Formal Theories of Language Acquisition: Practical and Theoretical Perspectives

Daniel N. Osherson
Center for Cognitive Science
M I T
Cambridge, MA 02139 U S A

Michael Stob
Department of Mathematics
Calvin College
Grand Rapids, MI 49506 U S A

Scott Weinstein
Department of Philosophy
University of Pennsylvania
Philadelphia, PA 19104 USA

ABSTRACT

Learning Theory is the study of systems that implement functions from evidential states to theories. The theoretical framework developed in the theory makes possible the comparison of classes of algorithms which embody distinct learning strategies along a variety of dimensions. Such comparisons yield valuable information to those concerned with inference problems in Cognitive Science and Artificial Intelligence. The present paper employs the framework of Learning Theory to study the design specifications of inductive systems which are of interest in the domain of language acquisition.

Section 0: Introduction

Learning Theory is tho investigation of systems that implement functions from evidential states to theories. Of central concern is the characterization of conditions under which such functions stabilize to accurate theories of a given environment. Within the theory, the informal notions of "evidence," "theory," "stabilization," "accuracy," and "environment" are replaced by precise definitions. Alternative formulations of these concepts yield alternative models within the theory. The vigorous development of Learning Theory began with a celebrated paper by Gold (1967). Angluin & Smith (1982) provide a valuable review of formal results.

Learning Theory is motivated by both scientific and technological concerns. Scientifically, the theory has proved useful in the analysis of human learning, particularly, language acquisition (see Osherson, Stob & Weinstein, forthcoming, for a review of issues). Technologically, the theory helps specify what is learnable 1n principle, and may thus guide the construction of practical systems of inductive inference.

Learning Theory yields potentially valuable insights about problems of inductive inference in the context of Cognitive Science and Artificial Intelligence. The theory provides the framework for systematic comparison of various learning algorithms. Such comparisons are particularly useful in determining the relative strength of classes of algorithms which embody distinct abstract learning strategies, in assessing their resource requirements, and in predicting their behavior in various environments. When combined with empirical studies of language acquisition, Learning Theory may provide constraints on the character of the learning strategies implemented by children, and reflect in turn on the character of the class of languages which may be acquired. Such studies in Cognitive Science may be of importance to system builders in Artificial Intelligence. They suggest that the search for ideal learning strategies' is not well motivated. Rather, by focussing on learners who embody different "styles" of learning, and by investigating their properties, the theory allows a comparison of the optimality of distinct approaches to learning along a multitude of dimensions. In addition, through the analysis of classes of algorithms that embody distinct learning strategies, this theoretical framework may provide a useful complement to studies of *ad hoc* systems built to perform inductive inference in problem domains of limited scope.

The present paper reviews some of our recent work on practical inference and relates it to problems in language acquisition. We consider design specifications for Inductive systems relevant to (a) the speed of Inference, (b) the simplicity of inferred theories, (c) the likelihood of inferential success, and (d) the resilience of such systems in environments subject to informational imperfection. Attention is restricted to learning paradigms in which only "positive information" is available about the language or data-set to be inferred; direct information about nonmembership is not offored to the learner; Angluin & Smith (1982) survey results relevant to learning paradigms in which both positive and negative information is assumed available.

Our exposition is organized as follows. The next section provides definitions and construals at the heart of contemporary learning theory. Section 2 exhibits theorems proper to the topics (a) - (d) listed above. Proofs of these theorems can be found in Osherson, Stob & Weinstein (1982, 1983d). In Section 3 we consider the relation between results reviewed here and language acquisition by children.

Section 1: The Gold modol

1.1 Sequences, languages, texts

N is the set of natural numbers. We take the notions finite sequence (in N) and infinite sequence (in N) to be basic. The set of all finite sequences is denoted: SEQ, For n £ N, and infinite sequence, t: rng(t) is the set of numbers appearing in t; t_n is the nth member of t; and t_n is the finite sequence of length n in t.

