Learning by Canonical Smooth Estimation, Part II: Learning and Choice of Model Complexity*†

Kevin L. Buescher‡ and P. R. Kumar§

Abstract

In this paper, we analyze the properties of a procedure for learning from examples. This "canonical learner" is based on a canonical error estimator developed in a companion paper.

In learning problems, we observe data that consists of labeled sample points, and the goal is to find a model, or "hypothesis," from a set of candidates that will accurately predict the labels of new sample points. The expected mismatch between a hypothesis' prediction and the actual label of a new sample point is called the hypothesis' "generalization error." We compare the canonical learner with the traditional technique of finding hypotheses that minimize the relative frequency-based empirical error estimate. We show that, for a broad class of learning problems, the set of cases for which such empirical error minimization works is a proper subset of the cases for which the canonical learner works. We derive bounds to show that the number of samples required by these two methods is comparable.

We also address the issue of how to determine the appropriate complexity for the class of candidate hypotheses. Many existing techniques solve this problem at the expense of requiring the evaluation of an absolute, a priori measure of each hypothesis' complexity. The method we present does not. It is based on an extension of the canonical learner and uses a natural, relative measure of each model's complexity. We show that this method will learn for a variety of common parametric hypothesis classes. Also, for a broad class of learning problems, we show that this method works whenever a certain conventional method for choosing model complexity does.

*Please address all correspondence to the first author.

†The research reported here has been supported in part by the U. S. Army Research Office under Contract No. DAAL-03-91-G-0182, the National Science Foundation under Grant No. ECS-92-16487, by AFOSR Contract No. AFOSR-89-0536, and by the Joint Services Electronics Program under Contract No. N00014-90-J1270.

‡Los Alamos National Laboratory, MS F645, Los Alamos, NM 87545. Email: klb@lanl.gov

§Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801. Email: prkumar@gauss.csl.uiuc.edu

1 Introduction

This paper develops and investigates the properties of a new class of learning procedures. These procedures are based on the canonical error estimation procedure developed in [1]. We also provide bounds on the number of samples required by a procedure. Additionally, we propose and analyze a method of selecting a hypothesis, or "model," with an appropriate degree of complexity.

Before we elaborate further we first review some of the terminology and results from [1]. As an example of a *learning problem*, imagine 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. More generally, we may observe labeled sample points (x_i, y_i) from $S = X \times Y$ drawn independently at random according to a fixed, unknown probability distribution P (P is assumed to lie in a known set of distributions, $\mathcal{P}$). The models that we can use, which we shall henceforth call *hypotheses*, are drawn from a class $\mathcal{H}$ of functions that map X to a set Z. In order to measure the performance of a hypothesis, we use a *loss function*, $L : Z \times Y \mapsto \mathbb{R}$. With each hypothesis $h \in \mathcal{H}$ we associate an *error* with respect to P, the "generalization error," $\mathrm{err}(P, h)$. This is the expected value of $L(h(x'), y')$ 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 $\mathcal{H}$.

A common approach to learning is to first find a means of estimating the error of each hypothesis in $\mathcal{H}$ *simultaneously* from the same labeled sample. Having solved this *simultaneous estimation* problem, one could then learn by simply choosing the hypothesis with the least *estimated* error, since this hypothesis will also have nearly the least *true* error. Inspired by the pioneering work of Vapnik and Chervonenkis ([2], [3], and [4]), much research has been done to determine when the empirical error estimate based on the labeled sample $\vec{s}(m) = (\vec{x}(m), \vec{y}(m))$,

$$f_{\mathrm{emp}}[\vec{s}(m), h] = \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i),$$

succeeds in simultaneously estimating the error (see [1] for references). In [1], we went beyond this empirical error estimation and examined simultaneous estimation problems that can be solved by estimators satisfying only a natural "smoothness" constraint. We developed a canonical estimator, f_{ce}, and showed that it is capable of smooth simultaneous estimation whenever such estimation is possible. In fact, for a broad class of problems the empirical estimator is smooth, so in these cases the canonical estimator, f_{ce}, will simultaneously estimate whenever f_{emp} does.

In this paper, we explore the properties of a *canonical learning procedure*, g_{cl}, which picks a hypothesis that minimizes f_{ce}. We compare g_{cl} to the common approach of minimizing the empirical error estimate, which we denote by g_{emp}. We show that, for a broad class of learning problems, g_{cl} learns whenever g_{emp} does. Further, we give an example where g_{cl} learns and g_{emp} does not.

We derive bounds on the "sample size" of g_{cl} and g_{emp}, the number of labeled sample points these learning procedures need to achieve a given level of performance. We show that, in general, the sample sizes of g_{cl} and g_{emp} are of the same order. Also, in some important special cases, we give bounds for the sample size of g_{cl} that are *smaller* than the available bounds for g_{emp}. Thus, the canonical learner does *not* gain its superiority over g_{emp} by virtue of using an excessive number of labeled samples.

We also address the issue of how to select $\mathcal{H}$. We must make $\mathcal{H}$ rich enough that it will contain a hypothesis with an acceptably low error level. A pitfall in learning with a rich $\mathcal{H}$ 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.

We focus on hypothesis classes that are *nested* in that $\mathcal{H}$ is the union of an increasing sequence of sets of hypotheses: $\mathcal{H}^1 \subseteq \mathcal{H}^2 \subseteq \ldots = \mathcal{H}$. 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 $\mathcal{H}^j$.

Many of the remedies for overfitting 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.

We show how the ideas underlying g_{cl} can be extended to provide a means of selecting a hypothesis with an appropriate level of complexity from a rich $\mathcal{H}$. This method 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. We give a condition that is sufficient for this procedure to learn and satisfied by many common parametric classes of hypotheses. Also, we show that this method will work whenever a certain conventional method of complexity selection works.

The remainder of this paper is organized as follows. In Section 2, we describe the canonical learner, g_{cl}, and show how it applies in a broader range of learning problems than the empirical error minimizer, g_{emp}. We show that the sample size bounds for g_{cl} compare favorably with those for g_{emp} in Section 3. We discuss the problem of selecting a hypothesis with an appropriate degree of complexity in Section 4 and describe some conventional approaches to this problem. We also present our approach to complexity selection, give conditions under which it works, and compare our method with one of the conventional approaches. In Section 5, we sum up the contributions of both this paper and [1] and indicate some directions for further research. Proofs of many of the results appear in the Appendix.

2 The Canonical Learner

Theorem 5.1 of [1] provides us with a new class of procedures that *learn* whenever $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable. When L is Lipschitz in h (see Definition 2.1 below), this is a weaker sufficient condition than $(\mathcal{P}, \mathcal{H})$ being simultaneously estimable by the empirical

estimator, f_{emp}. In this section, we prove a stronger result by comparing the ability of two procedures to learn pairs $(\mathcal{P}, \mathcal{H})$ that are not *necessarily* smoothly simultaneously estimable. We shall show that a particular *canonical learning procedure* based on the canonical estimator f_{ce} works whenever minimization of the empirical estimate does.

As in [1], in this paper we often assume that the loss criterion, L, is bounded and Lipschitz in h and that the range of the hypotheses, Z, is totally bounded. (In what follows, as in [1], d is the pseudometric on Z used to define smoothness for estimators.)

Assumption 2.1. *L is nonnegative and bounded by $\lambda < \infty$:*

$$0 \leq L(h(x), y) \leq \lambda \text{ for all } h \in \mathcal{H},\ x \in X, \text{ and } y \in Y.$$

Assumption 2.2. *Z is totally bounded under d (see [5]). That is, for each $\epsilon > 0$ there is a set of $K(\epsilon) < \infty$ points, $Z(\epsilon) \subseteq Z$, that form an ϵ-cover for Z with respect to d.*

Definition 2.1. *L is* Lipschitz *in h if there is $D < \infty$ such that*

$$|L(h(x), y) - L(h'(x), y)| \leq D\, d(h(x), h'(x)) \text{ for all } x \in X, y \in Y, \text{ and } h, h' \in \mathcal{H}.$$

As discussed in [1], we can base a learning procedure on f_{ce} by picking a hypothesis that minimizes the error estimate $f_{\text{ce}}[\vec{s}, h]$ over $\mathcal{H}$. However, note that the set of values that f_{ce} takes on the empirical cover, $\mathcal{H}_n(\vec{x}\,')$, is the same as those it takes on $\mathcal{H}$. Hence, we lose nothing by focusing on the class of learning procedures that minimize f_{ce} over the empirical cover instead of all of $\mathcal{H}$. The following procedure and theorem are direct consequences of Theorem 5.1 and Lemma 3.1 of [1].

Procedure 2.1. *A canonical learning procedure based on the canonical estimator.*

1. *Let $\epsilon_m \to 0$, $\epsilon_m \geq 0$.*
2. *Perform Procedure 5.2 of [1].*
3. *Let g_{cl} ϵ_m-approximately minimize $f_{\text{ce}}[\vec{s}(m), h]$ over $\mathcal{H}_n(\vec{x}\,')$: that is, choose $g_{\text{cl}}[\vec{s}(m)]$ as a member of*

$$\left\{ h \in \mathcal{H}_n(\vec{x}\,') : f_{\text{ce}}[\vec{s}(m), h] \leq \min_{h' \in \mathcal{H}_n(\vec{x}\,')} f_{\text{ce}}[\vec{s}(m), h'] + \epsilon_m \right\}.$$

(Note: In this and the learning procedures that follow, the choice of a minimizing hypothesis can be made definite by first establishing a priority ordering on $\mathcal{H}$.)

Theorem 2.1. *Under Assumptions 2.1 and 2.2,*

$(\mathcal{P}, \mathcal{H})$ is learnable by g_{cl} whenever $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable.

As with f_{ce}, g_{cl} represents a whole class of learning procedures determined by particular choices of the sequences ϵ_m, τ_n, μ_n, etc.

Before comparing g_{cl} to empirical error minimization, we first examine the similarities and differences between g_{cl} and some other procedures for learning in the literature. Also, we discuss how g_{cl} avoids overfitting the data.

Superficially, g_{cl} resembles the method of cross-validation (see [6]). However, the hypotheses selected by cross-validation are usually chosen by minimizing the empirical error on a subsample. The use of an empirical cover is clearly a different approach, since the labels of the first n points are not even used.

The learning procedure g_{cl} resembles the cover-based methods in [4], [7], [8], [9], and [10]. In these methods, knowledge of P or the structure of $\mathcal{P}$ is used to select a finite cover for $\mathcal{H}$ and empirical estimates are used to select the best element of the cover. Thus, there is an important difference: g_{cl} selects an empirical cover based on the *data*, whereas the covers in these other methods are fixed in advance.

Devroye examines a general structure for learning procedures in [11]. There, a class of candidate hypotheses is selected based on a "training set" ($\vec{s}\,'$ in our notation), and the hypothesis with the least empirical error on an independent "testing set" ($\vec{s}\,''$) is selected. Clearly, g_{cl} has this form. Here we provide concrete choices for the set of candidate hypotheses as well as a complete theoretical basis for these choices. We have provided estimates of the relative sizes of $\vec{s}\,'$ and $\vec{s}\,''$ as well as a characterization of a condition that is sufficient for this procedure to work.

