Concept Learning Using Complexity Regularization

Gábor Lugosi and Kenneth Zeger

***Abstract*—We apply the method of complexity regularization to learn concepts from large concept classes. The method is shown to automatically find a good balance between the approximation error and the estimation error. In particular, the error probability of the obtained classifier is shown to decrease as $O(\sqrt{\log n/n})$ to the achievable optimum, for large nonparametric classes of distributions, as the sample size n grows. We also show that if the Bayes error probability is zero and the Bayes rule is in a known family of decision rules, the error probability is $O(\log n/n)$ for many large families, possibly with infinite VC dimension.**

***Index Terms*— Learning theory, estimation, pattern recognition, classification.**

I. Introduction

In pattern recognition—or, as it has recently also been called, concept learning—the value of a $\{0, 1\}$-valued random variable Y is to be predicted based upon observing an $\mathcal{R}^d$-valued random variable X. A *prediction rule* (or *decision*) is a function $\phi: \mathcal{R}^d \to \{0, 1\}$, whose performance is measured by its error probability

$$\boldsymbol{P}\{\phi(X) \neq Y\}.$$

An optimal decision

$$g^*(x) = \begin{cases} 0, & \text{if } \boldsymbol{P}\{Y = 0 \mid X = x\} \geq \boldsymbol{P}\{Y = 1 \mid X = x\} \\ 1, & \text{otherwise} \end{cases}$$

requires the knowledge of the joint distribution of (X, Y). The error probability $L^* = \boldsymbol{P}\{g^*(X) \neq Y\}$ of g^* is called the Bayes risk. Assume that the distribution of (X, Y) is unknown, but a training sequence

$$D_n = ((X_1, Y_1), \cdots, (X_n, Y_n))$$

of independent, identically distributed random variables is available, where the (X_i, Y_i) have the same distribution as (X, Y), and D_n is independent of (X, Y). A *classifier* is a function $\phi_n: \mathcal{R}^d \times (\mathcal{R}^d \times \{0, 1\})^n \to \{0, 1\}$, whose error probability is the random variable

$$L(\phi_n) = \boldsymbol{P}\{\phi_n(X, D_n) \neq Y \mid D_n\}.$$

Manuscript received May 10, 1994; revised July 24, 1995. This work was supported in part by the National Science Foundation under Grants NCR-92-96231 and INT-93-15271. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Whistler, B.C., Canada, September 1995.

G. Lugosi is with the Department of Mathematics and Computer Science, Faculty of Electrical Engineering, Technical University of Budapest, Budapest, Hungary.

K. Zeger is with the Coordinated Science Laboratory, Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, IL 61801 USA.

Publisher Item Identifier S 0018-9448(96)00011-9.

In the theory of concept learning (see Valiant [20] and Blumer *et al.* [6]) it is often assumed that $L^* = 0$, and the Bayes decision g^* is known to be a member of a relatively small class $\mathcal{C}$ of decision functions, called concepts. The "smallness" of a class $\mathcal{C}$ may meaningfully be measured by its shatter coefficients and VC dimension [6], defined as follows.

Let $\mathcal{C}$ be a class of decisions $\phi: \mathcal{R}^d \to \{0, 1\}$, and denote by $\mathcal{A}$ the collection of subsets of $\mathcal{R}^d$ of the form $A = \{x: \phi(x) = 1\}$, where $\phi \in \mathcal{C}$. For $z_1, \cdots, z_n \in \mathcal{R}^d$, let $N_{\mathcal{A}}(z_1, \cdots, z_n)$ be the number of different sets in

$$\{\{z_1, \cdots, z_n\} \cap A;\ A \in \mathcal{A}\}$$

and define the nth *shatter coefficient* of $\mathcal{C}$ as

$$S(\mathcal{C}, n) = \max_{z_1, \cdots, z_n \in \mathcal{R}^d} N_{\mathcal{A}}(z_1, \cdots, z_n).$$

The largest integer $k \geq 1$ for which $S(\mathcal{C}, k) = 2^k$ is denoted by V, and it is called the *Vapnik–Chervonenkis dimension* (or VC dimension) of the class $\mathcal{C}$. If $S(\mathcal{C}, n) = 2^n$ for all n, then by definition, $V = \infty$.

The method of empirical risk minimization picks a classifier from $\mathcal{C}$ that minimizes the empirical error probability over $\mathcal{C}$. More precisely, define the empirical error probability of a decision ϕ by

$$\hat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i) \neq Y_i\}}$$

where I denotes the indicator function. Let $\tilde{\phi}_n$ denote a classifier chosen from $\mathcal{C}$ by minimizing $\hat{L}_n(\phi)$, i.e., $\hat{L}_n(\tilde{\phi}_n) \leq \hat{L}_n(\phi)$, $\phi \in \mathcal{C}$. Recently much attention has been paid to analyzing the error probability $L(\tilde{\phi}_n)$. If $\inf_{\phi \in \mathcal{C}} L(\phi) = 0$, then naturally, $\hat{L}_n(\tilde{\phi}_n) = 0$ almost surely, and for every n and $\epsilon > 0$

$$\begin{aligned} \boldsymbol{P}\{L(\tilde{\phi}_n) \geq \epsilon\} &\leq \boldsymbol{P}\left\{\sup_{\phi \in \mathcal{C}; \hat{L}_n(\phi) = 0} L(\phi) \geq \epsilon\right\} \\ &\leq 2S(\mathcal{C}, 2n)2^{-n\epsilon/2} \leq 2\left(\frac{2en}{V}\right)^V 2^{-ne/2} \end{aligned} \tag{1}$$

