Hypothesis Formation and Language Acquisition with an Infinitely-Often Correct Teacher

Sanjay Jain [*]
Dept. of Computer Science
University of Rochester
Rochester, New York 14627

Arun Sharma [†]
Dept. of Comp. and Inf. Sciences
University of Delaware
Newark, DE 19716

ABSTRACT

The presence of an "infinitely-often correct teacher" in scientific inference and language acquisition is motivated and studied. The treatment is abstract.

In the practice of science, a scientist performs experiments to gather experimental data about some phenomenon, and then tries to construct an explanation (or theory) for the phenomenon. A model for the practice of science is an inductive inference machine (a scientist) learning a program (an explanation) from the graph (set of experiments) of a recursive function (phenomenon). It is argued that this model of science is not an adequate one as scientists, in addition to performing experiments, make use of some approximate explanation (based on the "state of the art") about the phenomenon under investigation. An attempt has been made to model this approximate explanation as an additional information in the scientific process. It is shown that inference power of machines is improved in the presence of an approximate explanation. The quality of this approximate information is modeled using certain "density" notions. It is shown that additional information about a "better" quality approximate explanation enhances the inference power of learning machines as scientists more than a "not so good" approximate explanation.

[*] Supported by NSF grant CCR 832-0136.
[†] Supported by NSF grant CCR 871-3846.

Inadequacies in Gold's paradigm of language learning are investigated. It is argued that Gold's model fails to incorporate any additional information that children get from their environment. Children are sometimes told about some grammatical rule that enumerates elements of the language. These rules are some sort of additional information. Also, children are being given some information about what is not in the language. Sometimes, they are rebuked for making incorrect utterances or are told of a rule that enumerates certain non-elements of the language. An attempt has been made to extend Gold's model to incorporate both these kinds of additional information. It is shown that either type of additional information enhances the learning power of formal language learning devices.

INTRODUCTION

A model of scientific inference which involves learning of *predictive* explanations for phenomena [Gol67, BB75, CS83] may be described thus. Picture a scientist performing all the possible experiments (in any order) on a phenomenon, noting the result of each experiment while simultaneously, but algorithmically, conjecturing a succession of candidate explanations (programs) for predicting the results of all possible experiments. In this model, the set of all pairs of the form (*experiment, corresponding result*) associated with each phenomenon is taken to be coded by a function from $\mathcal{N}$ to $\mathcal{N}$, where $\mathcal{N}$ is the set of natural numbers. A criterion of success is for the scientist eventually to conjecture a program which he/she never gives up and which correctly predicts the results of all the possible experiments on the phenomenon, i.e., which correctly computes the function which codes the set of pairs associated with the phenomenon.

L. Blum and M. Blum [BB75] and Case and Smith [CS83] consider variations on the above criterion of success in which the final program is allowed to make up to a mistakes, where a is a natural number. The motivation for considering anomalies in [CS83] comes from the fact that physicists sometimes do employ explanations with anomalies.

This is a naive model of science. A scientist has more information available than just the result of experiments. C.S. Peirce [Pei58], [Rei70] argues that science is a nonterminating process of successive approximations. A scientist has some approximate explanation of the phenomenon based on the "state of the art" knowledge about that phenomenon. The model described above does not take in to account the presence of this additional information. In this paper, we make an attempt to model this additional information.

An *inductive inference machine* (IIM) is an algorithmic device which takes as its input a set of data given one element at a time, and which from time to time, as it is receiving its input, outputs programs. [KW80], [AS83], [Cas86], and [OSW86] contain surveys of work on inductive inference machines. Henceforth, we will concern ourselves with the problem in which an inductive inference machine is required to infer a program for a recursive function from its graph. This problem, as illustrated above, is analogous to the "naive" model of science. We describe below our approach to modeling additional information to a scientist.

An inductive inference machine is presented, as additional information, with a program which computes a partial function that (1) agrees infinitely often with the function being learned; and (2) does not contradict the function being learned. In other words, this additional information is an infinitely often correct teacher. However, the second restriction that this teacher not contradict the function being learned, we feel, makes our approach a simplistic one. We model the quality of this infinitely often correct teacher by using certain "density" notions from [Roy86].

A notion related to "scientific" inference of functions is the inductive inference of a type 0 grammar for a recursively enumerable language. To model language learning in children, Gold introduced the seminal notion of *identification* [Gol67]. We will use this paradigm as our model of language learning and refer to it as **TxtEx**-*identification* following [CL82]. According to this paradigm, a child (modeled as a machine) receives (in arbitrary order) all the well-defined strings of a language (a *text* for the language), and simultaneously, conjectures a succession of candidate grammars for the language being received. A criterion of success is for the child to eventually conjecture a correct grammar and to never change its conjecture thereafter. If, in this scenario, we replace the child machine by an arbitrary machine **M**, then we say that the machine **M** **TxtEx**-identifies the language. **TxtEx** is defined to be the class of sets $\mathcal{L}$ of r.e. languages such that some machine **TxtEx**-identifies each language in $\mathcal{L}$.

We study the effect of additional information in language learning. In this case, the language learning machine is provided with a grammar for a subset of the language being learned. It is also required that the "density" of difference between the two languages is no more than a certain, prespecified amount. The section on language learning contains an extensive discussion of the issues involved.

