ISRAEL JOURNAL OF MATHEMATICS, Vol. 50, No. 3, 1985

TOWARDS A NEW THEORY OF CONFIRMATION

BY

MOSHE KOPPEL

Department of Mathematics, Bar-Ilan University, Ramat Gan, Israel

ABSTRACT

Any sequence of events can be "explained" by any of an infinite number of hypotheses. Popper describes the "logic of discovery" as a process of choosing from a hierarchy of hypotheses the first hypothesis which is not at variance with the observed facts. Blum and Blum formalized these hierarchies of hypotheses as hierarchies of infinite binary sequences and imposed on them certain decidability conditions. In this paper we also consider hierarchies of infinite binary sequences but we impose only the most elementary Bayesian considerations. We use the structure of such hierarchies to define "confirmation". We then suggest a definition of probability based on the amount of confirmation a particular hypothesis (i.e. pattern) has received. We show that hypothesis confirmation alone is a sound basis for determining probabilities and in particular that Carnap's logical and empirical criteria for determining probabilities are consequences of the confirmation criterion in appropriate limiting cases.

You are tossing a rather mysterious coin. In the 1000 tosses you've already observed, the result has alternated between heads and tails — the odd-numbered tosses have all been heads, the even-numbered tosses have all been tails. Bets are being taken on the next toss.

You recall the criteria suggested by Carnap [2] for determining the probability of an event, say heads. The "empirical" criterion — the frequency of the event in past trials — yields a probability of $\frac{1}{2}$ for heads. The "logical" criterion — the proportion of possible states of the world for which the event holds — also yields a probability of $\frac{1}{2}$ for heads (assuming that "coming up heads" is an atomic predicate). Thus the entire "continuum of inductive methods" points to one claim — the probability of the next toss coming up heads is $\frac{1}{2}$.

Nevertheless, mindful of the coin's peculiarly obedient behavior, you put all your money on heads.

Received June 9, 1983 and in revised form November 1, 1984

Can we construct an inference method which would justify such a bet? This paper attempts just that. We take as primitive an ordering of the hypotheses explaining some observed phenomenon. We then define probability as a function of this ordering — an event is probable to the extent that the hypotheses which predict it are preferred. Thus in the example above if we choose an ordering of hypotheses which favors alternation, the probability of the next toss coming up heads is large. Moreover, we will show that regardless of the choice of ordering, Carnap's criteria are consequences of our definition for appropriate limiting cases.

Definitions

Any sequence of events, such as the results of the coin tosses described above, can be "explained" by any of an infinite number of hypotheses. Popper [5] describes the "logic of discovery" as a process of choosing from a hierarchy of hypotheses the first hypothesis which is not at variance with the observed facts. Blum and Blum [1] formalized these hierarchies of hypotheses as hierarchies of infinite binary sequences and imposed on them certain decidability conditions. In this paper we also consider hierarchies of infinite binary sequences but we impose only the most elementary Bayesian considerations. The properties of these hierarchies are given in detail in [3]. Here we review the basic definitions in order to introduce a new conception of confirmation and probability based on them.

Definition. An inference method is a mapping from the set of all finite binary strings of length $\geqq 1$ into the set of infinite binary strings such that:

(a) the image $H(S)$ of a string S has S as an initial segment,

(b) if some infinite string, α, is the image of its initial segment of length l, then α is the image of its initial segments of length greater than l.

Thus, H is a hierarchy of hypotheses in which preferred hypotheses are those which are images of shorter strings.

For a binary string S let $l(S)$ be the length of S, let $(S)_j$ be the jth bit of S and let $[S]_j$ be the initial segment of S of length j. If in some context the method H is given, $(S)_j$ and $[S]_j$ will be defined for $j > l(S)$ as $(H(S))_j$ and $[H(S)]_j$, respectively, and $L(S)$ (read: level of S) will be the smallest value of j such that $H([S]_j) = H(S)$. Thus $1 \leqq L(S) \leqq l(S)$.