(see Devroye and Wagner [13], Vapnik and Chervonenkis [24], Vapnik [21], Blumer *et al.* [6], and Lugosi [16] for different versions of the inequality). Clearly, this inequality is only useful if $V < \infty$. Unfortunately, as classes with finite VC dimension are always very small, the condition $\inf_{\phi \in \mathcal{C}} L(\phi) = 0$ is very restrictive. Vapnik and Chervonenkis [23], [24] proved distribution-free exponential inequalities for empirical error minimization. Following their work, several improvements have been proven. For most interesting values

of n and ϵ, one of the tightest bounds was given by Devroye [10], who showed that for every n and $\epsilon > 0$, and for all distributions of (X, Y), we have

$$\boldsymbol{P}\left\{\sup_{\phi\in\mathcal{C}}|\hat{L}_n(\phi) - L(\phi)| \geq \epsilon\right\} \leq 4e^8 S(\mathcal{C}, n^2)e^{-2n\epsilon^2} \quad (2)$$

and

$$\boldsymbol{P}\left\{L(\tilde{\phi}_n) - \inf_{\phi\in\mathcal{C}} L(\phi) \geq \epsilon\right\} \leq 4e^8 S(\mathcal{C}, n^2)e^{-n\epsilon^2/2}. \quad (3)$$

The strength of these inequalities is that they are valid for all distributions of (X, Y). One of the implications is that

$$\boldsymbol{E}L(\tilde{\phi}_n) - \inf_{\phi\in\mathcal{C}} L(\phi) \leq c\sqrt{\frac{V\log n}{n}}$$

where c is a universal constant (independent of the distribution). Thus the error probability of the empirically chosen decision is always within $O(\sqrt{\log n/n})$ of that of the best in $\mathcal{C}$. Unfortunately, if $V < \infty$, then for some distributions, $\inf_{\phi\in\mathcal{C}} L(\phi)$ may be arbitrarily far from the Bayes risk L^*. On the other hand, if $V = \infty$, then $L(\tilde{\phi}_n) - \inf_{\phi\in\mathcal{C}} L(\phi)$ will be large for some distributions [6], [12], [24].

A typical approach for resolving the conflict is the "method of sieves." Here one takes a sequence $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ of classes of classifiers such that all have finite VC dimension, yet they "grow" in the sense that for any distribution, the error probability of the best classifier in $\mathcal{C}^{(k)}$ converges to L^* as $k \to \infty$. Then we may obtain $\tilde{\phi}_n$ from the k_nth class $\mathcal{C}^{(k_n)}$ by minimizing the empirical error probability over $\mathcal{C}^{(k_n)}$, where the integer k_n is some prespecified function of the sample size n only. The integer k_n basically determines the complexity of the class from which the decision rule is selected. Typically, k_n should grow with n in order to assure convergence of the approximation error $\inf_{\phi\in\mathcal{C}^{(k_n)}} L(\phi) - L^*$, but it cannot grow too rapidly, for otherwise the estimation error $L(\tilde{\phi}_n) - \inf_{\phi\in\mathcal{C}^{(k_n)}} L(\phi)$ might fail to converge to zero. For the overall error of this rule, we clearly have

$$\boldsymbol{E}\{L(\tilde{\phi}_n)\} - L^* \leq c\sqrt{\frac{V_{k_n}\log n}{n}} + \left(\inf_{\phi\in\mathcal{C}^{(k_n)}} L(\phi) - L^*\right) \quad (4)$$

for a universal constant c. It can be shown that it is possible to choose the sequence of classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ and the sequence $\{k_n\}$ such that the error probability of the selected classifier converges to L^* with probability one for all distributions (see, e.g., Vapnik [22], Devroye [11], Faragó and Lugosi [14]). Ideally, to get the best performance, the two types of error should be about the same order of magnitude. Clearly, however, a prespecified choice of the complexity k_n cannot balance the two sides of the tradeoff for all distributions. Therefore, it is important to find methods such that the classifier is selected from a class whose index is automatically determined by the data D_n.

A possible solution to this problem may be derived from the idea of *structural risk minimization* (see Vapnik and Chervonenkis [24] and Vapnik [21]), also known as *complexity regularization* (see Barron [2], [3] and Barron and Cover [4]). The basic idea is to minimize the sum of the empirical error and a term corresponding to the "complexity" of the candidate classifier. In our application, this complexity is a simple function of the VC dimension of the class from which the candidate classifier is taken.

The idea of minimizing the sum of the empirical error and a term penalizing the complexity has been investigated in various statistical problems by, e.g., Akaike [1], Schwarz [19], Rissanen [17], [18], Barron [2], [3], Barron and Cover [4], Vapnik and Chervonenkis [24], and Vapnik [21].

