EXPONENTIAL INEQUALITIES IN NONPARAMETRIC ESTIMATION

Luc Devroye
Division of Statistics
University of California at Davis
Davis, CA. 95616

ABSTRACT. We derive exponential inequalities for the oscillation of functions of random variables about their mean. This is illustrated on the Kolmogorov-Smirnov statistic, the total variation distance for empirical measures, the Vapnik-Chervonenkis distance, and various performance criteria in nonparametric density estimation. We also derive bounds for the variances of these quantities.

1. Introduction.

Hoeffding (1963) showed that for independent random variables $X_1, X_2, \ldots, X_n$ with $a_i \le X_i \le b_i$,

$$\mathbf{P}\left\{\left|\sum_{i=1}^{n}(X_i - \mathbf{E}X_i)\right| > t\right\} \le 2e^{-2t^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \ , \ t > 0.$$

Perhaps the best known form of this inequality is obtained when the X_i's are i.i.d. Bernoulli (p) random variables. In that case, we obtain Chernoff's bound (1952) for the binomial distribution: if X is binomial (n, p), then

$$\mathbf{P}\{|X - np| > t\} \le 2e^{-2t^2/n} \ , \ t > 0.$$

Various extensions of these inequalities have been developed over the years. The generalization to martingales due to Hoeffding (1963) and Azuma (1967) has led to interesting applications in combinatorics and the theory of random graphs (for a survey, see McDiarmid, 1989). We have used it in density estimation (Devroye, 1988, 1989).

In this paper, we collect various extensions of Hoeffding's inequality and highlight their applications in the nonparametric estimation of densities and distribution functions. For completeness, the proofs of the inequalities are sketched as well. In the last section, we present new bounds for the variance of functions of independent random variables. The inequalities are illustrated on examples in nonparametric estimation, and are shown to be sharper than those obtained from the Efron-Stein inequality.

2. Inequalities for martingale difference sequences.

Hoeffding (1963) mentioned that his inequalities would also be valid when applied to martingale difference sequences. To make things a bit more precise, let us consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and a nested sequence $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$ of sub-σ-fields of $\mathcal{F}$. A sequence of integrable random variables $X_0, X_1, X_2, \ldots$ is a *martingale* if

$$\mathbf{E}\left\{X_{n+1} \mid \mathcal{F}_n\right\} = X_n \quad \text{a.s. ,} \quad \text{each } n \geq 0.$$

A sequence of integrable random variables $Y_1, Y_2, \ldots$ is a *martingale difference sequence* if for every $n \geq 0$,

$$\mathbf{E}\left\{Y_{n+1} \mid \mathcal{F}_n\right\} = 0 \quad \text{a.s. .}$$

Note that any martingale $X_0, X_1, X_2, \ldots$ leads to a natural martingale difference sequence by defining

$$Y_n = X_n - X_{n-1}, \quad n \geq 1.$$

And any martingale difference sequence $Y_1, Y_2, \ldots$ in turn yields a natural martingale by defining X_0 in an arbitrary fashion and setting

$$X_n = \sum_{i=1}^{n} Y_i + X_0.$$

For any nested sequence of sub-σ-fileds $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}$ and any integrable random variable X, we can define *Doob's martingale* by setting

$$X_n = \mathbf{E}\left\{X \mid \mathcal{F}_n\right\}, \quad n \geq 0.$$

Thus, $X_0 = \mathbf{E}X$, and if X is $\mathcal{F}_n$-measurable, then $X_n = X$, and

$$X - \mathbf{E}X = \sum_{i=1}^{n} (X_i - X_{i-1}) \ .$$

We begin with an inequality along the lines suggested by Hoeffding (1963) and Azuma (1967) (see McDiarmid, 1989):

THEOREM 1. *Let $\mathcal{F}_0 = \{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_n$ be a nested sequence of σ-fields. Let the integrable random variable X be $\mathcal{F}_n$-measurable, and define the Doob martingale $X_k = \mathbf{E}\left\{X \mid \mathcal{F}_k\right\}$. Assume that for $k = 1, \ldots n$, there exist random variables Z_k, $\mathcal{F}_{k-1}$-measurable, and constants c_k such that*

$$Z_k \leq X_k \leq Z_k + c_k.$$

Then for $t > 0$,

$$\mathbf{P}\left\{X - \mathbf{E}X \geq t\right\} \leq 2e^{-2t^2 / \sum_{i=1}^{n} c_i^2} \ , \ t > 0,$$
$$\mathbf{P}\left\{X - \mathbf{E}X \leq -t\right\} \leq 2e^{-2t^2 / \sum_{i=1}^{n} c_i^2} \ , \ t > 0.$$

