Pergamon

Pattern Recognition, Vol. 28, No. 7, pp. 1011–1018, 1995
Elsevier Science Ltd
Pattern Recognition Society
Printed in Great Britain.
0031-3203/95 $9.50 + .00

0031-3203(94)00141-3

LOWER BOUNDS IN PATTERN RECOGNITION AND LEARNING

LUC DEVROYE*† and GÁBOR LUGOSI‡

† School of Computer Science, McGill University, 3480 University Street, Montreal, H3A 2A7 Canada
‡ Department of Mathematics, Technical University of Budapest, 1521 Stoczek u. 2, Budapest, Hungary

(*Received 23 February 1994; revised 21 September 1994; received for publication 1 November 1994*)

Abstract—Lower bounds are derived for the performance of any pattern recognition algorithm, which, using training data, selects a discrimination rule from a certain class of rules. The bounds involve the Vapnik–Chervonenkis dimension of the class, and L, the minimal error probability within the class. We provide lower bounds when $L=0$ (the usual assumption in Valiant's theory of learning) and $L>0$.

Learning Nonparametric estimation Vapnik–Chervonenkis inequality Lower bounds
Pattern recognition

1. INTRODUCTION

In statistical pattern recognition (or classification), one is usually given a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$, which consists of n independent identically distributed $\mathscr{R}^d \times \{0,1\}$ valued random variables with the same distribution as (X, Y). Denote the probability measure of X by μ. The object is to guess Y from X and the training set. Let us formally call a given estimate (or pattern recognition rule) $g_n(X) = g_n(X; X_1, Y_1, \ldots, X_n, Y_n)$. The best possible rule, or the Bayes rule, is the one achieving the smallest (or Bayes) probability of error,

$$L^* \stackrel{\text{def}}{=} \inf_{g:\mathscr{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\}.$$

The object is to find rules g_n such that in a specified sense, the probability of error with g_n,

$$L_n \stackrel{\text{def}}{=} \mathbf{P}\{g_n(X) \neq Y \mid X_1, Y_1, \ldots, X_n, Y_n\},$$

is close to L^*.

Under the impetus of Valiant,[1] many people have recast the pattern recognition in the framework of learning. Originally this was done under two restrictions:

- $L^* = 0$: this happens only if with probability one, $\mathbf{P}\{Y = 1 \mid X\} \in \{0,1\}$. In pattern recognition, we speak of non-overlapping classes.
- One is interested in minimizing L_n over a given class of rules $\mathscr{G}$. That is, with the help of the training data, the designer picks a function from a given class of $\{0,1\}$-valued functions $\mathscr{G}$. (In the terminology of learning theory, elements of $\mathscr{G}$ are called concepts.) The error with the best rule in $\mathscr{G}$ is denoted by

$$L \stackrel{\text{def}}{=} \inf_{g \in \mathscr{G}: \mathscr{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\}.$$

In fact, it is assumed that $L = L^* = 0$, that is, that the Bayes rule is in $\mathscr{G}$.

Later, these requirements have been relaxed. To see what the limits are that one can achieve, minimax lower bounds for the quantity

$$\sup_{(X,Y):L=0} \mathbf{P}\{L_n - L \geq \varepsilon\}$$

are derived that are valid for all rules g_n. Needless to say, this provides us with information about the necessary sample size. An (ε, δ) learning algorithm in the sense of Valiant[1] is one for which we may find a sample size threshold $N(\varepsilon, \delta)$ such that for $n \geq N(\varepsilon, \delta)$:

$$\sup_{(X,Y):L=0} \mathbf{P}\{L_n - L \geq \varepsilon\} \leq \delta.$$

In this respect, $N(\varepsilon, \delta)$ may be considered as a measure of the appropriateness of the algorithm. Blumer *et al.*[2] showed that for any algorithm,

$$N(\varepsilon, \delta) \geq C\left(\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right) + V\right),$$

where C, is a universal constant and V is the Vapnik–Chervonenkis (or VC) dimension of $\mathscr{G}$, introduced by Vapnik and Chervonenkis.[3–5] We recall here that V is the largest integer n such that there exists a set $\{x_1, \ldots, x_n\} \subseteq \mathscr{R}^d$ that is shattered by $\mathscr{G}$. That is, for every subset $S \subseteq \{1, \ldots, n\}$, there exists $g \in \mathscr{G}$ such that $g(x_i) = 1$ when $i \in S$ an $g(x_i) = 0$ when $i \notin S$. In Ehrenfeucht *et al.*[6] the lower bound was partially improved to

$$N(\varepsilon, \delta) \geq \frac{V-1}{32\varepsilon}$$

* Author to whom correspondence should be addressed.

when $\varepsilon \leqslant 1/8$ and $\delta \leqslant 1/100$. It may be combined with the previous bound.

In the first part of this note we improve this bound further in constants. More importantly, the main purpose of this note is to deal also with the case $L > 0$, to tie things in with the more standard pattern recognition literature. In fact, we will derive lower tail bounds as above, as well as expectation bounds for

$$\sup_{(X,Y):L\text{fixed}} \mathbf{E}\{L_n - L\}.$$

For the $L = 0$ case it is shown in theorem 2 that

$$\sup_{(X,Y):L=0} \mathbf{P}\{L_n \geqslant \varepsilon\} \geqslant \frac{1}{2e\sqrt{\pi V}}\left(\frac{2e\varepsilon n}{V-1}\right)^{(V-1)/2} e^{-4n\varepsilon/(1-4\varepsilon)}.$$

Devroye and Wagner[7] showed that if g_n is a function that minimizes the empirical error

$$\sum_{i=1}^{n} I_{[g(X_i) \neq Y_i]}$$

over $\mathcal{G}$, and $L = 0$, then

$$\mathbf{P}\{L_n \geqslant \varepsilon\} \leqslant 4\left(\frac{2en}{V}\right)^V e^{-n\varepsilon/4}.$$

($I_{[A]}$ denotes the indicator of an event A.) Later this bound was improved by Blumer *et al.*[2] to

$$\mathbf{P}\{L_n \geqslant \varepsilon\} \leqslant 2\left(\frac{2en}{V}\right)^V e^{-n\varepsilon \log 2/2}.$$

Apart from the $\varepsilon^{(V-1)/2}$ term in the lower bound, and differences in constants, the lower bound and the upper bound have the same form.