$L(S)$ is the number of bits of S which are needed before a hypothesis is chosen which is not subsequently contradicted by the actual behavior of S. Given

S, the amount of confirmation had by a hypothesis $H(S)$ is defined to be $l(S)-L(S)$.

Let a universe of predicates, $U=[P_1,\ldots,P_k]$ consist of a set of logically independent predicates $P_1,\ldots,P_k$ — called *atomic predicates* — and all Boolean combinations of $P_1,\ldots,P_k$. If P' denotes either P or $-P$ then we call the conjunctions $P'_1\wedge\cdots\wedge P'_k$ the *elementary predicates* of U. Let S_P be a finite binary string representing as a characteristic function the extension of the predicate P. A world W manifesting U consists of a set of equally long binary strings $\{S_i\}$ where the i range over the predicates in U.

We think of the jth bits of all the strings in W as being simultaneous. Thus, if U contains k atomic predicates and the strings in W are of length l, we can visualize W as an $l\times 2^{2^k}$ matrix. Clearly the extensions of the atomic predicates determine the whole world — in particular if W is an $l\times 2^{2^k}$ world and B is a Boolean function then for every $i\leqq l$ we have

$$(*)\qquad (S_{B(P_1,\ldots,P_k)})_i = B((S_{P_1})_i,(S_{P_2})_i,\ldots,(S_{P_k})_i).$$

During each moment exactly one elementary predicate holds.

In this paper, we define probability as a function not of events but of worlds. The inference method H will operate on strings within the context of a world. $H(S_P)$ will depend on the string S_P but not on P itself — that is, the images of identical strings are identical even if they are the extensions of different Boolean functions of atomic predicates. We will be interested in those inference methods which predict worlds which satisfy $(*)$ for all $i>0$. Formally, we have

DEFINITION. An inference method H is *consistent* if for two strings S_1 and S_2 of equal length and any Boolean operation B,

$$H(B(S_1,S_2))_i = B((H(S_1))_i,(H(S_2))_i)\qquad \text{for all } i.$$

One immediate consequence of the consistency of H is that for any $n\geqq 1$, $H(1^n)=111\cdots1\cdots$. This is unsurprising since H is used to "predict" the behaviour of S_T where T is a tautology.

In this paper all inference methods referred to are consistent. The properties of consistent inference methods are discussed in detail in [3].

Since we are measuring the probabilities of worlds we need to represent the amount of confirmation received by a world. With each world W of length l, we associate a vector $V^H_W=(a_1,\ldots,a_l)$ where a_i is the number of strings, S, in W such that $L(S)\leqq i$. Then $2\leqq a_1\leqq a_2\leqq\cdots\leqq a_l=2^{2^k}$. The first inequality follows from the fact that every world contains $S_T=1^l$ and $S_F=0^l$ where T and F are tautology and contradiction and the fact that for any consistent inference method $L(1^n) = L(0^n) = 1$.

Axioms

In this section we give four axioms which constitute the definition of probability.

(a) Roughly speaking, a world is probable, relative to a given inference method, to the extent that the inference method predicts its occurrence. The extent to which method H predicts world W is expressed by V^H_W. If $V^H_W = (a_1, \dots, a_l)$ and $V^H_{W'} = (b_1, \dots, b_l)$ and $a_i \geqq b_i$ for all $1 \leqq i \leqq l$ then W is more predictable by H than W' and we write $V^H_W \geqq V^H_{W'}$.

Axiom I. $\mathcal{P}_H$ is a function mapping worlds into the real line $(0, 1)$ such that if W and W' are both $l \times 2^{2^k}$ worlds, then

$$V^H_W = V^H_{W'} \Rightarrow \mathcal{P}_H(W) = \mathcal{P}_H(W') \qquad \text{and} \qquad V^H_W > V^H_{W'} \Rightarrow \mathcal{P}_H(W) > \mathcal{P}_H(W').$$

For the rest of this paper, we will drop the subscript H unless we want to emphasize the dependence of $\mathcal{P}$ on H.

