International Journal of Foundations of Computer Science Vol. 8 No. 4 (1997) 409–424

LEARNING WITH NATURAL IMPRECISION

MARK A. CHANGIZI
Department of Applied Mathematics, University of Maryland, College Park, Maryland 20742, USA
e-mail: changizi@carnap.umd.edu

Received 17 June 1995
Revised 28 August 1997
Communicated by D. T. Lee

ABSTRACT

The theorems of inductive inference in computational learning theory may be interpreted as the ultimate theoretical constraints on the abilities of finite machines to fabricate hypotheses and make predictions from information or data. However, all of the models of learning in this inductive inference literature require the hypothesis to be exactly correct infinitely often, have no way of measuring how far off a prediction of a hypothesis is, and require the hypothesis to make predictions that are directly measurable. Furthermore, most of the models of learning do not allow the learning of uncomputable functions, and of those that are capable of learning uncomputable functions, the literature has not explicitly noticed this property. New notions of learning that rectify these deficiencies are introduced and examined. The first criterion considers a hypothesis successful if each prediction is within $\delta(x)$ of the function to be learned on x, $f(x)$. The second criterion considers a hypothesis successful if each prediction on x is equal to f on some domain element within $\epsilon(x)$ of x, so long as the nearby value of the function f is within $v(x)$ of $f(x)$. These new learning criteria respectively model (a) learning in science with imprecise hypotheses, and (b) learning in science with hypotheses that ignore noisy and bad data, smoothing in image recognition with radius of smoothing ϵ and threshold v, and learning languages (i.e., two-valued functions) with imprecise hypotheses.

Keywords: Gold model inductive inference, error, approximate learning, smoothing, extrapolation, imprecise theories.

1. Introduction

There are the following problems with the Gold model computational learning theory discipline of inductive inference thus far claiming to model the ultimate theoretical constraints on the abilities of finite machines to produce hypotheses (programs) and make predictions (program outputs) from information or data.

1. Each model requires a hypothesis to be exactly correct infinitely often.

2. Each model has no way to measure how far off a prediction of a hypothesis is.

3. No model allows for a hypothesis to make predictions that are not directly verifiable.

4. Few models allow the learning of uncomputable functions.

The central aim of this paper is to equip inductive inference with the tools needed to rectify these deficiencies.

1.1. Theories with Error

Learning with error is treated in the inductive inference literature in two ways. The first considers a learner to have successfully inferred a phenomenon if the hypothesis the learner produces makes correct predictions about the phenomenon ***with some probability***[7].[a] The second treatment of learning with error is with anomalies;[2,10,12] these criteria consider a learner to have successfully inferred a phenomenon even if the hypothesis is incomplete, or anomalous. Each of these treatments has no way of measuring how far off a prediction of a hypothesis is (a metric), and each forces the hypothesis to be correct infinitely often. Neither of these treatments is adequate in most natural situations. Science, for example, does not require a hypothesis' prediction to *ever* be exactly correct so long as it is always close enough. In language learning there may be stages at which every sentence a child utters is ungrammatical, yet close enough to be understood. The notions introduced in this paper supply just the sort of natural error criteria to handle such phenomena and do not conflict with the probabilistic or anomaly treatments. They may be used concurrently so that one may study probabilistic learning with anomalies *and* error.

1.2. Unmeasurable Predictions

Suppose a learner is equipped with an inch ruler (incapable of measurements more accurate than one inch) and is set upon measuring the width of pieces of paper coming off a production line in order to learn, in general, what the width of future pieces of paper off that same line will be. Suppose that each piece of paper has width exactly 8.5 inches. The function the learner is attempting to learn is the constant 8.5 inch function. Things are not so simple, however, since, as the ruler used is an inch ruler, the data the learner receives never says 8.5 inches. An example data stream the learner might receive is 8, 9, 9, 9, 8, 9, 8, 8, 9,.... A natural hypothesis for a real-world learner is the constant 8.5 inch function. However, this hypothesis' predictions are never directly verifiable by the measuring device at hand, the inch ruler. Previous notions of learning in the inductive inference literature cannot accommodate hypotheses with unmeasurable predictions. With the notions of error developed here it is possible to say that a learner infers the paper width function with error .5 inches via an everywhere 8.5 inch program.

[a]See also PAC learning, outside of inductive inference proper.[5]

1.3. Uncomputable Functions

There is no obvious reason, *a priori*, why all phenomena to be learned in nature should be assumed to be representable by a computable function. Nature might, after all, be uncomputable. However, nearly every learning criterion in the inductive inference literature cannot possibly learn uncomputable functions. There are learning criteria that are exceptions to this;[7,10,12] these criteria allow infinitely many errors, where informally, uncomputable sequences may have gone unnoticed.[b] The learning criteria introduced here allow the learning of uncomputable functions, and it is seen that larger sets of uncomputable functions become learnable as the amount of error is increased.

1.4. Dense vs. Not Dense Physical Properties

Suppose a scientist is using some fixed measurement apparatus to measure some physical property, and he has taken two distinct measurements. How many *different* physical values are there in the interval between the values of the two measurements? Finitely many or infinitely many? A somewhat simpler question is: How many different measurements are possible with that fixed measurement device within the interval between the values of the two measurements already made? Certainly, all actual devices have only finitely many discriminations possible between any two values. For example, a metric ruler is capable of measuring only finitely many different lengths less than one meter. And even an electron microscope can make only finitely many discriminations less than one meter. Nevertheless, scientists often take the stance that, although no measurement apparatus is capable of infinitely many discriminations, nature itself *does* have infinitely many values densely packed into every interval. Another philosophical stance is to say that nature's values are exactly those of the best possible measurement device—there are no unmeasurable values—and there are thus only finitely many values in any finite interval. Which stance is "correct" is a matter of debate, and which side of this debate one chooses to agree with directly affects how one models error in science. Notice that if one adopts the "dense" view of nature, then the measurement a plus or minus σ meters says that the length is one of infinitely many values, whereas taking the other view of nature leads to there being only finitely many possible values. Formally *defining* learning with error crucially depends on this decision. However the formal *results* presented here concerning learning criteria with error are completely independent of such a decision. It suffices, for example, to treat nature as if it is not dense, as far as the learnability issues discussed here are concerned.

2. Preliminaries

2.1. Notation

ω denotes the set of natural numbers, $\{0, 1, 2, 3, \ldots\}$, and Q the rationals. i, j,

[b] That these learning criteria are capable of learning uncomputable functions is not noticed.

