

A THEORY OF HIGHER ORDER PROBABILITIES¹

Haim Gaifman

Mathematics Department The Hebrew University Jerusalem, Israel
Currently Visitor at SRI

ABSTRACT

We set up a general framework for higher order probabilities. A simple HOP (Higher Order Probability space) consists of a probability space and an operation PR , such that, for every event A and every real closed interval Δ , $PR(A, \Delta)$ is the event that A 's "true" probability lies in Δ . (The "true" probability can be construed here either as the objective probability, or the probability assigned by an expert, or the one assigned eventually in a fuller state of knowledge.) In a general HOP the operation PR has also an additional argument ranging over an ordered set of time-points, or, more generally, over a partially ordered set of stages; $PR(A, t, \Delta)$ is the event that A 's probability at stage t lies in Δ . First we investigate simple HOPs and then the general ones. Assuming some intuitively justified axioms, we derive the most general structure of such a space. We also indicate various connections with modal logic.

¹A part of this paper has been included in a talk given in a NSF symposium on foundations of probability and causality, organized by W. Harper and B. Skyrms at UC Irvine, July 1985. I wish to thank the organizers for the opportunity to discuss and clarify some of these ideas.

Introduction

The assignment of probabilities is the most established way of measuring uncertainties on a quantitative scale. In the framework of subjective probability, the probabilities are interpreted as someone's (the agent's) degrees of belief. Since justified belief amounts to knowledge, the assignment of probabilities, in as much as it can be justified, expresses knowledge. Indeed, knowledge of probabilities, appears to be the basic kind of knowledge that is provided by the experimental sciences today.

This is knowledge of a partial, or incomplete, nature, but not in the usual sense of "partial". Usually we mean by "partial knowledge" knowledge of some, but not all, of the facts in a certain domain. But knowing that a given coin is unbiased does not enable one to deduce any non-tautological Boolean combination of propositions which describe outcomes in the next, say fifty tosses. And yet it constitutes very valuable knowledge about these very same outcomes. What is the objective content of this knowledge? What kind of fact is the fact that the true probability of "heads" is 0.5, i.e., that the coin is unbiased? I shall not enter here into these classical problems². I take it for granted that, among probability assignments, some are more successful, or better tuned to the actual world, than others. Consequently probability assignments are themselves subject to judgement and evaluation. Having, for example, to estimate the possibility of rain I might give it, going by the sky's appearance, 70%. But I shall be highly uncertain about my estimate and will adopt the different value given, five minutes later, in the weather forecast.

Thus we have two levels of uncertainty:

1. Uncertainty concerning the occurrence of a certain event - expressed through the assignment of probabilities.
2. Uncertainty concerning the probability values assigned in 1.

²My Salzburg paper [1983] has been devoted to these questions. The upshot of the analysis there has been that even a "purely subjective" probability implies a kind of factual claim, for one can assess its success in the actual world. Rather than two different kinds, subjective and objective probabilities are better to be regarded as two extremes of a spectrum.

When this second level is itself expressed by assigning probabilities we get second order probabilities. An example of a second order probability is furnished by a cartoon in "The New Yorker" showing a forecaster making the following announcement:

"There is now 60% chance of rain tomorrow, but, there is 70% chance that later this evening the chance of rain tomorrow will be 80%."

Just as we can iterate modal or epistemic operators, so in the system to be presented here we can iterate the probability-assignment operator to any depth. The goal of this paper is to present a general and adequate semantics for higher order probabilities and to obtain, via representation theorems, nice easily understood structures which give us a handle on the situation.

The basic structure to be defined here is a **HOP** (Higher Order Probability space). A *simple* HOP is based on a field of events, F , and on a binary operator $PR(,)$ which associates with every event A and every real closed interval Δ an event $PR(A, \Delta)$ in F . The intended meaning is that $PR(A, \Delta)$ is the event that A 's true probability lies in the interval Δ .

"True probability" can be understood here as the probability assigned by an ideal expert or by someone fully informed. It is however up to us (or to the agent) to decide what in the given context constitutes an "ideal expert" or "someone fully informed". If "full information" means knowing *all* the facts then, of course, the true (unknown to us) probability has only two values 0 and 1; this will make the HOP trivial in a certain sense. In the other extreme, the agent may regard himself as being already fully informed and this leads to the "opposite" trivialization of the HOP. Generally, the agent will regard the expert as being more knowledgeable than himself, but not omniscient; e.g., the expert might know the true bias of a coin but not the outcomes of future tossings, or he might have statistical information for estimating the bias, which the agent lacks.