Let *P0. P1*. . . . *pi,...* be a fixed list of all partial recursive functions of one variable, and assume the 1ist to be acceptable in the sense of Rogers (1967, Ch. 2). For i £ N, let W1 - domain *pi*, the recursively enumerable subset of N with *Index* i, Languages are Identified with nonempty members of {W1 € N). The collection of all languages is denoted: . For L £ , 1 C N, 1f L = W_i then 1 1s said to be for L.

A text for L £ &6, Is any Infinite sequence such that rng(t) = L, The class of all texts for L is denoted: T_L. Given a collection, *L*, **of languages. "T, denotes U_L € L , the class of all texts for languages 1n *L*.**

1.2 Learning functions

Let G be a fixed, computable isomorphism between SEQ and N. A *learning function* is any function from N into M; such a function will be thought of as operating on members of SEQ (via G), yielding indices for recursively enumerable sets. Learning functions may be total or partial, recursive or nonrecursive. The (partial) recursive learning functions are just Φ, Φ_1, . . . , Φ The class of all learning functions is denoted: F. The class of all recursive learning functions (partial or total) is denoted: F^{rec}.

1.3 Convergence, identification

Let $f \in \mathcal{F}$, $t \in T_{\mathcal{RE}}$. $i \in N$. We say that f converges to i on t just in case (a) $f(\sigma)$ is defined for all σ in t, and (b) for all but finitely many $n \in N$, $f(\bar{t}_n) = i$. Intuitively, f converges to i on t just in case (a) f never becomes "stuck" in examining ever longer finite sequences in t, and (b) f eventually conjectures i, and never departs from it thereafter.

Let $f \in \mathcal{F}$, $t \in T_{\mathcal{RE}}$. f is said to *identify* t just in case f converges to an index, i, on t such that $W_i = rng(t)$. f *identifies* $L \in \mathcal{RE}$ just in case f identifies every $t \in T_L$. f *identifies* $\mathcal{L} \subseteq \mathcal{RE}$ just in case f identifies every $t \in T_{\mathcal{L}}$; in this case $\mathcal{L}$ is said to be *identifiable*.

Gold's Theorem (1967): Let $\mathcal{L}$ include all finite languages and any infinite language. Then $\mathcal{L}$ is not identifiable.

Section 2: Practical learning

2.1 Efficient inference

Useful learning must not take too much time. This vague admonition can be resolved into two demands: (i) the learner must not examine too many inputs before settling for good on a correct hypothesis, and (11) the learner must not spend too long examining each input. Learners satisfying (i) will be called "text-efficient;" learners satisfying

(1t) will be called "time-efficient;" learners satisfying both (i) and (ii) will be called "efficient." In this section these requirements are formulated precisely and examined for their impact on Identifiability.

2.1.1 Text-efficiency

Following Gold (1967, Section 10), we define the partial functional CONV: $\mathcal{F} \times T_{\mathcal{RE}} \to N$ as follows. For all $f \in \mathcal{F}$, $t \in T_{\mathcal{RE}}$,

$$\mathrm{CONV}(f,t) = \mu n[(\forall m > n)(f(\bar{t}_m) = f(\bar{t}_{m-1}))].$$

CONV is defined on $f \in \mathcal{F}$, $t \in T_{\mathcal{RE}}$ if f converges on t, in which case CONV(f,t) is the length of the smallest $\sigma \in$ SEQ in t such that f's last revised conjecture in t is made on σ.

Now let $F \in \mathcal{F}$, $\mathcal{L} \subseteq \mathcal{RE}$. We say that f identifies $\mathcal{L}$ *text efficiently* just in case f identifies $\mathcal{L}$, and for all $g \in \mathcal{F}$ that identify $\mathcal{L}$,

(*) if $(\exists t \in T_{\mathcal{L}})(\mathrm{CONV}(g,t) < \mathrm{CONV}(f,t))$, then $(\exists s \in T_{\mathcal{L}})(\mathrm{CONV}(f,s) < \mathrm{CONV}(g,s))$.

f is said to identify $\mathcal{L}$ *text efficiently with respect to* $\mathcal{F}^{rec}$ just in case f identifies $\mathcal{L}$, and for all $g \in \mathcal{F}^{rec}$ that identify $\mathcal{L}$, (*) holds.