This Theorem uses a simple Lemma due to Hoeffding (1963):

LEMMA 1. *Let X be a random variable with $\mathbb{E}X = 0$, $a \le X \le b$. Then for $\lambda > 0$,*

$$\mathbb{E}\left\{e^{\lambda X}\right\} \le e^{\lambda^2 (b-a)^2/8}.$$

PROOF. Note that by convexity,

$$\begin{aligned}
e^{\lambda x} &\le \frac{x-a}{b-a} e^{\lambda b} + \frac{b-x}{b-a} e^{\lambda a} \ , \ a \le x \le b \ , \\
\mathbb{E}e^{\lambda X} &\le \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} \\
&= (1 - p + p e^{\lambda(b-a)}) e^{-p\lambda(b-a)} \ , \text{ where } p = \frac{a}{a-b} \in [0,1] \\
&\stackrel{\text{def}}{=} e^{\varphi(u)} \ ,
\end{aligned}$$

where $u = \lambda(b-a)$, $\varphi(u) = -pu + \log(1 - p + pe^u)$. But it is easy to see that

$$\begin{aligned}
\varphi'(u) &= -p + \frac{p}{p + (1-p)e^{-u}} \ , \\
\varphi''(u) &= \frac{p(1-p)e^{-u}}{(p + (1-p)e^{-u})^2} \le \frac{1}{4} \ , \\
\varphi(0) &= \varphi'(0) = 0 \ ,
\end{aligned}$$

and by Taylor's series expansion with remainder,

$$\varphi(u) \le \frac{u^2}{8} = \frac{\lambda^2 (b-a)^2}{8} \ . \ \square$$

PROOF. Set $Y_k = X_k - X_{k-1}$, $S_k = \sum_{i=1}^{k} Y_i = X_k - X_0$. Note that $S_n = X_n - \mathbb{E}X = X - \mathbb{E}X$. Also,

$$\begin{aligned}
\mathbb{P}\{X - \mathbb{E}X \ge t\} &= \mathbb{P}\{S_n \ge t\} \\
&\le e^{-\lambda t} \mathbb{E}\left\{e^{\lambda S_n}\right\} \quad \text{for} \lambda > 0 \quad \text{(by Chernoff's bounding method)} \\
&= e^{-\lambda t} \mathbb{E}\left\{e^{\lambda S_{n-1}} \mathbb{E}\left\{e^{\lambda Y_n} \middle| \mathcal{F}_{n-1}\right\}\right\} \\
&\le e^{-\lambda t} \mathbb{E}\left\{e^{\lambda S_{n-1}}\right\} e^{\lambda^2 c_n^2/8} \quad \text{(Lemma 1)} \\
&\le e^{-\lambda t} e^{(\lambda^2/8) \sum_{i=1}^{n} c_i^2} \quad \text{(iterate previous argument)} \\
&= e^{-2t^2 / \sum_{i=1}^{n} c_i^2} \quad (\text{take } \lambda = 4t / \sum_{i=1}^{n} c_i^2) \ .
\end{aligned}$$

The second inequality in Theorem 1 is obtained when we replace X by $-X$. $\square$

3. McDiarmid's extension of Hoeffding's inequality.

The following extension of Hoeffding's inequality is useful for random variables that are complicated functions of independent random variables, and that are relatively robust to individual changes in the values of the random variables.

THEOREM 2. *(McDiarmid, 1989) Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n \\ x'_1,\ldots,x'_n \in A}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \le c_i \ , \ 1 \le i \le n \ .$$

Then

$$\mathbb{P}\left\{|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \ge t\right\} \le 2e^{-2t^2 \big/ \sum_{i=1}^n c_i^2} \ .$$

PROOF. Define $Y = Y_n = f(X_1, \ldots, X_n)$, and let $\mathcal{F}_n$ be the σ-field generated by $X_1, \ldots, X_n$. Define

$$Y_k = \mathbb{E}\left\{Y \middle| \mathcal{F}_k\right\} \ , \ Z_k = \operatorname{ess\,inf}\left\{Y_k \middle| \mathcal{F}_{k-1}\right\} \ , \ W_k = \operatorname{ess\,sup}\left\{Y_k \middle| \mathcal{F}_{k-1}\right\} \ ,$$

so that $Z_k \le Y_k \le W_k$. We can apply Theorem 1 directly to Y_n if we can show that $W_k - Z_k \le c_k$. But this follows from

