Theory Discovery from Data with Mixed Quantifiers

by

Kevin T. Kelly, Clark Glymour

August 1988

Report CMU-PHIL-9

Philosophy
Methodology
Logic

Pittsburgh, Pennsylvania 15213-3890

Theory Discovery from Data with Mixed Quantifiers

Kevin T. Kelly

Carnegie Mellon University

Clark Glymour

Carnegie Mellon University

University of Pittsburgh

Table of Contents

Theory Discovery from Data with Mixed Quantifiers

Kevin T. Kelly
Carnegie Mellon University

Clark Glymour
Carnegie Mellon University
University of Pittsburgh

Abstract

Convergent realists desire scientific methods that converge reliably to informative, true theories over a wide range of theoretical possibilities. Much attention has been paid to the problem of induction from quantifier-free data. In this paper, we employ the techniques of formal learning theory and model theory to explore the reliable inference of theories from data containing alternating quantifiers. We obtain a hierarchy of inductive problems depending on the quantifier prefix complexity of the formulas that constitute the data, and we provide bounds relating the quantifier prefix complexity of the data to the quantifier prefix complexity of the theories that can be reliably inferred from such data without background knowledge. We also examine the question whether there are theories with mixed quantifiers that can be reliably inferred with closed, universal formulas in the data, but not without.

1. Introduction: Induction and Reliability

Scientific inquiry may be viewed as a process that receives increasing evidence and that periodically comes up with its current guess at a correct theory. This process is governed either by the natural physical dispositions of the theorizing agent, or by explicit methodological rules. For the purposes of this paper, we need not distinguish between the two cases.

A traditional criterion by which to judge a theorist or method of inquiry is its *reliability*. What we would really like is a method that always outputs a true, informative hypothesis on any data sampled from any theoretically possible world. Unfortunately, Hume's problem is the elementary observation that there can be no such method (assuming that the problem is properly inductive). A standard response to Hume's problem, called *convergent realism* requires only that the theorist *converge* to the truth in each theoretically possible world, no matter what order the data comes in. Converging to the truth is, of course, much easier than producing the truth all the time. From the convergent realist's point of view, an inductive problem is specified by a set of theoretically possible worlds, any one of which may be actual for all the theorist knows. He solves the problem just in case he converges to a useful, true hypothesis in each world the problem contains. The biggest inductive problem the theorist can solve (i.e. the set of all worlds in which he converges to the truth) is called his *inductive scope*. A theorist's scope is the formal correlate of its reliability. If one method's scope includes that of another, then the former is as reliable as the latter.

When we demand only convergence to the truth, Hume's problem no longer applies trivially to all inductive problems. It re-emerges for some problems (in the sense that no theorist can reliably converge to the truth in each possible world admitted by the problem), but it does not apply to others. It is a non-trivial project to map out where the dividing line falls. Moreover, we can weaken or strengthen the criteria of convergence, and draw the line between the solvable and unsolvable problems for each such definition. The result is a sort of topographical map of the intrinsic difficulties of inductive problems. We are happy when our inductive problems are easily solvable with respect to a stringent notion of convergence (i.e. one imposing strict limitations on numbers of mind-changes before convergence or imposing tight constraints on computational resource consumption in formulating hypotheses). But when difficult, pressing inductive problems arise, we still have to choose some approach to their solution (to sit on our hands and do nothing is just to choose a very bad approach to the problem). If no method can solve the problem under a strict criterion of success, it is time to ask whether the problem is solvable

under a slightly weaker notion of success. It is true that in the long run we are all dead; but for some inductive problems, all methods are hopelessly unreliable in the short run.

The reliabilist perspective on methods of inquiry is not new. Peirce, Reichenbach, Savage, and Putnam have all proposed analyses of the limiting correctness of proposed scientific methods. The statistical and Bayesian convergence theorems are also in this spirit. Under the influence of Chomsky, linguists, mathematicians and computer scientists have developed a rich formal study of the classes of grammars that can be correctly inferred in the limit. These analyses have been extended to curve-fitting [1], [2] to automatic computer programming [13], to concept learning [12], [7], [6], and to the inference of complete true theories in specified formal languages [5] [8], [9]. The general approach common to these studies is often referred to as *formal learning theory.* Formal learning theory is united by a simple, flexible formal picture of inquiry. The basic elements of this picture are as follows:

- An evidence language and an hypothesis language
- A space of theoretically possible worlds
- A formal protocol by which the theorist obtains evidence
- A theorist that produces hypotheses on the basis of the current evidence provided through the protocol.
- A criterion of hypothesis adequacy in a possible world
- A criterion of convergence of the theorist to an hypothesis on a given data presentation in a given possible world.

Any precise way of filling out these elements is called an inductive *paradigm.* The formal learning literature has examined many different paradigms, including those with stochastic theorists, stochastic worlds, noisy data, approximately true hypotheses, experimental theorists who construct questions for nature, computable theorists, quickly computable theorists, and a variety of criteria of convergence.

2. Induction from Data with Mixed Quantifiers

Despite the impressive variety of inductive paradigms examined, all previous studies in formal learning theory have assumed that the evidence formulas are quantifier-free. That is, a finite set of evidence formulas describes a finite number of relations holding over a finite subset of the universe under investigation. Background knowledge may be universal, but the data may not be.

The restriction to quantifier-free evidence sentences is undesirable in a general theory of reliable discovery methods. Many methodologists of the Nineteenth century, including William Whewell, proposed that theorists take empirical laws as their inputs and produce general theories to explain them. A reliable theorist should succeed over a broad range of possible worlds, in which different universal laws will have to be explained.

Bayesian convergence theorems do not assume that the data is quantifier-free [4]. Conditionalization is defined over all sentence types, and there is no reason why formal learning theory should not be developed at the same level of generality.

For a more homely example, a foreign language textbook typically includes particular examples of texts in the language together with general principles of grammar for the language in question. A reliable learner of foreign languages from textbooks should be able to make use of such principles when they are available.

Finally, some well known artificial intelligence learning systems tacitly depend upon universal data. This reliance on universal data is often called the "closed world hypothesis" by A.I. programmers. One version of the closed world hypothesis is this:

> If the teacher shows you an object and points out such and such parts, the object has no further parts that have not been pointed out.

For a standard artificial intelligence example, suppose that the teacher has shown you an arch, and has pointed out its parts, which are a lintel and two posts. In logical notation, he has provided the following data:

$Arch(a)$
$Part\text{–}of(a,b)$

$Part\text{–}of(a,c)$
$Part\text{–}of(a,d)$
$Lintel(d)$
$Post(b)$
$Post(c)$
$On\text{–}top(d,b)$
$On\text{–}top(d,c)$
$\neg Touches(b,c)$

$(\forall x)[(x \neq b \,\&\, x \neq c \,\&\, x \neq d) \rightarrow \neg Part\text{–}of(x,a)]$

The closed world hypothesis is an assumption about the kind of data presented to the theorist. The promise is that for each object x, a true, universal formula will eventually appear that tells you when you which objects are *not* parts of x. As we shall see, it is not accidental, but rather *necessary* that Winston's program receive these universal assurances if its performance is to be reliable.

Universal data of the form presented to Winston's program is available in many inductive applications. In the case of learning about classes of artifacts, such as carburettors and disc brakes, complete examples of such objects can be exhibited with the assurance that no parts are missing. The same is often true in the case of learning the rules of games just by watching others play. In this setting, the theorist may safely assume that nothing going on outside of the house is an essential part of a move or of a board configuration of the game being observed. In other settings, it is less obvious that the data includes universal assurances of the sort Winston's program receives. For example, a cell anatomist must always be prepared for the discovery of yet another level of structure or function that has not yet been noticed. Nature provides no assurance that every relevant structure of the cell has been observed.

As soon as we recognize that the data may contain universal quantifiers, the generalization to arbitrary combinations of quantifiers suggests itself. For example, imagine an application in which a concept must be learned over a domain of objects, some of which are artifacts and others of which are natural, like the cell. The teacher might tell the student that the object has no more parts when the object under discussion is discrete, and that each part has another part when the object can be analyzed into relevant parts to infinity. The latter sort of information involves mixed quantification (for all/there is).

Finally, induction from quantified data is interesting from a purely formal point of view. In computation theory, we know that some impossible problems are harder than others. If A can be solved given B but C

cannot be solved given B, then C is harder than A, even when both A and C are unsolvable. Similarly, if inductive problem A can be solved from data B and inductive problem C cannot be solved from data B, then C is harder than A, even if neither problem is solvable from the kind of data we actually have available. So the study of induction from quantified data may provide a formal topography of the intrinsic difficulties of unsolvable inductive problems.

In this paper we examine the following question: given that the data contains formulas up to a given quantifier complexity (i.e. number of alternations between universal and existential quantifiers) how much *more* complex (in terms of quantifiers) can a theory reliably inferred from this data be? The major result of the paper is the AE hierarchy theorem. Simply put, this theorem shows that without the help of background knowledge, a reliably inferred theory in a suitably rich language can involve one more quantifier alternation than the data, but not three. The question about two quantifier permutations remains open at every level but the first, which is settled in the negative.