The agent himself at some future time can be cast in the role of being "fully informed". Thus, if P is the forecaster's present probability function and if PR represents his state of knowledge later in the evening, then his announcement in "The New Yorker" cartoon can be summed up as follows, where $A =$ 'tomorrow it will rain':

$$P(A) = .6 \quad P(PR(A, [.8, .8])) = .7$$

In order to represent knowledge at different stages, we make PR into a 3-place operator: $PR(A, t, \Delta)$ is the event that the probability of A at stage t lies in Δ . The stages can be time-points, in which case t ranges over some ordered set. More generally, the set of stages is only partially ordered, where $s \leq t$ if the knowledge at stage t includes the knowledge at stage s . (Different agents may thus be represented in the structure.) This is how a HOP is defined in general. We shall first establish the

properties of simple HOPs, then use them to derive those of the more general spaces.

We shall also define, in a separate section, a formal logical calculus, to be called *probability logic*, which is naturally associated with simple HOPs. Various modalities can be reconstructed within this calculus. The general HOPs give rise to stage-dependent modalities whose calculus will be outlined at the end of the paper.

The import of the subject for various branches of philosophy and for the foundations of probability is obvious. Also obvious should be its bearing upon applied probabilistic reasoning in distributed networks, or upon efforts to incorporate such reasoning in AI systems. Mathematically, most of this paper is rather easy. Our goal has not been to prove difficult theorems, but to clarify some basic concepts and to outline a general, conceptually "clean", framework within which one can use freely and to good effect statements such as: 'With probability 0.7 Adam will know at stage 3 Bob's probability for the event A , with error ≤ 0.01 ' (where Adam and Bob are either people or processors). Statements of this form express intuitive thinking which may underly involved technical proofs; to use them openly and precisely can help us as a guide for finding and organizing our arguments.

A theoretic framework for higher order probabilities may also yield insights into systems of reasoning which employ non-probabilistic certainty measures. For when probability is itself treated like a random variable, we can use various methods of "safe" estimation which do not necessarily yield a probability measure. For example, define the *certainty measure* of an event A to be the largest α such that, with probability 1, the probability of A is $\geq \alpha$. This is only one, apparently the most conservative, measure among various measures that can be used.

Higher order probabilities have been considered by De-Finetti, but rejected by him owing to his extreme subjectivist views. Savage considered the possibility but did not take it up, fearing that the higher order probabilities will reflect back on the ground level, leading to inconsistencies. Instances of higher order probabilities figure in works of Good [1965] and Jaynes [1958]. More recent philosophical works are by Domotor [1985], Gärdenfors [1975] (for qualitative probabilities), Miller [1966], Skyrms [1980 A], [1980 B] - who did much to clarify matters, van-Frassen [1984], and others.

Due to limitations of space and deadline I have not entered into details of various proofs. Some of the material has been abridged; I have included some illustrative examples of simple HOPs, but not the more interesting ones of general HOPs (which arise naturally in distributed systems). Also the bibliography is far from complete.

Simple HOPs

Definition and Basic Properties

As in Kolmogoroff's framework [1933] we interpret propositions as subsets of some universal set, say W , and we refer to them as events. We can regard W as the set of all possible worlds. Thus we have $X =$ set of all worlds in which X is true and we get the following standard correspondence:

$$\begin{aligned} \vee \text{ (disjunction)} &\mapsto \cup \text{ (union)} \\ \wedge \text{ (conjunction)} &\mapsto \cap \text{ (intersection)} \\ \neg \text{ (negation)} &\mapsto - \text{ (complementation)} \end{aligned}$$

Terminology: A Boolean Algebra (of sets) is a class of sets closed under finite unions (and intersections) and under complementation (with respect to some presupposed universal set, in our case W). A field is a Boolean algebra closed under countable unions (what is known also as a σ -algebra). The field (Boolean algebra) generated by a class S of sets is the smallest field (Boolean algebra) which contains S as a subclass. Note that in generating a Boolean algebra we apply finitary operations only, whereas in generating a field infinitary countable operations are used. A field is countably generated if it has a countable set of generators. All probabilities are assumed here to be countably additive.

A **HOP** is a 4-tuple (W, F, P, PR) , where F is a field of subsets of W , to be called events, P is a probability over F and PR is a mapping associating with every $A \in F$ and every real closed interval Δ an event $PR(A, \Delta)$,

$$PR : F \times \text{set of closed intervals} \rightarrow F$$

As explained in the introduction $PR(A, \Delta)$ is the event that the true (or the eventual, or the expert-assigned) probability of A lies in Δ . P is the agent's current subjective probability.

Among the closed intervals, we include also the empty interval, \emptyset . The minimal and maximal elements of F are, respectively, $\mathbf{0}$ and $\mathbf{1}$; that is: $\mathbf{0} =$ empty subset of $W =$ False, $\mathbf{1} = W =$ True.

In the explanations I shall use "probability" both for the agent's current subjective probability as well as for the true, or eventual one; the contexts indicate the intended reading.

The following axioms are postulated for a HOP:

(I) $PR(A, [0,1]) = \mathbf{1}$ (For every A , the event that A 's probability lies in $[0,1]$ is W , i.e., true.)

(II) $PR(A, \emptyset) = \mathbf{0}$ (That A 's probability lies in the empty interval is the empty event, i.e., false.)