Fulk [Ful85, Ful80] and Jain and Sharma [JS89a] consider other approaches to modeling the presence of additional information in inductive learning. We now proceed formally.

NOTATIONS

$\mathcal{N}$ is the set of natural numbers. $\mathcal{I}^+$ is the set of positive integers. $*$ denotes any finite natural number. Unless otherwise specified, i, j, k, l, m, n denote integers. d, d_1, d_2 etc. denote real numbers between 0 and 1 (inclusive). a, b and c range over $(\mathcal{N} \cup \{*\})$. $\emptyset$ denotes the null set. $\text{card}(S)$ denotes the cardinality of the set S. max, min denote the maximum and minimum of a set respectively. $\subseteq$ denotes subset. $\subset$ denotes proper subset. For any two functions f_1 and f_2, $f_1 =^n f_2$ means that $\text{card}(\{x \mid f_1(x) \neq$

$f_2(x)\}) \leq n$. $f_1 =^* f_2$ means that card($\{x \mid f_1(x) \neq f_2(x)\}$) is finite. For any two sets S_1 and S_2, $S_1 =^n S_2$ means card($(S_1 - S_2) \cup (S_2 - S_1)$) $\leq n$. $S_1 =^* S_2$ means card($(S_1 - S_2) \cup (S_2 - S_1)$) is finite. δf and ρf denote the domain and range of the function f respectively.

L denotes a *recursively enumerable* subset of $\mathcal{N}$. $\mathcal{L}$ denotes a set of *recursively enumerable* (r.e.) languages. $\mathcal{E}$ denotes the class of all *recursively enumerable* languages. φ denotes a standard *acceptable* programming system [Rog58], [Rog67], [MY78]. φ_i denotes the function computed by program i in the φ-system. $W_i = \delta\varphi_i$. The set of all total recursive functions of one variable is denoted by $\mathcal{R}$. $\mathcal{S}, \mathcal{S}_1 \ldots$ denote subsets of $\mathcal{R}$. $2^{\mathcal{S}}$ denotes the power set of $\mathcal{S}$. $\langle i, j \rangle$ stands for an arbitrary computable one to one encoding of all pairs of natural numbers onto $\mathcal{N}$ [Rog67].

PRELIMINARIES

In this section, we briefly describe the fundamental paradigms that model language learning and scientific inference.

Definition 1 [Gol67] An *Inductive Inference Machine* (IIM) is an algorithmic machine which takes as its input a set of data given one element at a time, and which from time to time, as it is receiving its input, outputs programs.

IIMs have been used in the study of identification of programs for recursive functions as well as learning of grammars for languages [BB75] [CS83] [Che81] [Ful85] [Gol67] [OSW86] [Wie78]. For a survey of this work see [AS83], [OSW86], [KW80], and [Cas86].

Definition 2 If L is a language, i is a *grammar* for L iff $W_i = L$.

Definition 3 A *text* for a language L is a mapping t from $\mathcal{N}$ into $(\mathcal{N} \cup \{\#\})$ such that L is the set of natural numbers in the range of t.

Intuitively, a text for a language is an enumeration of the objects in the language with #'s representing pauses in the listing of such objects. Variables σ and τ, with or without subscripts, range over finite initial segment of texts. content(σ) $= \rho\sigma - \{\#\}$. $|\sigma|$ denotes the length of the finite initial segment σ. t, t' range over texts for languages. $\overline{t_n}$ denotes the initial segment of t with length n. $\sigma \subset t$ means σ is an initial segment of t. content(t) $= \rho t - \{\#\}$; intuitively it is a set of meaningful things presented in text t.

$\mathbf{M}(\sigma)$ is the last output of $\mathbf{M}$ after receiving input σ (note that σ can be encoded as a natural number). We will assume that $\mathbf{M}(\sigma)$ is always defined. $\mathbf{M}(t) \downarrow = i$ iff $(\overset{\infty}{\forall} n)[\mathbf{M}(\overline{t_n}) = i]$. We write $\mathbf{M}(t) \downarrow$ iff $(\exists i)[\mathbf{M}(t) \downarrow = i]$.

Definition 4 [Gol67] [CL82] $\mathbf{M}$ $\mathbf{TxtEx}^a$-*identifies* L (written: $L \in \mathbf{TxtEx}^a(\mathbf{M})$) iff for any text t for L, $\mathbf{M}(t)\downarrow$ and $W_{\mathbf{M}(t)} =^a L$.

Definition 5 $\mathbf{TxtEx}^a = \{\mathcal{L} \subseteq \mathcal{E} \mid (\exists \mathbf{M})[\mathcal{L} \subseteq \mathbf{TxtEx}^a(\mathbf{M})]\}$.

Essentially the concepts from Definitions 4 and 5 ($a = 0$ case) constitute Gold's influential language learning paradigm. The generalization of Gold's paradigm to the $a > 0$ case above was motivated by the fact that humans rarely learn a language perfectly. The $a > 0$ case in Definitions 4 and 5 is due to Case and Lynes [CL82]. Osherson and Weinstein [OW82b, OW82a] had independently introduced the $a = *$ case.