Quantifiers in the data make negative results about inductive inference much harder to prove than in the case of quantifier-free data. All negative results about learning from quantifier-free data can make use of the fact that for any stage of inquiry, only finitely many relations among finitely many individuals have yet been described. The "evil demon's" task is to mislead the theorist about the nature of her world infinitely often by making the world look first one way and then another. With quantifier-free data, the demon may freely reconfigure the relations over the unseen objects in complete isolation from what has been said about those that have already been described in the data. But when quantifiers occur in the data, these data impose constraints on all individuals at once. The demon must be far more facile if he is not to trip over these constraints as he constructs data that make the world look first one way and then another. To put it another way, quantifier-free data places only *local* constraints on what the demon can do to mislead you. Quantified data, on the other hand, place *global* constraints on what the demon can do to mislead you, so the demon must be very careful.

3. Formal Preliminaries

3.1. Languages

Both the hypothesis language H and the evidence language E are assumed to be fragments of some countable first-order language L. The set of all atoms or negated atoms of L is denoted BAS. BAS is, of course, the usual evidence language assumed in formal learning theory. We now proceed to define an infinite hierarchy of stronger evidence languages.

Let s be a formula of L. Formula s is said to be Π_n if and only if s is logically equivalent to a prenex-normal formula s' of L such that the quantifier prefix of s' begins with a universal quantifier and has at most n-1 alternations between blocks of universal and existential quantifiers. Formula s is Σ_n if s is logically equivalent to a prenex normal formula s' of L such that the prefix of s' begins with an existential quantifier and has at most n-1 alternations between existential and universal quantifiers. For example, formula

$$(\forall x)(\forall y)(\exists z)(\exists w)[P(x)]$$

is a Π_2 formula. It does not matter for the purposes of this classification that the existential quantifiers are vacuous. But it is also a Π_1 formula, for it is logically equivalent to the result of eliminating its vacuous quantifiers. To say that s is Π_n is not to say that it fails to be Π_{n-1}. To establish a lower bound, one must show that a formula does *not* belong to a given complexity class.

Now, let L' be a subset of L. We say that L' is Π_n [Σ_n] if and only if each formula of L' is a Π_n [Σ_n] formula. For example, if L is a monadic predicate language with unary function symbols but no identity, then L is Π_2 and Σ_2.[1] Finally, we define $\Pi_n(L)$ [$\Sigma_n(L)$] to be the greatest Π_n [Σ_n] subset of L. So long as L has a binary, non-logical predicate or a binary function symbol and some predicate, $\Pi_n(L)$ is Π_n but neither Π_{n-1} nor Σ_n. Similarly, $\Sigma_n(L)$ is Σ_n, but neither Σ_{n-1} nor Π_n. By convention, $\Pi_0(L) = \Sigma_0(L)$ = the quantifier-free formulas of L. So the paradigms that allow only quantifier-free data are at the bottom of an infinite hierarchy of paradigms, depending on how complex the evidence is permitted to be.

Finally, it is useful to define the *closure* $\overline{\Gamma}$ of a set Γ of formulas to be the set of all closed formulas in Γ, where a closed formula is a formula with no free variables. So for example, $\overline{\Pi}_n(L)$ is the set of all closed formulas in $\Pi_n(L)$.

---

[1]After renaming bound variables to make them distinct, the quantifier prefix may be shuffled at will until it either consists of universal quantifiers followed by all existential, or all existential quantifiers followed by all universal.

3.2. Theoretically Possible Worlds

From the point of view of first-order logic, we take "theoretical possibilities" to be relational structures. In this paper we restrict our attention to countable structures. A relational structure for L makes each sentence of L either true or false. For all the theorist knows, the actual world may be any structure in which his background knowledge is true. The stronger the theorist's knowledge, the fewer possibilities he need succeed over, so the easier his inductive task. Ideally, a theorist would like to be able to arrive at the truth no matter which model of his background knowledge is the actual world. It is an interesting question whether a given background theory generates a solvable inductive problem or not.

3.3. The Data Presentation

Let $\mathfrak{R}$ be a countable relational structure for L. Let the evidence language E be some arbitrary subset of L. An assignment function for E and $\mathfrak{R}$ is a map from the variables of E to the domain of $\mathfrak{R}$. Assignment function g for E and $\mathfrak{R}$ is *complete* if and only if it is *onto* the domain of $\mathfrak{R}$. That is, a complete assignment assures that the evidence language "mentions" each domain element of $\mathfrak{R}$. The E-data of $\mathfrak{R}$ with respect to assigment g is the set of all E-formulas satisfied by g in $\mathfrak{R}$ (i.e. $\{e \in E: \mathfrak{R} \models e[g]\}$). Let t be an ω-sequence of E-formulas. Let rng(t) denote the set of all formulas occurring in t. Then define:

> Sequence t is an E-environment for $\mathfrak{R}$ $\Leftrightarrow$ there is a *complete* assignment g for $\mathfrak{R}$, E such that rng(t) = the E-data of $\mathfrak{R}$ with respect to g.

That is, an E-environment for a possible world is an enumeration of the set of all E-formulas satisfied in $\mathfrak{R}$ by an assignment that makes sure everything in the structure's domain is mentioned. We will often need to refer to the set of all formulas occurring in an environment t. This set is denoted rng(t).

If E has only finitely many distinct variables, and the domain of $\mathfrak{R}$ is infinite, $\mathfrak{R}$ has no E-environments. This situation could be rectified in various ways. Here, we consider only evidence languages with infinitely many variables.

It is often convenient to refer to the finite initial segment of an environment so far available to the theorist. If e is an environment, then we denote the initial segment of e of length n by $\overline{e}_n$.

3.4. Theorists

Our theorists will be passive investigators, rather like positional astronomers who can watch, but never intervene in, the motions of the planets. That is, our theorists output theories periodically on the basis of the initial data segment so far provided at the whim of the environment. A theorist can't output an infinite theory or an infinite set of axioms for a theory; but he can output a finite decision procedure for the axioms of an infinite theory.

Formally, let SEQ be the set of all finite sequences of formulas in E. A theorist is an arbitrary function from SEQ to decision procedures for subsets of the hypothesis language H. In what follows, it simplifies the presentation to speak as though the theorist outputs the axioms themselves, rather than their decision procedure, and nothing in the results of the paper depends on the distinction.

The passivity of the theorist may seem an extreme limitation. We might prefer, for example, a more energetic theorist who sends questions to his lab, and who then receives answers in return. But in our present setting, the reliability results for passive observers hold for experimenters and conversely. This would not be the case if we were to consider uncountable structures so that the mentioned objects were a function of the theorist's choice. Nor would the equivalence between experimentation and passive observation hold if, as is actually the case, the properties of observed objects were to depend on the theorist's experimental intervention. The invariance between experiment and observation also breaks down for questions of complexity and convergence time. Clearly, the experimentalist can obtain a crucial datum whenever he wants, whereas the passive observer may have to wait arbitrarily long before it comes in. The examination of the differences between observation and experiment is outside the scope of this paper.

3.5. Convergence

So far we have specifed what we mean by possible worlds, investigators, and data presentations. These elements comprise the *ontology* of our inductive paradigm. It remains to specify the *normative* elements of the paradigm, which define success. From the point of view of reliability, the central notion is that of a theorist's convergence to a theory. Here we consider two different notions, one properly more stringent than the other.

> Theorist θ EA-converges to theory T on environment e if and only if there is an n such that for all h, for all $m > n$, $\theta(\overline{e}_m) \models h$ if and only if $h \in T$.

Theorist θ AE-converges to theory T on environment e if and only if for all h there is an n such that for all $m > n$, $\theta(\bar{e}_m) \models h \Leftrightarrow h \in T$.

The mnemonic "EA vs. AE" reflects the fact that the second definition results from the first definition if we permute the universal quantifier on h with the existential quantifier on n. Hence it is immediate that EA-convergence implies AE-convergence. The failure of the converse implication is presented in [8]. Intuitively, EA-convergence is convergence "all at once" to a theory. AE convergence is "piece-meal" convergence to a theory, in which parts of the theory keep trickling in, but the complete theory is never conjectured as such. Indeed, a theorist can AE-converge to a theory even when each conjecture he makes is inconsistent with the theory. Nonetheless, a theorist always converges to a unique theory. Moreover, for any particular prediction the AE-convergent scientist needs to make, there is a time after which it is made correctly. AE convergence is the sort of convergence proposed by several methodologists, including C.S. Peirce and Karl Popper. EA convergence is the sort of convergence usually examined in formal learning theory.

3.6. Identification

A theorist identifies a structure just in case he converges to a "good" theory of the structure no matter what order the data arrive in. We have just seen that convergence can come in different flavors. The "goodness" of a theory in a possible world can vary as well. Recall that H is our hypothesis language, which may be any subset of first-order language L. In this paper, we are very demanding, in that we require the theorist to find the set of all H-formulas valid in his world. That is, we require that he converge to the stongest possible true theory of his world that can be expressed in H, where free variables are viewed as bound by implicit universal quantifiers. We denote the set of all H-formulas valid in $\mathfrak{R}$ by $H(\mathfrak{R})$.

