# Learning by Canonical Smooth Estimation, Part I: Simultaneous Estimation*†

Kevin L. Buescher‡ and P. R. Kumar§

**Abstract**

This paper examines the problem of learning from examples in a framework that is based on, but more general than, Valiant's Probably Approximately Correct (PAC) model for learning. In our framework, the learner observes examples that consist of sample points drawn and labeled according to a fixed, unknown probability distribution. Based on this empirical data, the learner must select, from a set of candidate functions, a particular function, or "hypothesis," that will accurately predict the labels of future sample points. The expected mismatch between a hypothesis' prediction and the label of a new sample point is called the hypothesis' "generalization error."

Following the pioneering work of Vapnik and Chervonenkis, others have attacked this sort of learning problem by finding hypotheses that minimize the relative frequency-based *empirical* error estimate. We generalize this approach by examining the "simultaneous estimation" problem: When does *some* procedure exist for estimating the generalization error of *all* of the candidate hypotheses, simultaneously, from the *same* labeled sample? We demonstrate how one can learn from such a simultaneous error estimate and propose a new class of estimators, called "smooth estimators," that, in many cases of interest, contains the empirical estimator. We characterize the class of simultaneous estimation problems solvable by a smooth estimator and give a canonical form for the smooth simultaneous estimator.

*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

In decision and control, one often faces the problem of predicting the outcome of an event from measured variables, on the basis of several recorded observations. For example, given the measurements from several sensors, has a particular actuator failed? We may attempt to solve this problem by forming a *hypothesis* about the dependence of the event's outcome on the measured variables. For instance, "if the measurement from sensor A is low, and the measurement from sensor B is high, then the actuator has failed." We then hope that this hypothesis will *generalize* well in that it accurately predicts an event's outcome from the measured variables. In a word, our goal is to "learn." That is, we want to find a hypothesis, based on empirical data, that generalizes well.

This paper and its sequel ([1]) explore new techniques for learning. This paper focuses on a problem related to learning: "simultaneous estimation." In this problem, the goal is to estimate how well each hypothesis in a given set of candidate hypotheses will generalize. Before discussing the contributions and organization of this paper, we first elaborate on the notions of learning and simultaneous estimation.

## 1.1 Learning

In [2], Valiant proposed a simple model for the process of learning from examples that has been widely adopted and extensively studied by others. For reasons that will soon become apparent, this model has been dubbed the *Probably Approximately Correct* (PAC) model for learning ([3]). The model for learning we treat in this paper is more general, but we discuss the PAC model first to introduce some ideas and notation.

We first describe how the empirical data is gathered in the PAC model. Then we shall define an appropriate measure of how well a hypothesis generalizes. Let $X$ be a point set; $X$ represents quantities that we can measure. Assume that a fixed, unknown probability distribution $Q$ on $X$ determines the measurements we observe, and a fixed, unknown 0/1-valued function $c : X \mapsto \{0,1\}$ determines the outcome of the event we are studying. Thus, the observations that we make are labeled sample points (or examples) of the form

$(x_i, c(x_i))$, where each $x_i$ is drawn according to $Q$. Also, assume that the $x_i$ are drawn independently of one another. After $m$ such observations, our data consists of the *labeled sample* $(\vec{x}(m), c(\vec{x}(m)))$.

Suppose that we restrict ourselves to selecting a hypothesis from a set $\mathcal{H}$ of 0/1-valued functions. For each $h \in \mathcal{H}$, define the *generalization error* of $h$ as the expected value of $|h(x) - c(x)|$ when $x$ is drawn by $Q$:

$$\mathrm{err}(Q, c, h) := E_Q|h - c| = \int_X |h(x) - c(x)|Q(dx).$$

Note that, since $h$ and $c$ are 0/1-valued, $\mathrm{err}(Q, c, h) = Q[h(x) \neq c(x)]$. Thus, $\mathrm{err}(Q, c, h)$ is the expected error that we will make using $h$, if a new point $x$ is drawn by $Q$. When $\mathrm{err}(Q, c, h)$ is small, $h$ accurately predicts $c(x)$ from $x$.

In this terminology, we can describe the PAC model of learning as follows. Suppose that $c$ is a fixed, unknown member of a known set $\mathcal{C}$, and $\mathcal{C} \subseteq \mathcal{H}$. Based on a labeled sample $(\vec{x}, c(\vec{x}))$ generated by a fixed, unknown $Q$, we want to find an $h \in \mathcal{H}$ that has small error, i.e., $h$ is "approximately correct." Due to statistical fluctuations which may provide a poor sample, $\vec{x}$ may occasionally give us very little information about $c$. Hence, we will settle for finding only an $h$ that, with high probability, has small error. This justifies the terminology "probably approximately correct."

There are many ways to select $h$ based on the labeled sample $(\vec{x}(m), c(\vec{x}(m)))$. For instance, we could take $h$ to be a hypothesis that is consistent with the data, i.e.,

$$h(x_i) = c(x_i), \text{ for each } i \in \{1, \ldots, m\}.$$

In general, however, any sequence $g = \{g_m\}_{m=1}^{\infty}$ of mappings such that

$$g_m : [\vec{x}(m), c(\vec{x}(m))] \mapsto \mathcal{H} \text{ for each } m \geq 1, \vec{x}(m) \in X^m, \text{ and } c \in \mathcal{C}$$

is a candidate for learning. We say that the pair $(\mathcal{C}, \mathcal{H})$ is *PAC-learnable* if there is *some* $g$ such that $\mathrm{err}(Q, c, g_m[\vec{x}(m), c(\vec{x}(m))])$ approaches zero with probability approaching one as $m$ increases (regardless of the particular values of $c \in \mathcal{C}$ and $Q$).

The terminology "PAC-learnable" sometimes denotes learnability by a function that uses a reasonable (i.e., polynomial) number of samples. More commonly, this terminology implies that the candidate learning function is required to be an algorithm that uses a reasonable amount of computational resources. However, in this paper, we do not address these issues of "efficient learning."

Due to its simplicity, the PAC model has been exhaustively studied. Also, it is amenable to analysis when one takes issues of computational complexity into account (see [4], [5], and [6] and the references therein). Many researchers have generalized the PAC framework in an attempt to model more realistic situations; a partial list of these contributions is [7]–[22]. To motivate and illustrate the extension of the PAC model considered in this paper (which resembles that of [21]), we give the following example.

Consider, for example, a process control situation, where, to avoid waste and fill orders promptly, one wishes to predict the yield, in percent, of a certain chemical reaction. In this case, the observed outcome of the event of interest, $y$, takes a value in the set $Y = [0, 100] \subseteq \mathbb{R}$ (as opposed to just a binary value). Also, suppose that one attempts to predict $y$ by measuring the concentration of a certain reagent from a supplier; thus, $X = [0, \infty)$. This concentration varies randomly from day to day.

There are two reasons why it is unrealistic to assume that $y = c(x)$, i.e., $y$ is exactly equal to $c(x)$ for some function $c$, as in the PAC case. First, the observed values of $x$ and $y$ will probably differ from the true concentration and yield; there will be measurement noise. Second, there may be no functional relation between the true concentration and yield at all, since the yield may depend on other, unmeasured, variables, such as the ambient temperature. Thus, a probability distribution, $P$, on $X \times Y$ more adequately models the observations.

One may also have some prior knowledge about the possible values that $P$ can take. For instance, one may know bounds on the mean and variance of the daily fluctuations in the concentration of the supplied reagent, $x$. Additionally, physical principles may dictate that the expected value of $y$ given $x$ is monotone increasing as a function of $x$. In other words,

one may know that $P \in \mathcal{P}$ for some class of distributions $\mathcal{P}$.

Finally, the hypothesis one proposes for predicting $y$ from $x$ will be some function $h : X \mapsto [0, 100]$. We shall call the range of the hypotheses $Z$ to distinguish it from $Y$. To quantify how well $h(x)$ predicts $y$, one might choose to use a general *loss criterion*, $L : Z \times Y \mapsto \mathbb{R}$, rather than $|h(x) - y|$ and define the error as

$$\mathrm{err}(P, h) := E_P L(h(x), y) = \int_{X \times Y} L(h(x), y)\ P(dx, dy). \tag{1}$$

For example, if one is more concerned with filling orders quickly than wasting reagents, one might select

$$L(z, y) = \begin{cases} (z - y)^2 & \text{if } z \geq y \\ 0 & \text{otherwise} \end{cases},$$

so that no penalty is associated with underestimating the yield.

To summarize the discussion above, the model for learning considered in this paper involves four realistic modifications of the PAC model. First, the observations and hypotheses take values in sets $Y$ and $Z$, not just $\{0, 1\}$. Second, the observations are assumed to be generated by a probability distribution $P$ on $X \times Y$, instead of a distribution on $X$ and a concept $c$. Thus, the observations may be noisy. Third, we assume that there is some class $\mathcal{P}$ that models prior knowledge about the possible values of $P$. Lastly, we use a loss criterion of a general form, $L$, to measure how well a hypothesis generalizes.

In analogy with the PAC-learning case, we say that the pair $(\mathcal{P}, \mathcal{H})$ is *learnable* if there is some means of picking a hypothesis whose error approaches the optimum over $\mathcal{H}$ as the number of labeled samples increases.

## 1.2 Smooth Simultaneous Estimation

In this paper, we concentrate on the problem of *simultaneous estimation*: When is it possible to estimate the error of each hypothesis in a set $\mathcal{H}$ from the *same* labeled sample? If we can do so, we can then learn by simply choosing the hypothesis with the least *estimated* error, since this hypothesis will also have nearly the least *true* error.