For the case $L > 0$, several upper bounds for the performance of empirical error minimization were derived using Vapnik–Chervonenkis-type inequalities (see reference 8 for a survey). The best upper bounds have the form

$$c_1(n\varepsilon^2)^{c_2 V} e^{-2n\varepsilon^2},$$

(see references 9–10), which are much larger than the bounds for $L = 0$ for small ε. Among other inequalities, we show in theorem 5, that the ε^2 term in the exponent is necessary. In particular, for fixed $L \leqslant 1/4$

$$\sup_{(X,Y):L\text{fixed}} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{1}{4} e^{-4n\varepsilon^2/L}.$$

In general, we can conclude, that in the case $L > 0$, the number of samples necessary for a certain accuracy is much larger than in the usual learning theory setup, where $L = 0$ is assumed. This phenomenon was already observed by Vapnik and Chervonenkis[11] and Simon[12] who both proved lower bounds of the type

$$\sup_{(X,Y)\text{arbitrary}} \mathbf{E}\{L_n - L\} = \Omega\left(\sqrt{\frac{V}{n}}\right).$$

In terms of n, the order of magnitude of this lower bound is the same as those of upper bounds implied by the probability inequalities cited above. In our theorem 3 we point out that the lower bound on the expected value of any rule depends on L as

$$\sup_{(X,Y):L\text{fixed}} \mathbf{E}\{L_n - L\} = \Omega\left(\sqrt{\frac{LV}{n}}\right).$$

The results presented here can also be applied for classes with infinite VC dimension. For example, it is not hard to derive from theorem 1, what Blumer *et al.*[2] already pointed out, that if $V = \infty$, then for every n and g_n, there is a distribution with $L = 0$ such that

$$\mathbf{E}L_n \geqslant c$$

for some universal constant c. This generalizes the first theorem in reference 13, where Devroye showed a similar result if $\mathcal{G}$ is the class of all measurable discrimination functions. Thus, when $V = \infty$, distribution-free nontrivial performance guarantees for $L_n - L$ or $L_n - L^*$ do not exist.

Other general lower bounds for $L_n - L^*$ were also given in reference 13. For example, it is shown there that if $L^* < 1/2$, then for any sequence of rules g_n, and positive numbers $a_n \to 0$, there exists a fixed distribution such that $\mathbf{E}L_n \geqslant \min(L^* + a_n, 1/2)$ along a subsequence, that is, the rate of convergence to the Bayes-risk can be arbitrarily slow for some distributions. The difference with the minimax bounds given here is that the same distribution is used for all n, whereas the bad distributions for the bounds in this paper vary with n.

We note here that all results remain valid if we allow randomization in the algorithms g_n.

The case L = 0

We begin by quoting a result by Vapnik and Chervonenkis[11] and Haussler *et al.*[14]

Theorem 1. Let $\mathcal{G}$ be a class of discrimination functions with VC dimension V. Let χ be the set of all random variables (X, Y) for which $L = 0$. Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$, and $n \geqslant V - 1$,

$$\sup_{(X,Y)\in\chi} \mathbf{E}L_n \geqslant \frac{V-1}{2en}\left(1 - \frac{1}{n-1}\right).$$

We now turn to the probability bound. Our bound improves over the best bound that we are aware of thus far, as given in theorem 1 and corollary 5 of Ehrenfeucht *et al.*[6] In the case of $N(\varepsilon, \delta)$, the sample size needed for (ε, δ) learning, the coefficient is improved by a factor of 8/3.

Theorem 2. Let $\mathcal{G}$ be a class of discrimination functions with VC dimension $V \geqslant 2$. Let χ be the set of all random variables (X, Y) for which $L = 0$. Assume $\varepsilon \leqslant 1/4$. Define $v = [(V-1)/2]$, and assume $n \geqslant v$. Then for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{(X,Y)\in\chi} \mathbf{P}\{L_n \geq \varepsilon\} \geq \frac{1}{2e\sqrt{2\pi v}}\left(\frac{2ne\varepsilon}{V-1}\right)^{(V-1)/2} e^{-4n\varepsilon/(1-4\varepsilon)}.$$

In particular, when $\varepsilon \leq 1/8$ and

$$\log\left(\frac{1}{\delta}\right) \geq \left(\frac{4v}{e}\right)(2e\sqrt{2\pi v})^{1/v},$$

then

$$N(\varepsilon,\delta) \geq \frac{1}{8\varepsilon}\log\left(\frac{1}{\delta}\right).$$