More formally,

Theorist θ EA-identifies the H-theory of structure $\mathfrak{R}$ from E-data $\Leftrightarrow$ for each E-environment t for $\mathfrak{R}$, θ EA-converges to $H(\mathfrak{R})$.

Theorist θ AE-identifies the H-theory of structure $\mathfrak{R}$ from E-data $\Leftrightarrow$ for each E-environment t for $\mathfrak{R}$, θ AE-converges to $H(\mathfrak{R})$.

The requirement that the theorist converge to the complete H-theory of his world is ameliorated by the fact that H can be chosen as a fairly limited fragment of L.

3.7. Inductive Scope and Learning Problems

Recall that an inductive problem is posed by a set of theoretical possibilities, and that to solve a problem is to converge to the truth regardless of which possibility is actual. Accordingly, we define success over a set of possibilities in the following way.

> θ EA-identifies the H-theories of collection K of worlds $\Leftrightarrow$ θ EA-identifies the H-theory of each world $\mathfrak{R} \in$ K.
>
> θ AE-identifies the H-theories of collection K of worlds $\Leftrightarrow$ θ AE-identifies the H-theory of each world $\mathfrak{R} \in$ K.

Other things being equal, the larger K is, the harder the inductive problem K poses. Since we are considering only countable structures, the hardest problem that can arise is the set of all countable structures for (E $\cup$ H). We call this the *unrestricted theorizing problem for E and H.* The set of all countable relational structures for (E $\cup$ H) is denoted K(E $\cup$ H).

Unrestricted theorizing problems are the special inductive problems that arise when the theorist has no background knowledge. Therefore, unrestricted problems are of special epistemic interest. One natural epistemological project is: given E, solve for H such that the unrestricted learning problem for E and H is solvable. Another is to start with H and to solve for E so that the unrestricted learning problem for E and H is solvable. Finally, one can fix E and H both, and look for interesting problems K that are solvable with respect to E and H (there needn't exist a maximal one [10].) In this paper we pursue each of these projects.

For simplicity in stating our results, we define the notation

- AE(θ,H,E,K) $\Leftrightarrow$ θ AE-identifies the H-theory of each structure in K from E-data.
- AE(E,H,K) $\Leftrightarrow$ there is a theorist θ such that AE(θ,E,H,K).
- AE_{comp}(E,H,K) $\Leftrightarrow$ there is a *computable* theorist θ such that AE(θ,E,H,K)

- $AE(E,H) \Leftrightarrow AE(E,H,K(E \cup H))$
- $AE_{comp}(E,H) \Leftrightarrow AE_{comp}(E,H,K(E \cup H))$

The parallel definitions hold for EA identification.

4. Basic Relations

These results are all immediate consequences of the fact that we require hypotheses to be *valid* in $\mathfrak{R}$ but the evidence need only be *satisfied* in $\mathfrak{R}$ by some fixed interpretation.

L1 For all n, E, $[AE(E, \Sigma_n) \Leftrightarrow AE(E, \Pi_{n+1})]$

L2 For all n, H, $[AE(\Pi_n, H) \Leftrightarrow AE(\Sigma_{n+1}, H)]$

L3 For all n, E, $[AE(E, \Pi_n) \Leftrightarrow AE(E, \overline{\Pi}_n)]$

Proof: L1 says that since hypothesis formulas must be valid in $\mathfrak{R}$ to be correct, nothing is added or lost if we bind some of the free variables with universal quantifiers. L2 is the dual of L1. Since the data is viewed as satisfied by a fixed interpretation, the result of binding free variables with existential quantifiers adds no new information. And it does not decrease information, because Π_n is a subset of Σ_{n+1}, so we retain copies of the formulas with free variables. L3 is true for the same reason as L1. □

5. Universal data and the Osherson-Weinstein Condition

Osherson and Weinstein [11] have discovered an interesting characterization of AE identifiability when the evidence language is BAS. An examination of the proof of the theorem, however, reveals that they have actually demonstrated it for arbitrary evidence languages. All of the negative results in this paper will make use of their condition.

Let K be a collection of countable structures, let $s \in H$ be an hypothesis, and let E be our evidence

language.

K,s,E satisfy the Osherson-Weinstein condition $\Leftrightarrow$ for each $\mathfrak{R} \in K$, if $\mathfrak{R} \models s$ then there is a $\sigma \in SEQ(E)$ and a finite $g:var(\sigma) \rightarrow Dom(\mathfrak{R})$ such that for each $\mathfrak{R}'$ such that $\mathfrak{R}' \not\models s$ and for each complete assignment f such that $\mathfrak{R}' \models \sigma[f]$, there is a $\tau \in SEQ(E)$ such that $\mathfrak{R}' \models \tau[f]$ and for all assignments $g' \supseteq g$, $\mathfrak{R} \not\models \tau[g']$.

Osherson and Weinstein have shown that

Theorem OW1 (Osherson and Weinstein) [11]:

For each countable collection K of countable structures, AE(E,{s},K) $\Leftrightarrow$ E,s,K satisfy the Osherson-Weinstein condition and E, $\neg$ s, K satisfy the Osherson-Weinstein condition.

Proof: The proof of [11], proposition 31, suffices to show the theorem. □

An examination of Osherson and Weinstein's proof also reveals that the condition is necessary for *arbitrary* collections of countable structures. And since failure to AE identify the complete theory in a sub-language implies failure to identify the complete theory in a bigger language, we have as a corollary that

Theorem OW2 (Osherson and Weinstein) [11]: If AE(E,H,K) then for each $s \in H$, E,s,K satisfy the Osherson-Weinstein condition and E, $\neg$ s, K satisfy the Osherson-Weinstein condition.

Since we use the Osherson-Weinstein condition as a necessary condition in our negative proofs, and since the condition is complicated, it is useful to drive the negation through it now for future reference.

E, s, K *fail* to satisfy the Osherson-Weinstein condition $\Leftrightarrow$

there is a $\mathfrak{R} \in K$ such that

1. $\mathfrak{R} \models s$ and

2. for each $\sigma \in$ SEQ[E], for each g: freevar(σ) $\rightarrow$ Dom($\Re$), if $\Re \models \sigma[g]$ then there is an $\Re'$ $\in$ K and a surjection f: var(L) $\rightarrow$ Dom($\Re'$) such that

   a. $\Re' \mid \neq s$ and

   b. $\Re' \models \sigma[f]$ and

   c. for each $\tau \in$ SEQ[E] if $\Re' \models \tau[f]$ then there is a $g' \supseteq g$ such that $\Re \models \tau[g']$.

6. Quantified Data and no Background Knowledge: The AE Hierarchy Theorem

We now relate the quanatifier complexity of the data with the quantifier complexity of theories reliably inferred from such data without the help of background knowledge. We refer to the following two theorems collectively as the *AE hierarchy theorem*. First, we establish an easy upper bound on the data complexity required for reliable inductive inference over arbitrary countable structures.

Theorem 1: Let L be an arbitrary first-order language.

Then for all n, AE(Π_n(L), Π_{n+1}(L))

The theorist constructed in the proof of Theorem 1 employs the following technique. As the data increases, it considers ever larger fragments of its hypothesis language. At each stage, it weeds out the hypotheses refuted by the data, and then conjectures the greatest initial segment of the remainder that is consistent with the total data. The latter step ensures that the same false proposition is not conjectured infinitely often in different logical forms. The interesting part of the argument is to show that each truth is sure to be added by some time.

Proof of Theorem 1:

Choose an arbitrary first-order language L. Now suppose that σ is a sequence of L-formulas. Let lh(σ) denote the length of σ (or equivalently, the cardinality of the domain of σ, viewed as a

function). Also, let data(σ) be the result of conjoining the formulas occurring in σ and of binding all the free variables with existential quantifiers. Choose a fixed enumeration $\{s_1, s_2, \cdots, s_n, \cdots\}$ of $\Pi_{n+1}(L)$. The restriction of S to m is defined to be $\{s_i \in S: i \leq m\}$ Now we define two functions, POS and NEG from SEQ(Π_n) to finite subsets of Π_{n+1}.

- POS(σ) = $\{s_i \in \Pi_{n+1}: i \leq lh(\sigma$ & S_i is consistent with data(σ)$\}$
- NEG(σ) = $\{s_i \in \Pi_{n+1}: i \leq lh(\sigma$ & S_i is *not* consistent with data(σ)$\}$

Next, we define our theorist, θ, in terms of POS and NEG.

- θ(σ) = the restriction of POS(σ) to n, where n is least such that POS(σ) restricted to n entails no element of NEG(σ).

(Note that n always exists, since any set restricted to 0 is the empty set.) Let $s_i \in \Pi_{n+1}(L)$. Hence, there is a least j such that s_j is in prenex normal form, and $\models s_j \leftrightarrow s_i$. So s_j is of form

$(\forall x_1) \cdots (\forall x_n)[\Phi(x_1, \cdots, x_n, z_1, \cdots, z_m)]$

where only $z_1, \cdots, z_m$ occur free, and $\Phi(x_1, \cdots, x_n, z_1, \cdots, z_m) \in \Sigma_n(L)$.

Let $\mathfrak{R}$ be a relational structure for L, and let e be a Π_n environment for $\mathfrak{R}$ in virtue of complete assignment g. Now there are two cases. Either $\mathfrak{R} \models s_i$ or $\mathfrak{R} \not\models s_i$.