In this paper we analyze a method essentially due to Vapnik and Chervonenkis [24]. It is shown to produce a classifier ϕ_n^* that finds a nearly optimal balance between the approximation and the estimation error in the sense that the expected value of its error probability satisfies

$$\boldsymbol{E}\{L(\phi_n^*)\} - L^* \leq \inf_{k\geq 1}\left(\sqrt{\frac{16V_k\log n + 8(k+11)}{n}} + \left(\inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) - L^*\right)\right)$$

where V_k is the VC dimension of $\mathcal{C}^{(k)}$ (Theorem 1). To see what the above inequality means, observe that the essential improvement over (4) is the infimum over all k appearing on the right-hand side, as opposed to the predetermined k_n in (4). The first term inside the parentheses is essentially the same as the best distribution-free upper bound on the estimation error when the empirical error is minimized over the kth class $\mathcal{C}^{(k)}$. The second term is the approximation error in the same situation. Since the overall error is bounded from above by the infimum of these sums, we can say that the obtained distribution-free performance guarantee for ϕ_n^* is essentially the same as the best bound we could get if we had known the optimal k beforehand. We emphasize that the optimal k depends on the distribution, and the strength of this method lies in the fact that the above inequality is true for *all distributions* of (X, Y).

Now, it is not surprising that the method is strongly universally consistent (see our Corollary 1), i.e., for any distribution of (X, Y)

$$\lim_{n\to\infty} L(\phi_n^*) = L^* \text{ with probability one.}$$

Another corollary of the inequality above is that if in addition it is assumed that the Bayes decision g^* is a member of a class $\mathcal{C}^*$ which can be written as a countable union of classes each with finite VC dimension, then

$$\boldsymbol{E}L(\phi_n^*) - L^* \leq c\sqrt{\frac{\log n}{n}}$$

where the constant c depends on the distribution (Corollary 2). We emphasize that the rate $O(\sqrt{\log n/n})$ is achieved for every distribution in a very large *nonparametric* class of distributions, as the condition imposed on (X, Y) involves merely the form of the optimal decision g^*.

We also address the case when $g^* \in \mathcal{C}^*$ and $L^* = 0$, i.e., Y is a function of X (namely, $Y = g^*(X)$). This is the usual setup in Valiant's learning theory. We show in Theorem 2 that if $\mathcal{C}^*$ can be written as a countable union of classes with finite

VC dimension, then the method of minimizing a complexity penalized error estimate yields a classifier ϕ_n^*, with

$$\boldsymbol{E}L(\phi_n^*) \le c\frac{\log n}{n}.$$

Interestingly, in this case we have great freedom in defining the complexity penalty.

II. Structural Risk Minimization

Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ be a sequence of classifiers, from which we wish to select a classifier with the help of the training data D_n. The method of selecting a classifier that we describe next is based on minimizing the sum of the empirical error $\hat{L}_n$ and a complexity term over the union of the classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$.

For every n and j, we introduce the complexity penalty

$$r(j, n) = \sqrt{\frac{\log\left(4e^8 S(\mathcal{C}^{(j)}, n^2)\right) + j}{2n}}.$$

The penalty $r(j, n)$ is basically an estimate of the magnitude of the error caused by overfitting. Its choice is motivated by (2), and its usefulness will be apparent from the proof of Theorem 1. The algorithm is defined as follows: Let $\tilde{\phi}_{n,1}, \tilde{\phi}_{n,2}, \cdots$ be classifiers minimizing the empirical error $\hat{L}_n(\phi)$ over the classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$, respectively. For $\phi \in \mathcal{C}^{(j)}$, define the complexity-penalized error estimate

$$\tilde{L}_n(\phi) = \hat{L}_n(\phi) + r(j, n).$$

Finally, select the classifier ϕ_n^* minimizing the complexity penalized error estimate $\tilde{L}_n(\tilde{\phi}_{n,j})$ over $j \ge 1$. We will refer to ϕ_n^* as the *classification rule based on structural risk minimization.*

By a well-known inequality connecting shatter coefficients and the VC dimension, $S(\mathcal{C}, n) \le (ne/V)^V$ (e.g., Vapnik and Chervonenkis [23]), we see that the size of the complexity term $r(j, n)$ is approximately a constant times $\sqrt{(V_j \log n + j)/n}$, where V_j is the VC dimension of the class $\mathcal{C}^{(j)}$. In most typical applications the sequence $V_1, V_2, \cdots$ is strictly monotone increasing, therefore $V_j \ge j$, and the complexity is monotone increasing in j. The intuition—already suggested by Vapnik and Chervonenkis [24]—is that in larger classes the danger of overfitting the data is greater, and the complexity penalty is intended to compensate for the overfitting error. The main properties of the selected classifier are summarized in the following result.