In inference of programs for recursive functions by IIMs, the input sequence $\langle 0, f(0)\rangle, \langle 1, f(1)\rangle, ...$ is presented to the IIM. For all recursive functions f, $f|^n$ denotes the finite initial segment $((\langle 0, f(0)\rangle), (\langle 1, f(1)\rangle), \ldots, (\langle n, f(n)\rangle))$.

Definition 6 [Gol67] [BB75] [CS83] $\mathbf{M}$ $\mathbf{Ex}^a$-*identifies* f (written $f \in \mathbf{Ex}^a(\mathbf{M})$) iff both $\mathbf{M}(f)\downarrow$ and $\varphi_{\mathbf{M}(f)} =^a f$.

Definition 7 $\mathbf{Ex}^a = \{\mathcal{S} \subseteq \mathcal{R} \mid (\exists \mathbf{M})[\mathcal{S} \subseteq \mathbf{Ex}^a(\mathbf{M})]\}$.

The motivation for considering anomalies in the final program in Definitions 6 and 7 comes from the fact that physicists sometimes do employ explanations with anomalies [CS83]. The $a = *$ case was introduced by L. Blum and M. Blum [BB75] and the other $a > 0$ cases were introduced by Case and Smith [CS83].

Case and Smith [CS83] introduced another infinite hierarchy of identification criterion which we describe below. "**Bc**" stands for *behaviorally correct*. Barzdin [Bar74] independently introduced a similar notion. We now define these new criteria, both in the context of scientific inference and language learning.

Definition 8 [CS83] $\mathbf{M}$ $\mathbf{Bc}^a$-*identifies* f (written: $f \in \mathbf{Bc}^a(\mathbf{M})$) iff, $\mathbf{M}$ fed f outputs over time an infinite sequence of programs $p_0, p_1, p_2, \ldots$ such that $(\overset{\infty}{\forall} n)[\varphi_{p_n} =^a f]$.

Definition 9 [CS83] $\mathbf{Bc}^a = \{\mathcal{S} \subseteq \mathcal{R} \mid (\exists \mathbf{M})[\mathcal{S} \subseteq \mathbf{Bc}^a(\mathbf{M})]\}$.

Definition 10 [CL82] $\mathbf{M}$ $\mathbf{TxtBc}^a$-*identifies* L (written: $L \in \mathbf{TxtBc}^a(\mathbf{M})$) iff, for all texts t for L, $\mathbf{M}$ outputs over time an infinite sequence of grammars $g_0, g_1, g_2, \ldots$ such that $(\overset{\infty}{\forall} n)[W_{g_n} =^a L]$.

Definition 11 [CL82] $\mathbf{TxtBc}^a = \{\mathcal{L} \subseteq \mathcal{E} \mid (\exists \mathbf{M})[\mathcal{L} \subseteq \mathbf{TxtBc}^a(\mathbf{M})]\}$.

We usually write $\mathbf{Ex}$ for $\mathbf{Ex}^0$, $\mathbf{TxtEx}$ for $\mathbf{TxtEx}^0$, $\mathbf{Bc}$ for $\mathbf{Bc}^0$, and $\mathbf{TxtBc}$ for $\mathbf{TxtBc}^0$.

Theorem 1 just below states some of the basic hierarchy results about the $\mathbf{Ex}^a$ and $\mathbf{Bc}^a$ classes.

Theorem 1 *For all $n \in \mathcal{N}$,*
(a) $\mathbf{Ex}^n \subset \mathbf{Ex}^{n+1}$.
(b) $\bigcup_{n\in\mathcal{N}} \mathbf{Ex}^n \subset \mathbf{Ex}^*$.
(c) $\mathbf{Ex}^* \subset \mathbf{Bc}$.
(d) $\mathbf{Bc}^n \subset \mathbf{Bc}^{n+1}$.
(e) $\bigcup_{n\in\mathcal{N}} \mathbf{Bc}^n \subset \mathbf{Bc}^*$.
(f) $\mathcal{R} \in \mathbf{Bc}^*$.

Parts (a), (b), (d), and (e) are due to Case and Smith [CS83]. Part (f) is due to Harrington [CS83]. Blum and Blum [BB75] first showed that $\mathbf{Ex} \subset \mathbf{Ex}^*$. Barzdin [Bar74] independently showed $\mathbf{Ex} \subset \mathbf{Bc}$.

Theorem 2 just below states some of the basic results in language learning.

Theorem 2 [CL82] *For all $i, n \in \mathcal{N}$,*
(a) $\mathbf{TxtEx}^{n+1} - \mathbf{TxtEx}^n \neq \emptyset$.
(b) $\mathbf{TxtEx}^{2n+1} - \mathbf{TxtBc}^n \neq \emptyset$.
(c) $\mathbf{TxtEx}^{2n} \subset \mathbf{TxtBc}^n$.
(f) $\bigcup_n \mathbf{TxtBc}^n \subset \mathbf{TxtBc}^*$.

ADDITIONAL INFORMATION FOR FUNCTION INFERENCE

We define the following notions of "density" from [Roy86].

