1995). Beirlant and Mason (1995) developed a general method for deriving asymptotic normality of the L_p-errors of empirical functionals, which make use of the neighboring data at any point x of interest. Beirlant and Györfi (1998) continued this line of research, and established the asymptotic normality of L_2-error in partitioning regression estimation. To obtain asymptotic normality all these authors used the idea of Poissonization in conjunction with Bartlett's (1938) method of partial inversion for obtaining characteristic functions of conditional distributions.

To lighten notations, throughout the chapter we assume that X takes its value in a bounded region $\mathcal{X} \subset \mathcal{R}^d$.

3.1 Rate of Convergence

It is wellknown that no universal rate of convergence can be guaranteed for any classifiers if nothing is assumed about the distribution. There is no sequence of numbers a_n converging to zero such that there is a classification rule with error probability below $L^* + a_n$ for all distributions (Devroye (1982), Sec. 7.2 in Devroye, Györfi and Lugosi (1996)). In this section we show that the error probability of the partitioning classification rule converges to the Bayes error faster than $1/\sqrt{n}$ under certain conditions.

THEOREM 3.1. *(Pintér (2000)) Let R_n^* denote the Bayes error restricted to the partition $\mathcal{P}_n$, i.e.*

$$R_n^* = \sum_{j=1}^{m_n} \min\{\nu(A_{nj}), \mu(A_{nj}) - \nu(A_{nj})\},$$

Then

(i) $L_n \geq R_n^*$.

(ii) Assume that the partitions are finite, let m_n denote the number of cells in the partition. If Assumption 3.1 is satisfied, then

$$\mathbf{E}(L_n - R_n^*) \leq \frac{12}{ce}\frac{m_n}{n}.$$

PROOF

(i): Clearly $L_n \geq R_n^*$, because R_n^* is the Bayes error corresponding to the pair of random variables $(T_n(X), Y)$, where $T_n(x) = j$ if $x \in A_{nj}$, and since the partitioning classification rule $g_n(x)$ can be written as a function of $T_n(x)$, its error probability L_n cannot be smaller than R_n^*.

(ii): The decision of the partitioning rule in a cell is different from the Bayes decision exactly if either there are more ones in the cell than zeros, but $\nu(A_{nj})$, the probability that a label in the cell is 1, is less than $\mu(A_{nj}) - \nu(A_{nj})$, the probabilty that the label is 0, or vica versa. And in that case the difference between L_n and the Bayes error is the difference between the two probabilities: $|\mu(A_{nj}) - \nu(A_{nj}) - \nu(A_{nj})| = |\mu(A_{nj}) - 2\nu(A_{nj})|$.

$$\mathbf{E}(L_n - R_n^*) =$$
$$\mathbf{E}\left(\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| I_{\{sign(\mu_n(A_{nj}) - 2\nu_n(A_{nj})) \neq sign(\mu(A_{nj}) - 2\nu(A_{nj}))\}}\right)$$

where

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}} Y_i \quad \text{and} \quad \mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}}.$$

Thus

$$
\begin{aligned}
&\mathbf{E}(L_n - R_n^*) \\
&= \mathbf{E}\left(\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| I_{\{sign(\mu_n(A_{nj})-2\nu_n(A_{nj}))\neq sign(\mu(A_{nj})-2\nu(A_{nj}))\}}\right) \\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \\
&\qquad \mathbf{P}\left(sign(\mu_n(A_{nj}) - 2\nu_n(A_{nj})) \neq sign(\mu(A_{nj}) - 2\nu(A_{nj}))\right) \\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \Big[\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) > 0\right) I_{\{\mu(A_{nj})-2\nu(A_{nj})\leq 0\}} \\
&\qquad + \mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) \leq 0\right) I_{\{\mu(A_{nj})-2\nu(A_{nj})>0\}}\Big] \\
&= \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \\
&\quad [\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) > -(\mu(A_{nj}) - 2\nu(A_{nj}))\right) \\
&\qquad I_{\{\mu(A_{nj})-2\nu(A_{nj})\leq 0\}} \\
&+ \mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) \leq -(\mu(A_{nj}) - 2\nu(A_{nj}))\right) \\
&\qquad I_{\{\mu(A_{nj})-2\nu(A_{nj})>0\}}\Big] \\
&\leq \sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \\
&\quad [\mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) > |\mu(A_{nj}) - 2\nu(A_{nj})|\right) \\
&+ \mathbf{P}\left(\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj})) \leq -|\mu(A_{nj}) - 2\nu(A_{nj})|\right)] \\
&\leq 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| \\
&\quad \mathbf{P}\left(|\mu_n(A_{nj}) - 2\nu_n(A_{nj}) - (\mu(A_{nj}) - 2\nu(A_{nj}))| > |\mu(A_{nj}) - 2\nu(A_{nj})|\right)
\end{aligned} \tag{3.1}
$$

We can apply Bernstein's inequality with $\epsilon = |\mu(A) - 2\nu(A)|$ and $Z_i = I_{\{X_i \in A\}}(1 - 2Y_i) - (\mu(A) - 2\nu(A))$, because they are independent identically distributed zero mean random variables and $|Z_i| \leq 2$.

LEMMA 3.1. *(Bernstein (1946)) For independent identically distributed zero mean random variables Z_i with $|Z_i| \leq t$, and for all $\epsilon > 0$*

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \epsilon\right) \leq 2\exp\left(\frac{-n\epsilon^2}{2(\mathbf{E}(Z_i^2) + t\epsilon)}\right).$$

Thus

$$\mathbf{P}\left(|\mu_n(A) - 2\nu_n(A) - (\mu(A) - 2\nu(A))| > |\mu(A) - 2\nu(A)|\right)$$
$$\leq 2\exp\left(\frac{-n|\mu(A) - 2\nu(A)|^2}{2(\sigma_A^2 + 2|\mu(A) - 2\nu(A)|)}\right),$$

where $\sigma_A^2 = \mathbf{E}(Z_i^2) = \mu(A) - (\mu(A) - 2\nu(A))^2$.

From this and (3.1)

$$\begin{aligned}
&\mathbf{E}(L_n - R_n^*) \\
\leq\;& 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{\frac{-n|\mu(A_{nj})-2\nu(A_{nj})|^2}{2(\mu(A_{nj})-(\mu(A_{nj})-2\nu(A_{nj}))^2+2|\mu(A_{nj})-2\nu(A_{nj})|)}} \\
\leq\;& 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{\frac{-n|\mu(A_{nj})-2\nu(A_{nj})|^2}{2(3\mu(A_{nj}))}} \\
\leq\;& 2\sum_{j=1}^{m_n} |\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{-\frac{c}{6}n|\mu(A_{nj})-2\nu(A_{nj})|} \\
& \text{(because of Assumption 3.1)} \\
=\;& 2\frac{6}{cn}\sum_{j=1}^{m_n} \frac{c}{6}n|\mu(A_{nj}) - 2\nu(A_{nj})| 2e^{-\frac{c}{6}n|\mu(A_{nj})-2\nu(A_{nj})|} \\
\leq\;& 4\frac{6}{cn}m_n \max_z ze^{-z} \\
=\;& \frac{12}{ce}\frac{m_n}{n}
\end{aligned}$$

□

Corollary 3.1. *(Pintér (2000)) If in addition to Assumption 3.1*

$$\lim_{n\to\infty} \frac{m_n^2}{n} = 0, \tag{3.2}$$

then

$$\sqrt{n}\,(L_n - R_n^*) \to 0$$

in probability as $n \to \infty$.

Proof

It is well-known that if, for a nonnegative random variable Z_n, $\mathbf{E}Z_n \to 0$, then $Z_n \to 0$ in probability. Thus we have to prove that

$$\sqrt{n}\mathbf{E}(L_n - R_n^*) \to 0$$

as $n \to \infty$. From Theorem 3.1

$$\sqrt{n}\mathbf{E}\left(L_n - R_n^*\right) \le \frac{12e^{-1}}{c}\frac{m_n}{\sqrt{n}} \to 0$$

because of condition (3.2). □

3.2 Error Estimation

Estimating the error probability of a classification rule g_n is of great importance. The designer always wants to know what performance can be expected from a classifier. Since the distribution of the data is unknown, it is important to find and analyze error estimation methods that work well independently from the distribution of (X, Y).

One natural way of error estimation is to split the data into two parts, a training sample and a testing sample, and use the first for designing a classifier and the latter for the estimation. But when data are expensive, this constitutes a waste. There are well-known methods that use the same sample for training and testing: the resubstitution estimate, the deleted estimate etc.

3.2.1 Resubstitution Estimate

The resubstitution estimate $L_n^{(R)}$ counts the number of errors committed on the training sequence $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ by the classification rule, i.e. for a classifier g_n it is defined as

$$L_n^{(R)} = \frac{1}{n}\sum_{i=1}^{n} I_{\{g_n(X_i) \ne Y_i\}},$$

which for the partitioning classification rule can be written in the following form:

$$L_n^{(R)} = \sum_{j=1}^{m_n} \min\{\nu_n(A_{nj}), \mu_n(A_{nj}) - \nu_n(A_{nj})\}, \tag{3.3}$$

where

$$\nu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}} Y_i \quad \text{and} \quad \mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in A\}}.$$

$L_n^{(R)}$ is an estimate of the Bayes error R_n^* restricted to the partition $\mathcal{P}_n$.

Concerning the resubstitution error estimate for partitioning rule, the following inequalities are known (see Sec. 23.2 in Devroye, Györfi and Lugosi (1996)): for an arbitrary partition $\mathcal{P}_n$,

$$Var(L_n^{(R)}) \leq \frac{1}{n}$$

and

$$\mathbf{E}L_n^{(R)} \leq R_n^*.$$

For a finite partition of size m_n

$$R_n^* - \mathbf{E}L_n^{(R)} \leq \sqrt{\frac{2m_n}{n}}. \tag{3.4}$$

Combining this with the result of Theorem 3.1, we get

$$\mathbf{E}(L_n - L_n^{(R)}) \leq \sqrt{\frac{2m_n}{n}} + \frac{12}{ce}\frac{m_n}{n}.$$

The resubstitution estimate for the partitioning rule is asymptotically normal under certain conditions:

Theorem 3.2. *Assume that*

$$\lim_{n\to\infty} \max_j diam(A_{nj}) = 0, \tag{3.5}$$

$$\lim_{n\to\infty} \max_j \mu(A_{nj}) = 0 \tag{3.6}$$

and

$$\lim_{n\to\infty} \frac{m_n}{n} = 0, \tag{3.7}$$

If Assumption 3.1 is satisfied then there exists a sequence C_n such that

$$n^{1/2}\left(L_n^{(R)} - C_n\right)/\sqrt{\frac{1}{4}L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1).$$

The basics of our proof is the Poissonization. Introduce the notation N_n for a Poisson(n) random variable independent of $(X_1, Y_1), (X_2, Y_2), \ldots$, moreover put

$$n\tilde{\nu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}} Y_i$$

and

$$n\tilde{\mu}_n(A) = \sum_{i=1}^{N_n} I_{\{X_i \in A\}}.$$

Assume that the $\{A_{nj}\}$ are ordered according to non-decreasing distances of their centers from the origin. For $\alpha \in (0,1)$ choose the integer l_n such that

$$1 - \alpha_n = \sum_{j=1}^{l_n} \mu(A_{nj}) \leq 1 - \alpha < \sum_{j=1}^{l_n+1} \mu(A_{nj}).$$

Roughly speaking, $S_{\alpha_n} = \bigcup_{j=1}^{l_n} A_{nj}$ is approximately a sphere centered at the origin with $1 - \alpha \sim \mu(S_{\alpha_n})$. Because of (3.6)

$$0 \leq \alpha_n - \alpha \leq \mu(A_{n,m_n+1}) \leq \max_j \mu(A_{nj}) \to 0$$

as $n \to \infty$. Also define S_α as the sphere centered at the origin with the property $1 - \alpha = \mu(S_\alpha)$.

The following statement, a slight modification of the statements in Beirlant and Mason (1995) and Beirlant, Györfi and Lugosi (1994), is a key ingredient of the proof.

PROPOSITION 3.1. *Let g_{nj} be real measurable functions with*

$$\mathbf{E}\left\{g_{nj}\left(\tilde{\nu}_n(A_{nj}), \tilde{\mu}_n(A_{nj})\right)\right\} = 0 \quad (n, j \geq 1),$$

and let
$\tilde{M}_{n,\alpha} = \sum_{j=1}^{l_n} g_{nj}\left(\tilde{\nu}_n(A_{nj}), \tilde{\mu}_n(A_{nj})\right)$ *and* $M_{n,\alpha} = \sum_{j=1}^{l_n} g_{nj}\left(\nu_n(A_{nj}), \mu_n(A_{nj})\right)$.
Assume that

$$\begin{aligned} \Phi_{n,\alpha}(t,v) &= \mathbf{E}\left(\exp\left(it\tilde{M}_{n,\alpha} + iv\frac{N_n - n}{\sqrt{n}}\right)\right) \\ &\to e^{-(t^2\rho_\alpha^2 + 2tv\gamma_\alpha + v^2)/2}, \end{aligned}$$

with $\lim_{\alpha\to 0} \rho_\alpha = \rho_0$ *and* $\lim_{\alpha \to 0} \gamma_\alpha = \gamma_0$ *such that* $\gamma_0^2 < \rho_0^2$. *Then*

$$M_{n,\alpha} \Big/ \sqrt{\rho_0^2 - \gamma_0^2} \xrightarrow{\mathcal{D}} N(0,1).$$

PROOF. Clearly,

$$E\{\exp(itM_{n,\alpha})\} = E\{\exp(it\tilde{M}_{n,\alpha}) | N_n = n\}. \tag{3.8}$$

Obviously,

$$E\left(\exp\left(it\tilde{M}_{n,\alpha}+iuN_n\right)\right)=\sum_{k=0}^{\infty}e^{iuk}E\left(\exp\left(it\tilde{M}_{n,\alpha}\right)|N_n=k\right)p_n(k)$$

with

$$p_n(k)=P\{N_n=k\}=e^{-n}n^k/k!.$$

From this, by Fourier's inversion formula, it follows that

$$E\left(\exp\left(it\tilde{M}_{n,\alpha}\right)|N_n=n\right)=\frac{1}{2\pi p_n(n)}\int_{-\pi}^{\pi}e^{-iun}E\left(\exp\left(it\tilde{M}_{n,\alpha}+iuN_n\right)\right)du.$$

Stirling's formula gives that

$$2\pi p_n(n)=2\pi e^{-n}n^{n-1}/(n-1)!\sim(2\pi/n)^{1/2}\text{ as }n\to\infty.$$

Hence, substituting v for $u\sqrt{n}$, we get by (3.8) that

$$\begin{aligned}\Psi_{n,\alpha}(t) &= E\left(\exp\left(itM_n\right)\right)\\ &= (2\pi)^{-1/2}(1+o(1))\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}}\Phi_{n,\alpha}(t,v)dv.\end{aligned}$$

By the assumption of the proposition

$$\Phi_{n,\alpha}(t,v)\to e^{-(t^2\rho_\alpha^2+2tv\gamma_\alpha+v^2)/2},$$

as $n\to\infty$. By a variant of the dominated convergence theorem (see Rao (1973) p.136) it follows that

$$\Psi_{n,\alpha}(t)\to e^{-t^2(\rho_\alpha^2-\gamma_\alpha^2)/2},$$

as $n\to\infty$. Thus

$$M_{n,\alpha}\Big/\sqrt{\rho_\alpha^2-\gamma_\alpha^2}\xrightarrow{\mathcal{D}}N(0,1)\tag{3.9}$$

as $n\to\infty$. □

In the proof of Theorem 3.2 we will use the following lemmas.

Lemma 3.2. *(Györfi and Pintér (1998)) For a set A*

$$0\le\mathbf{E}|n\tilde{\tau}_n(A)|-|n\tau(A)|\le 2n\mu(A)e^{-\frac{n\tau(A)^2}{3\mu(A)}}.$$

PROOF. Without loss of generality assume that $\tau(A) > 0$, then

$$\begin{aligned}
\mathbf{E}|n\tilde{\tau}_n(A)| &= n\tau(A) + 2\mathbf{E}(-n\tilde{\tau}_n(A))^+ \\
&= n|\tau(A)| + 2\mathbf{E}(-n\tilde{\tau}_n(A))^+
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{E}(-n\tilde{\tau}_n(A))^+ \\
&= \sum_{k=0}^{\infty} \mathbf{E}\left\{(-n\tilde{\tau}_n(A))^+ \,|\tilde{\mu}_n(A) = \frac{k}{n}\right\} \\
&\qquad \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)} \\
&= \sum_{k=0}^{\infty} \mathbf{E}\left(-\sum_{i=1}^{k}(1-2Y_i(A))\right)^+ \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)} \\
&\le \sum_{k=1}^{\infty} k\mathbf{P}\{\sum_{i=1}^{k}(1-2Y_i(A)) < 0\} \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)} \\
&= n\mu(A) \sum_{k=0}^{\infty} \mathbf{P}\{\sum_{i=1}^{k+1}(1-2Y_i(A)) < 0\} \\
&\qquad \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)}.
\end{aligned}$$

By Hoeffding's inequality

$$\begin{aligned}
&\mathbf{P}\{\sum_{i=1}^{k+1}(1-2Y_i(A)) < 0\} \\
&= \mathbf{P}\{\frac{1}{k+1}\sum_{i=1}^{k+1}\left((1-2Y_i(A)) - \frac{\tau(A)}{\mu(A)}\right) < -\frac{\tau(A)}{\mu(A)}\} \\
&\le e^{-(k+1)\tau(A)^2/(2\mu(A)^2)} \le e^{-k\tau(A)^2/(2\mu(A)^2)},
\end{aligned}$$

thus

$$\begin{aligned}
&\mathbf{E}(-n\tilde{\tau}_n(A))^+ \\
&\le n\mu(A) \sum_{k=0}^{\infty} e^{-k\tau(A)^2/(2\mu(A)^2)} \frac{(n\mu(A))^k}{k!} e^{-n\mu(A)} \\
&\le n\mu(A) e^{e^{-\tau(A)^2/(2\mu(A)^2)} n\mu(A)} e^{-n\mu(A)} \\
&\le n\mu(A) e^{-n\tau(A)^2/(3\mu(A))}
\end{aligned}$$

since $\frac{\tau(A)^2}{\mu(A)^2} \le 1$ and for $0 \le x \le 1$ we have $-1 + e^{-x/2} \le -x/3$. □

LEMMA 3.3. *(Györfi and Pintér (1998)) If Z is a random variable with $\mathbf{E}|Z|^3 < \infty$ then*

$$\mathbf{E}||Z| - \mathbf{E}|Z||^3 < 27\mathbf{E}|Z - EZ|^3.$$

PROOF. If Z is any random variable with finite first moment, then first we show that

$$||Z| - \mathbf{E}|Z|| \leq |Z - \mathbf{E}Z| + 2\mathbf{E}\{|Z - \mathbf{E}Z|\} .$$

Assume without loss of generality that $\mathbf{E}Z > 0$. Then

$$||Z| - \mathbf{E}|Z|| \leq |Z - \mathbf{E}Z| + \mathbf{E}|Z| - \mathbf{E}Z .$$

This follows by considering four cases separately, $Z > \mathbf{E}|Z|$, $Z < -\mathbf{E}|Z|$, $\mathbf{E}Z < Z < \mathbf{E}|Z|$, and $\mathbf{E}Z > Z > -\mathbf{E}|Z|$. A drawing of $||z| - \mathbf{E}|Z||$ and $|z - \mathbf{E}Z|$ as a function of z is equally convincing. But

$$|Z| - Z = -2ZI_{Z<0} \leq -2ZI_{Z<0} + 2I_{Z<0}\mathbf{E}Z \leq 2|\mathbf{E}Z - Z|$$

and thus, $\mathbf{E}\{|Z| - Z\} \leq 2\mathbf{E}\{|Z - \mathbf{E}Z|\}$. Let $|Z - \mathbf{E}Z|$ have i-th moment μ_i. Then

$$\begin{aligned}
\mathbf{E}\{||Z| - \mathbf{E}|Z||^3\} &\leq \mu_3 + 6\mu_2\mu_1 + 12\mu_1^3 + 8\mu_1^3 \\
&\leq (1 + 6 + 12 + 8)\mu_3 \\
&= 27\mathbf{E}\{|Z - \mathbf{E}Z|^3\},
\end{aligned}$$

where the second inequality comes from the fact that
$(\mathbf{E}|Z - \mathbf{E}Z|^i)^{\frac{1}{i}}$ is monoton increasing. □

PROOF OF THEOREM 3.2. Because of $\min\{a, b\} = \frac{1}{2}(a + b - |a - b|)$

$$L_n^{(R)} = \frac{1}{2}(1 - J_n), \tag{3.10}$$

where

$$J_n = \sum_{j=1}^{m_n} |\tau_n(A_{nj})| \tag{3.11}$$

with

$$\tau_n(A) = \mu_n(A) - 2\nu_n(A),$$

therefore the statement of Theorem 3.2 is equivalent to

$$n^{1/2}\left(J_n - (1-2C_n)\right)/\sqrt{L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1). \tag{3.12}$$

In order to prove (3.12) let

$$1 - 2C_n = \sum_{j=1}^{m_n} \mathbf{E}|\tilde{\tau}_n(A_{nj})|,$$

where

$$\tilde{\tau}_n(A) = \tilde{\mu}_n(A) - 2\tilde{\nu}_n(A).$$

So we need to prove that

$$n^{1/2}\left(\sum_{j=1}^{m_n}\left(|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)\right)/\sqrt{L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1).$$

A result by Slutsky, saying that if Z_n and U_n are two sequences of random variables such that Z_n converges to the random variable Z in distribution aand U_n converges to a constant u_0 in probability then $Z_n + U_n$ converges to $Z + u_0$ in distribution, allows us to do this in the following two steps.
(1) We show that for $M_{n,\alpha} = n^{1/2}\sum_{j=1}^{l_n}\left(|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)$

$$M_{n,\alpha}/\sqrt{L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1),$$

as $n \to \infty$ and $\alpha \to 0$,
(2) then we show that for $M'_{n,\alpha} = \sqrt{n}\sum_{j=l_n+1}^{m_n}\left(|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)$

$$\lim_{\alpha\to 0}\lim_{n\to\infty}\mathbf{P}(|M'_{n,\alpha}| > \epsilon) = 0$$

for all $\epsilon > 0$.