An estimator is any sequence $f = \{f_m\}_{m=1}^{\infty}$ of mappings such that

$$f_m : [(\vec{x}(m), \vec{y}(m)), h] \mapsto \mathbb{R} \text{ for each } m \geq 1,\ \vec{x}(m) \in X^m,\ \vec{y}(m) \in Y^m, \text{ and } h \in \mathcal{H}.$$

(Henceforth, we shall refer to $f$ itself as a mapping and suppress the subscript $m$ in $f_m$.) For instance, we could take $f$ to be the empirical estimate of $\mathrm{err}(P, \mathcal{H})$ based on the labeled sample $(\vec{x}(m), \vec{y}(m))$,

$$f_{\mathrm{emp}}[(\vec{x}(m), \vec{y}(m)), h] := \frac{1}{m} \sum_{i=1}^{m} L(h(x_i), y_i) \tag{2}$$

(refer to (1)). We say that $(\mathcal{P}, \mathcal{H})$ is *simultaneously estimable* if there is *some* $f$ such that $f[(\vec{x}(m), \vec{y}(m)), h]$ approaches $\mathrm{err}(P, h)$ *simultaneously* over $h \in \mathcal{H}$ with probability approaching one as $m$ increases (regardless of the particular value of $P \in \mathcal{P}$).

Most existing results on learning in frameworks other than the PAC model fix attention only on pairs $(\mathcal{P}, \mathcal{H})$ for which $f_{\mathrm{emp}}$ is a simultaneous estimator, and thus involve finding a hypothesis that minimizes $f_{\mathrm{emp}}$ (see [23], [15], [16], [21], [24], and [22]). Such results are based on, or heavily influenced by, the pioneering work of Vapnik and Chervonenkis ([25], [26], and [23]), which provides necessary and sufficient conditions for $f_{\mathrm{emp}}$ to simultaneously estimate $(\mathcal{P}, \mathcal{H})$. In addition, Dudley ([27],[28]), Pollard ([29],[30]), Talagrand ([31]), Haussler ([15],[21]), and others ([32], [33]) have found useful conditions that are sufficient for $f_{\mathrm{emp}}$ to simultaneously estimate $(\mathcal{P}, \mathcal{H})$.

This paper goes beyond such "empirical error minimization" and investigates simultaneous estimation problems that can be solved by estimators satisfying only a "smoothness" constraint. A *smooth* estimator is one which gives similar error estimates for hypotheses that are close at all the sample points (i.e., $f[(\vec{x}(m), \vec{y}(m)), h] \approx f[(\vec{x}(m), \vec{y}(m)), h']$ when $h(x_i) \approx h'(x_i)$, $i \in \{1, \ldots, m\}$). As one would expect, for many choices of $L$, e.g., loss criteria that are "Lipschitz in $h$," the empirical estimator is smooth. This smoothness constraint excludes certain pathological pairs $(\mathcal{P}, \mathcal{H})$ from consideration and rules out overly complex estimators. By considering only smooth estimators, we hope to concentrate on "natural" pairs $(\mathcal{P}, \mathcal{H})$.

## 1.3 Contributions and Organization of this Paper

The foremost contribution of this paper is the concept of "smooth simultaneous estimability." This notion provides us with weaker sufficient conditions for learning than those based on

the simultaneous estimation of $(\mathcal{P}, \mathcal{H})$ by the empirical estimator. Also, it prompts us to consider learning procedures of a nonstandard form.

We show that smooth simultaneous estimation is possible if and only if $(\mathcal{P}, \mathcal{H})$ can be "empirically covered." An empirical cover for $\mathcal{H}$ is a set of hypotheses that is constructed from a sample so as to cover (and thus approximate) $\mathcal{H}$. We say that $(\mathcal{P}, \mathcal{H})$ can be *empirically covered* if it is likely that any two hypotheses that are close at all the sample points have similar errors.

We also present an explicit form for an estimator that serves as a smooth simultaneous estimator, whenever one exists. This "canonical estimator" takes the form of a two-step procedure. First, it uses a portion of the sample to determine a finite number of hypotheses that form an empirical cover for $\mathcal{H}$. Next, it estimates the error of all of the hypotheses in the cover using the remainder of the sample. The canonical estimator gains an advantage over the empirical estimator by explicitly using the structure of $\mathcal{H}$. In fact, we give an example where the canonical estimator simultaneously estimates and the empirical estimator does not.

Whenever smooth simultaneous estimation is possible, we can learn by simply selecting any hypothesis in $\mathcal{H}$ that minimizes the error as estimated by the canonical estimator. In a companion to this paper ([1]), we examine this "canonical learning procedure" in detail and compare it to empirical error minimization. We find that the canonical learner is, in many respects, superior to empirical error minimization. In [1], we also show how the ideas underlying this canonical learner can be used to pick a hypothesis of appropriate "complexity" from a hierarchy a hypothesis classes.

The rest of this paper is organized as follows. In Section 2, we define learnability and discuss some common special cases of our general framework. We define simultaneous estimability and show that it suffices for learning in Section 3. In Section 4, we motivate and define "smoothness." We also show that, in many cases, the class of smooth estimators contains the empirical estimator as a special case. Section 5 characterizes smooth simultaneous estimation in terms of whether or not $(\mathcal{P}, \mathcal{H})$ can be empirically covered. Also, we define

the notion of an empirical cover in Section 5 and use it to develop the canonical smooth simultaneous estimator. We summarize the results of this paper and preview those of [1] in Section 6.

# 2 Learning

Motivated by the informal description in the previous section, we now give a rigorous definition of "learning." Also, in this section we introduce notation and conventions that are used throughout this paper and its sequel, [1].

## 2.1 Learnability

We model the learning problem as follows. Let $X$, $Y$, and $Z$ be point sets, and 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 a vector of $m$ points from $S$ a *labeled sample* of length $m$,

$$\vec{s}(m) = [(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)].$$

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

Given $\vec{s}(m)$, 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 *hypothesis* $h : X \mapsto Z$ such that $h(x)$ is, on average, "close" to $y$ when $(x, y)$ is drawn by $P$. We quantify this closeness by a known *loss criterion*, $L : Z \times Y \mapsto \mathbb{R}$. When the label of $x$ is $y$, $L(h(x), y)$ represents the loss, or cost, of hypothesizing $h(x)$ as the label. The *error* of $h$ with respect to $P$ is then the expected value of $L(h(x), y)$ for another labeled sample point drawn by $P$:

$$\mathrm{err}(P, h) := E_P L(h(x), y) = \int_{X \times Y} L(h(x), y)\ P(dx, dy).$$

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

We now restate what learning entails: given $\mathcal{P}$, $L$, and $\vec{s}(m)$ drawn by an unknown $P \in \mathcal{P}$, find a hypothesis $h$ such that $\mathrm{err}(P,h)$ 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 $P^m$-probability near 1, a good hypothesis. (Note that our ability to find a good hypothesis depends on $\vec{s}(m)$, thus $P^m$, 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 $\mathcal{H}$,

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

This achievable level of error may differ greatly from the best possible error, $\mathrm{opt}(P,\mathcal{H}^*)$, where $\mathcal{H}^*$ represents all (measurable) functions from $X$ to $Z$.

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 $g[\vec{s}(m)]$ approaches the optimum as the number of samples, $m$, increases. We do not allow a candidate learning function $g$ to depend on $P$. This enforces the realistic constraint that we do not know the underlying distribution of the data. However, the set $\mathcal{P}$ models prior knowledge about the underlying distribution, and the choice of $g$ may reflect this. The noise-free case discussed in the next subsection provides an example of incorporating prior knowledge into $\mathcal{P}$.

We summarize the foregoing discussion with the following definition.

**Definition 2.1.** *$(\mathcal{P},\mathcal{H})$ is* learnable *if there is a mapping $g : \vec{s} \mapsto \mathcal{H}$ satisfying*

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

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

Throughout this paper, "$z_m \to 0$ in probability uniformly in $P \in \mathcal{P}$" means that, for any

given $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}} P^m[\vec{s}(m) : |z_m| > \epsilon] \to 0 \text{ as } m \uparrow \infty.$$

That is, for any given $\epsilon, \delta > 0$, there is $M(\epsilon, \delta) < \infty$ which does *not* depend on $P$ such that $P^m[|z_m| > \epsilon] < \delta$ for any $m \geq M(\epsilon, \delta)$. In the context of learning, $\epsilon$ and $\delta$ are parameters related to the relative *accuracy* of the hypothesis $g[\vec{s}]$ and to our *confidence* that $g[\vec{s}]$ has this property, respectively. If $g$ learns $(\mathcal{P}, \mathcal{H})$ and we take a large enough number of samples, we can be confident (with probability at least $1 - \delta$) that the error of $g[\vec{s}]$ lies $\epsilon$-close to the optimum, *regardless* of the particular value of $P$.

In these papers we do not address measurability issues. We leave it as an open problem to determine whether the results in these papers still hold, under some reasonable restrictions, when measurability is taken into account.

