General Characterizations of inductive Inference over Arbitrary Sets of Data Presentations

Kevin T. Kelly
Department of Philosophy
Carnegie Mellon University

1. Introduction

Formal learning theory is the abstract study of conditions under which a system, computable or otherwise, can converge reliably to a true or empirically correct conclusion about its world on the basis of an ever increasing stream of evidence about that world.[1] Reliability amounts to the demand that the scientist converge to a correct conclusion over each in a given class of possible evidence presentations describing a variety of possible worlds of study.

Among the more ambitious objectives of formal learning theory has been to characterize the collections of data presentations over which a scientist can reliably converge to correct views about a given subject matter. Angluin has isolated necessary and sufficient conditions for identifying formal languages over all complete, positive presentations of their elements [1]. Osherson and Weinstein have arrived at characterizations of reliable hypothesis investigation in model theoretic languages [17,18, 19]. Kelly and Glymour have isolated a set of characterizations concerning the investigation of hypotheses whose truth values change as a function of the state of the scientist [14].

Unfortunately, each of these characterizations is unbalanced in a way that makes it almost impossible to apply in practice. The problem is that these results apply only if we require the scientist to succeed over *each possible ordering* of the data true of a given world. They fail to apply if the inductive problem involves background knowledge that places any constraint whatsoever on evidence ordering.

[1] For a broad introduction to the field and for a good introductory bibliography, see [20].

A moment's reflection reveals that almost every real inductive inference problem involves background knowledge that restricts evidence ordering. The most obvious application is in physical science, where the background knowledge is a dynamical theory that does nothing but impose constraints on the order in which events may be observed. Even in the intended application of natural language learnability, the arbitrary data ordering requirement is inapplicable. The pragmatics of discourse impose non-trivial restrictions on the stream of sentences uttered in the vicinity of the infant.

The severity of requiring success over all possible presentations of a set is highlighted by the fact that convergence in the limit is a very weak notion of convergence. Negative results about a uniformly weak standard of success would be interesting. Positive results about a strong sense of success would be interesting. But the characterization theorems of learning theory are about an inductive paradigm that is both too strong and too weak, and therein lies the problem with its applicability. The reason why current characterizations require the scientist to succeed over all possible presentations of a given data set is that the proofs of these theorems invoke what Osherson, Stob and Weinstein have called the *locking sequence lemma* [1, 3, 17, 18, 19, 20], which applies only when this requirement is assumed.

The technique of this paper, following Gold [6] and Putnam [22], is to by-pass the locking sequence lemma by using the very definition of learning to obtain upper complexity bounds on what can be learned. By relativizing the notion of complexity to background knowledge, the technique yields characterizations valid over arbitrary sets of data presentations.

Another limitation of standard formal learning theory is that the results are paradigm specific. There are results about ascertaining the truth values of logical hypotheses over first-order structures, about identifying recursive functions, and about learning grammars for formal languages. It would be nice to have a more paradigm-independent approach behind our fundamental understanding of what is leamable and why. The approach taken here is to move the framework of learning theory to the more general setting of arithmetic and Borel classes of data presentations. This abstract setting allows us to separate the essence of learning theory from the accidents of particular paradigms. The general characterizations proved in this setting imply special case characterization theorems for language learnability, function identification, and logical

truth detection. They also imply characterizations of paradigms not yet considered, such as detecting function properties over uncountable sets of functions and detecting language properties over uncountable collections of languages.

The approach taken in this paper generalizes other dimensions of learning theory as well. For example, we eliminate from Angluin's characterization [1] of effective language learnability the restrictive assumption that the collection of languages to be learned be RE indexable. The amount of computable control one needs over hypotheses is captured exactly by the characterization theorem, and does not have to be stipulated as a hypothesis.

2. Borel Characterizations of Ineffective Inductive Inference

Let co® be the set of all infinite sequences of natural numbers. Let t *e* co®. Then define t_n = the item occurring at position n in t, and let t[n] denote the initial segment of t of length n.

Let co* be the set of all finite sequences of natural numbers. Let x *e* co*. B_T = {t 6 co®: x c. t}. Let fi = the set of all unions of sets B_T. Then JB = <co^{∞}, fi> is the *Baire topology* on co^{03}, and {B_T: x *e* co*} is a countable basis for JB. Now we may close fi under the operations of complementation and countable union to arrive at the least a-field 9_a generated by JB.

Baire topology is a suggestive setting for the study of inductive inference. Intuitively, an infinite sequence t may be thought of as an infinite evidence presentation, where t_n may be thought of as a code number for an observation made on this presentation at time n. t[n] may be thought of as the code numbers of observations made by time n. A hypothesis is said to be *empirically adequate* just in case it is consistent with all the data one will ever see in the future. Ideally, a hypothesis is empirically adequate just in case it is consistent with the actual, infinite data presentation. We may therefore think of a hypothesis as determining the set of all infinite data presentations for which it is empirically adequate. Indeed, for the purposes of our general setting for inductive inference, we may simply identify a hypothesis with the set of all data presentations for which it is empirically adequate.

Think of a scientist living on an infinite data presentation t *e* ω investigating hypothesis P ⊆ ω*[0]. This scientist examines the evidence t[n] seen at stage n of inquiry, and makes his best guess as to whether P is empirically adequate for the evidence that will ever occur in t. We can think of such a scientist as a function φ: ©* -> {1, 0}, where x e ω* is some initial segment t[n] of some data presentation t e ω®, and 1 is the guess that t e P, and 0 is the guess that t e P. We say that

scientist φ *semi-detects P* in the limit on t ⇔
(3n Vm>n φ(t[m]) = 1) ⇔ te P.

scientist $ *detects* P in the limit on t ⇔
(3n Vm>n [φ(t[m]) = 1 if t e P and φ(t[m]) = 0 otherwise].

scientist φ [semi-] detects P in the limit ⇔
Vt *e* ω[0]*, φ [semi-] detects P on t.

P is [semi-] detectable ⇔
3 scientist $ s.t. $ [semi-] detects P in the limit.

A semi-detectable set P is analogous to an RE set. If t G P then semi-detector $ eventually converges to 1 on t, but if t e P, then $ may fail to converge to any answer. A detectable set is analogous to a recursive set, in that a detector <$> converges to 1 if t e P and φ converges to 0 otherwise. A set is recursive if and only if it is both RE and co-RE. By analogy, it is easy to see that

Fact 2.1 (Osherson and Weinstein [18]):

P is detectable ⇔ P, $\overline{P}$ are both semi-detectable.

Proof: Let φ[+] semi-detect P and Let φ" semi-detect $\overline{P}$. Define

$$\phi(\sigma) = \begin{cases} 1 & \text{if \$" said 0 more recently than } \phi^+ \text{ on a} \\ 0 & \text{otherwise} \end{cases}$$

If t e P then φ[+] converges to 1 and φ either converges to 0 or says 0 infinitely often. So there is a time after which φ[+] always says 1 and at some time after this, φ" says zero. At this point $ converges to 1. The other case is similar. H

Hypotheses (viewed as sets of infinite data presentations) may have greater or lesser *topological* complexity. We measure the topological complexity of sets in JB_G by noting

when they are added in the construction of the Borel hierarchy. Using a familiar notation, we let $B = \{\Sigma^0_1, \Pi^0_1, \Sigma^0_2, \Pi^0_2, \ldots\}$, where the B indicates that we are talking about the Borel hierarchy over Baire space. Π^0_1 denotes the set of all complements of Σ^0_1 sets over ω^ω. Now, for each ordinal $\gamma > 1$, define

Σ^0_γ = the set of all countable unions of elements of $\bigcup_{\xi<\gamma} \Pi^0_\xi$ for $\xi < \gamma$.

Π^0_γ = the set of all countable intersections of elements of $\bigcup_{\xi<\gamma} \Sigma^0_\xi$ for $\xi < \gamma$.