And then we can use the following slightly amended version of Slutsky's theorem with $Z_{n,\alpha} = M_{n,\alpha}$, $U_{n,\alpha} = M'_{n,\alpha}$ and $u_0 = 0$ to complete the proof.

Lemma 3.4. *If $Z_{n,\alpha} \xrightarrow{\mathcal{D}} Z$ and $\lim_{\alpha\to 0}\lim_{n\to\infty}\mathbf{P}(|U_{n,\alpha}| > \epsilon) = u_0$ (a constant) then $Z_{n,\alpha} + U_{n,\alpha} \xrightarrow{\mathcal{D}} Z + u_0$.*

For the sake of completeness we present the proof, which differs from the textbook proof of Slutsky's theorem in Bickel and Doksum ((1976)) only in a small detail.

PROOF. Let $F_{(Z_{n,\alpha}+U_{n,\alpha})}$ and $F_{(Z+u_0)}$ denote the distribution function of $Z_{n,\alpha}+U_{n,\alpha}$ and $Z+u_0$ respectively. Let t be a continuity point of $F_{(Z+u_0)}$. A distribution function can have at most only countably many discontinuity points, thus for any t we can find an arbitrarily small $\epsilon > 0$ such that both $t+\epsilon$ and $t-\epsilon$ are continuity points of $F_{(Z+u_0)}$. Clearly,

$$\begin{aligned}
&F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) \\
&= \mathbf{P}(Z_{n,\alpha}+U_{n,\alpha} \le t, U_{n,\alpha} \ge u_0-\epsilon) + \mathbf{P}(Z_{n,\alpha}+U_{n,\alpha} \le t, U_{n,\alpha} < u_0-\epsilon) \\
&\le \mathbf{P}(Z_{n,\alpha} \le t-u_0+\epsilon) + \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon) \\
&\le F_{(Z_{n,\alpha}+u_0)}(t+\epsilon) + \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon).
\end{aligned}$$

Let Z_α be the random variable whose distribution is the limit distribution of $Z_{n,\alpha}$ as $n \to \infty$. Since $F_{(Z_{n,\alpha}+u_0)}(t) = \mathbf{P}(Z_{n,\alpha} \le t-u_0) = F_{Z_{n,\alpha}}(t-u_0)$, $Z_{n,\alpha}+u_0 \xrightarrow{\mathcal{D}} Z_\alpha + u_0$ as $n \to \infty$. Therefore

$$\begin{aligned}
\limsup_{n\to\infty} F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) &\le \lim_{n\to infty} F_{Z_{n,\alpha}}(t+\epsilon) + \lim_{n\to\infty} \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon) \\
&= F_{Z_\alpha}(t+\epsilon) + g(\alpha),
\end{aligned}$$

where $g(\alpha)$ stands for $\lim_{n\to\infty} \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon)$ which obviously depends only on α.

Similarly,

$$\liminf_{n\to\infty} F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) \ge \lim_{n\to\infty} F_{Z_{n,\alpha}}(t-\epsilon) = F_{Z_\alpha}(t-\epsilon),$$

as

$$\begin{aligned}
1 - F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) &= \mathbf{P}(Z_{n,\alpha}+U_{n,\alpha} > t) \\
&\le \mathbf{P}(Z_{n,\alpha} > t-u_0-\epsilon) + \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon) \\
&\le F_{(Z_{n,\alpha}+u_0)}(t-\epsilon) + \mathbf{P}(|U_{n,\alpha}-u_0| > \epsilon).
\end{aligned}$$

Thus, on the one hand,

$$\lim_{\alpha\to 0} \limsup_{n\to\infty} F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) \le \lim_{\alpha\to 0} F_{Z_\alpha}(t+\epsilon) + \lim_{\alpha\to 0} g(\alpha) = F_Z(t+\epsilon),$$

since $\lim_{\alpha\to 0} g(\alpha) = 0$ by assumption. While on the other hand,

$$\lim_{\alpha\to 0} \liminf_{n\to\infty} F_{(Z_{n,\alpha}+U_{n,\alpha})}(t) \ge \lim_{\alpha\to 0} F_{Z_\alpha}(t-\epsilon) = F_Z(t-\epsilon).$$

The result follows from the fact that ϵ can be arbitrarily small. □

STEP (1): We use the Proposition to establish the asymptotics for $\sqrt{n}\left(\sum_{j=1}^{l_n}\left(|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)\right)$. We choose the functions g_{nj} as

$$g_{nj}(u,v) = \sqrt{n}\left(|v-2u| - \mathbf{E}\left|\tilde{\mu}_n(A_{nj}) - 2\tilde{\nu}_n(A_{nj})\right|\right)$$

for $j = 1, \ldots, m_n$ and

$$\rho_\alpha = 1-\alpha \qquad \text{and} \qquad \gamma_\alpha = \int_{S_\alpha} |1-2\eta(x)|\mu(dx)$$

thus $\rho_0 = 1$, $\gamma_0 = \int_{S_\alpha} |1-2\eta(x)|\mu(dx) = \mathbf{E}|1-2\eta(X)| = 1-2L^*$ and

$$\rho_0^2 - \gamma_0^2 = L^*(4-L^*).$$

Now, we check the conditions of the Proposition. Introduce

$$\begin{aligned} S_n &= t\sqrt{n}\sum_{j=1}^{l_n}\left(|\tilde{\mu}_n(A_{nj}) - 2\tilde{\nu}_n(A_{nj})|\right. \\ &\quad \left. - \mathbf{E}|\tilde{\mu}_n(A_{nj}) - 2\tilde{\nu}_n(A_{nj})|\right) + v\frac{N_n-n}{\sqrt{n}} \\ &= t\sqrt{n}\sum_{j=1}^{l_n}\left(|\tilde{\tau}_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|\right) + v\frac{N_n-n}{\sqrt{n}} \end{aligned}$$

for which a central limit result is to be shown to hold. Remark that

$$\begin{aligned} Var(S_n) &= n\sum_{j=1}^{l_n}\left\{t^2 Var(|\tilde{\tau}_n(A_{nj})|)\right. \\ &\quad \left. + 2tv\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj}) - \mu(A_{nj}))\right)\right\} \\ &\quad + v^2, \end{aligned}$$

since

$$\begin{aligned} N_n - n &= \sum_{j=1}^{m_n} n\left(\tilde{\mu}_n(A_{nj}) - \mu(A_{nj})\right) \\ &= \sum_{j=1}^{l_n} n\left(\tilde{\mu}_n(A_{nj}) - \mu(A_{nj})\right) + \sum_{j=l_n+1}^{m_n} n\left(\tilde{\mu}_n(A_{nj}) - \mu(A_{nj})\right) \end{aligned}$$

Apply the following inequality for $Z_i = I_{\{X_i \in A\}}(1-2Y_i)$ and $p = 2$.

LEMMA 3.5. *(Beirlant, Mason (1995)) Let $\{Z_i\}$ i.i.d. random variables with $\mathbf{E}|Z_1|^p < \infty$, let N_n be a Poisson(n) random variable independent of $\{Z_i\}$, and N a standard normal random variable. Then there is a constant A_p (depending only on p) such that*

$$\left|\mathbf{E}\left|\frac{\sum_{i=1}^{N_n} Z_i - n\mathbf{E}Z_1}{\sqrt{n\mathbf{E}Z_1^2}}\right|^p - \mathbf{E}|N|^p\right| \leq \frac{A_p}{n^{1/2}}\left(\frac{\mathbf{E}|Z_1|^3}{(\mathbf{E}Z_1^2)^{3/2}} + 1\right)$$

$(1 \leq p \leq 3)$.

Then

$$\begin{aligned} |nVar(\tilde{\tau}_n(A)) - \mu(A)| &\leq \frac{A_2}{n^{1/2}}(\mu(A)^{1/2} + \mu(A)) \\ &\leq \frac{2A_2}{n^{1/2}}\mu(A)^{1/2}, \end{aligned} \tag{3.13}$$

therefore, on the one hand we get that

$$\begin{aligned} n\sum_{j=1}^{l_n} Var(\tilde{\tau}_n(A_{nj})) &\leq \sum_{j=1}^{l_n}\mu(A_{nj}) + \frac{2A_2}{n^{1/2}}\sum_{j=1}^{l_n}\mu(A_{nj})^{1/2} \\ &\leq 1 - \alpha + 2A_2\sqrt{\frac{l_n}{n}} \\ &\leq 1 - \alpha + 2A_2\sqrt{\frac{m_n}{n}} \\ &\to 1 - \alpha = \rho_\alpha^2 \end{aligned}$$

as $n \to \infty$ because of (3.7). Thus

$$\begin{aligned} &\limsup_n n\sum_{j=1}^{l_n} Var(|\tilde{\tau}_n(A_{nj})|) \\ &\leq \limsup_n n\sum_{j=1}^{l_n} Var(\tilde{\tau}_n(A_{nj})) \\ &\leq \rho_\alpha^2. \end{aligned}$$

We should prove that

$$\liminf_n n\sum_{j=1}^{l_n} Var(|\tilde{\tau}_n(A_{nj})|) \geq \rho_\alpha^2.$$

Obviously

$$\begin{aligned}
&\liminf_n n\sum_{j=1}^{l_n} Var(|\tilde{\tau}_n(A_{nj})|)\\
&\geq \liminf_n n\sum_{j=1}^{l_n} Var(\tilde{\tau}_n(A_{nj}))\\
&\quad -\limsup_n n\sum_{j=1}^{l_n}((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-(\mathbf{E}\tilde{\tau}_n(A_{nj}))^2),
\end{aligned}$$

thus we need that

$$\limsup_n n\sum_{j=1}^{l_n}((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-\tau(A_{nj})^2)=0.$$

By Lemma 3.2 for a set A

$$\begin{aligned}
&n((\mathbf{E}|\tilde{\tau}_n(A)|)^2-\tau(A)^2)\\
&=\frac{1}{n}((\mathbf{E}|n\tilde{\tau}_n(A)|)^2-(n\tau(A))^2)\\
&\leq \frac{1}{n}((2n\mu(A)e^{-n\tau(A)^2/(3\mu(A))}+n|\tau(A)|)^2-(n\tau(A))^2)\\
&=\frac{1}{n}(4n^2\mu(A)^2e^{-2n\tau(A)^2/(3\mu(A))}\\
&+4n^2\mu(A)|\tau(A)|e^{-n\tau(A)^2/(3\mu(A))})\\
&\leq 8n\mu(A)^2e^{-n\tau(A)^2/(3\mu(A))}.
\end{aligned}$$

Therefore by Assumption 3.1 and condition (3.7)

$$\begin{aligned}
&\sum_{j=1}^{l_n}\Big(n((\mathbf{E}|\tilde{\tau}_n(A_{nj})|)^2-\tau(A_{nj})^2)\Big)\\
&\leq \sum_{j=1}^{l_n} 8n\mu(A_{nj})^2e^{-n\tau(A_{nj})^2/(3\mu(A_{nj}))}\\
&\leq \sum_{j=1}^{l_n} 8n\mu(A_{nj})^2e^{-\frac{1}{3}nc^2\mu(A_{nj})}\\
&\leq 8\max_{z\geq 0} z^2e^{-z}\frac{9}{c^4}\frac{l_n}{n}\\
&\leq 8\max_{z\geq 0} z^2e^{-z}\frac{9}{c^4}\frac{m_n}{n}\to 0
\end{aligned}$$

To complete the asymptotics for $Var(S_n)$ it remains to show that

$$n\sum_{j=1}^{l_n}\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj})-\mu(A_{nj}))\right)\to\gamma_\alpha$$

as $n \to \infty$. For a set A

$$n\mathbf{E}\left(|\tilde{\tau}_n(A)|(\tilde{\mu}_n(A)-\mu(A))\right)$$
$$= n\mathbf{E}(|\tilde{\tau}_n(A)|\tilde{\mu}_n(A)) - n\mathbf{E}(|\tilde{\tau}_n(A)|)\mu(A)$$

For a set A let $Y(A)$ be a random variable with

$$\mathbf{P}(Y(A)\in B) = \mathbf{P}(Y\in B|X\in A)$$

then on the one hand

$$\begin{aligned} & n\mathbf{E}(|\tilde{\tau}_n(A)|\tilde{\mu}_n(A)) \\ &= \sum_{k=0}^{\infty}\mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\frac{k}{n}\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)} \\ &= \sum_{k=1}^{\infty}\mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\mu(A)\frac{(n\mu(A))^{k-1}}{(k-1)!}e^{-n\mu(A)} \\ &= \mu(A)\sum_{k=0}^{\infty}\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))|\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)}. \end{aligned}$$

On the other hand

$$\begin{aligned} & n\mathbf{E}(|\tilde{\tau}_n(A)|)\mu(A) \\ &= \mu(A)\sum_{k=0}^{\infty}\mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\frac{(n\mu(A))^k}{k!}e^{-n\mu(A)}, \end{aligned}$$

thus

$$\begin{aligned} & n\mathbf{E}\left(|\tilde{\tau}_n(A)|(\tilde{\mu}_n(A)-\mu(A))\right) \\ &= \mu(A)\sum_{k=0}^{\infty}\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right) \\ &\qquad \frac{(n\mu(A))^k}{k!}e^{-n\mu(A)}. \end{aligned}$$

Without loss of generality assume that $\tau(A) \geq 0$ then

$$\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right)$$

$$= \mu(A)\left(\left(\mathbf{E}\sum_{i=1}^{k+1}(1-2Y_i(A)) + 2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+\right)\right.$$

$$
\begin{aligned}
- &\quad \left(\mathbf{E}\sum_{i=1}^{k}(1-2Y_i(A)) + 2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+\right)\Bigg)\\
= \mu(A) &\left(\mathbf{E}(1-2Y_1(A)) + 2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+\right.\\
- &\quad \left. 2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+\right)
\end{aligned}
$$

Because of

$$
|2(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^+ - 2(-\sum_{i=1}^{k}(1-2Y_i(A)))^+| \le 2
$$

apply Hoeffding's inequality as in the proof of Lemma 3.2

$$
\begin{aligned}
&|\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right)\\
&-\mu(A)\mathbf{E}(1-2Y_1(A))|\\
&\quad = \quad |\mu(A)\left(\mathbf{E}|\sum_{i=1}^{k+1}(1-2Y_i(A))| - \mathbf{E}|\sum_{i=1}^{k}(1-2Y_i(A))|\right)\\
&\qquad\qquad -|\tau(A)||\\
&\quad = \quad |2\mathbf{E}(-\sum_{i=1}^{k+1}(1-2Y_i(A)))^{-}2\mathbf{E}(-\sum_{i=1}^{k}(1-2Y_i(A)))^+|\\
&\quad \le \quad 2\mu(A)\left(\mathbf{P}\{\sum_{i=1}^{k+1}(1-2Y_i(A)) < 0\}\right.\\
&\qquad\qquad\qquad \left.+\mathbf{P}\{\sum_{i=1}^{k}(1-2Y_i(A)) < 0\}\right)\\
&\quad \le \quad 2\mu(A)\left(e^{-(k+1)\tau(A)^2/(2\mu(A)^2)} + e^{-k\tau(A)^2/(2\mu(A)^2)}\right)\\
&\quad \le \quad 4\mu(A)e^{-k\tau(A)^2/(2\mu(A)^2)}.
\end{aligned}
$$

Summarizing the consequences of these inequlities we get that

$$
\begin{aligned}
&|n\sum_{j=1}^{l_n}\mathbf{E}\left(|\tilde{\tau}_n(A_{nj})|(\tilde{\mu}_n(A_{nj}) - \mu(A_{nj}))\right) - \sum_{j=1}^{l_n}|\tau(A_{nj})||\\
&\le \quad \sum_{j=1}^{l_n}4\mu(A_{nj})\\
&\qquad \sum_{k=0}^{\infty}e^{-k\tau(A_{nj})^2/(2\mu(A_{nj})^2)}\frac{(n\mu(A_{nj}))^k}{k!}e^{-n\mu(A_{nj})}\\
&\le \quad \sum_{j=1}^{l_n}4\mu(A_{nj})e^{-n\tau(A_{nj})^2/(3\mu(A_{nj}))} \to 0,
\end{aligned}
$$

therefore

$$n\sum_{j=1}^{l_n}\mathbf{E}\left(|\tilde\tau_n(A_{nj})|(\tilde\mu_n(A_{nj})-\mu(A_{nj}))\right)$$
$$=o(1)+\sum_{j=1}^{l_n}|\tau(A_{nj})|\to\gamma_\alpha.$$

To finish the proof of

$$S_n\stackrel{\mathcal{D}}{\to}N\left(0,t^2\rho_\alpha^2+2tv\gamma_\alpha+v^2\right)\ \text{ as } n\to\infty$$

by Lyapunov's central limit theorem, it suffices to prove that

$$\begin{aligned} n^{3/2}\Bigg(&\sum_{j=1}^{l_n}\mathbf{E}\left(|t\left(|\tilde\tau_n(A_{nj})|-\mathbf{E}|\tilde\tau_n(A_{nj})|\right)\right.\\ +&\quad \left.v(\tilde\mu_n(A_{nj})-\mu(A_{nj}))|^3\right)\\ +&\quad\sum_{j=l_n+1}^{m_n}\mathbf{E}|v(\tilde\mu_n(A_{nj})-\mu(A_{nj}))|^3\Bigg)\to 0,\end{aligned}$$

or, by invoking the c_r inequality, that

$$n^{3/2}\sum_{j=1}^{l_n}\mathbf{E}\left(||\tilde\tau_n(A_{nj})|-\mathbf{E}|\tilde\tau_n(A_{nj})||^3\right)\to 0,$$

and

$$n^{3/2}\sum_{j=1}^{m_n}\mathbf{E}\left(|\tilde\mu_n(A_{nj})-\mu(A_{nj})|^3\right)\to 0.$$

The second limit is shown in Beirlant, Györfi and Lugosi (1994). Concerning the first one apply Lemma 3.3 and Lemma 3.5 for $p=3$

$$\begin{aligned} &n^{3/2}\sum_{j=1}^{l_n}\mathbf{E}\left(||\tilde\tau_n(A_{nj})|-\mathbf{E}|\tilde\tau_n(A_{nj})||^3\right)\\ &\le\ 27n^{-3/2}\sum_{j=1}^{l_n}\mathbf{E}\left(|n\tilde\tau_n(A_{nj})-n\mathbf{E}\tilde\tau_n(A_{nj})|^3\right)\\ &\le\ 27n^{-3/2}\sum_{j=1}^{l_n}\Big(n^{3/2}\mu(A_{nj})^{3/2}\mathbf{E}|N|^3\\ &\qquad\qquad +n\mu(A_{nj})^{3/2}A_3(\mu(A_{nj})^{-1/2}+1)\Big)\\ &=\ 27\sum_{j=1}^{l_n}\left(\mu(A_{nj})^{3/2}\mathbf{E}|N|^3+n^{-1/2}A_3(\mu(A_{nj})+\mu(A_{nj})^{3/2})\right)\end{aligned}$$

$$
\begin{aligned}
&\leq 27\mathbf{E}|N|^3 \max_j \mu(A_{nj})^{1/2} \sum_{j=1}^{l_n} \mu(A_{nj}) + n^{-1/2} A_3 \sum_{j=1}^{l_n} 2\mu(A_{nj}) \\
&\leq 27\mathbf{E}|N|^3 \max_j \mu(A_{nj})^{1/2} + 2A_3 \frac{1}{\sqrt{n}} \to 0
\end{aligned}
$$

as $n \to \infty$ because of (3.6). So we proved that

$$
n^{1/2} \left(\sum_{j=1}^{l_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|) \right) / \sqrt{L^*(4 - L^*)} \xrightarrow{\mathcal{D}} N(0,1). \tag{3.14}
$$

STEP (2): Now we prove that for all $\epsilon > 0$

$$
\lim_{\alpha \to 0} \lim_{n \to \infty} \mathbf{P}(|M'_{n,\alpha}| > \epsilon) = 0,
$$

where $M'_{n,\alpha} = \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|)$.

$$
\begin{aligned}
&\mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|) \right| > \epsilon \right) \\
&= \mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|) \right.\right. \\
&\qquad \left.\left. + \sqrt{n} \sum_{j=l_n+1}^{m_n} (\mathbf{E}|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|) \right| > \epsilon \right) \\
&\leq \mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|) \right| \right. \\
&\qquad \left. + \left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (\mathbf{E}|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|) \right| > \epsilon \right) \\
&\leq \mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|) \right| > \frac{\epsilon}{2} \right) \\
&\qquad + \mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (\mathbf{E}|\tau_n(A_{nj})| - \mathbf{E}|\tilde{\tau}_n(A_{nj})|) \right| > \frac{\epsilon}{2} \right)
\end{aligned} \tag{3.15}
$$

The first probability can be upper bounded by Chebisev's inequality

$$
\begin{aligned}
&\mathbf{P}\left(\left| \sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|) \right| > \frac{\epsilon}{2} \right) \\
&\leq \frac{1}{\epsilon^2} Var\left(\sqrt{n} \sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|) \right)
\end{aligned}
$$

Now we apply Efron-Stein inequality for $Z_i = (X_i, Y_i)$ and $g(Z_1, Z_2, \ldots, Z_n) = \sqrt{n}\sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|)$.

