JOURNAL OF MULTIVARIATE ANALYSIS 12, 72–79 (1982)

Bounds for the Uniform Deviation of Empirical Measures

LUC DEVROYE*,†

McGill University

Communicated by M. Rosenblatt

If $X_1,..., X_n$ are independent identically distributed R^d-valued random vectors with probability measure μ and empirical probability measure μ_n, and if $\mathcal{A}$ is a subset of the Borel sets on R^d, then we show that $P\{\sup_{A\in\mathcal{A}} |\mu_n(A)-\mu(A)| \geqslant \varepsilon\} \leqslant cs(\mathcal{A}, n^2) e^{-2n\varepsilon^2}$, where c is an explicitly given constant, and $s(\mathcal{A}, n)$ is the maximum over all $(x_1,..., x_n) \in R^{dn}$ of the number of different sets in $\{\{x_1,..., x_n\} \cap A \mid A \in \mathcal{A}\}$. The bound strengthens a result due to Vapnik and Chervonenkis.

1. INTRODUCTION

The approximation of a probability measure μ on the Borel sets $\mathscr{B}$ of R^d by an empirical measure μ_n constructed from $X_1,..., X_n$, a sample of independent random vectors with common probability measure μ, has been of interest to statisticians for different applications. The classical empirical measure μ_n is defined by

$$\mu_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(X_1),$$

where I is the indicator function.

Let

$$U_n = \sup_{\mathcal{A}} |\mu_n(A) - \mu(A)|,$$

Received August 4, 1980.

AMS 1970 subject classification: 62H99.

Key Words and phrases: Random vector, empirical measure, probability inequality, uniform consistency.

* The author is with the School of Computer Science, McGill University, 805 Sherbrooke Street West, Montreal H3A 2K6, Canada.

† Research of the author was supported by the Air Force Office of Scientific Research under Grant AFOSR 76-3062 and 77-3385 while the author was at the University of Texas at Austin during the summers of 1979 and 1980.

0047-259X/82/010072–08$02.00/0

where $\mathscr{A}$ is a subclass of $\mathscr{B}$. Steele [12] gives necessary and sufficient conditions for the almost sure convergence to 0 of U_n. Dudley [4] studies the convergence in distribution of $\sqrt{n}\, U_n$, and Gaenssler and Stute [7] give a comprehensive survey of the literature on empirical measures. We want to find good upper bounds for

$$P\{U_n > \varepsilon\},$$

that do not depend upon μ. Obviously, $U_n = 1$ when $\mathscr{A} = \mathscr{B}$ and μ is absolutely continuous with respect to Lebesgue measure. Also, $U_n = 1$ when $\mathscr{A}$ is the class of all convex Borel sets, and μ puts its mass uniformly on the surface of the unit sphere (Rao [10]). These classes are too rich.

On the other hand, if $\mathscr{A} = \{A\}$ is a singleton set, then

$$P\{U_n > \varepsilon\} \leqslant 2e^{-2n\varepsilon^2} \tag{1.1}$$

by Hoeffding's inequality (Hoeffding, [15]). For the class of all left-infinite intervals on R^1, Dvoretzky *et al.* [5] showed that

$$P\{U_n > \varepsilon\} \leqslant ce^{-2n\varepsilon^2} \tag{1.2}$$

for some universal constant c not exceeding 611 (Devroye and Wise [3]). When $\mathscr{A} = \{(-\infty, a_1]x \cdots x(-\infty, a_d]; (a_1, ..., a_d) \in R^d\}$, Kiefer [8, 9] showed that for each $\alpha < 2$, there exists a constant $c(d, \alpha)$ such that

$$P\{U_n > \varepsilon\} \leqslant c(d, \alpha)\, e^{-\alpha n\varepsilon^2}. \tag{1.3}$$

Devroye [2] showed that for this class

$$P\{U_n > \varepsilon\} \leqslant 2e^2(2n)^d\, e^{-2n\varepsilon^2}, \qquad n\varepsilon^2 \geqslant d^2. \tag{1.4}$$

Bound (1.3) is a moderate deviation result ($n\varepsilon^2 \to \infty$ makes it go to 0) while (1.4) is a large deviation result ($n\varepsilon^2/\log n \to \infty$ makes it go to 0) that for fixed ε decreases more rapidly to 0 than (1.3).

Wolfowitz [14] discusses the behavior of U_n if $\mathscr{A}$ is the class of all linear halfspaces. For different classes of sets $\mathscr{A}$, a general method for obtaining upper bounds was developed by Vapnik and Chervonenkis [13].

