1 of 1

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36.

TITLE: *Selecting Model Complexity in Learning Problems*

AUTHOR(S): *Kevin L. Buescher, Los Alamos National Laboratory
P. R. Kumar, Coordinated Science Laboratory*

SUBMITTED TO: *32nd IEEE Conference on Decision and Control
(December 1993)*

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

RECEIVED
OCT 07 1993
OSTI

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy

Los Alamos MASTER
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Selecting Model Complexity in Learning Problems

Kevin L. Buescher *
Los Alamos National Laboratory
MS F645
Los Alamos, NM 87545
klb@lanl.gov

P. R. Kumar
Coordinated Science Laboratory
University of Illinois
1308 W. Main Street
Urbana, IL 61801-2307
prkumar@gauss.csl.uiuc.edu

Abstract

To learn (or generalize) from noisy data, one must resist the temptation to pick a model for the underlying process that overfits the data. Many existing techniques solve this problem at the expense of requiring the evaluation of an absolute, a priori measure of each model's complexity. We present a method that does not. Instead, it uses a natural, relative measure of each model's complexity. This method first creates a pool of "simple" candidate models using part of the data and then selects from among these by using the rest of the data.

1 Introduction

Suppose that we are given data that consists of noisy observations of a fixed function, and we want to find a model for this function that will accurately predict new observations. This is an example of a *learning problem*. If we decide to model this underlying function with, say, a polynomial, then it is important to fit the data to a polynomial of the proper order, or "complexity." Otherwise, we run the risk of picking a model that cannot adequately represent the underlying function or, worse yet, one that primarily fits the noise. In this paper, we investigate a means of selecting the proper complexity of models in learning problems.

The learning problems we treat are basically those addressed in [1]. The definition of "learning" we use is closely related to that in [2], which extends the notion of learning proposed in [3] and refined in [4].

In our learning framework, labeled sample points (x_i, y_i) from $S = X \times Y$ are drawn independently at random according to a fixed, unknown probability distribution P (P is assumed to lie in a known set of distributions, $\mathcal{P}$). In this paper, we will take $Y \subset \mathbb{R}^q$. The models that we can use, which we shall henceforth call

"hypotheses," are drawn from a class $\mathcal{H}$ of $\mathbb{R}^q$ -valued functions on X . With each hypothesis $h \in \mathcal{H}$ we associate an *error* with respect to P , the "generalization error." This is the expected value of the difference between $h(x')$ and y' at a new point (x', y') randomly drawn according to P . To *learn* from a sequence of labeled sample points, we must select a hypothesis that (with high probability) has nearly the least error of all hypotheses in $\mathcal{H}$.

We will focus on hypothesis classes that are *nested* in that $\mathcal{H}$ is the union of an increasing sequence of sets of hypotheses: $\mathcal{H}^1 \subseteq \mathcal{H}^2 \subseteq \dots = \mathcal{H}$. For instance, if $\mathcal{H}$ is the class of all multinomials, $\mathcal{H}^k$ might consist of k 'th order multinomials. It is natural to think of hypotheses that are in $\mathcal{H}^i$ but not in $\mathcal{H}^j$ as being more complex than the hypotheses in $\mathcal{H}^j$.

A pitfall in learning problems is the tendency to use a hypothesis that is overly complex and that *overfits* the data. That is, we may pick a hypothesis that agrees with the data and yet has a large generalization error. Many of the remedies for this problem in the current literature involve evaluating a measure of the complexity of each $\mathcal{H}^k$ in order to influence which $\mathcal{H}^k$ to pick a hypothesis from. Determining these complexities may involve a great deal of analysis, and often the procedure for influencing which $\mathcal{H}^k$ to use does not take the observed distribution of the data into account.

In this paper we shall present a learning procedure that picks hypotheses of appropriate complexity and yet does not involve any a priori, absolute measures of complexity. Instead, it uses the natural notion of complexity associated with the nested structure of $\mathcal{H}$ in order to compare hypotheses. This procedure first creates a pool of "simple" candidate hypotheses using part of the data and then selects from among these by using the rest of the data. In this way, the actual distribution of the data determines, in part, the complexity of the chosen hypothesis.

The rest of this paper is organized as follows. In Section 2, we develop the definition of learning. We discuss the problem of complexity selection in Section 3 and briefly describe some common solutions to this

*The research reported here has been partially supported by Contract No. AFSOR-89-0536, USARO Contract No. DAAL-03-91-G-0182, and JSEP Contract No. N00014-90-J-1270.

problem and their drawbacks. In Section 4, we present our learning procedure and show that it does learn, under mild assumptions.

2 Learning

We first make precise the notion of “learning.” Let X and Y be point sets. For simplicity, we will take $Y \subset \mathbb{R}^q$. Let $\mathcal{P}$ be a known class of probability measures on $S = X \times Y$. Let $P \in \mathcal{P}$ be a fixed, unknown probability measure on S . Information is gathered about this fixed, unknown probability measure $P \in \mathcal{P}$ by drawing points independently from S according to P . We call such a vector of n points from S a *labeled sample* of length n ,

$$\tilde{s}(n) = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)].$$

We write $\tilde{x}(n)$ for the *sample* $[x_1, x_2, \dots, x_n]$, $\tilde{y}(n)$ for the *labels* $[y_1, y_2, \dots, y_n]$, and (with some abuse of notation) $\tilde{s} = (\tilde{x}, \tilde{y})$. Further, we denote the probability associated with the sampling process, the n -fold product of P , by P^n .