Definition 12 [Roy86] *Density* of a set $A \subseteq \mathcal{N}$ in a finite and nonempty set B (denoted: $\mathbf{d}(A; B)$) is $\mathrm{card}(A \bigcap B)/\mathrm{card}(B)$.

Intuitively, $\mathbf{d}(A; B)$ can be thought of as the probability of selecting an element of A when choosing an arbitrary element from B.

Definition 13 [Roy86] *Density* of a set $A \subseteq \mathcal{N}$ (denoted: $\mathbf{d}(A)$) is $\lim_{n\to\infty} \inf\{\mathbf{d}(A; \{z \mid z \leq x\}) \mid x \geq n\}$.

Definition 14 [Roy86] The *asymptotic agreement* between two partial functions f and g (denoted: $\mathbf{aa}(f, g)$) is $\mathbf{d}(\{x \mid f(x) = g(x)\})$.

Definition 15 [Roy86] The *asymptotic disagreement* between two partial functions f and g (denoted: $\mathbf{ad}(f, g)$) is $1 - \mathbf{aa}(f, g)$.

We now describe our notion of additional information to an inductive inference machine learning a program from the graph of a recursive function. An IIM, trying to infer a program for a function f, is given as additional information, a program for a partial recursive function g which agrees with f to some extent. In Definition 16 just below, we precisely define what we mean by "a partial function g agrees with f to some extent".

Definition 16 Suppose d is a real number in the interval $[0,1]$. A partial function p is said to be *d-conforming* with a recursive function f iff, p satisfies the following two conditions:
(1) $p(x)\downarrow\Rightarrow p(x) = f(x)$, i.e., p does not contradict f.
(2) $\mathbf{d}(\{x \mid p(x) = f(x)\}) \geq d$.

Using Definition 16, we define below our new learning criterion for identification of a program from graph of a recursive function in the presence of an infinitely-often correct teacher.

Definition 17 Suppose d is a real number in the interval $[0,1]$. Suppose $a \in \mathcal{N} \cup \{*\}$. A machine $\mathbf{M}$ $\mathbf{Ap}^d\mathbf{Ex}^a$-*identifies* a function f (written: $f \in \mathbf{Ap}^d\mathbf{Ex}^a(\mathbf{M})$) iff when provided with a program for a partial function p which is d-conforming with f, $\mathbf{M}$ on f converges to a program i such that $\varphi_i =^a f$.

Definition 18 Suppose d is a real number in the interval $[0,1]$. Suppose $a \in \mathcal{N} \cup \{*\}$. $\mathbf{Ap}^d\mathbf{Ex}^a = \{\mathcal{S} \subseteq \mathcal{R} \mid (\exists \mathbf{M})[\mathcal{S} \subseteq \mathbf{Ap}^d\mathbf{Ex}^a(\mathbf{M})]\}$.

We similarly define the corresponding identification criterion for **Bc** inference.

Definition 19 Suppose d is a real number in the interval $[0,1]$. Suppose $a \in \mathcal{N} \cup \{*\}$. A machine $\mathbf{M}$ $\mathbf{Ap}^d\mathbf{Bc}^a$-*identifies* a function f (written: $f \in \mathbf{Ap}^d\mathbf{Bc}^a(\mathbf{M})$) iff when provided with a program for a partial function p which is d-conforming with f, $\mathbf{M}$ on f, outputs an infinite sequence of programs $p_1, p_2, \ldots$ such that $(\overset{\infty}{\forall} n)[\varphi_{p_n} =^a f]$.

Definition 20 Suppose d is a real number in the interval $[0,1]$. Suppose $a \in \mathcal{N} \cup \{*\}$. $\mathbf{Ap}^d\mathbf{Bc}^a = \{\mathcal{S} \subseteq \mathcal{R} \mid (\exists \mathbf{M})[\mathcal{S} \subseteq \mathbf{Ap}^d\mathbf{Bc}^a(\mathbf{M})]\}$.

In the above identification criteria, p — an approximation to f, is a good plausible additional information to a machine trying to learn a program for f from its graph. However, p may be a very bad approximator locally for large intervals which may be of importance. To overcome this situation, we use the notion of "uniform density" from [Roy86] to define a new identification criterion.

Definition 21 [Roy86] The *uniform density* of a set A in intervals of length $\geq n$ (denoted: $\mathbf{ud}_n(A)$) is $\inf(\{d(A; \{z \mid x \leq z \leq y\}) \mid x, y \in \mathcal{N}$ *and* $y - x \geq n\})$. *Uniform density* of A (denoted: $\mathbf{ud}(A)$) is $\lim_{n\to\infty} \mathbf{ud}_n(A)$.

Definition 22 [Roy86] The *asymptotic uniform agreement* between two partial functions f and g (denoted: $\mathbf{aua}(f,g)$) is $\mathbf{ud}(\{x \mid f(x)=g(x)\})$.

Definition 23 [Roy86] The *Asymptotic uniform disagreement* between two partial functions f and g (denoted: $\mathbf{aud}(f,g)$) is $1-\mathbf{aua}(f,g)$.

Using the notion of *uniform density* we define an improved learning criterion. Definition 24 just below is an analogous notion to Definition 16 for this new density notion.