## 2.2 Special Cases

To simply relate the ideas in this paper to previous work, we identify several special cases. When $Y = Z = \{0, 1\}$ and $L(z, y) = |z - y|$, we say that we are dealing with the *0/1-valued* case. Note that then

$$\text{err}(P, h) = P[(x, y) : y \neq h(x)].$$

This situation arises in the field of pattern recognition, where the goal is to find a "discrimination function" $(h(x))$ that correctly classifies points $x$ as belonging to either the set where $y$ is most likely 1 or most likely 0 (see [34], [35], and [36]). Because of its simplicity, the 0/1-valued case is the one most often addressed in the literature on computational learning theory. The *finite-valued* case, where $Y = Z = \{0, 1, \ldots, j\}$ for some fixed $j$, shares many of the properties of the 0/1-valued case (see [19] and [24]).

A third special case of interest arises when the data is generated by sample points drawn from a probability distribution $Q$ on $X$ and labeled by a deterministic function, $c : X \mapsto Y$. We refer to this as the *noise-free* case and think of the unknown function $c$ as the *concept* to be learned. We assume that $c$ is a member of a known set of functions, $\mathcal{C}$, and $Q$ is a member of a known set of distributions, $\mathcal{Q}$. Note that the distribution $P$ on $S = X \times Y$ is

induced from $Q$ and $c$:

$$P(A \times B) = \int_A 1\,(c(x) \in B)\, Q(dx)$$

for all (measurable) $A \subseteq X$ and $B \subseteq Y$. A triple, $(\mathcal{Q}, \mathcal{C}, \mathcal{H})$, thus characterizes the learning problem in this case. The relevant error measure is

$$\mathrm{err}(Q, c, h) := E_Q L(h(x), c(x)) = \int_X L(h(x), c(x))\; Q(dx),$$

and $\mathrm{opt}(Q, c, \mathcal{H})$ is defined analogously. The available data consists of the sample $\vec{x}(m)$ and its labeling by $c$,

$$c(\vec{x}(m)) = [c(x_1), c(x_2), \ldots, c(x_m)].$$

We say that $(\mathcal{Q}, \mathcal{C}, \mathcal{H})$ is learnable if there is $g : (\vec{x}, c(\vec{x})) \mapsto \mathcal{H}$ satisfying

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

in probability as $m \uparrow \infty$ uniformly in $Q \in \mathcal{Q}$ and $c \in \mathcal{C}$.

The fourth special case occurs when $P$ (or $Q$) is unrestricted, and $\mathcal{P}$ (or $\mathcal{Q}$) includes all possible probability distributions. We shall refer to this as the *distribution-free* case, and distinguish it by using $\mathcal{P}^*$ and $\mathcal{Q}^*$ in place of $\mathcal{P}$ and $\mathcal{Q}$, respectively. The requirement for distribution-free learnability is more stringent, and learnability is easier to characterize in the distribution-free case.

The case considered in [2] represents the intersection of many of the restrictions discussed above. This 0/1-valued, distribution-free, and noise-free case has come to be known as the *Probably Approximately Correct* (or PAC) case.

## 2.3 Other Conventions

At its inception in [2], the definition of learnability required that a candidate learning function be computable using a feasible amount of resources, e.g., computable in time polynomial in the size of the instances. Models for such "efficient learning" have been much refined and studied in the recent literature, especially in the PAC case. This has given rise to a field known as *computational* learning theory. See [37], [4] and [6] for details; for an accessible and

compelling account, see [5]. Because of the general learning framework under consideration here, such computational issues lie beyond the scope of this paper.

# 3 Simultaneous Estimation of Errors

Finding a useful condition that is both necessary and sufficient for learnability in the general framework of Section 2 presents a very difficult challenge. The most general cases for which such characterizations have been found are those related to $(\mathcal{P}^*, \mathcal{H})$ and $(\{Q\}, \mathcal{C}, \mathcal{C})$ in the 0/1-valued (or finite-valued) case (see [4], [9], [38], and [24], [39]). These are rather special cases. Also, a characterization for learnability has limited practical utility unless a viable learning procedure accompanies it. For these reasons, we restrict ourselves to investigating a particular technique for learning: estimating the error of each hypothesis and then picking a hypothesis with minimal estimated error.

In the remainder of this paper, we concentrate on the problem of how to estimate the error of all hypotheses *simultaneously*. In this section, we show that, to learn, it suffices to find such "simultaneous error estimates."

Suppose that $P$ is fixed and we have a real-valued function, $f$, such that $f[h]$ is close to $\mathrm{err}(P, h)$ for all $h \in \mathcal{H}$, i.e.,

$$\sup_{h \in \mathcal{H}} | \, f[h] - \mathrm{err}(P, h) \, | \text{ is small.} \tag{3}$$

Then, any $h^*$ with $f[h^*]$ close to $\inf_{h \in \mathcal{H}} f[h]$ will have $\mathrm{err}(P, h^*)$ close to $\mathrm{opt}(P, \mathcal{H})$. In other words, if (3) holds, we can find a near-optimal hypothesis by minimizing $f[h]$ over $h$. For the purpose of learning, we are interested in classifying pairs $(\mathcal{P}, \mathcal{H})$ in terms of whether such an $f$ can be found which approximates $\mathrm{err}(P, \cdot)$ from labeled samples picked according to $P \in \mathcal{P}$. We will not allow a candidate simultaneous estimator to depend on $P$ explicitly, but we entertain all other possibilities that may take advantage of the knowledge of $\mathcal{P}$ and $\mathcal{H}$.

**Definition 3.1.** $(\mathcal{P}, \mathcal{H})$ *is* simultaneously estimable *if there is a mapping $f$ from data of the form $[\vec{s}(m), h]$ to $\mathbb{R}$ with*

$$\sup_{h \in \mathcal{H}} | \, f[\vec{s}(m), h] - \mathrm{err}(P, h) \, | \to 0 \tag{4}$$

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

(Note: Traditionally, one would express (4) in words as "$f[\vec{s}, h]$ converges to $\mathrm{err}(P, h)$ *uniformly* over $\mathcal{H}$." Hence, "uniformly estimable" might seem like more proper terminology; however, we prefer to use "simultaneous" to avoid confusion with the uniformity over $\mathcal{P}$.)

An important feature of a simultaneous error estimate is that we do not need to use any further special information about $f$ in order to learn. For instance, we do not need to know how fast the convergence in (4) takes place. Let $\mathcal{H}(f, \epsilon, \vec{s})$ be the set of hypotheses that nearly "minimize" $f$ over $\mathcal{H}$ at $\vec{s}$:

$$\mathcal{H}(f, \epsilon, \vec{s}) := \left\{ h \in \mathcal{H} : f[\vec{s}, h] \leq \epsilon + \inf_{h' \in \mathcal{H}} f[\vec{s}, h'] \right\}.$$

**Definition 3.2.** *$g$* approximately minimizes $\boldsymbol{f}$ *if, for some sequence* $\epsilon_m \searrow 0$,

$$g[\vec{s}(m)] \in \mathcal{H}(f, \epsilon_m, \vec{s}(m)) \text{ for all } \vec{s}(m) \in S^m \text{ and } m \in \boldsymbol{N}\,. \tag{5}$$

(For a given sequence $\epsilon_m$, we say that $g$ $\epsilon_m$- approximately minimizes $f$ if (5) holds.)

**Definition 3.3.** $(\mathcal{P}, \mathcal{H})$ *is* learnable by minimizing $\boldsymbol{f}$ *if any $g$ that approximately minimizes $f$ learns* $(\mathcal{P}, \mathcal{H})$.

The next result follows easily from the definitions.

**Lemma 3.1.** $(\mathcal{P}, \mathcal{H})$ *is learnable by minimizing $f$ if $f$ simultaneously estimates* $(\mathcal{P}, \mathcal{H})$.

As in [21], Lemma 3.1 remains true even if $g$ selects a near-minimal hypothesis not necessarily for each $\vec{s}$, but with probability approaching 1 uniformly in $P \in \mathcal{P}$.

In the sequel to this paper ([1]), we shall make use of these definitions in examining the properties of learning procedures based on the simultaneous estimation scheme to be developed here.

# 4 Smooth Estimators

We say that an estimator is *smooth* if it assigns similar error estimates to hypotheses that are "close on the samples." We are motivated to restrict our attention to simultaneous

estimators that are smooth by the fact that some choices of $(\mathcal{P}, \mathcal{H})$ exist that are simultaneously estimable only by estimators of a rather pathological or contrived form. For instance, hints about the value of $\mathrm{err}(P, h)$ can sometimes be encoded in $h(\vec{x})$ (see Example 4.1). Not only do these "unnatural" cases make it difficult to characterize the class of simultaneously estimable $(\mathcal{P}, \mathcal{H})$, but the resulting estimators might be too complex to be of use. By constraining the estimators to be smooth, we obtain (in Section 5) a natural characterization for a large class of simultaneously estimable $(\mathcal{P}, \mathcal{H})$. Also, in Section 5 we show that there is a canonical smooth simultaneous estimator that has a simple form.

The notion of a pseudometric will be used repeatedly in our discussion of smoothness.

**Definition 4.1.** *Let $T$ be a set. A mapping $\rho : T \times T \mapsto [0, \infty)$ is a* pseudometric *on $T$ if for any $t_1, t_2, t_3 \in T$, $\rho$ satisfies*

*(i)* $\rho(t_1, t_2) \geq 0$ *(nonnegativity)*

*(ii)* $\rho(t_1, t_2) = \rho(t_2, t_1)$ *(symmetry)*

*(iii)* $\rho(t_1, t_3) \leq \rho(t_1, t_2) + \rho(t_2, t_3)$ *(triangle inequality)*

*If, in addition, $\rho(t_1, t_2) = 0 \Rightarrow t_1 = t_2$, $\rho$ is called a* metric.