Given $\tilde{s}(n)$, the goal is to “learn” the relationship between x and y when (x, y) is drawn according to P . We formulate this as a problem of selecting a model or *hypothesis* $h : X \rightarrow \mathbb{R}^q$ such that $h(x)$ is, on average, close to y when (x, y) is drawn by P . When the label of x is y , we take $\|h(x) - y\|$ to represent the loss, or cost, of hypothesizing $h(x)$ as the label. (Here, $\|\cdot\|$ is a known norm¹ on $\mathbb{R}^q$.) The *error* of h with respect to P is then the expected value of $\|h(x) - y\|$ for another labeled sample point drawn by P :

$$\text{err}(P, h) := E_P \|h(x) - y\| = \int_{X \times Y} \|h(x) - y\| dP(x, y).$$

This quantity is sometimes referred to as the *generalization error*.

We now restate what learning entails: given $\mathcal{P}$ and $\tilde{s}(n)$ drawn by an unknown $P \in \mathcal{P}$, find a hypothesis h such that $\text{err}(P, h)$ is small. There are two additional points to consider. First, since the information about P is obtained only through a sampling process, we cannot *guarantee* to find a good hypothesis; rather, we settle for finding, with P^n -probability near 1, a good hypothesis. (Note that our ability to find a good hypothesis depends on the particular labeled sample $\tilde{s}(n)$ that has been drawn. Thus P^n , and not P , is the appropriate probability measure to use.) Second, we would normally constrain the hypothesis to be from some known set $\mathcal{H}$, e.g., spline functions or feedforward neural networks. Thus, we seek a hypothesis in $\mathcal{H}$ whose error is close to the “optimum” over $\mathcal{H}$,

$$\text{opt}(P, \mathcal{H}) := \inf_{h \in \mathcal{H}} \text{err}(P, h).$$

¹A norm maps $\mathbb{R}^q$ to $[0, \infty)$ and satisfies (i) $\|v\| = 0$ iff $v = 0$; (ii) $\|\alpha v\| = |\alpha| \|v\|$ for all $\alpha \in \mathbb{R}$ and $v \in \mathbb{R}^q$; and (iii) $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in \mathbb{R}^q$ (the *triangle inequality*).

A pair $(\mathcal{P}, \mathcal{H})$ thus specifies a learning problem for us. To solve this problem, we must find a mapping from labeled samples to hypotheses (call it “ g ”) with the property that the error of $g[\tilde{s}(n)]$ approaches the optimum as the number of samples, n , increases. We do not allow a candidate learning function g to depend on P . However, the set $\mathcal{P}$ models prior knowledge about the underlying distribution, and the choice of g may reflect this.

We summarize the foregoing discussion with the following definition.

Definition 2.1. $(\mathcal{P}, \mathcal{H})$ is (nonuniformly) *learnable* if there is a mapping² $g : \tilde{s} \mapsto \mathcal{H}$ such that, for each $P \in \mathcal{P}$,

$$|\text{err}(P, g[\tilde{s}(n)]) - \text{opt}(P, \mathcal{H})| \rightarrow 0 \quad (2.1)$$

in probability as $n \uparrow \infty$, i.e., for any given $\epsilon > 0$,

$$P^n[\tilde{s}(n) : |\text{err}(P, g[\tilde{s}(n)]) - \text{opt}(P, \mathcal{H})| > \epsilon] \rightarrow 0$$

as $n \uparrow \infty$ for each $P \in \mathcal{P}$.

We use “nonuniform” to distinguish this from an alternate definition for which the convergence in (2.1) is uniform over $\mathcal{P}$. That is, *uniform* learnability would require that we achieve $|\text{err}(P, g[\tilde{s}(n)]) - \text{opt}(P, \mathcal{H})| < \epsilon$ with a certain probability by using a number of samples that is the same regardless of P . For the sort of $(\mathcal{P}, \mathcal{H})$ that we are interested in, uniform learning will frequently be impossible. However, we will present a learning procedure that is uniform over distributions that have the same “complexity.”

3 The Problem of Complexity Selection

Recall that our goal is to find a hypothesis, h , such that $\text{err}(P, h)$ is small. In the preceding section we restated this in terms of the problem of learning $(\mathcal{P}, \mathcal{H})$, yet we did not indicate how to choose $\mathcal{H}$. We can reduce the best attainable value of $\text{err}(P, h)$ (i.e., $\text{opt}(P, \mathcal{H})$) by including many functions in $\mathcal{H}$; that is, we may use a “rich” $\mathcal{H}$. For instance, we might let $\mathcal{H}$ be the set of all polynomials when $Y = \mathbb{R}$. With such a rich $\mathcal{H}$, given a finite amount of (noisy) data it is difficult to resist the temptation to overfit the data. We can always find an n ’th order polynomial that passes through n data points, but we would not expect such a hypothesis to have nearly the least possible generalization error. Rather, we should generally choose a less “complex” hypothesis, a polynomial of order less than n . The following question arises: How can we select the appropriate complexity for the amount of data we have? In this section, we shall describe some answers to this question from the literature that rely heavily

²We do not address the issues of computational complexity or measurability in this paper.

on a prior measure of a hypothesis' complexity. In the next section, we shall present an answer that depends solely on the data.

In this paper, we will concentrate on sets of hypotheses that are "nested." When $\mathcal{H}$ is such that

$$\mathcal{H} = \bigcup_{k=1}^{\infty} \mathcal{H}^k \text{ and } \mathcal{H}^1 \subseteq \mathcal{H}^2 \subseteq \dots,$$

we say that $\mathcal{H}$ is *nested*. Also, we say that $h \in \mathcal{H}^k$ is *simpler* than $h' \in \mathcal{H}$ if $h' \notin \mathcal{H}^k$. For example, $\mathcal{H}^k$ might consist of k 'th order multinomials. A nested structure allows us to compare the complexity of two hypotheses without reference to an absolute measure of complexity.

