The estimation problem of minimum mean squared error

January 6, 2003

Luc Devroye
László Györfi
Dominik Schäfer
Harro Walk

Abstract

Regression analysis of a response variable Y requires careful selection of explanatory variables. The quality of a set of explanatory features $X = (X^{(1)}, ..., X^{(d)})$ can be measured in terms of the minimum mean squared error

$$L^* = \min_f \mathbf{E}\{(Y - f(X))^2\}.$$

This paper investigates methods for estimating L^* from i.i.d. data. No estimate can converge rapidly for all distributions of (X, Y). For Lipschitz continuous regression function $\mathbf{E}\{Y|X = x\}$, two estimators for L^* are discussed: fitting a regression estimate to a subset of the data and assessing its mean residual sum of squares on the remaining samples, and a nearest neighbor cross-validation type estimate.

Key words: Cross-validation, mean squared error, nearest neighbor estimate, nonparametric regression estimation.

MSC 2000: primary 62G08, secondary 62G09.

1 Introduction

Let Y be a real valued random variable and let X be a d-dimensional random vector. The coordinates of X may have different types of distributions, some of them may be discrete (for example binary), others may be absolutely continuous. In the sequel we do not assume anything about the distribution of X. The task of regression analysis is to estimate Y given X, i.e., one aims to find a function f defined on the range of X such that $f(X)$ is "close" to Y. Typically, closeness is measured in terms of the **mean squared error** of f,

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

It is well-known that the mean squared error is minimized by the regression function m with

$$m(x) = \mathbf{E}\{Y \mid X = x\} \tag{1}$$

and a minimum mean squared error

$$L^* := \mathbf{E}\{(Y - m(X))^2\} = \min_f \mathbf{E}\{(Y - f(X))^2\}.$$

For each measurable function f, the mean squared error can be decomposed into

$$\begin{aligned} \mathbf{E}\{(f(X) - Y)^2\} &= \mathbf{E}\{(m(X) - Y)^2\} + \mathbf{E}\{(m(X) - f(X))^2\} \\ &= \mathbf{E}\{(m(X) - Y)^2\} + \int_{\mathbb{R}^d} \mid m(x) - f(x) \mid^2 \mu(dx), \end{aligned}$$

where μ denotes the distribution of X. The second term on the right hand side is called **excess error** or integrated squared error of the function f and will be denoted by

$$\|m - f\|^2 = \int_{\mathbb{R}^d} \mid m(x) - f(x) \mid^2 \mu(dx). \tag{2}$$

Clearly, the mean squared error of f is close to its minimum if and only if the excess error $\|m - f\|^2$ is close to zero.

The regression function cannot be calculated as long as the distribution of (X, Y) is unknown. Assume, however, that we observed data

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

consisting of independent and identically distributed copies of (X, Y). D_n can be used to produce an estimate $m_n = m_n(\cdot, D_n)$ of the regression function m. Since m arises from L_2 considerations, it is natural to study $L_2(\mu)$

convergence of the regression estimate m_n to m. In particular, the estimator m_n is called **weakly (strongly) universally consistent** if its excess error satisfies

$$\|m - m_n\|^2 \to 0 \text{ in probability (a.s.)}$$

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

Stone [18] first pointed out that there exist weakly universally consistent estimators. He considered local averaging estimates, i.e., estimates of the form

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x; X_1, \ldots, X_n) Y_i = \sum_{i=1}^{n} W_{ni}(x) Y_i,$$

where $W_{ni}(x)$ are the data-dependent weights governing the local averaging about x. Important local averaging estimates are partitioning, kernel and nearest neighbor estimates.

The **partitioning estimate** is defined by a partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \ldots\}$ of $\mathbb{R}^d$ and

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K_n(x, X_i)}{\sum_{i=1}^{n} K_n(x, X_i)},$$

where $K_n(x, u) = \sum_{j=1}^{\infty} I_{[x \in A_{n,j}, u \in A_{n,j}]}$. Results on weak and strong universal consistency can be found in Devroye and Györfi [5] and Györfi [9].