Roughly speaking, a smooth estimator is one that is nearly continuous in $h(\vec{x})$. To speak of continuity, we first must establish how we intend to measure the distance between points $h(\vec{x}(m)) \in Z^m$. Let $d$ be a pseudometric on $Z$. Note that the empirical estimate of $E_P d(h, h')$ based on the sample $\vec{x}(m)$,

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

is a pseudometric on both $\{h(\vec{x}(m)) : h \in \mathcal{H}\}$ and $\mathcal{H}$ . Accordingly, we call $\rho_{\vec{x}}(h, h')$ the *empirical distance* between $h$ and $h'$. We can now properly define smoothness. To justify our concept of smoothness, we need some additional definitions first.

For each fixed $s = (x, y)$,

$$\rho_{L,s}(h, h') := |L(h(x), y) - L(h'(x), y)| \tag{6}$$

is also a pseudometric on $\mathcal{H}$. A case of particular interest arises when $\rho_{L,s}(h,h')$ is small whenever $d(h(x),h'(x))$ is small, in which case we say that $L$ is *Lipschitz* in $h$.

**Definition 4.2.** *$L$ is* Lipschitz *in $h$ if there is $D < \infty$ such that*

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

We now note two situations in which $L$ is Lipschitz in $h$. When $Y = Z$, it is common to measure the discrepancy between $h(x)$ and $y$ by taking $L$ to be a metric. It is then logical to use $d \equiv L$, and we have as a consequence of the triangle inequality for such $L$ that $L$ is Lipschitz in $h$ with $D = 1$.

In the finite-valued case, where $|Z| < \infty$, if $d$ is a metric then any $L$ that is bounded is Lipschitz in $h$. That is, if there is $\beta < \infty$ such that

$$|L(z,y)| \leq \beta \text{ for all } z \in Z \text{ and } y \in Y,$$

then $L$ is Lipschitz in $h$. To see this, note that there is $\alpha > 0$ such that

$$z_1 \neq z_2 \Rightarrow d(z_1,z_2) > \alpha.$$

Thus, if $h(x) \neq h'(x)$,

$$\rho_{L,s}(h,h') \leq |L(h(x),y)| + |L(h'(x),y)| \leq 2\beta < \frac{2\beta}{\alpha} d(h(x),h'(x));$$

on the other hand, if $h(x) = h'(x)$, $\rho_{L,s}(h,h') = 0$. It follows that $L$ is Lipschitz in $h$ with $D = \frac{2\beta}{\alpha}$.

Let us consider what form of smoothness is appropriate for an estimator by concentrating on the case in which $L$ is Lipschitz in $h$. Recall the definition of $\rho_{L,s}$ in (6) and note that

$$\begin{aligned} |\operatorname{err}(P,h) - \operatorname{err}(P,h')| &\leq E_P \rho_{L,s}(h,h') \\ &\leq D\, E_P d(h,h'). \end{aligned}$$

Thus, $\operatorname{err}(P,h)$ is close to $\operatorname{err}(P,h')$ whenever $E_P d(h,h')$ is small, i.e., we need to consider only estimators whose value varies slowly with $h$ when $E_P d(h,h')$ measures the distance

between hypotheses. However, since we do not know $P$, we will use the empirical estimate of $E_P d(h, h')$, $\rho_{\vec{x}(m)}(h, h')$, to measure this distance. We can force small variations in $h$ with respect to this empirical distance to give only small variations in $f$ by requiring that $f$ be continuous in $h(\vec{x}(m))$ uniformly over $S^m$. That is, we could insist that for each $\theta > 0$ there is $\sigma(\theta, m) > 0$ such that

$$\rho_{\vec{x}}(h, h') < \sigma(\theta, m) \Rightarrow | f[\vec{s}(m), h] - f[\vec{s}(m), h'] | \leq \theta. \tag{7}$$

However, we weaken this condition (and, hence, broaden the class of estimators under consideration) by requiring only that (7) hold with probability near 1 for large $m$.

**Definition 4.3.** *$f$ is* smooth *over $(\mathcal{P}, \mathcal{H})$ if, for each fixed $\theta > 0$, there is $\sigma(\theta, m) > 0$ such that the probability that*

$$\sup_{h, h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}(m)}(h, h') < \sigma(\theta, m)} | f[\vec{s}(m), h] - f[\vec{s}(m), h'] | > \theta$$

*converges to 0 as $m \uparrow \infty$ uniformly in $P \in \mathcal{P}$.*

We use this definition of smoothness regardless of whether or not $L$ is Lipschitz in $h$, and we say that $(\mathcal{P}, \mathcal{H})$ can be *smoothly simultaneously estimated* if there is a smooth estimator that simultaneously estimates $(\mathcal{P}, \mathcal{H})$.

Note that smoothness is a property of the *estimator*. Imposing the condition of smoothness on an estimator does not directly place any "smoothness" constraints on the individual distributions or hypotheses in $(\mathcal{P}, \mathcal{H})$. The term "smooth" often carries connotations of differentiability; this clearly does not apply here.

**Example 4.1:** A nonsmooth simultaneous estimator.

Consider the noise-free case, with $X = Y = Z = [0, 1] \subset I\!R$; $\mathcal{Q} = \{Q\}$, where $Q$ is uniform on $[0, 1]$; $\mathcal{C} = \{c\}$, where $c(x) \equiv 0$; and $L(u, v) = d(u, v) = |u - v|$. Let $\mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1$, where $\mathcal{H}_0$ and $\mathcal{H}_1$ are defined as follows.

$$\mathcal{H}_1 = \{h_i^1 : i \in \boldsymbol{N}\}, \text{ where } h_i^1(x) = \begin{cases} \frac{1}{2^i} & \text{if } x \in [0, 1/2) \\ 1 & \text{otherwise} \end{cases},$$

and

$$\mathcal{H}_0 = \{h^0_{i,A} : i \in \boldsymbol{N}\,, A \subset [1/2,1], \text{ and } |A| < \infty\},$$

where

$$h^0_{i,A}(x) = \begin{cases} \frac{1}{2i+1} & \text{if } x \in [0,1/2) \\ 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} .$$

Let $\hat{x}(\vec{x})$ be the smallest component of $\vec{x}$. If $\hat{x}(\vec{x}) \in [0,1/2)$, we can determine $\mathrm{err}(Q,c,h) = E_Q h$ by examining $h(\hat{x}(\vec{x}))$. In fact, since $Q^m[\,\hat{x}(\vec{x}(m)) \in [0,1/2)\,]$ approaches 1 as $1-2^{-m}$, the following is a simultaneous estimator for $(\{Q\},\{c\},\mathcal{H})$:

$$f[(\vec{x},c(\vec{x})),h] = \begin{cases} \frac{h(\hat{x}(\vec{x}))}{2} & \text{if } \hat{x}(\vec{x}) < \frac{1}{2} \text{ and } \frac{1}{h(\hat{x}(\vec{x}))} \text{ is odd} \\ \frac{h(\hat{x}(\vec{x}))+1}{2} & \text{otherwise} \end{cases} .$$

The class of hypotheses $\mathcal{H}$ is rather unnatural, and this forces $f$ to ignore the bulk of the empirical data and rely on the encoding of $E_Q h$ in terms of $h(\hat{x}(\vec{x}))$. For small values of $h(\hat{x}(\vec{x}))$, $f$ varies wildly as $h(\hat{x}(\vec{x}))$ changes. In fact, $f$ is *not* smooth. To see this, let $A(\vec{x})$ be the set of points from $\vec{x}$ that fall in $[1/2,1]$. For any $i$ and any $\vec{x} = \vec{x}(m)$ such that $\hat{x}(\vec{x}) < 1/2$ (i.e., with probability $1-2^{-m}$),

$$\rho_{\vec{x}}(h^1_i, h^0_{i,A(\vec{x})}) = E_{\vec{x}}|h^1_i - h^0_{i,A(\vec{x})}| \le \frac{m}{m}\left(\frac{1}{2i} - \frac{1}{2i+1}\right) < \frac{1}{4i^2}\,,$$

but

$$|f[(\vec{x},c(\vec{x})),h^1_i] - f[(\vec{x},c(\vec{x})),h^0_{i,A(\vec{x})}]| > 1/2.$$

Since $i$ is arbitrary, $f$ can change significantly for hypotheses that are arbitrarily close on the data. Thus, $f$ is not smooth. We show in Section 5 that there is *no* smooth simultaneous estimator for $(\{Q\},\{c\},\mathcal{H})$. □

We can think of smooth estimators as a generalization of the *empirical estimator*,

$$f_{\mathrm{emp}}[\vec{s}(m),h] = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i),y_i). \tag{8}$$

In fact, when $L$ is Lipschitz in $h$, the class of smooth estimators includes the empirical estimator as a special case. This means that we can use the conditions mentioned in Subsection 2.1 (from [26], [23], [21], etc. ) as sufficient conditions for smooth simultaneous estimability.