Theorem 1: Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ be a sequence of classes of classifiers whose VC dimensions $V_1, V_2, \cdots$ are finite. Let ϕ_n^* be the classification rule based on structural risk minimization. Then for all n and k, and all $\epsilon \ge 4r(k, n)$, we have

$$\boldsymbol{P}\left\{L(\phi_n^*) - \inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) \ge \epsilon\right\} \le e^{-n\epsilon^2/2} + 4e^8 S(\mathcal{C}^{(k)}, n^2) e^{-n\epsilon^2/8}$$

and in particular, for all n

$$\boldsymbol{E}\{L(\phi_n^*)\} - L^* \le \inf_{k \ge 1}\left(\sqrt{\frac{16V_k \log n + 8(k+11)}{n}} + \left(\inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) - L^*\right)\right).$$

As explained in the Introduction, the above result indicates that our method finds a nearly optimal balance between the approximation and estimation errors.

Remark: Before stating a few corollaries of Theorem 1, let us briefly compare it to analogous results obtained by Barron and Cover [4] and Barron [3]. For example, a result in [3] states the following. For each n, let $\Gamma_n : \mathcal{R}^d \to \{0, 1\}$ be a countable list of classifiers and assign the complexity penalty $C(\phi, n)$ to each $\phi \in \Gamma_n$ such that they satisfy the Kraft-type inequality

$$\sum_{\phi \in \Gamma_n} 2^{-C(\phi, n)} < \infty.$$

If $\hat{\phi}_n$ is chosen from Γ_n to minimize the penalized error

$$\hat{L}_n(\phi) + c\sqrt{\frac{C(\phi, n)}{n}}$$

for an appropriate constant c, then

$$\boldsymbol{E}\{L(\hat{\phi}_n)\} - L^* = O\left(\inf_{\phi \in \Gamma_n}\left(\sqrt{\frac{C(\phi, n)}{n}} + (L(\phi) - L^*)\right)\right).$$

Barron calls the quantity within the parentheses on the right-hand side the *index of resolvability.* To make the comparison transparent, we may rewrite Theorem 1 as

$$\boldsymbol{E}\{L(\hat{\phi}_n)\} - L^* = O\left(\inf_{\phi \in \mathcal{C}^*}\left(\sqrt{\frac{C'(\phi, n)}{n}} + (L(\phi) - L^*)\right)\right)$$

where

$$\mathcal{C}^* = \bigcup_{k=1}^{\infty} \mathcal{C}^{(k)}$$

and

$$C'(\phi, n) = \log\left(S(\mathcal{C}^{(j)}, n)\right) + j$$

for each $\phi \in \mathcal{C}^{(j)}$. The significant difference between the two inequalities is that we can take the infimum over a much larger, *uncountable* set $\mathcal{C}^*$ of candidates. On the other hand, our result is less general, as the penalties are specifically defined in terms of the shatter coefficients. It is apparent from the proof that a Kraft-type summability is crucial in our case as well.

Next, we review some of the implications of this result. The first corollary states that the obtained classification rule is strongly universally consistent. The only conditions are that each class in the sequence has finite VC dimension, and the classes "approximate" the Bayes rule g^* well for all distributions.

Corollary 1: Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ be a sequence of classes of classifiers with finite VC dimensions $V_1, V_2, \cdots$ such that for any distribution of (X, Y)

$$\lim_{j \to \infty} \inf_{\phi \in \mathcal{C}^{(j)}} L(\phi) = L^*.$$

Then the classification rule ϕ_n^* based on structural risk minimization satisfies

$$\lim_{n \to \infty} L(\phi_n^*) = L^* \text{ with probability one}$$

for any distribution of (X, Y).

We remark here that the first conditions are satisfied by many sequences of classes $\mathcal{C}^{(j)}$, such as classes of histogram-type classifiers, generalized linear classifiers, neural networks, etc.

Corollary 1 shows that the method of structural risk minimization is universally consistent under very mild conditions on the sequence of classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$. This property, however, is shared by the minimization of the empirical error over the class $\mathcal{C}^{(k_n)}$, where k_n is a properly chosen function of the sample size n. The next special case displays the strength of structural risk minimization.

Corollary 2: Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ be a sequence of classes of classifiers such that the VC dimensions $V_1, V_2, \cdots$ are all finite. Assume further that the Bayes rule is contained in the union of these classes, i.e.

$$g^* \in \mathcal{C}^* \stackrel{\text{def}}{=} \bigcup_{j=1}^{\infty} \mathcal{C}^{(j)}.$$

Let K be the smallest integer such that $g^* \in C^{(K)}$. Then for every n and any $\epsilon > 4r(K, n)$, the error probability of the classification rule based on structural risk minimization ϕ_n^* satisfies

$$\boldsymbol{P}\{L(\phi_n^*) - L^* > \epsilon\} \le e^{-n\epsilon^2/2} + 4e^8 S(\mathcal{C}^{(K)}, n^2) e^{-n\epsilon^2/8}.$$

Furthermore

$$\boldsymbol{E}L(\phi_n^*) - L^* \le 4\sqrt{\frac{V_K \log n + (K/2) + 6}{n}}.$$

Corollary 2 shows that the rate of convergence is always of the order of $\sqrt{\log n/n}$, and the constant factor V_K depends on the distribution. The number V_K may be viewed as the inherent complexity of the Bayes rule for the distribution. The intuition is that the simplest rules are contained in $\mathcal{C}^{(1)}$, and more complex rules are added to the class as the index of the class increases. The bound on the error is about the same as if we had known K beforehand and minimized the empirical error over $\mathcal{C}^{(K)}$. One great advantage of structural risk minimization (similarly to minimum description length, automatic model selection, and other complexity regularization methods [1]–[4], [17]) is that it automatically finds where to look for the optimal classifier.