The **kernel estimate** is given by

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K_{h_n}(x - X_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)},$$

where $h_n > 0$ is a smoothing factor depending upon n, K is an absolutely integrable function (the kernel), and $K_{h_n}(x) = K(x/h_n)$. Under the conditions

$$h_n \to 0, \quad n h_n^d \to \infty$$

Devroye and Wagner [8], Spiegelman and Sacks [16], Devroye and Krzyżak [7] and Walk [19] proved consistency theorems for the kernel estimate.

For the **k-nearest neighbor estimate**, $W_{ni}(x; X_1, \ldots, X_n)$ is chosen to be $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and zero otherwise. Note in particular that $\sum_{i=1}^{n} W_{ni} = 1$. If

$$k_n \to \infty, \quad k_n/n \to 0 \tag{3}$$

then the consistency of the k-nearest neighbor estimate was established by Stone [18] and by Devroye et al. [6].

It is of great importance to be able to estimate the minimum mean squared error L^* accurately, even before one of the above regression estimates is applied: in a standard nonparametric regression design process, one considers a finite number of real-valued features $X^{(i)}$, $i \in I$, and evaluates whether these suffice to explain Y. In case they suffice for the given explanatory task, an estimation method can be applied on the basis of the features already under consideration, if not, more or different features must be considered. The quality of a collection of features $X^{(i)}$, $i \in I$, is measured by the minimum mean squared error

$$L^*(I) := \mathbf{E}\left|Y - \mathbf{E}\{Y|X^{(i)}, i \in I\}\right|^2$$

that can be achieved using the features as explanatory variables. $L^*(I)$ depends upon the unknown distribution of $(Y, X^{(i)} : i \in I)$. The first phase of any regression estimation process therefore heavily relies on estimates of L^* (even *before* a regression estimate is picked).

Accurate estimates of L^* for higher-dimensional feature vectors are indispensible because the problem of feature selection cannot be resolved by selection rules combining the features $X^{(i)}$ according to their "single feature" errors $L^*(\{i\})$. Indeed, an easy example shows that a combination of "good" single features may lead to a larger mean squared error than a combination of "worse" features. To see this, let $X^T := (X^{(1)}, X^{(2)}, X^{(3)})^T$ be jointly Gaussian with mean μ and variance-covariance matrix Σ, and let $Y = a^T X$ for some $a \in \mathbb{R}^3$ to be chosen later. For $I \subseteq \{1, 2, 3\}$ we have

$$L^*(I) = a^T \Sigma a - a^T \Sigma P^T (P \Sigma P^T)^{-1} P \Sigma a$$

(Shiryayev [15], §13, Theorem 2), where $P = P(I)$ contains the rows of the 3×3 identity matrix with row labels in I. Choose

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma := \begin{pmatrix} 1 & -0.7 & 0 \\ -0.7 & 1 & -0.7 \\ 0 & -0.7 & 1 \end{pmatrix} \quad \text{and} \quad a = \begin{pmatrix} 2 \\ 2.5 \\ 1 \end{pmatrix}$$

to obtain the following ordering of minimum mean squared errors:

$$L^*(\{1\}) = \frac{11}{16} > L^*(\{2\}) = \frac{59}{100} > L^*(\{3\}) = \frac{3}{16}$$

and

$$L^*(\{1,2\}) = \frac{2}{51} < L^*(\{1,3\}) = \frac{1}{8} < L^*(\{2,3\}) = \frac{8}{51}.$$

Thus, the two best single features ($X^{(2)}$ and $X^{(3)}$) become the worst two-dimensional feature, whereas the two worst single features ($X^{(1)}$ and $X^{(2)}$) jointly constitute the best two-dimensional feature.

Hence, the decision whether the features already under consideration will do, or whether other features should be targeted, requires accurate estimates of L^*. As will be seen, however, without additional assumptions on the distribution of (X, Y) one cannot trust any estimate of L^*. It is even futile to let the sample size tend to infinity because any estimate is doomed to converge arbitrarily slowly for some distribution of (X, Y) (Theorem 1). In practice, one can thus never claim to have a universally superior feature extraction or minimum error estimation method, no matter how many simulations are performed and no matter how large the sample sizes are. Error bounds or confidence bands for L^* can only be constructed under additional assumptions on the distribution of the data (Theorems 2 and 3).