(b) If U is a universe generated by k atomic predicates then there are $2^{k \cdot l}$ worlds of length l which could possibly manifest U. If we list these as $W_1, \dots, W_{2^{k \cdot l}}$ we certainly want $\sum_{i=1}^{2^{k \cdot l}} \mathcal{P}(W_i) = 1$. In fact, we want more than this. If W is a world of length l manifesting U, let $W \cdot E_i$ be the world of length $l + 1$ such that W manifests it for the first l moments and the elementary predicate E_i holds during the moment $l + 1$. Then we have

Axiom II. If $E_1, \dots, E_{2^k}$ are the elementary predicates of U then for all W

$$\sum_{i=1}^{2^k} \mathcal{P}(W \cdot E_i) = \mathcal{P}(W) \qquad \text{and} \qquad \sum_{i=1}^{2^k} \mathcal{P}(E_i) = 1$$

where $P(E_i)$ = probability of world of length 1 in which E_i holds.

We define the conditional probability of $W \cdot E_i$ relative to W as $\mathcal{P}(W \cdot E_i)/\mathcal{P}(W)$. The probability of an event $(S_P)_l = \theta$ (where $\theta = 0$ or 1) is equal to the sum of the probabilities of all worlds of length l in which $(S_P)_l = \theta$ holds. If the predicate $P = \bigvee_{i=1}^{j} E_i$, the conditional probability of an event $(S_P)_{l+1} = \theta$ relative to a world W of length l is equal to $\sum_{i=1}^{j} \mathcal{P}(W \cdot E_i)/\mathcal{P}(W)$.

(c) Consider the universe $U = [P_1, \dots, P_k]$. Let $U_i = [P_i]$. If W is a world manifesting U, let W_i be the restriction of W to the universe U_i. It is tempting to suggest that $\mathscr{P}(W) = \mathscr{P}(W_1) \cdot \mathscr{P}(W_2) \cdot \cdots \cdot \mathscr{P}(W_k)$. Upon further reflection, however, it becomes obvious that this is impossible because $\mathscr{P}(W)$ depends not only on S_{P_i} for $i = 1, \ldots, k$ but equally on $S_{B(P_1, P_2, \ldots, P_k)}$ where B is a Boolean operation. That is, although P_i and P_j are logically independent for $i \neq j$, they may be seen to be empirically dependent in W and thus the probabilities of their occurrences should be accordingly dependent.

For example, suppose $U = [P_1, P_2]$ and $S_{P_1} = S_{P_2} = 10010$. Assume that $L(10010) = 5$ so that $\mathscr{P}(W_1) = \mathscr{P}(W_2)$ is relatively low. To be exact $V^H_{W_1} = V^H_{W_2} = (2, 2, 2, 2, 4)$ which is the worst possible vector for a 5×4 world. Yet W itself includes no less than eight strings on level 1, namely the extensions of all those predicates implied by the predicate $P_1 \equiv P_2$ and their negations. Thus $V^H_W = (8, 8, 8, 8, 16)$ which dominates many possible vectors for 5×16 worlds. In fact, let W' be a different manifestation of U in which $S'_{P_1} = 10010$ but $S'_{P_2} = 11111$. Then $V^H_{W'_1} = (2, 2, 2, 2, 4)$ but $V^H_{W'_2} = (4, 4, 4, 4, 4)$, while $V^H_{W'} = (8, 8, 8, 8, 16)$. By Axiom I, $\mathscr{P}(W) = \mathscr{P}(W')$ and $\mathscr{P}(W'_2) > \mathscr{P}(W_2)$. Thus if $\mathscr{P}(W) = \mathscr{P}(W_1)\mathscr{P}(W_2)$ we have

$$\mathscr{P}(W'_1) \cdot \mathscr{P}(W'_2) > \mathscr{P}(W_1) \cdot \mathscr{P}(W_2) = \mathscr{P}(W) = \mathscr{P}(W'),$$

i.e. $\mathscr{P}(W'_1) \cdot \mathscr{P}(W'_2) \neq \mathscr{P}(W')$.

