Logique & Analyse **127-128** (1989), 285-295

HIERARCHICAL INDUCTIVE INFERENCE METHODS

Moshe KOPPEL

A fundamental question in the theory of inductive inference is this: Given an "observed" finite binary string, how should we predict its continuation?

In [1], Blum and Blum offer a definition of an inductive inference method using a hierarchy of hypotheses each of which is represented as an infinite binary string. According to this definition, an inference method consists of selecting as a hypothesis the first string in the hierarchy which has the observed string as an initial segment. The Blums' work has its philosophical roots in the ideas of Popper [6] but refines those ideas by imposing on the hierarchies considerations of computability. In the present article we consider those "consistent" hierarchies which satisfy certain elementary Bayesian considerations. It is shown that such considerations force a definite structure on inference hierarchies which allows us to better define such notions as "confirmation" and "complexity" of a world of events.

The Popper-Blum and Blum approach used here doesn't establish the probability of a given event but rather offers a (perhaps arbitrarily) preferred hypothesis for explaining observed events and hence predicting future events. Other approaches, notably that of Carnap in [2], differ from this one in that they don't consider the sequence of observations but rather the collection of observations. Hence they offer not explanatory hypotheses but rather probabilities.

Carnap describes criteria for predicting whether or not some predicate will hold in some trial. This "empirical" criterion is the frequency with which the predicate has held in previous trials. The "logical" criterion is the weight of the predicate (that is, its probability assuming that its constituent atomic predicates are independent with probability 1/2).

Our modification of the Popper-Blum and Blum scheme allows the development of a new criterion for determining the probability of an event relative to an inference method, namely, the extent to which the occurrence of that event conforms to the predictions of the inference method. In fact in [4] we show that, regardless of choice of inference method, under appropriate conditions this criterion alone allows for the satisfaction of both of Carnap's criteria as a consequence.

We outline the key results on hierarchical inference methods presented in this paper. First, given some binary string of length n including exactly m 1's, the ratio of consistent inference methods which predict a 1 in the next trial is m/n — this result is suggestive, albeit naively, of the "straight rule".

A precise definition is given of the predictability (or well-behavedness) of some world vis-a-vis a particular inference method. We show that for any world there is an optimal method (in the sense that no method would have had greater success than the optimal one at predicting that world) and that moreover there is a consistent optimal method. We show further that for any consistent inference method there is a worst-behaved (or unpredictable) world. We show that in such world there is a bound or the difference between the actual frequency with which some predicate holds and the weight (i.e., expected frequency) of that predicate.

Finally, in the conclusion we introduce our definition of probability.

II. *Inference Methods*

Definition. An *inference method* is a mapping from the set of all binary strings into the set of infinite binary strings s.t.

a) the image $E(S)$ of a finite string S has S as an initial segment.
b) if some infinite string is the image of its initial segment of length l, then it is also the image of all of its initial segments of length greater than l.
c) if S is infinite then $E(S) = S$.

Property b is known as "tenacity" and simply corresponds to the notion than if after l experiments we have been prompted to make a particular prediction, then we don't change that prediction so long as future events have confirmed, rather than contradicted, it.

Note that any method at all for predicting (including those not considered here) must isolate certain features of the world as relevant and be invariant over others. Here we assume that the relevant facts for prediction are the properties of the observed strings and not the predicates of which they are the extensions.

Such an inference method can be represented quite clearly with the use of the following figure which we call a *hierarchy*.

The first two infinite strings (marked Level 1) are the images of the one-

Figure 1

1 0 1 0 1	
0 0 0 1 0	Level 1
1 1 1 0 0	
0 1 1 0 0	Level 2
0 1 0 0 0	
0 0 1 0 1	
1 1 0 1 0	Level 3
1 0 0 1 0	
....	
....	
....	

bit strings 1 and 0, respectively. Consequently, by tenacity, they are also the images of all their initial segments. Level 2 consists of the images of the 2-bit strings 10 and 01. The other 2-bit strings 10 and 00 have already been accounted for on Level 1. Level 3 consists of the images of all 3-bit strings not accounted for on the first two levels. We continue constructing the chart ad infinitum.