A related problem is the estimation of conditional variances,

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

because of

$$L^* = \mathbf{E}\{\sigma^2(X)\}.$$

For nonparametric estimates of $\sigma^2(x)$ see, e.g., Müller and Stadtmüller [12], Neumann [14], Stadtmüller and Tsybakov [17], Kohler [11] and the literature cited there. Müller, Schick and Wefelmeyer [13] estimate L^* as the variance of an independent measurement error ϵ in the model

$$Y = m(X) + \epsilon.$$

2 Slow rate of convergence

In a first group of methods, L^* is estimated by an estimate $\widehat{L}_n$ of the error $L_n = \mathbf{E}\{(Y - m_n(X))^2 | D_n\}$ of some consistent regression estimate m_n. Clearly, if the estimate $\widehat{L}_n$ we use is consistent in the sense that $\widehat{L}_n - L_n \to 0$ with probability one as $n \to \infty$, and the rule $\{m_n\}_n$ is strongly consistent, then $\widehat{L}_n \to L^*$ with probability one. In other words, we have a consistent estimate of L^*.

The problem is, however, that even though for many estimates, $\widehat{L}_n - L_n$ can be guaranteed to converge to zero rapidly, regardless of what the distribution of (X, Y) is, the rate of convergence of $\mathbf{E}\{L_n\}$ to L^* for such a method may be arbitrarily slow (cf. Györfi et al. [10]). Thus, we can not expect a good performance for all distributions from such a method. The question remains whether it is possible to come up with another method of estimating L^* (by some function $\phi_n(X_1, Y_1, \ldots, X_n, Y_n)$ of the data) such that the difference $\phi_n(X_1, Y_1, \ldots, X_n, Y_n) - L^*$ converges to zero rapidly for all distributions.

Antos, Devroye, Györfi [3] proved that for pattern recognition the Bayes error cannot be estimated with guaranteed rate of convergence. Using their construction, the next theorem shows the same for regression estimation.

Theorem 1 *For any sequence $\{\phi_n\}$ of estimates and for any sequence $\{a_n\}$ of positive numbers converging to zero, a distribution of (X, Y) on $\{1, 2, 3, \ldots\} \times \{0, 1\}$ may be found such that*

$$\mathbf{E}\{|\phi_n - L^*|\} \geq a_n \quad \textit{infinitely often.}$$

Proof. Theorem 1 is a consequence of a result of Antos, Devroye and Györfi [3]. They considered the problem of estimating the Bayes error,

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

where

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

In their example, Y is $\{0, 1\}$-valued and $\eta(X) = m(X)$ may take the values $0, 1/2$ and 1. For such a distribution of (X, Y)

$$\mathbf{Var}\{Y|X\} = \eta(X) - \eta(X)^2 = \min(\eta(X), 1 - \eta(X))/2,$$

therefore

$$L^* = R^*/2.$$

Thus the corresponding result on R^* implies the theorem. □

3 Estimation of L^* by splitting the data based on a consistent regression estimate

Under additional assumptions on the distribution of (X, Y), estimates can be constructed that guarantee a good rate of convergence. In the sequel, we assume the regression function m to be Lipschitz continuous.

For the data $D_{2n} = \{(X_1, Y_1), ..., (X_{2n}, Y_{2n})\}$ let m_n be a regression estimate based on the first n observations $D_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ (training data). An estimate of the minimum mean squared error L^* can be obtained by the mean residual sum of squares of m_n on the remaining samples (test data). More precisely, we consider the estimate

$$\bar{L}_{2n} = \frac{1}{n} \sum_{i=n+1}^{2n} (m_n(X_i) - Y_i)^2.$$

Obviously, the rate of convergence of $\bar{L}_{2n}$ to L^* depends on the quality of the estimate m_n.

Theorem 2 *Assume that* $|Y| \leq B$, $|m_n(x)| \leq B$. *Then*

$$\mathbf{E}\left\{\int (m_n(x) - m(x))^2 \mu(dx)\right\} = O(n^{-2/(2+d)}) \tag{4}$$

implies that

$$\mathbf{E}\{|\bar{L}_{2n} - L^*|\} \leq const. \max\{n^{-1/2}, n^{-2/(2+d)}\}. \tag{5}$$