If on the other hand $n \geq 15$ and $n \leq (V-1)/(12\varepsilon)$, then

$$\sup_{(X,Y)\in\chi} \mathbf{P}\{L_n \geq \varepsilon\} \geq \frac{1}{20}.$$

Finally, for $\delta \leq 1/20$, and $\varepsilon < 1/2$,

$$N(\varepsilon,\delta) \geq \frac{V-1}{12\varepsilon}.$$

Proof. The idea is to construct a family $\mathscr{F}$ of 2^{V-1} distributions within the distributions with $L=0$ as follows: first find points $x_1,\dots,x_V$ that are shattered by $\mathscr{G}$. A member in $\mathscr{F}$ is described by $V-1$ bits, $\theta_1,\dots,\theta_{V-1}$. For convenience, this is represented as a bit vector θ. We write θ_{i-} and θ_{i-} for the vector θ in which the ith bit is set to 1 and 0, respectively. Assume $V-1 \leq n$. For a particular bit vector, we let $X = x_i (i < V)$ with probability p each, while $X = x_V$ with probability $1-p(V-1)$. Then set $Y = f_\theta(X)$, where f_θ is defined as follows:

$$f_\theta(x) = \begin{cases}\theta_i & \text{if } x = x_i, i < V;\\ 0 & \text{if } x = x_V.\end{cases}$$

Note that since Y is a function of X, we must have $L^* = 0$. Also, $L=0$, as the set $\{x_1,\dots,x_V\}$ is shattered by $\mathscr{G}$, i.e. there is a $g\in\mathscr{G}$ with $g(x_i) = f_\theta(x_i)$ for $1 \leq i \leq V$.

Observe that

$$L_n \geq p\sum_{i=1}^{V-1} I_{[g_n(x_i,X_1,Y_1,\dots,X_n,Y_n)\neq\theta_i]}.$$

Using this, for given θ,

$$\begin{aligned}&\mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\\ &\geq \mathbf{P}\left\{p\sum_{i=1}^{V-1} I_{[g_n(x_i,X_1,Y_1,\dots,X_n,Y_n)\neq\theta_i]}\right.\\ &\left.\geq \varepsilon \,\middle|\, X_1,Y_1,\dots,X_n,Y_n\right\}.\end{aligned}$$

This probability is either zero or one, as the event is deterministic. We now randomize and replace θ and Θ. For fixed $X_1,\dots,X_n$, we denote by J the collection $\{j: 1 \leq j \leq V-1, \cap_{i=1}^n [X_i \neq x_j]\}$. This is the collection of empty cells x_i. We bound our probability from below by summing over J only:

$$\begin{aligned}&\mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\\ &\geq \mathbf{P}\left\{p\sum_{i\in J} I_{[g_{ni}\neq\Theta_i]} \geq \varepsilon \,\middle|\, X_1,Y_1,\dots,X_n,Y_n\right\}\end{aligned}$$

where g_{ni} is shorthand for $g_n(x_i, X_1,\dots,Y_n)$. Conditionally, these are fixed members from $\{0,1\}$. The Θ_is with $i\in J$ constitute independent Bernoulli (1/2) random variables. Importantly, their values do not alter the g_{ni}s (this cannot be said for Θ_i when $i\notin J$). Thus, our lower bound is equal to

$$\mathbf{P}\{p\,\text{Binomial}(|J|,1/2) \geq \varepsilon \,|\, |J|\}.$$

We summarize:

$$\begin{aligned}&\sup_{(X,Y):L=0} \mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\\ &\geq \sup_{(X,Y)\in\mathscr{F}} \mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\\ &\geq \sup_\theta \mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\\ &\geq \mathbf{E}\{\mathbf{P}\{L_n \geq \varepsilon | X_1, Y_1,\dots,X_n,Y_n\}\}\\ &\geq \mathbf{E}\{\mathbf{P}\{p\,\text{Binomial}(|J|,1/2) \geq \varepsilon \,|\, |J|\}\}.\end{aligned}$$

As we are dealing with a symmetric binomial, it is easy to see that the last expression in the chain is at least equal to

$$\frac{1}{2}\mathbf{P}\{|J| \geq 2\varepsilon/p\}.$$

Assume that $\varepsilon < 1/2$. By the pigeonhole principle, $|J| \geq 2\varepsilon/p$ if the number of points X_i, $1 \leq i \leq n$, that are not equal to x_V does not exceed $V-1-2\varepsilon/p$. Therefore, we have a further lower bound:

$$\frac{1}{2}\mathbf{P}\{|J| \geq 2\varepsilon/p\} \geq \frac{1}{2}\mathbf{P}\{\text{Binomial}(n,(V-1)p) \leq V-1-2\varepsilon/p\}.$$

We consider two choice for p.

Choice A. Take $p = 1/n$, and assume $12n\varepsilon \leq V-1$, $\varepsilon < 1/2$. Note that for $n \geq 15$

$$\mathbf{E}|J| = (V-1)(1-p)^n \geq \frac{V-1}{e}\left(1-\frac{1}{n}\right) \geq \frac{V-1}{3}.$$

Also since $0 \leq |J| \leq V-1$, we have $\text{Var}|J| \leq (V-1)^2/4$. By the Chebyshev–Cantelli inequality,

$$\begin{aligned}&(1/2)\mathbf{P}\{|J| \geq 2n\varepsilon\}\\ &= (1/2)(1-\mathbf{P}\{|J| < 2n\varepsilon\})\\ &\geq (1/2)(1-\mathbf{P}\{|J| < (V-1)/6\})\\ &= (1/2)(1-\mathbf{P}\{|J|-\mathbf{E}|J| \leq (V-1)/6-\mathbf{E}|J|\})\\ &\geq (1/2)(1-\mathbf{P}\{|J|-\mathbf{E}|J| \leq -(V-1)/6\})\\ &\geq (1/2)\left(1-\frac{\text{Var}|J|}{\text{Var}|J|+(V-1)^2/36}\right)\\ &\geq (1/2)\left(1-\frac{(V-1)^2/4}{(V-1)^2/4+(V-1)^2/36}\right)\\ &= \frac{1}{20}.\end{aligned}$$

This proves the second inequality for $\sup \mathbf{P}\{L_n \geq \varepsilon\}$.