**Lemma 4.1.** *If $L$ is Lipschitz in $h$ ($\rho_{L,s}(h,h') \leq D\, d(h,h')$), then the empirical estimator is smooth with $\sigma(\theta,m) = \theta/D$. As a result, $(\mathcal{P},\mathcal{H})$ is smoothly simultaneously estimable if $f_{\text{emp}}$ simultaneously estimates $(\mathcal{P},\mathcal{H})$.*

**Proof:** By the triangle inequality on $\mathbb{R}$ and the definition of $f_{\text{emp}}$,

$$\begin{aligned} |\, f_{\text{emp}}[\vec{s}(m),h] - f_{\text{emp}}[\vec{s}(m),h']\,| &= \left| \frac{1}{m}\sum_{i=1}^{m} (\, L(h(x_i),y_i) - L(h'(x_i),y_i)\,) \right| \\ &\leq \frac{1}{m}\sum_{i=1}^{m} \rho_{L,s_i}(h,h') \\ &\leq \frac{D}{m}\sum_{i=1}^{m} d(h(x_i),h'(x_i)) \\ &= D\rho_{\vec{x}}(h,h'). \end{aligned}$$

Thus, $|\, f_{\text{emp}}[\vec{s}(m),h] - f_{\text{emp}}[\vec{s}(m),h']\,| < \theta$ whenever $\rho_{\vec{x}}(h,h') < \theta/D$. □

Another example of smoothness arises in the 0/1-valued case or, more generally, when $|Z|$ is finite. If $|Z|$ is finite and $d$ is a metric, there is $\alpha > 0$ such that $d(z_1,z_2) \leq \alpha \Rightarrow z_1 = z_2$. Thus, whenever $\rho_{\vec{x}(m)}(h,h') < \alpha/m$, $h(\vec{x}) = h'(\vec{x})$, and therefore $f[\vec{s},h(\vec{x})] - f[\vec{s},h'(\vec{x})] = 0$.

# 5 Smooth Simultaneous Estimation

In this section, we answer two important questions: When does a smooth simultaneous estimator exist and what form does such an estimator take? We present a probabilistic condition that characterizes when smooth simultaneous estimation is possible. We also develop a canonical smooth simultaneous estimator and give an example in which it succeeds and the empirical estimator $f_{\text{emp}}$ does not. First, we review some intermediate results and introduce some new definitions.

## 5.1 Preliminaries

In order to make statements about our ability to estimate $\text{err}(P,h)$, we must first place some restrictions on the loss criterion. An assumption we shall use throughout this paper and [1] is that $L(h(x),y)$ is bounded.

**Assumption 5.1.** *$L$ is nonnegative and bounded by $\lambda < \infty$:*

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

The assumption that $L$ is nonnegative is not restrictive, since we could just as well use $L$ plus a constant.

Now, suppose that $\mathcal{H} = \{h_1, h_2, \ldots, h_k\}$, so that $|\mathcal{H}| = k$. Assume that $L$ is Lipschitz in $h$ and recall, from Lemma 4.1, that then the empirical estimator is smooth. It is a simple consequence of Chebyshev's Inequality that, under Assumption 5.1, $f_{\text{emp}}$ simultaneously estimates $(\mathcal{P}, \mathcal{H})$. Thus, for our purposes, *any* $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable by $f_{\text{emp}}$ when $|\mathcal{H}| < \infty$. In fact, using Hoeffding's Inequality ([40]), we can give a bound on the number of samples used by $f_{\text{emp}}$ in this case. We state this result in a general form, since $f_{\text{emp}}$ serves as a building block for the estimation scheme developed in this section.

For our purposes, Hoeffding's inequality states that, under Assumption 5.1, for any $h \in \mathcal{H}$ and $\mu > 0$,

$$P^\ell \left[\vec{s}(\ell) : |f_{\text{emp}}[\vec{s}, h] - \text{err}(P, h)| > \mu\right] < 2\exp(-2\ell\mu^2/\lambda^2).$$

A simple argument then shows that the following lemma holds (see [21] for details).

**Lemma 5.1.** *Under Assumption 5.1, if $\mathcal{A}$ is a finite subset of $\mathcal{H}$, then for any $P \in \mathcal{P}$ and $\ell \geq (\lambda^2/2\mu^2)\ln[2|\mathcal{A}|/\nu]$,*

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

## 5.2 Empirical Coverings

The canonical estimator developed in this section will involve two approximations. One is the approximation of the error of a hypothesis by $f_{\text{emp}}$. The other is an approximation of the set $\mathcal{H}$ by an *empirical covering*.

Let $\rho$ be a pseudometric on $\mathcal{H}$. We say that a set $\mathcal{A}$ is an *$\epsilon$-cover for $\mathcal{H}$ with respect to $\rho$* if for every $h \in \mathcal{H}$ there is an $a(h) \in \mathcal{A}$ with $\rho(h, a(h)) < \epsilon$. We also say that such a pair $(\mathcal{A}, a(\cdot))$ is an *$\epsilon$-covering for $\mathcal{H}$ with respect to $\rho$*. A covering is *finite* if $|\mathcal{A}| < \infty$.

Let $d$ be the pseudometric on $Z$ used to define smoothness in Section 4. Recall that for each sample, $\vec{x}$, we can compute the *empirical distance* between $h$ and $h'$,

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

**Definition 5.1.** *An $\epsilon$-cover(ing) for $\mathcal{H}$ with respect to $\rho_{\vec{x}}$, the empirical distance, will be called an* empirical $\epsilon$-cover(ing) for $\mathcal{H}$ based on $\vec{x}$.

We will attempt to construct a simultaneous estimator for $(\mathcal{P},\mathcal{H})$ by first finding an empirical covering, $(\mathcal{H}_n(\vec{x}), a_{n,\vec{x}})$, and then estimating the error of $a_{n,\vec{x}}(h)$. For this approach to succeed, the error of $a_{n,\vec{x}}(h)$ must approximate that of $h$. That is, hypotheses that are close on the samples should have similar errors. The following condition ensures that this is the case.

**Definition 5.2.** *$(\mathcal{P},\mathcal{H})$ can be* empirically covered *if*

$$\sup_{h,h'\in\mathcal{H} \text{ s.t. } \rho_{\vec{x}(n)}(h,h')<\tau/n} | \operatorname{err}(P,h) - \operatorname{err}(P,h') | \to 0 \tag{9}$$

*in probability as $n \uparrow \infty$ and $\tau \downarrow 0$ uniformly in $P \in \mathcal{P}$.*

(Lemma A.1 in the Appendix clarifies the manner in which this convergence occurs. It turns out that, for any sequences $n_k \uparrow \infty$ and $\tau_k \downarrow 0$, the limit as $k$ increases will be the same.)

We can interpret the meaning of (9) in terms of coverings. Consider the pseudometric on $\mathcal{H}$ obtained by computing the difference of the error of two hypotheses:

$$\rho_{_P}(h,h') := |\operatorname{err}(P,h) - \operatorname{err}(P,h')|. \tag{10}$$

If we have access to finite coverings for $\mathcal{H}$ with respect to $\rho_{_P}$, we can simultaneously estimate $(\mathcal{P},\mathcal{H})$ by estimating the error of each cover element empirically. This approach cannot be implemented though, because it requires some knowledge of $P$ to produce the coverings, in general. However, when $(\mathcal{P},\mathcal{H})$ can be empirically covered, we can construct such coverings from a *sample* $\vec{x}$. By Definition 5.2 and (10), if $(\mathcal{P},\mathcal{H})$ can be empirically covered,

$$\sup_{h,h'\in\mathcal{H} \text{ s.t. } \rho_{\vec{x}(n)}(h,h')<\tau/n} \rho_{_P}(h,h') < \xi \tag{11}$$

will hold with probability approaching 1 for any fixed $\xi > 0$ as $n$ and $1/\tau$ increase. For $\vec{x}$ satisfying (11),

$$\rho_{\vec{x}(n)}(h, h') < \tau/n \Rightarrow \rho_{P}(h, h') < \xi.$$

That is, any empirical $\tau/n$-covering for $\mathcal{H}$ based on $\vec{x}$ is also a $\xi$-covering for $\mathcal{H}$ with respect to $\rho_{P}$. Briefly, if $(\mathcal{P}, \mathcal{H})$ can be empirically covered, it is likely that any empirical covering for $\mathcal{H}$ is a covering for $\mathcal{H}$ with respect to $\rho_{P}$ as well.

For our purposes, a vital property of an empirical cover is that it be finite. The following assumption guarantees that we can find such a cover.

**Assumption 5.2.** *$Z$ is totally bounded under the pseudometric $d$ (see [41]). That is, for each $\epsilon > 0$ there is a set of $K(\epsilon) < \infty$ points, $Z(\epsilon) \subseteq Z$, that form an $\epsilon$-cover for $Z$ with respect to $d$.*

As an example, when $Z \subseteq \mathbb{R}$ and $d(z_1, z_2) = |z_1 - z_2|$, Assumption 5.2 is satisfied whenever $Z \subseteq (-B, B)$ for some $B < \infty$. In this case, we could take

$$Z(\epsilon) = \{n\epsilon : n = 0, \pm 1, \ldots, \pm \lfloor B/\epsilon \rfloor\};$$

actually, about half this many points would suffice.

Although an empirical $\epsilon$-cover based on $\vec{x}(n)$ varies randomly along with $\vec{x}(n)$ (which is obtained by sampling according to $P$), its size can be bounded in terms of $\epsilon$ and $n$. In fact, the following construction yields an empirical $\epsilon$-covering of size at most $[K(\epsilon/2)]^n$.