Definition 24 Suppose d is a real number in the interval $[0,1]$. A partial function p is said to be *d-uniform conforming* with a recursive function f iff, p satisfies the following two conditions:

(1) $p(x)\downarrow\Rightarrow p(x)=f(x)$, i.e., p does not contradict f.

(2) $\mathbf{ud}(\{x \mid p(x)=f(x)\})\geq d$.

Definition 25 Suppose d is a real number in the interval $[0,1]$. Suppose $a\in\mathcal{N}\cup\{*\}$. A machine $\mathbf{M}$ $\mathbf{UAp}^d\mathbf{Ex}^a$-*identifies* a function f (written: $f\in\mathbf{UAp}^d\mathbf{Ex}^a(\mathbf{M})$) iff when provided with a program for a partial function p, which is d-uniform conforming with f, $\mathbf{M}$ on f converges to a program i such that $\varphi_i =^a f$.

Definition 26 $\mathbf{UAp}^d\mathbf{Ex}^a=\{\mathcal{S}\subseteq\mathcal{R}\mid(\exists\mathbf{M})[\mathcal{S}\subseteq\mathbf{UAp}^d\mathbf{Ex}^a(\mathbf{M})]\}$.

We similarly define the corresponding identification criterion for **Bc** inference.

Definition 27 Suppose d is a real number in the interval $[0,1]$. Suppose $a\in\mathcal{N}\cup\{*\}$. A machine $\mathbf{M}$ $\mathbf{UAp}^d\mathbf{Bc}^a$-*identifies* a function f (written: $f\in\mathbf{UAp}^d\mathbf{Bc}^a(\mathbf{M})$) iff when provided with a program for a partial function p, which is d-uniform conforming with f, $\mathbf{M}$ on f, outputs an infinite sequence of programs $p_1,p_2,\ldots$ such that $(\overset{\infty}{\forall} n)[\varphi_{p_n} =^a f]$.

Definition 28 $\mathbf{UAp}^d\mathbf{Bc}^a=\{\mathcal{S}\subseteq\mathcal{R}\mid(\exists\mathbf{M})[\mathcal{S}\subseteq\mathbf{UAp}^d\mathbf{Bc}^a(\mathbf{M})]\}$.

In what follows, we will refer to the two types of additional information as **Ap** and **UAp** type. Intuitively, **UAp** type additional information is a stronger type additional information, and hence we would expect the corresponding criteria of identification to be stronger. Since any $\mathbf{UAp}^d$ type additional information is also an $\mathbf{Ap}^d$ additional information we have the following two propositions.

Proposition 1 $(\forall a\in\mathcal{N}\cup\{*\})(\forall d\in[0,1])$ $[\mathbf{Ap}^d\mathbf{Ex}^a\subseteq\mathbf{UAp}^d\mathbf{Ex}^a]$.

Proposition 2 $(\forall a\in\mathcal{N}\cup\{*\})(\forall d\in[0,1])$ $[\mathbf{Ap}^d\mathbf{Bc}^a\subseteq\mathbf{UAp}^d\mathbf{Bc}^a]$.

Following theorems deal with the trade-offs between anomalies in the conjectured program, additional information, and types of identification criteria.

Theorem 3 $(\forall d \in (0,1])(\forall m \in \mathcal{N})$ $[\mathbf{UAp}^d\mathbf{Ex} - \mathbf{Ap}^1\mathbf{Bc}^m \neq \emptyset]$.

Theorem 3 says that there are classes of recursive functions that can be identified with some positive **UAp** type additional information but *cannot* be **Bc** identified with any predetermined number of anomalies allowed per program, and even the best possible **Ap** type additional information. In other words the best possible **Ap** type additional information and a more general criterion of inference cannot, in general, compensate for any **UAp** type additional information.

As a contrast, Theorem 4 below says that there are classes of recursive functions that can be **Ex**-identified with **Ap** type additional information but cannot be **Bc**-identified with any predetermined number of anomalies and **UAp** type additional information if the density associated with **Ap** type additional information is better than the one associated with **UAp** type additional information.

Theorem 4 $(\forall d_2 > d_1 \mid d_1, d_2 \in [0,1])(\forall k \in \mathcal{N})$ $[\mathbf{Ap}^{d_2}\mathbf{Ex} - \mathbf{UAp}^{d_1}\mathbf{Bc}^k \neq \emptyset]$.

Theorems 3 and 4 above together with Theorem 5 below give a complete picture about the relationship between different **Ex** and **Bc** identification criteria with **Ap** and **UAp** type additional information.

Theorem 5 $(\forall i \in \mathcal{N})$

1) $\mathbf{Ex}^{i+1} - \mathbf{UAp}^1\mathbf{Ex}^i \neq \emptyset$.
2) $\mathbf{Bc}^{i+1} - \mathbf{UAp}^1\mathbf{Bc}^i \neq \emptyset$.
3) $\mathbf{Ex}^* - \bigcup_i \mathbf{UAp}^1\mathbf{Ex}^i \neq \emptyset$.
4) $\mathbf{Bc} - \mathbf{UAp}^1\mathbf{Ex}^* \neq \emptyset$.

