Conditioning and the Sure-Thing Principle[*]

Robert J. Aumann[†] Sergiu Hart[‡] Motty Perry[§]

November 5, 2006

Abstract

This paper undertakes a careful examination of the concept of conditional probability and its use. The ideas are then applied to resolve a conceptual puzzle related to Savage's "Sure-Thing Principle."

1 Introduction

This paper undertakes a careful examination of the concept of conditional probability and its use (Section 3). The conclusions (Sections 4 and 5), while perhaps obvious once pointed out, have heretofore not been universally appreciated. The ideas are applied to resolve (Section 6) a conceptual puzzle (Section 2) related to Savage's "Sure-Thing Principle" (STP). We conclude (Section 7) by emphasizing that STP is a principle of rationality, *not* a logical principle.

[*] Research partially supported by grants of the Israel Science Foundation.

[†] Center for the Study of Rationality, and Institute of Mathematics, The Hebrew University of Jerusalem. *E-mail*: `raumann@math.huji.ac.il` *Web page*: `http://www.ma.huji.ac.il/~raumann`

[‡] Center for the Study of Rationality, Department of Economics, and Institute of Mathematics, The Hebrew University of Jerusalem. *E-mail*: `hart@huji.ac.il` *Web page*: `http://www.ma.huji.ac.il/hart`

[§] Center for the Study of Rationality, and Department of Economics, The Hebrew University of Jerusalem. *E-mail*: `motty@math.huji.ac.il` *Web page*: `http://economics.huji.ac.il/facultye/perry/cv.html`

2 The Sure-Thing Puzzle

The *Sure-Thing Principle* (Savage 1954) says that if a decision maker would take a certain action if he knew that an event E obtained, and also if he knew that its negation ˜E obtained, then he should take that action even if he knows nothing about E. Savage illustrates this as follows: "A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant. So, to clarify the matter to himself, he asks whether he would buy if he knew that the Democratic candidate were going to lose, and decides that he would. Similarly, he considers whether he would buy if he knew that the Republican candidate were going to lose, and again finds that he would. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event will obtain. ... It is all too seldom that a decision can be arrived at on the basis of this principle, but I know of almost no other extralogical principle governing decisions that finds such ready acceptance" (p. 21; slightly paraphrased). Indeed, in spite of attacks that have ensued throughout the years, the sure-thing principle does sound very compelling.

Consider now a similar story, differing from the above only in that there are *three* serious candidates—as indeed happened in 1912, when Theodore Roosevelt ran as an independent, against a Republican (Taft) and a Democrat (Wilson). Again, the businessman asks himself whether he would buy if he knew that the Democratic candidate was going to lose, and decides that he would; and he would also buy if he knew that the Republican candidate was going to lose. One of these two events will surely obtain, so as above, it would appear that he should buy even though he does not know which event will obtain.

But a closer look reveals that in this case, the reasoning is fallacious. Suppose the businessman assigns respective prior probabilities of $2/7, 3/7$, and $2/7$ to the Republican, Roosevelt, and the Democrat winning; and, that he believes that Roosevelt's progressive economic policies, if implemented, will lead to a period of prosperity that will make the value of the property rise dramatically. He decides to buy if and only if he thinks that Roosevelt's chances exceed 1/2. If he then learns that the Democrat will lose, his probability for Roosevelt winning goes to 3/5, so he buys. Similarly if he learns that the Republican will lose. But if he learns nothing, then his probability for Roosevelt winning remains at 3/7, so he does not buy.

To be sure, this situation is not quite the same as Savage's, because here, unlike above, *both* the Democrat *and* the Republican might lose; Roosevelt might win. But that seems irrelevant; on its face, the reasoning leading to the conclusion appears no less compelling than before.

Nevertheless, Savage's principle—when the events in question are disjoint—*does* seem compelling, whereas the example shows that when the events have a non-empty intersection, it is not. Why? What is the role of disjointness—when and why is it important? Can we formulate and justify a principle that will distinguish between these two cases?

3 Conditioning: The Question

For the moment, let us put aside the Sure-Thing Puzzle, and turn to a careful examination of conditional probabilities.

Suppose that people who test positive for HIV have a five-year survival probability[1] of, say,[2] 40%; in technical terminology, the five-year survival probability, *conditional* on testing positive for HIV, is 40%. Suppose further that your friend tells you that he just tested positive for HIV. What is your probability for his being alive in five years?

The answer that comes to mind is, "Well, 40%, of course; isn't that what you just told me? Isn't that what 'conditional probability' means?" But a closer examination casts doubt on this answer. You know *more* than that your friend tested positive for HIV; you know that he *told* you[3] so. That could signify, for example, that his situation is worse than that of the average HIV-

[1] For example, this would be the case if 40% of all people who test positive for HIV are alive five years after the test, and no significant medical advances are expected.