LEMMA 3.6. *(Efron and Stein (1981), Steele (1986)) Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables taking values in a set A, and let $g : A^n \to \mathcal{R}$ be a measurable function. If $Z_1', Z_2', \ldots, Z_n'$ form an independent copy of $Z_1, Z_2, \ldots, Z_n$, then*

$$Var(g(Z_1, \ldots, Z_n)) \le \frac{1}{2}\sum_{i=1}^{n} \mathbf{E}\left((g(Z_1, \ldots, Z_n) - g(Z_1, \ldots, Z_i', \ldots, Z_n))^2\right).$$

Thus

$$\begin{aligned} & Var\left(\sqrt{n}\sum_{j=l_n+1}^{m_n} (|\tau_n(A_{nj})| - \mathbf{E}|\tau_n(A_{nj})|)\right) \\ \le\ & \frac{1}{2}\sum_{i=1}^{n}\mathbf{E}\left\{\left(\sqrt{n}\sum_{j=l_n+1}^{m_n}\left(|\tau_n(A_{nj})| - |\tau_n^{(i)}(A_{nj})|\right)\right)^2\right\} \end{aligned}$$

where

$$\nu_n^{(i)}(A) = \frac{1}{n}\left(\sum_{\substack{k=1\\k\neq i}}^{n} I_{\{X_k\in A\}}Y_k + I_{\{X_i'\in A\}}Y_i'\right),$$

$$\mu_n^{(i)}(A) = \frac{1}{n}\left(\sum_{\substack{k=1\\k\neq i}}^{n} I_{\{X_k\in A\}} + I_{\{X_i'\in A\}}\right)$$

and

$$\tau_n^{(i)}(A) = \mu_n^{(i)}(A) - 2\nu_n^{(i)}(A).$$

It can be easily checked that $\sum_{j=l_n+1}^{m_n}\left(|\tau_n(A_{nj})| - |\tau_n^{(i)}(A_{nj})|\right)$ is either 0, $\frac{1}{n}$ or $\frac{2}{n}$, and that it is always 0 when both X_i and X_i' fall into S_{α_n}. Therefore

$$\begin{aligned} & \frac{1}{2}\sum_{i=1}^{n}\mathbf{E}\left\{\left(\sqrt{n}\sum_{j=l_n+1}^{m_n}\left(|\tau_n(A_{nj})| - |\tau_n^{(i)}(A_{nj})|\right)\right)^2\right\} \\ \le\ & \frac{1}{2}n\left(\frac{1}{2}\sqrt{n}\right)^2 (1 - \mathbf{P}(X_i \in S_{\alpha_n}, X_i' \in S_{\alpha_n})) \\ =\ & 2(1-\mu(S_{\alpha_n})^2) = 2(1-(1-\alpha_n)^2). \end{aligned}$$

Thus

$$
\begin{aligned}
\lim_{\alpha\to 0}\lim_{n\to\infty}\mathbf{P}\left(\left|\sqrt{n}\sum_{j=l_n+1}^{m_n}\left(|\tau_n(A_{nj})|-\mathbf{E}|\tau_n(A_{nj})|\right)\right|>\frac{\epsilon}{2}\right)& \\
\le \lim_{\alpha\to 0}\lim_{n\to\infty}\frac{1}{\epsilon^2}2(1-(1-\alpha_n)^2)& \\
=\lim_{\alpha\to 0}\frac{1}{\epsilon^2}2(1-(1-\alpha)^2)=0.&
\end{aligned}
$$

Now we show that the second term in (3.15) tends to 0 as well.

$$
\begin{aligned}
&\left|\sqrt{n}\sum_{j=l_n+1}^{m_n}\left(\mathbf{E}|\tau_n(A_{nj})|-\mathbf{E}|\tilde{\tau}_n(A_{nj})|\right)\right| \\
\le\ & \sqrt{n}\sum_{j=l_n+1}^{m_n}\left|\mathbf{E}|\tau_n(A_{nj})|-\mathbf{E}|\tilde{\tau}_n(A_{nj})|\right| \\
\le\ & \sqrt{n}\sum_{j=l_n+1}^{m_n}\mathbf{E}\left||\tau_n(A_{nj})|-|\tilde{\tau}_n(A_{nj})|\right| \\
& \qquad\text{(from Jensen's inequality)} \\
\le\ & \sqrt{n}\sum_{j=l_n+1}^{m_n}\mathbf{E}\left|\tau_n(A_{nj})-\tilde{\tau}_n(A_{nj})\right| \\
\le\ & \sqrt{n}\left(\sum_{j=l_n+1}^{m_n}\mathbf{E}\left|\mu_n(A_{nj})-\tilde{\mu}_n(A_{nj})\right|+2\sum_{j=l_n+1}^{m_n}\mathbf{E}\left|\nu_n(A_{nj})-\tilde{\nu}_n(A_{nj})\right|\right) \\
=\ & \sqrt{n}\left(\mathbf{E}\left|\mu_n(S_{\alpha_n}^c)-\tilde{\mu}_n(S_{\alpha_n}^c)\right|+2\mathbf{E}\left|\nu_n(S_{\alpha_n}^c)-\tilde{\nu}_n(S_{\alpha_n}^c)\right|\right) \\
=\ & \sqrt{n}\mathbf{E}\left|\mu_n(S_{\alpha_n}^c)-\mu(S_{\alpha_n}^c)\right|+\sqrt{n}\mathbf{E}\left|\mu(S_{\alpha_n}^c)-\tilde{\mu}_n(S_{\alpha_n}^c)\right| \\
& +2\sqrt{n}\mathbf{E}\left|\nu_n(S_{\alpha_n}^c)-\nu(S_{\alpha_n}^c)\right|+2\sqrt{n}\mathbf{E}\left|\nu(S_{\alpha_n}^c)-\tilde{\nu}_n(S_{\alpha_n}^c)\right|
\end{aligned}
$$

Now we show that each term can be upper bounded by $\sqrt{\mu(S_{\alpha_n}^c)}$.

$$
\begin{aligned}
\sqrt{n}\mathbf{E}\left|\mu_n(S_{\alpha_n}^c)-\mu(S_{\alpha_n}^c)\right| &\le \sqrt{n}\sqrt{\mathbf{E}\left(\mu_n(S_{\alpha_n}^c)-\mu(S_{\alpha_n}^c)\right)^2} \\
&= \sqrt{n}\sqrt{\frac{\mu(S_{\alpha_n}^c)(1-\mu(S_{\alpha_n}^c))}{n}} \\
&= \sqrt{\mu(S_{\alpha_n}^c)(1-\mu(S_{\alpha_n}^c))}\le\sqrt{\mu(S_{\alpha_n}^c)},
\end{aligned}
$$

where we used that $\mu_n(S_{\alpha_n}^c)$ is binomial with parameters n and $\mu(S_{\alpha_n}^c)$. Similarly, we can derive the same upper bound for the other three terms noting that $\nu_n(S_{\alpha_n}^c)$ is binomial, $\tilde{\mu}_n(S_{\alpha_n}^c)$ and $\tilde{\nu}_n(S_{\alpha_n}^c)$ are Poisson. Therefore

$$\begin{aligned}
\lim_{\alpha\to 0}\lim_{n\to\infty}\left|\sqrt{n}\sum_{j=l_n+1}^{m_n}(\mathbf{E}|\tau_n(A_{nj})|-\mathbf{E}|\tilde{\tau}_n(A_{nj})|)\right| &\leq 6\lim_{\alpha\to 0}\lim_{n\to\infty}\sqrt{\mu(S^c_{\alpha_n})}\\
&= 6\lim_{\alpha\to 0}\lim_{n\to\infty}\sqrt{\alpha_n}\\
&= 6\lim_{\alpha\to 0}\sqrt{\alpha}=0
\end{aligned}$$

□

THEOREM 3.3. *If, in addition to the conditions of Theorem 3.2, (3.2) is satisfied then*

$$n^{1/2}\left(L_n^{(R)}-R_n^*\right)/\sqrt{\frac{1}{4}L^*(4-L^*)}\xrightarrow{\mathcal{D}}N(0,1). \tag{3.16}$$

REMARK.
For cubic partitions with size h_n the conditions mean $h_n\to 0$ and $nh_n^{2d}\to\infty$ as $n\to\infty$. Although the second condition is stronger than the usual condition $nh_n^d\to\infty$ for the consistency of the partitioning rule, it does allow the usual good choices for h_n, i.e. choices that minimize the rate of convergence, e.g. $h_n=n^{-1/3}$. The condition might be weakened with a longer and more complicated proof.

For the proof we will need the following lemma.

LEMMA 3.7. *(Györfi and Pintér (1998)) Let $\{Z_i\}$ i.i.d. random variables with $\mathbf{E}|Z_1|^3<\infty$, let N_n be a Poisson(n) random variable independent of $\{Z_i\}$, and N a standard normal random variable. Then there is a constant A_1 such that*

$$\begin{aligned}
&\sup_a\left|\mathbf{E}\left|\frac{\sum_{i=1}^{N_n}Z_i-n\mathbf{E}Z_1}{\sqrt{n\mathbf{E}Z_1^2}}-a\right|-\mathbf{E}|N-a|\right|\\
&\leq\frac{A_1}{n^{1/2}}\frac{\mathbf{E}|Z_1|^3}{(\mathbf{E}Z_1^2)^{3/2}}
\end{aligned}$$

and

$$0\leq\mathbf{E}|N-a|-|a|\leq\sqrt{\frac{2}{\pi}}e^{-a^2/2}.$$

PROOF. We proceed as in Lemma 2.2 of Beirlant, Mason (1995). Let $\tilde{N}_i$ be independent Poisson(1) random variables such that

$$N_n = \sum_{j=1}^{n} \tilde{N}_j.$$

Put

$$K_0 = 1$$

and

$$K_i = K_{i-1} + \tilde{N}_i,\ i = 1, 2, \ldots$$

Introduce the following random variables

$$\tilde{Z}_i = \sum_{k=K_{i-1}}^{K_i - 1} Z_k,\ i = 1, 2, \ldots$$

(void sum is 0), then

$$\sum_{i=1}^{N_n} Z_i = \sum_{i=1}^{n} \tilde{Z}_i,$$

moreover $\{\tilde{Z}_i\}$ are i.i.d. random variables with

$$\mathbf{E}\{\tilde{Z}_i\} = \mathbf{E}\{Z_i\},\ Var(\tilde{Z}_i) = \mathbf{E}\{Z_i^2\}$$

and with a universal constant D

$$\mathbf{E}\{|\tilde{Z}_i - \mathbf{E}\{\tilde{Z}_i\}|^3\} \leq D\mathbf{E}\{|Z_i|^3\}.$$

Thus apply Lemma 5.8 in Devroye, Györfi (1985) for $\{\tilde{Z}_i - \mathbf{E}\tilde{Z}_1\}$

$$\begin{aligned} \sup_a \left| \mathbf{E} \left| \frac{\sum_{i=1}^n \tilde{Z}_i - n\mathbf{E}\tilde{Z}_1}{\sqrt{nVar(\tilde{Z}_1)}} - a \right| - \mathbf{E}|N - a| \right| \\ \leq \frac{c}{n^{1/2}} \frac{\mathbf{E}\{|\tilde{Z}_i - \mathbf{E}\{\tilde{Z}_i\}|^3\}}{Var(\tilde{Z}_1)^{3/2}} \end{aligned}$$

with some universal constant c, and

$$\mathbf{E}|N - a| = |a|\mathbf{P}\{|N| \leq |a|\} + \sqrt{\frac{2}{\pi}} e^{-a^2/2}.$$

Collecting these inequalities we get the lemma. □

Proof of Theorem 3.3.

The statement of Theorem 3.3 means that

$$n^{1/2}\left(J_n - J_n^*\right)/\sqrt{L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1), \tag{3.17}$$

where

$$J_n^* = \sum_{j=1}^{m_n} |\tau(A_{nj})|.$$

In order to prove this, note that (3.12) and (3.17) differ only in the centering constants. If we copy the previous proof with the other centering constant then we should only show that

$$n^{1/2}\sum_{j=1}^{l_n} |\mathbf{E}|\tilde{\tau}_n(A_{nj})| - |\tau(A_{nj})|| \to 0.$$

Lemma 3.7 implies that for a set A

$$\begin{aligned}
0 &\le n^{1/2}(\mathbf{E}|\tilde{\tau}_n(A)| - |\tau(A)|) \\
&= \sqrt{\mu(A)}\left(\mathbf{E}\left|\frac{n(\tilde{\tau}_n(A)-\tau(A))}{\sqrt{n\mu(A)}} - \frac{n\tau(A)}{\sqrt{n\mu(A)}}\right| \right. \\
&\qquad \left. - \left|\frac{n\tau(A)}{\sqrt{n\mu(A)}}\right|\right) \\
&\le \sqrt{\mu(A)}\sqrt{\frac{2}{\pi}}e^{-\frac{n\tau(A)^2}{2\mu(A)}} + \frac{A_1}{\sqrt{n}},
\end{aligned}$$

therefore by Assumption 3.1 and condition (3.2)

$$\begin{aligned}
&n^{1/2}\sum_{j=1}^{l_n} |\mathbf{E}|\tilde{\tau}_n(A_{nj})| - |\tau(A_{nj})|| \\
&\le \sqrt{\frac{2}{\pi}}\sum_{j=1}^{l_n}\sqrt{\mu(A_{nj})}e^{-\frac{n\tau(A_{nj})^2}{2\mu(A_{nj})}} + \frac{A_1 l_n}{\sqrt{n}} \\
&\le \sqrt{\frac{2}{\pi}}\sum_{j=1}^{l_n}\sqrt{\mu(A_{nj})}e^{-nc^2\mu(A_{nj})/2} + \frac{A_1 l_n}{\sqrt{n}} \\
&\le \sqrt{\frac{2}{\pi}}\max_z \sqrt{z}e^{-z}\frac{\sqrt{2}l_n}{c\sqrt{n}} + \frac{A_1 l_n}{\sqrt{n}} \to 0.
\end{aligned}$$

□

Under the conditions of Theorem 3.3, we can remove m_n from the upper bound in (3.4).

THEOREM 3.4. *(Pintér (2000)) For any distribution of (X, Y) and for n large enough for the estimate $L_n^{(R)}$ of the error probability of a partitioning rule satisfying the conditions of Theorem 3.3*

$$R_n^* - \mathbf{E}L_n^{(R)} \leq \frac{1.18}{\sqrt{n}}.$$

PROOF

From the result on the asymptotic normality

$$\mathbf{P}\{n^{1/2}|L_n^{(R)} - R_n^*| > \epsilon\} \to 1 - 2\Phi\left(\frac{\epsilon}{\sqrt{\frac{1}{4}L^*(4 - L^*)}}\right),$$

where Φ is the standard normal distribution function. Since $R_n^* - \mathbf{E}L_n^{(R)}$ is not random

$$\begin{aligned}
I_{\{\sqrt{n}|R_n^* - \mathbf{E}L_n^{(R)}| > \epsilon\}} &= \mathbf{P}\{\sqrt{n}|R_n^* - \mathbf{E}L_n^{(R)}| > \epsilon\} \\
&\leq \mathbf{P}\{\sqrt{n}(|L_n^{(R)} - \mathbf{E}L_n^{(R)}| + |L_n^{(R)} - R_n^*|) > \epsilon\} \\
&\leq \mathbf{P}\{\sqrt{n}|L_n^{(R)} - \mathbf{E}L_n^{(R)}| > \epsilon/2\} \\
&\quad + \mathbf{P}\{\sqrt{n}|L_n^{(R)} - R_n^*| > \epsilon/2\}.
\end{aligned}$$

For the first term we can apply McDiarmid's inequality with $Z_i = (X_i, Y_i)$ and $A = \mathcal{R}^d \times \{0, 1\}$.

LEMMA 3.8. *(McDiarmid (1989)) Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set A, and assume that the measurable function $F : A^n \to \mathcal{R}$ satisfies*

$$\sup_{z_1, \ldots, z_n, z_i'} |F(z_1, \ldots, z_i, \ldots, z_n) - F(z_1, \ldots, z_i', \ldots, z_n)| \leq c_i, \; 1 \leq i \leq n.$$

Then for all $\epsilon > 0$

$$\mathbf{P}\{|F(Z_1, \ldots, Z_n) - \mathbf{E}F(Z_1, \ldots, Z_n)| > \epsilon\} \leq 2e^{-2\epsilon^2/(\sum_{i=1}^n c_i^2)}.$$

It can be easily seen that in the case of the resubstitution estimate for partitioning rule $c_i = \frac{1}{n}$. Thus for all n and $\epsilon > 0$,

$$\mathbf{P}\{n^{1/2}|L_n^{(R)} - \mathbf{E}L_n^{(R)}| > \epsilon\} \leq 2e^{-2\epsilon^2}.$$

Therefore

$$\limsup_{n\to\infty} I_{\{\sqrt{n}|R_n^* - \mathbf{E}L_n^{(R)}|>\epsilon\}} \le 2e^{-\epsilon^2/2} + 1 - 2\Phi\left(\frac{\epsilon}{\sqrt{L^*(4-L^*)}}\right).$$

Obviously there exists such ϵ_0 for which the righthand side of the inequality is strictly smaller than 1, and then

$$I_{\{\sqrt{n}|R_n^* - \mathbf{E}L_n^{(R)}|>\epsilon_0\}} = 0$$

for sufficiently large n. Using the trivial upper bound

$$\Phi(\epsilon/\sqrt{L^*(4-L^*)}) \ge 1/2,$$

$\epsilon_0 = \sqrt{2\ln 2} \approx 1.18$ is a valid choice. □

The following result is a simple consequence of Corollary 3.1 and Theorem 3.3.

THEOREM 3.5. *(Pintér (2000)) Under the conditions of Theorem 3.3*

$$n^{1/2}\left(L_n - L_n^{(R)}\right)/\sqrt{\frac{1}{4}L^*(4-L^*)} \xrightarrow{\mathcal{D}} N(0,1).$$

Thus the resubstitution estimate is a good estimate of the true error probability of the partitioning classification rule, and using the result of the theorem a good estimate can be given to the probability that the difference between the resubstitution estimate and the error probability is greater than $\epsilon/\sqrt{n}$.

It is necessary to have some condition such as Assumption 3.1 that ensures that labels are far from being uniformly chosen.

PROPOSITION 3.2. *(Pintér (2000)) Assume that $\mu(A_{nj}) = 2\nu(A_{nj})$ for all cells A_{nj} of the partition, then*

$$\sqrt{n}\mathbf{E}(L_n - L_n^{(R)}) \ge \frac{1}{4}\sqrt{m_n}.$$

Thus in that case $\sqrt{n}\mathbf{E}(L_n - L_n^{(R)}) \to \infty$ as $n \to \infty$.