$k, l, m, n, s, t, u, v, w, x, y, z$ range over ω. a, b, c range over $\omega^* = (\omega \cup \{*\})$, where $(\forall n \in \omega)[n < *]$. $\in, \subseteq$, and $\subset$ denote respectively membership, containment, and proper containment of sets. $\text{card}(S)$ denotes the cardinality of set S. $\emptyset$ denotes the empty set. 2^S denotes the power set of set S. If X is a set of sets, then a *chain* of X is a subset Y of X such that for all $\alpha, \beta \in Y$, either $\alpha \subseteq \beta$ or $\beta \subseteq \alpha$. For functions f, $\text{graph}(f)$ denotes $\{(x, y) | y = f(x)\}$. $f \restriction S$ denotes f restricted to domain $S \subseteq \omega$. $\lambda x[f(x)]$ denotes f. T denotes the set of total functions from ω to ω and $R \subset T$ denotes the set of total recursive functions. T and ϑ range over subsets of T and R, respectively. f, g, h, range over T, ϕ, δ, ϵ range over R. e, p, q, r denote programs. For $S \subseteq \omega$ and a set of functions T, $\mathsf{T}\restriction S$ denotes $\{g | g = f \restriction S \text{ and } f \in \mathsf{T}\}$. The sequence $\langle \phi_i \rangle_{i \in \omega}$ denotes an arbitrary *acceptable programming system*[6,8,9,11] of the partial recursive functions: $\omega \to \omega$. It is useful to speak of i as the program for ϕ_i. Call f an *n-variant* of g (written $f =^n g$) iff $\text{card}(\{x | f(x) \neq g(x)\}) \leq n$. Say that $f =^* g$ iff $\text{card}(\{x | f(x) \neq g(x)\}) \in \omega$. Let $\delta \in R$. Then $f =_\delta g$ means that $|f(x) - g(x)| \leq \delta(x)$ for every x. Say that $f =_\delta^n g$ iff $\text{card}(\{x : |f(x) - g(x)| > \delta(x)\}) \leq n$. Say that $f =_\delta^* g$ iff $\text{card}(\{x : |f(x) - g(x)| > \delta(x)\}) \in \omega$. The quantifiers $\overset{\infty}{\exists}$ and $\overset{\infty}{\forall}$ respectively mean 'there are infinitely many' and 'for all but finitely many'. δ_2 *weakly majorizes* δ_1 (written $\delta_1 \prec \delta_2$) iff $[\overset{\infty}{\exists} x(\delta_1(x) < \delta_2(x))$ and $\overset{\infty}{\forall} x(\delta_1(x) \leq \delta_2(x))]$. $\lfloor m/n \rfloor$ denotes the greatest integer $\leq (m/n)$, and $\lceil m/n \rceil$ denotes the least integer $\geq (m/n)$. σ, τ range over finite sequences of natural numbers, i.e., σ, τ are functions from finite initial segments of ω into ω.

2.2. EX-identification

A learner is modeled by an *inductive inference machine.*

Definition 1 An *inductive inference machine* (IIM) is an algorithmic device with no *a priori* bounds on how much time or memory resource it shall use, that takes as its input the graph of a function an ordered pair at a time (in any order), and occasionally outputs computer programs.[4]

An IIM M is a function from finite sequences of graph elements to computer programs. In order that it always be defined, M of the empty sequence is set to 0. Also, it is useful to write $M(f) = p$ to mean that there is $\sigma \subset \text{graph}(f)$ such that for all σ' such that $\sigma \subseteq \sigma' \subset \text{graph}(f)$, $M(\sigma') = p$.

The following definition emanates from Gold and is the criterion by which an IIM is judged to have successfully learned a theory.

Definition 2 (a) An IIM M *EX_b^a-identifies* a function f (written $f \in EX_b^a(M)$) iff M, when fed the graph of f in any order, outputs over time computer programs, making at most b mind changes, the last of which computes a function g such that $f =^a g$.[4,1,2] (EX for "explains".) (b) $EX_b^a = \{F | (\exists M)[F \subseteq EX_b^a(M)]\}$

For ease of presentation, EX_b^0 is written EX_b. EX_* is written as EX.

This last definition is a learning criterion for learning accurate data with possibly *incomplete* hypotheses, since it is only when the theory fails to make a prediction at all on some input does the anomaly become a relevant one. For, if a program is converged to that makes only finitely many incorrect predictions, this will eventually

be discovered and may be patched in the limit. So, as it is, it is not an adequate model of a learner making false predictions. And further, in as much as it models a learner that produces false or anomalous hypotheses, the hypotheses are correct infinitely often. The learning criteria introduced in this paper are generalizations of this criterion, allowing an IIM to learn a function with a program that is imprecise.

3. EX^{δ}

Definition 3 (a) An IIM M $EX_b^{a,\delta}$-*identifies* a function f (written $f \in EX_b^{a,\delta}(M)$) iff M, when fed the graph of f in any order, outputs over time computer programs, making at most b mind changes, the last of which computes a function g such that $f =_{\delta}^{a} g$. (b) $EX_b^{a,\delta} = \{F | (\exists M)[F \subseteq EX_b^{a,\delta}(M)]\}$.

$EX_b^{a,\delta}$-identification is learning with at most b mind changes, with error δ over ω, with possibly a anomalies. For ease of presentation, $EX_b^{0,\delta}$ is written EX_b^{δ}.

Lemma 1 *Suppose $\delta, \delta_1, \delta_2$ are given. Let $h(x) \geq 2\delta(x) + 1$ for all x such that $\delta(x) > 0$. Let $\delta_1 \prec \delta_2$. Let $\vartheta_h = \{f | \phi_{f(0)} =^1 f \ \& \ (\forall x \neq 0)(f(x) \in \{0, h(x)\})$ $\& \ (\forall x \neq 0)(\delta_2(x) \leq \delta_1(x) \Rightarrow f(x) = 0)\}$. Then $\vartheta_h \in (EX_0^{1,\delta} - EX_*^{0,\delta})$.*

The proof of the lemma is a simple modification of the proof that $EX^1 - EX^0 \neq \emptyset$ by Case and Smith.[2]

Theorem 1 $(\forall b)(\forall \delta_1, \delta_2)\ [\delta_1 \prec \delta_2$ *iff* $EX_b^{\delta_1} \subset EX_b^{\delta_2}]$.

Proof. **Proof 1.** ($\Rightarrow$) Suppose $\delta_1 \prec \delta_2$. Noneffectively choose the least natural number x_0 such that $(\forall x > x_0)(\delta_1(x) \leq \delta_2(x))$. To show containment, suppose $\vartheta \in EX_b^{\delta_1}$. Then there is an M such that M $EX_b^{\delta_1}$-identifies every $f \in \vartheta$. Define IIM M^1 as follows:

For all $\sigma \subset \text{graph}(f)$, let

$$M^1(\sigma) = \begin{cases} e & \text{if } f\restriction\{x | x \leq x_0\} \subseteq \sigma \\ \uparrow & \text{otherwise} \end{cases}$$

where

$$\phi_e(x) = \begin{cases} f(x) & \text{if } x \leq x_0 \\ \phi_{M(\sigma)}(x) & \text{otherwise} \end{cases}$$

Thus, $\vartheta \in EX_b^{\delta_2}$.