To see this point in fuller generality consider the following. Let W be a world manifesting U, with S_{E_i} the extension of the elementary predicate E_i. Let W' be a world manifesting U, with S'_{E_i} the extension of the elementary predicate E_i, such that the set $\{S'_{E_i} : 1 \leqq i \leqq 2^k\}$ is identical with the set $\{S_{E_i} : i \leqq 1 \leqq 2^k\}$. Then we say W' is a permutation of W and have $V^H_W = V^H_{W'}$ for any H. Thus $\mathscr{P}(W) = \mathscr{P}(W')$— that is, probability is invariant over permutation of elementary predicates (compare Carnap [2], p. 14). But it can be shown that for any world W (except one with only constant strings) there is a permutation W' such that $\Pi^k_{i=1} \mathscr{P}(W_i) \neq \Pi^k_{i=1} \mathscr{P}(W'_i)$ as in the example above. We can capture the range of possible interdependence between atomic predicates by postulating the following inequality.

Axiom III. $$\min_{W'} \prod_i \mathscr{P}(W'_i) \leqq \mathscr{P}(W) \leqq \max_{W'} \prod_i \mathscr{P}(W'_i)$$

where min and max are taken over all permutations of W.

(d) Let W be a world manifesting a universe U with a single predicate P. Suppose $L(S_P) = l(S_P)$ — that is, the current best hypothesis $H(S_P)$ has received no confirmation. Noting that there are only two elementary predicates which can hold during moment $l + 1$, namely P and $-P$, and calling the hypothesis' prediction for $l+1$, E^{l+1}, we have by the first two axioms that the relative probability of $W \cdot E^{l+1}$ is greater than $\frac{1}{2}$. Nevertheless, if $l(S_P)$ (and hence $L(S_P)$) is very large then $H(S_P)$ is in any event not a particularly appealing hypothesis. In fact, observe that

$$V^H_{W \cdot -E^{l+1}} = (\overbrace{2,\dots,2}^{l},4) \qquad \text{while} \qquad V^H_{W \cdot E^{l+1}} = (\overbrace{2,\dots,2}^{l-1},4,4);$$

the vectors differ only for $a_{l(S_P)}$. In the limit (as l grows) the relative probability of the preferred event occurring during moment $l+1$ should tend down towards $\frac{1}{2}$.

Axiom IV. If W is an $l \times 4$ world manifesting $U = [P]$ such that $L(S_P) = l$ then

$$\frac{1}{2} < \frac{\mathcal{P}(W \cdot E^{l+1})}{P(W)} < \frac{1}{2} + \frac{f(l)}{2^{2^l}}$$

for some decreasing function f such that $\lim_{i\to\infty} f(i) = 0$.

The form $f(l)/2^{2^l}$, with $f(l) \to 0$, amounts to a requirement on the rate of convergence to $\frac{1}{2}$; it is the rate needed for the subsequent theorems.

If $\mathcal{P}$ is a function satisfying the first three axioms and satisfying Axiom IV for the function f then we say $\mathcal{P}$ is a probability of degree f.

Theorems

In this section we demonstrate that a probability which satisfies the four axioms presented above has many of the properties which one would hope for — regardless of the choice of inference method.

The kinds of general statements we will make about probabilities divide into two classes — those about small universes (generated by few atomic predicates) and those about large universes.