Corollary 2 may be rephrased as follows. Assume that the distribution of (X, Y) is such that the Bayes rule g^* is a member of a known class $\mathcal{C}^*$ that can be written as a countable union of classes with finite VC dimension. Then there exists a classification rule ϕ_n^* whose error probability converges to the Bayes error L^* at an $O(\sqrt{\log n/n})$ rate. This is a very fast rate of convergence for a huge class of distributions of (X, Y). The only condition on the joint distribution is that $g^* \in \mathcal{C}^*$. This is clearly not very severe, as no assumption is imposed on the distribution of X, and $\mathcal{C}^*$ can be a large class with infinite VC dimension. We emphasize that in order to achieve the $O(\sqrt{\log n/n})$ rate of convergence, we do not have to assume that the distribution is a member of a known finite-dimensional parametric family. The condition is imposed solely on the form of the Bayes classifier g^*. The only requirement is that $\mathcal{C}^*$ should be written as a countable union of classes of finite VC dimension. One can appreciate this guaranteed rate of convergence by recalling Devroye's [9] result, which states that for any sequence of classification rules there exists a distribution of (X, Y) such that the rate of convergence of the error probability to L^* is arbitrarily slow.

Remark: The empirical risk $\hat{L}_n$ of the classifier selected by empirical error minimization is usually an optimistically biased estimate of its error probability. However, from the proof of Theorem 1, we see the following by-product:

$$\boldsymbol{P}\{L(\phi_n^*) - \tilde{L}_n(\phi_n^*) \ge \epsilon\} \le e^{-2n\epsilon^2}.$$

This means that the penalized error estimate of the selected classification rule cannot be much larger than the actual error probability. In other words, the designer can be confident about not having much larger error probabilities than the estimated one.

A disadvantage of the method is that it requires thorough knowledge of the shatter coefficients (or at least the VC dimension) of the classes $\mathcal{C}^{(j)}$. For nested sequences, i.e., when $\mathcal{C}^{(1)} \subset \mathcal{C}^{(2)} \subset \cdots$, Buescher and Kumar [7], [8] proposed a general method which does not require any knowledge of the shatter coefficients. Their method, "simple empirical covering," has the universal consistency property, as in Corollary 1. On the other hand, their method seems to have a slower rate of convergence than structural risk minimization under the conditions of Corollary 2. Interestingly, as we will see in Theorem 2, in some situations we have a tremendous freedom in defining the complexity penalties.

III. Proof of Theorem 1

First we prove the probability inequality. Observe that

$$\begin{aligned}
&\boldsymbol{P}\left\{L(\phi_n^*) - \inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) \ge \epsilon\right\} \\
&\quad \le \boldsymbol{P}\left\{L(\phi_n^*) - \inf_{j \ge 1} \tilde{L}_n(\tilde{\phi}_{n,j}) \ge \frac{\epsilon}{2}\right\} \\
&\qquad + \boldsymbol{P}\left\{\inf_{j \ge 1} \tilde{L}_n(\tilde{\phi}_{n,j}) - \inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) \ge \frac{\epsilon}{2}\right\}. \qquad (5)
\end{aligned}$$

The first term on the right-hand side of the inequality may be bounded as follows:

$$\begin{aligned}
&\boldsymbol{P}\left\{L(\phi_n^*) - \inf_{j \ge 1} \tilde{L}_n(\tilde{\phi}_{n,j}) \ge \frac{\epsilon}{2}\right\} \\
&\quad = \boldsymbol{P}\left\{L(\phi_n^*) - \tilde{L}_n(\phi_n^*) \ge \frac{\epsilon}{2}\right\} \\
&\quad \le \boldsymbol{P}\left\{\sup_{j \ge 1} (L(\tilde{\phi}_{n,j}) - \tilde{L}_n(\tilde{\phi}_{n,j})) \ge \frac{\epsilon}{2}\right\} \\
&\quad = \boldsymbol{P}\left\{\sup_{j \ge 1} (L(\tilde{\phi}_{n,j}) - \hat{L}_n(\tilde{\phi}_{n,j}) - r(j, n)) \ge \frac{\epsilon}{2}\right\} \\
&\quad \le \sum_{j=1}^{\infty} \boldsymbol{P}\left\{|L(\tilde{\phi}_{n,j}) - \hat{L}_n(\tilde{\phi}_{n,j})| \ge \frac{\epsilon}{2} + r(j, n)\right\}
\end{aligned}$$

$$\le \sum_{j=1}^{\infty} 4e^8 S(\mathcal{C}^{(j)}, n^2) e^{-2n(\epsilon/2+r(j,n))^2} \text{ (by (2))}$$
$$\le \sum_{j=1}^{\infty} 4e^8 S(\mathcal{C}^{(j)}, n^2) e^{-2nr^2(j,n)} e^{-n\epsilon^2/2}$$
$$= e^{-n\epsilon^2/2} \sum_{j=1}^{\infty} e^{-j} \le e^{-n\epsilon^2/2}$$

