On nonparametric estimates of the expectation

András Antos *

Februar 10, 2000

Abstract

Let X be a real random variable with finite expectation m. Let $(X_1, \ldots, X_n)$ be an i.i.d. sample drawn from the distribution of X. Here we prove that without further conditions on the distribution of X, no rate-of-convergence results can be obtained for any sequence $\{m_n\}$ of estimates of m, that is, for any positive sequence $\{a_n\}$ converging to zero, a distribution may be found on the set of integers with finite mean such that

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|m_n - m|\}}{a_n} = \infty.$$

It is known that restricting ourselves to the class $\mathcal{D}^p$ of distributions with finite p^{th} absolute moment $\mathbf{E}\{|X|^p\}$ $(1 \le p \le 2)$, for $\widehat{m}_n = \sum_{i=1}^n X_i/n$ we get the rate

$$\mathbf{E}\{|\widehat{m}_n - m|\} = O\left(n^{-\frac{p-1}{p}}\right).$$

We prove that this is sharp, in the sense, that for any sequence $\{m_n\}$ of estimates and for any positive sequence $a_n = o(n^{-\frac{p-1}{p}})$, a distribution may be found in $\mathcal{D}^p$ such that

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|m_n - m|\}}{a_n} = \infty.$$

*A. Antos is with the Informatics Laboratory, Research Division, Computer and Automation Research Institute of the Hungarian Academy of Sciences, H-1518 Lágymányosi u.11, Budapest, Hungary. (email: antos@szit.bme.hu).

1 Introduction

Let X be a random variable having unknown (discrete) distribution $\mu(j)$ on the set $\mathcal{X}$ of integers. The expectation of X is

$$m \stackrel{\text{def}}{=} \mathbf{E}X = \sum_j j\mu(j).$$

We are given n i.i.d. observations

$$D_n = (X_1, \ldots, X_n),$$

drawn from the distribution of X. As a basic point of statistics, certainly it is of great importance to be able to estimate m accurately if it is finite, that is, find an estimate $m_n = m_n(D_n)$ of m such that $|m_n - m|$ is small.

An obvious estimate of m is the average

$$\widehat{m}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

By the Strong Law of Large Numbers, $\{\widehat{m}_n\}$ is a strongly universally consistent estimate, that is, for all distributions

$$\lim_{n\to\infty} \widehat{m}_n = m \quad \text{a.s..}$$

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

$$\lim_{n\to\infty} \mathbf{E}\{|\widehat{m}_n - m|\} = 0.$$

A natural question is whether there is a specified rate of convergence for the error of $\{\widehat{m}_n\}$, or at least for that of some other sequence $\{m_n\}$ of estimates. For example, if it is given that $\mathbf{E}\{X^2\}$ is finite, then

$$\mathbf{E}\{|\widehat{m}_n - m|\} \le \sqrt{\frac{\mathbf{Var}X}{n}}. \tag{1}$$

Here the constant $\sqrt{\mathbf{Var}X}$ depends on the distribution. This is necessary, because it is easy to see, that there does not exist sequence of estimates for which the expected error is uniformly less than, for example, $1/2$ (or any given sequence of real numbers) even for the class of all distributions with finite support. Just consider the following two distributions: the one of n^2B_n, and the one of $-n^2B_n$, where B_n is a binary random variable taking value 1 with probability $1/n^2$. The expectations are $+1$ and -1, respectively. Now for every n and for any estimate m_n, if all the samples fall onto 0 (which has probability $(1-1/n^2)^n = 1-o(1)$

for both distributions), then $|m_n(0,\ldots,0)-m| \geq 1$ for at least one of the two distributions, hence

$$\sup_{\mathcal{D}} \mathbf{P}\{|m_n - m| \geq 1\} \geq 1 - o(1),$$

where the supremum is taken over the class $\mathcal{D}$ of distributions with finite support.