[2] This number is purely hypothetical, for illustration only; the authors have no idea what the true probability is.

[3] Here we rule out the possibility that your friend lied. If this is admitted, the situation becomes even more complex.

positive testee, that he already has symptoms of AIDS, and that he needs your support here and now. Alternatively, it could conceivably signify other things, in the opposite direction. Either way, the simple answer of 40% is likely to be incorrect.

For another example, we turn to the celebrated text of Hodges and Lehman (1964), who illustrate the concept of conditional probability as follows: "Suppose a poker player happens, by accident, to catch a glimpse of the hand dealt to an opponent. The glimpse is too fleeting for individual cards to be distinguishable, but the player does perceive that all the cards are red. It is then certain that the opponent cannot have 'four of a kind,' as that would require him to have at least two black cards, and one also feels that the chance of the opponent holding a 'heart flush' is now higher than it was before the new information was obtained. Denoting by R the set of all poker hands formed of red cards only, what is now of interest is the conditional probability[4], given that R has occurred" (pp. 77–78; abbreviated and slightly paraphrased).

Again, that is wrong. You know more than that the opponent's hand is in R; you know that you caught a fleeting glimpse of his hand, just enough to allow you to conclude that his hand is in R. That could signify that he was careless because his hand was hopeless, or that he purposely allowed you to glimpse his hand, perhaps to discourage you from bidding against him, or numerous other possible reasons. In any event, it need *not* be the case that your probabilities are the original probabilities, conditioned on R.

The reader may now begin to feel uncomfortable. "Come," he may say, "you are quibbling. Whenever you get information, you must get it in some way. The formal theory abstracts away from such practical matters, and conditions only on the information itself, not on the way it was obtained. In practical applications, one must modify the conclusions to take these considerations into account, but the formal theory does not and cannot do so."

But that will not pass muster. There is no distinction between "information" and the way it is obtained. *The way it is obtained is part of the information*; ignoring it may greatly affect the probabilities, as we have seen.

[4] The original text here has the word "frequency."

How can we sort this out? Specifically, on what should one condition?

4 Conditioning: The Answer

Suppose you learn that an event E obtains. The *way* in which you learn it is called a *signal,* denoted s. In the HIV example, E is that your friend tested positive for HIV, and s is that he told you so. In the Hodges–Lehman poker example, E is that all your opponent's cards are red, and s is that by accident, you caught a fleeting glimpse of his hand and perceived that all the cards are red. Having received the signal s, you conclude that E obtains; s is a sufficient condition for E.

As we have seen, receiving the signal s does *not* in general justify conditioning on E. What, then, does? The answer is that to justify conditioning on E, the signal s must be necessary and sufficient for E.

Though s is sufficient for E in both our examples, it is necessary in neither. If your friend told you that he tested positive, then indeed he tested positive; but he could also have tested positive without telling you so. If you caught a glimpse of your opponent's hand and perceived it to be all red, then indeed it is all red; but it could also be all red without your catching a glimpse of it.

Signals that *are* necessary and sufficient for the corresponding events are quite common. If a lab technician routinely testing blood specimens finds a certain specimen to be HIV positive, then his probability is 40% that the person from whom the specimen was taken will live five years. The reason is that if the specimen had been negative, the lab technician would have known that, as well. More generally, the signal is likely to be necessary and sufficient if it is the result of a test that is applied indiscriminately, under all circumstances—as distinguished from information that one may or may not get.

When the signal s is only sufficient for E, but not necessary, then though one knows E, one cannot condition on E. Rather, one conditions on having received the signal s—i.e., on *everything* one knows. Of course this includes E, but it is not limited to E*:* since the signal is not necessary for E, one knows more than just E. This matter will be clarified in the next section, where we provide a formal treatment.

5 Formal Treatment

The standard epistemologic model (e.g., Aumann and Heifetz 2002) consists of a finite[5] set Ω, whose members are called *states of the world* or simply *states,* and whose subsets are called *events;* a probability measure P on Ω; and a partition $\mathcal{K}$ of Ω, whose atoms are called *information sets.* Here $\mathcal{K}$ represents the decision maker's information: if the true state of the world is ω, then he does not know that, but knows only that the true state is in the information set $\mathbf{K}(\omega) \in \mathcal{K}$ to which ω belongs; we say that the decision maker considers each member ν of $\mathbf{K}(\omega)$ to be *possible at* ω. The measure P is the decision maker's *prior* probability estimate on the state of the world—before he gets his information $\mathbf{K}(\omega)$.