Choice B. Take $p = 2\varepsilon/v$ and assume $\varepsilon \leqslant 1/4$. Assume $n \geqslant v$. Then the lower bound is

$$\frac{1}{2}\mathbf{P}\{\text{Binomial}(n, 4\varepsilon) \leqslant v\}$$

$$\geqslant \frac{1}{2}\binom{n}{v}(4\varepsilon)^v(1-4\varepsilon)^{n-v}$$

$$\geqslant \frac{1}{2}\frac{1}{\mathrm{e}\sqrt{2\pi v}}\left(\frac{4\mathrm{e}\varepsilon(n-v+1)}{v(1-4\varepsilon)}\right)^v(1-4\varepsilon)^n$$

$$\left(\text{since } \binom{n}{v} \geqslant \left(\frac{(n-v+1)\mathrm{e}}{v}\right)^v \times \frac{1}{\mathrm{e}\sqrt{2\pi v}} \text{ by Stirling's formula}\right)$$

$$\geqslant \frac{1}{2}\frac{1}{\mathrm{e}\sqrt{2\pi v}}\left(\frac{4\mathrm{e}\varepsilon(n-v+1)}{v}\right)^v(1-4\varepsilon)^n$$

$$\geqslant \frac{1}{2}\frac{1}{\mathrm{e}\sqrt{2\pi v}}\left(\frac{4\mathrm{e}n\varepsilon}{v}\right)^v(1-4\varepsilon)^n\left(1-\frac{v-1}{n}\right)^v$$

$$\geqslant \frac{1}{2}\frac{1}{\mathrm{e}\sqrt{2\pi v}}\left(\frac{2\mathrm{e}n\varepsilon}{v}\right)^v \mathrm{e}^{-4n\varepsilon/(1-4\varepsilon)} \quad (\text{since } n \geqslant 2(v-1))$$

(use $1 - x \geqslant \exp(-x/(1-x))$)

$$\geqslant \frac{1}{2}\frac{1}{\mathrm{e}\sqrt{2\pi v}}\left(\frac{2\mathrm{e}n\varepsilon}{v}\right)^v \mathrm{e}^{-8n\varepsilon} \quad (\text{since } \varepsilon \leqslant 1/8)$$

$$\geqslant \frac{(8\varepsilon)^v}{\log^v(1/\delta)} n^v \mathrm{e}^{-8n\varepsilon}$$

$$\left(\text{since we assume } \log\left(\frac{1}{\delta}\right) \geqslant \left(\frac{4v}{\mathrm{e}}\right)(2\mathrm{e}\sqrt{2\pi v})^{1/v}\right).$$

The function $n^v\mathrm{e}^{-8n\varepsilon}$ varies unimodally in n, and achieves a peak at $n = v/(8\varepsilon)$. For n below this threshold, by monotonicity, we apply the bound at $n = v/(8\varepsilon)$. It is easy to verify that the value of the bound at $v/(8\varepsilon)$ is always at least δ. If on the other hand, $(1/8\varepsilon)\log(1/\delta) \geqslant n \geqslant v/(8\varepsilon)$, the lower bound achieves its minimal value at $(1/8\varepsilon)\log(1/\delta)$, and the value there is δ. This concludes the proof.

The case $\mathrm{L} > 0$

In this section, we consider both expectation and probability bounds when $L > 0$. The bounds involve n, V and L jointly. The minimax lower bound below is valid for any discrimination rule, and depends upon n as $\sqrt{L(V-1)/n}$. As a function of n, this decrease as in the central limit theorem. Interestingly, the lower bound becomes smaller as L decreases, as should be expected. The largest sample sizes are needed when L is close to $1/2$. (Note that for $L = 1/2$, $N(\varepsilon, \delta) = 0$, since any random decision will give $1/2$ error probability.) When L is very small, we provide an $\Omega(1/n)$ lower bound, just as for the case $L = 0$. The constants in the bounds may be tightened at the expense of more complicated expressions.

Theorem 3. Let $\mathscr{G}$ be a class of discrimination functions with VC dimension $V \geqslant 2$. Assume that $n \geqslant 8(V-1)$. Let χ be the set of all random variables (X, Y) for which for fixed $L \in (0, 1/2)$,

$$L = \inf_{g \in \mathscr{G}} \mathbf{P}\{g(X) \neq Y\}.$$

Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$,

$$\sup_{(X,Y)\in\chi} \mathbf{E}(L_n - L) \geqslant \begin{cases} \sqrt{\dfrac{L(V-1)}{8n}}(1-2/n)^{2n} & \text{if } n \geqslant \dfrac{V-1}{2L}\max(4, 1/(1-2L)^2); \\ \dfrac{2(V-1)}{n}(1-8/n)^{2n} & \text{if } n \leqslant 2(V-1)/L \text{(this implies } L \leqslant 1/4). \end{cases}$$

Proof. Again we consider the finite family $\mathscr{F}$ from the previous section. The notation θ and Θ is also as above. X now puts mass p at x_i, $i < V$, and mass $1 - (V-1)p$ at x_V. This imposes the condition $(V-1)p \leqslant 1$, which will be satisfied. Next introduce the constant $c \in (0, 1/2)$. We no longer have Y as a function of X. Instead, we have a uniform $[0, 1]$ random variable U independent of X and define

$$Y = \begin{cases} 1 & \text{if } U \leqslant \frac{1}{2} - c + 2c\theta_i, X = x_i, i < V \\ 0 & \text{otherwise.} \end{cases}$$