In the 0/1-valued, noise-free case, a scheme akin to g_{cl} is used in [12] to transform a learning procedure for one triple $(\mathcal{P}, \mathcal{C}_1, \mathcal{C}_1)$ into a learning procedure for another triple

$(\mathcal{P}, \mathcal{C}_2, \mathcal{C}_2)$.

If $g : \vec{s} \mapsto \mathcal{H}$ picks a hypothesis that agrees with the data and yet g does not learn, we say that g *overfits* the data. There is a natural tendency to choose a hypothesis that agrees with the data. However, if the data is noisy and we select a hypothesis that fits the data (and hence the noise) too well, the generalization error of such a hypothesis will usually be larger than the optimum.

That one should resist the temptation to overfit the data is a well-known maxim. This principle is the basis of many estimation methods, such as Rissanen's Minimum Description Length Principle ([13]) and Vapnik's Principle of Structural Risk Minimization ([4]). These methods penalize a hypothesis' empirical error on the basis of the "complexity" of the class of hypotheses from which it is drawn (see Section 4 for more details). The learning procedure g_{cl} takes a different approach. It first constrains the class of candidate hypotheses on the basis of a portion of the data and then freely optimizes the empirical error over this class on the remainder of the data.

It may initially seem surprising that g_{cl} attempts to minimize $\mathrm{err}(P, h)$ over $\mathcal{H}$ by first approximating $\mathcal{H}$ with a finite set, $\mathcal{H}_n(\vec{x}\,')$. Such an approach may "underfit" the data, but it will not overfit the data. In fact, even in situations in which g_{cl} does not learn, g_{cl} will not overfit the data because $f_{\mathrm{emp}}[\vec{s}\,'', g_{\mathrm{cl}}[\vec{s}\,]]$ provides an accurate estimate of $\mathrm{err}(P, g[\vec{s}\,])$.

Let g_{emp} denote (the usual) class of learning procedure based on minimizing the empirical estimator, f_{emp}. Each such g_{emp} is determined by a sequence $\epsilon_m \to 0$, $\epsilon_m \geq 0$, and a mapping that ϵ_m-approximately minimizes f_{emp}.

We can compare g_{cl} with g_{emp} when L is Lipschitz in h. Under this condition, we know by Lemma 4.1 of [1] that f_{ce} smoothly simultaneously estimates $(\mathcal{P}, \mathcal{H})$ whenever f_{emp} does. However, smooth simultaneous estimability is only a sufficient condition for learnability. There are some $(\mathcal{P}, \mathcal{H})$ that are learnable by certain choices for g_{cl} or g_{emp} (recall the need to choose ϵ_m, τ_n, μ_n, etc.), but are not simultaneously estimable by f_{ce} or f_{emp}, respectively. For g_{cl}, this could occur if, by a fortunate choice, $\mathcal{H}_n(\vec{x}\,')$ always contains a hypothesis that

minimizes $\mathrm{err}(P,h)$. Similarly, even though the set of hypotheses that minimizes f_{emp} may contain many nonoptimal hypotheses, g_{emp} may make a fortuitous selection of an optimal hypothesis. We cannot expect to rely on such luck in practice. Thus, we should not compare g_{cl} and g_{emp} according to whether *some* choice of each learns, but we should rather compare them according to whether *every* possible choice of them learns. The following theorem and example show that, in this sense, g_{cl} is superior to g_{emp}.

Theorem 2.2. *Under Assumptions 2.1 and 2.2, if L is Lipschitz in h,*

$$\textit{Every choice of } g_{\mathrm{emp}} \textit{ learns } (\mathcal{P},\mathcal{H}) \Rightarrow \textit{ Every choice of } g_{\mathrm{cl}} \textit{ learns } (\mathcal{P},\mathcal{H}).$$

Proof: See the Appendix.

We close this section with an example of a pair $(\mathcal{P},\mathcal{H})$ that is learnable by g_{cl}, but not by g_{emp}.

Example 2.1.

Consider the real-valued, noise-free case (refer to Section 2 of [1]). Let $X = [0,1)$, $Y = Z = [0,1]$, and take L and d to be equal, with $L(z,y) = |z-y|$. Let $\mathcal{Q}$ be a singleton, $\{Q\}$, where Q is uniform on X. Further, let $\mathcal{C} = \{c\}$, where $c(x) \equiv 0$.

Suppose that $\mathcal{H} = \{\hat{h}\} \cup \bigcup_{i=1}^{\infty} h_i$, where $\hat{h}(x) \equiv 1/4$ and $h_i(x) = b_i(x)$, where $0.b_1(x)b_2(x)b_3(x)\ldots$ is the binary representation of $x \in X$. Note that $\hat{h}$ minimizes the error:

$$\mathrm{err}(Q,c,\hat{h}) = 1/4, \text{ and } \mathrm{err}(Q,c,h_i) = 1/2 \text{ for all } i.$$

A learning procedure for $(\{Q\},\{c\},\mathcal{H})$ should therefore pick $\hat{h}$.

We first note that g_{cl} learns $(\{Q\},\{c\},\mathcal{H})$, because, as shown next, this triple can be empirically covered (refer to Definition 5.2 and Theorem 5.1 of [1]). Any pair $h,h' \in \mathcal{H}$ will fall into one of three categories: $h = h' = \hat{h}$; $h = h_i$ and $h' = h_j$; or $h = \hat{h}$ and $h' = h_i$. In each of these cases, for any $\vec{x}$,

$$\rho_{\vec{x}}(h,h') = E_{\vec{x}}|h-h'| < 1/4 \Rightarrow |\mathrm{err}(Q,h) - \mathrm{err}(Q,h')| = 0.$$

Thus, $(\{Q\}, \{c\}, \mathcal{H})$ can be empirically covered, and therefore g_{cl} learns $(\{Q\}, \{c\}, \mathcal{H})$.

However, g_{emp} does not learn $(\{Q\}, \{c\}, \mathcal{H})$ (and thus f_{emp} does not simultaneously estimate it either). To see this, let the binary representation of $x \in X$ be $0.b_1(x)b_2(x)b_3(x)\ldots$. When x is drawn according to the uniform distribution Q, the b_i's are independent and identically distributed: $b_i(x) = 0$ with probability $1/2$. For a sample $\vec{x}(m)$, the probability that none of $b_1(\vec{x}), \ldots, b_\ell(\vec{x})$ is $\vec{0}$ (a vector of zeroes) is $(1 - 2^{-m})^\ell$. Thus, for any m and any fixed $\delta < 1$,

$$Q^m[\vec{x}(m) : \text{ there is } i \text{ such that } h_i(\vec{x}) = \vec{0}\,] \geq \delta.$$

So, with probability at least δ,

$$\min_{h_i,\, i=1,2,\ldots} f_{\text{emp}}[(\vec{x}, c(\vec{x})), h_i] = 0.$$

However, $f_{\text{emp}}[(\vec{x}, c(\vec{x})), \hat{h}] = 1/4$ for any $\vec{x}$. Since g_{emp} ϵ_m-approximately minimizes f_{emp}, for large m, g_{emp} will not pick the best hypothesis, $\hat{h}$, for a substantial portion of the samples. □

3 Sample Size Bounds

We have seen in [1] and the previous section that the canonical estimator and learner work whenever empirical error estimation and minimization do. One might suspect that these canonical procedures succeed only by using an exorbitant number of samples. In this section, we show that this is not the case. In fact, in some important special cases we give bounds on the number of samples needed by f_{ce} and g_{cl} that are less (by a constant factor) than the known bounds for f_{emp} and g_{emp}.

3.1 Preliminaries

The *sample size* of f_{emp}, $m_{f_{\text{emp}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$, is the least value of m such that

$$\sup_{P \in \mathcal{P}} P^n \left[\vec{s}(n) : \sup_{h \in \mathcal{H}} \mid f_{\text{emp}}[\vec{s}, h] - \text{err}(P, h) \mid > \epsilon \right] < \delta$$

for all $n \geq m$. For given values of $\mathcal{P}$ and $\mathcal{H}$, $m_{f_{\text{emp}}}$ is fixed. Note, however, that there is some freedom in choosing a canonical estimator, due to the leeway in choosing τ_n, μ_n, etc. In Procedure 5.2 of [1], we constrained the admissible values of these parameters so that, if $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable, any such choice of f_{ce} would be a smooth simultaneous estimator. However, we may want to choose other values for these parameters to improve the sample size of the resulting estimator.

Recall the important steps in f_{ce}: take n of the m samples, find a τ_n/n-empirical covering for $\mathcal{H}$, and use the rest of the samples to estimate the error of the cover elements. Thus, for a fixed number of samples, m, the form of f_{ce} depends only on two parameters: n and τ. We define the *minimum sample size* of f_{ce}, $\tilde{m}_{f_{\text{ce}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$, to be the least value of m (at each ϵ and δ) such that there are values of n and τ for which the mapping f_{ce} satisfies

$$\sup_{P \in \mathcal{P}} P^m \left[\vec{s}(m) : \sup_{h \in \mathcal{H}} \mid f_{\text{ce}}[\vec{s}(m), h] - \text{err}(P, h) \mid > \epsilon \right] < \delta. \tag{1}$$

Note that there is also some freedom in the choice of an empirical cover for each n, τ, and $\vec{s}$. Since we cannot rely on making "lucky" choices for the empirical cover, we require that $\tilde{m}_{f_{\text{ce}}}$ be such that (1) holds for all choices of the τ/n-empirical covering.

Let us consider what it means to compare $m_{f_{\text{emp}}}$ and $\tilde{m}_{f_{\text{ce}}}$. If we show that

$$\tilde{m}_{f_{\text{ce}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H}) \leq m_{f_{\text{emp}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H}) \text{ for each } \epsilon \text{ and } \delta, \tag{2}$$

this means that there is *some* choice of f_{ce} (that varies with ϵ and δ) whose sample size is less than that of f_{emp} at ϵ, δ. Note though that there need not be a single choice of f_{ce} (determined by fixed sequences $\{n_m\}$ and $\{\tau_m\}$) that requires fewer samples than f_{emp} for *all* values of ϵ and δ. However, in terms of comparing the number of samples needed by f_{emp} and f_{ce} for given values of ϵ and δ, it is valid to compare $m_{f_{\text{emp}}}$ and $\tilde{m}_{f_{\text{ce}}}$.

In analogy with $\tilde{m}_{f_{\text{ce}}}$, we could define a minimum sample size for the canonical learner, $\tilde{m}_{g_{\text{cl}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$. This would be the least m such that, for any $P \in \mathcal{P}$, the error of $g_{\text{cl}}[\vec{s}(m)]$ is within ϵ of $\text{opt}(P, \mathcal{H})$ with probability at least $1 - \delta$. If, for simplicity, we take $\epsilon_m \equiv 0$ in Procedure 2.1, then we can bound $\tilde{m}_{g_{\text{cl}}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$ by $\tilde{m}_{f_{\text{ce}}}(\epsilon/2, \delta, \mathcal{P}, \mathcal{H})$. To see this, note

that if $f_{\rm ce}$ is within $\epsilon/2$ of $\mathrm{err}(P,h)$ simultaneously over $\mathcal{H}$, then

$$\begin{aligned}\mathrm{err}(P, g_{\rm cl}[\vec{s}]) &\leq \epsilon/2 + \min_{h\in\mathcal{H}} f_{\rm ce}[\vec{s},h] \\ &\leq \epsilon/2 + \inf_{h\in\mathcal{H}} \mathrm{err}(P,h) + \epsilon/2 \\ &= \epsilon + \mathrm{opt}(P,\mathcal{H}).\end{aligned}$$

A similar comment holds for bounding the sample size of $g_{\rm emp}$ by $m_{f_{\rm emp}}$. Hence, in the following, we focus on sample size bounds for the estimators.