Observe that if only n moments have transpired in our universe — that is, the strings which we wish to extend are of length n — then we need make use only of the first n levels of our hierarchy. We call this part of the hierarchy an *inference n-set*.

Now let S be a finite or infinite binary string. Let S^T be a truncation of S. Let $l(S)$ be the length of S. Let $(S)_j$ be the j^{th} digit of S if $j \leq l(S)$ and in general let $(S)_j$ be the j^{th} digit of $E(S)$. Finally, let $L(S)$ be the level of S — that is, the level on which $E(S)$ appears. (If S is infinite and doesn't appear in the hierarchy, then $L(S) = \infty$.)

A number of facts about inference hierarchies are obvious. There are exactly 2^{k-1} strings on level k of the hierarchy, except for $k = 1$ which has two strings (one beginning with 1, the other with 0). Each S such that $l(S) = k$, appears (as an initial segment) exactly once in the first k levels and exactly 2^{n-k} times in the first n levels for $n > k$. Also, $L(S^T) \leq L(S)$.

Translating all this into an inference method simply entails predicting that some phenomenon represented by the finite binary string S will continue in accordance with $E(S)$. If subsequent evidence eliminates $E(S)$ as a candidate then we move up the hierarchy until a suitable string is found, etc. In light of this, a very natural measure of the *confirmation* of a hypothesis $E(S)$ can be given — namely, $C(S) = l(S) - L(S)$. This can be intuited in two ways. First, $C(S)$ represents the number of levels up the hierarchy which we have to move if we reject the optimal choice $E(S)$ in favor of the next best choice. Second, the first $L(S)$ terms can be thought of as *establishing* the hypothesis $E(S)$ (since $E(S)$ is not conjectured until $L(S)$ terms have appeared), thus leaving the next $l(S) - L(S)$ terms to confirm it.

III. *Worlds and Consistency*

Let $P_1, \ldots, P_k$ be a set of logically independent projectible predicates. A universe $U = \{P_1, \ldots, P_k\}$ is the set of Boolean combinations of the "generating" predicates $P_1, \ldots, P_k$. Let P'_i represent either the predicate P_i or the predicate $\sim P_i$. $E_1, \ldots, E_{2^k}$ is the set of elementary predicates of the form $\bigwedge_{i=1}^{k} P'_i$. Then every predicate in U is a disjunction of elementary predicates. If a predicate is the disjunction of m elementary predicates we call $m/2^k$ the Bernoulli distribution of P (written $B(P)$).

For a predicate P let S_P be a binary string such that $(S_P)_i = 1$ if and only if the predicate P holds at moment i. Let a *world* W of length l of the universe $U = \{P_1, \ldots, P_k\}$ be the set of 2^{2^k} strings of length l, $S_{B_i(P_1, \ldots, P_k)}$, where B_i ranges over all Boolean functions. For any given universe there are 2^{k-l} possible worlds of length l, since during each of l moments exactly one of the 2^k elementary predicates holds.

Observe that for any world of length l,

$(S_{B(P_1, \ldots, P_k)})_i = B((S_{P_1})_i, \ldots, (S_{P_k})_i)$ for each $i \leq l$.

We need to ensure that the world predicted by some inference method also satisfies this equality for each $i > l$.

Definition. An inference method is *consistent* if for every Boolean function B and any two strings S_1 and S_2 of equal length $E(B(S_1, S_2)) = B(E(S_1), E(S_2))$.

Similarly we say an inference n-method is consistent if for every Boolean function B and any two strings S_1 and S_2 *of length n*, $E(B(S_1, S_2)) = B(E(S_1), E(S_2))$.

Repetition Theorem: *An inference n-set* N *is consistent if and only if for each* j *there exists an* i $\leq$ n *such that for every string* S *in* N, $(S)_j = (S)_i$.

