Brian Skyrms

Three Ways to Give a Probability Assignment a Memory

Consider a model of learning in which we update our probability assignments by conditionalization; i.e., upon learning S, the probability of not-S is set at zero and the probabilities of statements entailing S are increased by a factor of one over the initial probability of S. In such a model, there is a certain peculiar sense in which we lose information every time we learn something. That is, we lose information concerning the initial relative probabilities of statements not entailing S.

The loss makes itself felt in various ways. Suppose that learning is meant to be corrigible. After conditionalizing on S, one might wish to be able to decide that this was an error and "deconditionalize." This is impossible if the requisite information has been lost. The missing information may also have other theoretical uses; e.g., in giving an account of the warranted assertability of subjunctive conditionals (Adams 1975, 1976; Skyrms 1980, 1981) or in giving an explication of "evidence E supports hypothesis H" (see the "paradox of old evidence" in Glymour 1980).

It is therefore of some interest to consider the ways in which probability assignments can be given a memory. Here are three of them.

I. *Make Like an Ordinal* (Tait's Suggestion)[1]: A probability assignment will now assign each proposition (measurable set) an ordered pair instead of a single number. The second member of the ordered pair will be the probability; the first member will be the memory. To make the memory work properly, we augment the rule of conditionalization. Upon learning P, we put the current assignment into memory, and put the result of conditionalizing on P as the second component of the ordered pairs in the new distribution. That is, if the pair assigned to a proposition by the initial distribution is (x, y), then the pair assigned by the final distribution is ((x, y), z), where z is the final probability of that proposition gotten by conditionalizing on P. (If P has initial probability zero, we go to the closest state in which it has positive probability to determine the ratios of final probabilities for propositions that entail P.)

This suggestion gives probability assignments a perfect memory. From a practical viewpoint, the price that is paid consists in the enormous amount of detail that is built into an assignment for a relatively old learning system of this kind, and the consequent costs in terms of capacity of the system.

II. *Don't Quite Conditionalize* (Probability Kinematics with or without Infinitesimals): Upon learning P, one might not quite give P probability one, but instead retain an itty-bitty portion of probability for its negation, distributing that portion among the propositions that entail not-P in proportion to their prior probabilities. There are two versions of this strategy, depending on whether the itty-bitty portion is a positive real magnitude or an infinitesimal one. Let us consider the first alternative. It should be noted that this may simply be a more realistic model of learning for some circumstances. But I am concerned here only with its value as a memory device. As such it has certain drawbacks. In the first place, the itty-bitty probabilities used as memory might be hard to distinguish from genuinely small current probabilities. In the second place, we get at best short-term memory. After a few learning episodes where we learn P_1 . . . P_n, the information as to the relative initial values of propositions that entailed not-P_1 & . . . & not-P_n is hopelessly lost. On the other hand, we have used no machinery over and above the probability assignment. This gives us a cheap, dirty, short-term memory.

The drawbacks disappear if we make the itty-bitty portion infinitesimal. The development of nonstandard analysis allows us to pursue this possibility in good mathematical conscience.[2] There is no danger of confusing an infinitesimal with a standard number. Furthermore, we can utilize *orders* of infinitesimals to implement long term-memory. (Two nonstandard reals are of the same *order* if their quotient is finite.) There is some arbitrariness about how to proceed, because there is no largest order of infinitesimals. (But, of course, arbitrariness is already present in the choice of a nonstandard model of analysis). Pick some order of infinitesimals to function as the largest working order. Pick some infinitesimal i of that order. On learning P, we update by probability kinematics on P; not-P, giving not-P final probability i. Successive updatings do not destroy information, but instead push it down to smaller orders of infinitesimals. For instance, if we now learn Q, the information as to the relative magnitude of the initial probabilities of propositions that entail not-P & not-Q lives in infinitesimals of the order i^2.

This strategy of probability kinematics with infinitesimals gives probability distributions a memory that is almost as good as that supplied by Tait's suggestion.[3] It has the advantage of a certain theoretical simplicity. Again it is only the probability assignment that is doing the work. Memory is implicit rather than something tacked on. This theoretical simplicity is bought at the price of taking the range of the probability function to be non-Archimedian in a way that reduces consideration of the practical exemplification of the model to the status of a joke.

III. *Keep a Diary:* Our system could start with a given probability distribution, and instead of continually updating, simply keep track of what it has learned. At any stage of the game its current probability distribution will be encoded as a pair whose first member is the original prior distribution, and whose second member is the total evidence to date. If it needs a current probability, it computes it by conditionalization on its total evidence. Such a *Carnapian* system has its memory structured in a way that makes error correction a particularly simple process—one simply deletes from the total evidence.

Information storage capacity is still a problem for an old Carnapian robot, although the problem is certainly no worse than on the two preceeding suggestions. Another problem is choice of the appropriate prior, providing we do not believe that rationality dictates a unique choice.