Consider the *empirical estimate* of $\text{err}(P, h)$ based on the labeled sample $\tilde{s}(n)$:

$$f_{\text{emp}}[\tilde{s}(n), h] := \frac{1}{n} \sum_{i=1}^n \|h(x_i) - y_i\|. \quad (3.1)$$

A common learning technique is to find a hypothesis from $\mathcal{H}$ that minimizes f_{emp} . Many researchers have investigated conditions under which this *empirical error minimization* results in learning (see [5], [1], [6], and [2]). In particular, the following is a sufficient condition for learning.

Definition 3.1. f_{emp} converges simultaneously over $(\mathcal{P}, \mathcal{H})$ if for any given $\epsilon > 0$

$$\sup_{P \in \mathcal{P}} P^n \left[\tilde{s}(n) : \sup_{h \in \mathcal{H}} |f_{\text{emp}}[\tilde{s}(n), h] - \text{err}(P, h)| > \epsilon \right] \rightarrow 0$$

as $n \uparrow \infty$; that is, $f_{\text{emp}}[\tilde{s}(n), h]$ converges to $\text{err}(P, h)$ in probability simultaneously over $\mathcal{H}$ and uniformly in $P \in \mathcal{P}$.

It is easy to see why simultaneous convergence of f_{emp} serves as a sufficient condition for empirical error minimization to result in learning. If $f_{\text{emp}}[\tilde{s}, h]$ is close to the actual error, $\text{err}(P, h)$, for all $h \in \mathcal{H}$, then any h^* that minimizes $f_{\text{emp}}[\tilde{s}, h]$ must have an error close to $\inf_{h \in \mathcal{H}} \text{err}(P, h)$.

In their pioneering work ([5] and [7]), Vapnik and Chervonenkis developed a necessary and sufficient condition for f_{emp} to converge simultaneously over $(\mathcal{P}, \mathcal{H})$. In the particular case where the hypotheses and labels are 0/1-valued and $\mathcal{P}$ is $\mathcal{P}^*$, the set of all probability measures on S , this convergence is characterized by a measure of the complexity of $\mathcal{H}$ known as the Vapnik-Chervonenkis dimension.

We now define the Vapnik-Chervonenkis dimension.

Let $\mathcal{B}$ be a class of 0/1-valued functions on X , and let $\tilde{x}(n)$ be an n -vector of points from X . We use $\Delta^{\mathcal{B}}(\tilde{x})$ to denote the number of distinct labelings of $\tilde{x}$ by functions from $\mathcal{B}$: $\Delta^{\mathcal{B}}(\tilde{x}) := |\{b(\tilde{x}) : b \in \mathcal{B}\}|$. Clearly, $\Delta^{\mathcal{B}}(\tilde{x}(n)) \leq 2^n$. We say that $\mathcal{B}$ *shatters* $\tilde{x}(n)$ if $\Delta^{\mathcal{B}}(\tilde{x}(n)) = 2^n$.

Definition 3.2. The Vapnik-Chervonenkis dimension of $\mathcal{B}$, $\text{VCdim}(\mathcal{B})$, is the largest n for which there is some $\tilde{x}(n) \in X^n$ shattered by $\mathcal{B}$. (If, for arbitrarily large n , there are $\tilde{x}(n)$ that are shattered by $\mathcal{B}$, we say that $\text{VCdim}(\mathcal{B}) = \infty$.)

Theorem 3.1. [Vapnik and Chervonenkis]

Let $\mathcal{P}^*$ be the set of all probability measures on S . The empirical error estimate f_{emp} converges simultaneously over $(\mathcal{P}^*, \mathcal{H})$ if and only if $\text{VCdim}(\mathcal{H}) < \infty$.

One can also define analogous notions of dimension when the hypotheses and labels are $\mathbb{R}^q$ -valued; see [1] and [2] for definitions. These dimensions can be used to give conditions that are sufficient, but not in general necessary, for f_{emp} to converge simultaneously over $(\mathcal{P}^*, \mathcal{H})$.

For many rich, interesting sets $\mathcal{H}$, f_{emp} does not converge simultaneously over $(\mathcal{P}, \mathcal{H})$, but it does so for each $(\mathcal{P}, \mathcal{H}^k)$, $k = 1, 2, \dots$. A useful feature of the VC dimension is that it can be used to give bounds on the number of labeled samples needed for

$$\sup_{h \in \mathcal{H}^k} |f_{\text{emp}}[\tilde{s}, h] - \text{err}(P, h)| < \epsilon$$

to occur with probability more than $1 - \delta$, in terms of $\text{VCdim}(\mathcal{H}^k)$. Because of this, the VC dimension can be used as a measure of the complexity of the class $\mathcal{H}^k$, since it determines the number of samples needed to learn from this class.