We shall bound $\tilde{m}_{f_{\rm ce}}$ by establishing bounds for three intermediate quantities. The first of these is the number of samples needed to achieve a good covering for $\mathcal{H}$. Let $N_{\rm ec}(\tau,\epsilon,\delta)$ be the least value of n for which

$$\sup_{P\in\mathcal{P}} P^n \left[\vec{s}(n) : \sup_{h,h'\in\mathcal{H} \text{ s.t. } \rho_{\vec{x}(n)}(h,h')<\tau/n} \mid \mathrm{err}(P,h) - \mathrm{err}(P,h') \mid > \epsilon \right] < \delta,$$

where

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

The second is the maximum size of an empirical τ/n-cover based on a sample of length n.

$$H_{\rm ec}(\tau,n) := \max_{\vec{x}(n)\in X^n} |\mathcal{H}_n(\vec{x})|,$$

where $\mathcal{H}_n(\vec{x})$ is an empirical τ/n-cover for $\mathcal{H}$ constructed by the method in Procedure 5.1 of [1]. The last quantity is $\ell(H,\epsilon,\delta)$, a number such that $f_{\rm emp}$ can simultaneously estimate the errors of a subset of $\mathcal{H}$ of size H to within ϵ with probability at least $1-\delta$ using ℓ or more samples.

As in the proof of Theorem 5.1 of [1], if we take $N_{\rm ec}(\tau,\epsilon/2,\delta/2)$ labeled samples, $\vec{s}'$,

$$\sup_{h\in\mathcal{H}} \mid \mathrm{err}(P, a_{n,\vec{x}'}(h)) - \mathrm{err}(P,h) \mid > \epsilon/2$$

with probability less than $\delta/2$. If we take $\ell(H_{\rm ec}(\tau, N_{\rm ec}(\tau,\epsilon/2,\delta/2)),\epsilon/2,\delta/2)$ more labeled samples, $\vec{s}''$,

$$\sup_{h\in\mathcal{H}} \left| f_{\rm emp}[\vec{s}'', a_{n,\vec{x}'}(h)] - \mathrm{err}(P,h) \right| > \epsilon$$

with probability less than δ. Therefore, $\tilde{m}_{f\text{ce}}$ can be bounded as

$$\tilde{m}_{f\text{ce}}(\epsilon, \delta, \mathcal{P}, \mathcal{H}) \leq N_{\text{ec}}(\tau, \epsilon/2, \delta/2) + \ell(H_{\text{ec}}(\tau, N_{\text{ec}}(\tau, \epsilon/2, \delta/2)), \epsilon/2, \delta/2). \tag{3}$$

Note that $H_{\text{ec}}(\tau, n)$ and $\ell(H, \epsilon, \delta)$ increase with n and H, respectively. Therefore, to bound $\tilde{m}_{f\text{ce}}$ it suffices to bound N_{ec}, H_{ec}, and ℓ. We are free to change τ.

In bounding $\tilde{m}_{f\text{ce}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$, we split ϵ and δ evenly between N_{ec} and ℓ. By other choices of this split, $\tilde{m}_{f\text{ce}}$ can usually be reduced. The simple analysis above suffices for our purposes.

3.2 The General Case

To compare the sample size of f_{emp} with $\tilde{m}_{f\text{ce}}$, we shall assume that f_{emp} simultaneously estimates $(\mathcal{P}, \mathcal{H})$, since otherwise $m_{f\text{emp}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$ is infinite for small enough values of ϵ and δ. Also, suppose that L is Lipschitz in h with a constant D.

Fix $\epsilon > 0$ and $\delta > 0$ and consider a value of $\vec{s} = (\vec{x}, \vec{y})$ such that $f_{\text{emp}}[\vec{s}, h]$ is within $\epsilon/5$ of $\text{err}(P, h)$ for all $h \in \mathcal{H}$. As in the proof of Lemma 4.1 in [1],

$$\rho_{\vec{x}}(h, h') < \tau/n \Rightarrow |f_{\text{emp}}[\vec{s}, h] - f_{\text{emp}}[\vec{s}, h']| < D\tau/n,$$

and therefore

$$\rho_{\vec{x}}(h, h') < \tau/n \Rightarrow |\text{err}(P, h) - \text{err}(P, h')| < D\tau/n + 2\epsilon/5 \leq D\tau + 2\epsilon/5.$$

Thus, $N_{\text{ec}}(\tau, D\tau + 2\epsilon/5, \delta/2) \leq m_{f\text{emp}}(\epsilon/5, \delta/2, \mathcal{P}, \mathcal{H})$. Choose $\tau = \frac{\epsilon}{10D}$. Then

$$N_{\text{ec}}(\tau, \epsilon/2, \delta/2) \leq m_{f\text{emp}}(\epsilon/5, \delta/2, \mathcal{P}, \mathcal{H}). \tag{4}$$

By Lemma 5.2 of [1], $H_{\text{ec}}(\tau, n) \leq [K(\tau/2n)]^n$. However, without knowing how $K(\tau/2n)$ changes with τ, we cannot give a definite bound for H_{ec} (see the remark that follows). Instead, let us ignore H_{ec} and note that we can bound $\ell(H, \epsilon, \delta)$ by $m_{f\text{emp}}(\epsilon, \delta, \mathcal{P}, \mathcal{H})$ for any H. From (3) and (4), we have

$$\tilde{m}_{f\text{ce}}(\epsilon, \delta, \mathcal{P}, \mathcal{H}) \leq m_{f\text{emp}}(\epsilon/5, \delta/2, \mathcal{P}, \mathcal{H}) + m_{f\text{emp}}(\epsilon/2, \delta/2, \mathcal{P}, \mathcal{H}).$$

Thus, if $m_{f\text{emp}}$ is polynomial in $1/\epsilon$ and $\ln(1/\delta)$, $\tilde{m}_{f\text{ce}}$ is as well. This tells us that $\tilde{m}_{f\text{ce}}$ is, at most, roughly the same order as $m_{f\text{emp}}$.

Remark 3.1. *We can capture the manner in which $K(\alpha)$ scales with α by the quantity*

$$\lim_{\alpha \searrow 0} \frac{\log K(\alpha)}{\log(\frac{1}{\alpha})}.$$

This is the metric dimension *of the set Z (see [5] and [14]). It is a way of defining the dimension of a metric space (even when it is not a vector space) by appealing to the notion of volume. For example, if Z is $\mathbb{R}^n$ and d is the Euclidean distance, $K(\alpha)$ is proportional to $1/\alpha^n$.*

3.3 The 0/1-valued, Distribution-Free Case

We now specialize to the case where the hypotheses and labels are 0/1-valued and $\mathcal{P} = \mathcal{P}^*$, the set of all probability distributions on $X \times \{0,1\}$. In [2], Vapnik and Chervonenkis introduced a property of a set of 0/1-valued functions $\mathcal{H}$ that determines when $(\mathcal{P}^*, \mathcal{H})$ is simultaneously estimable by f_{emp}. This property has come to be known as the "Vapnik-Chervonenkis dimension" of $\mathcal{H}$, and it plays a central role in the theory of PAC-learning (see [15]). We will compare the sample sizes of f_{emp} and f_{ce} in terms of the VC-dimension of $\mathcal{H}$.

Before defining the VC-dimension, we develop some notation. Let $\mathcal{B}$ be a class of 0/1-valued functions on a set V, and let $\vec{v}(n)$ be an n-vector of points from V. We use $\Delta^{\mathcal{B}}(\vec{v})$ to denote the number of distinct labelings of $\vec{v}$ by functions from $\mathcal{B}$: $\Delta^{\mathcal{B}}(\vec{v}) := |\{b(\vec{v}) : b \in \mathcal{B}\}|$. Clearly, $\Delta^{\mathcal{B}}(\vec{v}(n)) \leq 2^n$. We say that $\mathcal{B}$ *shatters* $\vec{v}(n)$ if $\Delta^{\mathcal{B}}(\vec{v}(n)) = 2^n$.

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

See [2], [16], [17], [4], [18], [15], [19], [20], [21], and [22] for examples of classes of finite VC-dimension.

Using results from [4] and [15], we can bound $m_{f_{\text{emp}}}$ as follows.

Lemma 3.1. *With $q = \text{VCdim}(\mathcal{H}), 1 \leq q < \infty$,*

$$m_{f_{\text{emp}}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H}) < \frac{q}{\epsilon^2}\left[20\ln(8/\epsilon) + 8\ln(6/\delta)\right]. \tag{5}$$

Proof: See the Appendix.

In [14], Haussler uses results from [23] to give a bound on $m_{f_{\text{emp}}}$ that is similar to (5) in that it is of order $q \ln(1/\epsilon)/\epsilon^2$ in ϵ and q.

The proof of the following result appears in the Appendix.

Theorem 3.1. *With $q = \text{VCdim}(\mathcal{H}), 1 \leq q < \infty$, $\tilde{m}_{f_{\text{ce}}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H})$ can be upper bounded by a number of order*

$$\frac{2}{\epsilon^2}\left[q \ln \frac{60\epsilon}{\epsilon} + \ln \frac{4}{\delta}\right]. \tag{6}$$

(Where "of order" means that we have retained only the highest order terms in $1/\epsilon$ and $1/\delta$.)

Note that (6) is of the same order as the bound for $m_{f_{\text{emp}}}$ in Lemma 3.1. In fact, the leading constant is several times less. We must keep in mind, however, that these are only *bounds* on $m_{f_{\text{emp}}}$ and $\tilde{m}_{f_{\text{ce}}}$.

It is interesting to note that N_{ec} is smaller than (6) by a factor of $1/\epsilon$. Thus, only a small fraction of the samples are needed for creating the empirical cover.

Lemma 3.1 gives a bound for $m_{g_{\text{emp}}}$ that is of order $\ln(1/\epsilon)/\epsilon^2$ in ϵ. In the special case where $\text{opt}(P, \mathcal{H}) = 0$ for each $P \in \mathcal{P}^*$, [15] and [14] give bounds for $m_{g_{\text{emp}}}$ that are of order $\ln(1/\epsilon)/\epsilon$. A separate analysis shows that there is a bound of order $\ln(1/\epsilon)/\epsilon$ for $\tilde{m}_{g_{\text{cl}}}$ in this case as well.

3.4 The Real-Valued, Distribution-Free Case

Here, we examine the case where the hypotheses and labels are real-valued and $\mathcal{P} = \mathcal{P}^*$, the set of all probability distributions on $X \times Y$. Building on the work of Dudley ([16] and [18]) and Pollard ([23] and [24]), Haussler has made much progress in finding conditions that are sufficient for f_{emp} to be a simultaneous estimator (see [25] and [14]). One of these conditions is that a certain "pseudodimension" be finite. This pseudodimension generalizes the VC-dimension to classes of real-valued functions. (Vapnik generalizes the VC-dimension in a different fashion in [4].)