Proof: In this context $S \cap T$ refers to the meet of S and T, $S \cup T$ refers to the join of S and T and $\sim S$ is the complement of S.

Consider the string 1^n. For any $j > n$ and any S of length n, $(1^n)_j = (S \cup \sim S)_j = (S)_j \cup (\sim S)_j = (S)_j \cup \sim((S)_j) = 1$. Thus, $E(1^n) = 1^\omega$ for all n and all consistent E.

Let D_i be the string of length n each bit of which is 0 except for the i^{th}

bit which is 1. Let B_1^n be a Boolean function on n strings such that the i^{th} bit of B_1^n $(S_1, \ldots, S_n)$ is 1 if and only if the i^{th} bit of *exactly* one of the strings $S_1, \ldots, S_n$ is 1.

For every $j > n$, B_1^n $((D_1)_j, (D_2)_j, \ldots, (D_n)_j) = (B_1^n (D_1, D_2, \ldots, D_n))_j = (1^n)_j = 1$. Thus for every $j > n$ there is exactly one $i_j \leq n$ such that $(D_{i_j})_j = 1$. We claim that for any S of length n, $(S)_j = 1$ if and only if $(S)_{i_j} = 1$. If $(S)_{i_j} = 1$ then for some string T, $(S)_j = (D_{i_j} \cup T)_j = (D_{i_j})_j \cup (T)_j = 1$. If $(S)_{i_j} = 0$ then $(\sim S)_{i_j} = 1$ and by the above $(\sim S)_j = 1$ and thus $(S)_j = 0$. This completes the proof that if an inference n-set is consistent then for each j there exists an $i \leq n$ (namely i_j) such that for every S of length n $(S)_j = (S)_i$.

Figure 2

```
0 0 0 0 0 0 0 ....
1 1 1 1 1 1 1 ....
------------------
0 1 0 1 0 1 0 ....
1 0 1 0 1 0 1 ....
------------------
0 0 1 0 1 0 0 ....
1 1 0 1 0 1 1 ....
0 1 1 1 1 1 0 ....
1 0 0 0 0 0 1 ....
------------------
0 0 0 1 0 1 0 ....
1 1 1 0 1 0 1 ....
0 1 0 0 0 0 0 ....
1 0 1 1 1 1 1 ....
0 0 1 1 1 1 0 ....
1 1 0 0 0 0 1 ....
0 1 1 0 1 0 0 ....
1 0 0 1 0 1 1 ....
```

To prove the converse, observe that for any Boolean function B and strings of length n, S_1 and S_2 we have $B((S_1)_{i_j}, (S_2)_{i_j}) = B((S_1)_{i_j}, (S_2)_{i_j})$ by hypothesis. But since $i_j \leq n$ it follows that $B((S_1)_{i_j}, (S_2)_{i_j}) = (B(S_1, S_2))_{i_j} = (B(S_1, S_2))_{i_j}$. QED

Note that a consistent inference method is an inference method which consists solely of consistent inference n-sets. Figure 2 contains an example of a consistent hierarchy.

Since for each $1 \leq n \leq 4$ the inference n-set here is consistent we have that the second column of Level 1 is a copy of the first, the third column of Levels 1 and 2 is a copy of one of the previous two, etc. It follows that 0^ω and 1^ω always appear on Level 1.

By the Repetition theorem we have that for any consistent inference n-set and every string of length n, there exist some $i = i(n) \leq n$ such that $(S)_{n+1} = (S)_i$. Moreover, for any consistent inference method H, the set $\bigcup_{n=1}^{\infty}\{i(n)\}$ completely determines H. Thus we have

Theorem. There is a one-to-one correspondence between consistent inference methods and functions $q{:}N \rightarrow N$ such that $q(n) \leq n$.

We now show that the ratio of consistent inference methods which predicate that a predicate will hold in some future trial is equal to the frequency with which that predicate has held in previous trials.
Let $S_1, S_2, \ldots, S_{2^n}$ be some ordering of the set of binary strings of length n.