We now describe two of the prevalent strategies in the literature for selecting the complexity of hypotheses. In the first, a measure of the complexity of $\mathcal{H}^k$ is used to determine a function $k(n)$ such that minimization of $f_{\text{emp}}[\tilde{s}(n), h]$ over $\mathcal{H}^{k(n)}$ based on n labeled samples will result in successful learning. These methods constrain the complexity of the class of candidate hypotheses in terms of the amount of available data; see [8], [6], and [9] for examples. In the second, for each k a hypothesis $h_k^* \in \mathcal{H}^k$ is found by minimizing $f_{\text{emp}}[\tilde{s}(n), h]$ over $\mathcal{H}^k$. Then, a penalty function $\text{pen}(n, k)$ (which increases with k) is added to $f_{\text{emp}}[\tilde{s}(n), h_k^*]$ and the resulting quantity is minimized over k to select the final hypothesis. The underlying intuition is that the penalty will tend to exclude overly complex hypotheses from consideration. See [10], [11], [1], [12], and [13] for examples.

The methods described in the preceding paragraph have two main drawbacks. First, in order to determine values for $k(n)$ or $\text{pen}(n, k)$, a great deal of analysis must be performed on the sets $\mathcal{H}^k$ (e.g., to determine $\text{VCdim}(\mathcal{H}^k)$). Second, these methods involve *a priori* measures of the complexity of $\mathcal{H}^k$. The learning procedure we present in the next section avoids these problems by using only the nested structure of $\mathcal{H}$ to compare the complexity of hypotheses and by making this comparison based on the observed distribution of the data.

4 Learning with a Simple Empirical Cover

In this section, we describe our learning procedure. It involves a “simple empirical cover,” an approximation of the set $\mathcal{H}$ by a set of simple hypotheses that is based on a portion of the samples. The particular form of this learning procedure is suggested by a canonical form for an error estimator that is more sophisticated than f_{emp} . Due to space constraints, we do not describe these results here; we refer the reader to [14], [15], and [16].

First, we list a few preliminary conventions and results.

Assumption 4.1. *The hypotheses and the labels are bounded: $\|h(x)\| < B$ and $\|h(x) - y\| \leq \lambda$ for all $h \in \mathcal{H}$, $x \in X$, and $y \in Y$.*

A simple argument based on Hoeffding’s inequality ([17]) shows that the following lemma holds (see [2] for details).

Lemma 4.1. *Under Assumption 4.1, if $\mathcal{A}$ is a finite subset of $\mathcal{H}$, then for any $\ell \geq \lambda^2 \ln[2|\mathcal{A}|/\nu]/(2\mu^2)$ and any $P \in \mathcal{P}$,*

$$P' \left[\tilde{s}(\ell) : \max_{a \in \mathcal{A}} |f_{\text{emp}}[\tilde{s}, a] - \text{err}(P, a)| > \mu \right] < \nu.$$

Thus, with probability at least $1 - \nu$, $f_{\text{emp}}[\tilde{s}, a]$ will be within μ of $\text{err}(P, a)$ for every $a \in \mathcal{A}$ when we take ℓ or more samples.

For any set of hypotheses $\mathcal{H}$, we define $\mathcal{H}^\Delta$ to be the following set of functions:

$$\mathcal{H}^\Delta := \{ \|h_1 - h_2\| : h_1, h_2 \in \mathcal{H} \}.$$

We use $\mu_n \searrow 0$ to mean that μ_n is a sequence of non-increasing, positive numbers that converges to zero. Also, let ψ be a function that maps X to $\mathbb{R}$. We use $E_{\mathcal{F}(n)}\psi$ to denote the empirical estimate of $E_P\psi$ based on $\mathcal{F}(n)$:

$$E_{\mathcal{F}(n)}\psi = \frac{1}{n} \sum_{i=1}^n \psi(x_i).$$

Definition 4.1. *We say that the empirical estimates converge simultaneously over $(\mathcal{P}, \mathcal{H}^\Delta)$ if*

$$\sup_{\psi \in \mathcal{H}^\Delta} |E_P\psi - E_{\mathcal{F}(n)}\psi| \rightarrow 0$$

in probability as $n \uparrow \infty$ uniformly in $P \in \mathcal{P}$.

4.1 The case of 0/1-valued hypotheses

In the case where the labels and hypotheses are 0/1-valued, the learning procedure, which we denote by g_{sec} , has a very simple form. In this case, we measure the error of h by $E_P|h(x) - y|$. We say that h is a *simplest* hypothesis in a set $\mathcal{G}$ if $h \in \mathcal{H}^k$ and no $h' \in \mathcal{G}$ belongs to $\mathcal{H}^{k-1}$.

Procedure 4.1. *A learning scheme for 0/1-valued hypotheses.*

Let $\mu_n \searrow 0$ and $\nu_n \searrow 0$. For a labeled sample $\tilde{s}(m)$ of m points, perform the following three steps.

1. Let n be the largest integer such that

$$n + \lceil \ln[2^{n+1}/\nu_n]/(2\mu_n^2) \rceil \leq m. \quad (4.1)$$

Split $\tilde{s}$ into two subsamples,

$$\tilde{s}(m) = (\tilde{s}'(n), \tilde{s}''(m - n)),$$

where $\tilde{s}'(n) = (\tilde{x}', \tilde{y}')$.

2. For each $i \in \{0, \dots, 2^n - 1\}$, let $\tilde{b}_i$ be the n -bit binary representation of i . Let $h_{\tilde{x}', i}$ be a simplest hypothesis satisfying $h_{\tilde{x}', i}(\tilde{x}') = \tilde{b}_i$. (There may be no such hypothesis; in this case we say “ $\tilde{b}_i$ is not feasible.”) Let

$$\mathcal{H}_n(\tilde{x}') = \{h_{\tilde{x}', i} : 0 \leq i \leq 2^n - 1 \text{ and } \tilde{b}_i \text{ is feasible}\}.$$

3. Use $\tilde{s}''$ to estimate $\text{err}(P, h_{\tilde{x}', i})$ empirically, and let g_{sec} minimize this over $\mathcal{H}_n(\tilde{x}')$:

$$g_{\text{sec}}[\tilde{s}] \in \left\{ h : f_{\text{emp}}[\tilde{s}'', h] = \min_{h' \in \mathcal{H}_n(\tilde{x}')} f_{\text{emp}}[\tilde{s}'', h'] \right\}.$$

In Step 1, it suffices to pick n such that n goes to infinity as m does and n satisfies (4.1) for large m . For example, if $\mu_n = 1/\sqrt{n}$ and $\nu_n = e^{1-n}$, we can use $n = \sqrt{m}$.