$$\begin{aligned}
W_k &= \operatorname{ess\,sup}\left\{\mathbb{E}\left\{f(X_1, \ldots, X_n)\middle|\mathcal{F}_k\right\} \middle| \mathcal{F}_{k-1}\right\} \\
&\le \sup_{x \in A} \mathbb{E}\left\{f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots, X_n)\middle|\mathcal{F}_k\right\} \\
&= \sup_{x \in A} \mathbb{E}\left\{f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots, X_n)\middle|\mathcal{F}_{k-1}\right\} \\
&\le \inf_{x \in A} \mathbb{E}\left\{f(X_1, \ldots, X_{k-1}, x, X_{k+1}, \ldots, X_n)\middle|\mathcal{F}_{k-1}\right\} + c_k \\
&\le \operatorname{ess\,inf}\left\{\mathbb{E}\left\{f(X_1, \ldots, X_n)\middle|\mathcal{F}_k\right\} \middle| \mathcal{F}_{k-1}\right\} + c_k \\
&= Z_k + c_k \ . \ \square
\end{aligned}$$

4. Applications.

4.1. Chernoff's bound. Let $X_1, \ldots, X_n$ be Bernoulli (p_1) , ..., Bernoulli (p_n) random variables, and consider $\sum_{i=1}^n X_i$. Clearly, the conditions of Theorem 2 are fulfilled with $c_i \equiv 1$. Thus,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - p_i)\right| \ge t\right\} \le 2e^{-2t^2/n} \ .$$

As a special case, $p_i \equiv p$ for all i, we obtain Chernoff's bound (1952) for a binomial (n, p) random variable X:

$$\mathbb{P}\left\{|X - \mathbb{E}X| \ge t\right\} \le 2e^{-2t^2/n} \ .$$

4.2. The Kolmogorov-Smirnov statistic. Let F_n be the standard empirical distribution function based upon an i.i.d. sample $X_1, \ldots, X_n$ drawn from a distribution function F on the real line. The Kolmogorov-Smirnov distance is

$$\sup_x |F_n(x) - F(x)| \ .$$

Note that changing one point X_i can increase or decrease F_n by at most $1/n$ over a certain interval. Thus the condition of Theorem 2 is fulfilled with $c_i \equiv 1/n$. We have

$$\mathbb{P}\left\{ \left| \sup_x |F_n(x) - F(x)| - \mathbb{E} \sup_x |F_n(x) - F(x)| \right| \geq \frac{t}{\sqrt{n}} \right\} \leq 2e^{-2t^2} \ .$$

In this respect, we note that Dvoretzky, Kiefer and Wolfowitz (1956) showed that

$$\mathbb{P}\left\{ \sup_x |F_n(x) - F(x)| \geq \frac{t}{\sqrt{n}} \right\} \leq Ce^{-2t^2}$$

for some $C > 0$, while Massart (1990) proved that one can take $C = 2$. Massart's bound and the inequality derived from Theorem 2 do not imply each other.

4.3. Nonparametric density estimation: the L_1 norm. Let f_n be the Parzen-Rosenblatt kernel estimate of a density f based upon an i.i.d. sample $X_1, \ldots, X_n$ drawn from f (Rosenblatt, 1956; Parzen, 1962):

$$f_n(x) = f_n(x; X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i) \ .$$

Here K is a given function (the kernel) integrating to one, $K_h(u) = \frac{1}{h} K(\frac{u}{h})$, and $h > 0$ is a smoothing factor. An important criterion for evaluating the performance of a density estimate is $\int |f_n - f|$. This random variable satisfies the conditions of Theorem 2 with $c_i \equiv 2\int |K|/n$ as we will now see. Take any numbers $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ with $x'_i = x_i$ except for $i = j$. Then,

$$\begin{aligned}
\Big| \int |f_n(x; x_1, \ldots, x_n) - f(x)| \, dx &- \int \left| f_n(x; x'_1, \ldots, x'_n) - f(x) \right| \, dx \Big| \\
&\leq \int \left| f_n(x; x_1, \ldots, x_n) - f_n(x; x'_1, \ldots, x'_n) \right| \, dx \\
&\leq \frac{1}{n} \int \left| K_h(x - x_j) - K_h(x - x'_j) \right| \, dx \\
&\leq \frac{2 \int |K|}{n} \ .
\end{aligned}$$

Thus, dropping the $(.)$ and dx, we have

$$\mathbb{P}\left\{ \left| \int |f_n - f| - \mathbb{E} \int |f_n - f| \right| > t \right\} \leq 2e^{-nt^2/2\int^2 |K|} \ .$$

This inequality improves over one first published in Devroye (1988), where the exponent had the constant 32 instead of 2. The present improvement was independently pointed out to me by Pinelis (1990).