**Procedure 5.1.** *Construction of a finite empirical covering.*

*For each fixed $\epsilon > 0$, $n$, and $\vec{x}(n) = [x_1, x_2, \ldots, x_n]$, we shall construct both an empirical $\epsilon$-cover for $\mathcal{H}$ based on $\vec{x}(n)$, $\mathcal{H}_n(\vec{x})$, and the associated mapping $a_{n,\vec{x}}$ from hypotheses to $\mathcal{H}_n(\vec{x})$.*

*By Assumption 5.2, there is an $\epsilon/2$-cover for $Z$ with respect to $d$ with at most $K(\epsilon/2)$ elements, $Z(\epsilon/2) \subseteq Z$. There are $[K(\epsilon/2)]^n$ possible mappings, $M_k$, from $i \in \{1, \ldots, n\}$ to cover elements: each $M_k(i) \in Z(\epsilon/2)$. Initially, let $\mathcal{A}_0$ be the empty set. Construct $\mathcal{H}_n(\vec{x})$ by iterating on $k \in \{1, \ldots, [K(\epsilon/2)]^n\}$ as follows. For each $k$, if there is $h_k \in \mathcal{H}$ such that*

$$d(h_k(x_i), M_k(i)) < \epsilon/2 \text{ at each } i \in \{1, \ldots, n\}, \tag{12}$$

*let* $\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{h_k\}$. *Otherwise, set* $\mathcal{A}_k = \mathcal{A}_{k-1}$. *Finally, let* $\mathcal{H}_n(\vec{x}) = \mathcal{A}_{[K(\epsilon/2)]^n}$.

*The proof of Lemma 5.2 shows that, for each* $h \in \mathcal{H}$, *there is* $h' \in \mathcal{H}_n(\vec{x})$ *with* $\rho_{\vec{x}}(h, h') < \epsilon$. *Set* $a_{n,\vec{x}}(h) = h'$.

**Lemma 5.2.** *Under Assumption 5.2, the pair* $(\mathcal{H}_n(\vec{x}), a_{n,\vec{x}}(\cdot))$ *produced by Procedure 5.1 is an empirical* $\epsilon$*-covering for* $\mathcal{H}$ *based on* $\vec{x}(n)$ *of size at most* $[K(\epsilon/2)]^n$.

**Proof:** See the Appendix.

In the next subsection, we assemble the facts and definitions given here into an explicit procedure for smooth simultaneous estimation.

## 5.3 A Canonical Smooth Simultaneous Estimator

We now describe a "canonical" estimator that will smoothly simultaneously estimate $(\mathcal{P}, \mathcal{H})$ whenever such estimation is possible. (Throughout this subsection, suppose that Assumptions 5.1 and 5.2 hold.) The estimator has two stages. It first approximates $\mathcal{H}$ with a finite empirical covering and then empirically estimates the error of all members of the cover simultaneously.

**Procedure 5.2.** *The Canonical Estimator.*

*Let* $\tau_n \searrow 0$, $\mu_n \searrow 0$, *and* $\nu_n \searrow 0$ ($\tau_n \searrow 0$ *means that* $\tau_n$ *is positive, nonincreasing, and converges to zero). For a labeled sample* $\vec{s}(m)$ *of* $m$ *points, perform the following three steps.*

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

$$n + \lceil (\lambda^2/2\mu_n^2) \ln[2H(n)/\nu_n] \rceil \leq m. \tag{13}$$

   *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, by Procedure 5.1, an empirical* $(\tau_n/n)$*-covering for* $\mathcal{H}$ *based on* $\vec{x}\,'$, $(\mathcal{H}_n(\vec{x}\,'), a_{n,\vec{x}'}(\cdot))$. *Consequently, for any* $h \in \mathcal{H}$, *there is* $a_{n,\vec{x}'}(h) \in \mathcal{H}_n(\vec{x}\,')$ *with* $\rho_{\vec{x}'}(a_{n,\vec{x}'}(h), h) < \tau_n/n$. *By Lemma 5.2, the size of* $\mathcal{H}_n(\vec{x}\,')$ *is at most* $H(n)$. *By*

*Lemma 5.1, with probability at least $1-\nu_n$, the empirical error estimate, $f_{\mathrm{emp}}[\vec{s}'', h']$, will be within $\mu_n$ of $\mathrm{err}(P, h')$ simultaneously for all $h' \in \mathcal{H}_n(\vec{x}')$.*

3. *The Canonical Estimator, $f_{\mathrm{ce}}$, is the composition of the map $a_{n,\vec{x}'}$ in Step 2 and the empirical error estimate:*

$$f_{\mathrm{ce}}[\vec{s}, h] = f_{\mathrm{emp}}[\vec{s}''(m-n), a_{n,\vec{x}'}(h)].$$

In Step 1, it suffices to pick $n$ such that $n$ goes to infinity as $m$ does and $n$ satisfies (13) for large $m$. For example, in the 0/1-valued case, $\lambda = 1$ and $H(n) \leq 2^n$. Thus, if $\mu_n = 1/\sqrt{n}$ and $\nu_n = e^{1-n}$, we can use $n = \sqrt{m}$.

We note that there is actually a whole class of canonical estimators. Each choice of the *sequences* $\tau_n$, $\mu_n$, $\nu_n$, and $\mathcal{H}_n(\vec{x})$ leads to an $f_{\mathrm{ce}}$. When we make statements such as "$f_{\mathrm{ce}}$ smoothly simultaneously estimates," we mean that *every* such $f_{\mathrm{ce}}$ smoothly simultaneously estimates.

Let us examine the form of $f_{\mathrm{ce}}$. Note that the labels $\vec{y}'$ of $\vec{x}'$ are not needed. This is an advantage if labeling the samples is costly. Unlike the empirical estimator, $f_{\mathrm{ce}}$ does make use of the structure of $\mathcal{H}$; however, it does *not* depend on $\mathcal{P}$. For each $h' \in \mathcal{H}_n(\vec{x}')$, $f_{\mathrm{ce}}$ is constant on the set of hypotheses that are mapped to $h'$ by $a_{n,\vec{x}'}$. Also, $f_{\mathrm{ce}}$ depends on $h$ only through $h(\vec{x}')$.

For each fixed $n$ and $\vec{x}'(n)$, $f_{\mathrm{ce}}$ takes on only finitely many (at most $H(n)$) values. Actually, $H(n)$ can itself be replaced by any number as large as $|\mathcal{H}_n(\vec{x}')|$ when choosing $n$ in Step 1 of Procedure 5.2. The size of $\mathcal{H}_n(\vec{x}')$ will frequently be much less than $H(n)$. We examine some such cases in [1] where $|\mathcal{H}_n(\vec{x}')|$ grows only polynomially in $n$, instead of exponentially.

Procedure 5.2, as stated, illustrates the form of a canonical estimator and is not intended to make the most efficient use of samples. For specific values of $\mathcal{P}$ and $\mathcal{H}$, some specific choice of $\tau_n$, $\mu_n$, $\nu_n$, and $\mathcal{H}_n(\vec{x})$ may result in an $f_{\mathrm{ce}}$ that produces better estimates for a given number of samples. We discuss this in more detail in [1].

## 5.4 Characterizing Smooth Simultaneous Estimation

The following theorem relates the major concepts in this section to smooth simultaneous estimability.

**Theorem 5.1.** *Under Assumptions 5.1 and 5.2, the following are equivalent.*

*(i)* $(\mathcal{P}, \mathcal{H})$ *is smoothly simultaneously estimable.*

*(ii)* $(\mathcal{P}, \mathcal{H})$ *can be empirically covered.*

*(iii)* $f_{\mathrm{ce}}$ *smoothly simultaneously estimates* $(\mathcal{P}, \mathcal{H})$.

**Proof:** See the Appendix.

Theorem 5.1 is useful for three reasons. First, it characterizes smooth simultaneous estimability in terms of the sampling-based criterion of (9). Examples 5.1 and 5.2 that follow and Example 2.1 in [1] illustrate how to use this characterization.

**Example 5.1.**

We can use Theorem 5.1 to show that $(\mathcal{P}, \mathcal{H})$ in Example 4.1 cannot be smoothly simultaneously estimated. There, for any $i$ and $\vec{x}$,

$$\rho_{\vec{x}}(h_i^1, h_{i,A(\vec{x})}^0) < \frac{1}{4i^2}\,.$$

However,

$$|\mathrm{err}(P, h_i^1) - \mathrm{err}(P, h_{i,A(\vec{x})}^0)| > 1/2.$$

Thus, $(\mathcal{P}, \mathcal{H})$ cannot be empirically covered, and there is no smooth estimator that simultaneously estimates $(\mathcal{P}, \mathcal{H})$. □

Theorem 5.1 not only tells us *when* smooth simultaneous estimation is possible, but it also tells us *how*. The Canonical Estimator, $f_{\mathrm{ce}}$, will suffice whenever such estimation is possible.

Lemma 4.1 implies that, for many natural choices of $d$ and $L$, $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable (and hence, by Lemma 3.1, *learnable*) if the empirical estimator

simultaneously estimates $(\mathcal{P}, \mathcal{H})$. Example 5.2 shows that the converse of this statement is not true. Thus, Theorem 5.1 provides us with sufficient conditions for learnability that are weaker than those based on simultaneous estimation by empirical estimates in [26], [23], [21], and [15].

**Example 5.2.** $f_{\text{ce}}$ may work in some cases where $f_{\text{emp}}$ does not.

(This example is derived from an example in [42] which, in turn, was inspired by [23].)