Intuitively, f identifies $\mathcal{L}$ text-efficiently [with respect to $\mathcal{F}^{rec}$] just in case no other [recursive] learning function that identifies $\mathcal{L}$ is strictly faster than f in terms of convergence delay. This notion of text efficiency yields:

Proposition 1: A collection, $\mathcal{L}$, of languages is identifiable if and only if some $f \in \mathcal{F}$ identifies $\mathcal{L}$ text efficiently.

Proposition 1 shows that text efficiency is not a restrictive design feature relative to the class of all learning functions. In contrast, the next proposition shows that text efficiency is restrictive relative to the class of recursive learning functions.

Proposition 2: There is some $\mathcal{L} \subseteq \mathcal{RE}$ such that (i) some $f \in \mathcal{F}^{rec}$ identifies $\mathcal{L}$, but (ii) no $f \in \mathcal{F}^{rec}$ identifies $\mathcal{L}$ text efficiently with respect to $\mathcal{F}^{rec}$.

2.1.2 Time efficiency

A learner is time efficient if it reacts quickly to new inputs. We formalize this notion only for the case of recursive learning functions. A computational complexity measure in the sense of Blum (1967a) is imposed upon our acceptable ordering of partial recursive functions. The measure justifies reference to the number of "steps" required for φ_i to halt on j (i,j $\in$ N). For i,s,k $\in$ N, we let $\varphi_{i,s}(k)$ be the output (if any) after s steps in the computation of $\varphi_i(k)$; "$|\varphi_i(k)|$" denotes $\mu s[\varphi_{i,s}(k)$ halts].

Now let $\mathcal{L} \subseteq \mathcal{RE}$, let $i \in N$, and let $h: N \to N$ be a total recursive function. We say that φ_i identifies $\mathcal{L}$ *h-time efficiently* just in case (i) φ_i identifies $\mathcal{L}$, and (ii) for every $t \in T_{\mathcal{L}}$, there is a $k \in N$ such that for all $j > k$:

$$|\varphi_i(\bar{t}_{j+1})| \leq |\varphi_i(\bar{t}_j)| + h(t_{j+1}).$$

Intuitively, φ_i identifies $\mathcal{L}$ h-time efficiently if for every $t \in T_{\mathcal{L}}$, φ_i eventually takes no more than $h(t_{n+1})$ additional steps to respond to $\bar{t}_{n+1}$ than to respond to $\bar{t}_n$: that is, except for a constant, the growth in φ_i's response time to ever longer initial segments of t is bounded by h. h-time efficiency turns out not to restrict the classes of languages identifiable by recursive learning function. This is the burden of the next proposition.

Proposition 3: There is a recursive function, h, such that for all $\mathcal{L} \subseteq \mathcal{RE}$, some $f \in \mathcal{F}^{rec}$ identifies $\mathcal{L}$ if and only if some $\varphi_i \in \mathcal{F}^{rec}$ identifies $\mathcal{L}$ h-time efficiently.

Indeed, any h such that $h(x) > x$ almost everywhere can be chosen in Proposition 3.

2.1.3 Efficiency

Let *I* C RE>6, and let h be a total recursive function. $_{\Phi i}$ is said to identify L *h-efficientlyy* just in case (i) Φi identifies L text efficiently with respect to F^{rec} (in particular, *Φi* identifies *L),* and (ii) Φj is h-time efficient for L. h-efficiency, that is. combines the virtues of text efficiency (with respect to F^{rec}) and h-time efficiency. The next proposition shows that, for any h, h-efficiency is more restrictive than text efficiency as a design feature of recursive learning functions.

Proposition 4: For every total recursive function, h, there is a collection, L, of languages such that (i) some f € F^{rec} identifies L text efficiently, but (ii) no Φ € Frec identifies L h-efficiently.

Proposition 5: For some total recursive function, h, there is L C RE such that (i) some f € F^{rec} identifies L text efficiently, (ii) some Φi identifies *L* h-time efficiently, but (iii) no Φj identifies L h-efficiently.