REMARK 1. We recall that $\sqrt{n}\mathbf{E}\int|f_n - f| \to \infty$ for the kernel estimate when one of these conditions holds:

(1) $\lim_{n\to\infty} h = 0$;
(2) The characteristic function for f has unbounded support;
(3) $\int\sqrt{f} = \infty$.

See e.g. Devroye and Györfi (1985) or Devroye (1988). When this is the case, a simple application of Chebyshev's inequality shows that

$$\frac{\int|f_n - f|}{\mathbf{E}\int|f_n - f|} \to 1 \quad \text{in probability.}$$

In other words, the L_1 eror behaves asymptotically like a deterministic sequence, just as averages do when the weak law of large numbers applies.

REMARK 2. For the standard histogram estimate, regardless of the bin width, we have

$$\mathbf{P}\left\{\left|\int|f_n - f| - \mathbf{E}\int|f_n - f|\right| > t\right\} \le 2e^{-nt^2/2} .$$

Thus, just as for the kernel estimate, we have an inequality that is valid for all f and n, and for all choices of the bin widths and the smoothing factors. The non-asymptotic character of the inequalities will undoubtedly make them useful tools for further applications.

REMARK 3. By the boundedness of $\int|f_n - f|$, we note that $\int|f_n - f| \to 0$ in probability if and only if $\mathbf{E}\int|f_n - f| \to 0$. But if these quantities tend to zero in the indicated senses, by the results of this section, for every $\epsilon > 0$ and $t > 0$, it is possible to find n_0 such that for $n > n_0$,

$$\mathbf{P}\left\{\int|f_n - f| > t\right\} \le 2e^{-n(1-\epsilon)t^2/2\int^2|K|} .$$

Thus, weak convergence of the L_1 error implies complete convergence. This observation is at the basis of the equivalence results of Devroye (1983), but the present proof is much shorter. We note that a sufficient condition for the weak (and thus complete) convergence for the kernel estimate is that $h \to 0$ and $nh \to \infty$ as $n \to \infty$ (see Devroye, 1987, where a proof is given based upon results of Scheffé (1947) and Glick (1974)). When the kernel K has at least one non-vanishing moment, then these conditions are necessary as well.

REMARK 4. The condition in Theorem 2 demands that each individual sample point have a limited influence on $\int|f_n - f|$. This may not be the case for some data-based methods of choosing h.

4.4. L_p **norms.** Define the L_p norm of g by $\|g\|_p = \left(\int |g|^p\right)^{1/p}$, where $p \geq 1$ is fixed. If f_n is shorthand for $f_n(x; x_1, \ldots, x_n)$, $g_n \equiv f_n(x; x'_1, \ldots, x'_n)$ and if $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ are sequences of numbers with $x_i = x'_i$ except for $i = j$, then

$$\begin{aligned} \left| \|f_n - f\|_p - \|g_n - f\|_p \right| &\leq \|f_n - g_n\|_p \\ &\leq \left\| \frac{1}{nh} K_h(x - x_j) - \frac{1}{nh} K_h(x - x'_j) \right\|_p \\ &\leq \frac{2}{nh^{1-1/p} \|K\|_p} \end{aligned}$$

by Minkowski's inequality. From Theorem 2, we then deduce the following inequality:

$$\mathbf{P}\left\{ \left| \|f_n - f\|_p - \mathbf{E}\,\|f_n - f\|_p \right| \geq t \right\} \leq 2e^{-nt^2 h^{2-2/p}/(2\|K\|_p^2)} \ .$$

The inequality remains formally valid even if $p = \infty$, in which case we obtain the supremum norm.