The set $\mathcal{H}_n(\tilde{x}')$ can be thought of as an approximation of $\mathcal{H}$. It “covers” $\mathcal{H}$ in the sense that, for any $h \in \mathcal{H}$, there is an $h' \in \mathcal{H}_n(\tilde{x}')$ with $h'(x'_i) = h(x'_i)$, $i = 1, \dots, n$.

Unlike empirical error minimization (see Section 3), g_{sec} first narrows the candidate hypotheses down to $\mathcal{H}_n(\tilde{x}')$, a finite set containing at most 2^n members. It then uses the rest of the data to pick one of the best candidates (the choice of n in Step 1 stems from Lemma 4.1 with $\lambda = 1$ and $|\mathcal{A}| = 2^n$.) Thus, g_{sec} will not overfit the data. One question remains: Does $\mathcal{H}_n(\tilde{x}')$ contain a hypothesis with small error? Under rather mild assumptions on $(\mathcal{P}, \mathcal{H})$, it does.

As follows, we can show that g_{sec} learns $(\mathcal{P}^*, \mathcal{H})$ when each $\mathcal{H}^k$ has finite VC dimension. $\mathcal{P}^*$ is the set of all probability measures on S .

Take any $P \in \mathcal{P}^*$, and fix $\epsilon > 0$. Because

$$\text{opt}(P, \mathcal{H}) = \inf_{h \in \mathcal{H}} \text{err}(P, h),$$

for some $k_\epsilon < \infty$ there is $h_\epsilon \in \mathcal{H}^{k_\epsilon}$ with

$$\text{err}(P, h_\epsilon) - \text{opt}(P, \mathcal{H}) < \epsilon. \quad (4.2)$$

By Step 2, for any $\tilde{x}'(n)$, there is some $h_{\tilde{x}'(n)}^* \in \mathcal{H}_n(\tilde{x}')$ such that

$$h_{\tilde{x}'(n)}^*(\tilde{x}'(n)) = h_\epsilon(\tilde{x}'(n)). \quad (4.3)$$

By construction, $h_{\tilde{x}'(n)}^*$ is the simplest such hypothesis, so $h_{\tilde{x}'(n)}^* \in \mathcal{H}^{k_\epsilon}$. Since $\text{VCdim}(\mathcal{H}^{k_\epsilon}) < \infty$, it is

straightforward to show that $\text{VCdim}((\mathcal{H}^k)^\Delta)$ is also finite. By a theorem analogous to Theorem 3.1 (see [1]), the empirical estimates converge simultaneously over $(\mathcal{P}^*, (\mathcal{H}^k)^\Delta)$. Thus,

$$|E_P[h_{\mathcal{H}'(n)}^* - h_\epsilon] - E_{\mathcal{H}'(n)}[h_{\mathcal{H}'(n)}^* - h_\epsilon]| \rightarrow 0 \quad (4.4)$$

in probability as $n \uparrow \infty$. By (4.3), $E_{\mathcal{H}'(n)}[h_{\mathcal{H}'(n)}^* - h_\epsilon] = 0$, and thus (4.4) implies that

$$E_P[h_{\mathcal{H}'(n)}^* - h_\epsilon] \rightarrow 0$$

in probability. Therefore,

$$|\text{err}(P, h_{\mathcal{H}'(n)}^*) - \text{err}(P, h_\epsilon)| \rightarrow 0 \quad (4.5)$$

in probability as well. By (4.2) and (4.5), for large n ,

$$\text{err}(P, h_{\mathcal{H}'(n)}^*) - \text{opt}(P, \mathcal{H}) < 2\epsilon \quad (4.6)$$

with probability approaching 1. That is,

$$P^n[\tilde{s}'(n) : \text{err}(P, h_{\mathcal{H}'(n)}^*) - \text{opt}(P, \mathcal{H}) < 2\epsilon] \rightarrow 1$$

as $n \uparrow \infty$. Thus, $h_{\mathcal{H}'(n)}^*$ is a good choice for a hypothesis.

By Lemma 4.1, $m - n$ is large enough that g_{sec} usually picks either $h_{\mathcal{H}'(n)}^*$ or a better hypothesis. In fact, for large m ,

$$\begin{aligned} \text{err}(P, g_{\text{sec}}[\tilde{s}(m)]) &\leq \min_i \text{err}(P, h_{\mathcal{H}',i}) + \epsilon \\ &\leq \text{err}(P, h_{\mathcal{H}'(n)}^*) + \epsilon \end{aligned} \quad (4.7)$$

with probability approaching 1. Hence, for large m ,

$$\text{err}(P, g_{\text{sec}}[\tilde{s}'(n), \tilde{s}''(m-n)]) - \text{opt}(P, \mathcal{H}) < 3\epsilon$$

with probability approaching 1. Since ϵ is arbitrarily small,

$$\text{err}(P, g_{\text{sec}}[\tilde{s}(m)]) \rightarrow \text{opt}(P, \mathcal{H})$$

in probability as $m \uparrow \infty$ for each $P \in \mathcal{P}^*$. That is, g_{sec} nonuniformly learns $(\mathcal{P}^*, \mathcal{H})$.