It is certainly to be expected that as a hypothesis receives increasing amounts of confirmation the probability that it will hold during the next moment approaches 1. Let $e_l = \mathcal{P}(W)$ where W is an $l \times 4$ world manifesting $U = [P]$ such that $L(S_P) = l$. Let $c_l = e_l - 2e_{l+1}$. Note that by Axioms I and II, $c_l > 0$. For any world let E^{l+1} be the elementary predicate predicted for moment $l+1$.

Confirmation Theorem. *If W is an $l \times 4$ world manifesting $U = [P]$ such that $L(S_P) = m$ then*

$$\frac{\mathcal{P}(W \cdot E^{l+1})}{\mathcal{P}(W)} > 1 - \frac{1}{2^l \cdot c_m}.$$

Thus if l gets large while m is fixed, i.e. some hypothesis on level m is confirmed *many* $(l \cdot m)$ *times, then the probability that it will hold during the next moment approaches* 1.

PROOF. Define $X_{i,m} = \mathcal{P}(W_{i,m})$ where $W_{i,m}$ is an $i \times 4$ world the non-trivial strings of which are on level m. Thus $e_m = X_{m,m}$. Note that by Axiom 1,

$$X_{i,m} = X_{i+1,m} + X_{i+1,i+1}.$$

Using this repeatedly, beginning from $X_{m,m}$ we obtain

$$\mathcal{P}(W) = X_{l,m} = e_m - \sum_{i=1}^{l} e_{m+1} \quad \text{and} \quad \mathcal{P}(W \cdot E^{l+1}) = X_{l+1,m} = e_m - \sum_{i=1}^{l+1} e_{m+1}.$$

Recall that $e_i < \frac{1}{2}e_{i-1}$. Thus,

$$\frac{\mathcal{P}(W \cdot E^{l+1})}{\mathcal{P}(W)} = 1 - \frac{e_{l+1}}{e_m - \sum_{i=1}^{l} e_{m+i}} > 1 - \frac{e_{l+1}}{e_m - 2e_{m+1}} > 1 - \frac{1}{2^l \cdot c_m}.$$

The counterpart to this theorem for unpredictable single-predicate universes is simply Axiom IV. It might be argued that while the theorem is intuitive, Axiom IV is not — in the sense that in focusing solely on the question of hypothesis confirmation it neglects such other factors as frequency. In fact, it is our contention that frequency is *not* an a priori determinant of probability, but rather a tool for calculating probabilities determined by hypothesis confirmation. This tool is useful under conditions to be outlined below — conditions not satisfied by small universes. Intuition about the relationship between frequency and probability is shaped by the fact that the range of human senses constitutes a large universe.

For example, suppose during 100 moments (trials) the extension of the predicate P is

$$\overbrace{111\cdots1}^{99}0.$$

The intuition that the P will hold during the next moment is to be reconstructed within my proposal as follows: There are many predicates independent of P, say $P_1, \ldots, P_n$, such that for all $1 \leqq i \leqq n$, P_i did not hold during moment 100. Thus although the hypothesis that "P always holds" has been discredited, we have that for each i the hypothesis "$-P_i \vee P$ always holds" (i.e. P_i implies P) has much confirmation since the string $S_{-P \vee P} = 1^{100}$ and hence $-P_i \vee P$ is likely to hold during the next moment. Since (again by considerations of confirmation) it is unlikely that none of the predicates P_i will hold during the next moment, it is likely that P will hold during the next moment.

We will formalize this argument shortly, yet it must already be clear that this reasoning can only work in a universe with sufficient predicates and that frquency should play no rule in a universe with a single predicate.

In a world containing "patterned" strings — that is, strings satisfying some low-level hypothesis — we've arranged for the probability of future events to be guided by considerations of entropy. But what of an "unpredictable" world — one in which strings are not patterned? We show that for such worlds, with appropriate dimensions, any probability satisfying the axioms will — in the limit — satisfy *both* of Carnap's criteria — the probability of a predicate holding approaches both the frequency of its past occurrence and its weight (where the weight of a predicate P which is the disjunction of m elementary predicates is defined to be $m/2^k$).