(III) If $\Delta_1 \cup \Delta_2$ is an interval then $PR(A, \Delta_1 \cup \Delta_2) = PR(A, \Delta_1) \cup PR(A, \Delta_2)$ (A's probability lies in the interval $\Delta_1 \cup \Delta_2$ iff it lies either in Δ_1 or in Δ_2)

In the following two axioms "n" is a running index ranging over $\{1, 2, \dots\}$.

(IV) $\bigcap_n PR(A, \Delta_n) = PR(A, \bigcap_n \Delta_n)$ (A's probability lies in every Δ_n iff it lies in their intersection).

(V) If, for all $n \neq m$, $A_n \cap A_m = \emptyset$, then $\bigcap_n PR(A, [\alpha_n, \beta_n]) \subset PR(\bigcup_n A_n, [\sum_n \alpha_n, \sum_n \beta_n])$ (For pairwise disjoint A_n 's, if A_n 's probability lies in $[\alpha_n, \beta_n]$, $n=1, 2, \dots$, then the probability of $\bigcup_n (A_n)$ lies in $[\sum_n \alpha_n, \sum_n \beta_n]$.)

Note that axioms (I)-(V) involve only W , F and PR . The crucial axiom which connects PR with P will be stated later.

THEOREM 1 For every HOP, $H=(W, F, P, PR)$ there is a mapping p which associates with every x in W a probability, p_x , over F such that

$$(1) \quad PR(A, \Delta) = \{x : p_x(A) \in \Delta\}$$

The mapping p is uniquely determined by (1) and can be defined by:

$$(2) \quad p_x(A) = \inf\{\alpha : x \in PR(A, [0, \alpha])\}$$

as well as by:

$$(2') \quad p_x(A) = \sup\{\alpha : x \in PR(A, [\alpha, 1])\}.$$

Vice versa, if, for every $x \in W$, p_x is a probability over F such that $\{x : p_x(A) \in \Delta\}$ is in F for all $A \in F$ and all real closed Δ , and if we use (1) as a definition of PR then Axioms (I)-(V) are satisfied.

We call p the kernel of the HOP.

The proof of Theorem 1 is nothing more than a straight-forward derivation of all the required details from the axioms, using (2) as the definition of p_x . (The "vice versa" part is even more immediate than the first part.)

We can now extend PR and define $PR(A, \Xi)$, for arbitrary subsets Ξ of reals, as $\{x : p_x(A) \in \Xi\}$. If Ξ is a Borel set then $PR(A, \Xi)$ is in F .

The meaning of p_x is obvious: it is the probability which corresponds to the maximal state of

knowledge in world x - the distribution chosen by the expert of that world.

Notation: For $\alpha \in [0,1]$, $PR(A,\alpha) =_{df} PR(A, [\alpha,\alpha])$.

The picture is considerably simpler in the discrete case, where W is countable. Assuming with no loss of generality that $\{x\} \in F$ for every $x \in W$, the probability of some $A \subset W$ is simply the sum of the probabilities of the worlds in A . In that case, we can eliminate the closed intervals and consider only the special cases $PR(A,\alpha)$ where α ranges over $[0,1]$; also our 5 axioms can be replaced by 3 simpler ones. Discrete cases arise in many situations and are very useful as illustrative examples. But to consider only the discrete case is highly restrictive.

Notation: For $x,y \in W$, $A \in F$, put: $p(x,A) =_{df} p_x(A)$ and (assuming $\{y\} \in F$) $p(x,y) =_{df} p(x,\{y\})$ and $P(y) =_{df} P(\{y\})$.

In the discrete case P is obviously determined by the values $P(x)$, $x \in W$. Thus, ordering W , we can represent P as a probability vector (a countable vector of non-negative entries which sum up to 1). Similarly the kernel p becomes a probability matrix (a countable square matrix in which every row is a probability vector). Examples (i) and (ii) in the **Examples** subsection can serve to illustrate the situation (the discussion there presupposes however the next subsection).

Mathematically, what we have got is a Markov process (with initial probability P and transition probabilities $p(x, \cdot)$, $x \in W$). But the interpretation is altogether different from the usual interpretation of such a structure. The connection between P and the kernel p is established in the sixth axiom.

Axiom (VI) And Its Consequences

Let $P(A|B)$ be the conditional probability of A , given B . It is defined in the case that $P(B) \neq 0$ as $P(A \cap B)/P(B)$. It is what the agent's probability for A should be had he known B .

Axiom (VI_w) If $P(PR(A, [\alpha,\beta])) \neq 0$ then $\alpha \leq P(A | PR(A, [\alpha,\beta])) \leq \beta$.