And for each $\gamma \geq 1$, define

$$\Delta^0_\gamma = \Sigma^0_\gamma \cap \Pi^0_\gamma.$$

Inductive inference usually occurs with respect to some prior background knowledge T. The effect of T is to exclude from possibility some set of infinite evidence presentations. To put it another way, T picks out for consideration a special subset K of data presentations, namely, those with respect to which T is empirically adequate. So given knowledge K, the scientist has an edge in trying to detect P.

scientist $ [semi-] detects P over K <=>
    Vt e K, $ [semi-] detects P on t.

P is [semi-] detectable over K <=>
    3 scientist $ s.t. $ [semi-] detects P over K.

When the scientist has background knowledge K, we are interested not in the intrinsic Borel complexity of a given hypothesis, but rather, its *conditional* Borel complexity *relative to* K. To investigate the complexity of inductive inference with respect to background knowledge, we must relativize the Borel hierarchy to K.

Let $K \subseteq \omega^\omega$. Define $B? = \mathcal{B}? \cap K$, and let τ^K = the set of all unions of sets B?. Then $\mathcal{B}^*$. $\langle K, \tau^K \rangle$ is the *induced Baire topology* on K, with countable basis $\{B?: \sigma \in \omega^{<\omega}\}$. As before, we may close τ^K under complementation and union, to form #5, the least sigma field containing τ^K. Finally, we define the K-relativized version of the Borel hierarchy.

$$\Sigma^{0,K}_1 = \mathcal{B}^K.$$

$\Pi_1^{B,K}$ = the set of all complements of elements of $_T\Delta_1^{B,K}$.

Now, for each ordinal $\gamma > 1$, define

$\Sigma_\gamma^{B,K}$ = the set of all countable unions of elements of $\bigcup_{\xi<\gamma}\Pi_\xi^{B,K}$.

$\Pi_\gamma^{B,K}$ = the set of all countable intersections of elements of $\bigcup_{\xi<\gamma}\Sigma_\xi^{B,K}$ for $\xi < \gamma$.

And for each $\gamma \geq 1$, define

$$\Delta_\gamma^{B,K} = \Sigma_\gamma^{B,K} \cap \Pi_\gamma^{B,K}.$$

Now a natural question arises. Can we characterize the detectable and semi-detectable hypotheses in terms of their relative Borel complexities? Indeed we can. We adapt a proof of Gold and Putnam from the context of computable functions to the context of arbitrary functionals over ω^ω.

Theorem 2.2, Characterization of Semi-Detectability over K:

P is semi-detectable over K $\Leftrightarrow$ P $\cap$ K $\in \Sigma_1^{1,K}$.

Proof: => Suppose that P is semi-detectable over K. Then we have that

$\exists \Phi: \omega^* \to \{0, 1\}\ \forall t \in K\ [(\exists n\ \forall m > n\ \Phi(t[m]) = 1) \Leftrightarrow t \in P]$.

Choose Φ'as promised. So

$P \cap K = \{t \in K: \exists n\ \forall m > n\ *^a(t[m]) = 1\}$.

Now define

$$R_n = \bigcup_{\substack{lh(a) = n \\ *(a) = 1}} B_\sigma^K$$

where lh(a) is the length of a. $R_n \in \Pi_1^{B,K}$ since $K \in \Pi_1^K$ and

$$R_n = K - \bigcup_{\substack{lh(a)=n \\ \phi(\sigma)\neq 1}} B_a^K$$

Observe that

$$P \cap K = \bigcup_{n\in\omega} \bigcap_{m>n} R_m$$

Since $\Pi_1^{B,K}$ is closed under countable intersections, we have that

$$\left(\bigcap_{m>n} R_m\right) \in \Pi_1^{B,K}$$

Hence Pn K is a countable union of $\Pi_1^{B,K}$ sets, and is therefore a $\Delta_2^{B,K}$ set.

<= Suppose that $P \cap K \in \Delta_2^{B,K}$. Then there is an enumerated collection {C1, C2, ..., C_n, ...} such that for each i e w, $C_j \in \Pi_1^{B,K}$ and

$$P \cap K = \bigcup_{n\in\omega} C_n$$

Let a G (o*. Let a- be the result of deleting the last item in a. Define

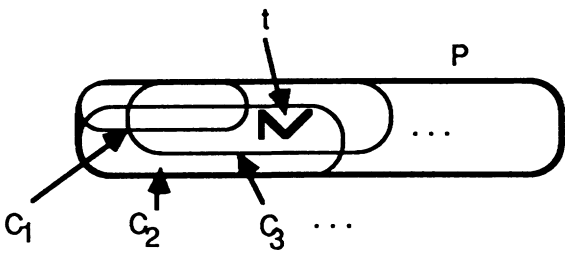
POINTER(o) = 0

POINTER(a) = $\begin{cases} \text{POINTER}(a^-) & \text{if } B^{?*}CPO^{*}TER^{W} \\ \text{POINTER}(a^-) + 1 & \text{otherwise} \end{cases}$

Let a- denote the result of deleting the last entry in a. Now define scientist

$$(f)(a) = \begin{cases} h & \text{if POINTER(a) = POINTER(a-)} \\ 0 & \text{otherwise} \end{cases}$$

Suppose $P \cap K$.

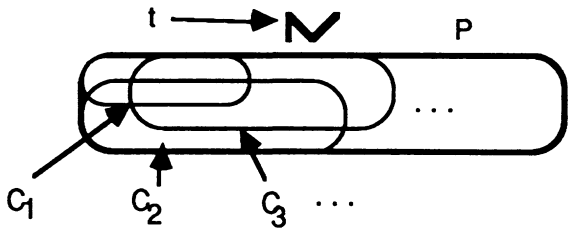


Then $\exists n\ t \in C_n$. Since $t \in C_n$, and since $C_n \in \Pi_1^{B, K}$, we have that $\overline{C_n} \in \Sigma_1^{B, K}$, so there is a $W \subseteq \omega^*$ such that

$$\overline{C_n} = \bigcup_{\sigma \in W} B_\sigma^K$$

If there is a $\sigma \subseteq t$ such that $B_\sigma^K \subseteq \overline{C_n}$ then $t \in \overline{C_n}$, which is a contradiction. Hence, for each $\sigma \subseteq t$, it is not the case that $B_\sigma^K \subseteq \overline{C_n}$. So $\forall k$ POINTER(t[k]) $\leq n$. Since POINTER never moves backward, $\exists n' \exists k \forall k' > k$ POINTER(t[k']) = n'. Hence ϕ converges to 1 as required.

Now suppose that $t \in K - P$.



Then for each $m \in \omega$, $t \notin C_m$ so for each m, $t \in \overline{C_m}$. Let $n \in \omega$. since

$$\overline{C_n} = \bigcup_{\sigma \in W} B_\sigma^K$$

there is some $\sigma \in W$ such that for some k, $t[k] = \sigma$. So for all $k' > k$, POINTER(t[k']) > n. So

(*) $\forall n \exists k \forall k' > k$ POINTER(t[k']) > n.

So there are infinitely many k such that $\phi(t[k]) = 0$, as required. ■

Corollary 2.2.a: Characterization of Detectability over K:

$$P \text{ is detectable over } K \Leftrightarrow P \cap K \in \Delta_2^{B, K}.$$

Proof: Theorem 2.2 and Fact 2.2. ■

Corollary 2.2.b: If $P \cap K$ is countable then P is semi-detectable in K.

Proof: Each singleton is $\Pi_1^{B, K}$, so $P \cap K$ can be built up as a countable union of singletons. Hence $P \cap K \in \Sigma_2^{B, K}$. ■

Corollary 2.2.c: If K is countable then P is detectable over K.