Case 1: $\mathfrak{R} \not\models s_i$. Then there is an assignment $h:\{x_1, \cdots, x_n, z_1, \cdots, z_m\} \rightarrow Dom(\mathfrak{R})$ such that $\mathfrak{R} \not\models \Phi(x_1, \cdots, x_n, z_1, \cdots, z_m)[h]$. Now we define a function REVERSE that negates a formula and then drives the negation in to quantifier-free formulas.

REVERSE(Φ) =

$\neg\ \Phi$ if Φ is quantifier-free

$(\exists\ x)$(REVERSE Ψ) if $\Phi = (\forall\ x)\Psi$

$(\forall\ x)$(REVERSE Ψ) if $\Phi = (\exists\ x)\Psi$

So $\mathfrak{R}$ |= REVERSE($\Phi(x_1, \cdots, x_n, z_1, \cdots, z_m)$[h]. Choose f: $\{x_1, \cdots, x_n, z_1, \cdots, z_m\} \rightarrow$ Variables(L) such that $g \circ f \supseteq h$. This is possible, since g is onto (since e is for $\mathfrak{R}$ in virtue of g). Set s = REVERSE($\Phi(f(x_1), \cdots, f(x_n), f(z_1), \cdots, f(z_m))$). So $\mathfrak{R}$ |= s[g]. Since $s \in \Pi_n(L)$, there is a j such that REVERSE($\Phi(f(x_1), \cdots, f(x_n), f(z_1), \cdots, f(z_m))$) = e_j (i.e. s is the jth item in environment e). Let j' > max(j,i). Then (*) $s_i \in NEG(\overline{e}_{j'})$. Hence, $\theta(\overline{e}_{j'}) \mid \neq s_i$, by the definition of θ.

Case 2: $\mathfrak{R} \mid = s_i$. Suppose j < i and $\mathfrak{R}$ |= s_j. Then for all $n \in \mathbf{N}$ s_j is consistent with data($\overline{e}_n$). Hence, for each n, $s_j \notin NEG(\overline{e}_n)$. Now suppose j' < i and $\mathfrak{R} \mid \neq s_{j'}$. By (*) in case 1, there is a k such that for all k' > k, $s_{j'} \in NEG(\overline{e}_{k'}) - POS(\overline{e}_{k'})$. Call the least such k for $s_{j'}$ the *modulus* of $s_{j'}$. Set m = MAX{w: $(\exists\ j' \leq i)$[w is the modulus of $s_{j'}$}. Let n > m. Then the greatest initial segment of $POS(\overline{e}_n)$ that entails no element of $NEG(\overline{e}_n)$ includes s_i, since each element of $POS(\overline{e}_n)|i$ is true and can therefore entail no falsehood. So s_i is included in $\theta(\overline{e}_n)$. Hence, for all but finitely many n, $\theta(\overline{e}_n)$ |= s_i.

□

A natural question is whether Theorem 1 can be strengthened to say anything about computable theorizing. We know that:

Proposition KG1:

If L has no function symbols, then $AE_{comp}(\Pi_0(L), \Pi_1(L))$.

Proof: Proposition 3, [8]. □

So Theorem 1 can be strengthened to AE_{comp} when n = 0. Since the theorist constructed in the proof of Theorem 2 must somehow find the greatest initial segment of non-refuted hypotheses consistent with the total data, and since this consistency test is not effective in general, we leave open the question whether Theorem 2 can be strengthened from AE to AE_{comp}.

An analysis of the computational properties of the above method would have other consequences. Osherson and Weinstein [11], Theorem 80, shows the following: Suppose we have a method ψ that converges to the correct truth value for an arbitrary input sentence $s \in H$, on the basis of the data for an arbitrarily chosen world $\Re \in K$. Then by means of essentially the same technique as that employed by our theorist, they show that there is a ϕ such that $AE(\phi,E,H,K)$. To strengthen the conclusion to $AE_{comp}(\phi,E,H,K)$ given that ψ is computable, we would again need to know whether the full, uncomputable consistency test against the data could be avoided. The question whether the uncomputable consistency test can be avoided would therefore seem to be of pivotal importance to the study of AE identification as a computational problem.

We cannot expect a lower bound that matches the upper bound of Theorem 1 for arbitrary languages, since some languages are such that for some n, $\Pi_{n+1}(L) - \Pi_n(L)$ is empty. Then clearly, Π_n data suffices to AE identify the complete, true L-theory. For example, in the special case in which n = 0, we can show the following proposition.

Proposition KG2:

Say that L is *simple* just in case

1. L has no predicates at all (including identity) or
2. L has no non-logical predicates of arity 2 or greater, no function symbols of arity 2 or greater, and no identity or
3. L has no non-logical predicates of arity 2 or greater and no function symbols (of any arity).

Then $AE(\Pi_0(L), L)$ if L is simple and

not AE(Π_0(L), Π_2(L)) if L is not simple.

Proof: The results of [8] amount to an exhaustive case argument for this proposition. □

On the other hand, we can ask whether there exist languages for which lower bounds approaching the upper bounds of Theorem 1 can be established. Call a language L *rich* just in case for each n, Π_{n-1}(L) - Π_n(L) is non-empty and Π_n(L) - Σ_n(L) is non-empty. For example, the language of number theory is rich in this sense.

Now we can establish an infinite hierarchy of lower bounds on data complexity for rich languages.

Theorem 2: Let L be rich. Then for all n ≥ 0, for all s ∈ L, if s ∉ Π_{n+2}(L), then not AE(Σ_n, {s}).

The idea of the proof is is based upon a model theoretic construction called Σ_n-chains. Let $\mathfrak{R}$, $\mathfrak{R}$' be structures for L. $\mathfrak{R}$ is said to be a Σ_n-substructure of $\mathfrak{R}$' if and only if $\mathfrak{R}$ is a substructure of $\mathfrak{R}$' and for each s ∈ Σ_n(L) and for each assignment g into the domain of $\mathfrak{R}$, if $\mathfrak{R}$ |= s[g] then $\mathfrak{R}$' |= s[g]. We also say in this case that $\mathfrak{R}$' is a Σ_n-extension of $\mathfrak{R}$. The indexed set {$\mathfrak{R}_\alpha$: α < β} is a Σ_n-chain if and only if for each γ, γ' < β, if γ < γ', then $\mathfrak{R}_\gamma$ is a Σ_n-substructure of $\mathfrak{R}_{\gamma'}$.

Keisler's n-sandwich theorem tells us that a Π_{n+1}-sentence has its truth preserved under unions of Σ_n-chains. We strengthen this to the preservation of truth under countable Σ_n-chains of countable structures. Under the supposition that L is rich, we can choose our hypothesis so that it is not Π_{n+2}. So there exists a countable Σ_{n+1}-chain such that each structure in the chain makes the hypothesis true, but the union of the structures in the chain is a countable structure in which the hypothesis is false. Since the data is Σ_n, both data formulas and their negations are Σ_{n+1}. We then use the properties of Σ_n chains to show that no learner can distinguish the structures in the chain on the basis of Σ_n data from the structure that results from the union of the chain.

Proof of Theorem 2: We require three lemmas.

Lemma 1: Let L be countable, and s ∈ L. If s is preserved under unions of countable Σ_{n-1}-chains of countable structures, then s ∈ Π_n.

Proof sketch for lemma: The proof is the result of several, straightforward applications of the generalized Lowenheim-Skolem theorem ([3], theorem 3.1.6) to the proof of Keisler's n-sandwich theorem, theorem 5.2.8 in [3]. Keisler's theorem has as a consequence that if s is preserved under unions of Σ_{n-1}-chains then $s \in \Pi_n$. To obtain the restriction to countable chains, as required in the lemma, we observe that the relevant parts of Keisler's proof require only countable chains. Several applications of the generalized Lowenheim Skolem theorem within Keisler's proof suffice to obtain the restriction to countable structures. □

Lemma 2: Let $\{\mathfrak{R}_\alpha: \alpha < \beta\}$ be a Σ_{n-1}-chain and let $\mathfrak{R} = \cup_{\alpha<\beta}\mathfrak{R}_\alpha$. Then for all $\alpha<\beta$, $\mathfrak{R}_\alpha$ is a Σ_{n-1}-substructure of $\mathfrak{R}$.