In summary: the results in this section give us corollaries that imply that both **Ap** and **UAp** type of additional information enhance scientific inference power of machines with respect to both **Ex** and **Bc** identification criteria. Also, in general **UAp** type of additional information results in a bigger enhancement as compared to a similar **Ap** type of additional information.

ADDITIONAL INFORMATION FOR LANGUAGE LEARNING

Formal language learning theory was originally motivated by the study of language learning in children. It relied on early claims of psycholinguists that children are rarely if ever informed of grammatical errors, instead they are only presented with strings in the language. Based on this, Gold [Gol67] developed the notion of **TxtEx**-identification. However, it turns out that the class **TxtEx**, which contains sets of *r.e.* languages that can be **TxtEx**-identified by some language learning machine, contains "small" classes of languages. For instance, none of the classes of languages in the

Chomsky hierarchy (regular, context free, context sensitive, and r.e.) are contained in **TxtEx**. This led Gold to two possible conclusions. One was that the class of natural languages is much "smaller" than previously thought, and the other was that children are being given additional information in some subtle way. Angluin [Ang80a] [Ang80b] and Wiehagen [Wie77] [KW80] address the first conclusion of Gold. We will concern ourselves, in this section, with the second conclusion of Gold.

It is not uncommon for an elder person (a parent or teacher) to tell a child some small grammatical rule that enables the child to enumerate a list of elements of the language. Basically, this additional information (the grammatical rule) enables the child to know certain elements of the language before it knows it by *text* presentation. This kind of additional information can be modeled in the Gold paradigm by requiring that in addition to a *text* for the language, the language learning device be provided with a grammar for a subset of the language. It turns out that this kind of additional information indeed increases the language learning power of learning machines. We further model the quality of this additional information by measuring the "density of agreement" of the language, whose grammar is provided as additional information, with the one being learned. Not surprisingly, a "better quality" additional information enhances the learning power more than a "not so good" additional information. We now define this "density" notion and the new language learning criteria.

Definition 29 The density of a language L_1 in an infinite language L_2 (denoted by $\mathbf{d}(L_1; L_2)$) is defined as follows: Let $x_1 < x_2 < x_3, \ldots$ be the elements of L_2. $\mathbf{d}(L_1; L_2) = \mathbf{d}(\{i \mid x_i \in L_1\})$.
Similarly, uniform density of L_1 in L_2 (denoted: $\mathbf{ud}(L_1; L_2)$) is $\mathbf{ud}(\{i \mid x_i \in L_1\})$.

Definition 30 Suppose d is a real number in the interval $[0, 1]$. A Language L' is said to be *d-language conforming* with a recursively enumerable language L iff, L' satisfies the following two conditions:
(1) $L' \subseteq L$,
(2) $\mathbf{d}(L'; L) \geq d$.

Definition 31 Suppose d is a real number in the interval $[0, 1]$. A Language L' is said to be *d-language uniform conforming* with a recursively enumerable language L iff, L' satisfies the following two conditions:
(1) $L' \subseteq L$,
(2) $\mathbf{ud}(L'; L) \geq d$.

Definition 32 Let $d \in [0, 1]$ and $a \in (\mathcal{N} \cup \{*\})$. A machine $\mathbf{M}$ $\mathbf{Ap}^d\mathbf{TxtEx}^a$-*identifies* a language L (written: $L \in \mathbf{Ap}^d\mathbf{TxtEx}^a(\mathbf{M})$) iff when provided with a grammar for a language L', which is *d-language conforming* with L, as an additional information, $\mathbf{M}$ on any text for L converges to a grammar i such that $W_i =^a L$.

Definition 33 $\mathbf{Ap}^d\mathbf{TxtEx}^a = \{\mathcal{L} \subseteq \mathcal{E} \mid (\exists \mathbf{M})[\mathcal{L} \subseteq \mathbf{Ap}^d\mathbf{TxtEx}^a(\mathbf{M})]\}$.

We can similarly define $\mathbf{UAp}^d\mathbf{TxtEx}^a$, $\mathbf{Ap}^d\mathbf{TxtBc}^a$, and $\mathbf{UAp}^d\mathbf{TxtBc}^a$ criteria of language learning. Clearly, these criteria are analogs of the similar criteria for function inference. All the theorems in function inference carry over to language learning.

Above, we were concerned with additional information that *supplements* the information a child is already receiving in the form of a *text* for the language. In other words, the additional information that we just modeled, is about what is in the language and not about what is not in the language. However, literature of speech language pathology and linguistics contains extensive refutations of the claim that children receive no negative data [BB64][Dal76]. Intuitively, it is clear that children are receiving information about the complement of the language they are trying to learn. If a child's utterances do not have the desired effect, it somehow works as a clue that the utterance is not in the language. An elder person (a parent or a teacher) either rebukes the child or tells it specifically that something is not in the language. Better still, an elder person can provide the child with a rule that enumerates a list of strings which are not members of the language. This kind of additional information can be modeled in the Gold's paradigm by requiring that the language learning device be provided with a grammar for a subset of the complement of the language being learned. It turns out that even this kind of additional information enhances the language learning power of learning devices.

