THE REVIEW OF SYMBOLIC LOGIC
Volume 1, Number 1, June 2008

PROBABILISTIC CONDITIONALS ARE ALMOST MONOTONIC

MATTHEW P. JOHNSON AND ROHIT PARIKH
City University of New York

Abstract. One interpretation of the conditional *If P then Q* is as saying that the probability of Q given P is high. This is an interpretation suggested by Adams (1966) and pursued more recently by Edgington (1995). Of course, this probabilistic conditional is nonmonotonic, that is, if the probability of Q given P is high, and R implies P, it need not follow that the probability of Q given R is high. If we were confident of concluding Q from the fact that we knew P, and we have stronger information R, we can no longer be confident of Q. We show nonetheless that *usually* we would *still be justified* in concluding Q from R. In other words, probabilistic conditionals are mostly monotonic.

1. Probabilistic conditionals. How should we interpret a conditional like 'If John comes to the party, then so will Mary,' that is, of the form *If A then B*? The standard interpretation used in mathematics is to treat it as equivalent to $A \to B$, that is, as $\neg A \vee B$ (see Sanford, 2003; Parikh, 2006).

But often, this does not fit our intuition. The following example due to Dorothy Edgington is instructive. Clearly if God does not exist, then he cannot answer our prayers. Consider the statement S:

> If God does not exist, then it is not the case that if I pray, my prayers will be answered.

Many who disagree about the existence of God will tend to accept S.

There are 2 *if*s in S. Can they both be expressed as material conditionals? Suppose we symbolize S as $\neg G \to \neg(P \to A)$, interpreting both implications, the main one on the outside and the subsidiary one inside the parentheses, as material implications. Also suppose I do not pray. Then, P is false and $P \to A$ is true. Hence, $\neg(P \to A)$ is false. But then for S to be true, $\neg G$ must be false and hence G must be true. *I can prove the existence of God simply by not praying!*

Even those who believe in God will find this argument fishy and will look more kindly on other ways, *beside the material conditional*, of interpreting the indicative conditional.[1]

One suggestion, associated with Adams (1966) and Edgington (1995) herself, is to treat the conditional probabilistically. Thus, asserting *If A then B* is tantamount to saying that the probability of B given A is high. Someone who says 'If John comes to the party, Mary will come too' is asserting that $p(M|J)$ is high, perhaps more than 0.90.

Received xxxxx, 200x.
Research supported in part by a grant from the PSC-CUNY grants program. Versions of this paper were given in colloquia at Boston University, City University of New York Graduate Center, and Institute for the History and Philosophy of Science and Technology in Paris. A preliminary version of this paper appeared as Parikh (2007).

[1] This particular variety of conditional is called the *indicative conditional* to distinguish it from the so-called subjunctives or counterfactuals.

doi:10.1017/S1755020308080106

Such probabilistic conditionals received a blow from the results of Lewis (1976), which showed that (on pain of triviality) such conditionals cannot be interpreted as *propositions*. In other words, there cannot be a proposition (a set of possible worlds) C such that $p(C) = p(B|A)$ for all probability measures p.

2. Near monotonicity: Finite setting. But let us leave that worry aside and ask about the *logic* of the probabilistic conditional interpreted not as an implication (i.e., as a connective) but instead as a consequence relation $\mid\!\sim$. This avoids the Lewis problem because we are *not saying* that *If A then B* is a proposition. Let us say that we accept $A \mid\!\sim B$ if $p(B|A)$ is high, where $\mid\!\sim$ represents the indicative conditional interpreted probabilistically.

Arlo-Costa and Parikh (2005) looked at probabilistic conditionals in the context of cores,[2] a notion investigated earlier by Fraassen (1995), and they looked at some conditions on nonmonotonic relations considered by Gabbay (1985) and Kraus *et al.* (1990).

Various rules of inference apply to such consequence relations. Thus, from $A \mid\!\sim B$ and $B \models C$, we can derive $A \mid\!\sim C$, where $\models$ represents the classical logical consequence. This rule, called right weakening, is sound. So is the rule (AND) which derives $A \mid\!\sim B \wedge C$ from $A \mid\!\sim B$ and $A \mid\!\sim C$ (provided we make some sacrifice in probability[3]). But a rule that does *not* hold is monotonicity (M) or *strengthening the antecedent*. This would be the rule, *from $A \mid\!\sim B$ and $C \models A$, we should be able to derive $C \mid\!\sim A$*. In particular, we would like to be able to derive $(A \wedge X) \mid\!\sim B$ from $A \mid\!\sim B$.

$$\boxed{(\mathrm{M})\quad \frac{A \mid\!\sim B}{A \wedge X \mid\!\sim B}}$$

Alas, the rule (M) is known not to be sound. The probability of B given A may be high, and the probability of B given A and X may be low. For instance, if our domain is integers up to 100, then the probability that n is odd given that it is prime is quite high. But the probability that it is odd given that it is prime *and* less than 4 is only 0.5.