Proof of lemma: This is a substitution instance of [3], Theorem 3.1.15. □

Lemma 3: Let $\mathfrak{R}$ be a Σ_n-substructure of $\mathfrak{R}'$. Let $s \in \Sigma_{n-1}$ and g:var(s) $\rightarrow$ Dom($\mathfrak{R}$). Then

$$\mathfrak{R} \models s[g] \Leftrightarrow \mathfrak{R}' \models s[g]$$

Proof of lemma: Assume the lemma's hypothesis. Since $s \in \Sigma_{n-1}$, $s \in \Sigma_n$. Suppose $\mathfrak{R} \models s[g]$, where g:var(s) $\rightarrow$ Dom($\mathfrak{R}$). Since $\mathfrak{R}$ is a Σ_n-substructure of $\mathfrak{R}'$, $\mathfrak{R}' \models s[g]$. Now suppose $\mathfrak{R}' \models s[g]$. Since $s \in \Sigma_{n-1}$, there is a $\delta \in \Pi_{n-1}$ such that $\models \sigma \leftrightarrow \neg\, \delta$. So $\mathfrak{R}' \mid\neq \delta$. But since $\delta \in \Pi_{n-1}$, $\delta \in \Sigma_n$. So if $\mathfrak{R}' \mid\neq \delta[g]$ then $\mathfrak{R} \mid\neq \delta[g]$, by the contrapositive of the definition of Σ_n-substructure. Hence, $\mathfrak{R} \mid\neq \delta[g]$, so $\mathfrak{R} \models s[g]$. □

Proof of theorem resumed:

Let $s \notin \Sigma_{n+2}(L)$. This is possible since L is rich. Let s' = -s. So $s' \notin \Pi_{n+2}(L)$. Then by Lemma 1, there is a countable Σ_{n+1}-chain $\{\mathfrak{R}_\alpha: \alpha<\beta\}$ s.t.

For all $\alpha < \beta$, $\mathfrak{R}_\alpha$ is countable and

For all $\alpha < \beta$, $\mathfrak{R}_\alpha \models s'$ and

$\cup_{\alpha < \omega} \mathfrak{R}_\alpha \not\models s'$.

Now define

Let $\mathfrak{R} = \cup_{\alpha < \omega} \mathfrak{R}_\alpha$, so $\mathfrak{R} \not\models s$.

Let $K = \{\mathfrak{R}_\alpha : \alpha < \omega\} \cup \{\mathfrak{R}\}$.

Hence,

For all $\alpha < \omega$, $\mathfrak{R}_\alpha \not\models s$.

$\mathfrak{R} \models s$.

Each structure in K is countable
(recall $\mathfrak{R}$ is a countable union of countable structures).

Now we show that not $AE(\Sigma_n, \{s\}, K)$ by showing that the Osherson-Weinstein condition fails to hold of K, h, and $\Sigma_n(L)$, and by applying theorem OW2.

INSERT FIGURE 1 HERE

First, choose an arbitrary $\sigma \in SEQ(\Sigma_n)$, and g: freevar(σ) $\rightarrow$ Dom($\mathfrak{R}$). Recall that $\mathfrak{R} \models s$. Suppose $\mathfrak{R} \models \sigma[g]$. Now choose α so that rng(g) $\subseteq$ Dom($\mathfrak{R}_\alpha$). There is one, since rng(g) is a finite subset of $\mathfrak{R} = \cup_{\alpha < \beta} \mathfrak{R}_\alpha$. Recall that $\mathfrak{R}_\alpha \not\models s$. By Lemma 2, $\mathfrak{R}_\alpha$ is a Σ_{n+1}-substructure of $\mathfrak{R}$. By Lemma 3, $\mathfrak{R}_\alpha \models \sigma[g]$. Pick surjection f:var $\rightarrow$ Dom($\mathfrak{R}_\alpha$) so that g $\subseteq$ f. There is one since

rng(g) $\subseteq$ Dom($\mathfrak{R}_\alpha$). Hence, $\mathfrak{R}_\alpha \models \sigma[f]$. Now suppose $\tau \in SEQ(\Sigma_n)$. Suppose $\mathfrak{R}_\alpha \models \tau[f]$. Since $\mathfrak{R}_\alpha$ is a Σ_{n+1}-substructure of $\mathfrak{R}$ and $\tau \in SEQ[\Sigma_n]$, and since rng(f) $\subseteq$ Dom($\mathfrak{R}_\alpha$), we have that $\mathfrak{R} \models \tau[f]$. But since $f \supseteq g$, there is an $g' \supseteq g$ such that $\mathfrak{R} \models \tau[g']$. So by Theorem OW2, we have that not AE(Σ_n, {s}, K). □

Theorem 1, together with propositions L1-L3 yields the following system of positive results:

Positive results

Let L be any first order language. Then

- **Cor. 2.1:** For all $n \geq 0$, AE($\Sigma_n(L)$, $\Pi_n(L)$)
- **Cor. 2.2:** For all $n \geq 0$, AE($\Sigma_n(L)$, $\Sigma_{n-1}(L)$)
- **Cor. 2.3:** For all $n \geq 0$, AE($\Pi_n(L)$, $\Pi_{n+1}(L)$)
- **Cor. 2.4:** For all $n \geq 0$, AE($\Pi_n(L)$, $\Sigma_n(L)$)

These results provide upper bounds on the difficulty of AE induction from different kinds of data.

Theorem 2 gives rise to the following system of negative results, which provide lower bounds on the difficulty of AE induction for rich languages.

Negative results

Let L be rich. Then

- **Cor. 3.1:** For all $n \geq 0$, not AE($\Sigma_n(L)$, $\Pi_{n+2}(L)$)
- **Cor. 3.2:** For all $n \geq 0$, not AE($\Sigma_n(L)$, $\Sigma_{n+1}(L)$)
- **Cor. 3.3:** For all $n \geq 0$, not AE($\Pi_n(L)$, $\Pi_{n+3}(L)$)
- **Cor. 3.4:** For all $n \geq 0$, not AE($\Pi_n(L)$, $\Sigma_{n+2}(L)$)

Proof: Corollary 3.2 follows from Theorem 3 and the fact that in a rich language, some $\Pi_{n+2}(L)$ formula is not $\Sigma_{n+2}(L)$. Corollary 3.1 follows from Corollary 3.1 and fact L1. Corollary 3.3 follows from Corollary 3.1 and fact L2. Corollary 3.4 follows from Corollary 3.2 and fact L2. □

Figure 2 summarizes these corollaries.

INSERT FIGURE 2 HERE.

As is evident from the figure, Theorems 1 and 2 leave a gap of one quantifier alternation with respect to the postive results. The gap is annoying, and we expect that Theorem 2 can be improved to close the gap. For example, Proposition KG2 closes the gap when n = 0.

6.1. Discussion

The theorist constructed to prove Theorem 1 is driven by counterexamples to universal hypotheses. A Π_n formula in the data can serve as a counterexample to a Π_{n+1} hypothesis because free variables in hypotheses are viewed as universally quantified (as is standard in mathematical logic) but free variables in the data are interpreted by a fixed interpretation, and so may be viewed as existentially quantified. This means that Π_n formulas in the data can be viewed as Σ_{n+1} sentences. Clearly, each Π_{n+1} formula has a Σ_{n+1} sentences equivalent to its negation. For example, $(\forall x)(\exists y)(\forall z)(P(xyzw))$ has $(\forall y)(\exists z)(\neg Pxyzw)$ as a counterexample, where x and w are interpreted as though existentially quantified. The first formula is Π_3 while the second is Π_2.

The strategy of the theorist is to wait for counterexamples to false hypotheses and to do a little juggling to ensure that refuted hypotheses are never again entailed by later conjectures. If Theorem 2 could be strengthened to close the "gap", in analogy to the base case, then its corollaries would tell us that such counterexamples *must* be a available for reliable success (in the AE, or piece-meal convergence sense) without background knowledge.

So far, all our results concern AE or "piece-meal" convergence. At this point, one might ask whether our positive results could be strengthened to EA or "all-at-once" convergence. We refute this conjecture in the case where n=0. That is,

Prop KG 3:

Let L be non-simple. Then not EA(Π_0(L), Π_1(L)).

Proof: [8], Propositions 7,8, and 9. □

We do not yet have the proof for n > 1.

Our results shed some light on the logical importance of the "closed world hypothesis" in artificial intelligence. Winston's program is intended to infer definitions with embedded existential quantifiers. Consider the following, mistaken definition of "arch".

$(\forall x)[Arch(x) \leftrightarrow$
$(\exists y)(\exists z)(\exists w)$
$[Part\text{–}of(y,x)\ \&$
$Part\text{–}of(z,x)\ \&$
$Part\text{–}of(w,x)\ \&$
$Post(y)\ \&\ Post(z)\ \&\ Beam(w)\ \&$
$On\text{–}top(w,y)\ \&\ On\text{–}top(w,z)\ \&\ no\text{–}touch(y,z)\ \&$
$Purple(w)]$

The definition errs in requiring that the lintel of each arch be purple. To simplify our discussion, let's abbreviate the definition as

$(\forall x)[Arch(x) \leftrightarrow \Pi(x,y,z,w)]$

The prenex normal form of the definition is

$(\forall x)(\forall y1)(\forall z1)(\forall w1)(\exists y)(\exists z)(\exists w)$
$[Arch(x) \rightarrow \Pi(x,y1,z1,w1)\ \&$
$\Pi(x,y,z,w) \rightarrow Arch(x)]$

So the left-to-right side of the definition is a Π_2 sentence and the right-to-left side is a Π_1 sentence. Hence, if the right-hand-side is not *sufficient* for arch-hood, Π_0 data will suffice to reveal this fact and refute the definition. But if the definition is false because it is too strong (i.e. the right-hand-side is not a necessary condition for arch-hood) then Π_1 data is necessary to provide a counterexample But this is just the sort of data the "closed world hypothesis" guarantees the theorist. For consider the case of an arch x' with a red rather than a purple lintel. Then the theorist eventually sees the data

$Arch(x'),Part\text{–}of(y',x'),Part\text{–}of(z',x'),Part\text{–}of(w',x'),$
$Post(y'),Post(z'),Beam(w'),On\text{–}top(w',y'),On\text{–}top(w',z'),$
$No\text{–}touch(y',z'),notPurple(w'),$

$(\forall u)[(u \neq y' \,\&\, u \neq w' \,\&\, u \neq z' \rightarrow \neg Part\text{–}of(u,x)]$

Since no other objects are part of x', no other object can satisfy Π(x',y,z,w), so this data refutes the necessity of the definiens for arch-hood.