Definition. If N is an inference n-set the p^{th} *projection* of N is the set of strings $\cup_{i=1}^{2n}\{T_i \mid T_i = S_i * (S_i)_p\}$ where * denotes concatenation.

Definition. A p^{th} projection is *consistent* if it is the p^{th} projection of a consistent inference n-set.

Lemma. Any consistent p^{th} projection is also a consistent j^{th} projection for any j.

As a result of the above Lemma which is an immediate consequence of the Repetition Theorem we can speak simply of consistent projections without specifying the number of the projection.

Definition. If S_i is a string of length n, the desirability of the prediction $(S_i)_p = 1$ (written $D((S_i)_p = 1)$) is equal to the proportion of all consistent projections of inference n-sets for which $(T_i)_{n+1} = 1$ (i.e., which contain $S_i * 1$).

Straight Rule Theorem. If S *is a string of length* n *containing* m 1*'s then for every* p, $D((S)_p = 1) = m/n$. *In particular, for any* n, $D((1^n)_p = 1) = 1$.

This follows from the fact that — from the Repetition Theorem — there is one consistent projection corresponding to each choice of $1 \leq i_j \leq n$ and for exactly m of these i_j, $(S)_{i_j} = 1$.

We now define a different type of prediction desirability. If H is an inference hierarchy and R is some set of finite strings, the desirability modulo H of the prediction $(R)_p = 1$ (written $D_H((R)_p = 1)$) is equal to the proportion of strings S in R such that $(S)_p = 1$ according to H. Roughly speaking, if we are using some fixed consistent hierarchy H and we don't know what string S is, though we do know $S \in R$, then $D_H((R)_p = 1)$ is the desirability of the prediction that $(S)_p = 1$.

Second Straight Rule Theorem. If R *consists of all strings of length* n *containing* m 1*'s, then for any consistent* *inference hierarchy* H *and any* p, *we have* $D_H((R)_p = 1) = m/n$.

This follows from the fact that of the $\binom{n}{m}$ strings in R exactly $\binom{n-1}{m-1}$ have 1 as the p^{th} bit.

IV. *Complexity*

In this section we offer a natural definition of the relative complexity (opposite: predictability) of a world vis-a-vis to a given inference method and, conversely, of the relative effectiveness of an inference method at predicting a given world.

Consider some $lx2^{2^k}$ world W and some inference method H. Let a_i denote the number of strings in W which H correctly predicts after seeing their first i bits — that is, which appear on Level i or below in H. Thus, for example $a_l = 2^{2^k}$. Let $V_W^H = \{a_1, \ldots, a_2\}$.

Definition. Given two inference methods H_1 and H_2 and an $lx2^{2^k}$ world W such that $V_W^{H_1} = (a_1, \ldots, a_l)$ and $V_W^{H_2} = (a_1', \ldots, a_1')$ we say that H_1 is *at least as effective* as H_2 for W (written $V_W^{H_1} > V^{H_2}{}_W$) if $a_i \geq a_i'$ for all $1 \leq i \leq l$.

Analogously, we have

Definition. Given two $lx2^{2^k}$ worlds W_1 and W_2 and an inference method H such that $V_{W_1}^H = (a_1, \ldots, a_l)$ and $V_{W_2}^H$ $(a_1', \ldots, a_1')$ we say that W_1 is *at least as predictable* (not more predictable than) W_2 vis-a-vis H if $a_i \geq a_i'$ for all $i \leq i \leq l$.

Observe that predictability is a property of worlds, not strings. Thus it captures the idea that although a particular string may not exhibit projectible patterns, if it is demonstrably interdependent with some other string in the world, the world is rendered more predictable.

We now give a complete characterization of all possible vectors V_W^H both in the instance where H ranges over all inference methods and where it ranges only over consistent methods.