Fulk [Ful85, Ful80] investigated a different approach to additional information about the complement of a language. He showed that being given *text* for a language and a grammar for the complement is equivalent to being given *text* for it and enumeration of a non-empty, finite sequence of grammars, the last of which is a grammar for the complement. However, we feel, a grammar for the complement of the language is too much additional information, and children certainly are not being given a rule that lists everything that is ungrammatical. We further employ the above density notions to differentiate a "good quality" additional information about the complement from a "not so good quality" additional information. As in the previous case, better the additional information, more is the enhancement achieved in learning power of language learning devices. We now define this notion.

Definition 34 Let $d \in [0,1]$. Let $a \in (\mathcal{N} \cup \{*\})$. A machine $\mathbf{M}$ $\mathbf{ACp}^d\mathbf{TxtEx}^a$-*identifies* a language L (written: $L \in \mathbf{ACp}^d\mathbf{TxtEx}^a(\mathbf{M})$) iff when provided with a grammar for a language L', which is *d-language conforming* with the complement of L (i.e. $\mathcal{N} - L$), as an additional information, $\mathbf{M}$ on any text for L converges to a grammar i such that $W_i =^a L$.

Definition 35 $\mathbf{ACp}^d\mathbf{TxtEx}^a = \{\mathcal{L} \subseteq \mathcal{E} \mid (\exists \mathbf{M})[\mathcal{L} \subseteq \mathbf{ACp}^d\mathbf{TxtEx}^a(\mathbf{M})]\}$.

We can similarly define $\mathbf{UACp}^d\mathbf{TxtEx}^a$, $\mathbf{ACp}^d\mathbf{TxtBc}^a$, and $\mathbf{UACp}^d\mathbf{TxtBc}^a$ criteria of language learning.

Finally, we define a language learning criteria that incorporates additional information both about elements of the language (positive information) and about elements of the complement of the language (negative information). This kind of additional information is better than just providing positive additional information or just providing negative additional information.

Definition 36 Let $d_1, d_2 \in [0,1], a \in (\mathcal{N} \cup \{*\})$. A machine $\mathbf{M}$ $\mathbf{Ap}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtEx}^a$-*identifies* a language L (written: $L \in \mathbf{Ap}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtEx}^a(\mathbf{M})$) iff when provided with grammars for languages L_1, which is d_1*-language conforming* with L, and L_2, which is d_2*-language conforming* with the complement of L (i.e. $\mathcal{N} - L$), as additional information, $\mathbf{M}$ on any text for L converges to a grammar i such that $W_i =^a L$.

Definition 37 $\mathbf{Ap}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtEx}^a = \{\mathcal{L} \subseteq \mathcal{E} \mid (\exists\mathbf{M})[\mathcal{L} \subseteq \mathbf{Ap}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtEx}^a(\mathbf{M})]\}$.

We can similarly define the following criteria of language learning.

1) $\mathbf{Ap}^{d_1}\mathbf{UACp}^{d_2}\mathbf{TxtEx}^a$,
2) $\mathbf{UAp}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtEx}^a$,
3) $\mathbf{UAp}^{d_1}\mathbf{UACp}^{d_2}\mathbf{TxtEx}^a$,
4) $\mathbf{Ap}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtBc}^a$,
5) $\mathbf{Ap}^{d_1}\mathbf{UACp}^{d_2}\mathbf{TxtBc}^a$,
6) $\mathbf{UAp}^{d_1}\mathbf{ACp}^{d_2}\mathbf{TxtBc}^a$,
7) $\mathbf{UAp}^{d_1}\mathbf{UACp}^{d_2}\mathbf{TxtBc}^a$.

All the results in function inference have a counterpart in language learning. These results along with the following theorems give us corollaries that imply that providing either positive or negative additional information enhances language acquisition power of formal devices with respect to both **TxtEx** and **TxtBc** identification criteria. Also, providing both positive and negative additional information to a language learning device is better than just providing one of them.

Theorem 6 *For all* $k \in \mathcal{N}$,

1) $\mathbf{TxtEx}^{k+1} - \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtEx}^k \neq \emptyset$,
2) $\mathbf{TxtBc}^{k+1} - \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtBc}^k \neq \emptyset$,
3) $\mathbf{TxtEx}^* - \bigcup_k \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtEx}^k \neq \emptyset$,
4) $\mathbf{TxtBc} - \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtEx}^* \neq \emptyset$,
5) $\mathbf{TxtEx}^{2k+1} - \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtBc}^k \neq \emptyset$,
6) $(\mathbf{U})\mathbf{Ap}^{d_1}(\mathbf{U})\mathbf{ACp}^{d_2}\mathbf{TxtEx}^{2k} \subseteq (\mathbf{U})\mathbf{Ap}^{d_1}(\mathbf{U})\mathbf{ACp}^{d_2}\mathbf{TxtBc}^k$,
7) $\mathcal{E} \notin \mathbf{UAp}^1\mathbf{UACp}^1\mathbf{TxtBc}^*$.

Theorem 7 $(\forall d > 0)[\mathbf{UAp}^d\mathbf{TxtEx} - \mathbf{Ap}^1\mathbf{UACp}^1\mathbf{TxtBc}^* \neq \emptyset]$.