Showing proper containment, let $h(x) = 2\delta_2(x) \dot{-} 1$, and consider ϑ_h as defined in Lemma 1. Let M be the machine that waits for $f(0)$ to be input and then outputs a program q such that $\phi_q(0) = f(0)$ and for all $x > 0$, $\phi_q(x) = \delta_2(x)$. Note that $\delta_2(x)$ is within $\delta_2(x)$ of 0 and $2\delta_2(x) \dot{-} 1$, for all x. So, $f =_{\delta_2} \phi_{M(f)}$ for every $f \in \vartheta_h$. Thus, $\vartheta_h \in EX_0^{\delta_2}$.

Suppose by way of contradiction that $\vartheta_h \in EX^{\delta_1}$. Then there is M such that M EX^{δ_1}-identifies every $f \in \vartheta_h$. Define M^2 as follows:

For all $\sigma \subset \text{graph}(f)$, let

$$M^2(\sigma) = \begin{cases} e & \text{if } M(\sigma)\downarrow \\ \uparrow & \text{otherwise} \end{cases}$$

where

$$\phi_e(x) = \begin{cases} f(x) & \text{if } x = 0 \\ 2\delta_2(x) \dot{-} 1 & \text{if } x > 0, \delta_1(x) < \delta_2(x) \text{ and } \phi_{M(\sigma)}(x) =_{\delta_1} 2\delta_2(x) \dot{-} 1 \\ 0 & \text{otherwise} \end{cases}$$

M^2 identifies each $f \in \vartheta_h$ *exactly*. So $\vartheta_h \in EX$, and this contradicts Lemma 1. The intuition is that ϑ_h is what EX^{δ_2} can just barely do.

($\Leftarrow$) Now the converse is proved. *Case 1.* Suppose not $\overset{\infty}{\exists} x(\delta_1(x) < \delta_2(x))$. Then $\overset{\infty}{\forall} x(\delta_2(x) \leq \delta_1(x))$. The argument used in this proof above to show containment may be used with δ_1 and δ_2 switched to show $EX_b^{\delta_2} \subseteq EX_b^{\delta_1}$. So, $EX_b^{\delta_1} \not\subset EX_b^{\delta_2}$.

Case 2. Suppose not $\overset{\infty}{\forall} x(\delta_1(x) \leq \delta_2(x))$. Then $\overset{\infty}{\exists} x(\delta_2(x) < \delta_1(x))$. Let $h(x) = 2\delta_1(x) \dot{-} 1$. Then $\vartheta_h \in (EX_0^{\delta_1} - EX_*^{\delta_2})$ by the argument used in this proof above to show proper containment but with δ_1 and δ_2 switched. Therefore, $EX_b^{\delta_1} \not\subset EX_b^{\delta_2}$.

Proof 2. ($\Rightarrow$) The argument to show containment is identical to the argument in Proof 1.

To show proper containment, let $\mathsf{T}=\{f|(\forall x \in \omega)(f(x) \in \{0, 2\delta_2(x) \dot{-} 1\})\}$. Let M be the machine that ignores its input and outputs the program r such that $\phi_r = \delta_2$. $f =_{\delta_2} \delta_2$ for every $f \in \mathsf{T}$. Thus, $\mathsf{T} \in EX_0^{\delta_2}$.

Suppose M' EX^{δ_1}-identifies every $f \in \mathsf{T}$. Then, for every $f \in \mathsf{T}$, $\phi_{M'(f)} =_{\delta_1} f$. Let $X = \{x|\delta_1(x) < \delta_2(x)\}$; X is recursive and infinite. So, $\phi_{M'(f)}\restriction X =_{\delta_1} f\restriction X$ for every $f \in \mathsf{T}$. Define M'' as follows:

For all $\sigma \subset \text{graph}(f)$, let

$$M''(\sigma) = \begin{cases} e & \text{if } M'(\sigma)\downarrow \\ \uparrow & \text{otherwise} \end{cases}$$

where

$$\phi_e(x) = \begin{cases} 2\delta_2(x) \dot{-} 1 & \text{if } \phi_{M'(\sigma)}(x) =_{\delta_1} 2\delta_2(x) \dot{-} 1 \\ 0 & \text{otherwise} \end{cases}$$

Intuitively, since $\delta_1(x) < \delta_2(x)$ for $x \in X$, $\phi_{M'(f)}(x)$ is within $\delta_1(x)$ of only one from the set $\{0, 2\delta_2(x) \dot{-} 1\}$, so one might as well output a program that is *exactly* correct (within X), i.e., has range equal to $\{0, 2\delta_2(x) \dot{-} 1\}$. So, $\phi_{M''(f)}\restriction X = f\restriction X$, for every $f \in \mathsf{T}$. $\phi_{M''(f)}\restriction X$ is, for each f, some recursive function over X. But there are only countably many such recursive functions, and $\text{card}(\mathsf{T}\restriction X)=2^\omega$. So there must be $f \in \mathsf{T}$ such that $\phi_{M''(f)}\restriction X \neq f\restriction X$, and this is a contradiction.

($\Leftarrow$) The converse is identical to Proof 1. □

The first proof above shows new recursive functions become identifiable as greater error is allowed, and the second shows that new non-recursive functions become identifiable as greater error is allowed.

It is possible to prove a similar result where error is over the rationals instead of the naturals, corresponding to the "dense" case discussed in the Introduction. Let

N be bijectively coded onto the rationals via $\rho : \omega \to Q$. Let $f \approx_\delta g$ mean that $|\rho(f(x)) - \rho(g(x))| \leq \rho(\delta(x))$, for all x. Notice that $\rho(f(x))$, $\rho(g(x))$, and $\rho(\delta(x))$ are rational numbers, for all $x \in N$. Say that $f \approx_\delta^n g$ iff $\text{card}(\{x : |\rho(f(x)) - \rho(g(x))| > \rho(\delta(x))\}) \leq n$. Say that $f \approx_\delta^* g$ iff $\text{card}(\{x : |\rho(f(x)) - \rho(g(x))| > \rho(\delta(x))\}) < \omega$. δ_2 *weakly* Q*-majorizes* δ_1 (written $\delta_1 \lhd \delta_2$) iff $[\overset{\infty}{\exists} x\ (\rho(\delta_1(x)) < \rho(\delta_2(x)))$ and $\overset{\infty}{\forall} x\ (\rho(\delta_1(x)) \leq \rho(\delta_2(x)))]$. The next definition is the EX^δ analog over the rationals.

Definition 4 (a) An IIM M $ratEX_b^{a,\delta}$*-identifies* a function f (written $f \in ratEX_b^{a,\delta}(M)$) iff M, when fed the graph of f in any order, outputs over time computer programs, making at most b mind changes, the last of which computes a function g such that $g \approx_\delta^a f$. (b) $ratEX_b^{a,\delta} = \{F | (\exists M)\ [F \subseteq ratEX_b^{a,\delta}(M)]\}$.

For ease of presentation, $ratEX_0^{0,\delta}$ is written $ratEX^\delta$.