Another question is whether a certain universal rate of convergence is achievable for some sequence of estimates with constants depending upon the distribution, as in (1). We prove that the answer is negative again: There is no method that guarantees a certain rate of convergence for all distributions with finite m, that is, the convergence of the error of *any* sequence of estimates can be arbitrary slow.

Consequently, we have to assume conditions on the distribution to get rate-of-convergence results. There are such results on the analogy of (1). We also prove that these results are sharp.

2 Slow rate of convergence

First we state the result for the convergence in probability:

Theorem 1 *For any sequence $\{m_n\}$ of estimates and for any sequence $\{a_n\}$ of positive numbers converging to zero, a distribution may be found on $\mathcal{X}$ to be the distribution of X, such that $|m| < \infty$ and for any $\epsilon > 0$*

$$\mathbf{P}\{|m_n - m| > a_n\} > \frac{1}{2} - \epsilon \quad \textit{infinitely often.}$$

Remark. Applied to the sequence $\{\sqrt{a_n}\}$ instead of $\{a_n\}$, Theorem 1 implies that for any $\{m_n\}$ there is a distribution on $\mathcal{X}$ with $|m| < \infty$ such that for any $K > 0$

$$\limsup_{n\to\infty} \mathbf{P}\{|m_n - m| > K a_n\} \geq \frac{1}{2}. \quad \square$$

Thus we get the slow-rate result for the expected error:

Corollary 1 *For any sequence $\{m_n\}$ of estimates and for any sequence $\{a_n\}$ of positive numbers converging to zero, a distribution may be found on $\mathcal{X}$ to be the distribution of X, such that $|m| < \infty$ and*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|m_n - m|\}}{a_n} = \infty. \quad \square$$

Remark. We note that similar global slow-rate-of-convergence results were obtained in pattern recognition (Cover [8], Devroye [9] and Devroye, Györfi and Lugosi [12]), regression estimation, density estimation (Devroye [10], [11] and Birgé [7]) and estimation of the Bayes-error of pattern recognition (Antos, Devroye and Györfi [3]) and other functionals (Antos [2]). $\square$

REMARK. The phrase "infinitely often" cannot be dropped from Theorem 1 and Corollary 1 (that is, the limit superior cannot be replaced by limit inferior or infimum). Indeed, there exist deterministic sequences $\{m_n\}$ with $|m_n - m| \le 2/\sqrt{n}$ infinitely often simultanously for every m. Just consider the dyadic sequence

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

and let $m_{2n-1} = m'_n$ and $m_{2n} = -m'_n$. Now for every m, for every i large enough, there is an element of the sequence in the first $2(1+4+\cdots+4^i) < 4 \cdot 4^i$ elements, whose distance from m is at most 2^{-i}. We thus obtain a very good estimate along an (unknown) subsequence for every value of m. (This can be generalized to certain finite dimensional spaces, see also Birgé [7].) $\square$

The argument in the proof uses a randomization over a parametrized subclass of distributions on $\mathcal{X}$. If u denotes the parameter, then $D_n(u)$ indicates how the samples depend on u. Denote the distribution of D_n by $\mu^{(n)}$. Let $\mathcal{N} = \{0, 1, 2, \ldots\}$.

Lemma 1 *Let F be a functional assigning a real value to each possible distribution of X, and let $F_n = F_n(D_n)$ be estimates of F. Consider a class $\{\mu_u : u \in \{-1,+1\}^{\mathcal{N}}\}$ of distributions parametrized by a sign vector $u = (u_0, u_1, u_2, \ldots)$ with the following properties: Let $A_{n,k} \subset \mathcal{X}^n$ such that if u and u' are two vectors coinciding in all but the k^{th} sign, then for every $x_1^n = (x_1, \ldots, x_n) \in A_{n,k}$*