We now make some observations about Procedure 4.1 as a preparation for the general case in the next section. We can weaken many of the assumptions we made above. For one, the hypotheses in $\mathcal{H}_n(\tilde{x}')$ need not be the simplest: it would suffice to have $h_{\mathcal{H}'(n),i} \in \mathcal{H}^{[Mk]}$, for some $M < \infty$, where k_i is the smallest integer such that there is $h \in \mathcal{H}^{k_i}$ with $h(\tilde{x}') = \tilde{b}_i$. This comment is useful because there are some cases for which finding the *simplest* hypothesis consistent with a labeling is much harder than finding one that is only reasonably simple (see [18]). Also, it suffices for $\mathcal{H}$ to be such that the empirical estimates converge simultaneously over $(\mathcal{P}^*, (\mathcal{H}^k)^\Delta)$ for each k , rather than to have $\text{VCdim}(\mathcal{H}^k) < \infty$ for each k .

4.2 The general case

In this section, we extend Procedure 4.1 to a much more general case. Also, we show that this learning procedure accomplishes more than nonuniform learning; it "learns" uniformly over distributions that have the same "complexity."

As in the 0/1-valued case, the learning procedure we give here will utilize an approximation of $\mathcal{H}$ whose members are "simple." In order to construct this approximation, we need to define a notion of distance between hypotheses.

Definition 4.2. The empirical distance between h and h' based on $\tilde{x}(n)$ is

$$\rho_{\tilde{x}(n)}(h, h') = \frac{1}{n} \sum_{i=1}^n \|h(x_i) - h'(x_i)\|.$$

Our aim is to "empirically cover" $\mathcal{H}$ with a set $\mathcal{H}_n(\tilde{x})$ in the sense that any $h \in \mathcal{H}$ is near a member of $\mathcal{H}_n(\tilde{x})$ when distance is measured by $\rho_{\tilde{x}}$. As in the remarks following Procedure 4.1, since finding a *simplest* hypothesis as in Step 2 of Procedure 4.1 may be very difficult, we will require the members of the approximating set to be only *nearly simple*. Hence, we make the following definition.

Definition 4.3. An M -simple empirical ϵ -cover for $\mathcal{H}$ based on $\tilde{x}(n)$ is a set $\mathcal{H}_n(\tilde{x}) \subseteq \mathcal{H}$ for which there is a mapping $a_{n,\tilde{x}} : \mathcal{H} \rightarrow \mathcal{H}_n(\tilde{x})$ such that, for each integer j and any $h \in \mathcal{H}^j$, both $\rho_{\tilde{x}}(a_{n,\tilde{x}}(h), h) < \epsilon$ and $a_{n,\tilde{x}}(h) \in \mathcal{H}^{[Mj]}$ are satisfied.

A key observation is that we can easily construct finite M -simple empirical coverings. In the following lemma, let $\|\cdot\|$ be the norm that, at a point $v = (v_1, v_2, \dots, v_q)$ in $\mathbb{R}^q$, takes the value

$$\|v\| = \sum_{i=1}^q |v_i|.$$

An analogous result holds for other norms.

Lemma 4.2. Under Assumption 4.1, for any $M \geq 1$ we can construct an M -simple empirical ϵ -cover for $\mathcal{H}$ based on $\tilde{x}(n)$ that has at most $\lceil 2Bq/\epsilon \rceil^{nq}$ elements.

Proof: See the Appendix for the required construction.

In the following procedure, let Assumption 4.1 be satisfied.

Procedure 4.2. Learning with an M -simple empirical cover.

Let $1 \leq M < \infty$, $\tau_n \searrow 0$, $\mu_n \searrow 0$, and $\nu_n \searrow 0$. For a labeled sample $\tilde{s}(m)$ of m points, perform the following three steps.

1. Let $H(n) = \lceil 2Bq/\tau_n \rceil^{nq}$. Let n be the largest integer such that

$$n + \lceil (\lambda^2/2\mu_n^2) \ln[2H(n)/\nu_n] \rceil \leq m.$$

Split $\tilde{s}$ into two subsamples,

$$\tilde{s}(m) = (\tilde{s}'(n), \tilde{s}''(m-n)),$$

where $\tilde{s}'(n) = (\tilde{x}', \tilde{y}')$.

2. Construct, as in the proof of Lemma 4.2, an M -simple empirical (τ_n) -cover for $\mathcal{H}$ based on $\tilde{x}'$, $\mathcal{H}_n(\tilde{x}')$.
3. Use $\tilde{s}''$ to estimate $\text{err}(P, h_{\tilde{x}', i})$ empirically, and let g_{sec} minimize this over $\mathcal{H}_n(\tilde{x}')$:

$$g_{\text{sec}}[\tilde{s}] \in \left\{ h : f_{\text{emp}}[\tilde{s}'', h] = \min_{h' \in \mathcal{H}_n(\tilde{x}')} f_{\text{emp}}[\tilde{s}'', h'] \right\}.$$

As in the 0/1-valued case, this learning procedure resists overfitting the data and, under mild conditions on $(\mathcal{P}, \mathcal{H})$, it picks a hypothesis with small error.

Theorem 4.1. *Under Assumption 4.1, if the empirical estimates converge simultaneously over $(\mathcal{P}, (\mathcal{H}^i)^\Delta)$ for each i , then g_{sec} nonuniformly learns $(\mathcal{P}, \mathcal{H})$.*

Proof: See the Appendix.