$ratEX^{a,\delta}$-identification is learning with error δ over Q, with possibly a anomalies. Note that for any x, there are infinitely many values that $\phi_{M(f)}(x)$ may attain. A prediction is correct if it is any one of some recursive infinite subset of ω.

Theorem 2 $(\forall b)(\forall \delta_1, \delta_2)\ [\delta_1 \lhd \delta_2$ *iff* $ratEX_b^{\delta_1} \subset ratEX_b^{\delta_2}]$.

The proof of this may be obtained from Proof 1 of Theorem 1 by the following modifications: (i) EX must be replaced by $ratEX$, (ii) every time a function is written in the context of referring to the metric size of the error, that function must be replaced by ρ of the function, and (iii) an analog of Lemma 1 must be cited. All other changes needed are minimal and straightforward.[c]

4. $(\epsilon, v)EX$

The EX^δ criterion of learning equips inductive inference for its application to science and many other natural scenarios. One application, however, that it does not address is learning languages with error. In language learning the object is to learn a two-valued function, $f : \omega \to \{0, 1\}$, where each $x \in \omega$ codes some expression and 1 (0) means the expression is (is not) a sentence in the language. Learning a language with error, where error is defined as in Definition 3, is useless since there is no interesting notion of metric possible over the set $\{0, 1\}$.

Another useful piece of equipment that inductive inference could use to increase its applicability to natural scenarios such as science is to allow learning of noisy data via smoothing, or learning with the option of disregarding certain data points and interpolating.

The next learning criterion is strictly more general than $EX_b^{a,\delta}$-identification, but allows one the ability to model a number of new things: smoothing, interpolation, noise ignoring, and language error. The following definition specifies when a function g is to be considered a *smoothing* of f.

Definition 5 Let $v : \omega \to \omega \cup \{*\}$ be a recursive function such that $v(x) > 0$ for all x. Suppose $g, f \in T$, $\epsilon, \delta \in R$, and $\delta(x) < v(x)$ for all x. g is $(\epsilon, v)^{a,\delta}$*-smooth* to f iff for all but a many x, there is y such that (i) $|x - y| \leq \epsilon(x)$, (ii) $|f(x) - f(y)| \leq v(x)$, and (iii) $|g(x) - f(y)| \leq \delta(x)$.

[c]For an extended philosophical discussion of the content of this section, see Changizi.[3]

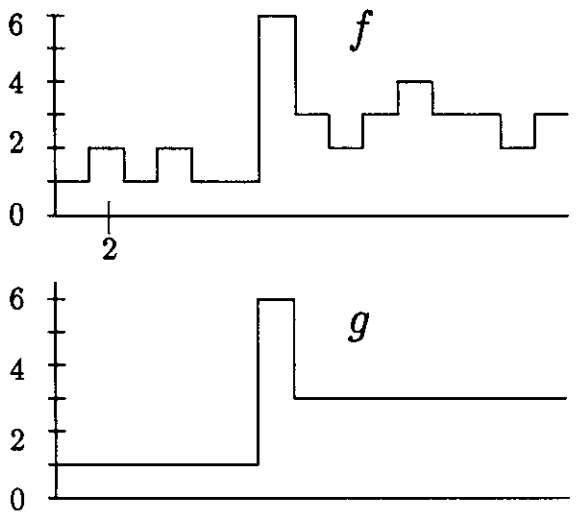

Fig. 1. g is $(\lambda x[1], \lambda x[1])$-smooth to f.

Definition 5 has three parts. (i) specifies the *radius of smoothing*; how far into the nearby domain one may look for a y with the appropriate properties (specified in parts (ii) and (iii)). (ii) specifies the *threshold*, an upper bound on how different from the actual value $f(x)$ the nearby domain element y can map to. The idea is that sometimes the value $f(x)$ is such a strong outlying data point that one may not want to treat it as noise or useless information. For example, in image recognition one wants to recognize that there is a boundary to a field, but ignore the small variation within the field. (iii) specifies the *error* allowed; it indicates that $g(x)$ must be within $\delta(x)$ of $f(y)$, for some y satisfying the previous two properties. Intuitively, $g(x)$ is required to be close (within $\delta(x)$) to what some nearby (within $\epsilon(x)$) domain element maps to ($f(y)$), so long as the value on x that g wrongly predicts ($f(x)$) is close enough (within $v(x)$) to what the nearby domain element maps to ($f(y)$).

It is supposed that $v(x) > \delta(x)$ for all x since, intuitively, otherwise there is never any reason to smooth, since all of the noise is being taken care of by the error δ. That is, smoothing is useful when there is some prediction that is wrong, i.e., not within $\delta(x)$, but is still inside the threshold ((ii) above). Allowing the threshold to be smaller than the error shunts this use. Also, it is required that $v(x) > 0$ for all x since if ever $v(x) = 0$, then $\epsilon(x)$ might as well be $= 0$. When $a = 0$ and $\delta = \lambda x[0]$, it is useful to abbreviate to 'g is (ϵ, v)-smooth to f.' Note that if $\epsilon = \lambda x[0]$, then 'g is $(\epsilon, v)^{a,\delta}$-smooth to f' means that $g =_\delta^a f$. Note also that the smooth relation is not symmetric. For example, in Figure 1, f is not $(\lambda x[1], \lambda x[1])$–smooth to g. Look at $f(2)$ and notice that there is no y such that (i) $|x - y| \leq 1$, (iii) $f(x) = g(y)$.

The following definition establishes what it means to learn via smoothing. It is

a generalization of Definition 3.

Definition 6 (a) An IIM M $(\epsilon, v)EX_b^{a,\delta}$*-identifies* a function f (written $f \in$ (ϵ, v) $EX_b^{a,\delta}$ (M)) iff M, when fed the graph of f in any order, outputs over time computer programs, making at most b mind changes, the last of which computes a function g such that g is $(\epsilon, v)^{a,\delta}$-smooth to f. (b) $(\epsilon, v)EX_b^{a,\delta} = \{F | (\exists M)[F \subseteq (\epsilon, v)EX_b^{a,\delta}(M)]\}$.

$(\epsilon, v)EX_b^{a,\delta}$-identification is learning with b mind changes via smoothing with radius ϵ, threshold v, error δ over ω, and possibly a anomalies. For ease of presentation, $(\epsilon, v)EX_b^{0,\delta}$ is written $(\epsilon, v)EX_b^{\delta}$, and when $\delta = \lambda x[0]$ it is written $(\epsilon, v)EX_b$. Many of the theorems are proved with $a = 0$ anomalies. Each may be generalized, but the proofs become less readable.

4.1. $(\epsilon, v)EX_b$-identification with a Threshold

The following theorem shows that tightening the radius of smoothing decreases the power of the learner for the special case that the threshold $v(x) < *$ for infinitely many x. Proving the theorem for the more general case will be seen to be impossible. Let $\epsilon_1 \prec \epsilon_2$ on X mean that ϵ_2 weakly majorizes ϵ_1 on the (possibly) restricted domain X.