Now it might have seemed an *ad hoc* maneuver on Winston's part to add just the right universal statements to the data to help his particular learning procedure to work. But in light of Theorem 2, and Proposition KG2, we see that in rich languages, nothing less suffices *regardless of the method employed.* Hence, Winston's technique is not to be faulted for relying on such data, and is therefore more interesting than it might at first have seemed.[2] This discussion illustrates a more general point. Many artificial intelligence programmers view negative theory as a sort of "kill-joy" or pessimistic pursuit. But the fact is, negative results can greatly enhance the interest and significance of positive programming work. It is one thing to write a program that does something under certain assumptions. It quite another to know that no program could succeed under weaker assumptions. Negative results may not build a better mousetrap, but without them, it is hard to tell what to *expect* out of a mousetrap.

7. Restricted Learning Problems

The above investigation was comprehensive only for unrestricted learning problems determined by very special kinds of evidence and hypothesis languages: i.e. $\Pi_n(L)$ and $\Sigma_n(L)$, for some L. Once we admit background knowledge or restrictions on hypothesis syntax other than quantifier complexity bounds, our negative results no longer apply. Once we restrict the evidence language in a way other than by quantifier complexity, our positive results no longer apply. Hence, many interesting questions remain.

For example, what if we require that no universal law in the data has any free variables? That is, what if $E = \overline{\Pi}_0(L) \cup \Pi_0(L)$? If we could close the gap between Theorems 1 and 2, then we would know that some problems are not solvable given such data. Is there an hypothesis language (possibly restricted) and an inductive problem for this language (also possibly restricted) that is not solvable without closed

[2]We do not want to suggest that Winston's program actually identifies a correct structural description. Algorithmic shortcuts in the program can cause it to fail even with the closed world assumption, and Winston made this clear in his paper.

universal data but that is solvable with it? This question is important, because the theorist constructed to prove Theorem 1, our basic positive result, uses Π_n formulas with free variables to serve as counterexamples to Π_{n+1} formulas. This theorist is therefore a sort of generalized and polished Popperian. The desired strengthening of Theorem 2 would say that the counterexamples must be available in order for unrestricted learning problems to be solvable over the kinds of hypothesis languages considered. But if the universal laws in the data have no free variables, then they cannot serve as counterexamples to properly Π_{n+1} hypotheses. So the question arises whether universal laws in the data can be of use in reliably inferring more complex laws than themselves.

The answer to the question is a qualified "yes".

Theorem 3: There is a language L and a problem K and an hypothesis $h \in \Pi_2(L)$ such that

1. not $AE[\Pi_0(L), \{h\}, K)$ but
2. $AE[(\overline{\Pi}_1(L) \cup \Pi_0(L), \{h\}, K)$

That is, there exists a problem and a single hypothesis for which closed universal laws make a difference. The example we produce in the proof of theorem 4 is contrived, but the proof provides a useful illustration of how learning theory and model theory can be combined to address questions about restricted learning problems with closed universal data.

To prove the theorem, we need to concoct a (possibly restricted) inductive problem that is "given away" by universal data but that cannot be solved without it. What we do is to construct an infinite sequence of closed universal sentences $\{s_1, \cdots, s_n, \cdots\}$ such that each properly entails its successor. The hypothesis is the dense-order postulate s. We construct a collection of worlds such that each world either makes some sentence in the sequence of entailments true and s false, or each sentence in the sequence false,and s true. The method relying on universal data need only wait until it sees a universal sentence in the data before conjecturing s. But a method without these universal clues must try to discover from quantifier-free data whether some universal sentence in the sequence is true. And this we have already shown to be an unsolvable problem [8], Proposition 9. The new twist to the argument is to ensure that the evidence concerning hypothesis s, which is phrased in a different vocabulary, does not make the

problem any easier.

Proof: Let L be the first order language with non-logical vocabulary {P, Q}, where P and Q are both binary predicates. L has no identity predicate. First, define

$$s = (\forall x_1)(\forall x_2)(\exists y)[Q(x_1,x_2) \rightarrow Q(x_1,y) \& Q(y,x_2)]$$

Now, for each n > 0, define

$$s_n = (\forall x_1) \cdots (\forall x_n)[P(x_1,x_2) \vee \cdots \vee P(x_{n-1},x_n)]$$

$T_n = \{s_n, \neg s\}$.

K_n = the countable models of T_n

T_n is consistent because s_n shares no non-logical vocabulary with s. Define

$T_\omega = \{s\} \cup \{\neg s_n : n \in \omega\}$

K_ω = the countable models of T_ω

$K = K_\omega \cup [\bigcup_{n \in \omega} K_n]$

$E = BAS \cup \overline{\Pi}_1(L)$.

T_ω is consistent, for s shares no non-logical vocabulary with any s_n, and each s_n is true in structure $\langle N, \{\langle i, j\rangle \varepsilon N^2 : j \neq i + 1\}\rangle$.

Next, define ϕ: SEQ(E) $\rightarrow \{s, \neg s\}$ as follows:

$$\phi(\sigma) = \begin{cases} s \text{ if for each } n < \omega,\ s_n \notin \mathrm{rng}(\sigma) \\ \neg\, s \text{ otherwise} \end{cases}$$

Let $\mathfrak{R} \in K$ and let e be for $\mathfrak{R}$ in virtue of complete assignment g. Suppose $\mathfrak{R} \models s$. Then for each $n \in \omega$, $\mathfrak{R} \mid\neq s_n$. So for each i, n, $s_n \notin \mathrm{rng}(\bar{e}_i)$. So each conjecture of ϕ entails s. Now suppose $\mathfrak{R} \mid\neq s$. Then there is an n such that $\mathfrak{R} \models s_n$. So there is an i such that for all $i' > i$, $s_n \in \mathrm{rng}(\bar{e}_{i'})$. So after stage i, ϕ never makes a conjecture that entails s. Hence, $AE(\phi, \overline{\Pi}_1, \{s\}, K)$.

Now we show that not $AE(\phi, BAS, \{s\}, K)$. Define

$$\mathfrak{R} = (\omega, \mathbf{P}, \mathbf{Q})$$

$$\mathbf{P} = \omega^2 - \{(1,2), (2,3), \cdots, (n, n+1), \cdots\}$$

$$\mathbf{Q} = \{(\langle x,y\rangle, \langle z,w\rangle): x/y < z/w\}$$

where $\langle x,y\rangle$ is the code number of y under the usual recursive bijection. Let $\sigma \in SEQ(\Pi_0)$ and finite g: $\mathrm{var}(\sigma) \to \omega$ be given. such that $\mathfrak{R} \models \sigma[g]$. Now define

$$\mathbf{P}' = \omega^2 - \{(i, i+1): i, i+1 \in \mathrm{rng}(g)\}$$

$$\begin{aligned} \mathbf{Q}' = \quad & \{(i, j): (i, j) \in \mathbf{Q} \,\&\, i, j \in \mathrm{rng}(g)\} \cup \\ & \{(i, j): i < j \,\&\, i, j \notin \mathrm{rng}(g)\} \cup \\ & \{(i, j): i \in \mathrm{rng}(g) \,\&\, j \notin \mathrm{rng}(g)\} \end{aligned}$$

$$\mathfrak{R}' = (\omega, \mathbf{P}', \mathbf{Q}')$$

INSERT FIGURE 3 HERE

$\mathfrak{R}' \models T_n$, for some n, and $\mathfrak{R} \models T_\omega$. Hence, both structures are in K, as required. Observe that

- (*) $\mathfrak{R}'$ restricted to rng(g) = $\mathfrak{R}$ restricted to rng(g)
- (**) f restricted to dom(g) = g.

Now define

$$f(x_0) = \begin{cases} g(x_0), \text{ if } g(x_0)\downarrow \\ 0 \text{ otherwise} \end{cases}$$

$$f(x_{n+1}) = \begin{cases} g(x_{n+1}), \text{ if } g(x_{n+1})\downarrow \\ (\mu\, j)[(\forall\, k \leq n)(f(x_k) \neq j)] \text{ otherwise} \end{cases}$$

$\mathfrak{R}' \models \sigma[f]$ by (*) and (**). Now let $\tau \in$ SEQ(BAS) be such that $\mathfrak{R}' \models \tau[f]$. We must construct $g' \supseteq g$ such that $\mathfrak{R} \models \tau[g']$. Let R = dom(g) $\cup$ var(τ). Partition R into equivalence classes such that each variable in a given class has the same image under f. Enumerate the equivalence classes according to the order **Q**' over the range of f. Let $C_1, \cdots, C_m$ be this enumeration. We define g' to have the same value for each element of a given class, so we may as well define the value to be assigned to each class.