Proof: If K is countable, then $P \cap K \in \Sigma_2^{B, K}$ $\overline{P} \cap K \in \Sigma_2^{B, K}$ by the previous corollary. Apply Theorem 2.2 and Fact 2.1. ■

Theorem 2.2 may be thought of as a normal form theorem. It says that each $\Sigma_2^{B, K}$ set has a special kind of $\Sigma_2^{B, K}$ definition, namely, a semi-detector ϕ. Hence, the interest of the theorem lies in showing that a certain strategy for constructing semi-detectors is *complete*, in the sense that the construction will work for any solvable semi-detection problem. On the other hand, the result is not very useful for obtaining particular negative results, for it is no harder to show that P cannot be defined with a semi-detector than it is to show that P cannot be defined by any $\Sigma_2^{B, K}$ definition. We must look elsewhere for a useful, complete technique for obtaining unsolvability results.

It is typical in learning theory to use diagonalization to establish unsolability for particular inductive problems. A diagonal argument consists of two stages, a *fooling stage* and a *closure stage*. In the fooling stage, we keep feeding a scientist a data sequence $t \in P$ until ϕ starts to converge to 1 at some stage n. Then we show ϕ some $t' \notin P$ that $t'[n] = t[n]$ until ϕ says 0 to prevent convergence to 1. Then we switch back to some presentation $t'' \in P$, and so forth.

In the closure stage, we must show that the limit point (with respect to the induced Baire topology) of K so constructed is in fact a member of P, else 4> was not responsible for converging to 1 on K, and the intended negative argument fails. In the following sequence of definitions, we generalize the Gold/Putnam diagonal strategy [5, 22] to our general perspective on semi-detectability over arbitrary sets of data presentations.

Say a is *in P* <=» 3t *e P* s.t. a c t .

£ is an *extension function for P* <=>

Vt *e* co®, if Vn e © 3m > n 3k > m s.t. t[m] is in P and t[k] = C(t[m]) then t e P.

K is a *P-tree* <=> P, $\bar{P}$ are dense in K.

P *semi-demon* in K <=> $3K^1$ Q K s.t.
(1) K^1 is a P-tree and
(2) 3£ s.t. £ is an extension function for P.

P has a *demon* in K <=» $3K^1$ c K s.t.
(1) K^1 is a P-tree and
(2) 3C s.t. C is an extension function for K^f.

Proposition 2.3:

If P has a semi demon in K then P is not semi-detectable over K

Proof: Suppose that P has a demon in K. $3K^1$ *Q* K s.t. K^f is a P-tree and 3£ s.t. C is an extension function for P n K\ Suppose for reductio that $ semi-detects P over K. We construct a t e P on which $ fails to converge to 1. Let us construct t in nested finite stages a[i].

a[0] = <>

a[n+1]:

If <t>(a[n]) = 1 then since K^1 is a P-tree, 3f e $\bar{P}$ s.t. a[n] $_c$ t\ Find the least k ≥ lh(a[n]) s.t. **4>(f[k])** - 0. There is one, else 4> converges to 1 on some *V e* P. Then let **σ[n+1] = ζ(t'[k]).**

If <t>(a[n]) = 0 3f e P s.t. a[n] c t^1. Find the least k ≥ lh(a([n]) s.t. 4>(f[k]) = 1. There is one, else $ fails to converge to 1 on t^1 e P. Then let a[n+1] = C(f[k]).

Define t = UJ$_e$ © a[i]

Observe that by construction,

Vn e co 3m > n 3k≥ m s.t. t[m] is in P and t[k] = C(t[m])

Since £ is an extension function for P in K\ we have that t *e* P. But by construction $ does not converge to 1 on P. Contradiction. So no ϕ can semi-detect P over K. The corollary follows from Theorem 2.1. H

Corollary 2.3.a: If P has a demon in K then P n K $\hat{}$ $\hat{}$ $^{V B, K}$ •

Proof: Proposition 2.3 and Theorem 2.2. **B**

Corollary 2.3.b:

(a) if P has a demon in K then P is not detectable over K.
(b) If P has a semi-demon in K or $\bar{P}$ has a semi-demon in K then P is not detectable over K.
(c) if P has a semi-demon in K or $\bar{P}$ has a demon in K, then P n K « $^{A}2^{B,K}$.
(d) if P has a demon in K then P n K e $^{A}2^{B, K}$.

Proof: (a) Same as the proof of Theorem 2.3, except that now it doesn't matter whether the data presentation is in P n K^1 or in $\bar{P}$ n K^f. (b-d) trivial. B

It would be very desirable to have an exact characterization of the relationship between diagonalization and $^{Y}2^{B, K}$ sets. But at present, we can give at least a partial converse to Proposition 2.3.

Proposition 2.4: If no subset of K is a P-tree then P is detectable over K.

Suppose that for each S c K, either P or $\bar{P}$ is not dense in S w.r.t. topology $3B^K$. Say that B£ is a *homogeneous fan* for P $\Leftrightarrow$ Vt G B^K_a, P(t). Then we have that K contains some homogeneous fan for P or for $\bar{P}$. Now we define an inductive procedure for stripping off homogeneous fans from K until K disappears. Then we use the fans as clues for a reliable inductive method.

P-fan[0] = 0
$\bar{P}$-fan[0] = 0
K[0] = K

P-fan[n+1] = {$B?^{nl}$: a G co* and$B?^{nl}$ * 0 and $Ba^{K[1]}_{11}$ is a homogeneous fan for P}
P-fan[n+1] = {B_a : a G co* and $B?^{-}$ * 0 and $^{B}<r$ is a homogeneous fan for P}
K[n+1] = K[n] - (P-fan[k+1] u $\bar{P}$-fan[k+1])

Since for each each S ^ K, either P or $\bar{P}$ is not dense in S w.r.t. topology JB^K, we have that at each stage n in the construction, either P-fan[n] * 0 or $\bar{P}$-fan[n] * 0. K[o)], the limit of this process, is the empty set, since each homogeneous $B^{K[n]}_a$ becomes homogeneous at some finite stage in the construction and is then removed.

Now we define an inductive method that semi-detects P over K.

CLUE(a) = the least x s.t. a c x and 3n s.t. $B?^{nl}{}_e$ (P-fan[n+1] u $\bar{P}$-fan[n+1]).

$$\phi(\sigma) = \begin{cases} i \text{ if } 3n \text{ s.t. } CLUE(a)i \text{ G } P\text{-fan}[n] \\ 0 \text{ otherwise.} \end{cases}$$

Let t G K. Then 3n, a s.t. t G $B?^{nl}$. So 3k CLUE(t[k]) = a = t[kj. Suppose t G P. Then $\bar{B}T$ e P-fan[n+1]. So Φ converges to 1 as required. Suppose *te* K - P. Then 3n $Bi^{\wedge 1}$ G $\bar{P}$-fan[n+1], So Φ converges to 0 as required. B

By proposition 2.4, the non-existence of a semi-demon is necessary and sufficient for semi-detectability *if* the answer to the following question is positive.

Question 2.5: If P n K e $\wedge^{B,\ K}_{t}$ then P has an extension function.

So a positive answer to this question would amount to a general completeness theorem for the Gold/Putnam diagonal argument as a way of proving a semi-detection problem unsolvable. It is noteworthy that the asymmetry between the solvable $\Sigma_2^{B,K}$ case and the unsolvable $\Pi_2^{B,K}$ case depends entirely on the existence of an extension function, and not on the presence of a P-tree, which is symmetric regarding the two cases.

3. Arithmetic Characterizations of Effective Inductive Inference

It is well known that the Borel hierarchy bears a close resemblance to the recursion-theoretic arithmetical hierarchy [8]. Many results that hold in the Borel hierarchy may be massaged gently to hold in the more restrictive, computational setting. This analogy suggests an analogously general characterization of semi-detectability and detectability by *computable* scientists.

P is *effectively* [semi-] detectable <=>
3 total recursive <f>: ω* -> {0, 1} s.t. 4> [semi-] detects P.