Theorem 3 *Suppose v is such that $X = \{x | v(x) < *\}$ is infinite. Then for all b, ϵ_1, and ϵ_2, $\epsilon_1 \prec \epsilon_2$ on X iff (ϵ_1, v) $EX_d \subset (\epsilon_2, v)$ EX_d.*

Proof. We prove only the $\Rightarrow$ direction. Suppose $\epsilon_1 \prec \epsilon_2$. The difficult task is to show proper containment. Notice that $\epsilon_1 \prec \epsilon_2$ implies that there are infinitely many x such that $\epsilon_1(x) < \epsilon_2(x)$. The goal is to describe a set of recursive functions that isolates the problem to those areas where $(\epsilon_2, v)EX_b$ is more powerful than $(\epsilon_1, v)EX_b$. Let $b^0 = 1$ and $b^{t+1} =$ the least $x[x > b^t + \epsilon_2(b^t)$ and $\epsilon_1(x) < \epsilon_2(x)$ and $v(x) < *]$. Let $c_s^0 = s$ and $c_s^{t+1} = c_s^t + 2v(b^t) + 3$.

Set $\vartheta =$

$\{f | \phi_{f(0)} =^1 f$ and
$(\forall t > 0)$ $(f(b^t) \in \{c_{f(0)}^t + v(b^t) + 1, c_{f(0)}^t + v(b^t) + 2\}$ and
$f(b^t + \epsilon_2(b^t)) = c_{f(0)}^t + v(b^t) + 1$ and
$(\forall x \in \{b^t + 1, \ldots, b^{t+1} - 1\} - \{b^t + \epsilon_2(b^t)\})(f(x) = c_{f(0)}^{t+1}))\}$.

Consider ϑ. Define $h(x) =$ the least t such that $x \geq b^t$. Let M be the IIM that outputs p such that

$$\phi_p(x) = \begin{cases} f(0) & \text{if } x = 0 \\ c_{f(0)}^{h(x)} + v(b^{h(x)}) + 1 & \text{if } x \in \{b^{h(x)}, b^{h(x)} + \epsilon_2(b^{h(x)})\} \\ c_{f(0)}^{h(x)} + 2v(b^{h(x)}) + 3 & \text{if } x \in \{b^{h(x)} + 1, \ldots, b^{h(x)+1} - 1\} - \\ & \quad \{b^{h(x)} + \epsilon_2(b^{h(x)})\}) \end{cases}$$

Pick $f \in \vartheta$. Pick $x > 0$ arbitrarily. If $x \in \{b^{h(x)} + 1, \ldots, b^{h(x)+1} - 1\} - \{b^{h(x)} + \epsilon_2(b^{h(x)})\}$, then $f(x) = \phi_p(x)$ via the third line in the definition of ϕ_p and the fourth line in the definition of ϑ. If $x = b^{h(x)} + \epsilon_2(b^{h(x)})$, then $f(x) = \phi_p(x)$ via the second line of the definition of ϕ_p and the third line of the definition of ϑ. If $x = b^{h(x)}$, then

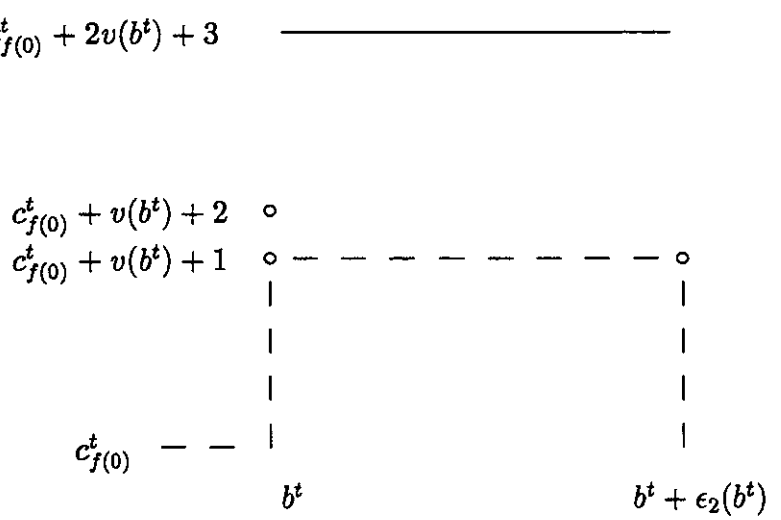

Fig. 2. ϑ consists of functions which have the property shown.

pick $y = b^{h(x)} + \epsilon_2(b^{h(x)})$. Then $f(y) = f(b^{h(x)} + \epsilon_2(b^{h(x)})) = c^{h(x)}_{f(0)} + v(b^{h(x)}) + 1$, so $\phi_p(x) = f(y)$, and also, $|f(y) - f(x)| \leq 1 \leq v(x)$ since $v(x) > 0$ for all x by supposition. So, ϕ_p is (ϵ_2, v)-smooth to f, for every $f \in \vartheta$. Figure 2 displays this. Therefore, $\vartheta \in (\epsilon_2, v)EX_0$.

It is needed to show that $\vartheta \notin (\epsilon_1, v)EX_d$. Suppose by way of contradiction that $\vartheta \in (\epsilon_1, v)EX_*$. Then since for each t, $\epsilon_1(b^t) < \epsilon_2(b^t)$, $b^t + \epsilon_2(b^t)$ is not within $\epsilon_1(b^t)$ of b^t, and this is the only domain element nearby that maps to a number within $v(b^t)$ of $f(b^t)$. Therefore, any IIM attempting to $(\epsilon_1, v)EX_*$-identify ϑ must get it *exactly* correct when guessing what b^t maps to, for each t. Thus, $\vartheta \in EX_*$. A theorem similar to Lemma 1 may be proven that shows that this is impossible. □

The next theorem shows how much the threshold must be raised in order for the power of learner to increase. It is supposed that there are infinitely many points for which the radius of smoothing is non-zero, since otherwise there are only finitely many points where a threshold can be employed, and in such cases, there is no increase in learnability due to the fact that finitely many data points can always be explicitly built into a hypothesis.

Theorem 4 *Suppose ϵ is such that $X = \{x|\epsilon(x) > 0\}$ is infinite. Then for all d, v_1 and v_2, $v_1 \prec v_2$ on X iff $(\epsilon, v_1)\ EX_d \subset (\epsilon, v_2)EX_d$.*

Proof. Suppose $v_1 \prec v_2$ on X. The proof of containment is similar to that in Theorem 1. To show proper containment, a set is described that is in (ϵ, v_2) $EX_d - (\epsilon, v_1)EX_d$. The set, intuitively, isolates the problem to those areas where $(\epsilon, v_2)EX_d$ is more powerful than $(\epsilon, v_1)EX_d$. Let $b^0 =$ the least $x[v_1(x) < v_2(x)$ and $\epsilon(x) > 0$ and $x > 0]$. Let $b^{t+1} =$ the least $x[x > b^t + \epsilon(b^t)$ and $v_1(x) < v_2(x)$ and $\epsilon(x) > 0]$. Let $c^0_s = s$ and $c^{t+1}_s = c^t_s + v_2(b^t) = s + \sum_{j=1}^{t} v_2(b^j)$.