$g''(C_n) =$

1. g(x), where $x \in C_n \cap$ dom(g), if there is such an x.
2. some <i,j> such that i"/j" > i/j > i'/j', and <i,j> $\neq$ <i',j'>+1 and <i",j"> $\neq$ <i,j>+1; where $g''(C_{n-1})$=<i',j'> and the next $C_{n'}$ such that $C_{n'} \cap$ dom(g) $\neq \varnothing$ has an element x such that g(x) = <i",j">, if there is an n' > n such that $C_{n'} \cap$ dom(g) $\neq \varnothing$ and the previous condition is not satisfied.
3. some <i,j> such that i/j > i'/j' and <i,j> $\neq$ <i',j'>+1 where $g''(C_{n-1})$ = <i',j'> otherwise.

The value of g" does not depend on the representative of $C_n \cap$ dom(g) chosen in cases 1 and 2 because if x,y $\in C_i \cap$ dom(g), then g(x) = g(y). Otherwise, x,y would not have the same f-value since f is 1-1 everywhere except possibly over dom(g). Note also, that for any two endpoints <i",j">, <i',j'>, there is some <i,j> such that i"/j" > <i,j> > <i',j'>. Moreover, there is such an <i,j> that is not equal to <i',j'>+1. For there are at most two code numbers <i,j> such that either <i,j> = <i',j'>+1, or if <i",j"> = <i,j>+1. But if i'/j' $\neq$ i"/j" then there are infinitely many other code numbers of distinct rationals to choose.

Now define g': (dom(g) $\cup$ var(τ)) $\rightarrow \omega$ in terms of g" as follows:

g'(x) = g"([x])

where [x] is the class C_i that contains x. By clause 1 of the definition of g" we have

(***) g $\subseteq$ g'

as required. Now, literals in τ can be of two forms: $\pm$ Q(x,y), and $\pm$ P(x,y). Case 1: Suppose $\Re' \models \pm Q(x,y)$[f]. Then $\Re \models \pm Q(x,y)$[g'], because g' takes the variables in τ to code numbers of pairs in such a way that these pairs are ordered by **Q'** just the way the values of the variables according to f are ordered by **Q**. Case 2: So now consider a literal of form $\pm P(x,y)$, and suppose $\Re' \models \pm P(x,y)$. Case A: Suppose g'(y) = g'(x)+1 By the definition of g', this can only happen if x,y $\in$ dom(g), since otherwise, we make sure to define g' so that g'(y) $\neq$ g'(x)+1. Since x,y $\in$ dom(g), (*) and (**) yield $\Re \models \pm P(x,y)$[g] $\Leftrightarrow \Re' \models \pm P(x,y)$[f]. So by (***) and the assumption that $\Re' \models \pm P(x,y)$[f], we have $\Re \models \pm P(x,y)$[g']. So this leaves Case B: g'(y) $\neq$ g'(x)+1. Case α: Suppose x,y $\in$ dom(g). Then by (*), (**), and (***) we have $\Re \models \pm P(x,y)$[g']. Case β: x $\notin$ dom(g): If y $\notin$ dom(g), then we are done by Case A. If y $\in$ dom(g), then by clause 2 in the definition of g", and by the definiton of $\Re$ we see that $\Re \models P(x,y)$[g'] and $\Re \mid\neq \neg P(x,y)$[g'], for only if g'(y) = 1+g'(x) will the atom be false (by the definition of $\Re$). But notice also that no pairs are missing from **Q'** except those pairs <i,i+1> such that i,i+1 $\in$ rng(g). If x $\notin$ dom(g) then f(x) is not involved in any such missing pair, so we have that $\Re' \models P(x,y)$[g'] and $\Re' \mid\neq$

$\neg P(x,y)$[g']. Case γ: y $\notin$ dom(g): Similar to Case β. So from these cases we may conclude that $\mathfrak{N} \models \tau$[g']. By Theorem OW2, we obtain not AE(ϕ, BAS, {s}, K). □

8. Open Questions

This paper is systematic with respect to the range of questions it addresses, but it addresses only a tiny portion of the turf that should be examined. In fact, so much remains to be done that we cannot hope to provide a comprehensive list of open questions here.

The first open question is whether the AE hierarchy can be refined to close the "gap" between Theorems 1 and 2 for the case in which the hypothesis has two more quantifier alternations than appear in the data. We suspect that, in analogy to the base case, it is never possible for the Π_{n+2} theory of a structure to be reliably inferred from Π_n data over arbitrary countable structures.

We have not fully addressed the the power of computable theorists in this paper. In Proposition KG1, we showed that there is a computable theorist ϕ such that AE(ϕ, Π_0, Π_1). But this result depended on the special model theoretic properties of Π_1 hypotheses. In Theorem 1 the theorist we constructed faces a non-computable consistency test. It is an interesting question whether computable theorists can duplicate the performance of the theorist constructed in the proof of Theorem 1. And if there are no such computable theorists, we would like to know how reliable a computable theorists can be compared to our maximally reliable non-computable ones. The determination of this question would also strengthen other results, such as Osherson and Weinstein's Theorem 80 [11], and shed important general light on the structure of AE inductive inference.

We would like to generalize our current setting to include uncountable relational structures as theoretical possibilities, and we are seeking a characterization of AE identifiability over uncountable collections of structures. The standard characterization theorems in formal learning theory lean heavily on the ability of a method to enumerate possible worlds, and to this extent they do not reflect in full generality of the nature of AE inductive inference.

From well-formed questions, we move to some more speculative projects. Theorem 2, our major lower bound result, was based upon a model theoretic "preservation theorem". There are other such theorems (e.g. a theory has a negation-free axiomatization if and only if its truth is preserved under structure

homomorphisms). For each such result there is a chance that it will provide a negative result for theorizing from the preserved type of data.

It would be interesting to study the impact of universal data for theorists who perform experiments. One way to do this is to have the theorist ask questions to the actual world and to receive contingent answers. Another is to have theorists that actually *select* the actual world by building it up as a sequence of experimental outcomes. Such theorists can be viewed as discovering experimentally necessary laws (i.e. laws true in any constructible world). The powers of such theorists could then be compared systematically with those of passive observers (of the sort studied in this paper).

Osherson and Weinstein have examined how standard methodological principles interact with reliability for computable theorists. These studies could be recapitulated for universal data. It would be interesting to discover whether the conflicts between these principles with reliability are relieved or exacerbated, when the data is quantified.

Finally, to obtain a representative perspective on the power of quantified data, it would be useful to extend the present inquiry to criteria of convergence other than AE and EA convergence. Possible candidates include convergence with high probability of a stochastic theorist, δ-ε convergence to theories with respect to an error measure, and "Probably approximately correct" convergence [7].

9. Conclusion

The results of this paper provide a basic perspective on AE identification from quantified data. Theorems 1 and 2 relate the quantificational complexity of a reliably inferred theory to the quantificational complexity of open evidence when no background knowledge is available. Theorem 3 provides an example of a restricted inductive problem that can be solved with closed universal data and open literals but not without the closed universal data.

The approach taken in this study is of interest for several reasons. First, its results provide a systematic picture of how reliability interacts with the expressive power of the hypothesis lanaguage and the evidence language. We think of these results as a small but nonetheless significant step toward the mapping out of the abstract topography of the *intrinsic* difficulty of the problem of induction. Unlike much epistemological work of the past, they focus on what any *possible* method could do, rather than on what our favorite, particular examples of methods can do.

Second, the approach of this paper provides a setting in which to begin to apply the rich insights and techniques of mathematical logic to the relatively underdeveloped but nonetheless exciting study of formal methods of empirical inquiry. Deductive logicians and recursion theorists have come to take systematic categorizations of the intrinsic difficulty of computational problems for granted in their discipline. There is no reason why empirical methodology should not aspire to an equally systematic and powerful understanding of its own subject matter.

Finally, the approach taken in this paper has application to actual developments in artificial intelligence. Artificial intelligence programmers are already committed to an engineering perspective on methodology, and their programs are often eminently suited to formal analysis. As we have seen, the results in this paper motivate the standard maneuver of invoking universal data in machine learning problems that cannot be solved reliably without it.

Acknowledgements

We would like to thank Stig Andur Pedersen for useful discussions and for comments on early drafts of this paper. We would also like to thank Dan Osherson and Scott Weinstein for patiently and carefully locating and correcting errors in a previous draft.

References

[1] Lenore and Manuel Blum.
Towards a Mathematical Theory of Inductive Inference.
Information and Control 28:125-55, June, 1975.

[2] J. Case and C. Smith.
Anomaly Hierarchies of Mechanized Inductive Inference.
In *Proceedings of the Tenth ACM Symp. on Theory of Computing*, pages 314-319. 1978.

[3] C. C. Chang and H. J. Keisler.
Model Theory.
North-Holland, Amsterdam, 1973.

[4] Haim Gaifman and Marc Snir.
Probabilities over Rich Languages, Testing and Randomness.
Journal of Symbolic Logic 47:495-548, 1982.

[5] Clark Glymour.
Inductive Inference in the Limit.
Erkenntnis 21:00-00, 1984.