In this formalism, conditioning appears as follows: If the true state of the world is ω, then after the information is received, the decision maker assigns probability 0 to states outside of $\mathbf{K}(\omega)$, and probability $\mathrm{P}(\{\nu\})/\mathrm{P}(\mathbf{K}(\omega))$ to states ν in $\mathbf{K}(\omega)$; that is, he assigns probability $\mathrm{P}(A \,|\, \mathbf{K}(\omega)) \equiv \mathrm{P}(A \cap \mathbf{K}(\omega))/\mathrm{P}(\mathbf{K}(\omega))$ to any event A. This is the *conditional probability of* A, *given* $\mathbf{K}(\omega)$.

It is often convenient to use the equivalent representation of information by "signals." A *signalling function* $\mathbf{s}$ is a function on Ω: if ω is the true state of the world, then the decision maker receives[6] the signal $\mathbf{s}(\omega)$. The information consists of the received signal. Therefore the information partition $\mathcal{K}$ is generated by $\mathbf{s}$; i.e., the atoms of $\mathcal{K}$ are the events $\mathbf{s}^{-1}(s) = \{\omega : \mathbf{s}(\omega) = s\}$ for each possible signal s in the range of the function $\mathbf{s}$. Indeed, if the decision maker considers ν to be possible at ω, then the signal $\mathbf{s}(\nu)$ at ν must be the same as the signal $\mathbf{s}(\omega)$ at ω (if the signals were different, they would distinguish between ν and ω). Conditioning on a specific signal s is thus

[5] Finiteness is assumed for simplicity only.

[6] It is possible, of course, that in some states no signal at all is received; but we shall simply call this the "null signal," and so, formally, get a signal for every state.

precisely the same as conditioning on the information set: at each state ω where $\mathbf{s}(\omega) = s$ we have $\mathbf{K}(\omega) = \mathbf{s}^{-1}(s)$, so $\mathrm{P}(A \,|\, \mathbf{K}(\omega)) = \mathrm{P}(A \,|\, \mathbf{s}^{-1}(s))$.

Let E be a distinguished event. In the examples above, E is the event that your friend tested positive for HIV, or your opponent holds a poker hand that is all red. The actual signal s that was received (your friend telling you that he tested positive, or you seeing that all the cards are red) is sufficient for the event E, but not necessary. In our formalism, that translates to $\mathbf{s}^{-1}(s) \subsetneqq E$; one computes $\mathrm{P}(A \,|\, \mathbf{s}^{-1}(s))$—*not* $\mathrm{P}(A \,|\, E)$.

To elaborate further, when there is just one way to know E—i.e., *only one* signal s such that $\mathbf{s}^{-1}(s) \subset E$—then $\mathbf{s}^{-1}(s) = K(E)$, which is the event of "knowing E"; in this case we condition on $K(E)$ rather than on E. If there is more than one way to know E—i.e., $\mathbf{s}^{-1}(s_i) \subset E$ for $i = 1, 2, ...$ (if your friend does not tell you, his spouse might)—then the conditioning is done on each $\mathbf{s}^{-1}(s_i)$ separately, according to the signal s_i actually received (in this case $K(E) = \cup_i \mathbf{s}^{-1}(s_i)$).

Summing up: To condition on an event, it is not enough simply to know that it obtains. One can condition only on one's information set—the set of all states that one considers possible, no more and no less. Equivalently, in terms of signals, once a signalling function is defined—i.e., the precise signal received is defined for *all* states—one conditions on the event that *one particular* signal is observed.[7]

6 Resolving the Sure-Thing Puzzle

At the outset, note that in Savage's formal treatment, the Sure-Thing Principle is not an axiom; it is derived from other, more fundamental postulates (1954, p. 24, Theorem 2). The current discussion refers not to this formal derivation, but rather to the principle considered on its own merits, as discussed by Savage in the passage cited at the beginning of Section 2 above.

We come now to the puzzle itself. Conditional decisions are analogous

[7] What needs to be specified precisely is in what states the signal s is received and in what states it is not—i.e., what is the set $\mathbf{s}^{-1}(s)$. In brief, the model must be complete and coherent.

to conditional probabilities: one must condition not on an event occurring, or even on knowing that the event occurs, but on *all* one's information, no more and no less.

The sure-thing principle is then based on the following, more fundamental principle: Suppose we have a process whereby a decision maker gathers information. Suppose that under *all* circumstances, the information that the decision-maker will obtain will lead him to take a certain decision d. Then he may decide on d without obtaining the information.[8]

In the three-candidate version of Savage's example (Section 2), this reasoning does not apply. It is true that either the Republican or the Democrat must lose; and after the elections, the businessman will know one of these two eventualities. But it is *not* true that after the elections, the businessman will know that the Republican lost, no more and no less, or will know that the Democrat lost, no more and no less.

For instance, if he listens to the radio, he will know exactly which candidate won; and depending on this, he will *not* invariably buy. Even if he is told either that the Republican lost (RL) or that the Democrat lost (DL), and nothing else, he will in fact know more. He will know that he was *told* RL or *told* DL. It is not possible that he is told RL if and only if RL occurs *and* that he is told DL if and only if DL occurs. Either the event *"RL"* of being told RL is strictly contained in the event RL, or the event *"DL"* of being told DL is strictly contained in DL, or both; the events "RL" and "DL" *are disjoint*, whereas the events RL and DL are not (see Figure 1).