Throughout this paper, we assume that

$$\text{(i)} \quad \sup_{A \in \mathscr{A}} |\mu_n(A) - \mu(A)|,$$

$$\text{(ii)} \quad \sup_{A \in \mathscr{A}} \mu_n(A)$$

and

$$\text{(iii)} \quad \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu'_m(A)|$$

are random variables where μ'_m is the empirical measure constructed from $X'_1,..., X'_m$, a sample of independent random vectors with common probability measure μ, and independent of $X_1,..., X_n$. For the classes $\mathcal{A}$ discussed below, this is the case (e.g., the products of closed left-infinite intervals; the products of intervals; the open spheres; the closed spheres; the open linear halfspaces; the closed linear halfspaces; the finite intersections of open (closed) linear halfspaces; the open convex sets; etc.).

THEOREM (Vapnik and Chervonenkis, [13]). *If* $N_{\mathcal{A}}(x_1,..., x_n)$ *is the number of* different *sets in*

$$\{\{x_1,..., x_n\} \cap A \mid A \in \mathcal{A}\},$$

and

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

then

$$P\{U_n > \varepsilon\} \leqslant 4s(\mathcal{A}, 2n)\, e^{-n\epsilon^2/8}, \qquad n\varepsilon^2 \geqslant 1. \tag{1.5}$$

We prove the following

THEOREM. *There exists a universal constant c such that*

$$P\{U_n > \varepsilon\} \leqslant cs(\mathcal{A}, n^2)\, e^{-2n\epsilon^2}. \tag{1.6}$$

The constant does not exceed $4e^{(4\epsilon + 4\epsilon^2)}$.

The proof of (1.6) is tailored to the proof of Vapnik and Chervonenkis [13]. A slightly different inequality is due to Devroye and Wagner [1].

Note. The quantity $s(\mathcal{A}, n)$ measures how "complex" the class $\mathcal{A}$ is. For example, we have

(1) $\mathcal{A} = \{A\}: s(\mathcal{A}, n) = 1$.

(2) $\mathcal{A} = \{(-\infty, a_1]x \cdots x(-\infty, a_d] \mid -\infty \leqslant a_1 \leqslant +\infty,..., -\infty \leqslant a_d \leqslant +\infty\}$:

$$s(\mathcal{A}, n) = (1 + n)^d.$$

(3) $\mathcal{A} = \{$all rectangles in $R^d\}$, where a rectangle is a d-fold product of intervals of the type $(a, b]$, (a, b), $[a, b)$, or $[a, b]$ with $-\infty \leqslant a \leqslant b \leqslant +\infty$:

$$s(\mathcal{A}, n) \leqslant \sum_{i=0}^{2d} \binom{n}{i} \leqslant 1 + n^{2d} \leqslant 2n^{2d},$$

$$s(\mathcal{A}, n) \leqslant \sum_{i=0}^{2d} \binom{n}{i} \leqslant \frac{2}{(2d-1)!} n^{2d}, \qquad n \geqslant 2d.$$

(4) $\mathcal{A} = \{$all linear halfspaces in $R^d\}$, where a linear half-space is a set of $(x_1, \dots, x_d) \in R^d$ satisfying

$$a_1 x_1 + \cdots + a_d x_d + a_0 > 0$$

or

$$a_1 x_1 + \cdots + a_d x_d + a_0 \geqslant 0$$

for some $(a_1, \dots, a_d, a_0) \in R^{d+1}$. We have:

$$s(\mathcal{A}, n) \leqslant 2 \sum_{i=0}^{d} \binom{n}{i} - 1 \leqslant 2n^d,$$

$$\leqslant \frac{4}{(d-1)!} n^d, \qquad n \geqslant d.$$

(5) $\mathcal{A} = \{$all closed or open l_2-spheres in $R^d\}$:

$$s(\mathcal{A}, n) \leqslant 2 \sum_{i=0}^{d+1} \binom{n}{i} - 1 \leqslant 2n^{d+1}.$$

The proofs of these inequalities use straightforward combinatorial arguments; most of them are summarized by Vapnik and Chervonenkis [13] and Feinhloz [6].

Note. For small ε, the bound in (1.6) becomes very close to $4s(\mathcal{A}, n^2)\, e^{-2n\varepsilon^2}$. For $\mathcal{A} = \{A\}$, it is just twice as large as Hoeffding's bound (1.1).