In fact, the proof of Theorem 4.1 reveals a stronger property. Consider the sets

$$\mathcal{P}_i = \left\{ P \in \mathcal{P} : \inf_{h \in \mathcal{H}^i} \text{err}(P, h) = \text{opt}(P, \mathcal{H}) \right\}.$$

We can think of each distribution in $\mathcal{P}_i$ as having the same “complexity” relative to the nested structure of $\mathcal{H}$. This complexity increases with i , and it seems reasonable to allow more labeled samples in order to learn in the face of more complex distributions. It turns out that, even though g_{sec} learns nonuniformly over $(\mathcal{P}, \mathcal{H})$, it learns *uniformly* over each $\mathcal{P}_i$. That is, the number of samples it requires to produce a hypothesis of a given quality is the same for distributions with the same complexity.

The utility of Theorem 4.1 stems from two facts. First, it applies in many cases of interest. Second, Procedure 4.2 does not utilize any special knowledge about $\mathcal{H}$ other than its decomposition into nested subsets. For instance, consider the case where the hypotheses and labels are real-valued and $\|\cdot\|$ is the absolute value. It can be shown (see [16]) that when $\text{VCdim}(\mathcal{H}^i)$ is finite for each i (as is the case for many parametric model classes), the empirical estimates converge simultaneously over $(\mathcal{P}^*, (\mathcal{H}^i)^\Delta)$ for each i . Yet, in this case, we do not need to know the *value* of these dimensions to apply Procedure 4.2.

5 Concluding Remarks

We have presented a technique that selects a hypothesis, or model, of appropriate complexity, avoiding overfitting the data. This technique has the advantage that it does not rely on *a priori* measures of each hypothesis’ complexity. Although the results in this paper were all presented in terms of $\mathbb{R}^q$ -valued hypotheses, we remark that analogous results hold in a much more general framework (see [16] for details).

Appendix

A.1 Proof of Lemma 4.2

For each fixed $\epsilon > 0$, $M < \infty$, n , and $\tilde{x}(n) = [x_1, x_2, \dots, x_n]$, we shall construct both an M -simple empirical ϵ -cover for $\mathcal{H}$ based on $\tilde{x}(n)$, $\mathcal{H}_n(\tilde{x})$, and the associated mapping $a_{n, \tilde{x}}$ from hypotheses to $\mathcal{H}_n(\tilde{x})$. We shall treat only the case where the hypotheses and labels are real valued and the norm used in computing $\text{err}(P, h)$ is the absolute value. The more general case is a straightforward extension of this.

By Assumption 4.1, $|h(x)| < B$ for every $h \in \mathcal{H}$ and $x \in X$. Consider slicing the interval $(-B, B)$ into $\lceil 2B/\epsilon \rceil$ subintervals, each of length less than ϵ . There are $\lceil 2B/\epsilon \rceil^n$ possible mappings, M_k , from $i \in \{1, \dots, n\}$ to subintervals: each $M_k(i)$ is a subinterval. Initially, let $\mathcal{A}_0$ be the empty set. Construct $\mathcal{H}_n(\tilde{x})$ by iterating on $k \in \{1, \dots, \lceil 2B/\epsilon \rceil^n\}$ as follows. Let us refer to j as the *index* of h if $h \in \mathcal{H}^j$ and $h \notin \mathcal{H}^{j-1}$. For each k , if there is no $h \in \mathcal{H}$ such that

$$h(x_i) \in M_k(i) \text{ at each } i \in \{1, \dots, n\}, \quad (\text{A.1})$$

then set $\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{h_k\}$. Otherwise let j^* be the index of a simplest hypothesis that passes through $M_k(i)$ at each i : j^* is the minimum of

$$\{j : \exists h \in \mathcal{H}^j \text{ with } h(x_i) \in M_k(i), i \in \{1, \dots, n\}\}.$$

Then let $\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{h_k\}$, where h_k is a member of $\mathcal{H}^{[Mj^*]}$ that passes through $M_k(i)$ at each i . Finally, let $\mathcal{H}_n(\tilde{x}) = \mathcal{A}_{\lceil 2B/\epsilon \rceil^n}$. Clearly, $|\mathcal{H}_n(\tilde{x})| \leq \lceil 2B/\epsilon \rceil^n$.

Now, take any $h \in \mathcal{H}$. This hypothesis must pass through some subinterval at each x_i . That is, there is some k' such that

$$h(x_i) \in M_{k'}(i), i \in \{1, \dots, n\}.$$

Let $h_{k'}$ be the hypothesis added to $\mathcal{H}^{k'-1}$ due to $M_{k'}$, and set $a_{n, \tilde{x}}(h) = h_{k'}$.