PROOF. Since $L_n \geq R_n^*$, and under the conditions of the proposition $R_n^* = 1/2$, we have to show, that

$$\mathbf{E}L_n^{(R)} \leq \frac{1}{2} - \frac{1}{4}\sqrt{\frac{m_n}{n}}.$$

$$\begin{aligned}
&\mathbf{E}L_n^{(R)} \\
&= \mathbf{E}\sum_{j=1}^{m_n} \min\{\nu_n(A_{nj}), \mu_n(A_{nj}) - \nu_n(A_{nj})\} \\
&= \mathbf{E}\sum_{j=1}^{m_n} \left(\frac{\mu_n(A_{nj})}{2} + \min\{\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}, \frac{\mu_n(A_{nj})}{2} - \nu_n(A_{nj})\}\right) \\
&= \mathbf{E}\sum_{j=1}^{m_n} \left(\frac{\mu_n(A_{nj})}{2} - \left|\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}\right|\right) \\
&= \frac{1}{2} - \mathbf{E}\sum_{j=1}^{m_n} \left|\nu_n(A_{nj}) - \frac{\mu_n(A_{nj})}{2}\right| \\
&= \frac{1}{2} - \sum_{j=1}^{m_n} \frac{1}{n}\mathbf{E}\left|n\nu_n(A_{nj}) - \frac{n\mu_n(A_{nj})}{2}\right| \\
&= \frac{1}{2} - \\
&\qquad \sum_{j=1}^{m_n} \frac{1}{n} \sum_{k=0}^{n} \mathbf{E}\left(\left|n\nu_n(A_{nj}) - \frac{n\mu_n(A_{nj})}{2}\right| |n\mu_n(A_{nj}) = k\right) \mathbf{P}(n\mu_n(A_{nj}) = k)
\end{aligned} \tag{3.18}$$

If $n\mu_n(A_{nj}) = k$, then $n\nu_n(A_{nj})$ is binomial with parameters k and $1/2$, therefore we are interested in the expected value $\mathbf{E}\left|Binom(k, \frac{1}{2}) - \frac{k}{2}\right|$. Let Z_i, $i = 1, \dots, k$ be random variables with distribution $\mathbf{P}(Z_i = 1) = \mathbf{P}(Z_i = -1) = 1/2$, then $\left|Binom(k, \frac{1}{2}) - \frac{k}{2}\right| = \frac{1}{2}\left|\sum_{i=1}^k Z_i\right|$. Now we use the following inequality.

LEMMA 3.9. *(Khinchine's inequality, see e.g. Devroye and Györfi (1985)) Let $Z_1, \dots, Z_k$ be independent random variables taking the values $+1$ and -1 with equal probability, and let $a_1, \dots, a_k$ be real numbers. Then there exist positive constants B_p and C_p depending upon $p > 0$ only such that*

$$B_p\sqrt{\sum_{i=1}^k a_i^2} \leq \mathbf{E}^{1/p}\left(\left|\sum_{i=1}^k a_i Z_i\right|^p\right) \leq C_p\sqrt{\sum_{i=1}^k a_i^2}.$$

The following values are optimal:

$$B_p = \begin{cases} 2^{1/2-1/p}, & \text{if } 0 < p \le p_0 \\ 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, & p_0 \le p < 2 \\ 1, & \text{if } 2 \le p \end{cases}$$

$$C_p = \begin{cases} 1, & \text{if } 0 < p \le 2 \\ 2^{1/2}\left(\frac{\Gamma((p+1)/2)}{\sqrt{\pi}}\right)^{1/p}, & 2 \le p \end{cases}$$

where $\Gamma(\cdot)$ *is the classical gamma-function extending the factorial and* $p_0 = 1.84742\cdots$ *is the solution in* $(1,2)$ *of* $\Gamma((p+1)/2) = \Gamma(3/2)$.

Applying this inequality to the above Z_i and $p = 1$, we get

$$\frac{1}{2\sqrt{2}}\sqrt{k} \le \mathbf{E}\left|Binom\left(k, \frac{1}{2}\right) - \frac{k}{2}\right| = \frac{1}{2}\mathbf{E}\left|\sum_{i=1}^{k} Z_i\right| \le \frac{1}{2}\sqrt{k}.$$

Thus from (3.18)

$$\begin{aligned}
&\mathbf{E}L_n^{(R)} \\
&\le \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}\sum_{k=0}^{n}\frac{1}{2\sqrt{2}}\sqrt{k}\binom{n}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-k} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{n}n\mu(A_{nj})\sum_{k=1}^{n}\frac{1}{2\sqrt{2}}\frac{1}{\sqrt{k}}\binom{n-1}{k-1}\mu(A_{nj})^{k-1}(1-\mu(A_{nj}))^{(n-1)-(k-1)} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\frac{1}{n}n\mu(A_{nj})\sum_{k=0}^{n-1}\frac{1}{\sqrt{k+1}}\binom{n-1}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-1-k} \\
&\le \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\mu(A_{nj})\frac{1}{\sqrt{\sum_{k=0}^{n-1}(k+1)\binom{n-1}{k}\mu(A_{nj})^k(1-\mu(A_{nj}))^{n-1-k}}} \\
&\quad \text{(from Jensen's inequality)} \\
&= \frac{1}{2} - \sum_{j=1}^{m_n}\frac{1}{2\sqrt{2}}\mu(A_{nj})\frac{1}{\sqrt{(n-1)\mu(A_{nj})+1}} \\
&\le \frac{1}{2} - m_n\frac{1}{2\sqrt{2}}\frac{\frac{1}{m_n}\sum_{j=1}^{m_n}\mu(A_{nj})}{\sqrt{(n-1)\frac{1}{m_n}\sum_{j=1}^{m_n}\mu(A_{nj})+1}} \\
&\quad \text{(from Jensen's inequality)} \\
&= \frac{1}{2} - \frac{1}{2\sqrt{2}}\frac{1}{\sqrt{(n-1)/m_n+1}} \\
&\le \frac{1}{2} - \frac{1}{4}\sqrt{\frac{m_n}{n}}
\end{aligned}$$

□

3.2.2 Deleted Estimate

The deleted estimate or cross-validation attempts to avoid the bias present in the resubstitution estimate. The method deletes (X_i, Y_i) from the training data and creates a classifier g_{n-1} using the remaining $n-1$ pairs. It tests for an error on (X_i, Y_i), and repeats this procedure for all n pairs of the training data D_n. Formally denote the training set with (X_i, Y_i) deleted by

$$D_{n,i} = ((X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)).$$

Then define

$$L_n^{(D)} = \frac{1}{n} \sum_{i=1}^{n} I_{\{g_{n-1}(X_i, D_{n,i}) \neq Y_i\}}.$$

Clearly, the deleted estimate is almost unbiased in the sense that

$$\mathbf{E} L_n^{(D)} = \mathbf{E} L_{n-1}.$$

Let g_{n-1} be the partitoning classification rule for $n-1$ samples, using partition $\mathcal{P}_n$. Concerning the deleted error estimate for partitioning rule the following is known (see Sec. 24.5 in Devroye, Györfi and Lugosi (1996)):

$$\mathbf{E}\left\{ \left(L_n^{(D)} - L_n \right)^2 \right\} \leq \frac{1 + 6/e}{n} + \frac{6}{\sqrt{\pi(n-1)}}.$$

The resubstitution estimate for any partitioning classification rule is smaller than the deleted estimate:

$$L_n^{(D)} \geq L_n^{(R)}, \tag{3.19}$$

since if (X_i, Y_i) is a mistake w.r.t. $L_n^{(R)}$, i.e. $Y_i \neq g_n(X_i)$, then the label Y_i is in the minority among the labels of the data falling into its cell. Then, of course, Y_i is in the minority of its cell w.r.t. $D_{n,i} = D_n \setminus \{X_i, Y_i\}$, which implies that (X_i, Y_i) is a mistake w.r.t. $L_n^{(D)}$.

The deleted estimate for partitioning classification rule is asymptotically normal under similar conditions as the resubstitution estimate:

Theorem 3.6. *(Pintér (2000)) Under the conditions of Theorem 3.3,*

$$n^{1/2} \left(L_n - L_n^{(D)} \right) / \sqrt{\frac{1}{4} L^*(4 - L^*)} \xrightarrow{\mathcal{D}} N(0,1).$$

Proof. It is easy to see that if U_n and V_n are random variables and $U_n \xrightarrow{\mathcal{D}} N(0,1)$ and $V_n \to 0$ in probability, then $U_n + V_n \xrightarrow{\mathcal{D}} N(0,1)$. So, since the conditions are the same as for Theorem 3.5, it suffices to show that

$$\sqrt{n}\left(L_n^{(D)} - L_n^{(R)}\right) \to 0$$

in probability as $n \to \infty$.

If for a nonnegative random variable Z_n, $\mathbf{E}Z_n \to 0$, then $Z_n \to 0$ in probability. Thus, since from (3.19) $L_n^{(D)} - L_n^{(R)} \geq 0$, we have to prove that

$$\sqrt{n}\mathbf{E}(L_n^{(D)} - L_n^{(R)}) \to 0 \quad (n \to \infty).$$

(X_i, Y_i) is a mistake w.r.t. $L_n^{(D)}$, i.e. $g_{n-1}(X_i, D_{n,i}) \neq Y_i$ if Y_i is in the minority of its cell w.r.t. $D_{n,i} = D_n \setminus \{X_i, Y_i\}$. If (X_i, Y_i) is a mistake, then each pair with the same label in that cell is also a mistake. The number of ones in a cell is $n\nu_n(A_{nj})$ and the number of zeros is $n(\mu_n(A_{nj}) - \nu_n(A_{nj}))$. Thus, for the partitioning rule we can write the deleted estimate as:

$$\begin{aligned} L_n^{(D)} &= \sum_{j=1}^{m_n} \Big(\nu_n(A_{nj}) I_{\{n\nu_n(A_{nj})-1 \leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}} \\ &\qquad + (\mu_n(A_{nj}) - \nu_n(A_{nj})) I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1 < n\nu_n(A_{nj})\}}\Big). \end{aligned}$$

The resubstitution estimate can be written in a similar form using (3.3)

$$\begin{aligned} L_n^{(R)} &= \sum_{j=1}^{m_n} \Big(\nu_n(A_{nj}) I_{\{n\nu_n(A_{nj}) \leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}} \\ &\qquad + (\mu_n(A_{nj}) - \nu_n(A_{nj})) I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj})) < n\nu_n(A_{nj})\}}\Big). \end{aligned}$$

Therefore

$$\begin{aligned} &\sqrt{n}\mathbf{E}(L_n^{(D)} - L_n^{(R)}) \\ &= \sqrt{n}\mathbf{E}\Bigg\{\sum_{j=1}^{m_n} \Big(\nu_n(A_{nj}) I_{\{n\nu_n(A_{nj})-1 \leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}} \\ &\qquad + (\mu_n(A_{nj}) - \nu_n(A_{nj})) I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1 < n\nu_n(A_{nj})\}} \\ &\qquad - \Big(\nu_n(A_{nj}) I_{\{n\nu_n(A_{nj}) \leq n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}} \\ &\qquad + (\mu_n(A_{nj}) - \nu_n(A_{nj})) I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj})) < n\nu_n(A_{nj})\}}\Big)\Big)\Bigg\} \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})-1\le n(\mu_n(A_{nj})-\nu_n(A_{nj}))<n\nu_n(A_{nj})\}}\right.\\
&\qquad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))-1<n\nu_n(A_{nj})\le n(\mu_n(A_{nj})-\nu_n(A_{nj}))\}}\Big)\right\}\\
&= \sqrt{n}\mathbf{E}\left\{\sum_{j=1}^{m_n}\Big(\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})=n(\mu_n(A_{nj})-\nu_n(A_{nj}))+1\}}\right.\\
&\qquad\qquad \left.+(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))=n\nu_n(A_{nj})\}}\Big)\right\}\\
&= \sqrt{n}\sum_{j=1}^{m_n}\Big(\mathbf{E}\left\{\nu_n(A_{nj})I_{\{n\nu_n(A_{nj})=n(\mu_n(A_{nj})-\nu_n(A_{nj}))+1\}}\right\}\\
&\qquad\qquad +\mathbf{E}\left\{(\mu_n(A_{nj})-\nu_n(A_{nj}))I_{\{n(\mu_n(A_{nj})-\nu_n(A_{nj}))=n\nu_n(A_{nj})\}}\right\}\Big)\\
&= \sqrt{n}\sum_{j=1}^{m_n}(T_{1,nj}+T_{2,nj}) \qquad (3.20)
\end{aligned}
$$

Consider the first term of the sum for the cell $A = A_{nj}$:

$$
\begin{aligned}
T_{1,n} &= \mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}\right\}\\
&= \sum_{k=0}^{n}\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=k\right\}\mathbf{P}\{n\mu_n(A)=k\}\\
&= \sum_{l=0}^{n/2}\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=2l+1\right\}\\
&\qquad \mathbf{P}\{n\mu_n(A)=2l+1\}
\end{aligned}
$$

Since if the number of pairs in a cell is even, then the indicator is trivially 0. $n\mu_n(A)$ is binomial with parameters n and $\mu(A)$, so

$$
\mathbf{P}\{n\mu_n(A)=2l+1\} = \binom{n}{2l+1}\mu(A)^{2l+1}(1-\mu(A))^{n-(2l+1)}.
$$

If $\mu_n(A) = 2l+1$, then $n\nu_n(A)$ is binomial with parameters $2l+1$ and $\nu(A)/\mu(A)$, thus

$$
\begin{aligned}
&\mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=n(\mu_n(A)-\nu_n(A))+1\}}|n\mu_n(A)=2l+1\right\}\\
&= \mathbf{E}\left\{\nu_n(A)I_{\{n\nu_n(A)=l+1\}}\right\}\\
&= \frac{l+1}{n}\mathbf{P}\{n\nu_n(A)=l+1\}\\
&= \frac{l+1}{n}\binom{2l+1}{l+1}\left(\frac{\nu(A)}{\mu(A)}\right)^{l+1}\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^{l}
\end{aligned}
$$

$$= \frac{2l+1}{n}\binom{2l}{l}\left(\frac{\nu(A)}{\mu(A)}\right)^{l+1}\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^{l}.$$

Therefore

$$\begin{aligned}
T_{1,n} &= \sum_{l=0}^{n/2}\frac{2l+1}{n}\binom{2l}{l}\nu(A)^{l+1}\left(\mu(A)-\nu(A)\right)^{l}\frac{1}{\mu(A)^{2l+1}} \\
&\qquad \binom{n}{2l+1}\mu(A)^{2l+1}(1-\mu(A))^{n-(2l+1)} \\
&= \sum_{l=0}^{n/2}\frac{2l+1}{n}\binom{2l}{l}\nu(A)^{l+1}(\mu(A)-\nu(A))^{l}\binom{n}{2l+1}(1-\mu(A))^{n-(2l+1)} \\
&\leq \sum_{l=0}^{n/2}2^{2l}\nu(A)^{l+1}(\mu(A)-\nu(A))^{l}\binom{n}{2l}(1-\mu(A))^{n-(2l+1)} \\
&\qquad\qquad \left(\text{since } \tfrac{2l+1}{n}\tbinom{2l}{l}\tbinom{n}{2l+1}=\tbinom{2l}{l}\tbinom{n-1}{2l}\leq 2^{2l}\tbinom{n}{2l}\right) \\
&= \frac{\nu(A)}{1-\mu(A)}\sum_{l=0}^{n/2}2^{2l}\nu(A)^{l}(\mu(A)-\nu(A))^{l}\binom{n}{2l}(1-\mu(A))^{n-2l} \\
&= \frac{\nu(A)}{1-\mu(A)}\sum_{l=0}^{n/2}\binom{n}{2l}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{2l}(1-\mu(A))^{n-2l} \\
&\leq \frac{\nu(A)}{1-\mu(A)}\sum_{k=0}^{n}\binom{n}{k}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{k}(1-\mu(A))^{n-k} \\
&= \frac{\nu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^{n} \\
&\leq \frac{\mu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^{n}. \qquad (3.21)
\end{aligned}$$

Similarly, we can get the same upper bound for the second term in (3.20):

$$\begin{aligned}
T_{2,n} &= \mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}\right\} \\
&= \sum_{k=0}^{n}\mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}|n\mu_n(A)=k\right\} \\
&\qquad \mathbf{P}\{n\mu_n(A)=k\} \\
&= \sum_{l=0}^{n/2}\mathbf{E}\left\{(\mu_n(A)-\nu_n(A))I_{\{n(\mu_n(A)-\nu_n(A))=n\nu_n(A)\}}|n\mu_n(A)=2l\right\} \\
&\qquad \mathbf{P}\{n\mu_n(A)=2l\} \\
&= \sum_{l=0}^{n/2}\frac{l}{n}\binom{2l}{l}\left(\frac{\nu(A)}{\mu(A)}\right)^{l}\left(\frac{\mu(A)-\nu(A)}{\mu(A)}\right)^{l}\binom{n}{2l}\mu(A)^{2l}(1-\mu(A))^{n-2l}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=0}^{n/2} \frac{l}{n} \binom{2l}{l} \nu(A)^l (\mu(A)-\nu(A))^l \binom{n}{2l} (1-\mu(A))^{n-2l} \\
&\le \frac{1}{n} \sum_{l=0}^{n/2} 2l 2^{2l} \nu(A)^l (\mu(A)-\nu(A))^l \binom{n}{2l} (1-\mu(A))^{n-2l} \\
&= \frac{1}{n} \sum_{l=0}^{n/2} 2l \binom{n}{2l} \left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{2l} (1-\mu(A))^{n-2l} \\
&\le \frac{1}{n} \sum_{k=0}^{n} k \binom{n}{k} \left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{k} (1-\mu(A))^{n-k}.
\end{aligned}
$$

The sum in the righthand side is the expected value of the binomial distribution with parameters n and $2\sqrt{\nu(A)(\mu(A)-\nu(A))}/(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A))$ if we normalize with the sum of $2\sqrt{\nu(A)(\mu(A)-\nu(A))}$ and $1-\mu(A)$. Thus

$$
\begin{aligned}
T_{2,n} &\le \frac{1}{n}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n \\
&\qquad \frac{1}{\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n} \\
&\qquad \sum_{k=0}^{n} k \binom{n}{k} \left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}\right)^{k} (1-\mu(A))^{n-k} \\
&= \frac{1}{n}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n \\
&\qquad \mathbf{E}\left\{Binom\left(n, \frac{2\sqrt{\nu(A)(\mu(A)-\nu(A))}}{2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)}\right)\right\} \\
&= \frac{1}{n}\frac{n2\sqrt{\nu(A)(\mu(A)-\nu(A))}}{2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)} \\
&\qquad \left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n \\
&\le \frac{\mu(A)}{1-\mu(A)}\left(2\sqrt{\nu(A)(\mu(A)-\nu(A))}+1-\mu(A)\right)^n.
\end{aligned}
$$

From this upper bound and from (3.20) and (3.21)

$$
\begin{aligned}
&\sqrt{n}\mathbf{E}(L_n^{(D)}-L_n^{(R)}) \\
&\le \sqrt{n}\sum_{j=1}^{m_n} 2\frac{\mu(A_{nj})}{1-\mu(A_{nj})}\left(2\sqrt{\nu(A_{nj})(\mu(A_{nj})-\nu(A_{nj}))}+1-\mu(A_{nj})\right)^n \\
&\le \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj}) e^{-n\left(\mu(A_{nj})-2\sqrt{\nu(A_{nj})(\mu(A_{nj})-\nu(A_{nj}))}\right)}
\end{aligned}
$$

$$= \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-n\mu(A_{nj})\left(1-2\sqrt{\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1-\frac{\nu(A_{nj})}{\mu(A_{nj}}\right)}\right)},$$

where the second inequality comes from the fact that $1-z \le e^{-z}$ if $z>0$, and clearly $\mu(A_{nj}) - 2\sqrt{\nu(A_{nj})(\mu(A_{nj})-\nu(A_{nj}))} > 0$, and we can assume that $\mu(A_{nj}) < 1/2$.

Assumption 3.1 means that

$$\left|\frac{1}{2} - \frac{\nu(A_{nj})}{\mu(A_{nj})}\right| > \frac{c}{2},$$

and because of this

$$\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1-\frac{\nu(A_{nj})}{\mu(A_{nj})}\right) \le \frac{1}{4} - \frac{c^2}{4}.$$

Thus, if we denote $1-2\sqrt{\frac{1}{4}-\frac{c^2}{4}} = 1-\sqrt{1-c^2}$ by $\delta(c)$ then

$$\begin{aligned}
\sqrt{n}\mathbf{E}(L_n^{(D)} - L_n^{(R)}) &\le \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-n\mu(A_{nj})\left(1-2\sqrt{\frac{\nu(A_{nj})}{\mu(A_{nj})}\left(1-\frac{\nu(A_{nj})}{\mu(A_{nj}}\right)}\right)} \\
&\le \sqrt{n}\sum_{j=1}^{m_n} 2\mu(A_{nj})e^{-\delta(c)n\mu(A_{nj})} \\
&= \sqrt{n}\sum_{j=1}^{m_n} \frac{2}{\delta(c)n}\delta(c)n\mu(A_{nj})e^{-\delta(c)n\mu(A_{nj})} \\
&\le \sqrt{n}\sum_{j=1}^{m_n} \frac{2}{\delta(c)n}\max_z ze^{-z} \\
&\le \sqrt{n}m_n\frac{2}{\delta(c)n} = \frac{2m_n}{\delta(c)\sqrt{n}} \to 0
\end{aligned}$$

because of condition (3.2). □

Thus, similarly to the resubstitution estimate, the deleted estimate is a good estimate of the true error probability of the partitioning classification rule, and using the result of the theorem a good estimate can be given to the probability that the difference between the deleted estimate and the error probability is greater than $\epsilon/\sqrt{n}$.