Proof. The theorem follows immediately from

$$\begin{aligned} \mathbf{E}\{|\bar{L}_{2n} - L^*|\} &\leq \mathbf{E}\{|\bar{L}_{2n} - \mathbf{E}\{\bar{L}_{2n}|D_n\}|\} + \mathbf{E}\{|\mathbf{E}\{\bar{L}_{2n}|D_n\} - L^*|\} \\ &\leq \frac{4B^2}{\sqrt{n}} + \mathbf{E}\left\{\int (m_n(x) - m(x))^2 \mu(dx)\right\}. \end{aligned}$$

□

If m is Lipschitz continuous, then the rate in (4) is the optimum rate of convergence. This rate is attained, e.g., for bounded X and bounded $\sigma^2(X)$, by the partitioning estimate with cubic partition with side length $h_n \sim n^{-1/(d+2)}$, by the kernel estimate with bandwidth $h_n \sim n^{-1/(d+2)}$ and the k_n-nearest neighbor estimate with $k_n \sim n^{-2/(d+2)}$ (cf. Györfi et al. [10]).

Weakening the boundedness condition on Y, for example, to the condition of $\sigma^2(X)$ being bounded poses problems. Indeed, in general this is not possible. To see this, consider the case where $m(X) = 0$ a.s., then $L^* = \mathbf{E}\{Y^2\}$. From Antos [1], [2], however, it is known that under the only condition of $\mathbf{E}\{Y^2\} < \infty$, arbitrarily slow rate of convergence can occur, which means that for any sequence of numbers $a_n > 0$ tending to zero and for any estimates $f_n(Y_1, \ldots, Y_n)$ there is a distribution of Y with $\mathbf{E}\{Y^2\} < \infty$ such that

$$\mathbf{E}\{|f_n(Y_1, \ldots, Y_n) - \mathbf{E}\{Y^2\}|\} > a_n$$

for infinitely many n.

4 Estimation by first nearest neighbor cross-validation

For dimensions $d = 1, 2$ the rate of convergence using the splitting-the-data technique cannot be improved. However, for $d \geq 3$ this can be achieved by application of cross-validation based on first nearest neighbor rule, which is a nonconsistent regression estimate.

Let $X_{i,n}$ be the first nearest neighbor of X_i from $\{X_1, \ldots, X_n\} \setminus \{X_i\}$,

$$X_{i,n} := X_j \quad \text{with} \quad j = \underset{k=1,\ldots,n; k \neq i}{\arg\min} |X_k - X_i|$$

(here and in the following $|\cdot|$ denotes the usual Euclidean norm for $d > 1$, the absolute value for $d = 1$). If X_k and X_ℓ are equidistant from X_i, i.e., $|X_k - X_i| = |X_\ell - X_i|$ for some $k \neq \ell$, then we have a tie. In this case we apply *tie breaking by indices*, such that for $k < \ell$, X_k is declared closer to X_i than X_ℓ.

Usually, functional estimates are derived from consistent function estimation. However, it is possible to get a better rate of convergence if the corresponding function estimate is not consistent. For example, in estimating the differential entropy one can have a fast estimate based on one-spacing, which is a density estimate such that the pointwise variance tends to infinity, causing a universal additive bias for the corresponding entropy estimate (cf. Beirlant et al. [4]). Similarly, here the estimate would be

$$\tilde{L}_n = \frac{1}{2n} \sum_{i=1}^{n} (Y_{i,n} - Y_i)^2,$$

$Y_{i,n}$ being the response variable that corresponds to $X_{i,n}$. One may expect some fast rate of convergence since

$$\frac{1}{2n} \sum_{i=1}^{n} Y_i^2 \approx \frac{1}{2} \mathbf{E} Y^2,$$

$$\frac{1}{2n} \sum_{i=1}^{n} Y_{i,n}^2 \approx \frac{1}{2} \mathbf{E} Y^2$$

and

$$\frac{1}{n} \sum_{i=1}^{n} Y_{i,n} Y_i \approx \mathbf{E} m(X)^2.$$

The problem, however, is that one data point X_i may be the first nearest neighbor of more other points, and we do not know how this affects the suspected fast rate of convergence of $\tilde{L}_n$. We therefore define a modified nearest neighbor structure such that any X_i is the (modified) nearest neighbor of exactly one other point. There is no fast rate of convergence of the estimate $\tilde{L}_n$ for tie breaking by indices, a good rate can only be achieved in case ties occur with probability zero. If, however, ties occur with non-zero probability, we apply *tie breaking with randomization*, i.e. we enhance the regressors X_i by an independent uniformly $[0,1]$ distributed coordinate. More precisely, we generate independent, identically distributed uniform$([0,1])$ random variables $U_1, ..., U_n$, independent of $(X_1, Y_1), ..., (X_n, Y_n)$, and replace X_i by (X_i, U_i). This does not affect the mean squared error, since by independence