Thus, when $X = x_i$, $i < V$, Y is 1 with probability $1/2 - c$ or $1/2 + c$. A simple argument shows that the best rule for θ is the one which sets

$$f_\theta(x) = \begin{cases} 1 & \text{if } x = x_i, i < V, \theta_i = 1; \\ 0 & \text{otherwise.} \end{cases}$$

Also, observe that

$$L = (V-1)p(1/2 - c). \tag{1}$$

We may then write, for fixed θ,

$$L_n - L \geqslant \sum_{i=1}^{V-1} 2pcI_{[g_n(x_i, X_1, Y_1, \ldots, X_n, Y_n) = 1 - f_\theta(x_i)]}.$$

It is sometimes convenient to make the dependence of g_n upon θ explicit by considering $g_n(x_i)$ as a function of $x_i, X_1, \ldots, X_n, U_1, \ldots, U_n$ (an i.i.d. sequence of uniform $[0, 1]$ random variables), and θ_i. The proof below is based upon Hellinger distances, and its methodology is essentially due to Assouad.[15] We replace θ by a uniformly distributed random Θ over $\{0, 1\}^{V-1}$. Thus,

$$\sup_{(X,Y)\in\mathcal{F}} \mathbf{E}\{L_n - L\} = \sup_{\theta} \mathbf{E}\{L_n - L\}$$
$$\geqslant \mathbf{E}\{L_n - L\} \quad \text{(with random } \Theta\text{)}$$
$$\geqslant \sum_{i=1}^{V-1} 2pc\mathbf{E}I_{[g_n(x_i, X_1, \ldots, Y_n) = 1 - f_\Theta(x_i)]}.$$

Fix $i < V$ and call the ith summand in the last expression E_i. Introduce the notation

$$p_\theta(x'_1, \ldots, x'_n, y_1, \ldots, y_n)$$
$$= \mathbf{P}\{\cap_{j=1}^n [X_j = x'_j, Y_j = y_j] \,|\, \Theta = \theta\}.$$

Clearly, this may be written as a Cartesian product:

$$p_\theta(x'_1, \ldots, x'_n, y_1, \ldots, y_n) = \prod_{j=1}^n p_\theta(x'_j, y_j),$$

where $p_\theta(x'_j, y_j) = \mathbf{P}\{X = x'_j, Y = y_j | \Theta = \theta\}$. Thus,

$$E_i = 2pc2^{-(V-1)} \sum_{\substack{(x'_1, \ldots, x'_n, y_1, \ldots, y_n) \\ \in \{x_1, \ldots, x_V\} \times \{0,1\})^n}} \sum_\theta I_{[g_n(x_i; x'_1, y_1, \ldots, x'_n, y_n) = 1 - f_\theta(x_i)]} \prod_{j=1}^n p_\theta(x'_j, y_j)$$

$$= 2pc2^{-(V-1)} \sum_{(x'_1, \ldots, x'_n, y_1, \ldots, y_n)} \sum_\theta \frac{1}{2}\left\{ I_{[g_n(x_i; x'_1, y_1, \ldots, x'_n, y_n) = 1]} \prod_{j=1}^n p_{\theta_{i-}}(x'_j, y_j) + I_{[g_n(x_i; x'_1, y_1, \ldots, x'_n, y_n) = 0]} \prod_{j=1}^n p_{\theta_{i+}}(x'_j, y_j) \right\}$$

$$\geqslant 2pc2^{-(V-1)} \sum_{(x'_1, \ldots, x'_n, y_1, \ldots, y_n)} \sum_\theta \min\left(\prod_{j=1}^n p_{\theta_{i+}}(x'_j, y_j), \prod_{j=1}^n p_{\theta_{i-}}(x'_j, y_j) \right)$$

$$\times \frac{1}{2}\{ I_{[g_n(x_i; x'_1, y_1, \ldots, x'_n, y_n) = 1]} + I_{[g_n(x_i; x'_1, y_1, \ldots, x'_n, y_n) = 0]} \}$$

$$= cp2^{-(V-1)} \sum_{(x'_1, \ldots, x'_n, y_1, \ldots, y_n)} \sum_\theta \min\left(\prod_{j=1}^n p_{\theta_{i+}}(x'_j, y_j), \prod_{j=1}^n p_{\theta_{i-}}(x'_j, y_j) \right)$$

$$\geqslant \frac{cp}{2} 2^{-(V-1)} \sum_\theta \left(\sum_{(x'_1, \ldots, x'_n, y_1, \ldots, y_n)} \sqrt{\prod_{j=1}^n p_{\theta_{i+}}(x'_j, y_j) \times \prod_{j=1}^n p_{\theta_{i-}}(x'_j, y_j)} \right)^2$$

$$= \frac{cp}{2} 2^{-(V-1)} \sum_\theta \left(\sum_{(x,y)} \sqrt{p_{\theta_{i+}}(x, y) p_{\theta_{i-}}(x, y)} \right)^{2n},$$

(by the identity $(\Sigma_i a_i)^n = \Sigma_{i_1, \ldots, i_n} a_{i_1} \cdots a_{i_n}$) where we used a discrete version of LeCam's inequality (reference 19; for example see page 7 of reference 18), which states that for positive sequences a_i and b_i, both summing to one,

$$\sum_i \min(a_i, b_i) \geqslant \frac{1}{2}\left(\sum_n \sqrt{a_i b_i} \right)^2.$$