Set $\vartheta =$

$\{f|\phi_{f(0)} =^1 f$ and
$(\forall x \in \{1, \ldots, b^0\})(f(x) = 0)$ and

$$(\forall t > 0)\ (f(b^t) \in \{c^t_{f(0)}, c^t_{f(0)} + v_2(b^t)\} \text{ and}$$
$$(\forall x \in \{b^t + 1, \ldots, b^{t+1} - 1\})(f(x) = c^t_{f(0)}))\}.$$

Define $h(x)$ = the least t such that $x \geq b^t$. Let M be the IIM that outputs p such that

$$\phi_p(x) = \begin{cases} f(0) & \text{if } x = 0 \\ 0 & \text{if } x \in \{1, \ldots, b^0\} \\ c^{h(x)}_{f(0)} & \text{otherwise} \end{cases}$$

Pick $f \in \vartheta$ arbitrarily. If $x \leq b^0$, then the first and second lines of the definitions of ϑ and ϕ_p show that $\phi_p(x) = f(x)$. If $x > b^0$ and $\exists t(x \in \{b^t + 1, \ldots, b^{t+1} - 1\})$, then the fourth line of the definition of ϑ and the third line of the definition of ϕ_p imply that $\phi_p(x) = f(x)$. If $x > b^0$ and $\exists t(x = b^t)$, then line three of the definition of ϕ_p shows that $\phi_p(x) = c^{h(x)}_{f(0)}$, but the third line of the definition of ϑ says that $f(x) \in \{c^{h(x)}_{f(0)}, c^{h(x)}_{f(0)} + v_2(b^{h(x)})\}$. If $f(x) = c^{h(x)}_{f(0)}$, then $\phi_p(x) = f(x)$. Otherwise $f(x) = c^{h(x)}_{f(0)} + v_2(b^{h(x)})$, and $\phi_p(x) \neq f(x)$. However, $\phi_p(x) = f(x+1)$, $\epsilon(x) > 0$ and $f(x+1)$ is inside the threshold. Thus, ϕ_p is (ϵ, v_2)–smooth to f, for every $f \in \vartheta$. So $\vartheta \in (\epsilon, v_2)EX_0$.

It must be shown that $\vartheta \notin (\epsilon, v_1)EX_d$. The proof of this is similar to that in Theorem 3. □

Proofs similar to Proof 2 of Theorem 1 can be used to show that larger sets of uncomputable functions may be learned via the strengthened learner in Theorem 3 and 4. Also, analogous notions of error over the rationals lead to the analogous theorems.

4.2. $(\epsilon, v)EX_b$-identification without a Threshold

Theorems 3 and 4 characterize the hierarchy when there is a threshold ($v(x) < *$) on infinitely many points. What happens when there is, almost everywhere, no threshold at all? There are reasons to ask this question. (i) When smoothing away noise or spurious data it may be desired to smooth a data point to a nearby value, *no matter how different.* (ii) It may be useful to allow interpolation between good data points, ignoring the bad data in between, without regard to how bad the bad data is. (iii) Modeling language learning with error requires looking only at {0,1}-valued functions—a special case of finite range functions. For finite range functions, the threshold may be larger than the largest element of the range, almost everywhere. And for two-valued functions, the notion of a threshold is not even useful. It is useful to know when new sets of finite range functions are learnable.

The following results are generalized to cover all of these cases. It is seen that this new question breaks the mold set by the previous three major error results, Theorems 1, 3 and 4. Informally, the results for language error become more difficult. It should be noted that there is no analogous threshold result to Theorem 4 here, as what is being considered now is the special case when there is no threshold almost everywhere.

A corollary of the next theorem shows that Theorem 3 does not hold for the case where there is no threshold almost everywhere. Let $T_a = \{f | \max(\text{range}$

$(f))-\min(\text{range}\ (f)) = a\}$. Let $R_a = T_a \cap R$.

Theorem 5 *Fix $a \in \omega \cup \{*\}$. Let v be the threshold function. Suppose $\overset{\infty}{\forall} x$ $(a < v(x))$. $(\forall b)$ $(\forall \epsilon_1, \epsilon_2)$ $[(\overset{\infty}{\forall} x(\epsilon_1(x) < \epsilon_2(x))) \not\Rightarrow (\exists \vartheta \subseteq R_a)$ $(\vartheta \in ((\epsilon_2, v)EX_b - (\epsilon_1, v)EX_b))]$.*

Proof. Fix a and v as in the statement of the theorem. Let $\epsilon_1 = \lambda x[x], \epsilon_2 = \lambda x[2x]$. Noneffectively choose the least natural number, x_0, such that $(\forall x > x_0)(a < v(x))$. Define M as follows:

For all $\sigma \subset \text{graph}(f)$, let

$$\phi_{M(\sigma)}(x) = \begin{cases} f(x) & \text{if } x \leq x_0 \\ f(0) & \text{otherwise} \end{cases}$$

Note that $R_a \in (\epsilon_1, v)EX_0$, since M just waits until the first x_0 data points come in and outputs its only guess. Intuitively, $\phi_{M(f)}$ is almost always allowed to look back all the way to $x = 0$ to decide where to smooth to, and so simply outputs $f(0)$ almost everywhere. Since $f(0)$ is within $v(x)$ of $f(x)$ for all $x > x_0$, then this works. □

Under what conditions, then, are new sets learnable for this case of no threshold almost everywhere? No necessary and sufficient conditions have yet been proven. However, the following theorem shows a very natural condition that is sufficient to increase learnability. The condition is that the radius of smoothing of each learner be constant almost everywhere; and that the constant of one learner be greater than the other. In lieu of a general theorem this is the next best thing, since it would likely be the most naturally interesting corollary to a general theorem.

Theorem 6 *Fix $a \in \omega \cup \{*\}$, $a \geq 2$. Let v be the threshold function and suppose $\overset{\infty}{\forall} x(a \leq v(x))$. Suppose $\overset{\infty}{\forall} x(\epsilon_1(x) = c)$ and $\overset{\infty}{\forall} x(\epsilon_2(x) = d)$. $(\forall b)(\forall \epsilon_1, \epsilon_2)$ [If $c < d$, then $(\epsilon_1, v)EX_b \subset (\epsilon_2, v)EX_b$ and $(\exists \vartheta \subseteq R_a)(\vartheta \in ((\epsilon_2, v)EX_0 - (\epsilon_1, v)EX_*))$].*

Proof. Fix a and v as in the statement of theorem. Suppose $c < d$. The proof of containment is similar to that in Theorem 3. To show proper containment, it is necessary to show that there is a set that is in $(\epsilon_2, v)EX_b - (\epsilon_1, v)EX_b$. Let x_0 be the least x such that for all $x' > x, \epsilon_1(x) = c$ and $\epsilon_2(x) = d$. Let $a_1 | a_2$ mean that a_1 divides a_2.