Assume for the sake of simplicity that K is a bona fide density. We claim that the relative stability result, i.e.,

$$\frac{\|f_n - f\|_p}{\mathbf{E}\,\|f_n - f\|_p} \to 1 \quad \text{in probability,}$$

holds whenever $h \to 0$, $nh \to \infty$ and $1 \leq p < 2$. Of course, for the norms to make sense, we have to additionally assume that $K, f \in L_p$. Assume for simplicity that $K \geq 0$. To prove the claim, we first havve to establish that for any density f, there exists a constant $c > 0$ such that

$$\mathbf{E}\,\|f_n - f\|_p \geq c \max\left(h^2, 1/\sqrt{nh}\right) \ .$$

Under smoothness and tail conditions on f, this result is rather standard. The generalization to all f requires some work. Back to the statement. It clearly suffices to show that the variance of $\|f_n - f\|_p$ is $o(\mathbf{E}^2\,\|f_n - f\|_p)$ by Chebyshev's inequality. The variance is $O(1/(\sqrt{n}h^{1-1/p}))$ — this follows from the exponential bound shown above. If $h \leq n^{-1/5}$, then the statement is easily verified since $h^{\frac{1}{2}-\frac{1}{p}} \to \infty$. If $h \geq n^{-1/5}$, then we need only verify that $\sqrt{n}h^{3-1/p} \to \infty$.

Interestingly, the first p for which the relative stability result fails is $p = 2$. We can only obtain it from the inequality shown above when $nh^5 \to \infty$, a condition that is known to yield suboptimal values for h for all densities (not just all smooth ones!). However, this does not mean that the relative stability result is not valid in L_2 for $h \sim n^{-1/5}$. Indeed, Hall (1982) proved that

$$\frac{\|f_n - f\|_2^2}{\mathbf{E}\,\|f_n - f\|_2^2} \to 1 \quad \text{in probability,}$$

under certain conditions on f, h and K.

4.5. Uniform deviation of empirical measures. An i.i.d. sample$X_1, \ldots, X_n$ with common probability measure μ on the Borel sets $\mathcal{B}$ of R^d induces an empirical probability measure μ_n by

$$\mu_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(X_i) \ ,$$

where I is the indicator function, and $B \in \mathcal{B}$. The total variation distance between μ_n and μ is

$$T_n \stackrel{\text{def}}{=} \sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| \ .$$

Clearly, $T_n \equiv 1$ if μ is nonatomic, so the total variation distance is rather restrictive. Vapnik and Chervonenkis (1971) considered instead

$$V_n \stackrel{\text{def}}{=} \sup_{A \in \mathcal{A}} |\mu_n(B) - \mu(B)|$$

where $\mathcal{A}$ is a suitable subclass of the Borel sets. For example, if $\mathcal{A} = \{(-\infty, x] : x \in R\}$ and $d = 1$, then V_n is the standard Kolmogorov-Smirnov distance discussed above. They showed in particular that

$$\mathbb{P}\{V_n \geq t\} \leq 4s(\mathcal{A}, 2n)e^{-nt^2/8} \ , \ nt^2 \geq 1,$$

where

$$s(\mathcal{A}, n) = \max_{x_1,\ldots,x_n) \in R^{dn}} N_{\mathcal{A}}(x_1, \ldots, x_n)$$

and $N_{\mathcal{A}}(x_1, \ldots, x_n)$ is the number of different sets in

$$\big\{\{x_1, \ldots, x_n\} \cap A \ \big| \ A \in \mathcal{A}\big\} \ .$$

For many families $\mathcal{A}$ that are not too rich, such as all halfspaces, all intersections of a finite number of halfspaces, all balls, etc., $s(\mathcal{A}, n) \leq n^D$ for a finite D (the "Vapnik-Chervonenkis dimension"), so that the Vapnik-Chervonenkis bound decreases exponentially with n. For extensions and discussions, see Devroye (1982), Gaenssler (1983), Pollard (1984) and Alexander (1984).

If we replace X_j in the sample by X_j' while holding all the other elements fixed, V_n changes by at most $1/n$, so that from Theorem 2,

$$\mathbb{P}\{|V_n - \mathbb{E}V_n| \geq t\} \leq 2e^{-2nt^2} \ .$$

This implies that $\sqrt{n}(V_n - \mathbb{E}V_n) = O(1)$ in probability. For the limit law theory of $\sqrt{n}V_n$, we refer to Dudley (1978).

4.6. Variable kernel estimates. Breiman, Meisel and Purcell (1977) introduced the variable kernel estimate

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_i}(x - X_i) \ ,$$

where H_i is a function of X_i and the data, $X_1, \ldots, X_n$. If H_i is a function of X_i and n only, then the inequality of section 4.3 applies unaltered.

A more interesting choice of H_i is that in which it becomes a function of the distance between X_i and its k-th nearest neighbor among the data. Replacing X_i by X_i' affects at most ck of the H_j's, where c is some universal constant depending upon the dimension only. This is seen by noting that X_i

can be among the k nearest neighbors of at most $c'k$ of the X_j's, where c' depends upon d only (Devroye and Penrod, 1986). Thus, $\int |f_n - f|$ changes by at most ck/n. Hence,

$$\mathrm{P}\left\{\left|\int |f_n - f| - \mathrm{E}\int |f_n - f|\right| > t\right\} \le 2e^{-2nt^2/(c^2k^2)} \ .$$

Thus, $\mathrm{V}\left\{\int |f_n - f|\right\} = O(k^2/n)$. Depending upon the choice of k, this can be used to establish the relative stability of the estimate.