For $r \in \mathbb{R}$, let sign$[r]$ be 1 if $r > 0$ and 0 otherwise. Also, let Φ be a class of real-valued functions on a set V. We have the following definition from [14] and [24].

Definition 3.3. *The* pseudodimension *of* Φ, psdim(Φ), *is the largest* n *for which there is some* $\vec{v}(n) \in V^n$ *and* $\vec{r}(n) \in \mathbb{R}^n$ *such that*

$$|\{sign[\phi(\vec{v}) + \vec{r}] : \phi \in \Phi\}| = 2^n.$$

Let $L_{\mathcal{H}}$ be the set of functions on $X \times Y$ given by $L_{\mathcal{H}} = \{L(h(\cdot), \cdot) : h \in \mathcal{H}\}$. We have the following bound on $m_{f\text{emp}}$ from [14].

Lemma 3.2. *Under Assumption 2.1,*

$$m_{f\text{emp}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H}) \leq \frac{64\lambda^2}{\epsilon^2}\left[2\text{psdim}(L_{\mathcal{H}}) \ln \frac{16e\lambda}{\epsilon} + \ln \frac{8}{\delta}\right]. \tag{7}$$

This bound can be compared with the following, which we prove in the Appendix.

Theorem 3.2. *Let* $Z = [0, B]$, $d(z_1, z_2) = |z_1 - z_2|$, *and* L *be Lipschitz in* h *with constant* D. *Then under Assumption 2.1,* $\tilde{m}_{f\text{ce}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H})$ *can be upper bounded by a number of order*

$$\frac{\lambda^2}{\epsilon^2}\left[3\text{psdim}(\mathcal{H}) \ln \frac{32eBD}{\epsilon} + 2 \ln \frac{8}{\delta}\right]. \tag{8}$$

(Where "of order" means that we have retained only the highest order terms in $1/\epsilon$ *and* $1/\delta$*.)*

Comparing (8) and (7), we see that these two are of the same size when psdim$(\mathcal{H})$ and BD are close to psdim$(L_{\mathcal{H}})$ and λ, respectively. The leading constants in (8) are smaller by a factor of more than thirty than those in (7). As an example, when $Y = [0, B]$ and $L(z, y) = |z - y|$, we have that $\lambda = B$, $D = 1$, and psdim$(\mathcal{H}) \leq$ psdim$(L_{\mathcal{H}})$, so $\tilde{m}_{f\text{ce}}$ compares favorably with $m_{f\text{emp}}$. Haussler gives some conditions for which psdim$(\mathcal{H}) =$ psdim$(L_{\mathcal{H}})$ on page 115 of [14].

4 Choice of Model Complexity

Thus far, we have only examined how to pick a hypothesis from a fixed set $\mathcal{H}$. We also have the option of selecting $\mathcal{H}$ itself. For instance, perhaps we should use a richer class $\mathcal{H}$ as the

amount of data we have increases in order to decrease $\mathrm{opt}(P, \mathcal{H})$. In this section, we address this issue by focusing on hypothesis classes that are *nested*:

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

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 be k'th order polynomials, or a neural network with k nodes.

Overfitting is hard to avoid when we attempt to learn from a finite amount of noisy data using a rich hypothesis class. That is, we may pick a complex hypothesis that agrees too well with the noisy data and therefore has a large generalization error. The question we must address is how to determine the appropriate complexity for our hypotheses or "models." Most of the proposed solutions for this problem require the evaluation of an a priori measure of the complexity of each $\mathcal{H}^k$ (e.g., $\mathrm{VCdim}(\mathcal{H}^k)$). However, these approaches often do not make use of the observed distribution of the data and may involve a great deal of analysis to calculate the complexity measures. We present a learning procedure in this section that does not suffer from these drawbacks.

For many rich, interesting classes $\mathcal{H}$, learning $(\mathcal{P}, \mathcal{H})$ will frequently be impossible because of the requirement that the number of samples used be *uniform* over $P \in \mathcal{P}$. Indeed, $\mathrm{VCdim}(\mathcal{H}) < \infty$ is *necessary* for learning $(\mathcal{P}, \mathcal{H})$ ([15]). Accordingly, we concentrate on *nonuniform* learnability in this section.

Definition 4.1. $(\mathcal{P}, \mathcal{H})$ *is* nonuniformly learnable *if there is a mapping* $g : \vec{s} \mapsto \mathcal{H}$ *such that, for each* $P \in \mathcal{P}$,

$$|\mathrm{err}(P, g[\vec{s}(m)]) - \mathrm{opt}(P, \mathcal{H})| \to 0 \tag{9}$$

in probability as $m \uparrow \infty$.

The learning procedure that we present here has the attractive property that it learns uniformly over distributions that have the same "complexity."

Before describing our approach to model selection, we review some common alternatives.

4.1 Some Conventional Methods for Choosing Model Complexity

We now briefly discuss some methods from the literature for selecting the complexity of hypotheses. These schemes fall into three categories: constrained complexity methods, penalty methods, and error estimation schemes.

First, consider constrained complexity methods. In these methods, a measure of the complexity of $\mathcal{H}^k$ is used to determine a function $k(n)$ such that minimization of $f_{\text{emp}}[\vec{s}(n), h]$ over $\mathcal{H}^{k(n)}$ based on n labeled samples will result in successful learning. A variety of methods use this same principle, and all of them rely on some connection between the convergence properties of f_{emp} and some prior measure of the complexity of hypotheses from $\mathcal{H}^k$ (see [26], [27], [28], [29], [30], and [31]). For instance, consider the case in which each $\mathcal{H}^k$ has finite VC-dimension. Inspection of Lemma 3.1 shows that, if we let $k(n)$ increase slowly enough, then f_{emp} simultaneously estimates $(\mathcal{P}^*, \mathcal{H}^{k(n)})$. Thus, we can nonuniformly learn $(\mathcal{P}^*, \mathcal{H})$ by minimizing $f_{\text{emp}}[\vec{s}(n), h]$ over $\mathcal{H}^{k(n)}$. However, to determine how fast to let $k(n)$ grow, we must know $\text{VCdim}(\mathcal{H}^k)$, a measure of the "complexity" of $\mathcal{H}^k$.

In penalty methods, for each k a candidate hypothesis $h_k^* \in \mathcal{H}^k$ is found by minimizing $f_{\text{emp}}[\vec{s}(n), h]$ over $\mathcal{H}^k$. Then, a penalty function $\text{pen}(n, k)$ (which increases with k) is added to $f_{\text{emp}}[\vec{s}(n), h_k^*]$ and the resulting quantity is minimized over k to select the final hypothesis. The underlying intuition is that we should trade some accuracy on the data in exchange for a "simpler" hypothesis. For examples of this method, see [32], [33], [34], [35], [4], [13], [36], [37], [38], and [39]. Again, this penalty is determined from some measure of complexity that is derived from the structure of $\mathcal{H}$ and must be carefully selected to ensure that learning occurs.

The two preceding methods 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, the *a priori* complexity measures strongly influence which $\mathcal{H}^k$ the final hypothesis comes from. The learning procedure we present in the next subsection avoids these problems by using only the nested structure of $\mathcal{H}$ to compare the

complexity of hypotheses and by using the observed distribution of the data to determine which $\mathcal{H}^k$ the final hypothesis comes from.

Error estimation schemes are used in applications for which the sample size is a critical issue. There are a number of techniques by which one can attempt to estimate the error of the candidate hypotheses $h_k^* \in \mathcal{H}^k$ and thereby choose the best value of k (see [40] for an overview of these methods). Most of these schemes involve withholding part of the samples and/or resampling the data in some fashion, as in cross-validation ([6] and [36]) and bootstrapping ([41]). A point that we should note here is that f_{emp} is still used in many of these methods to select the initial candidate hypotheses h_k^*. The learning procedure we present next, or at least the ideas it is based on, may provide a useful alternative to these methods.

4.2 Learning with a Simple Empirical Cover

We now show how a straightforward extension of the canonical learner, g_{cl}, can be used to address the problem of selecting model complexity. Recall that g_{cl} uses part of the data to approximate $\mathcal{H}$ by a finite subset, an empirical cover. When $\mathcal{H}$ is nested, we can modify this approach by requiring that the elements of the empirical cover be (nearly) as simple as possible.

Definition 4.2. *An* $\boldsymbol{M}$-simple empirical ϵ-covering for $\boldsymbol{\mathcal{H}}$ based on $\vec{\boldsymbol{x}}(\boldsymbol{n})$ *is a set* $\mathcal{H}_n(\vec{x})$ *and a mapping* $a_{n,\vec{x}} : \mathcal{H} \mapsto \mathcal{H}_n(\vec{x})$ *such that, for each* j *and any* $h \in \mathcal{H}^j$, *(i)* $\rho_{\vec{x}}(a_{n,\vec{x}}(h), h) < \epsilon$ *and (ii)* $a_{n,\vec{x}}(h) \in \mathcal{H}^{\lfloor Mj \rfloor}$.

Using a large M allows the elements of the cover to be less simple. In some cases, finding the *simplest* hypothesis consistent with a labeling is much harder than finding one that is only reasonably simple (see [42]). This would dictate using $M > 1$.

A key observation is that we can construct *finite* M-simple empirical coverings.

Lemma 4.1. *Under Assumption 2.2, for any* $M \geq 1$ *we can construct an* M*-simple empirical* ϵ*-covering for* $\mathcal{H}$ *based on* $\vec{x}(n)$ *that has at most* $[K(\epsilon/2)]^n$ *elements.*

Proof: See the Appendix.

The learning procedure that we now define is identical to the canonical learner, except that it uses M-simple empirical covers. In the following procedure, let Assumptions 2.1 and 2.2 be satisfied.

Procedure 4.1. *Learning with an M-simple empirical cover.*

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

1. *Let $H(n) = [K(\tau_n/2)]^n$. Let n be the largest integer such that*

$$n + \lceil (\lambda^2/2\mu_n^2) \ln[2H(n)/\nu_n] \rceil \leq m. \tag{10}$$

   Split $\vec{s}$ into two subsamples, $\vec{s}(m) = (\vec{s}\,'(n), \vec{s}\,''(m-n))$, where $\vec{s}\,'(n) = (\vec{x}\,', \vec{y}\,')$.

2. *Construct, as in the proof of Lemma 4.1, an M-simple empirical (τ_n)-covering for $\mathcal{H}$ based on $\vec{x}\,'$, By Lemma 5.1 of [1], with probability at least $1 - \nu_n$, the empirical error estimate, $f_{\text{emp}}[\vec{s}\,'', h']$, will be within μ_n of $\text{err}(P, h')$ simultaneously for all $h' \in \mathcal{H}_n(\vec{x}\,')$.*

3. *Let g_{sec} ϵ_m-approximately minimize $f_{\text{emp}}[\vec{s}\,''(m-n), h]$ over $\mathcal{H}_n(\vec{x}\,')$:*

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

In the following subsection, we examine two results that establish the usefulness of g_{sec}. The first gives a condition than many practically interesting $\mathcal{H}$ satisfy that is sufficient for g_{sec} to learn. The second shows that g_{sec} works whenever "constrained complexity" methods work.

4.3 Sufficient Conditions

In order to describe the conditions under which g_{sec} works, we develop some terminology.