2.2 Simple conjectures

To be useful, a learner should not only converge rapidly to a correct theory of its environment, it should also converge to a relatively simple theory: excessively complex theories, even if true, are of little practical use. To study the impact of such simplicity constraints on learnability, a total recursive size measure, S:N -• N, is now imposed on our acceptable ordering of partial recursive functions. Intuitively, S may be conceived as mapping indices to sizes, S(i) being the length of the program for *Φ1* corresponding to index i. The measure is governed by the following two axioms, due to Blum (1967b).

Axiom 1: For all i € N, there are only finitely many j € N such that S(j) • i.

Axiom 2: The predicate "j € S_l(1)," for i_fj € N, is decidable.

Define the function MS:°RE -> N as follows. For L € RE, RE(L) uJ[(3k)(W_k = L & S(k) = j]. Thus, MS(L) is the size of the smallest program that accepts L. Concern about simple conjectures may take the following form. Let g be a total recursive function, let f € F, and let L C RE. f is said to identify L *g-s1mply* just in case f identifies *L,* and for all t € T_L, f converges on t to an index, j, such that S(j) < g(MS(rng(t)). To exemplify, let g be \x.2x. Then f identifies *L* g-simply just in case f identifies L, and for all L € L and t € T_L, f converges on t to an index of size no greater than twice MS(L) (the size of the smallest program that accepts L * rng(t)).

Text efficiency and g-simplicity are more restrictive design features of recursive learning functions than either is alone. This is the content of the next proposition.

Proposition 6: There is *L* C RE such that (i) some f € F^{rec} identifies L text efficiently, (ii) for any total recursive function, g, such that g(x) > x for all x € N, some f € F^{rec} identifies L g-simply, but (iii) for every f € F^{rec}, and every total recursive function, h, if f identifies *L* text efficiently with respect to F^{rec}, then f does not identify L h-simply.

2.3 Learning in likely environments'[1]

In some environments each potential element of a language is associated with a fixed probability of occurrence, invariant through time. Such environments may be thought of as infinite sequences of stochastically independent events, the probability of a given element, e, appearing in the n+lst position being independent of the contents of positions 0 through n.

To study such environments, each L € RE is associated with a probability measure* m, on N such that for all x € N, x € L if and only 1f m_L({x}) > 0. (Recall that every L € RE 1s nonempty; see Section 1.1.) Next, we impose on ^{TRE} its Baire topology, RE; that 1s, for each *a* € SEQ, we take B_g • {t € ^{T}6bSI^a is in T to basic open set of F RE. For each L € R£, we define the (unique) complete probability measure, M_L, on $\mathcal{T}_{\mathfrak{L}}$ by stipulating that for all $\sigma \in SEQ$, $M_L(B_\sigma) = \Pi_{j < lh(\sigma)} m_L(\sigma_j)$. We now assume the existence of a fixed collection, $\{M_L | L \in \mathfrak{L}\}$ of measures on correponding members of $\{T_L | L \in \mathfrak{L}\}$. Intuitively, for measurable $S \subseteq T_L$, $M_L(S)$ is the probability that an arbitrarily selected text for L is drawn from S.

The following facts are easy to establish. For all $L, L' \in \mathfrak{L}$,

(I) if $L \neq L'$, then $M_L(T_{L'}) = 0$;

(II) for $f \in \mathcal{F}$, $M_L(\{t \in T_L | f$ identifies $t\})$ is defined.

In the stochastic context just discussed, the Gold definition of language identification seems needlessly restrictive. Rather than requiring identification of every text for a given language, L, it seems enough to require identification of any subset of T_L of sufficient probability. We are thus led to the following definition. Let $f \in \mathcal{F}$, $L \in \mathfrak{L}$. f is said to measure-one identify L just in case $M_L(\{t \in T_L | f$ identifies $t\}) = 1$. f measure-one identifies $\mathcal{L} \subseteq \mathfrak{L}$ just in case f measure-one identifies every $L \in \mathcal{L}$; in this case, $\mathcal{L}$ is said to be measure-one identifiable. The definition of measure-one identifiability is inspired by Waxier & Culicover (1980, Ch. 3).