Suppose, for example, that if Roosevelt wins, the businessman is told RL or DL with half-half probabilities. If he knows this, his conditional probability that Roosevelt won is in either case 3/7, so he will not buy the property. And if he does not know this, what exactly does he know? Without specifying what the businessman knows as a function of the circumstances, the model is incoherent.

In the two-candidate example, he can buy without knowing the outcome of the election. But in the three-candidate example, there is no way to set up the information-gathering process so that the businessman will buy under

[8] This is the formulation sought in the last sentence of Section 2.

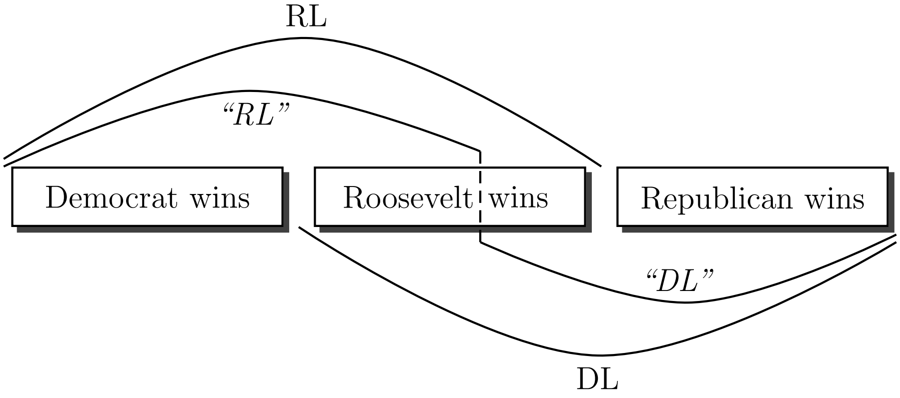

Figure 1: The three-candidates example

all circumstances.

Formally, an information-gathering process is modelled as a "signalling function." Thus in terms of signals, the sure-thing principle may be formulated as follows: *Given a signalling function, if the decision maker makes the same decision no matter what signal he gets, then he can make that same decision without getting any signal.*

7 The Logical Sure-Thing Principle

Suppose we wish to establish a proposition p. If q is another proposition, and we show that q entails p, and also that "not q" entails p, then p follows. Let us call this the *Logical Sure-Thing Principle* (LSTP); it is often used in mathematics, as well as in other applications of logic.

On the face of it, Savage's Sure-Thing Principle (STP) appears similar. But in fact, they are quite different. For one thing, LSTP is a logical truism; it follows from fundamental principles of logical reasoning. Specifically, it is a theorem of the propositional calculus. In contrast, STP is a desideratum of *rational* behavior, but it is not *logically* necessary. A decision maker who violates STP may be acting irrationally, but he is not acting illogically. Savage himself (1954, p. 21) refers to STP as an "*extra*logical principle."

For another thing, while disjointness is essential for STP, it is not needed for LSTP. Indeed, if p entails q, and p' entails q, then $p \vee p'$ entails q—whether p and p' are compatible or not. Equivalently, in set-theoretic terms, if $P \subset Q$ and $P' \subset Q$, then $P \cup P' \subset Q$—whether P and P' are disjoint or not. We suspect that one reason that people mistakenly extend STP to the non-disjoint case is that for LSTP this extension is legitimate.

To see that STP is not a logical truism, suppose that you consider reading a certain book. If you know it is written by A, you will read it, and if you know it is by B, you will read it. But that does not entail that you will read it if you know that it is by either A or B. Knowing the identity of the writer is important for appreciating the book (for instance, it brings to mind associations to other works by the same author).

To be sure, this is a little unusual, because the quality of the consequence (the reading of the book) is directly affected by the decision maker's knowledge—rather than the usual case, in which the knowledge only affects his estimate of the likelihood that this or that state obtains. The knowledge is important for its own sake: what matters is not *what* you know, but *that* you know it.

When that happens, Savage's STP indeed need not apply. But when the knowledge does not affect the consequence, as in most decision problems, STP does sound eminently reasonable.

References

Aumann, R. J. and A. Heifetz (2002), "Incomplete Information," in *Handbook of Game Theory, with Economic Applications,* Volume 3, edited by R. J. Aumann and S. Hart, Elsevier, Amsterdam, 2002, pp. 1665–1686.

Hodges, J. L., Jr., and E. L. Lehman (1964), *Basic Concepts of Probability and Statistics,* San Francisco: Holden-Day.

Savage, L. J. (1954), *The Foundations of Statistics,* New York: John Wiley.