Let $D(S)$ be the density of the string — that is the number of ones in S divided by $l(S)$. Let $B(P)$ be the weight of the predicate P. Let $\mathcal{P}(S\cdot 1)$ be the relative probability of the event $(S)_{l(S)+1}=1$ (that is, that the next bit of S will be 1).

DEFINITION. An $l\times 2^{2^k}$ world W is H-unpredictable if for any $l\times 2^{2^k}$ world W' we have $V^H_W \leqq V^H_{W'}$.

In [3] we have shown that H-unpredictable worlds exist for any H, k and l and have proved the following

LEMMA. *If for some H, W is an H-unpredictable $l\times 2^{2^k}$ world, then for any string S_P in W we have*

$$|B(P)-D(S_P)|<\left(1-\frac{2^k}{l}\right)B(P).$$

Note that for fixed k the density of $l\times 2^{2^k}$ worlds which are unpredictable for some H, tends to 1, as $l\to\infty$. Also given H, k, c, there exists m such that for all l and every non-trivial predicate P the number of $l\times 2^{2^k}$ worlds such that $L(S_P)<c$ is bounded by m. Nevertheless for fixed H and k most worlds are not H-unpredictable, though some non-zero fraction of them are.

LOGICAL CRITERION THEOREM. *If $\mathcal{P}_H$ is a probability of degree f and W is an H-unpredictable $l\times 2^{2^k}$ world with $l>k+2^k$ and $k>3$, then for any string S_P in W we have*

$$|\mathcal{P}_H(S_p\cdot 1)-B(P)|<f(l-2^k)\cdot B(P).$$

Thus for large $l-2^k$ — that is, the length of the world is large relative to the number of elementary predicates — the probability of a predicate holding is close to its Bernoulli distribution. We interpret this as meaning that although no two predicates are independent in a finite world, they can approach independence in the limit.

Using this theorem together with the lemma we get

EMPIRICAL CRITERION THEOREM. *If $\mathcal{P}_H$ is a probability of degree f and W is an H-unpredictable $l \times 2^{2^k}$ world with $l > k + 2^k$ and $k > 3$, then for any string S_p in W we have*

$$|\mathcal{P}_H(S_P \cdot 1) - D(S_P)| < \left[f(l - 2^k) + \left(1 - \frac{2^k}{l}\right)\right] B(P).$$

Thus if $l - 2^k$ is large and $1 - 2^k/l$ is small (e.g., $l = 2^k + ck$) the probability of an event approaches its frequency.

It is a weakness of this theorem that we require $k \to \infty$ for the theorem to hold whereas k is fixed in any actual universe. It would be interesting to consider sequences of worlds in which, through the addition of artificial predicates, k is made to increase with l.

PROOF OF LOGICAL CRITERION THEOREM. It is proved in [3] that if W is H-unpredictable then for every atomic predicate P, $L(P) > l - 2^k$. Using Axiom III and the notation from the proof of the Confirmation Theorem we get that for any elementary predicate E,

$$\mathcal{P}(W \cdot E) \leqq (X_{l+1, l-2^k})^k.$$

Also $\mathcal{P}(W) \geqq (X_{l,l})^k$. Recalling that $X_{l,l} = e_l$ and $X_{i,m} = X_{i+1,m} + e_{i+1}$ and $e_0 = 1$ and

$$\frac{1}{2} - \frac{f(i)}{2^{2^i}} < \frac{e_{i+1}}{e_i} < \frac{1}{2}$$

and letting $a_i = e_i / e_{i-1}$, $a = a_{l-2^k+1}$ we get

$$\frac{\mathcal{P}(W \cdot E)}{\mathcal{P}(W)} \leqq \left(\frac{X_{l+1,l-2^k}}{X_{l,l}}\right)^k = \left(\frac{e_{l-2^k} - \sum_{i=1}^{2^k+1} e_{l-2^k+i}}{e_l}\right)^k$$