Chapter 4

Scale-sensitive dimensions and skeleton estimates

In this chapter the classical binary classification (pattern recognition) problem is investigated when it is known in advance that the posterior probability function (or regression function) belongs to some class of functions. Very often, apart from the training sequence, some prior information is available about the joint distribution of (X, Y). For example, in some applications with $\mathcal{X} = \mathcal{R}^d$, it is known that η^* is a monotone function in all components of x. In other situations it may be known that η^* is a smooth function. In the basic PAC-learning setup Blumer et al. (1989), η^* is known to be the indicator function of one of the sets in a given class of sets. We assume throughout that η^* is a member of a known class of functions $\mathcal{F}$. Here we are interested in how this extra information can be exploited to obtain small probabilities of error. The methods we introduce and analyze here are based on minimizing the empirical risk over a carefully selected "skeleton" of the class of regression functions. The skeletons are coverings of the class based on metrics, especially fitted for classification. A new scale-sensitive dimension is introduced which is more suitable for the studied classification problem than other, previously defined, dimension measures. This fact is demonstrated by performance bounds for the skeleton estimates in terms of the new dimension.

4.1 Methods of empirical risk minimization

Perhaps the most natural way of exploiting the knowledge that η^* is in a known class of functions is the following: form the class of classifiers determined by the functions in $\mathcal{F}$, that is,

$$\left\{ g : g(x) = I_{\{\eta(x) \geq 1/2\}}, \eta \in \mathcal{F} \right\}.$$

The Bayes classifier is clearly in this class. Then we may select a member of the class by minimizing the empirical error $\hat{L}_n(g) = \frac{1}{n}\sum_{i=1}^n I_{\{g(X_i)\neq Y_i\}}$. Then an inequality of Vapnik and Chervonenkis (1974) (see also Anthony and Shawe-Taylor (1993)) implies that if $\tilde{g}_n$ is the selected classifier, then

$$\mathbf{E}\{L(\tilde{g}_n)\} - L^* \leq K \max\left(\sqrt{\frac{L^* \log\left(\mathbf{E}\left\{S_{\mathcal{F}}(X_1^n)\right\}\right)}{n}}, \frac{\log\left(\mathbf{E}\left\{S_{\mathcal{F}}(X_1^n)\right\}\right)}{n}\right). \tag{4.1}$$

Here, and in the rest of the chapter, K denotes a universal constant. It is well-known (Sauer (1972)) that the shatter coefficients may be uniformly bounded as

$$\sup_{x_1,\ldots,x_n} S_{\mathcal{F}}(x_1^n) \leq \left(\frac{ne}{\mathcal{V}_{\mathcal{F}}}\right)^{\mathcal{V}_{\mathcal{F}}}.$$

Therefore, we have

$$\mathbf{E}\{L(\tilde{g}_n)\} - L^* \leq K \max\left(\sqrt{\frac{L^*\mathcal{V}_{\mathcal{F}} \log n}{n}}, \frac{\mathcal{V}_{\mathcal{F}} \log n}{n}\right).$$

The method of minimizing the empirical error is extremely robust in the sense that the above upper bound remains true in a completely distribution-free setting if we replace L^* by the probability of error $\underline{L} = \inf_{\eta\in\mathcal{F}} L(\eta)$ of the best classifier in the class. In other words, if $\eta^* \notin \mathcal{F}$, the method still works reasonably well.

On the other hand, this result is optimal (up to a logarithmic factor) in a minimax sense: for *any* classification rule g_n, there exists a distribution of (X, Y) such that

$$\mathbf{E}L(g_n) - \underline{L} \geq K \max\left(\sqrt{\frac{\underline{L}\mathcal{V}_{\mathcal{F}}}{n}}, \frac{\mathcal{V}_{\mathcal{F}}}{n}\right), \tag{4.2}$$

see Devroye. Györfi and Lugosi (1996), Devroye and Lugosi (1995). However, if we exploit the additional information that $\eta^* \in \mathcal{F}$, we may be able to define classifiers with improved performance guarantees. It is precisely the goal of this chapter to explore this direction.

For simplicity, we have given bounds for the expected value of the probability of error $L(\tilde{g}_n)$. Alternatively, we may rephrase these results in terms of sample-size bounds. For example, another consequence of the above-mentioned result of Vapnik and Chervonenkis is that for any $\epsilon, \delta > 0$, $\mathbf{P}\{L(\tilde{g}_n) - L^* > \epsilon\} < \delta$ if the sample size n is at least

$$K \max\left(\frac{L^*\mathcal{V}_{\mathcal{F}}}{\epsilon^2}\log\frac{\mathcal{V}_{\mathcal{F}}}{\epsilon} + \frac{L^*}{\epsilon^2}\log\frac{1}{\delta}, \frac{\mathcal{V}_{\mathcal{F}}}{\epsilon}\log\frac{\mathcal{V}_{\mathcal{F}}}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right).$$

Note that unless $L^* < \epsilon$, the maximum is achieved by the first term.

The disadvantage of this method of empirical error minimization is that it only uses a part of the available information. Apart from the form of the classifiers defined by the functions in $\mathcal{F}$, no other information about the shape of these functions is exploited. For example, if $\mathcal{F}$ is the class of Lipschitz functions $\eta : [0,1] \to [0,1]$ (with Lipschitz constant 1), then clearly the class of classifiers defined by $\mathcal{F}$ is too large, and we have $\mathcal{V}_{\mathcal{F}} = \infty$.

Another approach to the classification problem is to directly estimate the regression function η^* by a function η_n'. For the probability of error of the corresponding classifier $g_n'(x) = I_{\{\eta_n'(x) \ge 1/2\}}$, we have from (1.11) that

$$\mathbf{E}L(g_n') - L^* \le 2\sqrt{\mathbf{E}\left\{|\eta^*(X) - \eta_n'(X)|^2\right\}}. \tag{4.3}$$

Observing that $\eta^*(x) = \mathbf{E}\{Y|X = x\}$ and therefore for any η

$$\mathbf{E}\left\{(\eta(X) - \eta^*(X))^2\right\} = \mathbf{E}\left\{(\eta(X) - Y)^2\right\} - \mathbf{E}\left\{(\eta^*(X) - Y)^2\right\},$$

leads to the idea to use the least squares regression estimate, that minimize the empirical L_2 error

$$\frac{1}{n}\sum_{i=1}^{n}(\eta(X_i) - Y_i)^2 \qquad \eta \in \mathcal{F},$$

and consider the classifier g_n' corresponding to the minimizing function η_n'. Then, exploiting that the regression function η^* is in $\mathcal{F}$, we may use a result of Lee, Bartlett, and Williamson (1996, Theorem 7) to conclude that for every $\epsilon > 0$,

$$\mathbf{E}L(g_n') - L^* \le K\left(\sup_{x_1,\ldots,x_n}\sqrt{\frac{\log N_1(\epsilon, x_1^n, \mathcal{F})}{n}} + \epsilon\right). \tag{4.4}$$

In Section 4.4 we present an alternative method with improved performance guarantees.

As proved by Alon, Ben-David, Cesa-Bianchi and Haussler (1993), the covering numbers in the above bound may be bounded in terms of the scale-sensitive dimension P_γ. They show that for all $\epsilon \in (0,1)$,

$$\sup_{x_1,\ldots,x_n} N_1(\epsilon, x_1^n, \mathcal{F}) \le 2\left(\frac{4n}{\epsilon^2}\right)^{P_{\epsilon/4}\log(2en/(P_{\epsilon/4}))}. \tag{4.5}$$

This inequality was subsequently improved by Bartlett and Long (1995).

Remark.
Another method of exploiting the information $\eta^* \in \mathcal{F}$ is *maximum likelihood*. Here one selects a function from $\mathcal{F}$ by maximizing the likelihood

$$\prod_{i=1}^n \eta(X_i)^{Y_i}(1-\eta(X_i))^{1-Y_i}.$$

It is shown in Devroye, Györfi and Lugosi (1996, Chapter 15) that maximum likelihood works well whenever the *bracketing metric entropy* of $\mathcal{F}$ is finite. □

Remark.
One might alternatively consider instead minimizing the empirical L_1 error

$$\frac{1}{n}\sum_{i=1}^n |\eta(X_i) - Y_i| \qquad \eta \in \mathcal{F}.$$

However, note that minimizing the empirical L_1 error may lead to catastrophically bad classifiers. To see this, consider the following simple example: let X be uniformly distributed on $[0,1]$, and let $\eta^*(x) = 1/4$ for all $x \in [0,1]$. Assume that $\mathcal{F} = \{\eta^*, \eta\}$, where $\eta(x) = I_{\{x>1/2\}}$. Then it is easy to see that

$$\mathbf{E}\{|\eta(X) - Y|\} = \frac{1}{2} < \mathbf{E}\{|\eta^*(X) - Y|\} = \frac{5}{8},$$

so the empirical L_1 minimizer will choose η instead of η^* with large probability. The probability of error of the obtained classifier does not even converge to L^* as $n \to \infty$!

4.2 Dimensions

So far, we have seen two different methods—one based on minimizing empirical misclassification and the other on empirical squared error minimization—

such that one of them guarantees small probability of error if the VC dimension $\mathcal{V}_{\mathcal{F}}$ is small, and the other when the scale-sensitive dimension P_ϵ is small. However, it is easy to see that no universal relationship exists between $\mathcal{V}_{\mathcal{F}}$ and P_ϵ. In fact, if $\mathcal{F}$ is the class of all functions on $\mathcal{R}$ whose value is in $[0,1/2]$ if $x<0$ and in $(1/2,1]$ if $x>0$ then $\mathcal{V}_{\mathcal{F}}=1$ but $P_\epsilon=\infty$ for every ϵ. On the other hand, if $\mathcal{F}$ contains every function defined on the positive integers such that $|\eta(x)-1/2|\le e^{-x}/2$, then $P_\epsilon=\lfloor -\log(2\epsilon)\rfloor$, but $\mathcal{V}_{\mathcal{F}}=\infty$. (The latter example is taken from Alon, Ben-David, Cesa-Bianchi and Haussler (1993).) Therefore, these two methods are not compatible. In fact, it is easy to show situations in which $\mathcal{V}_{\mathcal{F}}=\infty$, empirical error minimization fails but empirical squared error minimization provides a small probability of error, and vice versa when $P_\epsilon=\infty$. The reason is that these general methods are not designed for the specific classification problem we are studying.

In this chapter we look for methods which are universal in the sense that they unify the advantages of the above two principles, and in fact, may work well even if both previous methods fail. We discuss three different methods in detail. All of them are *skeleton estimates*, that is, first a finite subset of $\mathcal{F}$ is selected, and then the empirical error is minimized over this subset. The methods differ in the way the skeleton is formed. The first method—introduced in Section 4.3—assumes the knowledge of the distribution μ of X. We show that the rate of convergence of this method is upper bounded by a quantity determined by the *minimum* of $\mathcal{V}_{\mathcal{F}}$ and P_ϵ. In Section 4.4 we discuss methods which do not assume the knowledge of μ. Here the skeleton is formed in a data-dependent way. The first simple method has performance guarantees better than those of empirical squared error minimization. The second data-based skeleton estimate is specifically suited for the classification problem. The size of the error here is dominated by the covering numbers according to a new metric, and the performance bounds we obtain here are better than those obtainable for empirical error minimization. In Theorem 4.4 we relate these covering numbers to a new scale-sensitive dimension d_γ defined below, which is always smaller than the minimum of $\mathcal{V}_{\mathcal{F}}$ and P_γ.

Next we define a dimension for a class $\mathcal{F}$ of functions $\mathcal{X}\to[0,1]$.

DEFINITION 4.1. *Let $0 < \gamma \le 1/2$. We say that $\mathcal{F}$* γ-shatters *a finite set $A \subset \mathcal{X}$ if there exists some function $s : A \to \left[\frac{1}{2}-\gamma, \frac{1}{2}+\gamma\right]$ such that for every subset $E \subset A$ there is a function $\eta_E \in \mathcal{F}$ such that $\eta_E(x) \ge s(x)+\gamma$ if $x \in E$ and $\eta_E(x) < s(x)-\gamma$ if $x \in A - E$. The* γ-dimension d_γ *of $\mathcal{F}$ is defined as the largest positive integer n for which there exists a set A of cardinality n which is γ-shattered by $\mathcal{F}$. If for every n there is a set A which is γ-shattered by $\mathcal{F}$ then we say that $d_\gamma = \infty$.*

REMARK.
The asymmetry in the definition of γ-shattering is necessary for our purposes. It corresponds to the tie-breaking convention we apply, that is, to the fact that when $\eta(x) = 1/2$, the corresponding classification is always 1. If this asymmetry wasn't present, d_γ would be equal to the P_γ dimension of the set of "clipped" functions $\pi_{2\gamma}(\eta)$, $\eta \in \mathcal{F}$, where $\pi_{2\gamma}$ is defined for $y \in [0,1]$ as

$$\pi_{2\gamma}(y) = \begin{cases} \frac{1}{2}+2\gamma & \text{if } \ y \ge \frac{1}{2}+2\gamma \\ \frac{1}{2}-2\gamma & \text{if } \ y \le \frac{1}{2}-2\gamma \\ y & \text{otherwise.} \end{cases}$$

This class of functions also appears in Bartlett (1997) in a somewhat different setup. □

First note that d_γ is a monotone decreasing function of γ. To see this, let $\gamma_1 \ge \gamma_2$, and let A be a set which is γ_1-shattered by $\mathcal{F}$ according to some function $s : A \to \left[\frac{1}{2}-\gamma_1, \frac{1}{2}+\gamma_1\right]$ Then it is easy to see that the same set is γ_2-shattered by $\mathcal{F}$ according to the function s' defined by

$$s'(x) = \begin{cases} s(x) & \text{if } \ s(x) \in [1/2-\gamma_2, 1/2+\gamma_2] \\ 1/2-\gamma_2 & \text{if } \ s(x) < 1/2-\gamma_2 \\ 1/2+\gamma_2 & \text{if } \ s(x) > 1/2+\gamma_2. \end{cases}$$

Our main result concerning d_γ is Theorem 4.4, which is an upper bound for certain covering numbers appearing in the performance bound (Theorem 4.3) of a data-dependent skeleton estimate proposed in Section 4.4.

d_γ is closely related to the scale-sensitive dimension P_γ, whose usefulness have been demonstrated for learning "probabilistic concepts" and for more general regression function estimation problems, see Kearns and Shapire (1994), Alon, Ben-David, Cesa-Bianchi, and Haussler (1993), Bartlett, Long, and Williamson (1996), Bartlett and Long (1995), Bartlett (1997), Anthony and Bartlett (1994), Shawe-Taylor, Bartlett, Williamson, and Anthony (1996). The only difference is that in the definition of P_γ the range of the shattering function s is not restricted, it can take any value in $[0,1]$. Therefore, clearly, for every γ,

$$d_\gamma \le P_\gamma.$$

In fact, d_γ may be finite even if $P_\gamma = \infty$ for every γ. (Just consider the class of all functions $\eta : \mathcal{R} \to [0,1]$ such that $\eta(x) < 1/2$ if $x \le 0$ and $\eta(x) \ge 1/2$ if $x > 0$.) The restriction of the range of s is motivated by the fact that from the point of view of classification, only the behavior of the functions in $\mathcal{F}$ around $1/2$ matters. For some discussion on this we refer to Section 6.7 of Devroye, Györfi and Lugosi (1996).

Also clearly,

$$d_\gamma \le \mathcal{V}_\mathcal{F}$$

for each γ. Again, d_γ may be finite even if $\mathcal{V}_\mathcal{F} = \infty$. As a simple example, consider the class of Lipschitz functions on $[0,1]$. Then $d_\gamma \le P_\gamma = O(1/\gamma)$, but obviously $\mathcal{V}_\mathcal{F} = \infty$.

Since $d_\gamma \le \min(\mathcal{V}_\mathcal{F}, P_\gamma)$, we may interpret the new dimension as one that unifies the advantages of $\mathcal{V}_\mathcal{F}$ and the scale-sensitive dimension P_γ. On the other hand, it is important to note that d_γ may be finite even if $\min(\mathcal{V}_\mathcal{F}, P_\gamma) = \infty$. The simplest example is the class of all functions $\mathcal{R} \to [0, \frac{1}{2} + \frac{\gamma}{2}]$.

Several other dimensions have been introduced to measure the size of classes of functions. A partial survey is found in Alon, Ben-David, Cesa-Bianchi and Haussler (1993).

4.3 Skeleton estimates

Skeleton estimates first form a finite subclass of the class of functions $\mathcal{F}$ and then use some kind of empirical risk minimization to choose an estimate from this class. Different types of skeleton estimates were proposed and studied by Benedek and Itai (1988), Buescher and Kumar (1994a), (1994b), Dudley, Kulkarni, Richardson, and Zeitouni (1994), Kulkarni (1991), Vapnik (1982), and Devroye, Györfi, and Lugosi (1996, Chapter 28).

The "skeleton" is typically chosen as some kind of covering of the class $\mathcal{F}$, that is, for each function $\eta \in \mathcal{F}$, there should be a function in the skeleton which is sufficiently close to η in some sense. Different versions of skeleton estimates differ in the way this closeness is measured, and also in the way the classifier is selected from the skeleton. For example, assuming that the distribution μ of X is known, Vapnik (1982), and Benedek and Itai (1988) cover the class of classifiers in the metric $d_H(\eta, \eta') = \mathbf{P}\{g(X) \neq g'(X)\}$ and select by minimizing the empirical error over the covering. In Devroye, Györfi, and Lugosi (1996, Chapter 28), $\mathcal{F}$ is covered with respect to the supremum norm $\sup_{x \in \mathcal{X}} |\eta(x) - \eta'(x)|$, and again, the empirical error is minimized over the covering.

In this chapter we propose new skeleton estimates, specifically designed for the discussed classification problem, which perform better than previously discussed methods. The key ingredient of our method is a new way of selecting the skeleton. The "distance" according to which we measure closeness of members of $\mathcal{F}$ is motivated by (1.10).

For simplicity, assume first that the distribution of X is known. Then for $\epsilon > 0$, we may select a set of functions $\widehat{\mathcal{F}}_\epsilon$ such that for every $\eta \in \mathcal{F}$ there is an $\eta' \in \widehat{\mathcal{F}}_\epsilon$ such that

$$\mathbf{E}\left\{2|\eta(X) - \eta'(X)| I_{\{g(X) \neq g'(X)\}}\right\} < \epsilon.$$

(g and g' are the classifiers defined by η and η'.) Choose $\widehat{\mathcal{F}}_\epsilon$ such that it has minimal cardinality. Next, select a member of $\widehat{\mathcal{F}}_\epsilon$ which minimizes the empirical error

$$\hat{L}_n(g) = \frac{1}{n} \sum_{i=1}^{n} I_{\{g(X_i) \neq Y_i\}}$$

over all $g = I_{\{\eta \geq 1/2\}}$, $\eta \in \widehat{\mathcal{F}}_\epsilon$. Denote the minimizing function by $\widehat{\eta}_n$ and the corresponding classifier by $\widehat{g}_n$.

We have the following performance bound for this estimate:

THEOREM 4.1. *(Lugosi and Pintér (1998)) If $\eta^* \in \mathcal{F}$, then for every $\epsilon, \delta > 0$,*

$$\mathbf{P}\left\{L(\widehat{g}_n) - L^* > \delta + \epsilon\right\} \leq \left(|\widehat{\mathcal{F}}_\epsilon| + 1\right) e^{-\frac{3}{32}n\delta^2/(L^*+\delta+\epsilon)},$$

and

$$\mathbf{E}\left\{L(\widehat{g}_n)\right\} - L^* \leq \epsilon + \max\left[\sqrt{\frac{22(L^*+\epsilon)\log\left(n(|\widehat{\mathcal{F}}_\epsilon|+1)\right)}{n}}, \frac{22\log\left(n(|\widehat{\mathcal{F}}_\epsilon|+1)\right)}{n}\right].$$

A key ingredient of the proof is the following simple lemma, which will be useful also further in this chapter.