where we have substituted the defining expression for $r(j, n)$ to obtain the last equality. For the second term on the right-hand side of (5) we have

$$\boldsymbol{P}\left\{\inf_{j\ge 1} \tilde{L}_n(\tilde{\phi}_{n,j}) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) \ge \frac{\epsilon}{2}\right\}$$
$$\le \boldsymbol{P}\left\{\tilde{L}_n(\tilde{\phi}_{n,k}) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) \ge \frac{\epsilon}{2}\right\}$$
$$= \boldsymbol{P}\left\{\hat{L}_n(\tilde{\phi}_{n,k}) + r(k,n) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) \ge \frac{\epsilon}{2}\right\}$$
$$\le \boldsymbol{P}\left\{\hat{L}_n(\tilde{\phi}_{n,k}) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) \ge \frac{\epsilon}{4}\right\}$$
(since by assumption $r(k, n) \le \epsilon/4$)
$$\le \boldsymbol{P}\left\{\sup_{\phi\in\mathcal{C}^{(k)}} |\hat{L}_n(\phi) - L(\phi)| \ge \frac{\epsilon}{4}\right\}$$
$$\le 4e^8 S(\mathcal{C}^{(k)}, n^2) e^{-n\epsilon^2/8} \text{ (by (2))}$$

which proves the first inequality in Theorem 1. The inequality for the expected error probability follows from the previous inequality by the following simple argument. Note that

$$\boldsymbol{E}\{L(\phi_n^*)\} - L^* = \inf_{k\ge 1}\left(\left(\boldsymbol{E}\{L(\phi_n^*)\} - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi)\right) + \left(\inf_{\phi\in\mathcal{C}^{(k)}} L(\phi) - L^*\right)\right).$$

To bound the estimation error, fix k and write

$$\left(\boldsymbol{E}\{L(\phi_n^*)\} - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi)\right)^2$$
$$\le \boldsymbol{E}\left\{\left(L(\phi_n^*) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi)\right)^2\right\}$$
(by Jensen's inequality)
$$= \int_0^{\infty} \boldsymbol{P}\left\{\left(L(\phi_n^*) - \inf_{\phi\in\mathcal{C}^{(k)}} L(\phi)\right)^2 > t\right\} dt$$
$$\le u + \int_u^{\infty} (e^{-nt/2} + 4e^8 S(\mathcal{C}^{(k)}, n^2) e^{-nt/8})\, dt$$
$$\le u + \int_u^{\infty} (e^{-nt/2} + 4e^8 S(\mathcal{C}^{(k)}, n^2) e^{-nt/8})\, dt$$
(by the probability inequality, for any $u \ge 16r^2(k, n)$)
$$\le \frac{16V_k \log n + 8(k+11)}{n}$$

where we chose $u = 16r^2(k, n)$ and used the inequality $S(\mathcal{C}^{(k)}, n) \le n^{V_k}$. The theorem is thus proved. □

IV. Nonprobabilistic Concepts

In Valiant's [20] framework of learning theory, it is assumed that g^* is a member of a known class $\mathcal{C}$ of classifiers, and moreover, $L^* = 0$ (see also Blumer *et al.* [6]). In this case, we see from (1) that if $\mathcal{C}$ has a finite VC dimension V, then $\boldsymbol{E}\{L(\tilde{\phi}_n)\} \le c_0 V \log n/n$, where $\tilde{\phi}_n$ is a classifier minimizing the empirical error $\hat{L}_n$ over $\mathcal{C}$ (i.e., $\hat{L}_n(\tilde{\phi}_n) = 0$) and c_0 is a universal constant. It is also well known that if $V = \infty$, then there exists a universal constant c_1 such that for every n and every classification rule ϕ_n, $\boldsymbol{E}\{L(\phi_n)\} > c_1$ for some distribution (see Vapnik and Chervonenkis [24], and Haussler, Littlestone, and Warmuth [15]). Benedek and Itai [5] demonstrate a selection algorithm with a guaranteed rate of convergence of the expected error probability to zero. In this section, we demonstrated that the idea of complexity regularization can be applied in this setup as well. In particular, we show that if the Bayes rule g^* is contained in a known class that can be written as a union of classes with finite VC dimensions, then there is a classification rule ϕ_n^* such that $\boldsymbol{E}\{L(\phi_n^*)\} \le c \log n/n$, where the constant c depends (necessarily) on the distribution. This rate is always faster than that offered by the algorithm of Benedek and Itai. The solution technique is again complexity regularization, but this time the conditions on the penalty term are very mild.

Assume that $L^* = 0$ and $g^* \in \mathcal{C}^*$, where

$$\mathcal{C}^* = \bigcup_{j=1}^{\infty} \mathcal{C}^{(j)}$$

for some classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ with finite VC dimensions $V_1, V_2, \cdots$. Without loss of generality we assume that the classes are disjoint. Define

$$\mathcal{B}^{(i)} = \bigcup_{j=1}^{i} \mathcal{C}^{(j)}$$

so that the $\mathcal{B}^{(i)}$'s are nested, i.e., $\mathcal{B}^{(1)} \subseteq \mathcal{B}^{(2)} \subseteq \cdots$. As before, we define the classification rule ϕ_n^* as one minimizing a complexity penalized error estimate

$$\tilde{L}_n(\phi) = \hat{L}_n(\phi) + r(j, n) \tag{6}$$

over $\mathcal{C}^*$, where the penalty $r(j, n)$ is added to each $\phi \in \mathcal{B}^{(j)}$. The following result states that a wide variety of penalties provide a very fast convergence rate of the error probability of the selected classification rule.