It is straightforward to show that $\rho_{\tilde{x}}(a_{n, \tilde{x}}(h), h) < \epsilon$ and, if $h \in \mathcal{H}^j$, $a_{n, \tilde{x}}(h) \in \mathcal{H}^{[Mj]}$. $\square$

A.2 Proof of Theorem 4.1

Let

$$\mathcal{P}_i = \left\{ P \in \mathcal{P} : \inf_{h \in \mathcal{H}^i} \text{err}(P, h) = \text{opt}(P, \mathcal{H}) \right\}.$$

Fix any $\epsilon > 0$. Take any $P \in \mathcal{P}$. Note that $P \in \mathcal{P}_j$ for some j . Thus, there is $h^* \in \mathcal{H}^j$ with

$$\text{err}(P, h^*) - \text{opt}(P, \mathcal{H}) \leq \epsilon. \quad (\text{A.2})$$

By Step 2 of Procedure 4.2, $a_{n, \tilde{x}'}(h^*) \in \mathcal{H}^{[Mj]}$. Thus, since the empirical estimates converge simultaneously on $(\mathcal{P}, (\mathcal{H}^i)^\Delta)$ for any i , if n is large enough then $\tau_n < \epsilon$ and

$$E_P \|a_{n, \tilde{x}'}(h^*) - h^*\| < \rho_{\tilde{x}}(a_{n, \tilde{x}'}(h^*), h^*) + \epsilon < 2\epsilon \quad (\text{A.3})$$

with probability near one. That is, for any given $\delta > 0$, there is $N(\delta)$ such that

$$P^n[\tilde{s}'(n) : E_P \|a_{n, \tilde{x}'}(h^*) - h^*\| < 2\epsilon] > 1 - \delta$$

for any $n \geq N(\delta)$. (Note that $N(\delta)$ does not depend on P , as long as $P \in \mathcal{P}_j$.)

By the triangle inequality for the norm $\|\cdot\|$, whenever $\tilde{s}' = (\tilde{x}', \tilde{y}')$ is such that (A.3) holds,

$$|\text{err}(P, a_{n, \tilde{s}'}(h^*)) - \text{err}(P, h^*)| \leq 2\epsilon,$$

and hence, by (A.2),

$$\text{err}(P, a_{n, \tilde{s}'}(h^*)) - \text{opt}(P, \mathcal{H}) \leq \epsilon + 2\epsilon. \quad (\text{A.4})$$

By Lemma 4.1 and Step 1 of Procedure 4.2, $m - n$ is large enough that, with probability at least $1 - \nu_n$,

$$\begin{aligned} \text{err}(P, g_{\text{sec}}[\tilde{s}(m)]) &\leq \min_{h \in \mathcal{H}_n(\mathcal{P})} \text{err}(P, h) + \mu_n \\ &\leq \text{err}(P, a_{n, \tilde{s}'}(h^*)) + \mu_n. \end{aligned} \quad (\text{A.5})$$

Thus, for $\tilde{s}'$ such that (A.3) holds, we have by (A.4) and (A.5) that

$$\text{err}(P, g_{\text{sec}}[\tilde{s}', \tilde{s}''']) - \text{opt}(P, \mathcal{H}) \leq 3\epsilon + \mu_n \quad (\text{A.6})$$

with probability approaching one as $m \uparrow \infty$ (uniformly in $P \in \mathcal{P}_j$). That is,

$$\sup_{P \in \mathcal{P}_j} P^m[\tilde{s}(m) : \text{err}(P, g_{\text{sec}}[\tilde{s}]) - \text{opt}(P, \mathcal{H}) \leq 3\epsilon + \mu_n]$$

converges to 1. Since $\mu_n \searrow 0$ and ϵ is arbitrary, $\text{err}(P, g_{\text{sec}}[\tilde{s}]) \rightarrow \text{opt}(P, \mathcal{H})$ in probability as $m \uparrow \infty$, uniformly in $P \in \mathcal{P}_j$. Therefore, g_{sec} nonuniformly learns $(\mathcal{P}, \mathcal{H})$, and it does so uniformly over each $\mathcal{P}_j$. $\square$

References

- [1] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer Verlag, 1982.
- [2] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, vol. 100, pp. 78-150, 1992.
- [3] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, pp. 1134-1142, November 1984.
- [4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *Journal of the ACM*, vol. 36, no. 3, pp. 939-965, 1989.
- [5] V. N. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications*, vol. 16, no. 2, pp. 264-280, 1971.
- [6] V. N. Vapnik and A. Chervonenkis, "Minimization of expected risk based on empirical data," in *Proceedings of the First World Congress of the Bernoulli Society*, vol. 2, pp. 821-832, VNU Science Press, Utrecht, The Netherlands, 1987.
- [7] V. N. Vapnik and A. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Probability and Its Applications*, vol. 26, no. 3, pp. 532-553, 1981.
- [8] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.
- [9] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, no. 5, pp. 535-549, 1990.
- [10] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716-723, 1974.
- [11] V. N. Vapnik and A. Chervonenkis, "Ordered risk minimization, I," *Automation and Remote Control*, vol. 35, pp. 1226-1235, August 1974.
- [12] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416-431, 1983.
- [13] A. R. Barron, "Complexity regularization with applications to artificial neural networks," in *Proceedings of the NATO ASI on Nonparametric Functional Estimation*, Kluwer Academic Publishers, August 1990.
- [14] K. L. Buescher and P. R. Kumar, "Learning with respect to general loss functions by smooth simultaneous estimation," in *Proceedings of the 26th Annual Conference on Information Sciences and Systems*, (Princeton, New Jersey), 1992.
- [15] K. L. Buescher and P. R. Kumar, "Learning stochastic functions by smooth simultaneous estimation," in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 272-279, The Association for Computing Machinery, New York, NY, 1992.
- [16] K. L. Buescher, "Learning and smooth simultaneous estimation of errors based on empirical data," Technical Report UILU-ENG-92-2246 (Coordinated Science Laboratory Report DC-144), University of Illinois, Urbana, IL, November 1992. Ph.D. dissertation, Department of Electrical and Computer Engineering.
- [17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13-30, 1963.
- [18] D. Haussler, "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework," *Artificial Intelligence*, vol. 36, no. 2, pp. 117-222, 1988.

**DATE
FILMED**

3 / 9 / 94

END