LEMMA 4.1. *(Lugosi and Pintér (1998)) Let $\mathcal{F}'$ be a finite set of functions $\mathcal{X} \to [0,1]$, possibly depending on D_m. Let η_n minimize the empirical error*

$$\hat{L}_{n-m}(\eta) = \frac{1}{n-m}\sum_{i=m+1}^{n} I_{\{g(X_i)\neq Y_i\}}$$

over $\eta \in \mathcal{F}'$ and let $\widehat{\eta}$ be minimize the probability of error $L(\widehat{\eta})$ in $\mathcal{F}'$. Let g_n and $\widehat{g}$ denote the corresponding classifiers. Then

$$\mathbf{P}\left\{L(g_n) - L(\widehat{\eta}) > 2\delta \,|\, D_m\right\} \leq \left(|\mathcal{F}'| + 1\right) e^{-\frac{3}{8}(n-m)\delta^2/(L(\widehat{\eta})+2\delta)}. \tag{4.6}$$

PROOF. If $L(g_n) - L(\widehat{\eta}) > 2\delta$, then there exists an $\eta \in \mathcal{F}'$ such that $L(\eta) > L(\widehat{\eta}) + 2\delta$ and $\hat{L}_{n-m}(\eta) \leq \hat{L}_{n-m}(\widehat{\eta})$. Thus,

$$\begin{aligned}
&\mathbf{P}\{L(g_n) - L(\widehat{\eta}) > 2\delta | D_m\} \\
&\quad\leq \mathbf{P}\left\{\min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} \hat{L}_{n-m}(\eta) < \hat{L}_{n-m}(\widehat{\eta}) \,\middle|\, D_m\right\} \\
&\quad\leq \mathbf{P}\left\{\min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} \hat{L}_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m\right\} + \mathbf{P}\{\hat{L}_{n-m}(\widehat{\eta}) > L(\widehat{\eta}) + \delta | D_m\},
\end{aligned}$$

so we need to show that

$$\begin{aligned}
&\mathbf{P}\left\{\min_{\eta: L(\eta) > L(\widehat{\eta})+2\delta} \hat{L}_{n-m}(\eta) < L(\widehat{\eta}) + \delta \,\middle|\, D_m\right\} \\
&\quad\leq \mathbf{P}\left\{\max_{\eta\in\mathcal{F}'} \frac{L(\eta) - \hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \,\middle|\, D_m\right\}.
\end{aligned} \tag{4.7}$$

But if

$$\max_{\eta\in\mathcal{F}'} \frac{L(\eta)-\hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} \le \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}},$$

then for each $\eta\in\mathcal{F}'$

$$\hat{L}_{n-m}(\eta) \ge L(\eta) - \delta\sqrt{\frac{L(\eta)}{L(\widehat{\eta})+2\delta}}.$$

If, in addition, η is such that $L(\eta) > L(\widehat{\eta})+2\delta$, then by the monotonicity of the function $x - c\sqrt{x}$ (for $c>0$ and $x > c^2/4$),

$$\hat{L}_{n-m}(\eta) \ge L(\widehat{\eta})+2\delta - \delta\sqrt{\frac{L(\widehat{\eta})+2\delta}{L(\widehat{\eta})+2\delta}} = L(\widehat{\eta})+\delta,$$

and (4.7) follows. Thus, we have

$$\begin{aligned}
&\mathbf{P}\{L(g_n)-L(\widehat{\eta}) > 2\delta | D_m\}\\
&\le \mathbf{P}\{\hat{L}_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\}\\
&\qquad + \mathbf{P}\left\{ \max_{\eta\in\mathcal{F}_\epsilon} \frac{L(\eta)-\hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \,\middle|\, D_m\right\}.
\end{aligned}$$

Next, we bound both terms on the right-hand side of the above inequality. First, since given D_m, the conditional distribution of $(n-m)\hat{L}_{n-m}(\widehat{\eta})$ is binomial with parameters $n-m$ and $L(\widehat{\eta})$, we have by Bernstein's inequality (Lemma 3.1) that

$$\mathbf{P}\{\hat{L}_{n-m}(\widehat{\eta}) - L(\widehat{\eta}) > \delta | D_m\} \le e^{-(n-m)\delta^2/(2L(\widehat{\eta})+\frac{2}{3}\delta)}.$$

(Recall that $\widehat{\eta}$ minimizes the probability of error in $\mathcal{F}'$.) On the other hand, clearly

$$\begin{aligned}
&\mathbf{P}\left\{ \max_{\eta\in\mathcal{F}'} \frac{L(\eta)-\hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \,\middle|\, D_m\right\}\\
&\le |\mathcal{F}'| \max_{\eta\in\mathcal{F}'} \mathbf{P}\left\{ \frac{L(\eta)-\hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta})+2\delta}} \,\middle|\, D_m\right\}.
\end{aligned}$$

For any fixed η, the probability on the right-hand side is zero if $\gamma \stackrel{\text{def}}{=} \delta/\sqrt{L(\widehat{\eta})+2\delta} > \sqrt{L(\eta)}$. Otherwise, if $\gamma \le \sqrt{L(\eta)}$, then again by Bernstein's inequality,

$$
\begin{aligned}
&\mathbf{P}\left\{ \frac{L(\eta) - \hat{L}_{n-m}(\eta)}{\sqrt{L(\eta)}} > \frac{\delta}{\sqrt{L(\widehat{\eta}) + 2\delta}} \right| D_m \Bigg\} \\
&= \mathbf{P}\left\{ L(\eta) - \hat{L}_{n-m}(\eta) > \gamma\sqrt{L(\eta)} \right| D_m \Big\} \le e^{-\frac{(n-m)\gamma^2 L(\eta)}{2L(\eta) + 2\sqrt{L(\eta)}\gamma/3}} \\
&\le e^{-\frac{3}{8}(n-m)\gamma^2} \\
&= e^{-\frac{3}{8}(n-m)\delta^2/(L(\widehat{\eta})+2\delta)},
\end{aligned}
$$

which completes the proof of the lemma. □

We have the following easy corollary:

COROLLARY 4.1. *(Lugosi and Pintér (1998))*

$$
\begin{aligned}
&\mathbf{E}\left\{ L(g_n) - L(\widehat{\eta}) \middle| D_m \right\} \\
&\le \max\left[\sqrt{\frac{22 L(\widehat{\eta}) \log(n(|\mathcal{F}'|+1)) + 1}{n-m}}, \frac{22\log(n(|\mathcal{F}'|+1)) + 1}{n-m} \right].
\end{aligned}
$$

PROOF. Observe that Lemma 4.1 implies that for all $\delta > 0$,

$$
\mathbf{P}\left\{ L(g_n) - L(\widehat{\eta}) > \delta \middle| D_m \right\} \le (|\mathcal{F}'|+1)\max\left[e^{-\frac{3}{64}(n-m)\delta^2/L(\widehat{\eta})}, e^{-\frac{3}{64}(n-m)\delta} \right].
$$

But for every u,

$$
\mathbf{E}\left\{ L(g_n) - L(\widehat{\eta}) \middle| D_m \right\} \le u + \mathbf{P}\left\{ L(g_n) - L(\widehat{\eta}) > u \middle| D_m \right\}.
$$

Choosing

$$
u = \max\left[\sqrt{\frac{64 L(\widehat{\eta}) \log(n(|\mathcal{F}'|+1))}{3(n-m)}}, \frac{64\log(n(|\mathcal{F}'|+1)) + 1}{3(n-m)} \right]
$$

yields the corollary. □

PROOF OF THEOREM 4.1. Note that by the inequality (1.10),

$$
\min_{\eta \in \widehat{\mathcal{F}}_\epsilon} L(\eta) - L^* \le \epsilon.
$$

The Theorem now is a direct consequence of Lemma 4.1 and Corollary 4.1 (by taking $m = 0$). □

Observe that a straightforward consequence of the definition of the covering $\widehat{\mathcal{F}}_\epsilon$ is that $|\widehat{\mathcal{F}}_\epsilon| \leq \min(N_1(\epsilon/2, \mu, \mathcal{F}), N_H(\epsilon/2, \mu, \mathcal{F}))$. Therefore, the probability of error of the selected classifier is guaranteed to be small if *at least* one of these two covering numbers is controlled. With a bit of work, we may relate these covering numbers to the VC dimension and the Kearns-Shapire scale-sensitive dimension of $\mathcal{F}$. This implies the most important corollary of Theorem 4.1, which states that the obtained classifier has a small guaranteed probability of error whenever the *minimum* of the VC dimension $\mathcal{V}_\mathcal{F}$ and the scale-sensitive dimension $P_{\epsilon/256}$ is finite:

COROLLARY 4.2. *(Lugosi and Pintér (1998)) If $\eta^* \in \mathcal{F}$, then for every $\epsilon, \delta > 0$, and $\gamma > \epsilon$ (recall that ϵ is a parameter of the algorithm),*

$$\mathbf{P}\left\{L(\widehat{g}_n) - L^* > \gamma\right\} < \delta$$

for

$$n \geq K\frac{1}{\gamma^2}\left(\min(\mathcal{V}_\mathcal{F}, P_{\epsilon/256}) \log^2 \frac{1}{\epsilon} + \log \frac{1}{\delta}\right).$$

PROOF. As we noted it before, $|\widehat{\mathcal{F}}_\epsilon| \leq \min(N_1(\mu, \epsilon/2, \mathcal{F}), N_H(\mu, \epsilon/2, \mathcal{F}))$. The basic idea is that whenever deviations of empirical averages from their theoretical counterparts are uniformly small, empirical covering numbers are, in some sense, close to covering numbers measured by the corresponding expected distance. (This idea is quite standard, see Vapnik and Chervonenkis (1981), Nobel (1992), Bartlett, Kulkarni and Posner (1996)). Clearly, if

$$\sup_{\eta,\eta' \in \mathcal{F}} d_H(\eta, \eta') - d_{H,n}(\eta, \eta') < \frac{\epsilon}{4},$$

then $d_H(\eta, \eta') < \epsilon/2$ for every pair $\eta, \eta' \in \mathcal{F}$ such that $d_{H,n}(\eta, \eta') < \epsilon/4$. This implies that

$$\begin{aligned}
&\mathbf{P}\left\{N_H(\epsilon/4, X_1^n, \mathcal{F}) < N_H(\epsilon/2, \mu, \mathcal{F})\right\} \\
&\quad\leq \mathbf{P}\left\{\sup_{\eta,\eta' \in \mathcal{F}} d_H(\eta, \eta') - d_{H,n}(\eta, \eta') \geq \frac{\epsilon}{4}\right\} \\
&\quad\leq 4\mathbf{E}S_{\mathcal{F}-\mathcal{F}}(X_1^{2n})e^{-n\epsilon^2/128}
\end{aligned}$$

by the Vapnik-Chervonenkis inequality (Lemma 2.1), where $\mathcal{F}-\mathcal{F}$ denotes class of functions $I_{\{g(x)\neq g'(x)\}}$, $\eta,\eta' \in \mathcal{F}$. But it is easy to see that, for example, $S_{\mathcal{F}-\mathcal{F}}(x_1^{2n}) \le S_{\mathcal{F}}(x_1^{2n})^4$, so we have that

$$\mathbf{P}\left\{N_H(\epsilon/4, X_1^n, \mathcal{F}) < N_H(\epsilon/2, \mu, \mathcal{F})\right\} \le 4\left(\frac{2ne}{\mathcal{V}_{\mathcal{F}}}\right)^{4\mathcal{V}_{\mathcal{F}}} e^{-n\epsilon^2/128}.$$

The right-hand side of the inequality is strictly smaller than one if n is larger than a constant times $(\mathcal{V}_{\mathcal{F}}/\epsilon^2)\log(1/\epsilon)$. For such n's, there exist n points $x_1,\dots,x_n$ such that $N_H(\epsilon/4, x_1^n, \mathcal{F}) \ge N_H(\epsilon/2, \mu, \mathcal{F})$. But by Sauer's lemma, if $\epsilon \ge 4/n$,

$$\sup_{x_1,\dots,x_n} N_H(\epsilon/4, x_1^n, \mathcal{F}) \le \sup_{x_1,\dots,x_n} S_{\mathcal{F}}(x_1^n) \le \left(\frac{ne}{\mathcal{V}_{\mathcal{F}}}\right)^{\mathcal{V}_{\mathcal{F}}},$$

which means that for such n, $|\widehat{\mathcal{F}}_\epsilon| \le N_H(\epsilon/2, \mu, \mathcal{F}) \le \left(\frac{ne}{\mathcal{V}_{\mathcal{F}}}\right)^{\mathcal{V}_{\mathcal{F}}}$.

To obtain an analogous upper bound for $|\widehat{\mathcal{F}}_\epsilon|$ in terms of the scale-sensitive dimension P_γ, we proceed similarly:

$$\begin{aligned}\mathbf{P}\left\{N_1(\epsilon/4, X_1^n, \mathcal{F}) < N_1(\epsilon/2, \mu, \mathcal{F})\right\} &\le \mathbf{P}\left\{\sup_{\eta,\eta'\in\mathcal{F}} d_1(\eta,\eta') - d_{1,n}(\eta,\eta') \ge \frac{\epsilon}{4}\right\} \\ &\le 8\mathbf{E}N_1(\epsilon/32, X_1^n, \mathcal{F}-\mathcal{F})e^{-n\epsilon^2/2048}\end{aligned}$$

by Lemma 2.2, where $\mathcal{F}-\mathcal{F}$ denotes the class of functions $\eta(x)-\eta'(x)$, $\eta,\eta' \in \mathcal{F}$. But an elementary argument shows that

$$N_1(\epsilon/32, X_1^n, \mathcal{F}-\mathcal{F}) \le N_1(\epsilon/64, X_1^n, \mathcal{F})^2.$$

We may bound these covering numbers by the key result (4.5), according to which

$$\mathbf{P}\left\{N_1(\epsilon/4, X_1^n, \mathcal{F}) < N_1(\epsilon/2, \mu, \mathcal{F})\right\} \le 4\left(\frac{2^{14}n}{\epsilon^2}\right)^{2P_{\epsilon/256}\log(2en/(P_{\epsilon/256}))} e^{-n\epsilon^2/2048}.$$

The right-hand side is strictly smaller than one if n is larger than a constant times

$$\frac{P_{\epsilon/256}}{\epsilon^2}\log^2\frac{1}{\epsilon}.$$

For such n's therefore we have that

$$\begin{aligned} |\widehat{\mathcal{F}}_\epsilon| &\leq N_1(\epsilon/2, \mu, \mathcal{F}) \\ &\leq \sup_{x_1,\ldots,x_n} N_1(\epsilon/4, x_1^n, \mathcal{F}) \\ &\leq 2\left(\frac{2^{14}n}{\epsilon^2}\right)^{2P_{\epsilon/256}\log(2en/(P_{\epsilon/256}))}, \end{aligned}$$

where again we used (4.5).

But Theorem 4.1 implies that for $\gamma > \epsilon$,

$$\mathbf{P}\left\{L(\widehat{g}_n) - L^* > \gamma\right\} < \delta \quad \text{if} \quad n \geq K\frac{1}{\gamma^2}\left(\log|\widehat{\mathcal{F}}_\epsilon| + \log\frac{1}{\delta}\right),$$

so by putting the pieces together, the proof of the Corollary is finished. □

The corollary shows that the rate of convergence of the proposed algorithm is determined by the *minimum* of the two dimensions that appeared in the two algorithms obtained by two different philosophies—namely, empirically minimizing misclassification and estimating regression functions. The skeleton estimate investigated here seems to unify the advantages of the two approaches. The constants and the power of the logarithmic factors have not been optimized. In this qualitative discussion we sacrifice sharp results for simplicity. The bound of Theorem 4.1 is better than those that can be obtained for the above-mentioned skeleton estimates of Vapnik (1982), Benedek and Itai (1988), and Devroye, Györfi, and Lugosi (1996, Chapter 28).

On the other hand, the algorithm may work even if $\min(\mathcal{V}_\mathcal{F}, P_\epsilon) = \infty$ for every ϵ. Consider the example in which $\mathcal{F}$ contains all functions with values in $[0, \frac{1}{2} + \frac{\epsilon}{4}]$. Then $\min(\mathcal{V}_\mathcal{F}, P_\epsilon) = \infty$, and it is easy to see that both empirical error minimization and empirical squared error minimization may give huge probabilities of error. On the other hand, $|\widehat{\mathcal{F}}| = 1$, since the set $\mathcal{F}$ is covered by the single function $\eta \equiv \frac{1}{2} - \frac{\epsilon}{4}$. We conjecture that $\min(\mathcal{V}_\mathcal{F}, P_{\epsilon/256})$ in Corollary 4.2 can be replaced by $d_{K\epsilon}$, but we haven't been able to prove this.

The main disadvantage of the algorithm discussed here is that it assumes knowledge of the underlying distribution μ of X. In the next section we propose and analyze methods which do not use such information. It is interesting to note, however, that knowledge of μ usually does not help much. In fact, the minimax lower bound (4.2) remains valid even if μ is allowed to be known.

4.4 Data-based skeleton estimates

In this section we present classification rules that first form a finite skeleton of $\mathcal{F}$ based on a part of the training data, and then use the other half of the data to select the empirically best candidate from the skeleton. The idea of such empirical covering is due to Buescher and Kumar (1994b), and was further explored by Lugosi and Nobel (1996). The key difference here is that the metric according to which the covering is chosen is different from what is used in selecting a classifier from the covering. For example, in the simplest situation, one may first form an L_1-cover of $\mathcal{F}$, and then (instead of minimizing the empirical L_1 error) minimize the empirical misclassification over the skeleton. This hybrid approach seems to prove fruitful.

The precise definition of the classifier described above is as follows: first, the data sequence D_n is split into two parts:

$$D_m = ((X_1, Y_1), \ldots, (X_m, Y_m))$$

and

$$T_{n-m} = ((X_{m+1}, Y_{m+1}), \ldots, (X_n, Y_n)).$$

The first part, D_m, is used to create a skeleton of $\mathcal{F}$, that is, a finite set $\mathcal{F}_\epsilon^\circ = \mathcal{F}_\epsilon^\circ(D_m)$ such that each member of $\mathcal{F}$ is closely approximated by a function from $\mathcal{F}_\epsilon^\circ$ in the empirical L_1 metric. More precisely, given $\epsilon > 0$, let $\mathcal{F}_\epsilon^\circ$ be a class of functions of minimal cardinality satisfying the property that for each $\eta \in \mathcal{F}$ there exists an $\bar{\eta} \in \mathcal{F}_\epsilon^\circ$ such that

$$\frac{2}{m} \sum_{i=1}^{m} |\eta(X_i) - \bar{\eta}(X_i)| < \epsilon.$$

Clearly, $|\mathcal{F}_\epsilon^\circ| = N_1(\epsilon/2, X_1^m, \mathcal{F})$. The second part of the data, T_{n-m}, is used to test all classifiers defined by functions in $\mathcal{F}_\epsilon^\circ$, and to select one with minimal empirical error. In other words, we minimize the empirical error

$$\hat{L}_{n-m}(g) = \frac{1}{n-m} \sum_{i=m+1}^{n} I_{\{g(X_i) \neq Y_i\}}$$

among all classifiers $g = I_{\{\eta \geq 1/2\}}$ in $\mathcal{F}_\epsilon^\circ$. Denote the obtained classification rule by g_n° (and the corresponding regression function by η_n°). Note that g_n° ignores the values of $Y_1, \ldots, Y_m$, and therefore it may make efficient use of additional unlabeled samples, if available. The first half of the sample is only used to obtain information about μ. For simplicity, we only consider splitting the data into two equal halfs (i.e., $m = n/2$). This is a safe, but not necessarily optimal choice. The following performance bound for the obtained classifier is given for the case $m = n/2$, but the proof contains the general inequalities.

THEOREM 4.2. *(Lugosi and Pintér (1998)) If $\eta^* \in \mathcal{F}$, then for any n, ϵ, and $\delta \geq 6\epsilon$,*

$$\begin{aligned}
&\mathbf{P}\left\{L(g_n^\circ) - L^* > 3\delta\right\} \\
&\leq \; 2\mathbf{E}N_1(\epsilon/2, X_1^{n/2}, \mathcal{F})e^{-3n\delta^2/(32L^*)} + 10\mathbf{E}N_1(\epsilon/32, X_1^{n/2}, \mathcal{F})e^{-n\delta/512}.
\end{aligned}$$

In particular,

$$\begin{aligned}
&\mathbf{E}\left\{L(g_n^\circ)\right\} - L^* \leq 6\epsilon + \\
&3\max\left[\sqrt{\frac{11L^*\log(2n\mathbf{E}N_1(\epsilon/2, X_1^{n/2}, \mathcal{F}))}{n}}, \frac{512\log(10n\mathbf{E}N_1(\epsilon/32, X_1^{n/2}, \mathcal{F}))}{n}\right].
\end{aligned}$$

The bound obtained here is better than that of (4.4) for the method of squared error minimization in that it reflects the right dependence on L^* and that it involves the expected random covering numbers instead of their supremum. We believe that the same type of bound is also true for squared error minimization and we doubt that the above skeleton estimate is inherently better than squared error minimization. We included this method because of its simplicity, conceptual reasons, and comparison purposes.