We note next that for $x = x_j$ $1 \leqslant j \leqslant V, j \neq i$,

$$p_{\theta_{i+}}(x, y) = p_{\theta_{i-}}(x, y) = p_\theta(x, y).$$

For $x = x_i$, $i < V$, we have

$$p_{\theta_{i+}}(x, y) p_{\theta_{i-}}(x, y) = p^2\left(\frac{1}{4} - c^2\right).$$

Resubstitution in the previous chain of inequalities yields the following:

$$E_i \geqslant \frac{cp}{2} 2^{-(V-1)} \times \sum_\theta \left(\sum_{(x,y)} \sqrt{p_{\theta_{i+}}(x, y) p_{\theta_{i-}}(x, y)} \right)^{2n}$$

$$= \frac{cp}{2} 2^{-(V-1)} \sum_\theta \left(\sum_{\substack{(x,y) \\ x \neq x_i}} p_\theta(x, y) + 2p\sqrt{1/4 - c^2} \right)^{2n}$$

$$= \frac{cp}{2} 2^{-(V-1)} \sum_\theta \left(\sum_{(x,y)} p_\theta(x, y) + 2p\sqrt{1/4 - c^2} - p_\theta(x_i, 1) - p_\theta(x_i, 0) \right)^{2n}$$

$$= \frac{cp}{2} 2^{-(V-1)} \sum_\theta (1 + p\sqrt{1 - 4c^2} - p)^{2n}$$

$$= \frac{cp}{2} (1 + p\sqrt{1 - 4c^2} - p)^{2n}.$$

As the right-hand-side does not depend upon i, the overall bound becomes

$$\sup_{(X,Y)\in\mathcal{F}} \mathbf{E}(L_n - L) \geqslant \sum_{i=1}^{V-1} E_i$$
$$\geqslant \frac{(V-1)cp}{2} (1 + p\sqrt{1 - 4c^2} - p)^{2n}.$$

A rough asymptotic analysis shows that the best asymptotic choice for c is given by

$$c = \frac{1}{\sqrt{4np}}.$$

This leaves us with a quadratic equation in c. Instead of solving this equation, it is more convenient to take $c = \sqrt{(V-1)/(8nL)}$. If $2nL/(V-1) \geqslant 4$, then $c \leqslant 1/4$.

With this choice for c, the lower bound is

$$\sup_{(X,Y)\in\mathscr{F}} \mathbf{E}(L_n - L) \geqslant \frac{Lc}{1-2c}(1-4pc^2)^{2n}$$

(since $L = (V-1)p(1/2 - c)$ and $\sqrt{1-x} - 1 \geqslant -x$ for $0 \leqslant x \leqslant 1$)

$$\geqslant Lc(1 - 1/(n(1-2c)))^{2n}$$

(by our choice of c and the expression for L above)

$$\geqslant \sqrt{\frac{(V-1)L}{8n}}(1-2/n)^{2n}$$

(since $c \leqslant 1/4$).

The condition $p(V-1) \leqslant 1$ implies that we need to ask that $n \geqslant (V-1)/(2L(1-2L)^2)$.

Assume next that $2nL/(V-1) \leqslant 4$. Then we may put $p = 8/n$. Assume that $n \geqslant 8(V-1)$. This leads to a value of c determined by $1 - 2c = nL/4(V-1)$. In that case, as $c \geqslant 1/4$, the overall lower bound may be written as

$$\frac{(V-1)cp(1-p)^{2n}}{2} \geqslant (1-8/n)^{2n}\frac{2(V-1)}{n}.$$

This concludes the proof of theorem 3.

From the expectation bound in theorem 3, we may derive a probabilistic bound by a rather trivial argument. Unfortunately, the bound thus obtained only yields a suboptimal estimate for $N(\varepsilon, \delta)$.

Theorem 4. Let $\mathscr{G}$ be a class of discrimination functions with VC dimension $V \geqslant 2$. Assume that $n \geqslant 8(V-1)$. Let χ be the set of all random variables (X, Y) for which for fixed $L \in (0, 1/2)$,

$$L = \inf_{g\in\mathscr{G}} \mathbf{P}\{g(X) \neq Y\}.$$

Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$, and any $\varepsilon \leqslant A/2$,

$$\sup_{(X,Y)\in\chi} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{A}{2-A},$$

where

$$A = \begin{cases} \sqrt{\dfrac{L(V-1)}{8n}}(1-2/n)^{2n} & \text{if } n \geqslant \dfrac{V-1}{2L}\max(4, 1/(1-2L)^2); \\ \dfrac{2(V-1)}{n}(1-8/n)^{2n} & \text{if } n \leqslant 2(V-1)/L. \end{cases}$$

Also,

$$N(\varepsilon, \delta) \geqslant \frac{L(V-1)\mathrm{e}^{-10}}{32} \times \min\left(\frac{1}{\delta^2}, \frac{1}{\varepsilon^2}\right).$$

Proof. Assume that we have $\mathbf{E}(L_n - L) \geqslant A$ (as in theorem 3). Then a simple bounding argument yields, for $\varepsilon \leqslant A$,

$$\mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{A-\varepsilon}{1-\varepsilon}.$$

For $\varepsilon \leqslant A/2$, the lower bounds is at least $A/(2-A)$.