Axiom (VI_w) (the weak form of the forthcoming Axiom (VI)) is a generalization of Miller's Principle to the case of interval-based events. Rewritten in our notation, Miller's Principle is: $P(A | PR(A,\alpha)) = \alpha$. Axiom (VI_w) appears to be the following rule: My probability for A should be no less than α and no more than β , were I to know that in a more informed state my probability for A will be within these bounds. Plausible as it sounds, the use of the hypothetical "were I to know that..." needs in this context some clarification. Now a well-known way of explicating conditional probabilities is through conditional bets. Using such bets van-Frassen [1984] gives a Dutch-book argument for the Principle: Its

violation makes possible a system of bets (with odds in accordance with the agent's probabilities) in which the agent will incur a net loss in all circumstances. In this argument $PR(A, \alpha)$ is interpreted as the event that the agent's probability for A at a certain future time will be α , in which case he should accept at that time bets with odds α . The same kind of Dutch-book can be constructed if Axiom (VI_w) is violated. (Here it is crucial that we use an interval, the argument fails if we replace $[\alpha, \beta]$ by a non-convex Borel set.)

Axiom (VI) is the interval-based form of the stronger version of Miller's Principle which was suggested by Skyrms [1980 A].

Axiom (VI) *If C is a finite intersection of events of the form $PR(B, \Delta)$, and if $P(C \cap PR(A, [\alpha, \beta])) \neq 0$, then*

$$\alpha \leq P(A | C \cap PR(A, [\alpha, \beta])) \leq \beta$$

The same intuition which prescribes (VI_w) prescribes (VI) ; also here the violation of the axiom makes possible a Dutch-book against the agent. What is essential is that events of the form $PR(B, \Delta)$ be, in principle, knowable to the agent, i.e., be known (if true) in the maximal states of knowledge as defined by our structure.³

In what follows integrating a function $f(t)$ with respect to a probability m is written as $\int f(t) \cdot m(dt)$.

Lemma 1 *Axiom (VI_w) implies that the following holds for all $A \in F$:*

$$(3) \quad P(A) = \int p(x, A) \cdot P(dx)$$

The proof consists in applying the formula $P(A) = \sum_i P(A | B_i) \cdot P(B_i)$, where the B_i 's form a partition, passing to the limit and using the definition of an integral.

The implication $(3) \Rightarrow (VI_w)$ is not true in general. Note that in the discrete case (3) becomes:

$$(3_d) \quad P(x) = \sum_y p(x, y) \cdot P(y)$$

which means that the probability vector is an eigen-vector of the kernel.

Definition Call two worlds $x, y \in W$ epistemically equivalent, (or, for short, equivalent) and denote it by $x \simeq y$, if $p_x = p_y$. For S - a class of events, define K/S to be the field generated by all events of the form $PR(A, \Delta)$, $A \in S$, Δ - a real closed interval.

³It is important to restrict C in Axiom (VI) to an intersection of such events. The removal of this restriction will cause the p_x 's to be two-valued functions, meaning that all facts are known in the maximal knowledge states.

Epistemic equivalence means having the same maximal knowledge. Evidently $x \simeq y$ iff, for all A and all Δ , $x \in PR(A, \Delta) \Leftrightarrow y \in PR(A, \Delta)$; this is equivalent to: for all $C \in K[F]$, $x \in C \Leftrightarrow y \in C$. If $K[F]$ is generated by the countably many generators X_n , $n=0,1,\dots$ then the equivalence classes are exactly all non-empty intersections $\bigcap_n X_n'$ where each X_n' is either X_n or its complement. Hence the equivalence classes are themselves in $K[F]$, they are exactly the atoms of this field. The next lemma shows that the condition that $K[F]$ be countably generated is rather mild, for it holds whenever F itself is countably generated (which is the common state of affairs):

Lemma 2 *If S is either countable or a countably generated field, then $K[S]$ is countably generated.*

(As generators for $K[S]$ one can take all $PR(A, \Delta)$, $A \in S$, Δ - a rational closed interval; the second claim is proved by showing that if S' is a Boolean algebra that generates the field S then $K[S'] = K[S]$.)

Terminology: A 0-set is a set of probability 0. Something is said to hold for almost all x if it holds for all x except for a 0-set. The probability in question is P , unless specified otherwise.

Theorem 2 *If F is countably generated then axiom (VI) is equivalent to each of the following conditions:*

(A) (3) holds (for all A) and the following is true: Let C_x be the epistemic equivalence class to which x belongs, then

$$p_x(C_x) = 1 \quad \text{for almost all } x.$$

(B) (3) holds and, for almost all x , for all A :

$$(4) \quad p_x(A) = \int p_y(A) \cdot p_x(dy)$$

The proof that axiom (VI) is equivalent to (A) and implies (B) uses only basic measure theory. The present proof of (B) \Rightarrow (A) relies on advanced ergodic theory⁴ and I do not know if this can be avoided. Fortunately the rest of this paper does not rely on this implication (except the corresponding implication in Theorem 3). Note that in the discrete case (4) is equivalent to:

$$(4_d) \quad p(x, z) = \sum_y p(x, y) \cdot p(y, z)$$

(4_d) means that the kernel, as a matrix, is equal to its square.