Set $\vartheta =$

$\{f | \phi_{f(0)} =^1 f$ and
$(\forall x > 0)(f(x) \in \{0, 1\})$ and
$(\forall 1 \leq x \leq x_0)(f(x) = 0)$ and
$(\forall x > x_0)(2d - 1 | x \Rightarrow f(x) = 0)$ and
$(\forall x, y)[(x_0 \leq n(2d-1) < x, y < (n+1)(2d-1)) \Rightarrow f(x) = f(y)]\}$.

Note that $\vartheta \subseteq R_2 \subseteq R_a$. Let M be the IIM that outputs p such that $\phi_p = \lambda x[0]$. Pick $f \in \vartheta$. Pick x arbitrarily. If $1 \leq x \leq x_0$, then $f(x) = \phi_p(x) = 0$. Suppose that $x > x_0$. If $2d - 1 | x$, then $f(x) = \phi_p(x) = 0$. If not $2d - 1 | x$, then for some $n \in \omega$, $n(2d-1) < x < (n+1)(2d-1)$, and x is within d of either $n(2d-1)$ or $(n+1)(2d-1)$. Therefore, ϕ_p is (ϵ_2, v)-smooth to f. Thus, $\vartheta \in (\epsilon_2, v)EX_0$.

It is necessary now to show that $\vartheta \notin (\epsilon_1, v)EX_b$. Suppose by way of contradiction that $\vartheta \in (\epsilon_1, v)EX_*$. $\epsilon_1(x) = c < d = \epsilon_2(x)$, for all $x > x_0$, so every $x > x_0$ directly in between the multiples of $2d-1$ is not within c of any multiple of $2d-1$. That is, if there is $x > x_0$ such that, for some n, $x = n(2d-1)+d$, then such an x is not within c of any multiple of $2d-1$. There are infinitely many such points, and any IIM attempting to $(\epsilon_1, v)EX_*$-identify a function $f \in \vartheta$ must be *exactly* correct on these domain elements, as $\overset{\infty}{\forall} x(\epsilon_1(x) = c)$. Therefore, $\vartheta \in EX_*$ and a theorem similar to Lemma 1 can be proved that says that this is impossible. □

It is also possible to prove that there are new uncountable sets that become learnable, and thus new uncomputable functions.

5. Order Type

Let $\Delta = \{EX^\delta | \delta \in R\}$ be ordered by $\subseteq$. The following theorem helps to pin down the order type of Δ. Since $\delta_1 \prec \delta_2$ iff $EX^{\delta_1} \subset EX^{\delta_2}$, then the results below for Δ also hold for the analogous threshold of smoothing hierarchy, and, when $\overset{\infty}{\exists} x(v(x) < *)$, the radius of smoothing hierarchy.

Theorem 7 *1. Every chain of Δ is dense (i.e., between any two learning criteria there is another learning criteria.)*

2. Let δ_1 be recursive and $\overset{\infty}{\exists} x(\delta_1(x) > 0)$. Then there is δ such that $EX^{\delta_1} - EX^\delta \neq \emptyset$ and $EX^\delta - EX^{\delta_1} \neq \emptyset$. In fact, card$(\{\delta | EX^{\delta_1} - EX^\delta \neq \emptyset$ and $EX^\delta - EX^{\delta_1} \neq \emptyset\}) = \omega$ (i.e., there are incomparable learning criteria in Δ.)

3. Suppose $EX^{\delta_1}, EX^{\delta_2} \in \Delta$. There is $EX^\delta \in \Delta$ such that
(a) $EX^{\delta_1} \subseteq EX^\delta$,
(b) $EX^{\delta_2} \subseteq EX^\delta$, and
(c) For all $EX^\epsilon \in \Delta$ such that $EX^{\delta_1} \subseteq EX^\epsilon$ and $EX^{\delta_2} \subseteq EX^\epsilon$, $EX^\epsilon \subseteq EX^\delta$ iff $EX^\epsilon = EX^\delta$.
(i.e., for every finite set of criteria there is a unique weakest learning criterion that is more powerful than each criterion in the set.)

4. The proposition obtained by replacing every occurrence of "$\subseteq$" by "$\supseteq$" in number three is true (i.e., for every finite set of criteria there is a unique strongest learning criterion that is less powerful than each criterion in the set.)

5. There is a set A such that if $EX^\delta \supseteq EX^\phi$ for all $EX^\phi \in A$, then there is $EX^\epsilon \in \Delta$ such that $EX^\epsilon \supseteq EX^\phi$ for all $EX^\phi \in A$ and $EX^\epsilon \subset EX^\delta$ (i.e., there are sets of criteria for which there is no least upper bound.) Also, the analogous proposition for greatest lower bound is true.

Proof. (1) Suppose $EX^{\delta_1} \subset EX^{\delta_2}$. Then $\delta_1 \prec \delta_2$. Let $X = \{x | \delta_1(x) < \delta_2(x)\}$ and note that X is recursive. Let $(x_1, x_2, \ldots)$ be an enumeration of X. Let

$$\delta(x) = \begin{cases} \delta_2(x) & \text{if } x = x_i \text{ for some even } i \\ \delta_1(x) & \text{otherwise} \end{cases}$$

Now, $\delta_1 \prec \delta$. so $EX^{\delta_1} \subset EX^\delta$, and $\delta \prec \delta_2$. so $EX^\delta \subset EX^{\delta_2}$.

(2) Let $X = \{x|\delta_1(x) > 0\}$ and note that X is recursive. Let $(x_1, x_2, \ldots)$ be an enumeration of X. Let

$$\delta(x) = \begin{cases} \delta_1(x) + 1 & \text{if } x = x_i \text{ for some even } i \\ \delta_1(x) - 1 & \text{otherwise} \end{cases}$$

Then $\overset{\infty}{\exists} x(\delta_1(x) < \delta(x))$ and $\overset{\infty}{\exists} x(\delta(x) < \delta_1(x))$. In fact, for $R(i)$ any recursive relation such that $\text{card}(\{i|R(i)\}) = \text{card}(\{i|\text{not } R(i)\}) = \omega$, let

$$\delta_R(x) = \begin{cases} \delta_1(x) + 1 & \text{if } x = x_i \text{ for some i, and } R(i) \\ \delta_1(x) - 1 & \text{otherwise} \end{cases}$$

Then $\overset{\infty}{\exists} x(\delta_1(x) < \delta_R(x))$ and $\overset{\infty}{\exists} x(\delta_R(x) < \delta_1(x))$ So, given EX^{δ_1} such that $\overset{\infty}{\exists} x(\delta_1(x) > 0)$, there are infinitely many mutually incomparable learning criteria.