Consider the real-valued, noise-free case. Let $Y = Z = [0, 1]$, and take $L$ and $d$ to be equal, with $L(z, y) = |z - y|$. (Note that the assumptions for Theorem 5.1 are satisfied.) Let $\mathcal{Q}$ be a singleton, $\{Q\}$, and $\mathcal{C} = \{c\}$, where $c(x) \equiv 0$. Suppose that $\mathcal{H}$ is composed of an infinite sequence of hypotheses that are normalized to $\gamma > 0$:

$$E_Q h_i = \gamma, \; i = 1, 2, \ldots,$$

and that the hypotheses are at least $\epsilon$ apart:

$$E_Q |h_i - h_j| \geq \epsilon > 0 \text{ for every } i \neq j. \tag{14}$$

For instance, when $X = [0, 1) \subset \mathbb{R}$, let $0.b_1(x)b_2(x)b_3(x)\ldots$ be the binary representation of $x \in [0, 1)$ and define $h_i$ by $h_i(x) = b_i(x)$. Then $\gamma = 1/2$ and $\epsilon = 1/2$.

Since $c \equiv 0$, $\text{err}(Q, c, h_i) = \gamma$ for all $i$, and therefore $(\{Q\}, \{c\}, \mathcal{H})$ can be empirically covered. Therefore, by Theorem 5.1, the canonical estimator $f_{\text{ce}}$ smoothly simultaneously estimates $(\{Q\}, \{c\}, \mathcal{H})$.

However, $f_{\text{emp}}$ does *not* simultaneously estimate $(\{Q\}, \{c\}, \mathcal{H})$. Suppose, on the contrary, that it does. Then by results in [26] or [23], $f_{\text{emp}}$ simultaneously estimates $(\{Q\}, \{c\}, \mathcal{H}^\Delta)$, where

$$\mathcal{H}^\Delta = \{|h_i - h_j| : h_i, h_j \in \mathcal{H}\}.$$

That is, $E_{\vec{x}}|h_i - h_j|$ converges to $E_P|h_i - h_j|$ in probability simultaneously over $h_i, h_j \in \mathcal{H}$.

Define the metric $\rho_1$ on $Z^n$ by

$$\rho_1(\vec{u}, \vec{v}) = \frac{1}{n} \sum_{k=1}^{n} |u_k - v_k|,$$

and note that $\rho_1(h_i(\vec{x}), h_j(\vec{x})) = E_{\vec{x}}|h_i - h_j|$. By (14), for large $n$ the probability that $\vec{x}(n)$ is such that

$$\rho_1(h_i(\vec{x}), h_j(\vec{x})) > \epsilon/2 \text{ for all } i \neq j \tag{15}$$

is near 1. Let $\mathcal{N}(\epsilon, \mathcal{H}(\vec{x}(n)), \rho_1)$ denote the size of the smallest $\epsilon$-cover of

$$\mathcal{H}(\vec{x}) = \{h(\vec{x}) : h \in \mathcal{H}\}$$

with respect to $\rho_1$. (I.e., the size of the smallest set $U \subset Z^n$ with the property that for each $\vec{v} \in \mathcal{H}(\vec{x})$ there is $\vec{u} \in U$ with $\rho_1(\vec{v}, \vec{u}) < \epsilon$.) For $\vec{x}(n)$ satisfying (15), $\mathcal{N}(\epsilon/4, \mathcal{H}(\vec{x}(n)), \rho_1)$ must be infinite. Otherwise, there would be some point $\vec{u} \in Z^n$ and $\vec{v}_1, \vec{v}_2 \in \mathcal{H}(\vec{x})$ such that $\rho_1(\vec{u}, \vec{v}_1) < \epsilon/4$ and $\rho_1(\vec{u}, \vec{v}_2) < \epsilon/4$, implying $\rho_1(\vec{v}_1, \vec{v}_2) < \epsilon/2$, which contradicts (15). Thus, for any $T < \infty$,

$$Q^n[\vec{x}(n) : \mathcal{N}(\epsilon/4, \mathcal{H}(\vec{x}(n)), \rho_1) > T] \to 1.$$

By results in [26] or [23], this contradicts the assumption that $f_{\text{emp}}$ simultaneously estimates $(\{Q\}, \{c\}, \mathcal{H})$. □

# 6 Concluding Remarks

We have considered the problem of learning by simultaneous error estimates in a realistic framework. We have defined a natural notion of smoothness for estimators and given necessary and sufficient conditions for the existence of a smooth simultaneous estimator. We have developed a simple canonical form for these estimators. For most loss criteria, the common empirical estimator is smooth. Thus these results provide weaker sufficient conditions for a broad class of learning problems.

In a companion paper ([1]), we examine learning procedures based on selecting a hypothesis that minimizes the error as estimated by the canonical smooth simultaneous estimator. This "canonical learner" works whenever smooth simultaneous estimation is possible. We give evidence that this canonical learner enjoys certain advantages over empirical error minimization. When the loss criterion, $L$, is Lipschitz, we show that the canonical learner works whenever empirical error minimization does. In fact, we exhibit a pair $(\mathcal{P}, \mathcal{H})$ that can be

learned by the canonical learner but not by empirical error minimization. We derive bounds on the number of labeled sample points needed by each method and find that the bounds for the canonical learner are in general comparable to, and in certain special cases are less than, the bounds for empirical error minimization. We also show how the canonical learner can be readily applied to learning problems that involve a nested hierarchy of hypothesis classes. In such problems, choosing a hypothesis of appropriate "complexity" may be necessary to avoid fitting the noise in the data.

# Appendix

**Proof of Lemma 5.2** Clearly, $|\mathcal{H}_n(\vec{x}\,)| \leq [K(\epsilon/2)]^n$.

We need to show that $\mathcal{H}_n(\vec{x}\,)$ is an empirical $\epsilon$-cover for $\mathcal{H}$: for any $h \in \mathcal{H}$ there is an $h' \in \mathcal{H}_n(\vec{x}\,)$ such that $\rho_{\vec{x}}(h, h') < \epsilon$. Note that $h(x_i)$ will be close to some element of the cover $Z(\epsilon/2)$ at each $i$. In the notation of Procedure 5.1, for some $k'$, $h(x_i)$ is within $\epsilon/2$ of $M_{k'}(i)$ at each $i$:

$$d(h(x_i), M_{k'}(i)) < \epsilon/2,\ i \in \{1, \ldots, n\}. \tag{16}$$

Thus, there is *some* $h_{k'} \in \mathcal{H}_n(\vec{x}\,)$ that satisfies (12) with $k = k'$. Set $h' = h_{k'}$. By (16) and the triangle inequality for $d$,

$$d(h_{k'}(x_i), h(x_i)) \leq d(h_{k'}(x_i), M_{k'}(i)) + d(h(x_i), M_{k'}(i)) < \epsilon$$

for $i \in \{1, \ldots, n\}$, and so $\rho_{\vec{x}}(h', h) < \epsilon$, as claimed.

**Note:** This remark is will be useful in [1]. We could alter the construction in Procedure 5.1 so that $\mathcal{H}_n(\vec{x}\,)$ is an $\epsilon/2$-separated empirical $\epsilon$-cover. (A set $\mathcal{A}$ of hypotheses is $\epsilon/2$-separated if $\rho_{\vec{x}}(h, h') > \epsilon/2$ for any $h, h' \in \mathcal{A}$ such that $h \neq h'$.) To do this, we would have to alter the rule for adding an $h_k$ to $\mathcal{A}_{k-1}$: for each $k$, *search* for an $h_k$ that satisfies (12) and is $\epsilon/2$-separated from each member of $\mathcal{A}_{k-1}$. If there is one, add it to $\mathcal{A}_{k-1}$. Otherwise, move to the next $k$. Clearly, $\mathcal{H}_n(\vec{x}\,)$ constructed in this manner is $\epsilon/2$-separated. Also, for any $h$, there is $k'$ such that (16) holds. Then either there is some $h_{k'}$ in $\mathcal{H}_n(\vec{x}\,)$ that satisfies

(12) with $k = k'$, or there is $h'$ in $\mathcal{H}_n(\vec{x}\,)$ with $\rho_{\vec{x}}(h, h') \leq \epsilon/2$. In either case, there is an element of $\mathcal{H}_n(\vec{x}\,)$ that is within $\epsilon$ of $h$. □

**Proof of Theorem 5.1** We first note that, because $d(\cdot, \cdot) \geq 0$ and the samples are drawn independently, the following lemma is true.

**Lemma A.1.** *Let $W(m, \tau, P, \epsilon)$ be the set of $\vec{s}(m) = (\vec{x}, \vec{y})$ for which $\vec{x}$ is such that*

$$\sup_{h,h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}(m)}(h,h') < \tau/m} | \operatorname{err}(P, h) - \operatorname{err}(P, h') | > \epsilon.$$

*Then $P^m[W(m, \tau, P, \epsilon)]$ is nonincreasing in both $m$ and $1/\tau$. Thus, regardless of how $(m, 1/\tau) \to (\infty, \infty)$, the following limit is defined and has the same value:*

$$\lim_{(m,1/\tau) \to (\infty,\infty)} \sup_{P \in \mathcal{P}} P^m[W(m, \tau, P, \epsilon)].$$

We need to establish that that the following three statements are equivalent under Assumptions 5.1 and 5.2.

(i) $(\mathcal{P}, \mathcal{H})$ is smoothly simultaneously estimable.

(ii) $(\mathcal{P}, \mathcal{H})$ can be empirically covered.

(iii) $f_{\text{ce}}$ smoothly simultaneously estimates $(\mathcal{P}, \mathcal{H})$.

**(iii) ⇒ (i):** This is immediate from the definitions.

**(i) ⇒ (ii):** We show the contrapositive. Let us suppose that $(\mathcal{P}, \mathcal{H})$ *cannot* be empirically covered. We will show that no smooth estimator simultaneously estimates $(\mathcal{P}, \mathcal{H})$.