Gold and Putnam originally characterized limiting recursion in terms of the arithmetic hierarchy of sets of natural numbers [6, 21]. The only difference between their characterization and the one taken here is that we will be working in the arithmetic hierarchy of sets of *functions* on the natural numbers [8, 23]. This hierarchy is based upon the notion of a *partial recursive functional.* The recursive functional <f>[t, n] may be viewed as a functional computed by the following sort of Turing machine Mj. Mj reads the natural number n as an ordinary input. Mj cannot read t as an ordinary input, because t can't be read in a finite amount of time, so by that fact alone, Mfs computation would never halt. But Mj can be viewed as having t as an input if Mj scans only a finite chunk of t before making an output. So we may view a partial recursive functional as a functional computed by a Turing machine that scans at most a finite initial segment of t before making an output.

We may think of the behavior of a computable scientist as a total recursive functional 4>[t, n]. The input t is the infinite data sequence, and the input n tells the scientist what stage of inquiry he is at. At stage n, <fj scans as much of t as is necessary to produce his next output. Such a scientist may be turned into a computable map y:©* -> {1, 0} just by having y repeat <f>s last conjecture and feed new data to <fj until <fj comes up with a

new conjecture. The behaviors of <j and y are not identical, but they are the same up to convergence in the limit. Thus, thinking of scientists as partial recursive functionals allows us to plug computable inductive inference problems directly into the arithmetic functional hierarchy. This is an extremely useful move, for the arithmetic function hierarchy automatically mixes together the topological and computational aspects of inductive inference in just the right way to permit simple characterizations of computerized inductive inference.

Say that a relation has *type* <j, k> just in case it has j function places and k natural number places. Henceforth, when we speak of a relation, we imply that the relation has type <j, k> for some finite j and k. Then

$^{v}\mathbf{O}^{0}_{1}$ = the set of all relations with partial recursive positive tests.

$\mathbf{TT}^{0}_{1}$ = = the set of all complements of relations in $\mathbf{T}^{0}_{1}$.

$\mathbf{Z}^{0}_{n+1}$ = {3xi, -i $3x_n$ R : R e $^{n}\mathbf{n}^{0}$ and xi, . . . , x_n are number variables}.

$^{n}\mathbf{n}^{0}_{+i}$ = {Vx1, ..., Vxn R: R e $\mathbf{£}^{0}_{n}$ and x-|,..., x_n are number variables}.

The arithmetic hierarchy is here defined in a way that emphasizes its similarity to the Borel hierarchy. The differences may be summarized by the difference in definition for $\mathbf{£}^{0}_{1}$, and in the replacement of countable union and intersection with existential and universal quantification.

As in our Borel characterization, we would like to relativize the arithmetic hierarchy to background knowledge. In recursion theory, it is usual to speak of relative computation as computation with queries to an oracle for some non-recursive set of numbers [8], [23]. But this sort of relativization is inappropriate for our application. Background knowledge over data presentations is not an oracle that permits the scientist to decide a non-recursive set of numbers; it is, rather, a restriction on the set of infinite sequences the scientist may possibly encounter. Hence, we need to relativize the arithmetic hierarchy to a subset of the set of all possible data presentations. This is exactly what we did in the Borel case. Accordingly, define

Mj is a *partial recursive positive test for* type <j, k> relation R mod K G ©*^{0}
⇔

$\forall t_1, \ldots, t_j \in K\ \forall x_1, \ldots, x_k \in \omega[R(t_1, \ldots, t_j, x_1, \ldots, x_k) <*$
$M_j[t_1, \ldots, t_j, x_1, \ldots, x_k]$ halts]

Z_1 = the set of all relations with partial recursive positive tests mod K.

$zfj^{1\,K}$, $11^{\wedge K}$, and $AJ\}^{1K}$ are all defined in terms of $*V^{\Omega, K}$ in the same manner as before.
Now we can state the second characterization theorem.

Theorem 3, Characterization of Effective Semi-Detectability over K:

P is effectively semi-detectable over K w P n K $\in \wedge^{o, K}$ •

Proof: =» Suppose P is effectively semi-detectable over K. Then by definition, there is a total recursive <|>: ©* -» {0, 1} s.t.

$\forall t \in K\ [(\exists n\ \forall m > n\ \langle\rangle(t[m]) = 1) \Rightarrow t \in P]$.

Sb

$P \cap K = \{t \in K: \exists n\ \forall m > n\ \langle\rangle(t[m]) = 1\}$.

The relation R(t, m) *=> <|)(t[m]) = 1 is RE since $ is total recursive. Hence, the relation 3n Vm > n R(t, m) is in $Z_2^{J>, K}$ •

«= Suppose that P n K is a type <1, 0> relation in $Z_2^{o, K}$ • Then there is a relation R e $Z_1^{o, K}$ such that

P(t) & K(t) « 3y -nR(t, y),

where y is an n-vector of first-order variables. Let Mj[t, x] be a machine that halts if and only if R(t, x), since R e $\wedge^{o, K}$. We construct a total recursive analogue of the topological POINTER method. Let <n> be a recursive bijection from co to co^n. Let a e CD*. Define

$T_j[a, \langle x \rangle, u] \Leftrightarrow M_j[a, \langle x \rangle_1, \ldots, \langle x \rangle_n]^J$- in u steps.

Observe that $M_j[o, \langle x \rangle, u]$ may fail to halt either because too few computational steps are allocated by u or because Mj's queries run off the end of a. Define Φ_a* -* {0, 1} as follows:

$$POINTER(o) = 0$$

$$POINTER(\sigma) = \begin{cases} (POINTER(a-) \text{ if } -.T_i(\langle r, \langle POINTER(a-) \rangle, lh(a)) \\ |POINTER(a-) + 1 \text{ otherwise} \end{cases}$$

$$\langle t \rangle(a) = \begin{cases} \ldots . \; POINTER(a) = POINTER(o-) \\ |0 \text{ otherwise} \end{cases}$$

Suppose t G **P n K.** Then 3y -iR(t, y). Let x e co be least such that $\langle x \rangle = y$. Since -iR(t, $\langle x \rangle$), we have that for each u, -iTj(t[u], $\langle x \rangle$, u). Since for each u, -iTj(t[u], $\langle x \rangle$, u)$_f$ it follows that $Vu^1 > u$, POINTER(t[u']) ≤ x. So since POINTER never goes backwards, Φ converges to 1 as required.

So suppose that t e K - P. Then Vy R(t, y). So for each x, R(t, $\langle x \rangle$). So For each x, there is a u such that Tj(t[u], $\langle x \rangle$, u). Hence, for each x there is a u such that POINTER(t[u]) > x. So Φ produces infinitely many 0's on t, as required.

Finally, Φ is easily seen to be total recursive. ▪

Corollary 3.a: Characterization of Effective Detectability over K:

P is effectively detectable over K « P n K e $^{AA^{0}}2^{K}$.

Proof: Theorem 3 and the obvious, effective analogue of Fact 1. **B**

4. Applications of the Theorems

In light of the previous results, learning theoretic questions reduce to standard topological and recursion theoretic questions about membership in $^{B}2^{K}$ and $^{\Theta K}2$.

respectively. Conversely, learning-theoretic techniques may be viewed as techniques for establishing membership and non-membership in these interesting classes.

Example (Philosophy of Mind): Let P = {t: t is a total recursive function}, K = ω<ω. This can be thought of as a formalization of the inductive problem facing one who would determine whether or not an unknown system (e.g. a human) is a computer. P ∈ $\Sigma^0_3 - \Pi^0_3$ [23, p. 356]. Hence, P is not effectively semi-detectable over ω®, by Theorem 3.

But it is easy to verify that P has a demon in K. First, each finite initial segment of a function is extendable both by a recursive function and by a non-computable function, so K is a P-tree. Second, the identity function is trivially an extension function for K, since every possible sequence is in K. Hence, by Corollary 2.3.b, P is not even semi-detectable in the limit by a god with no computational limitations but with spatio-temporally localized sensory apparatus.