$$\mu_u^{(n)}(x_1^n) = \mu_{u'}^{(n)}(x_1^n). \tag{2}$$

For such coincident u and u' assume moreover that

$$|F(u) - F(u')| \ge f(k), \tag{3}$$

where $F(u) = F(\mu_u)$. Then for any sequence $\{F_n(D_n(u))\}$ of estimates, for any sequence $\{a_n\}$ of positive numbers converging to zero and for any subsequence $\{n_k\}_{k\in\mathcal{N}}$ of $\{n\}_{n\in\mathcal{N}}$,

$$\sup_u \limsup_{n\to\infty} \mathbf{P}\{|F_n(D_n(u)) - F(u)| > a_n\} \ge \frac{1}{2} \limsup_{k\to\infty} (I_{\{f(k)>2a_{n_k}\}} \inf_u \mu_u^{(n_k)}(A_{n_k,k})).$$

PROOF. We use randomization such that u is replaced by a sequence $U = (U_0, U_1, U_2, \ldots)$ of independent zero mean random signs, independent of X.

Let U_+^k and U_-^k denote the vector U with the difference that U_+^k forces the k^{th} sign to be 1 and U_-^k forces the k^{th} sign to be -1. Introduce the notation

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

Now for $n, k \in \mathcal{N}$

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

Conditioning on $D_n(U)$ in $A_{n,k}$,

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

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

$$\begin{aligned}
&\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| > a_n|D_n(U)\}\mathbf{P}\{U_k = -1|D_n(U)\} \\
&\quad + \mathbf{P}\{|F_n(D_n(U)) - F(U_+^k)| > a_n|D_n(U)\}\mathbf{P}\{U_k = 1|D_n(U)\} \\
&= \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| > a_n|D_n(U)\} + \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_+^k)| > a_n|D_n(U)\} \\
&\geq \frac{1}{2}\mathbf{P}\{|F_n(D_n(U)) - F(U_-^k)| + |F_n(D_n(U)) - F(U_+^k)| > 2a_n|D_n(U)\} \\
&\geq \frac{1}{2}\mathbf{P}\{|F(U_-^k) - F(U_+^k)| > 2a_n|D_n(U)\} \\
&\geq \frac{I_{\{f(k)>2a_n\}}}{2} \qquad \text{(using (3))}.
\end{aligned}$$

For $n = n_k$ taking expectation on $A_{n_k,k}$

$$\mathbf{E}\{R_{n_k}(U)\} \geq \frac{I_{\{f(k)>2a_{n_k}\}}}{2}\mathbf{P}\{D_{n_k}(U) \in A_{n_k,k}\} \geq \frac{I_{\{f(k)>2a_{n_k}\}}}{2}\inf_u \mu_u^{(n_k)}(A_{n_k,k}).$$

Since $R_n(U) \leq 1$, by Fatou's lemma

$$\sup_u \limsup_{n\to\infty} R_n(u) \quad\geq\quad \mathbf{E}\{\limsup_{n\to\infty} R_n(U)\}$$

$$
\begin{aligned}
&\geq \limsup_{n\to\infty} \mathbf{E}\{R_n(U)\} \\
&\geq \limsup_{k\to\infty} \mathbf{E}\{R_{n_k}(U)\} \\
&\geq \frac{1}{2}\limsup_{k\to\infty}(I_{\{f(k)>2a_{n_k}\}} \inf_u \mu_u^{(n_k)}(A_{n_k,k})). \quad \square
\end{aligned}
$$

Proof of Theorem 1. Given a distribution vector $\{q_0, q_1, \ldots\}$, we define a "hypercube" subclass $\{\mu_u : u \in \{-1,+1\}^{\mathcal{N}}\}$ of distributions parametrized by a sign vector $u = (u_0, u_1, u_2, \ldots)$. Let K be a random variable on $\mathcal{N}$ with $\mathbf{P}\{K = k\} = q_k$. We find a sequence $\{l_0 = 0, l_1, l_2, \ldots\}$ of distinct positive integers to be specified later, and we consider the sets $B_k = \{-l_k, l_k\}$. For a vector u, the distribution μ_u of X is defined by $X = u_K l_K$. (Now $\mu_u(B_k) = q_k$ independently of u.)