Measure-one identification of a language differs from ordinary identification only by a set of measure zero. The next proposition reveals the significance of this small difference.

Proposition 7 $\mathfrak{L}$ is measure-one identifiable.

Let $\mathcal{L} = \{L_i | i \in N\}$ be an indexed collection of languages, and let $\{p_i | i \in N\}$ be the corresponding measures on them. $\mathcal{L}$ is said to be uniformly measured just in case the predicates "$x \in L_y$" and "$p_x(\{y\}) = z$" are decidable (the decidability of the latter predicate actually implies that of the former). Minor modifications in the proof of Proposition 7 yield the following.

Proposition 8: Let L be a uniformly measured collection of languages. Then, some $f \in \mathcal{F}^{rec}$ measure-one identifies *L*.

Thus, in contrast to Gold's Theorem (Section 1.3, above), any uniformly measured collection of languages consisting of all finite sets and any infinite set is measure-one identifiable.

2.4 Imperfect environments

2.4.1 Noisy texts

A noisy text for a language, L, is any text for a language of the form $L \cup D$, where D is a finite set. Thus, a noisy text for a language, L, can be pictured as a text for L into which any number of intrusions from a finite set have been inserted. Since the empty set is finite, texts for L count as noisy texts for L. We say that a learning function, f, identifies a language, L, on noisy text just in case f converges to an index for L *on* every noisy text for L. A learning function, f, identifies a collection, L, *of languages* on *noisy* text just in case f identifies every language in *L* on noisy text.

It is clear that noisy text renders impossible the identification of the collection of all finite languages. The following proposition provides a less obvious example of the disruptive effects of such environments for recursive learning functions.

Proposition 9: There is a collection, L, of languages such that (a) every language in L is infinite and disjoint from every other language in *L*, (b) some recursive learning function identifies L, and (c) no recursive learning function identifies *L* on noisy text.

2.4.2 Incomplete texts

An incomplete text for a language, L, is defined to be a text for $L - D$, where D is any finite set. An incomplete text for a language, L, can be pictured as a text for L from which all occurrences of a given finite set of sentences have been removed. Texts for L count as incomplete texts for L. We say that a learning function, f. *identifies a language,* L, *on incomplete* text just in case f converges to an index for L on every incomplete text for L. Identifiability of collections of languages on incomplete text is defined straightforwardly.

Proposition 10: There is a collection, L, of languages such that (a) every language in L is infinite and disjoint from every other language in *L,* (b) some recursive learning function identifies 1, and (c) no recursive learning function identifies *L* on incomplete text.

2.4.3 Noisy and incomplete text compared

Let $\mathcal{C}_{\mathcal{N}}$ be the family of all collections, $\mathcal{L}$, of languages such that $\mathcal{L}$ can be identified (by arbitrary learning function) on noisy text. Define $\mathcal{C}_{\mathcal{J}}$ similarly with respect to incomplete text. We have:

Proposition 11: $\mathcal{C}_{\mathcal{N}}$ is a proper subset of $\mathcal{C}_{\mathcal{J}}$.

2.4.4 Finite-difference identification on imperfect text

Let L, L' ∈ $\mathfrak{RE}$. L is said to be a *finite variant* of L' just in case (L - L') ∪ (L - L') is finite. A learning function, f, is said to *finite-difference identify a language,* L, on *noisy* text just in case for every noisy text, t, for L, f converges on t to an index for a finite-variant of L; f *finite-difference identifies* a collection, $\mathcal{L}$, *of* languages on noisy text just in case for every L∈$\mathcal{L}$, f finite-difference identifies L on noisy text. Finite-difference identification on incomplete text is defined similarly.

Given the margin of error tolerated in finite-difference identification, one might doubt that imperfection restricts this kind of learning. It is thus natural to conjecture:

L is finite-difference identifiable [by recursive learning function] if and only if 1 is finite-difference identifiable on noisy text [by recursive learning function]; and

L is finite-difference identifiable [by recursive learning function] if and only if L is finite-difference identifiable on incomplete text [by recursive learning function].