For the bound on $N(\varepsilon, \delta)$ assume that $\mathbf{P}\{L_n - L > \varepsilon\} < \delta$. Then clearly, $\mathbf{E}\{L_n - L\} \leqslant \varepsilon + \delta$. Thus, when n is large enough to satisfy the assumptions of theorem 3, we have

$$\sqrt{\frac{L(V-1)}{8n}}\left(1 - \frac{2}{n}\right)^{2n} \leqslant \varepsilon + \delta.$$

Note that

$$(1-2/n)^{2n} \geqslant \exp\left(-\frac{4}{1-2/n}\right) \geqslant e^{-5}$$

when $n \geqslant 10$. We have

$$n \geqslant \frac{L(V-1)}{8\mathrm{e}^{10}(\varepsilon+\delta)^2} \geqslant \frac{L(V-1)}{32\mathrm{e}^{10}} \times \min\left(\frac{1}{\delta^2}, \frac{1}{\varepsilon^2}\right).$$

It is easy to see that $\sup_{(X,Y)} \inf_{g_n} \mathbf{P}\{L_n - L > \varepsilon\}$ is monotone decreasing in n, therefore, small values of n cannot lead to better bounds for $N(\varepsilon, \delta)$.

Theorem 5. Let $\mathscr{G}$ be a class of discrimination functions with VC dimension $V \geqslant 2$. Let χ be the set of all random variables (X, Y) for which for fixed $L \in (0, 1/4)$,

$$L = \inf_{g\in\mathscr{G}} \mathbf{P}\{g(X) \neq Y\}.$$

Then, for every discrimination rule g_n based upon $X_1, Y_1, \ldots, X_n, Y_n$, and any $\varepsilon \leqslant L$,

$$\sup_{(X,Y)\in\chi} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{1}{4}\mathrm{e}^{-4n\varepsilon^2/L},$$

and in particular, for $\varepsilon \leqslant L \leqslant 1/4$,

$$N(\varepsilon, \delta) \geqslant \frac{L}{4\varepsilon^2}\log\frac{1}{4\delta}.$$

Proof. The argument here is similar to that in the proof of theorem 3. Using the same notation as there, it is clear that

$$\sup_{(X,Y)\in\chi} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \mathbf{E}I_{[\sum_{i=1}^{V-1} 2pcI_{[g_n(x_i, X_1, \ldots, Y_n) = 1 - f_{\Theta}(x_i)]} \geqslant \varepsilon]}$$

$$= 2^{-(V-1)} \sum_{\substack{(x'_1, \ldots, x'_n, y_1, \ldots, y_n) \\ \in(\{x_1, \ldots, x_V\} \times \{0,1\})^n}} \sum_{\theta} I_{[\sum_{i=1}^{V-1} 2pcI_{[g_n(x_i, x'_1, y_1, \ldots, x'_n, y_n) = 1 - f_\theta(x_i)]} \geqslant \varepsilon]} \prod_{j=1}^{n} p_\theta(x'_j, y_j).$$

Now, observe, that if $\varepsilon/(2pc) \leqslant (V-1)/2$ (which will be called condition (*) below), then

$$I_{[\sum_{i=1}^{V-1} 2pcI_{\{g_n(x_i, x_1, y_1, \ldots, x_n, y_n) = 1 - f^{\theta}(x_i)\}} \geqslant \varepsilon]} + I_{[\sum_{i=1}^{V-1} 2pcI_{\{g_n(x_i, x_1, y_1, \ldots, x_n, y_n) = 1 - f_{\theta^c}(x_i)\}} \geqslant \varepsilon]} \geqslant 1,$$

where θ^c denotes the binary vector $(1-\theta_1, \ldots, 1-\theta_{V-1})$, that is, the complement of θ. Therefore, for $\varepsilon \leqslant pc(V-1)$, the last expression in the lower bound above is bounded from below by

$$\frac{1}{2^{V-1}} \sum_{\substack{(x'_1, \ldots, x'_n, y_1, \ldots, y_n) \\ \in (\{x_1, \ldots, x_V\} \times \{0,1\})^n}} \sum_{\theta} \frac{1}{2} \min\left(\prod_{j=1}^{n} p_\theta(x'_j, y_j), \prod_{j=1}^{n} p_{\theta^c}(x'_j, y_j) \right)$$

$$\geqslant \frac{1}{2^{V+1}} \sum_{\theta} \left(\sum_{\substack{(x'_1, \ldots, x'_n, y_1, \ldots, y_n) \\ \in (\{x_1, \ldots, x_V\} \times \{0,1\})^n}} \sqrt{\prod_{j=1}^{n} p_\theta(x'_j, y_j) \times \prod_{j=1}^{n} p_{\theta^c}(x'_j, y_j)} \right)^2 = \frac{1}{2^{V+1}} \sum_{\theta} \left(\sum_{(x,y)} \sqrt{p_\theta(x,y) p_{\theta^c}(x,y)} \right)^{2n},$$

It is easy to see that for $x = x_V$

$$p_\theta(x,y) = p_{\theta^c}(x,y) = \frac{1-(V-1)p}{2},$$

and for $x = x_i$, $i < V$,

$$p_\theta(x,y) p_{\theta^c}(x,y) = p^2 \left(\frac{1}{4} - c^2 \right).$$

Thus, we have the equality

$$\sum_{(x,y)} \sqrt{p_\theta(x,y) p_{\theta^c}(x,y)} = 1 - (V-1)p + 2(V-1)p\sqrt{\frac{1}{4} - c^2}.$$