[6] David Haussler.
Bias, Version Spaces and Valiant's Learning Framework.
In *Proceedings of the Fourth International Workshop on Learning*, pages 324-334. Morgan Kaufmann, Los Altos, Ca., 1987.

[7] M. Kearns, M. Li, L. Pitt, and L. Valiant.
Recent Results on Boolean Concept Learning.
In *Proceedings of the Fourth International Workshop on Learning*, pages 337-352. Morgan Kaufmann, Los Altos, Ca., 1987.

[8] Kevin T. Kelly and Clark Glymour.
Convergence to the Truth and Nothing But the Truth.
Philosophy of Science 56:, 1989.

[9] Osherson, D. and Weinstein, S.
Identification in the Limit of First Order Structures.
Journal of Philosophical Logic 15:55-81, 1986.

[10] Osherson, D., Stob, M., and Weinstein, S.
Systems that Learn.
M.I.T. Press, Cambridge, Mass., 1986.

[11] Osherson, D. and Weinstein, S.
Paradigms of Truth Detection.
1988.

[12] L. Pitt and L.G. Valiant.
Computational Limitations on Learning from Examples.
Technical Report TR-05-86, Center for Research in Computing Technology, 1986.

[13] Ehud Y. Shapiro.
Inductive Inference of Theories from Facts.
Research Report 192, Yale University: Department of Computer Science, February, 1981.

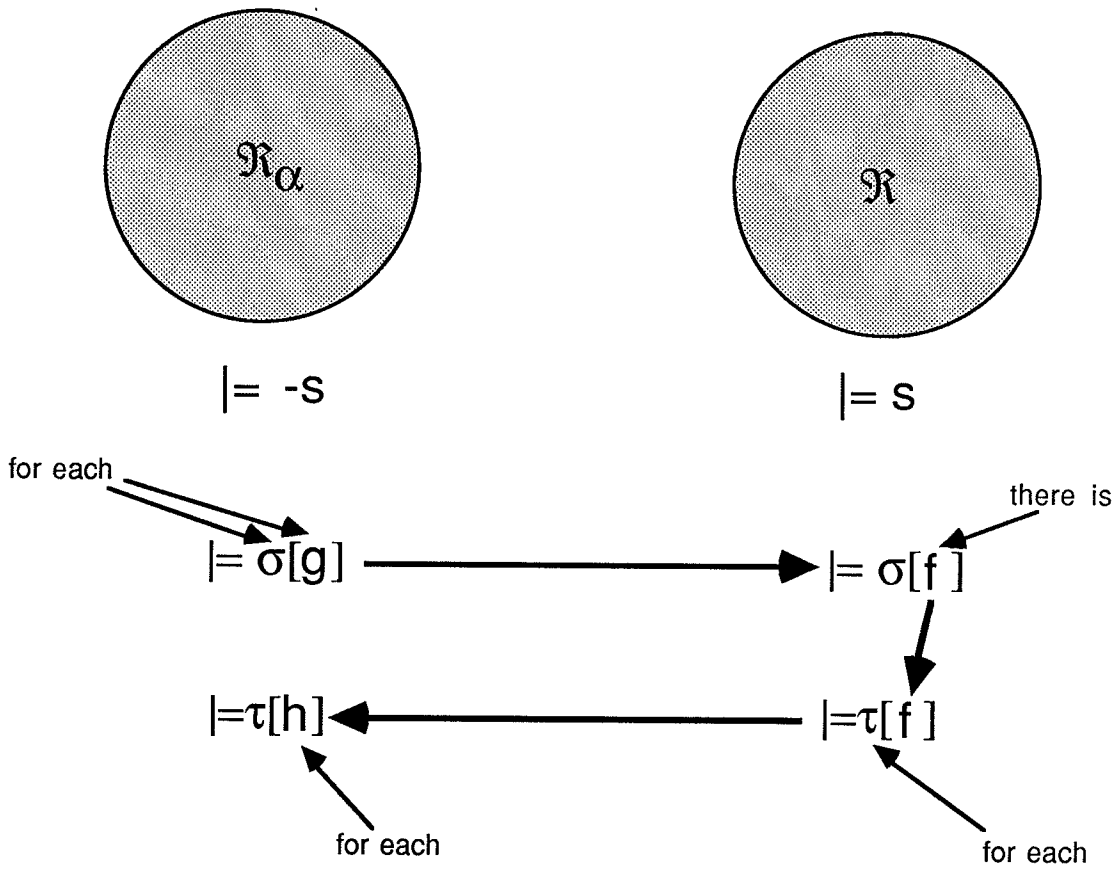
$\mathfrak{K}_\alpha$
$\mathfrak{K}$
|= -s
|= s
for each
there is
|= σ[g]
|= σ[f]
|=τ[h]
|=τ[f]
for each
for each

Figure 1

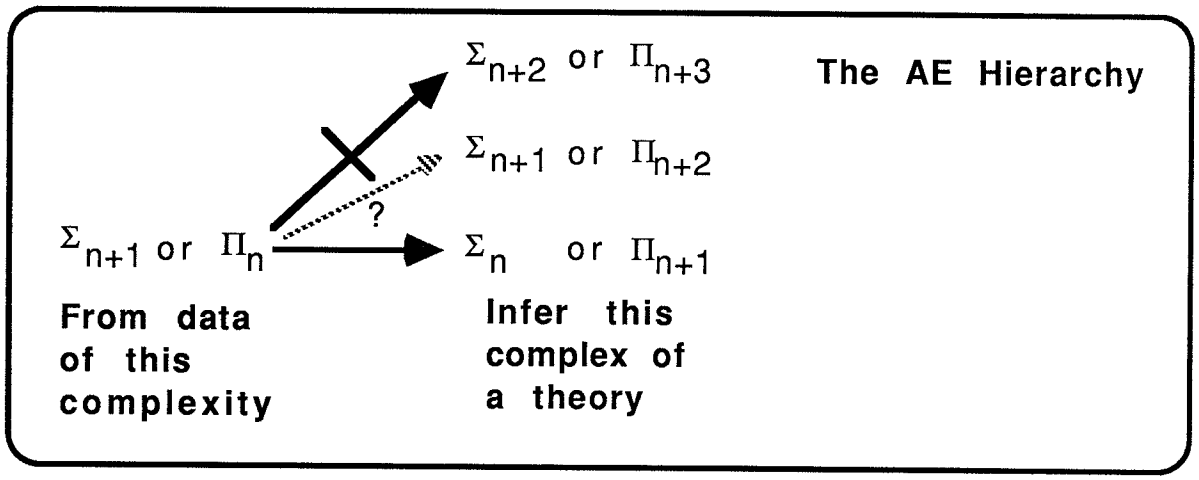
Σ_{n+2} or Π_{n+3}
The AE Hierarchy
Σ_{n+1} or Π_{n+2}
?
Σ_{n+1} or Π_n
Σ_n or Π_{n+1}
From data of this complexity
Infer this complex of a theory

Figure 2

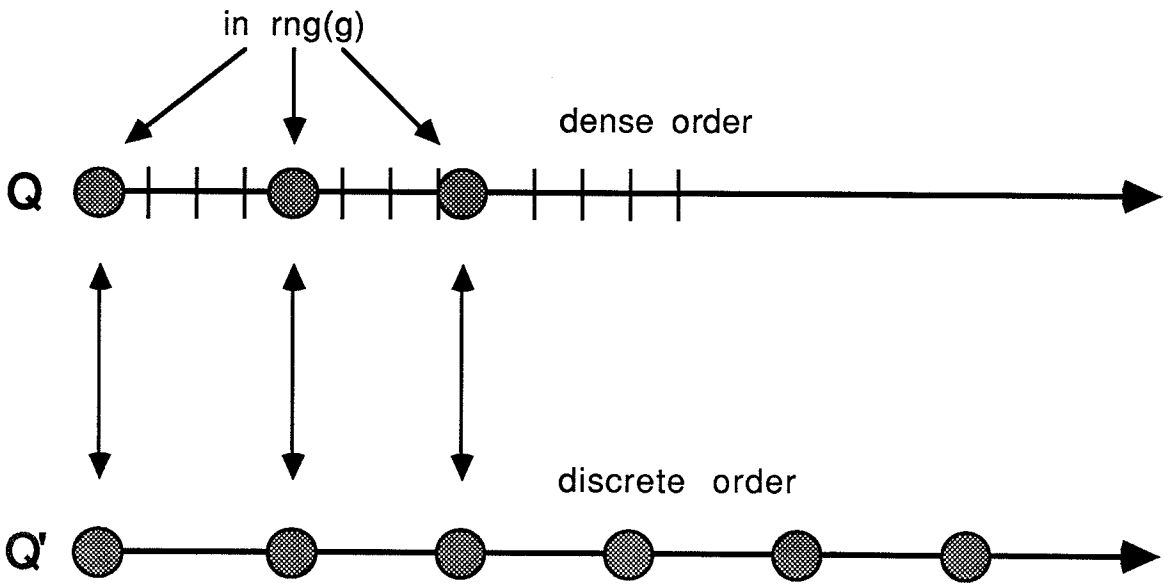

Figure 3