Definition 4.3. *Let Φ be a class of functions that map X to Z. We say that $\boldsymbol{E}_{\vec{x}}$* converges simultaneously over $(\boldsymbol{\mathcal{P}}, \boldsymbol{\Phi})$ *if*

$$\sup_{\phi \in \Phi} |E_P \phi - E_{\vec{x}(m)} \phi| \to 0$$

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

Also, let $d(\mathcal{H}^i, \mathcal{H}^i) := \{d(h, h') : h, h' \in \mathcal{H}^i\}$.

Under our usual assumptions, we can show that g_{sec} *nonuniformly* learns $(\mathcal{P}, \mathcal{H})$ whenever $E_{\vec{x}}$ converges simultaneously over $(\mathcal{P}, d(\mathcal{H}^i, \mathcal{H}^i))$ for each i. We can, in fact, prove a much stronger result. We shall show that g_{sec} can learn *uniformly* over distributions that have the same "complexity." Consider the nested classes $\mathcal{P}_i$ given by

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

If the true distribution comes from $\mathcal{P}_i$, we only need to consider hypotheses from $\mathcal{H}^i$. Thus, we can think of the distributions 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 can learn *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.

Theorem 4.1. *Suppose that Assumptions 2.1 and 2.2 hold and* L *is Lipschitz in* h. *If* $E_{\vec{x}}$ *converges simultaneously over* $(\mathcal{P}, d(\mathcal{H}^i, \mathcal{H}^i))$ *for each* i, *then* g_{sec} *nonuniformly learns* $(\mathcal{P}, \mathcal{H})$, *and it does so uniformly over each* $\mathcal{P}_i$. *That is, for each* i,

$$|g_{\text{sec}}[\vec{s}(m)] - \text{opt}(P, \mathcal{H})| \to 0$$

in probability as $m \uparrow \infty$ *uniformly over* $P \in \mathcal{P}_i$.

Proof: See the Appendix.

The utility of Theorem 4.1 stems from two facts. First, Procedure 4.1 does not utilize any special knowledge about $\mathcal{H}$ other than its decomposition into nested subsets. Second, it applies in many cases of interest. For instance, consider the case where the hypotheses and labels are real-valued and the distance measure $d(z_1, z_2) = |z_1 - z_2|$. Using results from [4], it is straightforward to show (see [43]) that $E_{\vec{x}}$ converges simultaneously over $(\mathcal{P}, d(\mathcal{H}^i, \mathcal{H}^i))$

whenever it does so over $(\mathcal{P}, \mathcal{H}^i)$. Thus, when $\mathrm{VCdim}(\mathcal{H}^i)$ or $\mathrm{psdim}(\mathcal{H}^i)$ is finite for each i (as is the case for many parametric model classes–see [19], [14], [20], [21] and [22]), $E_{\vec{x}}$ converges simultaneously over $(\mathcal{P}^*, d(\mathcal{H}^i, \mathcal{H}^i))$ for each i. Yet, in this case, we do not need to know the *value* of these dimensions to apply Procedure 4.1.

4.4 A Comparison with Constrained Complexity Methods

We could give a variety of conditions that are sufficient for either g_{sec} or the methods mentioned in Subsection 4.1 to work. It is more satisfactory to directly compare g_{sec} with one of these others, as we compared g_{cl} and g_{emp} in Section 2. In this subsection, we show that g_{sec} works whenever "constrained complexity" methods do.

Let g_{cc} denote the class of constrained complexity learning procedures described previously. Given a function $k(n) : \boldsymbol{N} \mapsto \boldsymbol{N}$ and a sequence $\epsilon_n \to 0$, $g_{\mathrm{cc}}[\vec{s}(n)]$ ϵ_n-approximately minimizes $f_{\mathrm{emp}}[\vec{s}(n), h]$ over $\mathcal{H}^{k(n)}$:

$$g_{\mathrm{cc}}[\vec{s}(n)] \in \left\{ h \in \mathcal{H}^{k(n)} : f_{\mathrm{emp}}[\vec{s}, h] \leq \min_{h' \in \mathcal{H}^{k(n)}} f_{\mathrm{emp}}[\vec{s}, h'] + \epsilon_n \right\}. \tag{11}$$

As with the previous learning procedures, there are various *choices* of g_{cc} determined by the choice of the sequence ϵ_n and the "minimizing" hypothesis in (11). Likewise, there are various choices of g_{sec} determined by τ_n, μ_n, $\mathcal{H}_n(\vec{x}\,')$, etc.

In order for g_{cc} to learn, $k(n)$ must not increase too fast with n. A $k(n)$ that works must be determined from the structure of $\mathcal{P}$ and $\mathcal{H}$ (e.g., by knowing $\mathrm{VCdim}(\mathcal{H}^i)$). The procedure g_{sec} has an advantage in this regard, since whenever there is *some* $k(n)$ such that g_{cc} works, g_{sec} (with $M = 1$) works as well.

Theorem 4.2. *Suppose that Assumptions 2.1 and 2.2 hold and L is Lipschitz in h. If there is $k(n)$ such that every choice of g_{cc} nonuniformly learns $(\mathcal{P}, \mathcal{H})$, then every choice of g_{sec} with $M = 1$ nonuniformly learns $(\mathcal{P}, \mathcal{H})$.*

Proof: See the Appendix.

As a trivial example of a case where g_{sec} works and g_{cc} (as well as many of the methods in Subsection 4.1) does not, consider the situation in Example 2.1 and set $\mathcal{H}^i \equiv \mathcal{H}$.

5 Concluding Remarks

In this paper and its predecessor ([1]), we have introduced new paradigms for estimating errors and learning that have advantages over empirical error-based methods. In [1], we established the generality of our canonical estimator by showing that it solves any simultaneous estimation problem that can be solved by an estimator that satisfies a natural smoothness constraint. For a broad class of learning problems (i.e., those for which Assumptions 2.1 and 2.2 hold and L is Lipschitz in h), we have shown that the canonical estimator and learner work whenever the corresponding empirical error-based schemes work. We have demonstrated, by examples, that the converse of this is not true. We have derived bounds on the sample sizes of our canonical procedures that are comparable (and in some special cases, smaller than) the available bounds for the empirical error-based methods. Finally, we have used the ideas behind the canonical learner to create a scheme for selecting a hypothesis of appropriate complexity from a rich, nested hypothesis class. Unlike many existing methods, this method uses the observed data to directly influence the complexity of the final, chosen hypothesis and does not require the evaluation of an a priori measure of the complexity of hypotheses. We have proved a result which indicates that this method works for many common parametric hypothesis classes. Also, under our usual assumptions (Assumptions 2.1 and 2.2 and L being Lipschitz in h), we have shown that our method will work whenever the conventional "constrained complexity" methods do. Although much work has been done, there are still many unanswered questions.

The one real advantage of the empirical error-based schemes is their simplicity. It remains to be seen whether there are computationally efficient ways of finding a good hypothesis with the canonical learner. For a specific problem, we may be able to use some additional structure to search for an optimum hypothesis in the empirical cover without actually constructing the cover itself.

It would be useful to find sample size bounds for the complexity selection method presented in Section 4. As we saw in Section 3, this sort of analysis can help us decide how to split the samples between those used for creating the empirical cover and the "test set."

In addition, a fertile area for experimental work is to use real data to compare the actual sample sizes of the canonical procedures with those of the empirical error-based methods.

There are a number of modifications that could be made to the learning framework. One of particular interest is to relax the requirement that the labeled samples are drawn independently and are identically distributed. A few papers in the literature do address more general situations; see [44], [30], [45], and [46]. Also, the learning framework could be extended to a nonparametric setting by allowing the hypothesis class to vary with the observed data (as in nearest neighbor classification).

It might prove useful to find a natural interpretation of the class of $(\mathcal{P}, \mathcal{H})$ that are learnable by the canonical learner. Specifically, does this class contain only $(\mathcal{P}, \mathcal{H})$ that are, in some sense, "smoothly learnable"?

Finally, we note that the canonical estimator and learner make essential use of the knowledge of $\mathcal{H}$ (in the construction of an empirical cover), but they do not take advantage of the knowledge of $\mathcal{P}$. In practical problems, some prior knowledge can replace a large amount of empirical data. A general learning procedure that capitalizes on the structure of $\mathcal{P}$ would be a powerful tool.

Appendix

Proof of Theorem 2.2

Suppose that Assumptions 2.1 and 2.2 hold, every choice of g_{emp} learns $(\mathcal{P}, \mathcal{H})$, and L is Lipschitz in h with constant D. Take any choice of g_{cl}: fix the mapping $g_{\text{cl}}[\vec{s}(m)]$ as well as the sequences τ_n, μ_n, ν_n, $\mathcal{H}_n(\vec{x}\,')$, and ϵ_m. We want to show that this choice of g_{cl} learns $(\mathcal{P}, \mathcal{H})$.

Recall that f_{ce} picks a τ_n/n-empirical covering, $(\mathcal{H}_n(\vec{x}\,'), a_{n,\vec{x}'})$, for each n and $\vec{x}\,'(n)$. Let g be a mapping that minimizes $f_{\text{emp}}[\vec{s}\,', h]$ over the finite set $\mathcal{H}_n(\vec{x}\,')$:

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

Because $\rho_{\vec{x}'}(a_{n,\vec{x}'}(h), h) < \tau_n/n$ and L is Lipschitz in h, we have, as in Lemma 4.1 of [1], that

$$| f_{\text{emp}}[\vec{s}', a_{n,\vec{x}'}(h)] - f_{\text{emp}}[\vec{s}', h] | < D\tau_n/n.$$

Thus,

$$\begin{aligned} f_{\text{emp}}[\vec{s}', g[\vec{s}']] &= \inf_{h \in \mathcal{H}} f_{\text{emp}}[\vec{s}', a_{n,\vec{x}'}(h)] \\ &\leq \inf_{h \in \mathcal{H}} f_{\text{emp}}[\vec{s}', h] + \frac{D\tau_n}{n}. \end{aligned} \tag{12}$$

We see by (12) that g is a $D\tau_n/n$-approximate minimizer of f_{emp}. Note also that g is a choice of g_{emp}. By assumption, g learns $(\mathcal{P}, \mathcal{H})$: there is a sequence $\gamma_n \searrow 0$ such that

$$\sup_{P \in \mathcal{P}} P^n \left[\vec{s}'(n) : |\text{err}(P, g[\vec{s}'(n)]) - \text{opt}(P, \mathcal{H})| > \gamma_n\right] < \gamma_n. \tag{13}$$

By the choice of n in Step 1 of Procedure 5.2 in [1], we have by Lemma 5.1 of [1] that

$$\sup_{P \in \mathcal{P}} P^{(m-n)} \left[\vec{s}''(m-n) : \max_{h' \in \mathcal{H}_n(\vec{x}')} | f_{\text{emp}}[\vec{s}''(m-n), h'] - \text{err}(P, h') | > \mu_n\right] < \nu_n. \tag{14}$$

Thus, for any $P \in \mathcal{P}$, the following chain of inequalities holds with probability at least $1 - \gamma_n - \nu_n$:

$$\begin{aligned} \text{err}(P, g_{\text{cl}}[\vec{s}(m)]) &\leq f_{\text{emp}}[\vec{s}'', g_{\text{cl}}[\vec{s}]] + \mu_n && (15) \\ &\leq \min_{h' \in \mathcal{H}_n(\vec{x}')} f_{\text{emp}}[\vec{s}'', h'] + \mu_n + \epsilon_m && (16) \\ &\leq \min_{h' \in \mathcal{H}_n(\vec{x}')} \text{err}(P, h') + 2\mu_n + \epsilon_m && (17) \\ &\leq \text{err}(P, g[\vec{s}']) + 2\mu_n + \epsilon_m && (18) \\ &\leq \text{opt}(P, \mathcal{H}) + \gamma_n + 2\mu_n + \epsilon_m. && (19) \end{aligned}$$

Inequalities (17) and (19) follow from (14) and (13), respectively. The remaining inequalities hold for the following reasons.