4.7. Recursive kernel density estimates. The bound of section 4.2 remains valid for density estimates of the form

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_i}(x - X_i) \ ,$$

where h_i depends upon i only; this is an estimate attributed to Wolverton and Wagner (1969). Consider estimates of the form

$$f_n(x) = \sum_{i=1}^{n} p_i K_{h_i}(x - X_i) \ ,$$

where $(p_1, \ldots, p_n)$ is a probability weight vector, and both p_i and h_i depend upon i and possibly n. This general form of estimate goes back to Deheuvels (1973). The condition of Theorem 2 holds with $c_i \equiv 2p_i \int |K|$. Clearly,

$$\sum_{i=1}^{n} c_i^2 = 4\sum_{i=1}^{n} p_i^2 \left(\int |K|\right)^2 \le 4 \max_{1\le i\le n} p_i \left(\int |K|\right)^2 \ .$$

This in turn can be used in the upper bound of Theorem 2.

Deheuvels (1973) proposed the latter estimate, based upon a fixed sequence $h_1, h_2, \ldots$, with $p_j = h_j / \sum_{i=1}^{n} h_i$. Assume furthermore that h_n oscillates so slightly that $\sum_{i=1}^{n} h_i^2 \le Anh_n^2$ and $\sum_{i=1}^{n} h_i \ge Bnh_n$ for some positive constants A, B. Then we see that upon proper substitution of the various bounds,

$$\mathrm{P}\left\{\left|\int |f_n - f| - \mathrm{E}\int |f_n - f|\right| > t\right\} \le 2e^{-Bnt^2/(2A\int^2 |K|)} \ .$$

Deheuvels' estimate is relatively stable for all f, whenever $h \to 0$, $nh \to \infty$, K has at least one non-vanishing moment, and h satisfies the regularity condition mentioned above.

5. Inequalities for the variance.

In many applications, one would like to obtain information about the variance, or the oscillation, of a random variable of the form $f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are i.i.d. random vectors. Often, $f(\cdot)$ is a rather complicated function of the data (see the examples in the previous section). One of the first general tools in this respect is the Efron-Stein inequality (Efron and Stein, 1981; see also Vitale, 1984).

THE EFRON-STEIN INEQUALITY. *Let f be a symmetric function of its n arguments, and let $X_1, \ldots, X_{n+1}$ be i.i.d. random vectors. Define*

$$\begin{aligned} S_i &= f(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}) \ , \\ S &= S_{n+1} = f(X_1, \ldots, X_n) \ , \end{aligned}$$

and

$$\overline{S} = \frac{1}{n+1} \sum_{i=1}^{n+1} S_i \ .$$

Then

$$\mathbb{V}\{S\} \le \sum_{i=1}^{n+1} \mathbb{E}\left\{ \left(S_i - \overline{S}\right)^2 \right\} = (n+1)\mathbb{E}\left\{ \left(S - \overline{S}\right)^2 \right\} \ .$$

When the right-hand-side of the inequality is worked out, and some further bounding is used, we obtain the following result:

$$\begin{aligned} \mathbb{V}\{S\} &\le (n+1)\mathbb{E}\left\{ \left(\frac{1}{n+1}\right)^2 \left(\sum_{i=1}^{n+1} (S - S_i)\right)^2 \right\} \\ &\le \frac{1}{n+1}\mathbb{E}\left\{ (n+1) \sum_{i=1}^{n+1} (S - S_i)^2 \right\} \\ &= \sum_{i=1}^{n+1} \mathbb{E}\left\{ (S - S_i)^2 \right\} \\ &= \sum_{i=1}^{n} \mathbb{E}\left\{ (S - S_i)^2 \right\} \\ &= n\mathbb{E}\left\{ (S_2 - S_1)^2 \right\} \ . \end{aligned}$$

Assume next that a condition similar to that of Theorem 2 holds:

$$\sup_{\substack{x_1, \ldots, x_n \\ x'_1, \ldots, x'_n \in A}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \le c \ , \ 1 \le i \le n \ . \qquad (*)$$

Then $|S_2 - S_1| \le c$, and thus, $\mathbb{V}\{S\} \le nc^2$.