A careful inversion of Definition 5.2 using Lemma A.1 shows that there are $\epsilon', \delta' > 0$ such that, for any $m$ and $\tau$, there is $P_{m,\tau} \in \mathcal{P}$ with the property that, with probability greater than $\delta'$, the labeled sample $\vec{s}(m) = (\vec{x}, \vec{y})$ is such that there are $h_{\vec{x}}, h'_{\vec{x}} \in \mathcal{H}$ with

$$\rho_{\vec{x}}(h_{\vec{x}}, h'_{\vec{x}}) < \tau/m, \text{ and} \tag{17}$$

$$| \operatorname{err}(P_{m,\tau}, h_{\vec{x}}) - \operatorname{err}(P_{m,\tau}, h'_{\vec{x}}) | > \epsilon'. \tag{18}$$

Take any candidate estimator $f$ that is smooth with parameter $\sigma(\theta, m)$. Let us set $\tau(m) = m\sigma(\epsilon'/3, m)$, fix $\theta$ at $\epsilon'/3$, and leave $m$ free. Because $f$ is $\sigma(\theta, m)$-smooth, (17) implies that $f$ will not change rapidly between $h_{\vec{x}}$ and $h'_{\vec{x}}$ for most $\vec{s}$. In fact, we have from the definition of smoothness that, for large enough $m$, all of (17), (18), and

$$| f[\vec{s}, h_{\vec{x}}] - f[\vec{s}, h'_{\vec{x}}] | \leq \epsilon'/3 \tag{19}$$

are true simultaneously with $P^m_{m,\tau(m)}$-probability greater than $\delta'/2$. For $\vec{s}$ such that (18) and (19) are both true, either

$$| f[\vec{s}, h_{\vec{x}}] - \mathrm{err}(P_{m,\tau(m)}, h_{\vec{x}}) | > \epsilon'/3$$

or

$$| f[\vec{s}, h'_{\vec{x}}] - \mathrm{err}(P_{m,\tau(m)}, h'_{\vec{x}}) | > \epsilon'/3.$$

Thus, for large $m$, there is $P_{m,\tau(m)} \in \mathcal{P}$ such that

$$\sup_{h \in \mathcal{H}} | f[\vec{s}, h] - \mathrm{err}(P_{m,\tau(m)}, h) | > \epsilon'/3$$

with probability greater than $\delta'/2$. That is, $f$ does not simultaneously estimate $(\mathcal{P}, \mathcal{H})$.

**(ii) $\Rightarrow$ (iii):** Assume that $(\mathcal{P}, \mathcal{H})$ can be empirically covered. Thus, we have by Lemma A.1 that, for any given $\epsilon > 0$,

$$\lim_{n \uparrow \infty} \lim_{\tau \downarrow 0} \sup_{P \in \mathcal{P}} P^n[W(n, \tau, P, \epsilon)] = 0. \tag{20}$$

We will first show that $f_{\mathrm{ce}}$ simultaneously estimates $(\mathcal{P}, \mathcal{H})$. In Step 2 of Procedure 5.2, we choose an $a_{n,\vec{x}'}(h)$ such that

$$\{(h, a_{n,\vec{x}'}(h)) : h \in \mathcal{H}\} \subseteq \{(h, h') : h, h' \in \mathcal{H} \text{ and } \rho_{\vec{x}'}(h, h') < \tau_n/n\}.$$

Since $\tau_n$ converges to zero, (20) implies that, when $n$ is large,

$$\sup_{h \in \mathcal{H}} | \mathrm{err}(P, a_{n,\vec{x}'}(h)) - \mathrm{err}(P, h) | \tag{21}$$

will be small with high probability. Also, $n$ in Step 1 of Procedure 5.2 is chosen so that, by Lemma 5.1,

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

is less than $\nu_n$. (Here we use the fact that $\vec{s}'$ and $\vec{s}''$ are drawn independently; $\mathcal{H}_n(\vec{x}')$ is fixed relative to $\vec{s}''$.) Since $a_{n,\vec{x}'}(h)$ is always in $\mathcal{H}_n(\vec{x}')$,

$$\sup_{h \in \mathcal{H}} | f_{\text{emp}}[\vec{s}'', a_{n,\vec{x}'}(h)] - \text{err}(P, a_{n,\vec{x}'}(h)) | \leq \mu_n \tag{22}$$

with probability at least $1 - \nu_n$. By (21), (22), and the triangle inequality for real numbers,

$$\sup_{h \in \mathcal{H}} | f_{\text{emp}}[\vec{s}'', a_{n,\vec{x}'}(h)] - \text{err}(P, h) |$$

is small with high probability for large $n$. This is true uniformly for $P \in \mathcal{P}$. Thus, $f_{\text{ce}} = f_{\text{emp}}[\vec{s}'', a_{n,\vec{x}'}(h)]$ simultaneously estimates $(\mathcal{P}, \mathcal{H})$.

Now we need to show that $f_{\text{ce}}$ is smooth. Fix $\theta > 0$. Because $f_{\text{ce}}$ simultaneously estimates $(\mathcal{P}, \mathcal{H})$ and (20) holds, we have that for $\tau(\theta, m)$ small enough, both

$$\sup_{h \in \mathcal{H}} | f_{\text{ce}}[\vec{s}(m), h] - \text{err}(P, h) | \leq \theta/3$$

and

$$\sup_{h, h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}(m)}(h, h') < \tau(\theta, m)/m} | \text{err}(P, h) - \text{err}(P, h') | \leq \theta/3$$

hold with probability approaching 1 uniformly over $\mathcal{P}$ as $m \uparrow \infty$. Thus, the probability that

$$\sup_{h, h' \in \mathcal{H} \text{ s.t. } \rho_{\vec{x}(m)}(h, h') < \tau(\theta, m)/m} | f_{\text{ce}}[\vec{s}, h] - f_{\text{ce}}[\vec{s}, h'] | \leq \theta$$

approaches 1 uniformly over $\mathcal{P}$ as $m \uparrow \infty$. That is, $f_{\text{ce}}$ is smooth over $(\mathcal{P}, \mathcal{H})$ with $\sigma(\theta, m) = \tau(\theta, m)/m$. □

# References


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

[2] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, pp. 1134–1142, November 1984.

[3] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, no. 4, pp. 319–342, 1988.

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

[6] D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth, "Equivalence of models for polynomial learnability," *Information and Computation*, vol. 95, pp. 129–161, 1991.

[7] D. Angluin and P. D. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, no. 4, pp. 343–370, 1988.

[8] G. M. Benedek and A. Itai, "Nonuniform learnability," in *Proceedings of the Fifteenth International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science, #317, pp. 81–92, Springer-Verlag, July 1988.

[9] 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.

[10] M. Linial, Y. Mansour, and R. L. Rivest, "Results on learnability and the Vapnik-Chervonenkis dimension," in *29th Annual IEEE Symposium on Foundations of Computer Science*, pp. 120–129, 1988.

[11] B. K. Natarajan, "Learning over classes of distributions," in *Computational Learning Theory: Proceedings of the 1988 Workshop*, pp. 408–409, Morgan Kaufmann, San Mateo, CA, 1988.

[12] R. Sloan, "Types of noise in data for concept learning," in *Computational Learning Theory: Proceedings of the 1988 Workshop*, pp. 91–96, Morgan Kaufmann, San Mateo, CA, 1988.

[13] B. K. Natarajan, "On learning sets and functions," *Machine Learning*, vol. 4, no. 1, pp. 67–97, 1989.

[14] 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.

[15] 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.

[16] M. J. Kearns and R. E. Schapire, "Efficient distribution-free learning of probabilistic concepts," in *31st Annual IEEE Symposium on Foundations of Computer Science*, pp. 382–391, 1990.

[17] A. T. Ogielski, "Information, probability, and learning from examples," unpublished course notes, Bell Communications Research, Morristown, NJ 07960, May 1990.

[18] 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.

[19] B. K. Natarajan, "Probably approximate learning of sets and functions," *SIAM Journal of Computing*, vol. 20, no. 2, pp. 328–351, 1991.

[20] B. K. Natarajan, *Machine Learning: A Theoretical Approach.* San Mateo, CA: Morgan Kaufmann, 1991.

[21] 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.

[22] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning," in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 341–352, The Association for Computing Machinery, New York, NY, 1992.

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

[24] S. Ben-David, N. Cesa-Bianchi, and P. M. Long, "Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions," in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pp. 333–340, The Association for Computing Machinery, New York, NY, 1992.

[25] 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.

[26] 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.

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

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

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

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

[31] M. Talagrand, "The Glivenko-Cantelli problem," *The Annals of Probability*, vol. 15, no. 3, pp. 837–870, 1987.

[32] P. Gaenssler, *Empirical Processes*, vol. 3 of *IMS Lecture Notes–Monograph Series*. Hayward, California: Institute of Mathematical Statistics, 1983.

[33] E. Gine and J. Zinn, "Some limit theorems for empirical processes," *The Annals of Probability*, vol. 12, no. 4, pp. 929–989, 1984.

[34] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[35] 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.

[36] K. Fukunaga, *Statistical Pattern Recognition*. Academic Press, second ed., 1990.

[37] L. Pitt and L. G. Valiant, "Computational limitations on learning from examples," *Journal of the ACM*, vol. 35, no. 4, pp. 965–984, 1988.

[38] 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.

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

[40] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

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

[42] 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.