(15): $g_{\text{cl}}[\vec{s}] \in \mathcal{H}_n(\vec{x}')$ and (14).

(16): g_{cl} ϵ_m-approximately minimizes f_{ce} over $\mathcal{H}_n(\vec{x}')$.

(18): $g[\vec{s}']$ is in $\mathcal{H}_n(\vec{x}')$.

Note that $\mathrm{err}(P, g_{\mathrm{cl}}[\vec{s}]) \geq \mathrm{opt}(P, \mathcal{H})$, and therefore (19) implies that

$$|\,\mathrm{err}(P, g_{\mathrm{cl}}[\vec{s}]) - \mathrm{opt}(P, \mathcal{H})\,| \leq \gamma_n + 2\mu_n + \epsilon_m$$

with probability approaching 1 uniformly in $P \in \mathcal{P}$. That is, this instance of g_{cl} learns $(\mathcal{P}, \mathcal{H})$. □

Proof of Lemma 3.1

From page 151 of [4] and Proposition A2.1 of [15], we have that

$$\sup_{P \in \mathcal{P}^*} P^n \left[\vec{s}(n) : \sup_{h \in \mathcal{H}} |\, f_{\mathrm{emp}}[\vec{s}, h] - \mathrm{err}(P, h)\,| > \epsilon \right] < 6 \left(\frac{2en}{q} \right)^q \exp(-n\epsilon^2/4)$$

when $1 \leq q = \mathrm{VCdim}(\mathcal{H}) < \infty$. Thus, we can bound $m_{f_{\mathrm{emp}}}$ by an n which satisfies

$$\frac{n\epsilon^2}{4} - q \ln(2en/q) - \ln(6/\delta) \geq 0. \tag{20}$$

Note that the derivative of (20) with respect to n is nonnegative when $n \geq 4q/\epsilon^2$. To prove Lemma 3.1, it suffices to show that some n with

$$4q/\epsilon^2 \leq n \leq \frac{q}{\epsilon^2} [20 \ln(8/\epsilon) + 8 \ln(6/\delta)] \tag{21}$$

satisfies (20). We can accomplish this by showing that

$$n' := \frac{q}{\epsilon^2} \left[10 \ln(20e/\epsilon^2) + 8 \ln(6/\delta) \right]$$

satisfies (20), because n' satisfies (21). We have

$$q \ln \frac{2en'}{q} + \ln \frac{6}{\delta} = q \ln \left[\frac{20e}{\epsilon^2} \ln \frac{20e}{\epsilon^2} \; + \; \frac{16e}{\epsilon^2} \ln \frac{6}{\delta} \right] + \ln \frac{6}{\delta} \; . \tag{22}$$

It is easy to show that, for $a, b \geq 2$, $a + b \leq ab$. Each of the terms inside the "ln[·]" in (22) exceeds 2, and therefore (22) is at most

$$q \ln \left[\frac{20e}{\epsilon^2} \ln \frac{20e}{\epsilon^2} \right] + q \ln \left[\frac{16e}{\epsilon^2} \ln \frac{6}{\delta} \right] + \ln \frac{6}{\delta} \; . \tag{23}$$

Since $\sqrt{a} > \ln a$, (23) does not exceed

$$\frac{5q}{2} \ln \left[\frac{20e}{\epsilon^2} \right] + 2q \ln \frac{6}{\delta} \; ,$$

which equals $n'\epsilon^2/4$. Thus, n' satisfies (20). □

Proof of Theorem 3.1

We first prove a useful lemma.

Lemma A.1. *Let* $\mathcal{H}^\Delta = \{|h - h'| : h, h' \in \mathcal{H}\}$. *Then*

$$\mathrm{VCdim}(\mathcal{H}^\Delta) \leq 10\mathrm{VCdim}(\mathcal{H}).$$

Proof of Lemma A.1

This is trivial when $\mathrm{VCdim}(\mathcal{H}) = 0$. Assume that $q := \mathrm{VCdim}(\mathcal{H}) \geq 1$. By Proposition A2.1 of [15],

$$\Pi_{\mathcal{H}}(n) := \max_{\vec{x} \in X^n} |\{h(\vec{x}(n)) : h \in \mathcal{H}\}| \leq \left(\frac{en}{q}\right)^q.$$

Let "$\oplus$" denote the component-wise exclusive-or operation on binary vectors. Since

$$\begin{aligned} |\{\hat{h}(\vec{x}(n)) : \hat{h} \in \mathcal{H}^\Delta\}| &= |\{h(\vec{x}(n)) \oplus h'(\vec{x}(n)) : h, h' \in \mathcal{H}\}| \\ &\leq |\{h(\vec{x}(n)) : h \in \mathcal{H}\}|^2 \\ &\leq [\Pi_{\mathcal{H}}(n)]^2, \end{aligned}$$

we have $\Pi_{\mathcal{H}^\Delta}(n) \leq [\Pi_{\mathcal{H}}(n)]^2 \leq (en/q)^{2q}$. Now, if $(en/q)^{2q} < 2^n$, then we certainly have $\Pi_{\mathcal{H}^\Delta}(n) < 2^n$; that is, $\mathrm{VCdim}(\mathcal{H}^\Delta) < n$. We can choose n to be linear as a function of q and satisfy $(en/q)^{2q} < 2^n$; with $n = 10q$,

$$2^{10q} = (32)^{2q} > (10e)^{2q}.$$

□

Let $\mathcal{H}$ have $\mathrm{VCdim}(\mathcal{H}) = q \geq 1$.

First, we bound N_{ec}. To simplify this task, choose $\tau = 1$. Note that

$$|\mathrm{err}(P, h) - \mathrm{err}(P, h')| \leq E_P|h - h'|.$$

Since the hypotheses are 0/1-valued, $\rho_{\vec{x}(n)}(h, h') < 1$ if and only if $\rho_{\vec{x}(n)}(h, h') = 0$. Thus, to bound N_{ec}, it suffices to find bounds on n for which

$$\sup_{h,h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}(n)}(h,h')=0} E_P|h - h'| > \epsilon \tag{24}$$

with probability less than δ.

Let $\mathcal{H}^\Delta = \{|h - h'| : h, h' \in \mathcal{H}\}$, and let $\vec{0}$ be a vector of all zeroes. We can rewrite Equation (24) as

$$\sup_{\hat{h} \in \mathcal{H}^\Delta \text{ s.t. } \hat{h}(\vec{x}(n)) = \vec{0}} E_P \hat{h} > \epsilon. \tag{25}$$

Let $q^\Delta = \text{VCdim}(\mathcal{H}^\Delta)$. By Lemma A.1, q^Δ is finite if q is. Also, it is true that $q^\Delta \geq 1$ if $q \geq 1$. We have from [47] that (25) holds with probability less than δ for any $P \in \mathcal{P}^*$ when n is at least

$$\frac{3}{\epsilon}\left(q^\Delta \ln(12/\epsilon) + \ln(2/\delta)\right). \tag{26}$$

Thus, (26) serves as a bound on $N_{\text{ec}}(1, \epsilon, \delta)$. Replacing ϵ and δ by $\epsilon/2$ and $\delta/2$ in (26) gives

$$N_{\text{ec}}(1, \epsilon/2, \delta/2) \leq \frac{6}{\epsilon}\left(q^\Delta \ln(24/\epsilon) + \ln(4/\delta)\right).$$

By Proposition A2.1 of [15], $H_{\text{ec}}(1, n) \leq (en/q)^q$. Also, we can bound ℓ by using Lemma 5.1 of [1] with $\lambda = 1$, $\mu = \epsilon/2$, and $\nu = \delta/2$:

$$\ell(H, \epsilon/2, \delta/2) \leq \frac{2}{\epsilon^2}(\ln H + \ln(4/\delta)).$$

Putting these bounds together,

$$\ell(H_{\text{ec}}(1, N_{\text{ec}}(1, \epsilon/2, \delta/2)), \epsilon/2, \delta/2) \leq \frac{2q}{\epsilon^2} \ln\left(\frac{6e}{q\epsilon}\left[q^\Delta \ln\frac{24}{\epsilon} + \ln\frac{4}{\delta}\right]\right) + \frac{2}{\epsilon^2}\ln\frac{4}{\delta}.$$

Using Lemma A.1 and rearranging, we have

$$\ell(H_{\text{ec}}(1, N_{\text{ec}}(1, \epsilon/2, \delta/2)), \epsilon/2, \delta/2) \leq \frac{2}{\epsilon^2}\left[q \ln\frac{60e}{\epsilon} + \ln\frac{4}{\delta}\right] + \frac{2q}{\epsilon^2}\ln\left(\ln\frac{24}{\epsilon} + \frac{1}{q^\Delta}\ln\frac{4}{\delta}\right).$$

Note that N_{ec}, the number of samples needed to achieve a good cover, is smaller than ℓ by a factor of $1/\epsilon$. Thus, if we keep only the highest-order terms in $1/\epsilon$ and $1/\delta$, we have by (3) that $\tilde{m}_{f_{\text{ce}}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H})$ is of order

$$\frac{2}{\epsilon^2}\left[q \ln\frac{60e}{\epsilon} + \ln\frac{4}{\delta}\right].$$

□

Proof of Theorem 3.2

We shall assume that $Z = [0, B]$, $d(z_1, z_2) = |z_1 - z_2|$, and L is Lipschitz in h with a constant D. Before proceeding with the proof, we introduce some notation and intermediate results.

For an n-vector $\vec{x}(n)$, let $h(\vec{x}(n)) = [h(x_1), \ldots, h(x_n)]$ and let $\mathcal{H}(\vec{x}(n))$ denote the set $\{h(\vec{x}(n)) : h \in \mathcal{H}\}$. Thus, when $Z = [0, B]$, $\mathcal{H}(x) \subseteq [0, B]$.

Consider the following pseudometric on $\mathbb{R}^n$ induced by d:

$$\rho_1(\vec{r}^{\,1}, \vec{r}^{\,2}) := \frac{1}{n} \sum_{i=1}^{n} |r_i^1 - r_i^2| \,.$$

A set of vectors $\vec{r}^{\,1}, \ldots, \vec{r}^{\,k}$ from $\mathbb{R}^n$ is said to be *ϵ-separated* if $\rho_1(\vec{r}^{\,i}, \vec{r}^{\,j}) > \epsilon$ for each $i \neq j$. We denote the size of the largest ϵ-separated subset of $\mathcal{H}(\vec{x})$ by $\mathcal{M}(\epsilon, \mathcal{H}(\vec{x}), \rho_1)$. A set $U \subseteq Z^n$ is an *ϵ-cover* for $\mathcal{H}(\vec{x}(n))$ if for each $\vec{r} \in \mathcal{H}(\vec{x})$ there is $\vec{u} \in U$ with $\rho_1(\vec{r}, \vec{u}) < \epsilon$. We denote the smallest ϵ-cover of $\mathcal{H}(\vec{x})$ by $\mathcal{N}(\epsilon, \mathcal{H}(\vec{x}), \rho_1)$. When $\mathcal{M}(\epsilon, \mathcal{H}(\vec{x}), \rho_1)$ is finite for every $\epsilon > 0$, as will always be the case when $Z = [0, B]$, the following inequalities hold (see [5]).