The derivation given here is that of Devroye (1987), where it was used to show that for the kernel estimate, regardless of the choice of h or the nature of f,

$$\mathbb{V}\left\{ \int |f_n - f| \right\} \le \frac{4 \int^2 |K|}{n} \ .$$

In 1986, Steele obtained a related inequality:

STEELE'S INEQUALITY. *Let f be an arbitrary function of its n arguments, and let $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ be i.i.d. random vectors. Define*

$$S_i = f(X_1, \ldots, X_{i-1}, X'_i, X_{i+1}, \ldots, X_n) ,$$

and

$$S = f(X_1, \ldots, X_n) .$$

Then

$$\mathbb{V}\{S\} \le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left\{ (S - S_i)^2 \right\} .$$

Note that the symmetry of f is no longer a requirement. Also, under condition (*), $\mathbb{V}\{S\} \le nc^2/2$. This yields an improvement by a factor of 2 over the Efron-Stein based bound. It is possible to improve these results even further, as in Theorem 3 below.

THEOREM 3. *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to R$ satisfies*

$$\sup_{\substack{x_1,\ldots,x_n \\ x'_1,\ldots,x'_n \in A}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n)| \le c_i , \ 1 \le i \le n .$$

Then

$$\mathbb{V}\{f(X_1, \ldots, X_n)\} \le \frac{1}{4} \sum_{i=1}^{n} c_i^2 .$$

PROOF. Let $\mathcal{F}_i$ be the σ-algebra generated by $X_1, \ldots, X_i$. Let $Y = Y_n = f(X_1, \ldots, X_n)$. Then $Y_i = \mathbb{E}\left\{ Y \mid \mathcal{F}_i \right\}$ forms a Doob martingale process. We formally set, as usual, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, so that $Y_0 = \mathbb{E}Y$. Thus,

$$\begin{aligned}
\mathbb{V}\{Y\} &= \mathbb{E}\left\{ (Y - Y_0)^2 \right\} \\
&= \mathbb{E}\left\{ \left(\sum_{i=1}^{n} (Y_i - Y_{i-1}) \right)^2 \right\} \\
&= \mathbb{E}\left\{ \sum_{i=1}^{n} (Y_i - Y_{i-1})^2 \right\} + 2 \sum_{1 \le i < j \le n} \mathbb{E}\left\{ (Y_i - Y_{i-1})(Y_j - Y_{j-1}) \right\} \\
&= \mathbb{E}\left\{ \sum_{i=1}^{n} (Y_i - Y_{i-1})^2 \right\} ,
\end{aligned}$$

where we used the martingale property to show that the cross product terms are zero: for $i < j$, we have

$$\begin{aligned}
&\mathbb{E}\left\{ (Y_i - Y_{i-1})(Y_j - Y_{j-1}) \mid \mathcal{F}_{j-1} \right\} \\
&\quad = (Y_i - Y_{i-1}) \left(\mathbb{E}\left\{ Y_j | \mathcal{F}_{j-1} \right\} - Y_{j-1} \right) = 0 \quad \text{almost surely.}
\end{aligned}$$

Theorem 3 follows from the above result if we can show that

$$\mathbf{E}\left\{(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}\right\} \le c_i^2/4 \ .$$

To see this, we observe that if

$$Z_i = \operatorname{ess\,inf}\left\{Y_i - Y_{i-1} \middle| \mathcal{F}_{i-1}\right\} \ , \ W_i = \operatorname{ess\,sup}\left\{Y_i - Y_{i-1} \middle| \mathcal{F}_{i-1}\right\} \ ,$$

then, as shown in the proof of Theorem 2, $W_i \le Z_i + c_i$, and thus, given $\mathcal{F}_{i-1}$, $Y_i - Y_{i-1}$ is a zero mean random variable taking values in the set $[Z_i, Z_i + c_i]$. But an arbitary random variable X taking values in a set $[a, b]$ has variance not exceeding $\mathbf{E}(X - (a+b)/2)^2 \le (b-a)^2/4$, so that

$$\mathbf{E}\left\{(Y_i - Y_{i-1})^2 | \mathcal{F}_{i-1}\right\} \le c_i^2/4 \ .$$

This concludes the proof of Theorem 3. $\square$

Remark 1. For the kernel estimate, we obtain

$$\mathbf{V}\left\{\int |f_n - f|\right\} \le \frac{\int^2 |K|}{n} \ ,$$

which is an improvement by a factor of 4 over the inequality shown in Devroye (1987), which was based upon the Efron-Stein bound. This improvement was suggested to me by Pinelis (1990), who mentions a range of inequalities in a much more general framework.