2. Proof of the Theorem

Define $n' = n^2 - n$, $T = (X_1, \dots, X_n)$, $V = (X_{n+1}, \dots, X_{n+n'})$, where $X_1, \dots, X_{n^2}$ are independent identically distributed random vectors from R^d with probability measure μ. Let μ_T and μ_V be the classical empirical measures for T and V, respectively. For each Borel subset A of R^d, let

$$\rho_A = |\mu_V(A) - \mu_T(A)|,$$

and define

$$\rho = \sup_{A \in \mathcal{A}} \rho_A,$$

$$\sigma = \sup_{A \in \mathcal{A}} |\mu(A) - \mu_T(A)|.$$

Also, let P, P_T and P_V be the probability measures induced by the overall sample (T, V), T and V in $R^{n^2 d}$, R^{nd} and $R^{n'd}$. We will first show that for $0 < \alpha < 1$, $\varepsilon > 0$,

$$P\{\rho > (1-\alpha)\varepsilon\} \geqslant \left(1 - \frac{1}{4\alpha^2\varepsilon^2 n'}\right) P\{\sigma > \varepsilon\}.$$

Indeed, notice that $\sigma > \varepsilon$ implies that $|\mu(A^*) - \mu_T(A^*)| > \varepsilon$ for some $A^* \in \mathcal{A}$ (depending upon T), and that on $\{\sigma > \varepsilon\}$, $\{|\mu_V(A^*) - \mu(A^*)| \leqslant \alpha\varepsilon\} \subseteq \{\rho_{A^*} > (1-\alpha)\varepsilon\} \subseteq \{\rho > (1-\alpha)\varepsilon\}$. Thus,

$$\begin{aligned} P\{\rho > (1-\alpha)\varepsilon\} &= \int_{R^{n^2 d}} I_{[\rho > (1-\alpha)\varepsilon]}\, dP \\ &= \int_{R^{nd}} dP_T \int_{R^{n'd}} I_{[\rho > (1-\alpha)\varepsilon]}\, dP_V \\ &\geqslant \int_{[\sigma > \varepsilon]} dP_T \int_{R^{n'd}} I_{[\rho > (1-\alpha)\varepsilon]}\, dP_V \\ &\geqslant P_T\{\sigma > \varepsilon\} \cdot \inf_{A \in \mathcal{A}} P\{|\mu_V(A) - \mu(A)| \leqslant \alpha\varepsilon\} \\ &\geqslant P\{\sigma > \varepsilon\} \cdot \left(1 - \frac{1}{4\alpha^2\varepsilon^2 n'}\right). \end{aligned}$$

Let (T_i, V_i) denote one of the possible $n^2!$ permutations of (T, V), and let $\rho_A(i)$, $\rho(i)$ be defined as ρ_A, but with (T_i, V_i) replacing (T, V). Two sets A and B from R^d are equivalent for (T, V) if

$$A \cap \{X_1, \ldots, X_{n^2}\} = B \cap \{X_1, \ldots, X_{n^2}\}.$$

For such equivalent sets, we have of course $\mu_{V_i}(A) = \mu_{V_i}(B)$, $\mu_{T_i}(A) = \mu_{T_i}(B)$, all $i = 1, \ldots, n^2!$

Proceeding as in Vapnik and Chervonenkis [13], we have

$$\begin{aligned}
&\frac{1}{n^2!}\sum_{i=1}^{n^2!} I_{[\rho(i)>(1-\alpha)\epsilon]} \\
&\quad = \frac{1}{n^2!}\sum_{i=1}^{n^2!} \sup_{A\in\mathcal{A}} I_{[\rho_A(i)>(1-\alpha)\epsilon]} \\
&\quad \leqslant \frac{1}{n^2!}\sum_{i=1}^{n^2!} \sum_{A\in\mathcal{A}_{(T,V)}} I_{[\rho_A(i)>(1-\alpha)\epsilon]} \\
&\quad \leqslant \sum_{A\in\mathcal{A}_{(T,V)}} \frac{1}{n^2!}\sum_{i=1}^{n^2!} I_{[\rho_A(i)>(1-\alpha)\epsilon]} \\
&\quad \leqslant N_{\mathcal{A}}(X_1,\dots,X_{n^2})\, e^{-2n\epsilon^2+4\alpha n\epsilon^2+4\epsilon^2} \\
&\quad \leqslant s(\mathcal{A},n^2)\, e^{-2n\epsilon^2+4\alpha n\epsilon^2+4\epsilon^2}, \qquad (2.1)
\end{aligned}$$

where $\mathcal{A}_{(T,V)} \subseteq \mathcal{A}$ is a subclass from $\mathcal{A}$ with the properties

(i) $A, B \in \mathcal{A}_{(T,V)}$ implies tat A and B are not equivalent for (T, V),

(ii) for every $A \in \mathcal{A}$, there exists a $B \in \mathcal{A}_{(T,V)}$ that is equivalent to A for (T, V).

Thus, $\mathcal{A}_{(T,V)}$ cannot have more than $s(\mathcal{A}, n^2)$ sets. Let us now explain the third inequality in (2.1).