Theorem 2: Assume that the Bayes risk L^* is zero and the Bayes decision g^* lies in

$$\mathcal{C}^* = \bigcup_{j=1}^{\infty} \mathcal{C}^{(j)}$$

for some sequence of disjoint classes $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \cdots$ with finite VC dimensions $V_1, V_2, \cdots$. Assume that the penalty function $r(j, n)$ satisfies the following conditions:

1) $r(j, n)$ is strictly monotone increasing in j; and
2) for each $j > 1$

$$\lim_{n\to\infty} (r(j, n) - r(j-1, n)) = 0.$$

Then the error probability of the classification rule ϕ_n^* defined in (6) satisfies

$$\boldsymbol{E}\{L(\phi_n^*)\} = O\left(\frac{\log n}{n}\right).$$

Remark: We note that for Condition 2 to hold it is sufficient that $\lim_{n\to\infty} r(j, n) = 0$ for each $j \geq 1$.

Proof: Let k be the smallest number such that $\inf_{\phi\in\mathcal{B}^{(k)}} L(\phi) = 0$. Denote

$$a_j = \inf_{\phi\in\mathcal{B}^{(j)}} L(\phi).$$

Clearly, $a_{k-1} > 0$. Assume (using Condition 2) that n is sufficiently large such that

$$r(k, n) - r(k-1, n) < a_{k-1}/2.$$

Observe that with probability 1,

$$\inf_{\phi\in\mathcal{B}^{(j)}} \hat{L}_n(\phi) = 0$$

for all $j \geq k$. With some abuse of notation, we write $\phi_n^* \in \mathcal{C}^{(j)}$, if the selected classifier is a member of the jth class $\mathcal{C}^{(j)}$. First we show that if n is large enough, then with very high probability $\phi_n^* \in \mathcal{C}^{(k)}$, and then we use (1) to conclude the proof. Thus

$$\begin{aligned}
&\boldsymbol{P}\{\phi_n^* \in \mathcal{B}^{(k-1)}\} \\
&\quad \leq \boldsymbol{P}\left\{\inf_{\phi\in\mathcal{B}^{(k-1)}} \tilde{L}_n(\phi) \leq \inf_{\phi\in\mathcal{C}^*-\mathcal{B}^{(k-1)}} \tilde{L}_n(\phi)\right\} \\
&\quad \leq \boldsymbol{P}\left\{\inf_{\phi\in\mathcal{B}^{(k-1)}} \hat{L}_n(\phi) + r(k-1, n) \leq r(k, n)\right\} \\
&\qquad \text{(since for } \phi \in \mathcal{B}^{(k-1)},\ \tilde{L}_n(\phi) \geq \hat{L}_n(\phi) + r(k-1, n) \\
&\qquad \text{by Condition 1, and } \inf_{\phi\in\mathcal{C}^*-\mathcal{B}^{(k-1)}} \tilde{L}_n(\phi) \leq r(k, n), \\
&\qquad \text{again by Condition 1)} \\
&\quad \leq \boldsymbol{P}\left\{\sup_{\phi\in\mathcal{B}^{(k-1)}} |\hat{L}_n(\phi) - L(\phi)| > a_{k-1}/2\right\} \\
&\qquad \text{(by the definition of } a_{k-1}, \text{ and} \\
&\qquad \text{since } n \text{ is large enough)} \\
&\quad \leq 4e^8(ne)^{2W_{k-1}} e^{-na_{k-1}^2/2}
\end{aligned}$$

by (2), where

$$W_j \leq \sum_{i=1}^{j} V_i < \infty.$$

This follows from the fact that since

$$\mathcal{B}^{(j)} = \bigcup_{i=1}^{j} \mathcal{C}^{(i)}$$

$$S(\mathcal{B}^{(j)}, n) \leq \prod_{i=1}^{j} S(\mathcal{C}^{(i)}, n) \leq \prod_{i=1}^{j} (ne/V_i)^{V_i} \leq (ne)^{W_j}.$$

Thus with very large probability

$$\phi_n^* \in \bigcup_{j=k}^{\infty} \mathcal{C}^{(j)}.$$

Notice, however, that also $\phi_n^* \in \mathcal{C}^{(k)}$, since for each $j \geq k$, $\inf_{\phi\in\mathcal{B}^{(j)}} \hat{L}(\phi) = 0$, and $r(k, n)$ is monotone increasing in k. Therefore, for every $\epsilon > 0$

$$\begin{aligned}
\boldsymbol{P}\{L(\phi_n^*) > \epsilon \mid \phi_n^* \in \mathcal{B}^{(k-1)}\} &\leq \boldsymbol{P}\left\{\sup_{\phi\in\mathcal{C}^{(k)}:\hat{L}_n(\phi)=0} L(\phi) > \epsilon\right\} \\
&\leq 2\left(\frac{2en}{V_k}\right)^{V_k} 2^{-n\epsilon/2}.
\end{aligned}$$