Remark 2. For the histogram estimate, we obtain

$$\mathbf{V}\left\{\int |f_n - f|\right\} \le \frac{1}{n} \ ,$$

Remark 3. Without further work, we also have

$$\mathbf{V}\left\{\sup_x |F_n(x) - F(x)|\right\} \le \frac{1}{4n}$$

for the Kolmogorov-Smirnov distance. Similarly, borrowing the notation of section 4.5, we have

$$\mathbf{V}\left\{\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|\right\} \le \frac{1}{4n} \ .$$

6. Acknowledgements.

The author's research was supported by NSERC grant A3456.

7. References.

K. S. Alexander, "Probability inequalities for empirical processes and a law of the iterated logarithm ," *Annals of Probability* , vol. 12, pp. 1041–1067, 1984.

K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal*, vol. 37, pp. 357–367, 1967.

L. Breiman, W. Meisel, and E. Purcell, "Variable kernel estimates of multivariate densities," *Technometrics*, vol. 19, pp. 135–144, 1977.

H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations ," *Annals of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

P. Deheuvels, "Sur une famille d'estimateurs de la densité d'une variable aléatoire," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 276, pp. 1013–1015, 1973.

P. Deheuvels, "Sur l'estimation séquentielle de la densité," *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 276, pp. 1119–1121, 1973.

L. Devroye, "Bounds for the uniform deviation of empirical measures," *Journal of Multivariate Analysis*, vol. 12, pp. 72–79, 1982.

L. Devroye, "The equivalence of weak, strong and complete convergence in L1 for kernel density estimates," *Annals of Statistics*, vol. 11, pp. 896–904, 1983.

L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

L. Devroye and C. S. Penrod, "The strong uniform convergence of multivariate variable kernel estimates," *Canadian Journal of Statistics*, vol. 14, pp. 211–219, 1986.

L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston, 1987.

L. Devroye, "An application of the Efron-Stein inequality in density estimation," *Annals of Statistics*, vol. 15, pp. 1317–1320, 1987.

L. Devroye, "Asymptotic performance bounds for the kernel estimate," *Annals of Statistics*, vol. 16, pp. 1162–1179, 1988.

L. Devroye, "The kernel estimate is relatively stable," *Probability Theory and Related Fields*, vol. 77, pp. 521–536, 1988.

L. Devroye, "The double kernel method in density estimation," *Annales de l'Institut Henri Poincaré*, vol. 25, pp. 533–580, 1989.

R. M. Dudley, "Central limit theorems for empirical measures," *Annals of Probability*, vol. 6, pp. 899–929, 1978.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of a sample distribution function and of the classical multinomial estimator," *Annals of Mathematical Statistics*, vol. 33, pp. 642–669, 1956.

B. Efron and C. Stein, "The jackknife estimate of variance," *Annals of Statistics*, vol. 9, pp. 586–596, 1981.

P. Gaenssler, *Empirical Processes*, Lecture Notes-Monograph Series, vol. 3, Institute of Mathematical Statistics, Hayward, CA., 1983.

N. Glick, "Sample-based classification procedures related to empiric distributions," *IEEE Transactions on Information Theory*, vol. IT-22, pp. 454–461, 1976.

P. Hall, "Limit theorems for stochastic measures of the accuracy of density estimators," *Stochastic Processes and Applications*, vol. 13, pp. 11–25, 1982.

W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

C. McDiarmid, "On the method of bounded differences," Technical Report, Institute of Economics and Statistics, Oxford University, 1989.

E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics*, vol. 33, pp. 1065–1076, 1962.

I. F. Pinelis, "To the Devroye's estimates for distributions of density estimators," Technical Report, 1990.

D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, New York, 1984.

M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics*, vol. 27, pp. 832–837, 1956.

H. Scheffé, "A useful convergence theorem for probability distributions," *Annals of Mathematical Statistics*, vol. 18, pp. 434–458, 1947.

J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *Annals of Statistics*, vol. 14, pp. 753–758, 1986.

V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

R. A. Vitale, "An expansion for symmetric statistics and the Efron-Stein inequality," in: *Inequalities in Statistics and Probability*, (edited by Y. L. Tong), pp. 112–114, IMS, Hayward, CA., 1984.

C. T. Wolverton and T. J. Wagner, "Recursive estimates of probability densities," *IEEE Transactions on Systems. Science and Cybernetics*, vol. 5, p. 307, 1969.