Example (Kant): Let P = {t: t has infinitely many occurrences of 1} and let K = {t: t is Boolean}. This problem corresponds to Kant's second conflict of pure reason, concerning the question whether matter be infinitely divisible. A data sequence is generated by picking a cutting instrument and applying it to the remaining half of a banana. If the cutting instrument fails, one must chooses a more costly and refined instrument (i.e. when we reach the level of atoms, particle accelerators must be used. After that, who knows? Each time a cut is achieved, we write down a 1, and each time a cutting instrument fails, we write down a 0. For all we know at the outset of the experiment, any sequence may arise. So let K be the set of all Boolean co-sequences. Kant thought that the infinite divisibility of bananas cannot be determined even on the basis of all possible experience [9]. One reading of this claim is that P cannot be detected in the limit over K. But this is correct, for K is a P-tree, and the function £(a) = a*(01) is an extension function for P. So by Corollary 2.3.b, P is not even semi-detectable over K. But by definition, it is clear that P n K $\in \Pi^0_2$. Hence, P n K $\in \Pi^0_2 - \Sigma^0_2$ •

Suppose, now, that we have background knowledge K\ ensuring that our experiment either converges to 0 or converges to 1. That is, either we reach a stage after which all of our attempted cuts work, or we reach a stage at which no attempted cut works. (This is entirely implausible, but we consider the example for logical reasons). Then K^1 = the set of all convergent Boolean co-sequences. So by Corollary 2.2.c, P is detectable over

K', so $P \cap K' \in \Delta_2^{B, K'}$. So background knowledge K' collapses the relative Borel complexity of P sufficiently to make P not only semi-detectable but detectable in the limit.

Observe that an effective scientist can detect P over K'. Just define $\phi(\sigma)$ = the last item occurring in σ. Hence, we also have that $P \cap K \in \Delta_2^{0, K'}$.

Example (Bounded Rationality): It is interesting to consider the kinds of inductive problems that are detectable or semi-detectable, but not by any machine. In light of the results of the last two sections, these are just the problems <K, P> such that $P \cap K \in \Delta_2^{B, K} - \Delta_2^{0, K}$ or $P \in \Sigma_2^{B, K} - \Sigma_2^{0, K}$, respectively. Such problems are topologically easy in light of "clues" that are hard for a computer to decipher. For example, consider $K = \omega^\omega$ and $P = \{t: W_{t_1} \text{ is infinite}\}$. Clearly, $P \in \Delta_1^B$. But $Inf = \{x: W_x \text{ is infinite}\}$ is complete in Π_2^0 [23, p. 326] and hence is not Σ_2^0. So suppose that $P \in \Sigma_2^0$. Then we have some total recursive semi-detector ϕ of P over ω^ω, by Theorem 3. Let t be a total recursive function. Define $Inf(x) \Leftrightarrow \exists n\ \forall m>n\ \phi(x^*t[m]) = 1$, where x^*t is the result of tacking x onto the front of t. Hence, $Inf \in \Sigma_2^0$, contradiction. So $P \in \Delta_1^B - \Sigma_2^0$ and hence is detectable (with no mind-changes) by an ineffective scientist, but cannot even be semi-detected by a computable scientist.

Example (Logical Hypotheses): Several papers have been published on the detectability of hypotheses in a logical language over a complete enumeration of the diagram of a relational structure [10, 11, 12, 14, 17, 18, 19, 24].

Let L be a first-order logical language, $h \in L$, $T \subseteq L$. Let STR(L) = the set of all countable relational structures for L. Let EC(T) be the set of all in STR(L) in which T is true. Let $\mathfrak{A} \in$ STR(L). Let v be a variable assignment for $\mathfrak{A}$. $\Delta(\mathfrak{A}, v) = \{\pm e: e \text{ is a closed L-atom and } \mathfrak{A} \models \pm e\}$. $K(\mathfrak{A}, v) = \{t: t \text{ enumerates } \Delta(\mathfrak{A}, v)\}$. $K(\mathfrak{A}) = \{t: \exists v \text{ s.t. } t \text{ enumerates } \Delta(\mathfrak{A}, v) \text{ and } v \text{ is onto } |\mathfrak{A}|\}$. $K(T) = \cup\{K(\mathfrak{A}): \mathfrak{A} \in E.C.(T)\}$. So a logical inductive inference problem <T, h> automatically generates a topological inference problem <K(T), K(h)>. Say that a formula $\phi \in \Sigma_n$ if and only if ϕ is logically equivalent to a formula of L with at most n blocks of adjacent quantifiers of the same type. $\phi \in \Pi_n$ if and only if ϕ is logically equivalent to the negation of a Σ_n formula. Say that $\phi \in \Sigma_n^T$ [Π_n^T] just in case $\exists \psi \in \Sigma_n$ [Π_n] such that $T \models \phi \leftrightarrow \psi$. Finally, $\phi \in \Delta_n^T$ just in case $\phi \in \Sigma_n^T$

and $\phi \in \Pi^0_n$. Then we have the following relationship between logical hypotheses and the Borel sets of data presentations they generate. Since our logical formulas are finitary, the correspondence works only up to ω.

Proposition 4.1: Let $T \subseteq L$. Let h be a sentence in L. Then $\forall n \in \omega$

(1) $h \in \Sigma^0_n \Rightarrow K(h) \cap K(T) \in \Sigma^{0,K(T)}_n$ and

(2) $h \in \Pi^0_n \Rightarrow \overline{K}(h) \cap K(T) \in \Pi^{0,K(T)}_n$

Proof: Choose a countable, recursive set Const of constants and form the expanded language Lconst that results from adding Const to L. Then let STR(L, Const) be the set of all countable structures for L in which each domain element is named by some $c \in$ Const. Since the structures in STR(L) are all countable, each $\alpha \in$ STR(L) is the reduct that results from eliminating the specification of denotations for the constants in Const from some $\mathfrak{B} \in$ STR(L, Const). Say then that α is the Const-reduct of $\mathfrak{B}$.

Suppose $\mathfrak{B} \in$ STR(L, Const). Define $K(\mathfrak{B}) = K(\alpha)$, where α is the Const-reduct of $\mathfrak{B}$. So no elements of Const occur in the elements of $K(\mathfrak{B})$. If some $c \in$ Const occurs in h, then define $K(h) = \cup\{K(\mathfrak{B}): \mathfrak{B} \in$ STR(L, Const) and $\mathfrak{B} \models h\}$. Now we establish the proposition by showing something slightly different:

Let $T \subseteq L$. Let h be a sentence in Lconst. Then

(1) $h \in \Sigma \Rightarrow K(h) \cap K(T) \in \Sigma^{0,K(T)}_1$ and

(2) $h \in \Pi \Rightarrow K(h) \cap K(T) \in \Pi^{0,K(T)}_1$

Base case: (1) Let $h \in \wedge^T$. Then $T \models h \leftrightarrow \exists x \langle\phi\rangle$, where $\langle\phi\rangle$ is quantifier-free. Let α be a finite data sequence. Say that α *verifies* $\exists x \langle\phi\rangle \Leftrightarrow \exists$ bijective assignment $\%$ of constants in Const to variables in α s.t. the result of applying $\%$ uniformly to α and conjoining the resulting literals entails $\exists x \langle\phi\rangle$.