Summarizing, since $L = p(V-1)(1/2 - c)$, we have

$$\sup_{(X,Y) \in \chi} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{1}{4} \left(1 - \frac{L}{\frac{1}{2} - c} \left(1 - \sqrt{1 - 4c^2}\right) \right)^{2n}$$

$$\geqslant \frac{1}{4} \left(1 - \frac{L}{\frac{1}{2} - c} 4c^2 \right)^{2n}$$

$$\geqslant \frac{1}{4} \exp\left(- \frac{16nLc^2}{1-2c} \bigg/ \left(1 - \frac{8Lc^2}{1-2c} \right) \right),$$

where again, we used the inequality $1 - x \geqslant e^{-x/(1-x)}$. We may choose c as $\varepsilon/(2L + 2\varepsilon)$. It is easy to verify that condition (*) holds. Also, $p(V-1) \leqslant 1$. From the condition $L \geqslant \varepsilon$ we deduce that $c \leqslant 1/4$. The exponent is the expression above may be bounded as

$$\frac{\frac{16nLc^2}{1-2c}}{1 - \frac{8Lc^2}{1-2c}} = \frac{16nLc^2}{1 - 2c - 8Lc^2}$$

$$= \frac{\frac{4n\varepsilon^2}{L+\varepsilon}}{1 - \frac{2\varepsilon^2}{L+\varepsilon}}$$

(by substituting $c = \varepsilon/(2L + 2\varepsilon)$)

$$\leqslant \frac{4n\varepsilon^2}{L}.$$

Thus,

$$\sup_{(X,Y) \in \chi} \mathbf{P}\{L_n - L \geqslant \varepsilon\} \geqslant \frac{1}{4} \exp(-4n\varepsilon^2/L),$$

as desired. Setting this bound equal to δ provides the bound on $N(\varepsilon, \delta)$.

Theorems 4 and 5 may be course be combined. They show that $N(\varepsilon, \delta)$ is bounded from below by terms like $(1/\varepsilon^2)\log(1/\delta)$ (independent of V) and $(V-1)/\varepsilon^2$. As δ is typically small, the main term is thus not influenced by the VC dimension. This is the same phenomenon as in the case $L = 0$.

Acknowledgements—We thank Dr Simon for sending us his 1993 paper on lower bounds. The first author's research was sponsered by NSERC grant A3456 and FCAR grant 90-ER-0291.

REFERENCES

1. L. G. Valiant, A theory of the learnable, *Commun. ACM* **27**, 1134–1142 (1984).
2. A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, *J. ACM* **36**, 0–0 (1989).
3. V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Prob. Appl.* **16**, 264–280 (1971).
4. V. N. Vapnik and A. Ya. Chervonenkis, Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data, *Automat. Remote Cont.* **32**, 207–217 (1971).
5. V. N. Vapnik and A. Ya. Chervonenkis, Necessary and sufficient conditions for the uniform convergence of means to their expectations, *Theory Probab. Appl.* **26**, 532–553 (1981).
6. A. Ehrenfeucht, D. Haussler, M. Kearns and L. Valiant, A general lower bound on the number of examples needed for learning, *Inform. Comput.* **82**, 247–261 (1989).
7. L. Devroye and T. J. Wagner, Nonparametric discrimination and density estimation, Technical report 183, Electronics Research Center, University of Texas (1976).
8. L. Devroye, Automatic pattern recognition: a study of the probability of error, *IEEE Trans. Pattern Anal. Mach. Intell.* **10**, 530–543 (1988).
9. K. Alexander, Probability inequalities for empirical processes and a law of the iterated logarithm, *Ann. Probab.* **12**, 1041–1067 (1984).
10. M. Talagrand, Sharper bounds for Gaussian and empirical processes, *Ann. Probab.* **22**, 28–76 (1994).
11. V. N. Vapnik and A. Ya. Chervonenkis, *Theory of Pattern Recognition*, Nauka, Moscow (1974).
12. H. U. Simon, General lower bounds on the number of examples needed for learning probabilistic concepts, *Proc.*

6th Annual Workshop on Computational Learning Theory, pp. 402–412 (1993).
13. L. Devroye, Any discrimination rule can have an arbitrarily bad probability of error for finite sample size, *IEEE Trans. Pattern Anal. Mach. Intell.* **4**, 154–157 (1982).
14. D. Haussler, N. Littlestone and M. Warmuth, Predicting {0, 1} functions from randomly drawn points *Proc. 29th IEEE Symp. on the Foundations of Computer Science*, 100–109 (1988).
15. P. Assouad, Deux remarques sur l'estimation, *C.R.Acad. Sci. Paris*, **296**, 1021–1024 (1983).
16. L. Birgé, Approximation dans les espaces métriques et théorie de l'estimation, *Z. Wahrscheinlich. Ver. Gebiete* **65**, 181–237 (1983).
17. L. Birgé, On estimating a density using Hellinger distance and some other strange facts, *Probab. Theory Related Fields*, **71**, 271–291 (1986).
18. L. Devroye, *A Course in Density Estimation*. Birkhäuser, Boston (1987).
19. L. LeCam, Convergence of estimates under dimensionality restrictions, *Ann. Stat.* **1**, 38–53 (1973).

About the Author— LUC DEVROYE was born in Belgium on 6 August, 1948. He obtained his doctoral degree from the University of Texas at Austin in 1976 and joined McGill University in 1977, where he is Professor of Computer Science. He is interested in probability theory, probabilistic analysis of data structures and algorithms, random search, pattern recognition, nonparametric estimation, random variate generation, and font design.

About the Author—GÁBOR LUGOSI was born on 13 July, 1964 in Budapest, Hungary. He is an associate professor at the Department of Mathematics and Computer Science, Technical University of Budapest. He received his degree in electrical engineering from the Technical University of Budapest in 1987, and his Ph.D. from the Hungarian Academy of Sciences in 1991. His main research interests include pattern recognition, information theory, and nonparametric statistics.