Lemma A.2.

$$\mathcal{M}(2\epsilon, \mathcal{H}(\vec{x}), \rho_1) \leq \mathcal{N}(\epsilon, \mathcal{H}(\vec{x}), \rho_1) \leq \mathcal{M}(\epsilon, \mathcal{H}(\vec{x}), \rho_1).$$

Finally, we have this result (Theorem 6 of [14]).

Lemma A.3. *If $\mathcal{H}(x) \subseteq [0, B]$ and $\operatorname{psdim}(\mathcal{H}) = q$ for some $1 \leq q < \infty$, then for each $0 < \epsilon \leq B$,*

$$\mathcal{M}(\epsilon, \mathcal{H}(\vec{x}(n)), \rho_1) < 2 \left(\frac{2eB}{\epsilon} \ln \frac{2eB}{\epsilon} \right)^q .$$

First, we bound N_{ec}. As in the 0/1-valued case, we note that

$$|\mathrm{err}(P, h) - \mathrm{err}(P, h')| \leq D E_P |h - h'|. \tag{27}$$

Thus, to bound N_{ec} it suffices to find bounds on when

$$\sup_{h, h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}}(h, h') < \tau/n} E_P |h - h'| > \epsilon. \tag{28}$$

Let $\mathcal{H}^\Delta = \{|h - h'| : h, h' \in \mathcal{H}\}$. We can rewrite (28) as

$$\sup_{\hat{h} \in \mathcal{H}^\Delta \text{ s.t. } E_{\vec{x}}\hat{h} < \tau/n} E_P \hat{h} > \epsilon. \tag{29}$$

Let $\epsilon > 0$. Then

$$d_\epsilon(r_1, r_2) := \frac{|r_1 - r_2|}{r_1 + r_2 + \epsilon} \tag{30}$$

is a metric on $[0, \infty) \subseteq I\!R$. Using a result from [14], we can find bounds on when

$$\sup_{\hat{h} \in \mathcal{H}^\Delta} d_\epsilon(E_{\vec{x}}\hat{h}, E_P\hat{h}) \leq 1/2, \tag{31}$$

which will give bounds on when (29) occurs. Note that if we set $\tau/n = \epsilon$, then for any $\vec{s} = (\vec{x}, \vec{y})$ for which (31) holds and any $\hat{h}$ for which $E_{\vec{x}}\hat{h} < \tau/n$,

$$E_P\hat{h} - E_{\vec{s}}\hat{h} \leq |E_P\hat{h} - E_{\vec{s}}\hat{h}| \leq \epsilon/2 + (E_P\hat{h} + E_{\vec{s}}\hat{h})/2 \leq \epsilon + E_P\hat{h}/2;$$

hence $E_P\hat{h} \leq 4\epsilon$. Thus,

$$P^n\left[\vec{s}(n) : \sup_{\hat{h} \in \mathcal{H}^\Delta \text{ s.t. } E_{\vec{x}}\hat{h} < \tau/n} E_P\hat{h} > 4\epsilon\right] \leq P^n\left[\vec{s}(n) : \sup_{\hat{h} \in \mathcal{H}^\Delta} d_\epsilon(E_{\vec{x}}\hat{h}, E_P\hat{h}) > 1/2\right].$$

By (27), $\mathcal{H}^\Delta(x) \subseteq [0, 2B]$, and Theorem 2 of [14] (attributed to Pollard), we have that $N_{\text{ec}}(\epsilon n, 4\epsilon D, \delta) \leq k$, where k is a number such that

$$4E_{P^k}(\mathcal{N}(\epsilon/16, \mathcal{H}^\Delta(\vec{x}(k)), \rho_1))\exp(-\epsilon k/128B) < \delta.$$

Now, we need to find a bound on $\mathcal{N}(\epsilon, \mathcal{H}^\Delta(\vec{x}(k)), \rho_1)$. It is easy to show that

$$\mathcal{N}(\epsilon, \mathcal{H}^\Delta(\vec{x}(k)), \rho_1) \leq [\mathcal{N}(\epsilon/2, \mathcal{H}(\vec{x}(k)), \rho_1)]^2$$

by using the triangle inequality for ρ_1. Also, we have from Lemmas A.2 and A.3 that, with $q = \text{psdim}(\mathcal{H})$,

$$\mathcal{N}(\epsilon/2, \mathcal{H}(\vec{x}(k)), \rho_1) \leq 2\left(\frac{4eB}{\epsilon}\ln\frac{4eB}{\epsilon}\right)^q.$$

Putting all of these inequalities together, we can bound $N_{\text{ec}}(\epsilon n, 4\epsilon D, \delta)$ by a k that satisfies

$$16\left(\frac{2^6eB}{\epsilon}\ln\frac{2^6eB}{\epsilon}\right)^{2q}\exp(-\epsilon k/2^7B) < \delta,$$

or

$$k > \frac{2^7 B}{\epsilon}\left[2q\ln\left(\frac{2^6 eB}{\epsilon}\ln\frac{2^6 eB}{\epsilon}\right)+\ln\frac{16}{\delta}\right].$$

Changing ϵ to $\epsilon/8D$, δ to $\delta/2$, and using $\sqrt{a} > \ln a$ with $a = (2^9 eBD/\epsilon)$,

$$N_{\rm ec}(\epsilon n/8D, \epsilon/2, \delta/2) \le \frac{2^{10}BD}{\epsilon}\left[3q\ln\frac{2^9 eBD}{\epsilon}+\ln\frac{32}{\delta}\right]+1. \tag{32}$$

Although the constants in this bound are rather large, we will see that $N_{\rm ec}$ is smaller (in order) by a factor of $1/\epsilon$ compared to ℓ.

Now we turn to bounding $H_{\rm ec}$ and ℓ. Procedure 5.1 of [1] can be altered so that the resulting empirical ϵ-cover is $\epsilon/2$-separated (see the note at the end of the proof of Lemma 5.2 in [1]). Thus, if we construct $\mathcal{H}_n(\vec{x}\,)$ in this manner, $H_{\rm ec}(\epsilon n, n)$ is at most $\mathcal{M}(\epsilon/2, \mathcal{H}(\vec{x}(n)), \rho_1)$, the size of the largest $\epsilon/2$-separated subset of $\mathcal{H}$. By Lemma A.3,

$$\mathcal{M}(\epsilon/2, \mathcal{H}(\vec{x}(n)), \rho_1) < 2\left(\frac{4eB}{\epsilon}\ln\frac{4eB}{\epsilon}\right)^q,$$

and we have

$$H_{\rm ec}(\epsilon n/8D, n) < 2\left(\frac{32eBD}{\epsilon}\ln\frac{32eBD}{\epsilon}\right)^q$$

for any n. By Lemma 5.1 of [1],

$$\ell(H, \epsilon/2, \delta/2) \le \frac{2\lambda^2}{\epsilon^2}(\ln H + \ln(4/\delta)),$$

and thus, using $\sqrt{a} > \ln a$ with $a = (32eBD/\epsilon)$,

$$\ell(H_{\rm ec}(\epsilon n/8D, N_{\rm ec}(\epsilon n/8D, \epsilon/2, \delta/2)), \epsilon/2, \delta/2) \le \frac{\lambda^2}{\epsilon^2}\left[3q\ln\frac{32eBD}{\epsilon}+2\ln\frac{8}{\delta}\right]. \tag{33}$$

Note that the bound on $N_{\rm ec}$ in (32) is smaller than ℓ by a factor of $1/\epsilon$. Thus, if we keep only the highest-order terms in $1/\epsilon$ and $1/\delta$, we have by (3) that $\tilde{m}_{f_{\rm ce}}(\epsilon, \delta, \mathcal{P}^*, \mathcal{H})$ is of order

$$\frac{\lambda^2}{\epsilon^2}\left[3{\rm psdim}(\mathcal{H})\ln\frac{32eBD}{\epsilon}+2\ln\frac{8}{\delta}\right].$$

□

Proof of Lemma 4.1

This is a trivial modification of the proof of Lemma 5.2 in [1]. The construction in that proof involves adding *some* hypothesis h_k to the cover if

$$d(h_k(x_i), M_k(i)) < \epsilon/2 \text{ at each } i \in \{1, \ldots, n\} \tag{34}$$

(refer to (16) in [1]). We shall add a constraint on the choice of h_k.

We refer to k as the *index* of $\mathcal{H}^k$. Let j^* be the index of the simplest hypothesis within $\epsilon/2$ of $M_k(i)$:

$$j^* = \min\{j : \text{ there is } h \in \mathcal{H}^j \text{ with } d(h, M_k(i)) < \epsilon/2, i \in \{1, \ldots, n\}\}.$$

Then an appropriate constraint on an h_k added to $\mathcal{A}_{k-1}$ is that it be a member of $\mathcal{H}^{\lfloor Mj^* \rfloor}$. Now, we need to check that this construction does give an M-simple empirical ϵ-cover. Take any h; assume that $h \in \mathcal{H}^j$. For some k', h is within $\epsilon/2$ of $M_{k'}(i)$ at each i. Thus, $j^{*\prime}$, the index of the simplest hypothesis within $\epsilon/2$ of $M_{k'}(i)$ at each i, must satisfy $j^{*\prime} \leq j$. By construction, the cover element added to $\mathcal{A}_{k-1}$ due to the labeling $M_{k'}$ (which is the cover element associated with h by the mapping a) has index at most $\lfloor Mj^{*\prime} \rfloor \leq \lfloor Mj \rfloor$. As in the proof of Lemma 5.2, $\rho_{\vec{x}}(a(h), h) < \epsilon$ is also satisfied. □

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 $\beta > 0$ and any $j \in \boldsymbol{N}$. Take any $P \in \mathcal{P}_j$. There is $h^* \in \mathcal{H}^j$ with

$$\text{err}(P, h^*) - \text{opt}(P, \mathcal{H}) \leq \beta. \tag{35}$$

By Step 2 of Procedure 4.1, $a_{n,\vec{x}'}(h^*) \in \mathcal{H}^{\lfloor Mj \rfloor}$. Thus, since $E_{\vec{x}}$ converges simultaneously on $(\mathcal{P}, d(\mathcal{H}^i, \mathcal{H}^i))$ for each i, if we pick n large enough then $\tau_n < \beta$ and

$$E_P d(a_{n,\vec{x}'}(h^*), h^*) < \rho_{\vec{x}}(a_{n,\vec{x}'}(h^*), h^*) + \beta < 2\beta \tag{36}$$

with probability near one. (The required size of n does not depend on P, as long as $P \in \mathcal{P}_j$.) Since L is Lipschitz in h for some $D < \infty$, whenever $\vec{s}\,'(n) = (\vec{x}\,', \vec{y}\,')$ is such that (36) holds,

$$|\mathrm{err}(P, a_{n,\vec{x}'}(h^*)) - \mathrm{err}(P, h^*)| \leq 2\beta D,$$

and hence, by (35),

$$\mathrm{err}(P, a_{n,\vec{x}'}(h^*)) - \mathrm{opt}(P, \mathcal{H}) \leq \beta + 2\beta D. \tag{37}$$