Let $\{E_u : u \in U\}$ be the family of epistemic equivalence classes, with different indices attached to

⁴I am thankful to my colleagues at the Hebrew University H. Furstenberg, I. Katzenelson and B. Weiss for their help in this item. Needless to say that errors, if any, are my sole responsibility.

different classes. Let P_u be the common p_x for $x \in E_u$; let m be the probability, defined for all $V \subset U$ such that $\bigcup_{u \in U} E_u \in F$, by:

$$m(V) = P(\bigcup_{u \in V} E_u)$$

Then (A) is equivalent to the following condition:

(C) For all A in F

$$P(A) = \int_U P_u(A) \cdot m(du)$$

and for almost all (with respect to m) u

$$P_u(E_u) = 1$$

The first equality in (C) is a recasting of (3); it can be equivalently described by saying that P is a mixture of the P_u 's with weight function m . Altogether (C) means that we have here what is known as the disintegration of the probability space. It makes for a rather transparent structure.

For W -countable the disintegration means the following: After deleting from the kernel-matrix rows and columns which correspond to some set of probability 0, the rest decomposes into submatrices around the main diagonal in each of which all rows are equal, with 0's in all other places; P itself is a mixture of these rows. Such HOPs are exactly those that can be constructed as follows (hence this is the method for setting up higher order probabilities which avoid a Dutch book):

- Chose a partition $\{E_u : u \in U\}$ of W into non-empty disjoint sets, with different u 's marking different sets.
- Chose for each u in U a probability, P_u , on W such that $P_u(E_u) = 1$ for all $u \in U'$, where U' is some non-empty subset of U .
- Chose a probability, m , on U such that $m(U') = 1$, and let P be the mixture of the P_u 's with weight function m .
- For each $u \in U$ and each $x \in E_u$ put $p_x = P_u$ and define $PR(A, \Delta)$ to be $\{x : p_x(A) \in \Delta\}$.

The construction is essentially the same for a general W (with a countably generated F); some additional stipulations of measurability should be included in order to make possible the formation of the mixture and to ensure that the $PR(A, \Delta)$'s are in F .

Definition Call P_u and its corresponding equivalence class, E_u , ontological if $P_u(E_u) = 1$, call it and its corresponding class coherent if P_u is a mixture of ontological P_u 's. Call a world ontological (coherent) if it belongs to an ontological (coherent) equivalence class.

An ontological class is of course coherent. A coherent class which is not ontological must get the value 0 under its own P_u . It represents a state of knowledge in which the agent knows for sure that his

eventual probability function will be different from his current one (and that it will be an ontological one).

The set of ontological worlds gets the value 1 under P and under each p_x where x is coherent. It is referred to as the ontological part of the HOP. Together with the structure induced by the original HOP it forms by itself a simple HOP. Similarly we define the coherent part of the HOP as the set of all coherent worlds (together with the induced structure). As far as calculating probabilities goes, only the ontological part matters. Coherent non-ontological worlds are useful as representatives of transitory states of knowledge.

Examples

Example 1: $W = \{w1, w2, w3\}$ $P = (1/3, 1/3, 1/3)$ and the kernel matrix is:

$$\begin{matrix} .5 & .5 & 0 \\ 0 & .5 & .5 \\ .5 & 0 & .5 \end{matrix}$$

The agent's current probability assigns each world the value 1/3. Eventually, in world $w1$ he will know that he is not in $w3$ and he will assign each of the worlds $w1, w2$ the value 0.5. This is the meaning of the first row. The other rows are similarly interpreted.

By direct checking one can verify that (VI_w) is satisfied. (The checking of all cases in this example is easy because $PR(A, \alpha) \neq \emptyset$ only for $\alpha = 0.5, 1$.) However the matrix is not equal to its square, hence Axiom (VI) is violated, as indeed the following case shows: Put $A = \{w1\}$, $C = PR(\{w2\}, 0.5)$. Then $C = \{x : p(x, w2) = 0.5\} = \{w1, w2\}$ and similarly $PR(A, 0.5) = \{w1, w3\}$. Hence $A = PR(A, 0.5) \cap C$ implying $P(A | PR(A, 0.5) \cap C) = 1 \neq 0.5$. This can be used to construct a Dutch book against the agent.

Note also that the epistemic equivalence classes are $\{w1\}, \{w2\}$ and $\{w3\}$ and that non is ontological; hence also there are no coherent worlds here.

Example 2 $W = \{w1, w2, \dots, w8\}$, P is: $(.1, .2, .2, .1, .4, 0, 0, 0)$ and the kernel matrix is:

.2	.4	.4							
.2	.4	.4							
.2	.4	.4							
			.2	.8					
			.2	.8					
					1				
.05	.2	.2	.05	.2	.5	0	0		
.2	.1	.1	.2	.1	.1	.1	.1	.1	

where all undisplayed entries are 0. The sets $\{w_1, w_2, w_3\}$, $\{w_4, w_5\}$ and $\{w_6\}$ are equivalence classes which are ontological. P is a mixture of these 3 types of rows, with weights 0.5, 0.5, 0, respectively. Hence condition (C) is satisfied, therefore also Axiom (VI). w_7 is a coherent non-ontological world, because the 7th row is a mixture of the first three types (with weights .25, .25, .5) w_8 is not coherent. The ontological part consists of the upper left 6×6 matrix and the coherent part of the 7×7 one.