In the proof we apply an inequality of Pollard (1986), sharpened by Haussler (1992). In particular, the following corollary is used, which was obtained by Buescher and Kumar (1994b) in a slightly different form. The form given here is found in Lugosi and Nobel (1996).

LEMMA 4.2. *Let $\mathcal{H}$ be class of functions on $\mathcal{X}$ such that $h(x) \in [0, A]$ for every $h \in \mathcal{H}$ and every $x \in \mathcal{X}$. Let $X_1, \ldots, X_m \in \mathcal{X}$ be i.i.d. random vectors. Then for each $\delta > 0$ and, $\epsilon > 0$,*

$$\mathbf{P}\left\{\sup_{h\in\mathcal{H}:\frac{1}{m}\sum_{i=1}^m h(X_i)<\epsilon} \mathbf{E}\{h(X)\} > \delta + 3\epsilon\right\} \leq 4\mathbf{E}\left\{|\mathcal{H}_{\epsilon/16}|\right\} e^{-m(\delta+\epsilon)/(64A)},$$

where $\mathcal{H}_\epsilon$ is any set of functions satisfying the property that for each $h \in \mathcal{H}$ there is a $h' \in \mathcal{H}_\epsilon$ with

$$\frac{1}{m}\sum_{i=1}^m |h(X_i) - h'(X_i)| < \epsilon.$$

PROOF OF THEOREM 4.2. Write

$$L(g_n^\circ) - L^* = \left(L(g_n^\circ) - \min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta)\right) + \left(\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^*\right),$$

so that

$$\mathbf{P}\{L(g_n^\circ)-L^* > 3\delta\} \leq \mathbf{P}\left\{L(g_n^\circ) - \min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) > 2\delta\right\}+\mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \delta\right\}. \tag{4.8}$$

First we bound the second probability. Introduce the notation $h_\eta(x) = 2|\eta(x) - \eta^*(x)|$,

$$J_m(\eta) = \frac{1}{m}\sum_{i=1}^m h_\eta(X_i) \quad \text{and} \quad J(\eta) = \mathbf{E}\{J_m(\eta)\} = \mathbf{E}\{h_\eta(X)\}$$

for all $\eta \in \mathcal{F}$. Recall that by (1.10), $L(\eta) - L^* \leq J(\eta)$, and that $\min_{\eta\in\mathcal{F}_\epsilon^\circ} J_m(\eta) < \epsilon$ by the definition of $\mathcal{F}_\epsilon^\circ$ and by the assumption $\eta^* \in \mathcal{F}$. Therefore,

$$\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* \leq \min_{\eta\in\mathcal{F}_\epsilon^\circ} J(\eta) \leq \sup_{\eta\in\mathcal{F}:J_m(\eta)<\epsilon} J(\eta).$$

Therefore, for $\delta \geq 6\epsilon$,

$$\begin{aligned}
\mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \delta\right\} &\leq \mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \frac{\delta}{2} + 3\epsilon\right\} \\
&\leq \mathbf{P}\left\{\sup_{\eta\in\mathcal{F}:J_m(\eta)<\epsilon} J(\eta) > \frac{\delta}{2} + 3\epsilon\right\} \\
&\leq 4\mathbf{E}\left\{N_1(\epsilon/16, X_1^m, \mathcal{H})\right\} e^{-m\delta/256} \qquad (4.9)
\end{aligned}$$

by Lemma 4.2, where $\mathcal{H}$ denotes the class of functions h_η, $\eta \in \mathcal{F}$. By noting that for every η, η', and x, $|h_\eta(x) - h_{\eta'}(x)| \le 2|\eta(x) - \eta'(x)|$, we see that $N_1(\epsilon/16, X_1^m, \mathcal{H}) \le N_1(\epsilon/32, X_1^m, \mathcal{F})$, so for $\delta \ge 6\epsilon$ we have

$$\mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \delta\right\} \le 4\mathbf{E}\left\{N_1(\epsilon/32, X_1^m, \mathcal{F})\right\} e^{-m\delta/256}.$$

The first probability on the right-hand side of (4.8) may directly be bounded by applying Lemma 4.1. Summarizing, we get

$$\begin{aligned}
&\mathbf{P}\left\{L(g_n^\circ) - L^* > 3\delta\right\} \\
&\le \mathbf{P}\left\{L(g_n^\circ) - \min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) > 2\delta\right\} + \mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \delta\right\} \\
&\le \mathbf{P}\left\{L(g_n^\circ) - \min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) > 2\delta \,\Big|\, \min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* \le \delta\right\} \\
&\quad + 2\mathbf{P}\left\{\min_{\eta\in\mathcal{F}_\epsilon^\circ} L(\eta) - L^* > \delta\right\} \\
&\le 2\mathbf{E}\left\{N_1(\epsilon/2, X_1^m, \mathcal{F})\right\} e^{-\frac{3}{8}(n-m)\delta^2/(L^*+3\delta)} \\
&\quad +8\mathbf{E}\left\{N_1(\epsilon/32, X_1^m, \mathcal{F})\right\} e^{-m\delta/256} \\
&\quad \text{(by Lemma 4.1)} \\
&\le 2\mathbf{E}\left\{N_1(\epsilon/2, X_1^m, \mathcal{F})\right\} \max\left[e^{-\frac{3}{16}(n-m)\delta^2/L^*}, e^{-(n-m)\delta/16}\right] \\
&\quad +8\mathbf{E}\left\{N_1(\epsilon/32, X_1^m, \mathcal{F})\right\} e^{-m\delta/256} \\
&\le 2\mathbf{E}\left\{N_1(\epsilon/2, X_1^m, \mathcal{F})\right\} e^{-\frac{3}{16}(n-m)\delta^2/L^*} + 10\mathbf{E}\left\{N_1(\epsilon/32, X_1^m, \mathcal{F})\right\} e^{-m\delta/256},
\end{aligned}$$

so the proof of the probability inequality is finished. As for the bound on the expected value of the probability of error, observe that for any $u \ge 6\epsilon$,

$$\begin{aligned}
\mathbf{E}\left\{L(g_n^\circ)\right\} - L^* &\le 3u + \mathbf{P}\left\{L(g_n^\circ) - L^* > 3u\right\} \\
&\le 3u + 2\mathbf{E}\left\{N_1(\epsilon/2, X_1^m, \mathcal{F})\right\} e^{-\frac{3}{16}(n-m)u^2/L^*} \\
&\quad +10\mathbf{E}\left\{N_1(\epsilon/32, X_1^m, \mathcal{F})\right\} e^{-mu/256},
\end{aligned}$$

and recalling that $m = n/2$, choose

$$\begin{aligned}
&u = 6\epsilon + \\
&\max\left[\sqrt{\tfrac{11L^*\log(2n\mathbf{E}N_1(\epsilon/2, X_1^{n/2}, \mathcal{F}))}{n}}, \tfrac{512\log(10n\mathbf{E}N_1(\epsilon/32, X_1^{n/2}, \mathcal{F}))}{n}\right].
\end{aligned}$$

□

Again, we may rephrase the probability inequality in terms of sample sizes and the scale-sensitive dimension P_γ. The proof of the next corollary is direct by applying (4.5):

COROLLARY 4.3. *(Lugosi and Pintér (1998)) If $\eta^* \in \mathcal{F}$, then for every $\epsilon, \delta > 0$, and $\gamma > 18\epsilon$ (recall that ϵ is a parameter of the algorithm),*

$$\mathbf{P}\left\{L(g_n^\circ) - L^* > \gamma\right\} < \delta$$

for

$$n \geq K \max\left(\frac{L^* P_{\epsilon/8}}{\gamma^2}\log^2\frac{P_{\epsilon/8}}{\epsilon} + \frac{L^*}{\gamma^2}\log\frac{1}{\delta}, \frac{P_{\epsilon/128}}{\gamma}\log^2\frac{P_{\epsilon/128}}{\epsilon} + \frac{1}{\gamma}\log\frac{1}{\delta}\right).$$

Next we apply the ideas of Section 4.3 to define a data-based skeleton estimate better suited for classification purposes. The main idea is to empirically cover $\mathcal{F}$ with respect to the empirical version of the "metric" $\mathbf{E}\left\{2|\eta(X) - \eta'(X)|I_{\{g(X)\neq g'(X)\}}\right\}$. More precisely, given $\epsilon > 0$, let $\overline{\mathcal{F}}_\epsilon$ be a class of functions of minimal cardinality satisfying the property that for each $\eta \in \mathcal{F}$ there exists an $\bar{\eta} \in \overline{\mathcal{F}}_\epsilon$ such that

$$\frac{2}{m}\sum_{i=1}^{m}|\eta(X_i) - \bar{\eta}(X_i)|I_{\{g(X_i)\neq\bar{g}(X_i)\}} < \epsilon. \tag{4.10}$$

Then, just like in the case of the previous classifier, the empirical error

$$\hat{L}_{n-m}(g) = \frac{1}{n-m}\sum_{i=m+1}^{n} I_{\{g(X_i)\neq Y_i\}}$$

is minimized among all classifiers $g = I_{\{\eta\geq 1/2\}}$ in $\overline{\mathcal{F}}_\epsilon$. Denote the obtained classification rule and the corresponding regression function by $\overline{g}_n$ and $\overline{\eta}_n$. We have the following property:

THEOREM 4.3. *(Lugosi and Pintér (1998)) If $\eta^* \in \mathcal{F}$, then for any n, ϵ, and $\delta \geq 6\epsilon$,*

$$\begin{aligned}\mathbf{P}\left\{L(\overline{g}_n) - L^* > 3\delta\right\} \;\leq\;& 2\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}e^{-3n\delta^2/(32L^*)} + 2\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}e^{-n\delta/32}\\ &+8\mathbf{E}N_1(\epsilon/16, X_1^{n/2}, \mathcal{G})e^{-n\delta/512},\end{aligned}$$

where $\mathcal{G}$ is the class of functions $|2\eta^*(x)-1|I_{\{g(x)\neq g^*(x)\}}$, $\eta \in \mathcal{F}$. In particular,

$$\mathbf{E}\{L(\overline{g}_n)\} - L^* \leq 6\epsilon + 3\max\left[\sqrt{\frac{11L^*\log\left(2n\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}\right)}{n}}, \frac{512\log\left(10n\max\left[\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}, \mathbf{E}N_1(\epsilon/16, X_1^{n/2}, \mathcal{G})\right]\right)}{n}\right].$$

PROOF. The proof is completely analogous to that of Theorem 4.2, so we omit it. □

Thus, the rate of convergence of the error of the selected classifier is determined by the logarithm of the expected value of the covering number $|\overline{\mathcal{F}}_\epsilon|$. The lower-order term involves the expected L_1 covering numbers of the class $\mathcal{G}$. Unfortunately, this covering number cannot be bounded by $|\overline{\mathcal{F}}_\epsilon|$, and in fact, in some situations it can be much larger. However, note that for any $\eta, \eta' \in \mathcal{F}$,

$$\begin{aligned}\Big||2\eta^*(x)-1|I_{\{g(x)\neq g^*(x)\}} - |2\eta^*(x)-1|I_{\{g'(x)\neq g^*(x)\}}\Big| &= I_{\{g(x)\neq g'(x)\}}|2\eta^*(x)-1| \\ &\leq I_{\{g(x)\neq g'(x)\}}.\end{aligned}$$

Therefore, $N_1(\epsilon/16, X_1^{n/2}, \mathcal{G})) \leq N_H(\epsilon/16, X_1^{n/2}, \mathcal{F}))$, and in particular, for $\epsilon \geq 32/n$,

$$N_1(\epsilon/16, X_1^{n/2}, \mathcal{G})) \leq S_{\mathcal{F}}(X_1^{n/2}) \leq \left(\frac{ne}{2\mathcal{V}_{\mathcal{F}}}\right)^{\mathcal{V}_{\mathcal{F}}}. \tag{4.11}$$

Thus, the finiteness of the VC dimension and $L^* > 0$ guarantees that for sufficiently large sample sizes the maximum is taken by the first term, involving $\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}$. If $\mathcal{V}_{\mathcal{F}} = \infty$, we cannot guarantee any nontrivial bound for $\mathbf{E}\{L(\overline{g}_n)\}$ (see also the remark below).

However, since $\mathbf{E}\left\{N_1(\epsilon/16, X_1^{n/2}, \mathcal{G})\right\}$ only appears in the lower-order term, in most cases it has a minor importance. The main message is that the expected size of the error is of the order of $\sqrt{L^*\log\left(\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}\right)/n}$, unless the Bayes error L^* is very small.

Below in Theorem 4.4 we relate the new scale-sensitive dimension d_γ, introduced in Section 4.2, to the covering numbers appearing in Theorem 4.3.

It is worth comparing $\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}$ to other quantities relevant for analyzing the performance of different classification rules. For example, it is easy to see that for all $\epsilon \geq 4/n$,

$$\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\} \leq \mathbf{E}\left\{S_{\mathcal{F}}(X_1^{n/2})\right\}.$$

(Note that "good" choices for ϵ are always larger, and in some cases much larger than $4/n$.) Thus, for the data-dependent skeleton classifier, Theorem 4.3 guarantees better performance than the inequality (4.1) for empirical risk minimization. The reason why it is possible to improve on these bounds is that one can make use of the additional information provided by the knowledge of the form of the possible regression functions.

We must remark here, however, that in order to obtain Theorem 4.3 (as well as Theorems 4.1, 4.2), we needed to assume that $\eta^* \in \mathcal{F}$, that is, the "true" regression function is the member of the class $\mathcal{F}$. The bounds obtained for empirical risk minimization do not require this assumption. Empirical risk minimization is, therefore, much more robust than the skeleton estimate introduced here.

It is also easy to see that

$$\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\} \leq \mathbf{E}\left\{N_1(\epsilon/2, X_1^{n/2}, \mathcal{F})\right\}.$$

Therefore, the rate of convergence achieved by the classifier $\overline{g}_n$ is usually better than that of empirical squared error minimization (or than that of the first data-dependent skeleton classifier introduced in this section). Consider, for example, the following trivial case: let $\mathcal{F}$ contain *all* functions $\eta : \mathcal{R} \to [0,1]$ such that $\eta(x) < 1/2$ if $x < 0$ and $\eta(x) \geq 1/2$ if $x \geq 0$. Then there is only one classifier induced by these functions, and accordingly, $|\overline{\mathcal{F}}_\epsilon| = 1$. On the other hand, this class is clearly too large for obtaining meaningful bounds for regression function estimation.

Remark.

Note, however, that $\overline{g}_n$ may fail in some situations when empirical squared error minimization and g_n° do not. We have the following counterexample: Let $\mathcal{F}$ contain all functions on $[0,1]$ which take their values between $1/2-\epsilon/4$

and $1/2+\epsilon/4$, and the constant zero function. Assume that $\eta^*(x)=0$ for all $x \in [0,1]$. Then $|\overline{\mathcal{F}}_\epsilon| = 1$, since $\mathcal{F}$ can be covered by the single function which equals $1/2-\epsilon/4$ if x is one of the first $n/2$ data points, and $1/2+\epsilon/4$ otherwise. If the distribution of X is absolutely continuous, $L(\overline{g}_n)=1$ with probability one, while $L^*=0$. On the other hand, $\mathbf{E}N_1(\epsilon/32, X_1^{n/2}, \mathcal{F}) = 2$, so, for example, g° is guaranteed to give a small error probability. □

Next we relate $|\overline{\mathcal{F}}_\epsilon|$ to the new scale-sensitive dimension d_γ defined in Section 4.2:

THEOREM 4.4. *(Lugosi and Pintér (1998)) For any value of* $X_1,\dots,X_m$, *if* $12m \geq \lceil 1/\epsilon \rceil$, $\epsilon \leq 1$, *then*

$$|\overline{\mathcal{F}}_\epsilon| \leq 2(6m)^{d_{\epsilon/10}\log_2(2em/d_{\epsilon/10})+1}.$$

This result is the analogue of (4.5) for the new dimension introduced here. The proof is based on the proof of Lemma 3.4 of Alon, Ben-David, Cesa-Bianchi, and Haussler (1993). Just like there, we also begin with "discretizing." First we introduce a discrete analogue of the γ-dimension, related to the "strong dimension" of Alon et al. (1993).

Let b be a positive even integer, and let $\mathcal{G}$ be a class of functions $\mathcal{X} \to \{1,2,\dots,b\}$. We say that $\mathcal{G}$ *b-shatters* a finite set $A \subset \mathcal{X}$ according to a function $s: A \to \{b/2, b/2+1\}$ if for every subset E of A, there exists a function $f_E \in \mathcal{G}$ such that

$$f_E(x) \begin{cases} \leq s(x)-1 & \text{if } x \in E \\ \geq s(x)+1 & \text{if } x \in A-E. \end{cases}$$

We say that $\mathcal{G}$ *b-shatters* A if $\mathcal{G}$ *b-shatters* A according to some s. The *b-dimension* Δ_b of $\mathcal{G}$ is the largest integer n such that there exists a set A, shattered by $\mathcal{G}$, with $|A|=n$. If there is no such larger integer, then we say that $\Delta_b = \infty$.

Let $\gamma > 0$. The γ-discretization of a function $\eta: \mathcal{X} \to [0,1]$ is defined by $\eta^\gamma(x) = \text{round}(\eta(x)/\gamma)+1$, where

$$\text{round}(z) = \begin{cases} \lfloor z \rfloor & \text{if } z - \lfloor z \rfloor < 1/2 \\ \lfloor z \rfloor + 1 & \text{if } z - \lfloor z \rfloor \geq 1/2. \end{cases}$$

Note that with $\gamma = 1/(b-1)$, and $\eta \in [0,1]$, the γ-discretization $\eta^\gamma(x)$ of η maps $\mathcal{X}$ to $\{1, \ldots, b\}$.

Lemma 4.3. *(Lugosi and Pintér (1998)) Assume that $\gamma = 1/(b-1)$ for some positive even integer b, and let $\mathcal{F}^\gamma = \{\eta^\gamma : \eta \in \mathcal{F}\}$ denote the class of γ-discretizations of functions in $\mathcal{F}$. If Δ_b denotes the b-dimension of $\mathcal{F}^\gamma$, then*

$$\Delta_b \leq d_{\gamma/2},$$

where $d_{\gamma/2}$ is defined in Definition 4.1 for $\mathcal{F}$.

Proof. We show that if $\mathcal{F}^\gamma$ b-shatters a set A, then $\mathcal{F}$ $\gamma/2$-shatters A. Let $s : A \to \{b/2, b/2+1\}$ be the function which is used by $\mathcal{F}^\gamma$ to b-shatter A. Then for every $E \subset A$ there is a function $\eta_E \in \mathcal{F}$ such that

$$\eta_E^\gamma(x) \begin{cases} \leq s(x) - 1 & \text{if } x \in E \\ \geq s(x) + 1 & \text{if } x \in A - E. \end{cases}$$

Then clearly, if $s'(x) = \gamma s(x) - \gamma$, then

$$\eta_E(x) \begin{cases} < s'(x) - \gamma/2 & \text{if } x \in E \\ \geq s'(x) + \gamma/2 & \text{if } x \in A - E, \end{cases}$$

so $\mathcal{F}$ clearly $\gamma/2$-shatters A. $\square$

Next we relate $|\overline{\mathcal{F}}_\epsilon|$ to certain packing numbers of the class of discretizations of functions in $\mathcal{F}$. Let $\{x_1, \ldots, x_m\} \subset \mathcal{X}$. We say that a subset $\mathcal{F}'$ of $\mathcal{F}$ is *ϵ-separated* if for any $\eta_1, \eta_2 \in \mathcal{F}'$,

$$\frac{1}{m} \sum_{i=1}^{m} 2I_{\{g_1(x_i) \neq g_2(x_i)\}} |\eta_1(x_i) - \eta_2(x_i)| \geq \epsilon.$$

The maximal size $M(\epsilon, \mathcal{F})$ of such an ϵ-separated set is called the *ϵ-packing number* of $\mathcal{F}$. Now consider a class $\mathcal{G}$ of functions $\mathcal{X} \to \{1, \ldots, b\}$, where b is an even positive integer. We say that $\mathcal{G}' \subset \mathcal{G}$ is *2-separated* if for any $f_1, f_2 \in \mathcal{G}'$,

$$\max_{i=1,\ldots,m} |f_1(x_i) - f_2(x_i)| I_{\{u(f_1(x_i)) \neq u(f_2(x_i))\}} \geq 2,$$

where the function u is defined by

$$u(a) = \begin{cases} 1 & \text{if } a > b/2 \\ 0 & \text{if } a \le b/2. \end{cases}$$

The maximal size of a 2-separated subset of $\mathcal{G}$ is denoted by $M_b(2, \mathcal{G})$. The proof of the next lemma is trivial.