If $Y_1,\dots, Y_{n^2}$ is a permutation of $y_1,\dots, y_{n^2}$, a sequence of 0's and 1's, with $Y_i = I_{[X_i\in A]}$, then

$$\begin{aligned}
&\frac{1}{n^2!}\sum_{i=1}^{n^2!} I_{[\rho_A(i)>(1-\alpha)\epsilon]} \\
&\quad = P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n'}\sum_{i=1}^{n'} Y_{n+i}\right| > (1-\alpha)\varepsilon\right\} \\
&\quad = P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n'}\left(n^2\mu_{(T,V)}(A) - \sum_{i=1}^{n} Y_i\right)\right| > (1-\alpha)\varepsilon\right\} \\
&\quad = P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu_{(T,V)}(A)\right| > (1-\alpha)\varepsilon\,\frac{n'}{n^2}\right\} \\
&\quad \leqslant 2\exp\left\{-2n(1-\alpha)^2\varepsilon^2\left(\frac{n'}{n^2}\right)^2\right\} \\
&\quad \leqslant 2\exp\{-2n\varepsilon^2 + 4\alpha n\varepsilon^2 + 4\varepsilon^2\},
\end{aligned}$$

where $\mu_{(T,V)}$ is the classical empirical measure for (T, V), and where we used Hoeffding's inequality for sampling without replacement from n^2 binaryvalued elements with sum $n^2\mu_{(T,V)}(A)$ (Hoeffding [15]; Serfling [11]). Taking expectations on both sides of (2.1) gives

$$P\{\rho > (1-\alpha)\varepsilon\} \leqslant s(\mathcal{A}, n^2)\, e^{-2\epsilon^2 + 4\alpha n\epsilon^2 + 4\epsilon^2}.$$

Collecting bounds yields

$$\begin{aligned} P\{\sigma > \varepsilon\} &\leqslant 2s(\mathcal{A}, n^2)\frac{1}{1-(1/4\alpha^2\varepsilon^2 n')}\, e^{(4\alpha n\epsilon^2 + 4\epsilon^2)}\, e^{-2n\epsilon^2} \\ &\leqslant 2e^{(4\epsilon/\gamma + 4\epsilon^2)}\frac{1}{1-\gamma^2/2}\, s(\mathcal{A}, n^2)\, e^{-2n\epsilon^2}, \end{aligned}$$

when $\alpha = 1/\gamma n\varepsilon$, $n \geqslant 2$, $0 < \gamma < \sqrt{2}$. For $\gamma = 1$, we obtain

$$P\{\sigma > \varepsilon\} \leqslant 4e^{(4\epsilon + 4\epsilon^2)} s(\mathcal{A}, n^2)\, e^{-2n\epsilon^2}.$$

Note. We have in fact shown that

$$P\{U_n > \varepsilon\} \leqslant 4e^{(4\epsilon + 4\epsilon^2)} e^{-2n\epsilon^2} E\{N_{\mathcal{A}}(X_1, \ldots, X_{n^2})\}. \tag{2.2}$$

In many cases, this bound is considerably smaller than (1.6).

References

[1] Devroye, L. P. and Wagner, T. J. (1976). Nonparametric discrimination and density estimation, Technical Report 183, Information Systems Research Laboratory, University of Texas.

[2] Devroye, L. P. (1977). A uniform bound for the deviation of empirical distribution functions. *J. Multivar. Anal.* **7** 594–597.

[3] Devroye, L. P. and Wise, G. L. (1979). On the recovery of discrete probability densities from imperfect measurements. *J. Franklin Inst.* **307** 1–20.

[4] Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929.

[5] Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **33** 642–669.

[6] Feinholz, L. (1979). Estimation of the performance of partitioning algorithms in pattern classification. Thesis, Department of Mathematics, McGill University, Montreal.

[7] Gaenssler, P., and Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Ann. Probab.* **7** 193–243.

[8] Kiefer, J. and Wolfowitz, J. (1958). On the deviations of the empiric distribution function of vector chance variables. *Trans. Amer. Math. Soc.* **87** 173–186.

[9] Kiefer, J. (1961). On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm. *Pacific J. Math.* **11** 649–660.

[10] Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.* **33** 659–680.

[11] SERFLING, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *Ann. Statist.* **2** 39–48.

[12] STEELE, J. M. (1978). Empirical discrepancies and subadditive processes. *Ann. Probab.* **6** 118–127.

[13] VAPNIK, V. N. AND CHERVONENKIS, A. YA. (1971). On the uniform convergence of the relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

[14] WOLFOWITZ, J. (1960). Convergence of the empiric distribution function on half-spaces. *Contributions in Probability and Statistics*, I (Olkin et al., Eds.), pp. 504–507, Stanford University Press, Stanford, Calif.

[15] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.