For this distribution, it is easy to verify that

$$
m = m(u) = \sum_{j=1}^{\infty} u_j q_j l_j,
$$

and thus $|m(u)| \leq \sum_{j=1}^{\infty} q_j l_j$.

If u and u' differs only in the k^{th} sign, then $|m(u) - m(u')| = 2q_k l_k$. Observe that if no sample falls on B_k, then the estimate has no information about u_k, which gives $2q_k l_k$ contribution to the error.

We apply Lemma 1 for $A_{n,k} = \{x_1^n : \nexists i, x_i \in B_k\}$ and $f(k) = 2q_k l_k$. Now

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

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

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

Hence by Lemma 1, we get that for any sequence $\{m_n\}$ of estimates, for any sequence $\{a_n\}$ of positive numbers converging to zero and for any subsequence $\{n_k\}$,

$$
\sup_u \limsup_{n\to\infty} \mathbf{P}\{|m_n(D_n(u)) - m(u)| > a_n\} \geq \frac{1}{2}\limsup_{k\to\infty}(I_{\{q_k l_k > a_{n_k}\}}(1 - q_k)^{n_k}).
$$

We can choose $\{n_k\}$ such that $\sum_k a_{n_k} < \infty$, and choose the distribution vector $\{q_k\}$ such that $q_k = o(1/n_k)$ as $k \to \infty$ and $\sum_k k q_k < \infty$ (e.g., $q_k = \min(a_{n_k}/n_k, 2^{-k})$, $q_0 = 1 - \sum_{k=1}^{\infty} q_k$ is suitable). Setting

$$
l_k = \min\{z \in \mathcal{N} : z > a_{n_k}/q_k, z \neq l_1, \ldots, l_{k-1}\},
$$

we get $a_{n_k}/q_k < l_k \leq a_{n_k}/q_k + k$, so

$$
a_{n_k} < q_k l_k \leq a_{n_k} + k q_k.
$$

Hence, on the one hand $|m(u)| \le \sum_{k=1}^{\infty} a_{n_k} + \sum_{k=1}^{\infty} k q_k < \infty$, that is, $m(u)$ is finite, on the other hand

$$\sup_u \limsup_{n\to\infty} \mathbf{P}\{|m_n(D_n(u)) - m(u)| > a_n\} \ge \frac{1}{2} \limsup_{k\to\infty} (1-q_k)^{n_k} = \frac{1}{2}.$$

So for all $\{a_n\}$, there is $\{q_k\}$, $\{l_k\}$ and u such that for any $\epsilon > 0$

$$\mathbf{P}\{|m_n(D_n(u)) - m(u)| > a_n\} > \frac{1}{2} - \epsilon \quad \text{infinitely often.} \quad \square$$

3 Rate of convergence

According to the results of the previous section, in order to get rate-of-convergence results, we have to assume conditions on the distribution.

Examine the absolute moment parameter $m^{(p)} = \mathbf{E}\{|X|^p\}$ $(p > 1)$ of the distributions in the class used in the proof of Theorem 1. For every distribution in the proof (by rough calculation)

$$m^{(p)} = \sum_k q_k l_k^p = \sum_k (q_k l_k)^p / q_k^{p-1} \asymp \sum_k a_{n_k}^p n_k^{p-1}/o(1) + \sum_k k^p o(1)/n_k.$$

The latest sums can be made finite by the choice of $\{n_k\}$ if and only if $a_n^p n^{p-1} \to 0$, that is, if $a_n = o(n^{-\frac{p-1}{p}})$. This suggests that maybe the rate $\{n^{-\frac{p-1}{p}}\}$ can be achieved if we restrict ourselves for the $m^{(p)} < \infty$ case.