The example can be made more concrete by the following scenario. A number is to be chosen from $\{1, 2, 3\}$. For $i=1, 2, 3$, the number chosen in w_i is i , but in each of these 3 worlds the maximal knowledge consists in assigning probabilities 0.2, 0.4, 0.4 to the 3 possibilities. In w_4 the number chosen is 1 and in w_5 it is 2; in either of these worlds the maximal knowledge consists in assigning the probabilities 0.2, 0.8. In w_6 the number is 2 and it is also assigned probability 1. In the agent's current state he assigns probability 0 to finding himself eventually in the third state of maximal knowledge, and equal probabilities to the first and second states. World w_7 represent a similar situation but with different weights. We can imagine 3 lotteries for choosing the number; in each equivalence class the maximal knowledge is knowledge of the chosen lottery.

Example 3: Let H be the probability of "heads" of some given coin of unknown bias. Treat H as a random variable. The agent's knowledge is represented by a probability distribution for H . Say it is the uniform distribution over $[0, 1]$. The expert does not know the value of H but he has some additional information. Say his additional information is the value of N - the number of "heads" in 50 independent tosses. Then our worlds can be regarded as pairs (h, n) , such that in (h, n) the event $H=h \cap N=n$ is true; here h is a real number in $[0, 1]$ and n an integer between 0 and 50. The field F is generated by the sets $[\alpha, \beta] \times \{n\}$, $0 \leq \alpha \leq \beta \leq 1$, $n=0, \dots, 50$.

Given $H=h$, we get the binomial distribution $b_{h, 50}$ for N . This fact, together with the agent's

uniform distribution for H , determines his probability P over F . The expert's probability in world (h, n) is obtained by conditioning on his information, it is $P(N=n)$. There are 51 equivalence classes which correspond to the 51 possible values of N and all worlds are ontological.

As is well known, different values of N give rise to different conditional distributions of H . Therefore the events $N=n$ are in the field generated by the events⁵ $PR(H \in [\alpha, \beta], \Delta)$. The whole field F is therefore generated by events which are either of the form $H \in [\alpha, \beta]$ or obtained from these by applying the operator PR . Consequently we can give an abstract description of this HOP which does not mention the fifty tosses. The only function of the tosses is to affect the distribution of H ; in our framework such changes in distribution constitute themselves events which can be treated directly, without having to bring in their causes.

The Case of a General Field

The restriction that F be countably generated is a mild one. The probability spaces which commonly appear in theory, or in applications, are essentially of this nature⁶. Usually we are interested in properties that involve only countably many generators. We will first show that for studying such properties we can always restrict ourselves to the case where the underlying field is countably generated.

Definition Given a simple HOP (W, F, P, PR) and given $S \subset F$, define $H[S]$ as the smallest field containing S and closed under PR (i.e., $A \in H[S] \Rightarrow PR(A, \Delta) \in H[S]$ for every real closed interval Δ).

$H[S]$, together with the restrictions of P and PR to it, forms a subHOP, where this notion is defined in the obvious way.

Lemma 3 *If S is a Boolean algebra and, for every A in S and every rational closed interval Δ , $PR(A, \Delta)$ is in S , then $H[S]$ is the field generated by S .*

This means that, once we have a Boolean algebra closed under $PR(, \Delta)$ for all Δ with rational endpoints, we get all the rest by countable Boolean operations without using PR .

Corollary *If S is either countable, or a countably generated field, then $H[S]$ is countably generated.*

⁵ Actually there are 51 real numbers α_n such that the event $N=n$ is the same as $PR(H \leq 1/2, \alpha_n)$.

⁶ They are *separable*, i.e., for some countably generated field every event in the space differs from a set in the field by a 0-set.

Using this we can derive from Theorem 2 an analogous result for general fields:

Theorem 3 *Axiom (VI) is equivalent to each of the following conditions:*

(A') (3) holds and for every C in $K[F]$, for almost all x : $p_x(C)=1$ if $x \in C$, $p_x(C)=0$ otherwise.

(B') (3) holds and for every A in F (4) is true for almost all x .

(B') differs from the analogous (B) of Theorem 2 in that the exceptional 0-set for (4) can depend on A .

Say that A is equal a.e. to B if $A-B$ and $B-A$ are 0-sets. Say that two classes of sets are equal modulo 0-sets if every member of one is equal a.e. to some member of the other.