Suppose $\mathfrak{B} \in$ STR(L, Const) s.t. $\mathfrak{B} \models T \cup \{h\}$ & $t \in K(\mathfrak{B})$. Then $\exists n$ s.t. $t[n]$ verifies $\exists x \langle\phi\rangle$. In the other direction, suppose $t \in K(T)$ and for some n, $t[n]$ verifies $\exists x \langle\phi\rangle$. Since

t[n] verifies $\exists \mathbf{x}\phi$, there is an assignment χ of variables to constants in Const so that $t[n]\chi \models \exists \mathbf{x}\phi$. Since $t \in K(T)$ and no $c \in$ Const occurs in T, $\exists \mathfrak{A} \in STR(L)$ s.t. $\mathfrak{A} \models T$ and $\mathfrak{A}$ satisfies all the literals occurring in t according to some variable assignment v onto the domain of $\mathfrak{A}$. Now expand $\mathfrak{A}$ to $\mathfrak{B} \in STR(L, Const)$ by interpreting each constant $c \in$ Const that occurs in h by the domain element $v(\chi^{-1}(c))$. Now $\mathfrak{B} \models T \cup \{h\}$ and $t \in K(\mathfrak{B})$, so $t \in K(h)$. Hence,

(*) If $t \in K(T)$ then

$\exists n$ s.t. t[n] verifies $\exists \mathbf{x}\phi \Leftrightarrow$

$\exists \mathfrak{B} \in STR(L, Const)$ s.t. $\mathfrak{B} \models T \cup \{h\}$ & $t \in K(\mathfrak{B}) \Leftrightarrow$

$t \in K(\exists \mathbf{x}\phi) \Leftrightarrow$

$t \in K(h)$

Now define $S = \{\sigma: \exists t \in K(T)$ s.t. $\sigma \subseteq t$ and t verifies $h\}$. Then $K(h) \cap K(T) = \bigcup_{\sigma \in S} B_\sigma^{K(T)}$, by (*). So $K(h) \cap K(T) \in \Sigma_1^{B, K(T)}$. The dual argument works for (2).

Induction:

(1) Let $h \in \Sigma_{n+1}^T$. Then there is a $\phi \in \Pi_n^T$ s.t. $T \models h \leftrightarrow \exists \mathbf{x}\phi(\mathbf{x})$ where $\mathbf{x}$ is an n-vector of variables. Let $\mathbf{c} \in Const^n$. Then by the induction hypothesis, $K(\phi(\mathbf{c})) \in \Pi_n^{B, K(T)}$.

Note that for each $t \in K(T)$,

$t \in K(h) \Leftrightarrow$

$t \in \cup\{K(\mathfrak{A}): \mathfrak{A} \in STR(L, Const)$ and $\mathfrak{A} \models h\} \Leftrightarrow$

$t \in \cup\{K(\mathfrak{A}): \mathfrak{A} \in STR(L, Const)$ and $\exists \mathbf{c} \in Const^n$ s.t. $\mathfrak{A} \models \phi(\mathbf{c})\} \Leftrightarrow$

$t \in \cup\{K(\phi(\mathbf{c}): \mathbf{c} \in Const^n\}$

So $K(h) \cap K(T) \in \Sigma_{n+1}^{B, K(T)}$.

The induction for (2) is dual. Finally, since the proposition holds for each sentence of L_{Const}, it holds for each sentence of L. ■

On the other side, we can show:

Proposition 4.2: If h e Δ_2^T then K(h) is not semi-detectable over K(T).

Proof: Suppose that h e Δ_2^T. By application of Chang and Keisler Theorem 3.1.6 at various stages in the proof of Chang and Keisler's theorem 3.2.3 [4] we obtain that if h e Σ_2^T then there is a countable elementary chain of countable models of T $\{\&J: i \in \omega\}$ s.t. Vi e co 8j |* h and the countable model & = $u_{j\in G}$> Sj |= h.

Suppose Φ can semi-detect K(h) over K(T). Then we may diagonalize as follows. Present evidence from S until $ says 1, which must eventually happen. Then there is an n such that S_n l= the evidence so far presented, since S is the union of the &j and the evidence presented so far is finite and the evidence is quantifier free. Start presenting evidence from S_n until Φ says 0. Then since fl is the union of an elementary chain, we have by Chang and Keisler's Theorem 3.1.13 [4] that & satisfies the evidence presented so far out of fl_n. If we are careful to present new evidence from a fixed t e K(&) each time we return to S in this construction, we shall have presented complete evidence for some model S ofT u {h} on which Φ fails to converge. Contradiction. M

Corollary 4.2.a: Let L be a first-order language and let h be a sentence of L Then

(1) K(h) is semi-detectable over K(T) « h e Σ_2^T.

(2) K(h) is detectable over K(T) » h *e* Δ_2^T.

Proof: (1) Propositions 4.1, 4.2, (2) Fact 2.1. ■

Furthermore, we can show the following surprising fact about effective hypothesis investigation. Together with Proposition 4.2, it implies that computers are just as good as Greek gods at truth detection of first-order sentences. This is less surprising when one considers that determining whether a hypothesis is syntactically verified or refuted by data is decidable and existential quantification is a relatively "clean" way of taking countable unions over Borel sets. Recall that in the proof of Theorem 3, it was these two features of $\Sigma_2^{vo, K}$ sets that enabled us to construct a computable learner for each such set.

Proposition 4.3:

(1) K(h) is effectively semi-detectable over K(T) <=> h e $\hat{2}_2^T$.

(2) K(h) is effectively detectable over K(T) » h e A_2^T.

Proof: **(1) => Proposition 4.2.**

<= 3\|/eni s.t. T |= h <=> 3xy(x). $ uses a pointer-and-enumeration method just as in Theorem 3, except that the pointer advances over an effective enumeration of formulas $v(^{x}i)$» $v(^{x}2)$» ••• where xi, X2, ... are vectors variables of L The pointer advances whenever the universal formula Vy£(xj, y) it points to is directly refuted by the data, in the sense that some set S of literals occur in the data such that S is propositionally inconsistent with Vy£(xj, y) under the substitution of some vector y^1 for y. This refutation test is recursive. Hence, the resulting scientist $ is total recursive over initial data segments. (2) follows from an effective version of Fact 2.1. B

Using techniques described in [10, 14], we can generalize Proposition 4.3 to cases in which the evidence formulas are A_n instead of A1.

Proposition 4.4:

(1) K(h) is effectively semi-detectable over K(T) from A_n-data <=> he $^{Y}A_{n+1}^{T}$.

(2) K(h) is effectively detectable over K(T) from A_n-data <=> h e $^{AA}n+1^{T}$.

Proof: **Similar to Proposition 3.3, except we use Chang and Keisler Theorem 5.2.8. (the Keisler n-sandwich theorem) instead of 3.2.3. C.f. [10, 13]. B**

Osherson and Weinstein [17] have constructed a universal inductive inference machine, that for any T and h, detects K(h) over K(T) just in case it is possible for any machine to do so. In our framework, their machine may be viewed roughly like this. Using an oracle for T, the machine isolates \p. Then it proceeds to use y just the way our $ does.

Many of the positive results of [12] follow directly from these propositions. For example, suppose that h is expressed in a monadic language. Each monadic sentence h is

A2, so <p v -ip, h> induces the problem <K(p v -ip), K(h)>, where K(h) e $^{*2}_{.B, K(pv^p)}$. Similarly, we demonstrated in [11] that no hypothesis with ineliminable quantifier alternations is detectable over arbitrary logical data presentations. This is because such a hypothesis generates a set of presentations that is not $^{A}_{A}B_f$ K(pv-.p)^{2}.

This perspective on the first-order paradigm is useful because it essentially segregates the topological, the computational, and the model-theoretic aspects of logical inductive inference problems. The intricate model theoretic arguments required in studies of first-order logical inductive inference can now be seen as just one special way to induce a Borel set of a given complexity over data sequences. Once we have plugged the logical framework into the Borel hierarchy, we arrive at epistemic insights that are at once simpler and more general.

5. **Characterizations of Identifiability**

Formal learning theory began with the study of language identification. Instead of trying to decide the truth value of a given hypothesis in the limit, an identifier is required to produce a correct theory for the data presentation he is on. So while detection problems are intended to assess strategies for hypothesis evaluation, identification problems are intended to assess strategies for discovering new theories.

The setting for identifiability assumed here is more general than usual. We assume that each data presentation has a hypothesis that is empirically adequate for it. No further structure is assumed on the notion of empirical adequacy. In particular, we do not make the common learning-theoretic assumption that two hypotheses are either empirically adequate for exactly the same data presentations, or are empirically adequate for disjoint sets of data presentations..