In Step 1 of Procedure 4.1, $m - n$ is chosen to be large enough that, with probability at least $1 - \nu_n$,

$$\mathrm{err}(P, g_{\mathrm{sec}}[\vec{s}\,', \vec{s}\,'']) \leq \min_{h \in \mathcal{H}_n(\vec{x}')} \mathrm{err}(P, h) + \mu_n + \epsilon_m \leq \mathrm{err}(P, a_{n,\vec{x}'}(h^*)) + \mu_n + \epsilon_m. \tag{38}$$

Thus, for $\vec{s}\,'$ such that (36) holds, we have by (37) and (38) that

$$\mathrm{err}(P, g_{\mathrm{sec}}[\vec{s}\,', \vec{s}\,'']) - \mathrm{opt}(P, \mathcal{H}) \leq \beta + 2\beta D + \mu_n + \epsilon_m \tag{39}$$

with probability approaching one (uniformly in $P \in \mathcal{P}_j$) as $n \uparrow \infty$. Clearly, since $\mu_n \searrow 0$ and $\beta > 0$ is arbitrary, we can make the right-hand side of (39) arbitrarily close to 0 with probability near one for n large enough. Thus, g_{sec} nonuniformly learns $(\mathcal{P}, \mathcal{H})$, and it does so uniformly over each $\mathcal{P}_j$. □

Proof of Theorem 4.2 This proof closely parallels that of Theorem 2.2.

Suppose that Assumptions 2.1 and 2.2 hold, there is $k(n)$ such that every choice of g_{cc} nonuniformly learns $(\mathcal{P}, \mathcal{H})$, and L is Lipschitz in h with constant D. Take any choice of g_{sec} with $M = 1$: fix the mapping $g_{\mathrm{sec}}[\vec{s}(m)]$ as well as the sequences τ_n, μ_n, ν_n, $\mathcal{H}_n(\vec{x}\,')$, and ϵ_m. We want to show that this choice of g_{sec} nonuniformly learns $(\mathcal{P}, \mathcal{H})$.

Recall that g_{sec} picks a τ_n-empirical covering, $(\mathcal{H}_n(\vec{x}\,'), a_{n,\vec{x}'})$, for each n and $\vec{x}\,'(n)$. Let g be a mapping that minimizes $f_{\mathrm{emp}}[\vec{s}\,', h]$ over the finite set $\mathcal{H}_n(\vec{x}\,') \cap \mathcal{H}^{k(n)}$. As in the proof of Theorem 2.2,

$$|\, f_{\mathrm{emp}}[\vec{s}\,', a_{n,\vec{x}'}(h)] - f_{\mathrm{emp}}[\vec{s}\,', h]\,| < D\tau_n. \tag{40}$$

Because $M = 1$, for any $h \in \mathcal{H}^{k(n)}$ there is $a_{n,\vec{x}'}(h) \in \mathcal{H}_n(\vec{x}\,')$ such that $a_{n,\vec{x}'}(h) \in \mathcal{H}^{k(n)}$ as well. Thus,

$$\{a_{n,\vec{x}'}(h) : h \in \mathcal{H}^{k(n)}\} \subseteq \{h' : h' \in \mathcal{H}_n(\vec{x}\,') \cap \mathcal{H}^{k(n)}\}.$$

This, along with (40), gives

$$\begin{aligned} f_{\text{emp}}[\vec{s}', g[\vec{s}']] &= \inf_{h' \in \mathcal{H}_n(\vec{x}') \cap \mathcal{H}^{k(n)}} f_{\text{emp}}[\vec{s}', h'] \\ &\leq \inf_{h \in \mathcal{H}^{k(n)}} f_{\text{emp}}[\vec{s}', a_{n,\vec{x}'}(h)] \\ &\leq \inf_{h \in \mathcal{H}^{k(n)}} f_{\text{emp}}[\vec{s}', h] + D\tau_n. \end{aligned} \tag{41}$$

We see by (41) that g is a $D\tau_n$-approximate minimizer of f_{emp} over $\mathcal{H}^{k(n)}$. Thus, g is a choice of g_{cc}, and so g nonuniformly learns $(\mathcal{P}, \mathcal{H})$. The remainder of the proof is identical to that of Theorem 2.2, except that there is no uniformity over $\mathcal{P}$. □

References

[1] K. L. Buescher and P. R. Kumar, "Learning by canonical smooth estimation, part I: Simultaneous estimation." Submitted to the *IEEE Transactions on Automatic Control*.

[2] 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.

[3] 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.

[4] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York: Springer Verlag, 1982.

[5] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in functional spaces," *American Mathematical Society Translations (Series 2)*, vol. 17, pp. 277–364, 1961.

[6] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society (B)*, vol. 36, pp. 111–147, 1974.

[7] G. M. Benedek and A. Itai, "Learnability by fixed distributions," in *Computational Learning Theory: Proceedings of the 1988 Workshop*, pp. 80–90, Morgan Kaufmann, San Mateo, CA, 1988.

[8] S. R. Kulkarni, "On metric entropy, Vapnik-Chervonenkis dimension, and learnability for a class of distributions," Technical Report CICS-P-160, Center for Intelligent Control Systems, M.I.T., Cambridge, MA, October 1989.

[9] S. R. Kulkarni, "Problems of computational and information complexity in machine vision and learning," Technical Report CICS-TH-298, Center for Intelligent Control Systems, M.I.T., Cambridge, MA, July 1991. Ph.D. dissertation, Department of Electrical Engineering and Computer Science.

[10] S. R. Kulkarni and M. Vidyasagar, "Learning decision rules for pattern classification under a family of probability measures." Preprint, October, 1993.

[11] L. Devroye, "Automatic pattern recognition: A study of the probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-10, no. 4, pp. 530–543, 1988.

[12] S. Ben-David, G. M. Benedek, and Y. Mansour, "A parameterization scheme for classifying models of learnability," in *Computational Learning Theory: Proceedings of the Second Workshop*, pp. 285–302, Morgan Kaufmann, San Mateo, CA, 1989.

[13] 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.

[14] 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.

[15] 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.

[16] R. M. Dudley, "Central limit theorems for empirical measures," *The Annals of Probability*, vol. 6, no. 6, pp. 899–929, 1978.

[17] R. S. Wenocur and R. M. Dudley, "Some special Vapnik-Chervonenkis classes," *Discrete Mathematics*, vol. 33, pp. 313–318, 1981.

[18] R. M. Dudley, "A course on empirical processes," in *Lecture Notes in Mathematics*, vol. 1097, Springer Verlag, 1984.

[19] E. Baum and D. Haussler, "What size net gives valid generalization?," *Neural Computation*, vol. 1, pp. 151–160, 1989.

[20] A. Macintyre and E. D. Sontag, "Finiteness results for sigmoidal neural networks," in *Proceedings of the Twenty-Fifth Annual ACM Syposium on the Theory of Computing*, pp. 325–334, Association for Computing Machinery, 1993.

[21] P. W. Goldberg and M. R. Jerrum, "Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers." Submitted to *Machine Learning*. Also in *Proceedings of the Sixth Annual ACM Conference on Computational Learning Theory*, 1993.

[22] W. Maass, "Bounds for the computational power and learning complexity of analog neural nets," in *Proceedings of the Twenty-Fifth Annual ACM Syposium on the Theory of Computing*, pp. 335–344, Association for Computing Machinery, 1993.

[23] D. Pollard, *Convergence of Stochastic Processes*. New York: Springer Verlag, 1984.

[24] D. Pollard, *Empirical Processes: Theory and Applications*. Hayward, California: Institute of Mathematical Statistics, 1990.

[25] D. Haussler, "Generalizing the PAC model: Sample size bounds from metric dimension-based uniform convergence results," in *30th Annual IEEE Symposium on Foundations of Computer Science*, pp. 40–45, 1989.

[26] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.

[27] D. D. Cox, "Approximation of least squares regression on nested subspaces," *The Annals of Statistics*, vol. 16, no. 2, pp. 713–732, 1988.

[28] 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.

[29] V. N. Vapnik, "Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures)," in *Computational Learning Theory: Proceedings of the Second Workshop*, pp. 3–21, Morgan Kaufmann, San Mateo, CA, 1989.

[30] H. White, "Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings," *Neural Networks*, vol. 3, no. 5, pp. 535–549, 1990.

[31] G. Lugosi and K. Zeger, "Nonparameteric estimation via empirical risk minimization." Submitted to the *IEEE Transactions on Information Theory*, February, 1993.

[32] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. AC-19, no. 6, pp. 716–723, 1974.

[33] V. N. Vapnik and A. Chervonenkis, "Ordered risk minimization, I," *Automation and Remote Control*, vol. 35, pp. 1226–1235, August 1974.

[34] V. N. Vapnik and A. Chervonenkis, "Ordered risk minimization, II," *Automation and Remote Control*, vol. 35, pp. 1403–1412, September 1974.

[35] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[36] K. C. Li, "Asymptotic optimality for C_P, C_L, cross-validation and generalized cross-validation: Discrete index set," *The Annals of Statistics*, vol. 15, no. 3, pp. 958–975, 1987.

[37] 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.

[38] A. R. Barron, "Approximation and estimation bounds for artificial neural networks," in *Computational Learning Theory: Proceedings of the Fourth Annual Workshop*, pp. 243–249, Morgan Kaufmann, San Mateo, CA, 1991.

[39] A. R. Barron and T. M. Cover, "Minimum complexity density estimation," *IEEE Transactions on Information Theory*, vol. IT-37, no. 4, pp. 1034–1054, 1991.

[40] S. M. Weiss and C. A. Kulikowski, *Computer Systems That Learn.* San Mateo, CA: Morgan Kaufmann, 1991.

[41] B. Efron, "Estimating the error rate of a prediction rule: Improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, pp. 316–331, June 1983.

[42] D. Haussler, "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework," *Artificial Intelligence*, vol. 36, no. 2, pp. 117–222, 1988.

[43] 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.

[44] D. Aldous and U. Vazirani, "A Markovian extension of Valiant's learning model," in *31st Annual IEEE Symposium on Foundations of Computer Science*, pp. 392–404, October 1990.

[45] A. B. Nobel, "Evaluating the performance of a simple inductive procedure in the presence of overfitting error," in *Computational Learning Theory: Proceedings of the Fourth Annual Workshop*, pp. 267–274, Morgan Kaufmann, San Mateo, CA, 1991.

[46] A. B. Nobel, "On uniform laws of averages," Technical Report 81, Stanford University, Department of Statistics, Stanford, CA, June 1992. Ph.D. dissertation, Department of Electrical Engineering.

[47] M. Anthony and N. Biggs, *Computational Learning Theory.* Cambridge Tracts in Computer Science, Cambridge: Cambridge University Press, 1992.