$$= \left(\frac{a_1 \cdots a_{l-2^k} - \sum_{i=1}^{2^k+1} a_1 \cdots a_{l-2^k+1}}{a_1 \cdots a_l}\right)^k$$

$$= \left(\frac{1 - \sum_{j=1}^{2^k+1} \prod_{i=1}^{j} a_{l-2^k+i}}{\prod_{i=1}^{2^k} a_{l-2^k+i}}\right)^k \leqq \left(\frac{1 - \sum_{j=1}^{2^k+1} a^j}{a^{2^k}}\right)^k$$

$$= \left(\frac{1 - \dfrac{a - a^{2^{k+2}}}{1-a}}{a^{2^k}} \right)^k = \left(\frac{1 - 2a + a^{2^{k+2}}}{(1-a)a^{2^k}} \right)^k$$

$$\leqq \left(\frac{2(1-2a)}{a^{2^k}} + 2a^2 \right)^k \leqq \left(\frac{1}{2} + \frac{4f(l-2^k)}{2^{l-2^k} a^{2^k}} \right)^k$$

$$\leqq \left(\frac{1}{2} + \frac{4f(l-2^k)}{2^{2^{k+1}} \left(\frac{1}{2} - \frac{f(l-2^k)}{2^{2^{k+1}}} \right)^{2^k}} \right)^k < \left(\frac{1}{2} + \frac{4f(l-2^k)}{2^{2^{k+1}} \left(\frac{1}{2^{2^k}} - \frac{2^k f(l \cdot 2^k)}{(2^{2^k-1})(2^{2^{k+1}})} \right)} \right)^k$$

$$< \left(\frac{1}{2} + \frac{4f(l-2^k)}{2^{2^k}-1} \right)^k < \frac{1}{2^k} + \frac{8f(l-2^k)}{2^{2^k}-1} < \frac{1}{2^k} + \frac{f(l-2^k)}{4^k}.$$

In the penultimate inequality we use the fact that $(\frac{1}{2} + d)^k < (\frac{1}{2})^k + 2d$ and assume that

$$\frac{4f(l-2^k)}{2^{2^k}-1} < \frac{1}{2}.$$

In the case where this assumption fails the theorem is trivial.

Thus

$$\frac{\mathcal{P}(W \cdot E)}{\mathcal{P}(W)} < \frac{1}{2^k} + \frac{f(l-2^k)}{4^k}$$

even for the most probable elementary predicate E. Note that by Axiom II

$$\sum_{i=1}^{2^k} \frac{\mathcal{P}(W \cdot E_i)}{\mathcal{P}(W)} = 1$$

where the sum is taken over all elementary predicates. Suppose that for some elementary predicate E_j we have

$$\frac{\mathcal{P}(W \cdot E_j)}{\mathcal{P}(W)} \leqq \frac{1}{2^k} - \frac{f(l-2^k)}{2^k}.$$

Then

$$\sum_{i=1}^{2^k} \frac{\mathcal{P}(W \cdot E_i)}{\mathcal{P}(W)} < \sum_{\substack{i=1 \\ i \neq j}}^{2^k} \frac{\mathcal{P}(W \cdot E_i)}{\mathcal{P}(W)} + \frac{1}{2^k} - \frac{f(l-2^k)}{2^k}$$

$$< (2^k - 1)\left(\frac{1}{2^k} + \frac{f(l-2^k)}{4^k} \right) + \frac{1}{2^k} - \frac{f(l-2^k)}{2^k}$$

$$< 1.$$

Thus for all elementary predicates E we have

$$\left|\frac{\mathcal{P}(W \cdot E)}{\mathcal{P}(W)} - \frac{1}{2^k}\right| < \frac{f(l-2^k)}{2^k}.$$