$$\mathbf{E}\{Y_i - \mathbf{E}\{Y_i|X_i\}\}^2 = \mathbf{E}\{Y_i - \mathbf{E}\{Y_i|(X_i,U_i)\}\}^2.$$

Moreover,

$$|(X_i,U_i) - (X_j,U_j)|^2 = |X_i - X_j|^2 + |U_i - U_j|^2$$

is the sum of two independent random variables one of which is absolutely continuous. It follows that $|(X_i,U_i)-(X_j,U_j)|^2$ is absolutely continuous, and hence ties occur with probability zero.

We shall set up a modified neighboring structure such that each X_i is the neighbor of exactly one of the X_j's. The modified neighboring structure will be based on a (data-dependent) permutation $j(\cdot)$ on $\{1, ..., n\}$,

$$j(i) = j(i, X_1, ..., X_n) \neq i,$$

such that

$$\sum_{i=1}^{n} |X_i - X_{j(i)}|^2 \leq 4 \sum_{i=1}^{n} |X_i - X_{i,n}|^2. \tag{6}$$

We say that $X_{j(i)}$ is the modified nearest neighbor of X_i. Note that each X_i has exactly one modified nearest neighbor and is the modified nearest neighbor of exactly one of the X_j's. The latter does not hold for the first nearest neighbors without tie breaking. As shall be seen later, the modified nearest neighboring structure makes the bias of the estimate vanish.

The following *algorithm* can be applied to construct a permutation j satisfying (6): Observe that

$$G = \{(X_i, X_j) : X_j \text{ is the first nearest neighbor of } X_i, i = 1, ..., n\}$$

defines a directed graph on $D_n = \{X_1, ..., X_n\}$ with each X_i having exactly one direct successor because we apply the tie breaking by indices. We represent G by directed edges

$$X_i \longrightarrow X_j \qquad \text{if } (X_i, X_j) \in G.$$

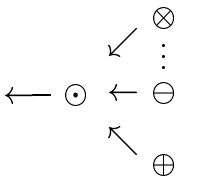
The directed graph consists of a number of connected components, each of which has precisely one cycle of nodes. The following steps are applied to each connected component:

Step 1: Let the nodes of the cycle be denoted by

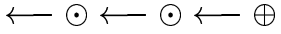
$$X_{i_1} \longrightarrow X_{i_2} \longrightarrow \cdots \longrightarrow X_{i_k} \longrightarrow X_{i_1}$$

with different $i_1, ..., i_k$ and $k \geq 2$. For each X_{i_ℓ}, let G_ℓ be the subgraph consisting of all direct and indirect predecessors of the node X_{i_ℓ} not in the cycle (i.e., all nodes from which we can reach X_{i_l} in finitely many moves along directed edges). Note that $G_1, ..., G_k$ are pairwise disjoint.

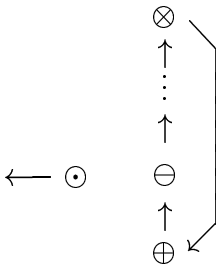
Step 2: This step applies if for all ℓ, G_ℓ is either void or only consists of nodes not connected by directed edges (in other words, X_{i_ℓ} has at most one level of predecessors). Fix ℓ. If G_ℓ is void the algorithm concerning G_ℓ stops, otherwise the nodes of G_ℓ are partitioned into levels, where the level of a node is its path distance to the inner cycle. Repeat the following procedure from the furthest level until level two: The nodes $\oplus, \ominus, ..., \otimes$ of the highest level (i.e., furthest from X_{i_ℓ} in terms of the number of directed edges needed to connect both) are of one of the following forms:



or



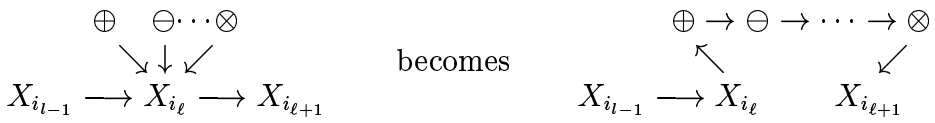
$\odot$ represents nodes of lower order. Modifying these subgraphs to become



and

respectively, we split off cycles and thus reduce the number of levels in G_ℓ by at least one. With the triangle inequality and the inequality $(a+b)^2 \leq 2(a^2+b^2)$, it is easily established that the sum of the squared distances along the resulting graph does not exceed four times the sum of the squares of the distances of the original graph. At the end of this, we have a number of disconnected cycles, and the leftover original component with all nodes either in the cycle or at level one.

Step 3: For these nodes we do the following:



Again, the sum of squared distances in the resulting graph exceeds the sum of squared distances in the original graph by at most a factor 4. Choosing $j(i)$ as the label of the direct successor of X_i in the modified graph yields the desired permutation.

The algorithm, applied to one connected component, is illustrated in Figure 1.

Now, with the modified nearest neighbor structure, we estimate the minimum mean squared error L^* by a nearest neighbor cross validation type estimate

$$\widehat{L}_n = \frac{1}{2n}\sum_{i=1}^{n}(Y_{j(i)} - Y_i)^2.$$

The following theorem provides the rate of convergence for $\widehat{L}_n$.

Theorem 3 *Let (X,Y) be an $\mathbb{R}^d \times \mathbb{R}$-valued random vector with $d \geq 3$ and $|Y| \leq B$ almost surely. Assume that the regression function m is Lipschitz continuous. Then*

$$\mathbf{E}\{|\widehat{L}_n - L^*|\} \leq \begin{cases} C \cdot n^{-1/2} & \text{for } d \in \{3,4\} \\ C \cdot n^{-2/d} & \text{for } d \geq 5 \end{cases}$$

with some constant C depending upon d, B and the Lipschitz constant of m.

Note how the rate of convergence improved in comparison with (5) for $d \geq 3$. The crucial point determining the rate is the speed of convergence of $\mathbf{E}|X_i - X_{i,n}|^2$ given in (7) below. For $d < 3$, (7) only holds under additional

assumptions on the distribution of X (Györfi et al. [10], Problem 6.7). We then obtain

$$\mathbf{E}\{|\widehat{L}_n - L^*|\} \leq C \cdot n^{-1/2},$$

which is no improvement in comparison with (5). Hence, for Lipschitz continuous m and for $d = 1$ or 2, the best rate can be achieved by splitting the sample, whereas for $d \geq 3$ a good rate can be achieved by first nearest neighbor cross-validation.

Proof. Because $(X_{j(1)}, X_{j(2)}, \ldots, X_{j(n)})$ is a permutation of $(X_1, X_2, \ldots, X_n)$,