Assuming Axiom (VI) we get:

Corollary *If $S \subset F$, then: (i) The fields $K[S]$, $K[K[S]]$ and $K[H[S]]$ are equal modulo 0-sets. (ii) If S is a boolean algebra then $H[S]$ is equal modulo 0-sets to the field generated by $S \cup K[S]$.*

(To show, for example, that $K[S]=K[K[S]]$ modulo 0-sets, consider $C \in K[S]$; by Theorem 3, $\{x : p_x(C) \in \Delta\}$ is equal a.e. to C if $\Delta=[1,1]$, is equal a.e. to $W-C$ if $\Delta=[0,0]$, and is a 0-set if $0,1 \notin \Delta$. Hence, for all Δ , $PR(C, \Delta)$ is equal a.e. to one of: C , $W-C$, W , \emptyset . Since $K[K[S]]$ is generated by such sets, the claim follows.)

Roughly speaking, (ii) means that, modulo 0-sets, nested applications of PR reduce to non-nested applications. A stronger, syntactical version of this is given in the next section.

Probability Logic

Let Ξ be a set of reals such that $0,1 \in \Xi$. Call an interval with end-points in Ξ a Ξ -interval.

Let PRL_{Ξ} be the calculus obtained by adjoining sentential operants, $PR(, \Delta)$, to the propositional calculus, where Δ ranges over all closed Ξ -intervals. Here, for the sake of convenience, I use 'PR' for the syntactical operant, as well as for the operation in HOPs. Given some class $\{X_i : i \in I\}$ of sentential variables, the class of all wffs (well formed formulas) of PRL_{Ξ} is the smallest such that:

- Every sentential variable is a wff
- If ϕ and ψ are wffs, so are $\neg\phi$ and $\phi * \psi$ where $*$ is any of the standard binary connectives.

• If ϕ is a wff and Δ is a closed Ξ -interval then $PR(\phi, \Delta)$ is a wff.

Let $H = (W, F, P, PR)$ be a simple HOP and let τ be a mapping which maps each sentential variable to a member of F . Then the value $|\phi|_{H, \tau}$ of the wff ϕ is defined by interpreting the sentential connectives as the corresponding Boolean operations and each syntactic operant $PR(, \Delta)$ as the operation $PR(, \Delta)$ of the HOP.

Definition A wff ϕ is *p-valid*, to be denoted $\models_p \phi$, if, for every simple HOP H which satisfies Axiom (VI) and every τ , the probability of $|\phi|_{H, \tau}$ is 1. Two wffs ϕ, ψ are *p-equivalent* if $\phi \leftrightarrow \psi$ is p-valid.

Call ϕ a *PC-formula* if it is a wff of the propositional calculus, i.e., does not contain any *PR*.

Theorem 4 Every wff of PRL_{Ξ} is p-equivalent to a Boolean combination of PC-formulas and formulas of the form $PR(\sigma, \Delta)$ in which σ ranges over PC-formulas.

This means that as far as probabilities are concerned (i.e., if we disregard 0-sets) we need not use nested *PR*'s.

Theorem 5 Translate into PRL_{Ξ} the wffs of propositional modal logic with the necessity operant *N*, by replacing each $N(\psi)$ by $PR(\psi, [1,1])$. Let ϕ^* be the translation of ϕ . Then

$$S5 \vdash \phi \quad \text{iff} \quad \models_p \phi^*$$

Thus *S5* becomes a fragment of PRL_{Ξ} . This relation becomes more explicit if we rewrite ' $PR(\psi, \Delta)$ ' as ' $N_{\Delta}(\psi)$ '.

It can be shown that for $\Xi =$ set of rationals the set of p-valid wffs is recursive. Also PRL_{Ξ} can be provided with a natural set of formal axioms so that, with modus ponens as derivation rule, p-validity coincides with provability.

Some Questions

Other validity notions can be considered (e.g., that $|\phi|_{H, \tau}$ always contains all coherent worlds in the HOP), as well as other interpretations of the necessity operant (e.g., as $\phi \wedge PR(\phi, [1,1])$). What modal logics are thereby obtained?

General HOPs

In general, a HOP is a structure of the form:

$$(W, F, P, T, PR)$$

where, as before, (W, F, P) is a probability space, $T = (T, <)$ is a partially ordered set and where

$$PR : F \times T \times \text{set of closed intervals} \rightarrow F$$

$PR(A, t, \Delta)$ is the event that the probability of A at stage t lies in Δ . If the stages coincide with time points then the partial ordering of T is total. As before, P is the current subjective probability; here "current" is earlier (i.e., less than or equally informative) than the stages in T . Put:

$$PR_t(A, \Delta) =_{df} PR(A, t, \Delta)$$

The first five axioms (I*) - (V*) in this setting are the obvious generalizations of our previous axioms (I) - (V). Namely, we replace 'PR' by 'PR_t' and require that the condition hold for all t in T .

Theorem 1 generalizes in the obvious way and we get, for each $t \in T$ and each $x \in W$, a probability $p_{t,x}$ which determines PR_t ; it represents the maximal state of knowledge at stage t in world x .