Lemma 4.4. *(Lugosi and Pintér (1998)) Let b be a positive even integer, and assume that $\gamma = 1/(b-1) \le \epsilon/4$. Then*

$$|\overline{\mathcal{F}}_\epsilon| \le M(\epsilon, \mathcal{F}) \le M_b(2, \mathcal{F}^\gamma).$$

The key of the proof of Theorem 4.4 is the following combinatorial lemma:

Lemma 4.5. *(Lugosi and Pintér (1998)) Let $\mathcal{X}$ be a set of cardinality m, and let $\mathcal{G}$ be a class of functions $\mathcal{X} \to \{1, \ldots, b\}$, where b is an even positive integer and $m \ge b/6$. Then*

$$M_b(2, \mathcal{G}) \le 2(6m)^{\lceil \log_2 y \rceil}, \qquad \textit{where} \quad y = \sum_{i=1}^{\Delta_b} \binom{m}{i} 2^i$$

and Δ_b is the b-dimension of $\mathcal{G}$.

Proof. We may assume that $b \ge 4$ since otherwise there are no two 2-separated functions in $\mathcal{G}$ and the statement is trivial. Let $A \subset \mathcal{X}$ and $s : A \to \{b/2, b/2+1\}$. We say that $\mathcal{G}$ b-shatters the pair (A, s) if it b-shatters A according to s. To any $k \ge 2$ and $m \ge 1$, define $t(k, m)$ as the largest integer t such that if $\mathcal{H}$ is any 2-separated class of functions with $|\mathcal{H}| = k$, then $\mathcal{H}$ b-shatters at least t distinct pairs (A, s). If no such $\mathcal{H}$ exists, then we say that $t(k, m) = \infty$. (Recall that $m = |\mathcal{X}|$.)

Clearly, the number of possible pairs (A, s) such that $|A| \le d$ is at most $y = \sum_{i=1}^{d} \binom{m}{i} 2^i$. Thus, if $t(k, m) > y$ for some k, then $M_b(2, \mathcal{G}) < k$ whenever $\Delta_b \le d$. Therefore, we need to show that $t\left(2(6m)^{\lfloor \log_2 y \rfloor}, m\right) > y$ for all $d \ge 1$, $m \ge 1$.

We see immediately that $t(2, m) = 1$ for all $m \ge 1$. Next we show that

$$t(12km, m) \ge 2t(2k, m-1). \tag{4.12}$$

If there is no 2-separated class with size $12km$, then the left-hand side of (4.12) is ∞, and the inequality is trivially true. Thus, assume that there is a 2-separated class $\mathcal{H}$ with size $12km$. Split $\mathcal{H}$ into $6km$ pairs of functions. For each such pair (h_1, h_2),

$$\max_{x \in \mathcal{X}} |h_1(x) - h_2(x)| I_{\{u(h_1(x)) \neq u(h_2(x))\}} \geq 2$$

that is, for each such pair there exists an $x \in \mathcal{X}$ such that $|h_1(x) - h_2(x)| \geq 2$ and $u(h_1(x)) \neq u(h_2(x))$. Since $|\mathcal{X}| = m$, there exists an $x \in \mathcal{X}$ such that this property holds for at least $6k$ pairs. Order all these pairs such a way that $h_1(x) \geq h_2(x)$. For $j \in \{1, \ldots, b\}$, define

$$\tau(j) = \begin{cases} 1 & \text{if } j < b/2 \\ 2 & \text{if } j = b/2 \\ 3 & \text{if } j = b/2 + 1 \\ 4 & \text{if } j > b/2 + 1. \end{cases}$$

Since for all pairs $h_1(x) - h_2(x) \geq 2$ and $u(h_1(x)) \neq u(h_2(x))$, there are three possible values of te pair $(\tau(h_1(x)), \tau(h_2(x)))$ (namely $(3,1)$, $(4,1)$, and $(4,2)$). By the pigeonhole principle, there are at least $6k/3 = 2k$ pairs (h_1, h_2) for which the $(\tau(h_1(x)), \tau(h_2(x)))$ is the same. Then it follows that there are two subclasses $\mathcal{H}_1, \mathcal{H}_2 \subset \mathcal{H}$ and indices $i, j \in \{1, 2, 3, 4\}$ with $|\mathcal{H}_1| = |\mathcal{H}_2| = 2k$ such that for each $h_1 \in \mathcal{H}_1$, $\tau(h_1(x)) = i$, for each $h_2 \in \mathcal{H}_2$, $\tau(h_2(x)) = j$, and $i \geq j + 2$. Clearly, the members of $\mathcal{H}_1$ are 2-separated on $\mathcal{X} - \{x\}$, and the same is true for $\mathcal{H}_2$. Thus, according to the definition of $t(k, m)$, $\mathcal{H}_1$ and $\mathcal{H}_2$ both b-shatter $t(2k, m-1)$ pairs (A, s) with $A \subset \mathcal{X} - \{x\}$.

Clearly, $\mathcal{H}$ b-shatters every pair (A, s) which is b-shattered by either $\mathcal{H}_1$ or $\mathcal{H}_2$. Also, if a pair (A, s) is b-shattered by both $\mathcal{H}_1$ and $\mathcal{H}_2$, then it is easy to see that $\mathcal{H}$ b-shatters the pair $(A \cup \{x\}, s')$, where $s'(z) = s(z)$ if $z \in A$, and

$$s'(x) = \begin{cases} b/2 + 1 & \text{if } j = 2 \\ b/2 & \text{if } i = 3 \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Therefore, $\mathcal{H}$ b-shatters at least as many (A, s) pairs as the sum of the numbers of pairs shattered by $\mathcal{H}_1$ and $\mathcal{H}_2$, so (4.12) is proved.

Let now $n = 2(6m)(6(m-1))\cdots(6(m-r+1))$, where $r \le m$. Then by repeated application of (4.12), we obtain

$$t(n, m) \ge 2^r t(2, n-r) = 2^r.$$

Since t is monotone in its first argument, for all $r \le m$,

$$t(2(6m)^r, m) \ge 2^r.$$

Take $r = \lceil \log_2 y \rceil$. If $r \le m$, then

$$t(2(6m)^{\lceil \log_2 y \rceil}, m) \ge 2^{\lceil \log_2 y \rceil} > y,$$

as desired. If $r > m$, then $2(6m)^r > 2(6m)^m > b^m$ by the condition $6m \ge b$. But b^m is the number of all functions from $\mathcal{X}$ to $\{1, \ldots, b\}$, so there is no 2-separated class larger than this, hence $t(2(6m)^{\lceil \log_2 y \rceil}, m) = \infty > y$, establishing the lemma. □

PROOF OF THEOREM 4.4. Let $b = 4\lceil \frac{1}{\epsilon} + \frac{1}{2} \rceil$ and $\gamma = \frac{1}{b-1}$. Then it is easy to check that $\gamma \le \epsilon/4$—so that the hypothesis of Lemma 4.4 is satisfied—and $\gamma/2 \ge \epsilon/10$. Then

$$\begin{aligned}
|\overline{\mathcal{F}}_\epsilon| &\le M_b(2, \mathcal{F}^\gamma) \quad \text{(by Lemma 4.4)} \\
&\le 2(6m)^{\lceil \log_2 (\sum_{i=1}^{\Delta_b} \binom{m}{i} 2^i) \rceil} \quad \text{(by Lemma 4.5)} \\
&\le 2(6m)^{\left\lceil \log_2 \left(\sum_{i=1}^{d_{2\gamma}} \binom{m}{i} 2^i\right) \right\rceil} \quad \text{(by Lemma 4.3)} \\
&\le 2(6m)^{\left\lceil \log_2 \left(\frac{2em}{d_{\gamma/2}}\right)^{d_{\gamma/2}} \right\rceil} \\
&\quad \text{(since } \binom{m}{i} 2^i \le \binom{2m}{i} \text{ and } \textstyle\sum_{i=1}^{k} \binom{n}{i} \le (ne/k)^k) \\
&\le 2(6m)^{d_{\epsilon/10} \log_2 (2em/d_{\epsilon/10}) + 1}
\end{aligned}$$

as desired. □

We may combine the result of Theorem 4.4 in a straightforward way with Theorem 4.3 and (4.11) to obtain the following sample-size bound for the data-dependent skeleton estimate $\overline{g}_n$:

COROLLARY 4.4. *(Lugosi and Pintér (1998)) Assume that $\eta^* \in \mathcal{F}$. For every $\epsilon, \delta > 0$, and $\gamma > 18\epsilon$ (recall that ϵ is a parameter of the algorithm),*

$$\mathbf{P}\left\{L(\overline{g}_n) - L^* > \gamma\right\} < \delta$$

for

$$n \geq K \max\left(\frac{L^* d_{\epsilon/20}}{\gamma^2} \log^2 \frac{d_{\epsilon/20}}{\epsilon} + \frac{L^*}{\gamma^2} \log \frac{1}{\delta}, \frac{\mathcal{V}_{\mathcal{F}}}{\gamma} \log^2 \frac{\mathcal{V}_{\mathcal{F}}}{\epsilon} + \frac{1}{\gamma} \log \frac{1}{\delta}\right).$$

Therefore, if $\mathcal{V}_{\mathcal{F}} < \infty$ and $L^* > 0$, for sufficiently small γ's the sample size is determined by $d_{\epsilon/20}$ instead of the VC dimension as for empirical error minimization.

Finally, we indicate some points where the bound of Theorem 4.4 is loose. First of all, the theorem provides an upper bound for the maximal possible value of $|\overline{\mathcal{F}}_\epsilon|$, whereas the interesting quantity in Theorem 4.3 is its expected value $\mathbf{E}\left\{|\overline{\mathcal{F}}_\epsilon|\right\}$. In certain cases, the difference may be significant: Devroye, Györfi, and Lugosi (1996, Theorem 13.13) provides such an example.

If $\mathcal{F}$ is a class of indicator functions, then $|\overline{\mathcal{F}}_\epsilon|$ is just the random shatter coefficient $S_{\mathcal{F}}(X_1^m)$ for $\epsilon = 2/m$. In this case Sauer's lemma (1972) implies that $\log|\overline{\mathcal{F}}_\epsilon| \leq \mathcal{V}_{\mathcal{F}} \log m$, whereas Theorem 4.4 only gives $\log|\overline{\mathcal{F}}_\epsilon| = O(\mathcal{V}_{\mathcal{F}} \log^2 m)$.

If $\mathcal{F}$ is the class of all Lipschitz functions (with Lipschitz constant 1) on $[0,1]$, then it is easy to see that $\log|\overline{\mathcal{F}}_\epsilon| = O(1/\epsilon)$. However, it is also easy to see that $d_\epsilon = O(1/\epsilon)$, and therefore Theorem 4.4 only implies $\log|\overline{\mathcal{F}}_\epsilon| = O\left(\frac{\log^2 m}{\epsilon}\right)$. However, the practical importance of the log factors is minor, so Theorem 4.4 may be a useful tool to bound $|\overline{\mathcal{F}}_\epsilon|$.

4.5 Concluding remarks

Several interesting questions have been left unanswered. Perhaps the most interesting one is the characterization of "learnability." We suspect that the finiteness of d_γ is sufficient and necessary. More precisely, our conjecture is the following: given a class $\mathcal{F}$ of regression functions, a sample size of the order of (ignoring logarithmic factors)

$$\max\left(\frac{L^* d_\epsilon}{\gamma^2} + \frac{L^*}{\gamma^2}\log\frac{1}{\delta}, \frac{d_\epsilon}{\gamma} + \frac{1}{\gamma}\log\frac{1}{\delta}\right)$$

is sufficient and necessary to guarantee the existence of a classifier g_n such that for all possible $\eta^* \in \mathcal{F}$ and for all distributions μ the excess probability of error $L(g_n) - L^*$ be smaller than γ with probability larger than $1-\delta$ for all δ and $\gamma > c\epsilon$ for some constant c. We do not know the answer even if μ is allowed to be known. The bounds given here for classifiers $\widehat{g}_n$ of Section 4.3 and $\overline{g}_n$ of Section 4.4 both fall short of this performance guarantee. In the case of known μ, $\widehat{g}_n$ is a candidate to prove the sufficiency part. However, we haven't been able to replace in Corollary 4.2 the minimum of $\mathcal{V}_\mathcal{F}$ and P_ϵ by d_ϵ. The main technical difficulty comes from the fact that the "metric" which defines the covering does not satisfy the triangle inequality. In the case of unknown distributions, as we pointed it out, $\overline{g}_n$ fails in some cases when d_γ is finite. If our conjecture is true, a new classifier is to be found. Also, unlike in the completely distribution-free setup, the answers in the two cases (i.e., known and unknown μ) may be very different.

Bibliography

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1993). Scale-sensitive dimensions, uniform convergence, and learnability. *Technical Report 143-95, DSI, University of Milan, Italy.* An extended abstract appeared in the Proceedings of the 1993 IEEE Symposium on the Foundations of Computer Science IEEE Press.

Anthony, M. and Bartlett, P. L. (1994). Function learning from interpolation. Technical Report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia.

Anthony, M. and Shawe-Taylor, J. (1993). A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217.

Bartlett, M. S. (1938). The characteristic function of a conditional statistic. *J. London Math. Soc.*, 13:62–67.

Bartlett, P. (1997). The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory.*

Bartlett, P., Kulkarni, S., and Posner, S. (1996). Covering numbers for real-valued function classes. Technical Report, Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia.

Bartlett, P. and Long, P. (1995). More theorems about scale-sensitive dimensions and learning. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 392–401. Association for Computing Machinery, New York.

Bartlett, P., Long, P., and Williamson, R. (1996). Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52:434–452.

Beirlant, J. and D.Mason (1995). On the asymptotic normality of l_p-norms of empirical functionals. *Mathematical Methods of Statistics*, 4:1–19.

Beirlant, J. and Györfi, L. (1998). On the asymptotic l_2-error in partitioning regression estimation. *J. statistical Planning and Inference*, 71:93 – 107.

Beirlant, J., Györfi, L., and Lugosi, G. (1994). On the asymptotic normality of the l_1- and l_2- errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309–318.

Benedek, G. M. and Itai, A. (1988). Learnability by fixed distributions. In *Computational Learning Theory: Proceedings of the 1988 Workshop*, pages 80–90. Morgan Kaufman, San Mateo, CA.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Technical Report, University of California, Berkeley.

Berlinet, J., Devroye, L., and Györfi, L. (1995). Asymptotic normality of l_1-error in histogram density estimation. *Statistics*, 26:329–343.

Bernstein, S. N. (1946). *The Theory of Probabilities.* Gastehizdat Publishing House, Moscow.

Bickel, P. J. and Doksum, K. A. (1976). *Mathematical statistics. Basic ideas and selected topics.* Holden-Day Series in Probability and Statistics. Holden-Day, Inc., San Francisco, Calif.-Düsseldorf-Johannesburg.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrica*, 66:429–436.

Buescher, K. L. and Kumar, P. R. (1994a). Learning by canonical smooth estimation, Part I: Simultanious estimation. *To appear in IEEE Transactions on Automatic Control.*

Buescher, K. L. and Kumar, P. R. (1994b). Learning by canonical smooth estimation, Part II: Learning and choice of model complexity. *To appear in IEEE Transactions on Automatic Control.*

Carbonez, A., Györfi, L., and van der Meulen, E. C. (1995). Partitioning-estimates of a regression function under random censoring. *Statistics and Decisions*, 13:21–37.

Cox, D. R. (1972). Regression models and life tables. *J. Royal Statistical Society Ser. B*, 33:187–202.

Cox, D. R. (1975). Partial likelihood. *Biometrica*, 62:269–276.

Dabrowska, D. M. (1987). Nonparametric regression with censored data. *Scandinavian J. Statistics*, 14:181–197.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional kaplan-meier estimate. *Annals of Statistics*, 17:1157–1167.

Devroye, L. (1982). Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61:467–481.

Devroye, L. and Györfi, L. (1983). Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, Konecny, F., Mogyoródi, J., and Wertz, W., editors, pages 67–76. Akadémiai Kiadó, Budapest, Hungary.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

Devroye, L., Györfi, L., Krzyzak, A., and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York.

Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23:71–82.

Devroye, L. and Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition*. To appear.

Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8:231–239.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique*, 10:5–12.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929.

Dudley, R. M., Kulkarni, S., Richardson, T., and Zeitouni, O. (1994). A metric entropy bound is not sufficient for learnability. *IEEE Transactions on Information Theory*, 40:883–885.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Annals of Statistics*, 9:586–596.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

Glick, N. (1973). Sample based multinomial classification. *Biometrics*, 29:241–256.

Gordon, L. and Olshen, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Annals of Statistics*, 6:515–533.

Gordon, L. and Olshen, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10:611–627.

Györfi, L. (1991). Universal consistencies of a regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics*, Roussas, G., editor, pages 329–338. NATO ASI Series, Kluwer Academic Publishers, Dordrecht.

Györfi, L. and Pintér(Horváth), M. (1998). On the asymptotic normality of the resubstitution error estimate for partitioning classification rule. In *Advances in Data Science and Classification*, Rizza, A., Vichi, M., and Bock, H. H., editors, pages 197–204. Springer.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150.

Horváth, L. (1981). On nonparametric regression with randomly censored data. In *Proc. Third Pannonian Symp. on Math. Statist. Visegrad, Hungary*, Mogyoródi, J., Vincze, I., and Wertz, W., editors, pages 105–112. Akadémiai Kiadó, Budapest, Hungary.

James, I. R. and Smith, P. J. (1984). Consistency results for linear regression with censored data. *Annals of Statistics*, 19:590–600.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, 53:457–481.

Kearns, M. and Shapire, R. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer Systems Sciences*, 48:464–497.

Kohler, M. (1997). On the universal consistency of a least squares spline regression estimator. *Math. Methods of Statistics*, 6:349–364.

Kohler, M. (1999). Universally consistent regression function estimation using hierarchical b-splines. *Journal of Multivariate Analysis*, 67:138–164.

Kohler, M. and Krzyżak, A. (2001). Nonparametric regression estimation using penalized least squares. *IEEE Trans. Information Theory.* to appear.

Kohler, M., Máthé, K., and Pintér, M. (2000). Prediction from randomly right censored data. *Journal of Multivariate Analysis.* to appear.

Kulkarni, S. R. (1991). *Problems of computational and information complexity in machine vision and learning.* PhD Thesis, Department of Electrical Engineering and Computer Science, MIT.

Lai, T. L. and Ying, Z. (1991). Large–sample theory of a modified buckley–james estimator. *Annals of Statistics*, 19:1370–1902.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. on Information Theory*, 42:2118–2132.

Lugosi, G. and Nobel, A. (1996). Adaptive model selection using empirical complexities. submitted to *Annals of Statistics.*

Lugosi, G. and Pintér(Horváth), M. (1998). Scale-sensitive dimensions and skeleton estimates for classification. *Discrete Applied Mathematics*, 85:37–61.

Lugosi, G. and Zeger, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41:677–678.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, Cambridge.

Nobel, A. B. (1992). *On uniform laws of averages.* PhD Thesis, Department of Statistics, Stanford University, Stanford, California.

Peterson, A. V. (1977). Expressing the kaplan-meier estimator as a function of empirical subsurvival functions. *J. Amer. Statist. Assoc.*, 72:854–858.

Pintér, M. (2000). On the rate of convergence of error estimates for the partitioning classification rule. *Theoretical Computer Science.* to appear. An extended abstract appeared in the Proceedings of the 4th European Conference on Computational Learning Theory (EUROCOLT'99).

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Pollard, D. (1986). Rates of uniform almost sure convergence for empirical processes indexed by unbounded classes of functions. Manuscript.

Rényi, A. (1968). *Valószínűségszámítás (Probability Theory).* Tankönyvkiadó, Budapest, second edition.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Annals of Statistics*, 18:303–328.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147.

Shawe-Taylor, J., Bartlett, P., Williamson, R., and Anthony, M. (1996). A framework for structural risk minimization. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 68–76. Association of Computing Machinery, New York.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, 8:240–246.

Steele, J. M. (1975). *Combinatorial entropy and uniform limit laws.* PhD Thesis, Stanford University, Stanford, CA.

Steele, J. M. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758.

Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 8:1348–1360.

Stute, W. and Wang, J. L. (1993). The strong law under random censorship. *Annals of Statistics*, 21:1591–1607.

Susarla, V., Tsai, W. Y., and Ryzin, J. V. (1984). A buckley-james-type estimator for the mean with censored data. *Biometrika*, 71:624–625.

Tukey, J. W. (1947). Nonparametric estimation ii. statistically equivalent blocks and tolerance regions. *Annals of Mathematical Statistics*, 18:529 – 539.

Vapnik, V. N. (1982). *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280.

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.

Vapnik, V. N. and Chervonenkis, A. Y. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26:821–832.