Let R(t, i) be a relation of type <1, 1>. Define

dom(R) = {t: 3i e © s.t. R(t, i)}

mg(R) = {i: at R(t, i)}.

R|K = {<t, i>: t E K & R(t, i)}.

Ki^{R}= {t: R(t, i)}.

R(t, i) *Is an adequacy relation* for K <=> K c dom(R).

Let R be an adequacy relation over K.

R' *is a simplification of* R in K <=>
Vke rng(R') K ^ K[R] and
Vt *e* K 3i G co s.t. t G K?

$ identifies R *over* K » Vt G K 3i R(t, i) and 3n Vm > n <t>(t[m]) = j.

R *is identifiable over* K <=> 3<t>: co* -> co: <t> identifies R over K.

R *is effectively identifiable over* K <=>
3 partial recursive <t>: co* -» co: such that 4> identifies R over K.

Proposition 5.1: A Characterization of Ineffective Identifiability over K:
Let R be an adequacy relation over K.

R is identifiable over K <=> Vi, [R]Kj n K G $Z_2^{B_* K}$ •

Proof: => an R-identifier is a semi-detector for each K_j^p, j e co. Apply Theorem 2.2.

<= By Theorem 2.2, we may assume a semi-detector Φ^j for each Kj^R, j e co. Define <t>(a) to output the first i such that Φ^j has not said 0 any more recently than any other Φ^i on a. ■

Theorem 5.2: A Characterization of Effective Identifiability over K.
Let R be an adequacy relation over K.

R is effectively identifiable over K »

3R' s.t. R^f is a simplification of R in K and R'|K e $F_2^{\circ>^K}$.

Proof: <= Suppose that R is effectively identifiable over K by total recursive $. Then

Vt *e* K 3i 3n s.t. R(t, i) and Vm > n <t>(t[m]) = i.

Define

R'(t, i) » 3n Vm > n <|)(t[m]) = i.

The relation R(t, m, i) <=> <|>(t[m]) = i is RE since $ is partial recursive. Hence the relation R'(t, i)|K G $^{o, K}_{Z2}$. We know that Vt e K 3k s.t. t e Kk, else 4 fails to converge on some t G K, which is absurd. Finally, we know that Vk G rng(R') Kk = Kk, for otherwise, 3t e Kk on which $ does not converge to k, which contradicts the definition of R\

=> Suppose that there is a simplification R^1 of R such that R'|K € $^{J>, K}_{^2}$. Then we have Vt G K, R'(t, i) « 3x -iS(t, i, x), where x is an n-vector of first order variables, and where S(t, i, x) G $^{0, K}_{i}$. Let Mj[t, i, x] be a machine that halts if and only if S(t, i, x), since S e $^{o, K}_{t}$ Let <x> be a recursive bijection from co to co^{n+1}. (Think of <x>i as encoding i, and of <x>2,..., $<x>_n$+i as encoding xi. . . . x_n. The inconvenience is required to dovetail a search on all of these arguments in parallel). Let a G co*. Define

Tj[a, <x>, u] <=> Mj[a, <x>i, <x>2. . . . .$<x>_n$]^ in u steps.

Let <> denote the empty sequence. Define

POINTER(o) = 0

$$\text{POINTER}(a) = \begin{cases} \text{POINTER}(<\text{H if-Ti}(a, <\text{POINTER}(a\text{-})>, \text{lh}(a)) \\ \text{POINTER}(a\text{-}) + 1 \text{ otherwise} \end{cases}$$

4>(a) = <POINTER(a)>i.

Let t G K. Suppose z is such that -iR'(t, <z'>i). Then since R^f(t, i) <=> 3x —iS(t, i, x), we know that S(t, <z>). So 3k Mi(t, <z>) halts in k steps after asking questions only about initial segment t[k] of t. So 3k such that Tj[t[k], <z>, k]. Therefore,

(*) if-.R^f(t, <z>i) then Vk^1 > k, POINTER(t[k^f]) > z.

Since R^1 is a simplification of naming relation R for K, there is an i such that R^f(t, i) and R(t, i). So choose w to be the least w^1 G CO such that -iS(t, <w^f>). Hence Mj(t, <w>) never halts. Hence, for each k, we have —**iTi(t[k],** <z>, k). So Vm G CO POINTER(tfm]) ≤ w. So either POINTER converges to w, or by (*), POINTER converges to some w^1 < w such that R'(t, <w'>i). But then 4 converges to some i such that R(t, i), as required.

Finally, it is easy to see that $ is total recursive. •

Fact 5.3: [Total recursive] $ detects P over K »
<$ identifies naming relation {<t, 1>: t e P} u {<t, 0>: t e P} over K.

Corollary: Corollary 3.a Corollary 2.2.a are special cases of Fact 5.3. ^

Example (Language Learning): A new characterization theorem for effective language learnability paradigm drops out of theorem 5.3 if we think of our collection K of languages as inducing an empirical adequacy relation over data presentations. Let L e RE. Define, for each i s.t. Wj e L,

$R_L(t, i) = \{^{<l}\text{»}^{i>:}$ Wj e L and t is an enumeration of Wj.

$K_L = \{t: \exists i \text{ s.t. } R_L(t, i)\}$.

Then <K^, R^> is identification problem generated by language identification problem L. Angluin [1] has established the following characterization theorem for effective language learnability from positive data:

L is *RE-indexable* <=>there is a type <0, 2> relation Q(i, x) e $\hat{}^0_1$ such that for each L e there is an i e co such that L = {x: Q(i, x)}.

Define Lj = {x: Q(i, x)}

If L is RE-indexable then

L is identifiable <=> 3 type <0, 2> relation P(i, k) e $\hat{}^0_1$ such that for each i e co,

(1) {k: P(i, k)} is finite and
(2) {k: P(i, k)} Q L| and
(3) Vk G co, {k: P(i, k)} G Lj => L4, *ct* Lj.

Angluin's theorem assumes that L is RE-indexable. This assumption is used to enumerate all and only the relevant hypotheses in order to produce a successful $ whenever the right-hand-side of the theorem is true. Theorem 5.2 does not require this assumption. It shows that one can succeed by means of a mechanical enumeration of

parts of hypotheses even when one cannot mechanically enumerate the hypotheses themselves.

The negative side of Angluin's theorem is proved by a variant of the locking sequence technique, and this sort of argument depends essentially upon the requirement that ϕ be forced to converge to the truth on every enumeration of an RE set. Theorem 5.2 does not impose this requirement. This generalization is achieved by replacing the locking sequence argument with an argument that uses the complexity of definition of learning to bound the complexity of solvable inductive problems.

Finally, the language learning paradigm assumes that the relation of empirical adequacy induces a partition over data presentations, so that each two hypotheses either have exactly the same data presentations or share no data presentations. Theorem 5 shows that this assumption is also unnecessary for the characterization of language learnability.

Example (Recursive Function Identification):

Theorems 5.1 and 5.2 may be viewed as theorems about function identification. Let K be a set of functions to be identified. The function identification problem K induces the following adequacy relation:

$$R_K = \{<t, i>: i \in \omega \text{ and } t \in K \text{ and } t = \phi_i\}.$$

Since the set Rec of total recursive functions is countable, we have by Theorem 5.1 and Corollary 2.2.c that each $K \subseteq$ Rec is ineffectively identifiable.

Gold has shown [5, 6] that Rec is not effectively identifiable. So according to Theorem 5.2, there is no simplification R' of R_{Rec} such that $R'|Rec \in \Sigma_2^{0, Rec}$.

6. Detectability and Bayesian Convergence to Certainty

A pressing question for formal learning theorists is to specity the relation of learning theoretic results to the standard probabilistic convergence theorems. For example, consider the following special case of a convergence theorem reported in [7].