Recall that the relative probability of an event

$$(S_P \cdot 1) = \sum_{VE_i = P} \frac{(W \cdot E_i)}{(W)}$$

where the sum is taken over those elementary predicates which are disjuncts of P. There are precisely $B(P) \cdot 2^k$ such disjuncts. Thus

$$|P(S_P \cdot 1) - B(P)| \leqq \left|\frac{\mathcal{P}(W \cdot E)}{\mathcal{P}(W)} - \frac{1}{2^k}\right| (B(P) \cdot 2^k) \leqq \frac{f(l-2^k)}{2^k} \cdot B(P) \cdot 2^k$$
$$= f(l-2^k) \cdot B(P). \qquad \square$$

Conclusions

In order to make plausible our contention that the probability described above is, in skeleton form, something like what people really mean when they speak of probability, it will be necessary both to sharpen and to generalize our results. We need to show, for example, that in the absence of total evidence, partial evidence is also a useful, if imperfect, guide for determining probabilities and in particular that as partial evidence approaches total evidence, our predictions approach, in some continuous way, the predictions we'd make from total evidence.

Our ace in the hole is the fact that we obtained results without specifying an inference method. By restricting the choice of inference method we can make stronger statements about the resultant probability measures. For example, consider the following formulation of the problem of partial evidence. If W is a world including the string S_P which contains n ones, let W^P be the world of length n consisting of the restriction of W to those moments i for which $(S_P)_i = 1$ and those strings $S_{P'}$ for which P' is logically independent of P. If W is H-unpredictable, will W^P be H-unpredictable? If W exhibits strong patterns relative to inference method H will these be preserved in W^P? The results of Martin-Löf [4] suggest that by choosing H such that higher level strings are of higher computation complexity, we might under certain conditions answer these question affirmatively. In fact, Solomonoff [6] has shown how computational complexity can be used to define probability in a way very similar to ours, but the domain of his probability function is the set of events, i.e., single binary strings while the domain of our probability function is the set of worlds, i.e., of collections of strings closed under Boolean operations.

We also need to extend our results to a more general class of worlds. Since the "real world" manifests more than a single predicate and since it is not completely unpredictable (by our definition) neither the Confirmation Theorem nor the Criterion Theorems apply to it. It might be sufficient however to extend these results to "quasi-unpredictable worlds" — that is, worlds in which all strings are either very high level or very low level (i.e., all confirmed causal relations are relatively simple). We propose that if W is a quasi-unpredictable world then the Criterion Theorems can be extended to W where the predicates are clumped into equivalence classes and the constraints on l and k translated appropriately, and that some form of the Confirmation Theorem can be applied to the low level strings of W.

Such an extension of our results might be achieved through the use of another additivity axiom, namely the following:

Let W be a world manifesting the universe $U = [P_1, \ldots, P_k]$ and let W' be a world manifesting the universe $U' = [P_1, \ldots, P_{k+1}]$. If W is the restriction of W' to the universe U, then we call W' a refinement of W. Then $\mathcal{P}(W) = \Sigma \mathcal{P}(W')$ where the sum is over all refinements of W, i.e., over all possible extensions of P_{k+1}.

Acknowledgment

The author thanks Haim Gaifman for helpful suggestions.

References

1. L. Blum and M. Blum, *Toward a mathematical theory of inductive inference*, Inf. Control **28** (1975), 125–155.
2. R. Carnap, *The Continuum of Inductive Methods*, Univ. of Chicago Press, Chicago, 1950.
3. M. Koppel, *Hierarchal inductive inference methods*, Philosophia (submitted).
4. P. Martin-Löf, *The definition of random sequences*, Inf. Control **9** (1966), 602–619.
5. K. Popper, *The Logic of Scientific Discovery*, 2nd edn., Harper Torchbooks, New York, 1968.
6. R. J. Solomonoff, *A formal theory of inductive inference*, Inf. Control 7 (1964), 1–22, 224–254.