Summarizing, we have shown that if n is sufficiently large such that $r(k, n) - r(k-1, n) < a_{k-1}/2$, then for all $\epsilon > 0$

$$\begin{aligned}
\boldsymbol{P}\{L(\phi_n^*) > \epsilon\} &\leq \boldsymbol{P}\{\phi_n^* \in \mathcal{B}^{(k-1)}\} \\
&\quad + \boldsymbol{P}\{L(\phi_n^*) > \epsilon \mid \phi_n^* \in \mathcal{B}^{(k-1)}\} \\
&\leq 4e^8 n^{2W_{k-1}} e^{-na_{k-1}^2/2} + 2\left(\frac{2en}{V_k}\right)^{V_k} 2^{-n\epsilon/2}.
\end{aligned}$$

The statement for $\boldsymbol{E}\{L(\phi_n^*)\}$ now follows easily: If n is sufficiently large such that the above probability inequality is satisfied, then for all $t \in (0, 1)$

$$\begin{aligned}
&\boldsymbol{E}\{L(\phi_n^*)\} \\
&\quad \leq t + \boldsymbol{P}\{L(\phi_n^*) > t\} \\
&\quad \leq t + 4e^8 n^{2W_{k-1}} e^{-na_{k-1}^2/2} + 2\left(\frac{2en}{V_k}\right)^{V_k} 2^{-nt/2} \\
&\quad = 4e^8 n^{2W_{k-1}} e^{-na_{k-1}^2/2} + \frac{(2V_k+2)\log n + (2e)^{V_k}}{n} \\
&\qquad \text{(choose } t = (2V_k+2)\log n/n)
\end{aligned}$$

which concludes the proof. □

ACKNOWLEDGMENT

The authors wish to thank the two reviewers and A. Nobel for helpful suggestions.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[2] A. R. Barron, "Logically smooth density estimation," Tech. Rep. TR 56, Dept. of Statistics, Stanford Univ., Stanford, CA, 1985.
[3] ——, "Complexity regularization with application to artificial neural networks," in G. Roussas, Ed., *Nonparametric Functional Estimation and Related Topics* (NATO ASI ser.). Dordrecht, The Netherlands: Kluwer, 1991.
[4] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.
[5] G. M. Benedek and A. Itai, "Nonuniform learnability," *J. Comput. Syst. Sci.*, vol. 48, pp. 311–323, 1994.
[6] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik–Chervonenkis dimension," *J. ACM*, vol. 36, pp. 929–965, 1989.
[7] K. L. Buescher, "Learning and smooth simultaneous estimation based on empirical data," Ph.D. dissertation, Univ. of Illinois, Urbana-Champaign, 1992.
[8] K. L. Buescher and P. R. Kumar, "Learning by canonical smooth estimation, Part II: Learning and choice of model complexity," to appear in *IEEE Trans. Automat. Contr.*, 1995.
[9] L. Devroye, "Any discrimination rule can have an arbitrarily bad probability of error for finite sample size," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-4, pp. 154–157, 1982.
[10] ——, "Bounds for the uniform deviation of empirical measures," *J. Multivariate Anal.*, vol. 12, pp. 72–79, 1982.
[11] ——, "Automatic pattern recognition: A study of the probability of error," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, pp. 530–543, 1988.

[12] L. Devroye and G. Lugosi, "Lower bounds in pattern recognition and learning," *Pattern Recogn.*, 1995, to appear.
[13] L. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Tech. Rep. 183, Electronics Res. Ctr., Univ. of Texas, Austin, 1976.
[14] A. Faragó and G. Lugosi, "Strong universal consistency of neural network classifiers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1146–1151, 1993.
[15] D. Haussler, N. Littlestone, and M. Warmuth, "Predicting $\{0, 1\}$ functions from randomly drawn points," in *Proc. 29th IEEE Symp. on Foundations of Computer Science.* Los Alamitos, CA: IEEE Computer Soc. Press, 1988, pp. 100–109.
[16] G. Lugosi, "Improved upper bounds for probabilities of uniform deviations," *Statist. Probability Lett.*, vol. 25, pp. 71–77, 1995.
[17] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *Annals Statist.*, vol. 11, pp. 416–431, 1983.
[18] ——, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.
[19] G. Schwarz, "Estimating the dimension of a model," *Annals Statist.*, vol. 6, pp. 461–464, 1978.
[20] L. G. Valiant, "A theory of the learnable," *Commun. ACM*, vol. 27, pp. 1134–1142, 1984.
[21] V. N. Vapnik, *Estimation of Dependencies Based on Empirical Data.* New York: Springer, 1982.
[22] ——, "Inductive principles of the search for empirical dependencies (methods based on weak convergence of probability measures," in *Proc. 2nd Annual Workshop on Computational Learning Theory*, 1989, pp. 3–24.
[23] 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.
[24] ——, *Theory of Pattern Recognition.* Moscow: Nauka, 1974 (in Russian); German translation: *Theorie der Zeichenerkennung.* Berlin: Akademie Verlag, 1979.