Theorem 8 $(\forall d > 0)[\mathbf{UACp}^d\mathbf{TxtEx} - \mathbf{UAp}^1\mathbf{ACp}^1\mathbf{TxtBc}^* \neq \emptyset]$.

Theorem 9 $(\forall d_1, d_2 \mid d_2 > d_1)[\mathbf{Ap}^{d_2}\mathbf{TxtEx} - \mathbf{UAp}^{d_1}\mathbf{UACp}^1\mathbf{TxtBc}^* \neq \emptyset]$.

Theorem 10 $(\forall d_1, d_2 \mid d_2 > d_1)[\mathbf{ACp}^{d_2}\mathbf{TxtEx} - \mathbf{UAp}^1\mathbf{UACp}^{d_1}\mathbf{TxtBc}^* \neq \emptyset]$.

University of Rochester Technical Report No. 282 [JS89b] contains a detailed account of this paper.

ACKNOWLEDGEMENTS

We would like to thank John Case and Mark Fulk for advice and encouragement. Zuzana Dobes, Lata Narayanan, and Rajeev Raman provided helpful discussions. Department of CS at SUNY Buffalo and the Xerox University Grants Program to University of Rochester provided the equipment support. Sanjay Jain was supported by NSF grant CCR 832-0136 and Arun Sharma was supported by NSF grant CCR 871-3846.

References

[Ang80a] D. Angluin. Finding patterns common to a set of strings. *Journal of Computer and System Science*, 21:46–62, 1980.

[Ang80b] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, 45:117–135, 1980.

[AS83] D. Angluin and C. Smith. A survey of inductive inference: theory and methods. *Computing Surveys*, 15:237–289, 1983.

[Bar74] J. A. Barzdin. Two theorems on the limiting synthesis of functions. *Latv. Gos. Univ. Uce. Zap.*, 210:82–88, 1974.

[BB64] R. Brown and U. Bellugi. Three processes in the child's acquisition of syntax. *Harvard Educational Review*, 34:133–151, 1964.

[BB75] L. Blum and M. Blum. Toward a mathematical theory of inductive inference. *Information and Control*, 28:125–155, 1975.

[Cas86] J. Case. Learning machines. In W. Demopoulos and A. Marras, editors, *Language Learning and Concept Acquisition*, Ablex Publ. Co., 1986.

[Che81] K. Chen. *Tradeoffs in Machine Inductive Inference.* PhD thesis, SUNY/ Buffalo, 1981.

[CL82] J. Case and C. Lynes. Machine inductive inference and language identification. *Lecture Notes in Computer Science, Springer-Verlag, Berlin*, 140, 1982.

[CS83] J. Case and C. Smith. Comparision of identification criteria for machine inductive inference. *Theoretical Computer Science*, 25:193–220, 1983.

[Dal76] P. Dale. *Language Development, Structure and Function.* Holt, Reinhart, and Winston, New York, 1976.

[Ful80] M. Fulk. Inductive inference with additional information. *Journal of Computer and System Science*, to appear, 1980.

[Ful85] M. Fulk. *A Study of Inductive Inference machines.* PhD thesis, SUNY/ Buffalo, 1985.

[Gol67] E.M. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[JS89a] S. Jain and A. Sharma. *Knowledge of an Upper Bound on Grammar Size Helps Language Learning.* Technical Report 283, University of Rochester, 1989.

[JS89b] S. Jain and A. Sharma. *Learning with an Infinitely Often Correct Teacher.* Technical Report 282, University of Rochester, 1989.

[KW80] R. Klette and R. Wiehagen. Research in the theory of inductive inference by gdr mathematicians–a survey. *Information Sciences*, 22:149–169, 1980.

[MY78] M. Machtey and J. Young. *An Introduction to the General Theory of Algorithms.* North Holland, New York, 1978.

[OSW86] D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn, An Introduction to Learning Theory for Cognitive and Computer Scientists.* MIT Press, Cambridge, Mass., 1986.

[OW82a] D. Osherson and S. Weinstein. Criteria of language learning. *Information and Control*, 52:123–138, 1982.

[OW82b] D. Osherson and S. Weinstein. A note on formal learning theory. *Cognition*, 11:77–88, 1982.

[Pei58] C.S. Peirce. In A.W.Burks, editor, *Collected Papers*, Harvard University Press,Cambridge Mass., 1958.

[Rei70] F. E. Reilly. In *Charles Pierce's Theory of Scientific Method*, Fordham University Press,New York, 1970.

[Rog58] H. Rogers. Godel numberings of partial recursive functions. *Journal of Symbolic Logic*, 23:331–341, 1958.

[Rog67] H. Rogers. *Theory of Recursive Functions and Effective Computability.* McGraw Hill, New York, 1967.

[Roy86] J. Royer. Inductive inference of approximations. *Information and Control*, 70:156–178, 1986.

[Wie77] R. Wiehagen. Identification of formal languages. *Lecture Notes in Computer Science, Springer-Verlag, Berlin*, 53:571–579, 1977.

[Wie78] R. Wiehagen. Characterization problems in the theory of inductive inference. *Automata, Languages and Programming 1978, Lectures Notes in Computer Science*, 62:494–508, 1978.