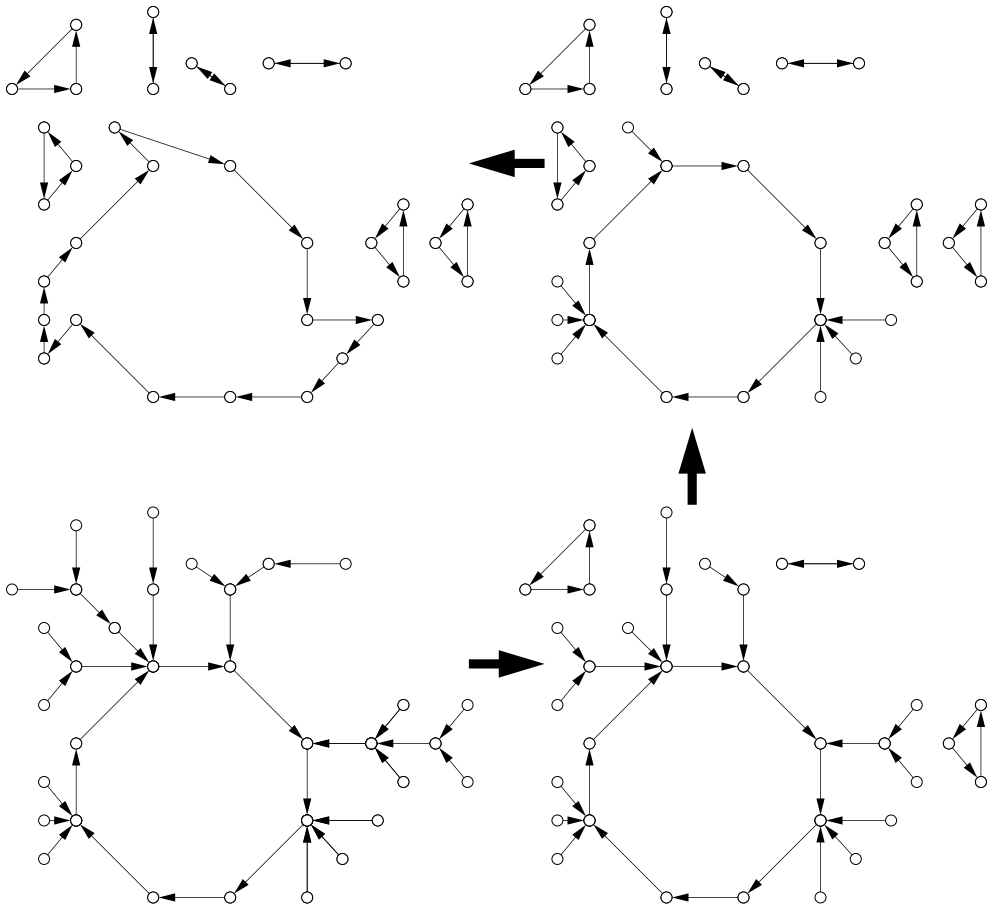


Figure 1: Illustration of the algorithm for a connected component

we have that

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Y_{j(i)}^2$$

and

$$\sum_{i=1}^{n} m(X_i)^2 = \sum_{i=1}^{n} m(X_{j(i)})^2,$$

therefore

$$\begin{aligned}
\widehat{L}_n &= \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - m(X_i)^2) \\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\left(m(X_{j(i)}) - m(X_i)\right)^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(m(X_{j(i)})m(X_i) - Y_{j(i)}Y_i) \\
&= L_{1,n} + L_{2,n} + L_{3,n}.
\end{aligned}$$

Then

$$\mathbf{E}|\widehat{L}_n - L^*| \leq \mathbf{E}|L_{1,n} - L^*| + \mathbf{E}L_{2,n} + \mathbf{E}|L_{3,n}|$$

Obviously

$$\mathbf{E}|L_{1,n} - L^*| \leq \sqrt{\mathbf{E}|L_{1,n} - L^*|^2} \leq \frac{B^2}{2\sqrt{n}}.$$

For the second term we use the Lipschitz property

$$|m(x) - m(z)| \leq K|x - z|$$

and (6) to obtain

$$\begin{aligned}
2\mathbf{E}L_{2,n} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(m(X_{j(i)}) - m(X_i)\right)^2 \\
&\leq \frac{K^2}{n}\sum_{i=1}^{n}\mathbf{E}|X_{j(i)} - X_i|^2 \\
&\leq \frac{4K^2}{n}\sum_{i=1}^{n}\mathbf{E}|X_i - X_{i,n}|^2.
\end{aligned}$$

Now,

$$\mathbf{E}|X_i - X_{i,n}|^2 \leq const.\, n^{-2/d} \tag{7}$$

for $d \geq 3$ (Györfi et al. [10], Lemma 6.4), so that

$$\mathbf{E}L_{2,n} \leq const.\ n^{-2/d}.$$

As to the third term, put

$$\Delta_{i,j(i)} = m(X_{j(i)})m(X_i) - Y_{j(i)}Y_i.$$

If for integers i and k the integers $i, j(i), k, j(k)$ are pairwise different (note that by construction $i \neq j(i)$ and $k \neq j(k)$), then

$$\mathbf{E}\{\Delta_{i,j(i)}\Delta_{k,j(k)}|X_i, X_{j(i)}, X_k, X_{j(k)}\} = 0$$

and therefore

$$\begin{aligned} n^2\mathbf{E}L_{3,n}^2 &= \mathbf{E}\left\{\sum_{i,k}\Delta_{i,j(i)}\Delta_{k,j(k)}\right\} \\ &\leq 4B^4\left(\sum_{i,k}I_{[j(i)=j(k)]} + \sum_{i,k}I_{[i=j(k)]} + \sum_{i,k}I_{[j(i)=k]} + \sum_{i,k}I_{[i=k]}\right). \end{aligned}$$