The next two propositions show that both versions of both conjectures are false.

Proposition 12: There is a collection, L, of languages such that some recursive learning function identifies L, but no learning function (recursive or not) finite-difference identifies *L* on noisy text.

Proposition 13: There 1s a collection, JL, of languages such that some recursive learning function identifies L, but no learning function (recursive or not) finite-difference identifies L on incomplete text.

Section 3: Language acquisition and formal models of Inference

The circumstances of normal language acquisition by children appear to share a fundamental feature with the inference paradigms discussed above. Infants apparently have no direct access to the nonsentences (so labeled) of the target language. This assertion rests on evidence that children are seldom corrected for ungrammatical utterances *per* se, nor do they communicate more successfully with grammatical than with ungrammatical sentences (Brown & Hanlon, 1970). In short, children learn their language on texts.

As a consequence of this shared environmental feature, the propositions adduced above are relevant to the study of language acquisition. Thus: (a) Propositions 1 - 4 reveal some of the consequences of efficient learning by children; if language acquisition proceeds efficiently in the senses earlier defined, the class of learnable languages is narrower on this account, (b) If the class of natural languages Is infinite, then a corollary of results in Section 2.2 shows that infinitely many grammars conjectured by children are wildly oversized (by any reasonable measure of size), (c) If each sentence in the target language can be associated with a lower bound on the probability of Its occurrence in the child's linguistic environment, then Proposition 7 shows that the set of languages that can be learned with certainty is very large. And (d) Propositions 9 and 10 reveal the surprising consequences for language acquisition of even mild imperfections in children's linguistic environments. These kinds of connections between Formal Learning Theory and language acquisition by children have become increasingly central to linguistic theory and developmental psycholinguistics (see Wexler & Culicover, 1980; Osherson, Stob & Weinstein 1983a).

The results reviewed in this paper represent only one of several perspectives on language acquisition offered by Formal Learning Theory. Thus, in addition to efficiency, soveral other learning strategies plausibly attributed to children have been investigated from the learning theoretic point of view (Osherson, Stob & Weinstein 1983b). These include such response tendencies as (a) restriction to hypotheses compatible with available data, (b) gradual shifts in hypotheses rather than large leaps, (c) perseveration on conjectures that predict the available linguistic data, (d) restriction to grammars with nontrivial "recursive" rules, and (e) exclusion of long-past data in hypothesis selection. Additionally, criteria of successful acquisition less stringent than identification have been formulated and studied in the context of contemporary linguistic theory and language acquisition. These issues are discussed in Osherson & Weinstein (1983c).

REFERENCES

Angluin, D. & Smith, C.H., A survey of inductive inference: Theory and methods, Yale University Technical Report 250, October 1982.

Blum, M., On the size of machines, Information & Control. 11(3). 1967a. pp. 257-285.

Blum, M., A machine-Independent theory of the complexity of the recursive functions, JACM, 14(2). 1967b, pp. 322-336.

Gold, E. M.. Language Identification In the limit, Information & Control, 10. 1967, 447-474.

Osherson, f). . Stob, M. & Weinstein. S., Environments for learning, manuscript, 1982.

Osherson, D., Stob, M. & Weinstein, S., Analysis of a language acquisition paradigm, manuscript, 1983a.

Osherson, D., Stob, M. & Weinstein, S., Learning strategies, Information A Control, 1983b.

Osherson, D. & Weinstein, S., Criteria of learning, Information & Control, 1983c.

Osherson, D. , Stob, M. & Weinstein, S., Three topics in the theory of practical learning, manuscript, 1983d.

Osherson, D. , Stob, M. & Weinstein, S., Introduction to formal learning theory, (forthcoming).

Rogers, H. , Theory of recursive functions and effective computabil1ty, McGraw-Hill. 1967.

Wexler, K. & Culicover, P., Formal principles of language acquisition, MIT Press, 1980.