In certain tractible cases, however, this storage problem is greatly simplified by the existence of *sufficient statistics*.[4] Suppose I am observing a Bernoulli process, e.g., a series of independent flips of a coin with unknown bias. Each experiment or "observation" will consist of recording the outcome of some finite number of flips. Now instead of writing down the whole outcome sequence for each experiment, I can summarize the experiment by writing down (1) the number of trials and (2) the number of heads observed. The ordered pair of (1) and (2) is a sufficient statistic for the experiment. Conditioning on this summary of the experiment is guaranteed to give you the same results as conditioning on the full description of the experiment. Where we have sufficient statistics, we can save on memory capacity by relying on statistical summaries of experiments rather than exhaustive descriptions of them.

In our example, we can do even better. Instead of writing down an ordered pair (x, y) for each trial, we can summarize a totality of n trials by writing down a single ordered pair.

$$\left(\sum_{i=1}^{i=n} x_i, \sum_{i=1}^{i=n} y_i\right).$$

We have here a *sufficient statistic of fixed dimension*, which can be gotten by component-by-component addition from sufficient statistics for the individual experiments. (Such sufficient statistics of fixed dimension can be shown—under certain regularity conditions—to exist if and only if the common density of the individual outcomes is of exponential form.) Where such sufficient statistics exist, we need only store one vector of fixed dimension as a summary of our evidence.

The existence of sufficient statistics of fixed dimension can also throw some light on the other problem, the choice of an appropriate prior. In our example, the prior can be represented as a probability distribution over the bias of the coin; the actual physical probability of heads. Denote this parameter by "w." Suppose that its prior distribution is a beta distribution; i.e., for some α and β greater than zero, the prior probability density is proportional to $w^{\alpha-1}(1 - w)^{\beta-1}$. Then the posterior distribution of w will also be a beta distribution. Furthermore, the posterior distribution of w depends on the prior distribution and the summary of the evidence in an exceptionally simple way. Remembering that x is the number of trials and y is the number of heads, we see that the posterior beta distribution has parameters α' and β', where $\alpha' = \alpha + y$ and $\beta' = \beta + x - y$. The family of beta distributions is called a *conjugate* family of priors for random samples from a Bernoulli distribution. It can be shown that whenever the observations are drawn from a family of distributions for which there is a sufficient statistic of fixed dimension, there exists a corresponding family of conjugate priors. (There are conjugate priors for familiar and ubiquitous distributions such as Poisson, Normal, etc.) Random sampling, where the observation is drawn from an exponential family and where the prior is a member of the conjugate family, offers the ultimate simplification in data storage and data processing. The diary need only include the family of priors, the parameters of the prior, and the current value of the sufficient statistic of fixed dimension. For these reasons, Raiffa and Schlaifer (1961) recommend, in the case of vague knowledge of priors, to *choose* a member of the relevant family of conjugate priors that fits reasonably well.

We are not always in such a nice situation, where sufficient statistics do so much work for us; but the range of cases covered or approximated by the exponential families is not inconsiderable. In these cases, keeping a diary (in shorthand) is not as hopeless a strategy for a quasi-Carnapian robot as it might first appear.

One might wonder whether these techniques of diary-keeping have some application to an Austinian robot which, on principled grounds, never learns anything with probability one. I think that they do, but this question goes outside the scope of this note. (See Field 1978, Skyrms 1980b, and Skyrms forthcoming).

Notes

[1] Proposed by Bill Tait in conversation with myself and Brian Ellis.

[2] For a thumbnail sketch see appendix 4 of Skyrms (1980a). For details see the references listed there.

[3] There is this difference. Suppose we think that we learn P, but then decide it was a mistake and in fact learn not-P. On the infinitesimal approach traces of the "mistake" are wiped out, while on Tait's suggestion they remain on the record.

[4] On the Bayesian conception of sufficiency, sufficient statistics of fixed dimension, and conjugate priors, see Raiffa and Schlaifer (1961).

References

Adams, E. 1975. *The Logic of Conditionals*. Dordrecht, Reidel.

—. 1976. Prior Probabilities and Counterfactual Conditionals. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, ed. W. Harper and C. Hooker. Dordrecht: D. Reidel.

Field H. 1978. A Note on Jeffrey Conditionalization. *Philosophy of Science* 45: 171-85.

Glymour C. 1980. *Theory and Evidence*. Princeton: Princeton University Press.

Raiffa, H. and Schlaifer, R. 1961. *Applied Statistical Decision Theory*. Cambridge, Mass.: Harvard Business School. Paperback ed. Cambridge, Mass.: MIT Press, 1968.

Skyrms, B. 1980a. *Causal Necessity*. New Haven, Conn.: Yale University Press.

Skyrms, B. 1980b. Higher Order Degrees of Belief. In *Prospects for Pragmatism*, ed. D. H. Mellor. Cambridge, England: Cambridge University Press.

Skyrms, B. 1981. The Prior Propensity Account of Subjunctive Conditionals. In *Ifs*, ed. W. Harper, R. Stalnaker, and G. Pearce. Dordrecht: Reidel.

Skyrms, B. Forthcoming. "Maximum Entropy Inference as a Special Case of Conditionalization." *Synthèse*.