(3) Let $\delta(x) = \max(\{\delta_1(x), \delta_2(x)\})$ for all $x \in \omega$. Suppose $\mathsf{T} \in EX^{\delta_1}$. Then there is M such that M EX^{δ_1}-identifies every $f \in \mathsf{T}$. So, for every $f \in \mathsf{T}$, $\phi_{M(f)} =_{\delta_1} f$, which means that for every $f \in \mathsf{T}$, $|\phi_{M(f)}(x) - f(x)| \leq \delta_1(x)$ for all x. But $\delta_1(x) \leq \delta(x)$ for all x since $\delta(x) = \max(\{\delta_1(x), \delta_2(x)\})$ for all x. So $\phi_{M(f)} =_{\delta} f$ for every $f \in \mathsf{T}$ and therefore $\mathsf{T} \in EX^{\delta}$. So $EX^{\delta_1} \subseteq EX^{\delta}$. An identical argument shows $EX^{\delta_2} \subseteq EX^{\delta}$. (a) and (b) are satisfied by EX^{δ}.

To show that (c) is satisfied, suppose $EX^{\delta_1} \subseteq EX^{\epsilon}$, $EX^{\delta_2} \subseteq EX^{\epsilon}$ and $EX^{\epsilon} \subseteq EX^{\delta}$. Want to show that $EX^{\delta} \subseteq EX^{\epsilon}$. Suppose by way of contradiction that $EX^{\delta} \not\subseteq EX^{\epsilon}$. Then $\overset{\infty}{\exists} x(\epsilon(x) < \delta(x))$. Let $X = \{x|\epsilon(x) < \delta(x)\}$ and note that X is infinite. But since $\delta(x) = \max(\{\delta_1(x), \delta_2(x)\})$ for all x, then for any x, $\epsilon(x) < \delta(x)$ iff $[\epsilon(x) < \delta_1(x)$ or $\epsilon(x) < \delta_2(x)]$. Letting $X_1 = \{x|\epsilon(x) < \delta_1(x)\}$ and $X_2 = \{x|\epsilon(x) < \delta_2(x)\}$, note that $X = X_1 \cup X_2$. Since X is infinite, at least one of the sets X_1, X_2 is infinite. Without loss of generality suppose that X_1 is infinite. Then $\overset{\infty}{\exists} x(\epsilon(x) < \delta_1(x))$. Since $EX^{\delta_1} \subseteq EX^{\epsilon}$, then $\overset{\infty}{\forall} x(\delta_1(x) \leq \epsilon(x))$. But then not $\overset{\infty}{\exists} x(\epsilon(x) < \delta_1(x))$, a contradiction.

(4) A similar proof to the proof of (3) may be given to show that any two criteria have a greatest lower bound.

(5) Let $\langle p_i \rangle_{i \in \omega}$ be an enumeration of the primes. For all $i \in \omega$, let

$$f_i(x) = \begin{cases} 1 & \text{if } x = p_i^n \text{ for some } n \in \omega \\ 0 & \text{otherwise} \end{cases}$$

Let $A = \{EX^{\phi} | \phi = f_i \text{ for some } i \in \omega\}$. Pick EX^{δ} arbitrarily such that $EX^{\delta} \supseteq EX^{\phi}$ for all $EX^{\phi} \in A$. So, $\overset{\infty}{\forall} x(\delta(x) \geq f_i(x))$ each $i \in \omega$. Let ϵ be defined as follows:

$$\epsilon(x) = \begin{cases} \delta(x) - 1 & \text{if } x = p_i^n \text{ for some } i, n \in \omega \text{ such that } \delta(p_i^n) > 0 \text{ and} \\ & \text{for all } m < n, \delta(p_i^m) = 0. \\ \delta(x) & \text{otherwise} \end{cases}$$

Note that $\epsilon \prec \delta$, so $EX^{\epsilon} \subset EX^{\delta}$. Pick i arbitrary. $f_i(x) = 1$ iff $x = p_i^n$ for some $n \in \omega$. Pick n least such that $\delta(p_i^n) > 0$. Then $\epsilon(p_i^n) = \delta(p_i^n) - 1$, and for all $k > n, \epsilon(p_i^k) = \delta(p_i^k)$. So, $\overset{\infty}{\forall} x(f_i(x) \leq \epsilon(x))$, and this holds for each $i \in \omega$. Therefore, $EX^{\epsilon} \supseteq EX^{\phi}$ for all $EX^{\phi} \in A$. A satisfies the theorem. □

6. Conclusion

Two new notions of error have been analyzed that make the theorems of inductive inference more easily interpreted as the ultimate constraints on machine learning of science, language, and learning scenarios in general. The first notion of error is useful for science (EX^{δ}) but is not helpful for applications to language learning, smoothing, or interpolation. The second notion of error ($((\epsilon, v)EX^{\delta}$) can be used in two distinct ways. In the first case there is a threshold on infinitely many points, and this models smoothing and interpolation with a threshold, but not language learning (i.e., learning two-valued functions). In the second case there is a threshold on only finitely many points, and this models language learning, smoothing and interpolation without a threshold. The hierarchy of learning criteria behaves similarly for the first notion and the first case of the second notion of error. This similarity breaks down for the second case of the second learning criterion, and a general theorem for this has not yet been proved. Also, it has been noted that one's view concerning the number, finite or infinite, of different possible measurements between any two measurements does not affect the results.

Acknowledgements

Thanks must be given to Professors Carl Smith and Bill Gasarch for their comments and criticisms. I also thank the referees at IJFCS for their corrections, criticisms and suggestions.

References

1. L. Blum and M. Blum, "Toward a mathematical theory of inductive inference," *Information and Control*, **28** (1975) 125–155.
2. J. Case and C. Smith, "Comparison of identification criteria for machine inductive inference," *Theoretical Computer Science*, **25(2)** (1983) 193–220.
3. M. Changizi, "Function identification from noisy data with recursive error bounds," *Erkenntnis*, **45** (1996) 91–102.
4. E. M. Gold, "Language identification in the limit," *Information and Control*, **10** (1967) 447–474.
5. W. Maass, "Efficient agnostic pac-learning with simple hypotheses," in *Proceedings of the Workship on Computational Learning Theory*, 1994.
6. M. Machtey and P. Young, An Introduction to the General Theory of Algorithms (North-Holland, 1978).
7. L. Pitt, "Probabilistic inductive inference," *Journal of the ACM*, **36(2)** (1976) 383–433.
8. H. Rogers, Jr., "Gödel numberings of partial recursive functions," *Journal of Symbolic Logic*, **23** (1958) 331–341.
9. H. Rogers, Jr., Theory of Recursive Functions and Effective Computability (McGraw Hill, New York, 1967).
10. J. S. Royer, "Inductive inference of approximations," *Information and Control*, **70(2/3)** (1986) 156–178.
11. C. Smith, A Recursive Introduction to the Theory of Computation (Springer-

Verlag, 1994).

12. C. H. Smith and M. Velauthapillai, "On the inference of approximate programs," *Theoretical Computer Science*, **77** (1990) 249–266.