Here, $\sum_{i,k} I_{[j(i)=j(k)]} = \sum_i \sum_k I_{[j(i)=j(k)]} = \sum_i 1 = n$ since $j(\cdot)$ is a permutation on $\{1, ..., n\}$. Arguing analogously for the other double sums, we end up with

$$\sqrt{\mathbf{E}L_{3,n}^2} \leq 4B^2 n^{-1/2}.$$

Summarizing the previous results, we obtain

$$\mathbf{E}\{|\widehat{L}_n - L^*|\} \leq C \max\{n^{-2/d}, n^{-1/2}\},$$

where C depends on d, B and the Lipschitz constant K. □

Acknowledgements

The authors would like to thank the two anonymous referees for their comments and suggestions.

References

[1] Antos, A. (1999). Performance limits of nonparametric estimators. *PhD Thesis*, Technical University of Budapest.

[2] Antos, A. (2002). On nonparametric estimates of expectation. In *Limit Theorems in Probability and Statistics*, eds. I. Berkes, E. Csáki, M. Csörgö, pp. XXX–XXX. J. Bolyai Mathematical Society, Budapest.

[3] Antos, A. Devroye, L. and Györfi, L. (1999). Lower bounds for Bayes error estimates. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21, pp. 643–645.

[4] Beirlant, J., Dudewicz, E. J., Györfi, L. and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *International J. Math. Stat. Sci.*, 6, pp. 17–39.

[5] Devroye, L. and Györfi, L. (1983). Distribution-free exponential upper bound on the L_1 error of partitioning estimates of a regression function. In *Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics*, eds. Konecny F., Mogyoródi, J. and Wertz, W. , pp. 67-76. Akadémiai Kiadó, Budapest.

[6] Devroye, L., Györfi, L., Krzyżak, A. and Lugosi, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22, pp. 1371–1385.

[7] Devroye, L. and Krzyżak, A. (1989). An equivalence theorem for L1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, 23, pp. 71–82.

[8] Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Annals of Statistics*, 8, pp. 231–239.

[9] Györfi, L. (1991). Universal consistencies of regression estimate for unbounded regression functions. In *Nonparametric Functional Estimation and Related Topics*, ed. G. Roussas, pp. 329–338. Kluwer Academic Publishers, Dordrecht.

[10] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

[11] Kohler, M. (2002). Nonparametric regression with additional measurement errors in the dependent variable, *Preprint 2002-14 Math. Inst. A, Universität Stuttgart.*

[12] Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis, *Annals of Statistics*, 15, pp. 610-625.

[13] Müller, U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic, *Statistics*, to appear.

[14] Neumann, M.-H. (1994). Fully data-driven nonparametric variance estimators, *Statistics*, 25, pp. 189-212.

[15] Shiryayev, A. N. (1984). *Probability.* Springer, New York.

[16] Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Annals of Statistics*, 8, pp. 240–246.

[17] Stadtmüller, U. and Tsybakov, A. (1995). Nonparametric recursive variance estimation, *Statistics*, 27, pp. 55-63.

[18] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5, pp. 595–645.

[19] Walk, H. (2002). Almost sure convergence properties of Nadaraya-Watson regression estimates. In *Modeling Uncertainty. An Examination of its Theory, Methods and Applications*, eds. M. Dror, P. L'Ecuyer and F. Szidarovszky, pp. 201-223. Kluwer Academic Publishers, Dordrecht.

Authors

Luc Devroye
School of Computer Science
McGill University
Montréal PQ
Canada H3A 2K6
luc@cs.mcgill.ca

László Györfi
Department of Computer Sciences and Information Theory
Budapest University of Technology and Economics

Stoczek u. 2
H-1521 Budapest
Hungary
gyorfi@szit.bme.hu

Dominik Schäfer and Harro Walk
Fachbereich Mathematik
Universität Stuttgart
Pfaffenwaldring 57
D-70511 Stuttgart
Germany
schaefdk@mathematik.uni-stuttgart.de
walk@mathematik.uni-stuttgart.de