Indeed, the theorem below implies the upper bound of $O(n^{-\frac{p-1}{p}})$ for the L_1 error of $\{\widehat{m}_n\}$:

Theorem 2 *(*BAHR AND ESSEEN [6], SEE THEOREM 2.6.20 IN PETROV [13]*) For $1 \le p \le 2$, if $\mathbf{E}\{|X|^p\} < \infty$, then for $\widehat{m}_n = \sum_{i=1}^n X_i/n$,*

$$\mathbf{E}\{|\widehat{m}_n - m|\} \le (\mathbf{E}\{|\widehat{m}_n - m|^p\})^{1/p} \le \frac{(2\mathbf{E}\{|X|^p\})^{1/p}}{n^{1-1/p}}. \quad \square$$

This means that for the class $\mathcal{D}^p$ of distributions with finite p^{th} absolute moment, $\{n^{-\frac{p-1}{p}}\}$ is a rate of convergence of the average $\{\widehat{m}_n\}$ with constants depending on the actual distribution. (As we seen in Section 1, no rate of convergence is possible with a fixed constant.) We show that the rate above is sharp, that is, for *any* $\{m_n\}$ the rate of convergence of the error can be arbitrary close to $\{n^{-\frac{p-1}{p}}\}$:

Theorem 3 *For any sequence $\{m_n\}$ of estimates and for any sequence $a_n = o(n^{-\frac{p-1}{p}})$ of positive numbers, a distribution may be found on $\mathcal{X}$ in $\mathcal{D}^p$ to be the distribution of X such that for any $\epsilon > 0$*

$$\mathbf{P}\{|m_n - m| > a_n\} > \frac{1}{2} - \epsilon \quad \textit{infinitely often.}$$

REMARK. Applied to the sequence $\left\{\sqrt{a_n n^{\frac{p-1}{p}}}\right\}$ instead of $\{a_n\}$, Theorem 3 implies that for any $\{m_n\}$ there is a distribution on $\mathcal{X}$ in $\mathcal{D}^p$ such that for any $K > 0$

$$\limsup_{n\to\infty} \mathbf{P}\{|m_n - m| > K a_n\} \geq \frac{1}{2}. \quad \Box$$

REMARK. The combination of Theorems 2 and 3 implies that for $\mathcal{D}^p$ and for the L_1 error one cannot get better rate of convergence than $\{\widehat{m}_n\}$. The only gain for another estimate would be with the multiplicative constant. $\Box$

Thus we get the analogous result for the expected error:

Corollary 2 *For any sequence $\{m_n\}$ of estimates and for any sequence $a_n = o(n^{-\frac{p-1}{p}})$ of positive numbers, a distribution may be found on $\mathcal{X}$ in $\mathcal{D}^p$ to be the distribution of X such that*

$$\limsup_{n\to\infty} \frac{\mathbf{E}\{|m_n - m|\}}{a_n} = \infty. \quad \Box$$

REMARK. We note that similar individual rate-of-convergence results were obtained in density estimation (Birgé [7]), regression estimation (Antos, Györfi and Kohler [4]) and pattern recognition (Antos and Lugosi [5] and Antos [1], see also Antos [2]). $\Box$

REMARK. Examine also the parameter $l^{(p)} = \mathbf{E}\{|X|\log_2^p |X|\}$ of the distributions in the class used in the proof of Theorem 1. For every distribution (by rough calculation)

$$\begin{aligned} l^{(p)} &= \sum_k q_k l_k \log_2^p l_k = \sum_k q_k l_k (\log_2(q_k l_k) - \log_2 q_k)^p \\ &\asymp \sum_k a_{n_k}\left(\log_2 a_{n_k} + \log_2 \frac{n_k}{o(1)}\right)^p \asymp \sum_k a_{n_k}\log_2^p a_{n_k} + \sum_k a_{n_k}\log_2^p \frac{n_k}{o(1)}. \end{aligned}$$

The latest sums can be made finite by the choice of $\{n_k\}$ if and only if $a_n \log_2^p n \to 0$, that is, if $a_n = o(\frac{1}{\log^p n})$. This suggests that the rate $\{\frac{1}{\log^p n}\}$ can be achieved if we restrict ourselves for the $l^{(p)} < \infty$ case. $\Box$

PROOF OF THEOREM 3. Let $a_n = b_n n^{-\frac{p-1}{p}}$, where $b_n \to 0$. The proof is the same as that of Theorem 1, but we assure that the class of distributions is a subclass of $\mathcal{D}^p$.