Let W be a $2^{2^k}xl$ world. Recall that during any moment t_i exactly one elementary predicate $E_{h(i)}$ holds in W. If in W, $h(i) = h(j)$ we call the moments M_i and M_j *partners*. If $h(i) \neq h(j)$ for all $j < i$, we call M_i a *free moment*. Let $c(i)$ denote the number of different elementary predicates which hold in the first i moments of W (equivalently: the number of free moments among the first i moments of W). Let α be the number of moments succeeding the last non-free moment of W and let $t_i = a_i/2^{2^k} - c(l)$.

Theorem 1. For any $lx2^{2^k}$ universe W, there exists an inference method H such that $V_W^H = (a_1, \ldots, a_l)$ if and only if
i) t_i is an integer for all $1 \leq i \leq l$
ii) $t_i \leq t_{i+1} \leq t_i + 2^{c(i)}$ for all $1 \leq i \leq l$
iii) $t_i \leq 2^{c(i)}$ for all $i < l - \alpha$; $t_i = 2^{c(i)}$ for all $i \geq l - \alpha$.

Theorem 2. For any $lx2^{2^k}$ universe W, there exists a consistent inference method H such that $V_W^H = (a_1, \ldots, a_l)$ if and only if
i) $t_1 = 2$
ii) $t_{i+1} = t_i$ or $t_{i+1} = 2t_i$ for all $1 \leq i \leq l$
iii) $t_i \leq 2^{c(i)}$ for all $i < l - \alpha$; $t_i = 2^{c(i)}$ for all $i \geq l - \alpha$.

V. *Optimal Methods and Unpredictable Worlds*

It might be objected that the restriction to consistent inference methods is a handicap — that for some world, some inconsistent inference method might nevertheless be more effective than any consistent one. We prove that this cannot occur.

For a world W, call an inference method H_0 optimal if $V_W^{H_0} \geq V_W^H$ for any (consistent or inconsistent) inference method H. Note that effectiveness relative to W is *not* a linear ordering on inference methods. Nevertheless we have

Theorem. For any world W, *there exists a consistent, optimal inference method* H.

Proof. From Theorem 2 of Section IV it follows that there exists a consistent inference method such that $t_i = 2^{c(i)}$ for all $i \leq l$ and from Theorem 1 it follows that this method is optimal.

We have thus far considered the effectiveness of different inference methods for some given world. We now consider the complexity (or well-behavedness) of different worlds with respect to some given inference method. If some $lx2^{2^k}$ world W_0 is such that $V_{W_0}^H \leq V_W^H$ for every $lx2^{2^k}$ world W, then we call W_0, "H-unpredictable".

Lemma. For any positive integers l, k there exists an $lx2^{2^k}$ world W which is H-unpredictable for every consistent inference method H.

Distinct Strings Lemma. If W is an $lx2^{2^k}$ world with $l \geq 2^k$ then there exists a consistent inference method H such that W is H-unpredictable if and only if the string 1^l appears exactly once in W.

Proof of Distinct Strings Lemma. If the string 1^l appears more than once in W then for any H, V_W^H will be such that $a_1 > 2$. But for any H there is an $lx2^{2^k}$ world W' such that $V_{W'}^H$ has $a_1 = 2$. Therefore W can't be H-unpredictable. Conversely, if 1^l appears only once in W then by Theorem 2 we can choose H such that V_W^H has $a_i = 2$ if $i \leq l - 2^k$ and $a_i = 2^{i+2^k-l}$ otherwise. QED

Density Theorem. If there exists a consistent inference method H *such that* W *is* H-*unpredictable then for any* S_p *in* W, $|$ D(S_p) - B(P) $|$ < c·B(P).

Surprisingly, the notion "unpredictable" as defined here is closely related to the common notion of "random" in that the frequency with which a predicate in an unpredictable world holds, is close to its Bernoulli distribution.