The "correct" generalization of Axiom (VI) is not as obvious, but is not difficult to find:

Axiom (VI*) For each $t \in T$ the following holds: If C is a finite intersection of events of the form $PR_s(B, \Delta)$ where every s is $\leq t$, and $P(C \cap PR_t(A, [\alpha, \beta])) \neq 0$, then

$$\alpha \leq P(A | C \cap PR_t(A, [\alpha, \beta])) \leq \beta$$

The argument for this axiom is the same as the argument for Axiom (VI). The essential point is that if $s \leq t$ then true events of the form $PR_s(B, \Delta)$ are known at stage t . The same Dutch book argument works for Axiom (VI*).

As before, we consider fields generated by knowable events and define epistemic equivalence; but now these concepts depend on the stage parameter, to be displayed here as an additional subscript. Thus we put:

$$x \simeq_t y \iff_{df} p_{t,x} = p_{t,y}$$

Then $x \simeq_t y$ iff $x \in A \iff y \in A$, for all $A \in K_t[F]$.

Theorem 6 Assume F to be countably generated, then Axiom (VI*) is equivalent to the conjunction of:

(D) For each $t \in T$ the simple HOP (W, F, P, PR_t) satisfies Axiom (VI).

and

(E) For each $s \leq t$, $x \approx_s y \Rightarrow x \approx_t y$, for almost all x, y (i.e., for all $x, y \in W'$ where $P(W') = 1$).

(E) means that, as we pass to more progressive stages, almost everywhere epistemic equivalence is the same or becomes stronger; the partition into equivalence classes can change only by becoming more refined.

Like Theorem 2 the last theorem has a version that applies to general fields but I shall not enter here into it. In the following theorem F is assumed to be countably generated.

Theorem 7 Assume Axiom (VI*). Let $s \leq t$. Then, for almost all x , $p_{s,x}$ is a mixture of $p_{t,y}$'s (where y ranges over W). Consequently, for almost all x , $(W, F, p_{s,x}, PR_t)$ is a simple HOP satisfying Axiom (VI).

Logic of HOPs and Stage Dependent Modalities

Fix a partially ordered set $T = (T, <)$. The logic $PRL_{E,T}$ (which corresponds to HOPs with set of stages T) is defined in the same way as PRL_E , except that PR has an additional argument ranging over T . As before we employ a systematically ambiguous notation. Define ϕ to be p-valid if it gets probability 1 in all HOPs in which the set of stages is T .

Now consider a propositional modal language, M_T , in which we have, instead of a single necessity operant, an indexed family $N_t, t \in T$. $N_t\phi$ states that ϕ is necessary at stage t , i.e., necessary by virtue of the maximal knowledge available at that stage.

For $\phi \in M_T$, let ϕ^* be the wff obtained by replacing each $N_t\psi$ by $PR_t(\psi, [1,1])$. It can be shown that the set of all ϕ in M_T such that ϕ^* is p-valid is exactly the set of wffs derivable, by modus ponens and the rule: if $\vdash \phi$ then $\vdash N_t\phi$, from the following axioms:

- (i) All tautologies. (ii) For each $t \in T$, the axiom schemas of S5, with N replaced by N_t , and
- (iii) $N_s\phi \rightarrow N_t\phi$, for each $s \leq t$.

Note that (iii) accords well with the intended meaning of the N_t 's: If something is necessary at stage s it is also necessary at later stages. On the other hand, something not necessary at stage s can be necessary later.

REFERENCES

- Domotor, Z. 1981 *Higher Order Probabilities* (Manuscript).
- Gaifman, H. 1983 *Towards a Unified Concept of Probability* Manuscript of invited lecture to the 1983 International Congress for Logic Philosophy and Methodology of Science, Salzburg. To appear in the proceedings of the congress, North Holland.
- Gärdenfors, P. 1975
Qualitative Probability as Intentional Logic Journal of Philosophical Logic 4, pp.171-185.
- Good, I.J. 1965 *The Estimation of Probabilities*, Cambridge Massachusets.
- Jaynes, E.T. 1958 *Probability Theory in Science and Engineering*, Dallas.
- Kolmogoroff, A.N. 1933
Grundbegriffe der Wahrscheinlichkeit Ergebnisse der Mathematik und ihrer Grenzgebiete, No 2, Berlin.
- Miller, D. 1966 *A Paradox of Information* British Journal for the Philosophy of Science 17.
- Skyrms, B. 1980A *Higher Order Degrees of Belief* in "Prospects for Pragmatism" Essays in honor of F.P. Ramsey, ed. D.H. Mellor.
- Skyrms, B. 1980B *Causal Necessity (Appendix 2)*, Yale New-Haven Conn.
- van Fraassen, B. 1984
Belief and the Will Journal of Philosophy, No 81 pp. 235-256.

The list is far from being complete. Some important papers not mentioned in the abstract are to be found in Iffs, W. Harper ed. Boston Reidel, 1981. Material related in a less direct way: non-probabilistic measures of certainty (e.g., the Dempster-Shafer measure), expert systems involving reasoning with uncertainties, probabilistic protocols and distributed systems, has not been included, but should be included in a fuller exposition.