Choose $\{n_k\}$ such that

$$\sum_k b_{n_k} < \infty \quad , \quad \sum_k b_{n_k}/n_k \leq 1 \quad , \quad \sum_k k^p b_{n_k}/n_k < \infty,$$

choose the distribution vector $\{q_k\}$ such that for $k \geq 1$, $q_k = b_{n_k}/n_k$ ($q_0 = 1 - \sum_{k=1}^{\infty} q_k$), and choose l_k the same way as in Theorem 1. Now we have

$$l_k^p \leq (a_{n_k}/q_k + k)^p = (a_{n_k} n_k / b_{n_k} + k)^p = (n_k^{1/p} + k)^p \leq 2^{p-1}(n_k + k^p),$$

so

$$q_k l_k^p \leq 2^{p-1} q_k (n_k + k^p) = 2^{p-1}(b_{n_k} + k^p b_{n_k}/n_k).$$

Hence, on the one hand

$$m^{(p)} \leq 2^{p-1} \left(\sum_{k=1}^{\infty} b_{n_k} + \sum_{k=1}^{\infty} k^p b_{n_k}/n_k \right) < \infty,$$

that is, $m^{(p)}$ is finite, on the other hand

$$\sup_u \limsup_{n\to\infty} \mathbf{P}\{|m_n(D_n(u)) - m(u)| > a_n\} \geq \frac{1}{2} \limsup_{k\to\infty} (1-q_k)^{n_k} = \frac{1}{2}.$$

So for all $\{a_n\}$, there is $\{q_k\}$, $\{l_k\}$ and u such that for any $\epsilon > 0$

$$\mathbf{P}\{|m_n(D_n(u)) - m(u)| > a_n\} > \frac{1}{2} - \epsilon \quad \text{infinitely often.} \quad \square$$

References

[1] A. Antos. Lower bounds on the rate of convergence of nonparametric pattern recognition. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99, Proceedings*, volume 1572 of *LNAI/LNCS*, pages 241–252, Berlin, 1999. Springer. Submitted to Theoretical Computer Science.

[2] A. Antos. *Performance limits of nonparametric estimators.* PhD thesis, Technical University of Budapest, H-1521 Stoczek u. 2, Budapest, Hungary, 1999.

[3] A. Antos, L. Devroye, and L. Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:643–645, 1999.

[4] A. Antos, L. Györfi, and M. Kohler. Lower bounds on the rate of convergence of nonparametric regression estimates. *Journal of Statistical Planning and Inference*, 83:91–100, 2000.

[5] A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Machine Learning*, 30:31–56, 1998.

[6] B. Von Bahr and C.G. Esseen. Inequalities for the rth absolute moment of a sum of independent random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.*, 36:299–303, 1965.

[7] L. Birgé . On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, 71:271-291, 1986.

[8] T. M. Cover. Rates of convergence for nearest neighbor procedures. In *Proceedings of the First Annual Hawaii International Conference on Systems Sciences*, pages 413–415, Honolulu, HI, 1968.

[9] L. Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4:154–157, 1982.

[10] L. Devroye. On arbitrarily slow rates of global convergence in density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 62:475-483, 1983.

[11] L. Devroye. Another proof of a slow convergence result of Birgé. *Statistics and Probability Letters*, 23:63–67, 1995.

[12] L. Devroye, L. Györfi, and G. Lugosi. *Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[13] V.V. Petrov. *Limit theorems of probability theory.* Clarendon Press, Oxford, 1995.