Proof of Density Theorem. Let W be some unpredictable world generated by the strings $S_1, \ldots, S_k$. Let S'_i denote either S_i or $\sim S_i$. We first prove that if $S = S'_{i_1} \cap \ldots \cap S'_{i_m}$ then $D(S) \geq (1 - c)/2^m$. Suppose that $D(S) < (1 - c)/2^m$. If p is the number of 1's in S then $p = D(S) \cdot l \leq [(1 - c)/2^m] \cdot l = 2^{k-m}$. Let $S_{i_{m+1}}, \ldots, S_{i_k}$ be the generators of W which don't appear in the definition of S. There are precisely 2^{k-m} different strings of the form $S'_{i_{m+1}} \cap \ldots \cap S'_{i_k}$ and since $p < 2^{k-m}$ one of these strings, say T, must be 0 during each moment for which S is 1. But then the string $\sim(S \cap T) = 1^l$ contradicting the Distinct Strings Lemma. To prove the theorem for arbitrary $S = f(S_1, \ldots, S_k)$ note that S can be written in disjunctive normal form with disjoint disjuncts $R_1, \ldots, R_j$. Then $D(S) = \Sigma D(R_i) \geq \Sigma(1 - c)\, B(R_i) = (1 - c)\, B(S)$. Finally we derive the inequality $D(S) \leq (1 + c)\, B(S)$ by replacing S with $\sim S$ in the first inequality.

VIII. *Conclusions*

We have been careful to use the term "desirability of prediction" rather than "probability". It is certainly not the case that if some predicate has been true in every previous trial that the probability is 1 that it will be true in the next trial. Nevertheless, if we had to make some prediction, that is certainly what we'd predict.

What we have said, in short, is that if it is the case that "x percent of all observed A's have been B's" then x percent of all consistent inference methods predict that given some future A, it will be a B. While one needn't *a priori* give equal weight to all consistent inference methods, we have suggested that it is more fruitful to apply a principle of indifference to possible inference methods than to, say, possible trial results. Hence the "straight rule".

We have shown that the consistency condition does not adversely effect the potential for predicting accurately in general, and that the resultant notion of unpredictability corresponds in essential ways to the intuitive notion of randomness.

Despite the obvious limitations of this method for purposes of prediction, it can be shown that these ideas can be used as a basis for a new definition of probability. Let W_1 and W_2 be two possible continuations of the world W. (That is, $l(W_1) = l(W_2) > l(W)$ but W_1, W_2 and W are identical for the first $l(W)$ moments.) Then we say that (with respect to H) W_1 is a more probable continuation of W than W_2, if $V^H_{W_1} > V^H_{W_2}$. That is the probability of some continuation of a world is the extent to which it conforms to the expectations of some given inference method. It is shown in [4] that many of the properties of such a probability are invariant over the choice of consistent H and that these properties are precisely those we would want not only in a theory of probability but in a theory of confirmation as well.

The treatment in this paper leaves wide open the question of choice of hierarchy. Ultimately if the approach outlined here is to be useful, some additional criteria will have to be employed for selecting appropriate hierarchies. The computability considerations of Blum and Blum, for example, are certainly not precluded by our consistency constraint. Also, the complexity considerations of Chaitin [3] especially as exploited by Martin-Löf [5], are appropriate in this framework.

Bar-Ilan University

REFERENCES

[1] Blum, L. and Blum, M. Toward a Mathematical Theory of Inductive Inference, Information and Control 28, 125-155 (1975).

[2] Carnap, R. *The Continuum of Inductive Methods*, Univ. of Chicago Press, Chicago (1950).

[3] Chaitin, G. Algorithmic Information Theory, IBM J. Res. and Development 21, 350-359 (1977).

[4] Koppel, M. Towards a New Theory of Confirmation, Israel J. Math. 50, 207-218 (1985).

[5] Martin-Löf, P. The Definition of Random Sequences, Information and Control 9, 602-619 (1966).

[6] Popper, K. *The Logic of Scientific Discovery*, (2nd ed.), Harper Torchbooks, New York (1968).