Theorem (Halmos [7]): Let ji be an *arbitrary* probability measure on JB_a.

V ^-measurable P a^-measurable S s.t.

$$H(S) = 0$$

$$\forall t \in \bar{S},\ n(P|t[n]) \xrightarrow[n]{} \begin{cases} 1 \text{ if } t \in P \\ 0 \text{ otherwise} \end{cases}$$

The convergence of the conditional measure n(Jt[n]) to 1 or 0 induces a scientist that detects P over $\bar{S}$ as follows:

$$\phi(\sigma) = \begin{cases} 1 \text{ if } \mu(P|\sigma) > .5 \\ 0 \text{ otherwise} \end{cases}$$

So whenever *i(P|t[n]) converges correctly to 0 or 1 for each data presentation in K, some scientist <(>(t[n]) detects P over K. Hence, by Corollary 2.a, we have immediately that $P \cap K \in {}^{A}_{A}2^{B,K}$.

What does this say about the statistical convergence theorem? Just this. The Bayesian inductive method of conditionalization certainly works, but this is because each hypothesis restricted to K is a $^{A}2^{B\,»\,K}$ hypothesis, where K is the complement of a set of data presentations of measure 0. The Bayesian can reliably infer "everything" using his probabilities, but that is just because the Bayesian is probabilistically *certain* that "everything" is just what a learning-theoretic detector can reliably infer without probabilities, namely, the $^{A}2^{B,K}$ hypotheses.

The relationship between Bayesianism and learning theory is not so clear in the computable case. We may think of the Bayesian account of induction as being based on two parts, a *method* (conditionalization) and a *standard* (a unit probability of convergence to an empirically adequate hypothesis).

The relationship between Bayesianism and learning theory is not so clear in the computable case. We may think of the Bayesian account of induction as being based on two parts, a *method* (conditionalization) and a *standard of success* (a unit probability of convergence to an empirically adequate hypothesis).

Concerning the Bayesian standard of success, an interesting question is whether the Halmos result can be strengthened so that each μ-measurable set can be detected with probability one by a *computable* scientist. It is easy to see that this is not the case for a wide class of measures, however.

Proposition 6: $\exists$ countably additive probability measure μ over $\mathfrak{B}_\sigma$, $\exists$ μ-measurable $P \subseteq \omega^\omega$ s.t. $\forall$ μ-measurable $S \subseteq \omega^\omega$, if $\mu(S) = 0$ then $P \notin \Delta_2^{0, \omega^\omega - S}$.

Proof: Once again, consider $P = \{t: W_{t_1}$ is infinite$\}$. For each $i \in \omega$, let $i0$ denote the infinite sequence beginning with i and having 0 in each successive position. Let $P' = \{i0: i \in \omega\}$. Let μ be the unique, countably additive measure s.t. $\forall i \in \omega$, $\mu(\{i0\}) = 2^{-i}$. Then $\omega^\omega - P'$ is the largest event of measure 0 under μ. Now suppose that ϕ is a total recursive scientist who detects P over P'. ϕ can be turned into a total recursive ψ that detects P over all of ω^ω as follows: define $\psi(\sigma) = \phi(\tau)$ where τ is the finite sequence that begins with the first item i occurring in σ and that adds length(σ)-1 0's onto i. Hence, $P \in \Delta_2^0$. But this is absurd, by the argument given in the bounded rationality example in Section 4 above. ■

This negative result raises a new and more interesting question. Can we characterize the properties and measures that give rise to effectively solvable detection problems?

Concerning Bayesian method, we have the further question whether the performance of each computable scientist can be matched in power by a computable Bayesian conditionalizer armed with some countably additive prior measure. This question is somewhat finicky to set up, as one must define in a suitable way what it is for a conditionalizer to be effective.

7. Conclusion

In this paper, characterizations were established for semi-detectability, for detectability, and for identifiability, both for effective and for ineffective agents. These characterizations yield special case results for language learning, function identification and logical hypothesis detection. This permits us to factor the essential features of inductive inference from the accidental details of different inductive paradigms. The

results also apply over arbitrary collections of data presentations. The characterizations do not assume enumerability or effective enumerability of possible worlds. Finally, they illustrate an exact relationship between the statistical and the learning-theoretic convergence theorems on the one hand, and between learning theory and standard topology and recursion theory on the other.

Acknowledgments

This research was supported, in part, by a generous grant from the Buhl foundation. I would like to thank Cory Juhl and Teddy Seidenfeld for comments on early drafts. Special thanks are due to Clark Glymour, for many useful comments and suggestions.

References

[1] Angluin, D., "Inductive Inference of Formal Languages from Positive Data", *Information and Control* 45 (1980), pp. 117-135.

[2] Angluin, D. and C. Smith, "A Survey of Inductive Inference Methods", Technical Report 250, Yale University, October (1982).

[3] Blum, M. and L Blum, "Toward a Mathematical Theory of Inductive Inference", *Information and Control,* 28 (1975).

[4] Chang, C. C. and H. J. Keisler, Model Theory, New York: North Holland (1973).

[5] Gold, E. M., "Language Identification in the Limit", *Information and Control* 10 (1967) 447-474.

[6] Gold, E. M., "Limiting Recursion", *Journal of Symbolic Logic,* 30: 1 (1965). pp. 27- 48.

[7] Halmos, P. R., *Measure Theory,* New York: Springer (1974).

[8] Hinman, P. G., *Recursion-Theoretic Hierarchies,* New York: Springer (1978).

[9] Kant, I., *Prolegomena to any Future Metaphysics,* L Beck, Trans., Indianapolis: Bobbs-Merrill (1950).

[10] Kelly, K. "Induction from the General to the More General", *Proceedings of the Second Annual Workshop on Computational Learning Theory,* R. Rivest, D. Haussler, and M. Warmuth, San Mateo: Morgan-Kaufmann (1989).

[11] Kelly, K. "Theory Discovery and the Hypothesis Language", *Prceedings of the Fifth International Conference on Machine Learning,* San Mateo: Morgan Kaufmann, (1988).

[12] Kelly, K. and C. Glymour, "Convergence to the Truth and Nothing but the Truth", *Philosophy of Science* 56:2 (1989).

[13] Kelly, K. and C. Glymour, "Inductive Inference from Theory-Laden Data", Tech report CMU-LCL-89-5, Laboratory for Computational Linguistics, Carnegie Mellon University (1989).

[14] Kelly, K. and C. Glymour, "Theory Discovery from Data with Mixed Quantifiers", *Journal of Philosophical Logic,* Forthcoming.

[15] Kugel, P., "Induction, Pure and Simple", *Information and Control,* 33 (1977), pp. 276-336.

[16] Levy, A., *Basic Set Theory,* New York: Springer (1979).

[17] Osherson, D., and S. Weinstein, "A Universal Inductive Inference Machine", Unpublished manuscript, (1989).

[18] Osherson, D., and S. Weinstein, "Identification in the Limit of First Order Structures", *Journal of Philosophical Logic,* 15 (1986) pp. 55-81

[19] Osherson, D., and S. Weinstein, "Paradigms of Truth Detection", *Journal of Philosophical Logic,* 18 (1989), pp. 1-42.

[20] Osherson, D., M. Stob, and S. Weinstein, *Systems that Learn,* Cambridge: MIT Press (1986).

[21] Putnam, H. "'Degree of Confirmation[1] and Inductive Logic", in *The Philosophy of Rudolph Carnap,* A. Schilpp (ed), Lasalle, Illinois: Open Court (1963).

[22] Putnam, H. "Trial and Error Predicates and a Solution to a Problem of Mostowski". *Journal of Symbolic Logic,* 30: 1 (1965). pp. 49-57.

[23] Rogers, H., *The Theory of Recursive Functions and Effective Computability,* Cambridge: MIT Press (1987).

[24] Shapiro, E. "Inductive Inference of Theories from Facts", Report YLU 192, Department of Computer Science, Yale University (1981).