A well-known example involves birds. Given that Tweety is a bird, there is a high probability that it flies. But given that Tweety is a bird and a penguin, the probability drops to 0.

What we show below is that the rule (M) is *mostly* sound. That is, provided that A is large enough, for *most* X, the conclusion continues to hold.

Probabilistic conditionals are *mostly* monotonic. And this is good news, for clearly, while accepting that the monotonicity condition does not hold *universally*, we do want it to hold *usually*.

For, consider birds. If the dictum ‘Birds fly’ could be destroyed at the drop of a hat, it would be useless. We could not conclude that female birds fly, that blue birds fly, or that the bird sitting on the window sill is likely to fly. It is almost always the case that when we know that some creature is a bird, we also have *some* additional information X. And usually, we do not drop the dictum ‘Birds fly’ when we have some additional information. Thus, it must be the case that the dictum is reasonably sturdy. Information like ‘It is a penguin’ is *unusual*. It is this sturdiness that we will prove below.

[2] Arlo-Costa and Parikh (2005) interpret ‘high’ as 1.

[3] If $p(B|A) > 0.95$ and $p(C|A) > 0.95$, then $p(B \wedge C|A) > 0.9$ since $p(B \wedge C|A) = p(B|A) + p(C|A) - p(B \vee C|A) > 0.95 + 0.95 - 1 = 0.90$.

2.1. A concrete example. We are representing propositions as sets of possible worlds. One of us has already objected this identification (Parikh, 2005). But the representation is commonly accepted and a result that uses it ought to have some relevance. The following theorem is stated rather loosely but will be followed up by a more precisely stated theorem.

Theorem 1. *Suppose that W is a finite space with all points equally likely. Suppose that A and B are sufficiently large subsets of W and $p(B|A) \geq 0.95$. Then, for most randomly selected $X \subseteq W$, $p(B \cap X|A \cap X) = p(B|A \cap X) > 0.948$.*

Of course the set-theoretic operation $\cap$ corresponds to the logical operation $\wedge$.[4] In terms of $|\!\sim$, it means that if you know $A \,|\!\sim B$ and want to know if $A \wedge X \,|\!\sim B$, the answer will be 'Most likely'.

We have used the numbers 0.95 and 0.948 and the uniform distribution on W for convenience, but of course the result holds more generally, as will be evident from the proof. What 'sufficiently large' B means will be made more explicit below. The technique of proof requires the central limit theorem and the binomial distribution's Gaussian approximation. It is quite accessible.

Proof. We start by noting some simplifications. Since we are taking probabilities relative to A or its subsets, the points in W that are not in A play no role. So we shall assume that $A = W$. This automatically implies that $B \subseteq A$, an assumption that we could have justified independently.

We require the sets to be large, so assume for concreteness that the set $A - B$ has cardinality 10,000 and B has cardinality 190,000, in which case $p(B|A) = 0.95$. A random subset X of A has 2 parts: X_B, which is simply $X \cap B$, and the remaining part X_R of X, which is $X \cap (A - B)$.

The expected size of X_B is 95,000 (half of 190,000), but it could be more or less. But by the central limit theorem (Billingsley, 1995), the standard deviation σ of the size of X_B is $0.5 \times \sqrt{190,000}$, which is approximately 217.9. Thus, $95,000 - 3\sigma$ is more than 94,346. It is unlikely that the actual value differs from the expected value by more than 3σ. Indeed, using standard tables, the probability that X_B has size more than 94,346 exceeds 0.9987. Similarly, the set X_R has expected size 5,000, but the standard deviation σ' is 50. Thus, with the same probability 0.9987, X_R has size less than 5,150. Therefore, with probability greater than 0.9974, the ratio $|X_B|/|X_B \cup X_R|$ is greater than $94,346/(94,346 + 5,150)$, which is 0.9482 or very nearly 0.95. (The figure 0.9974 comes from the fact that even if both errors of 0.0013 $(1 - 0.9987)$ were to add up, we would still only get an error of 0.0026.)

This means that if the sets A and B are both large, and $p(B|A) > 0.95$, then (when a random subset X of A is chosen) with probability greater than 0.9974, we have $p(B \cap X|A \cap X) > 0.9482$. □

We can show that similar results will hold if the random set X is chosen in some other way, for example, if we toss a die for each point of A and put a point in X only if the die shows a 6.

One could ask if the rule (M) can be called sound if it holds only for *most* X. Note, however, that $A \,|\!\sim B$ does *not* say that if A is true, then B is also true 100% of the time. It only says that if A holds, then B is very likely to hold. If the rule applies 99.74% of the

[4] Since our propositions are sets, we will use whichever notation seems more appropriate within a context.

time, and the premise (which we accepted) only ‘applies’ 95% of the time, then it is hard to justify the premise while rejecting the rule.

2.2. The limiting case. We now state a more precise result, which actually generalizes the observation above to the case where the probability β of B relative to A is positive but not necessarily close to 1.[5]

The intuitive idea is that we can think of the set X as a random *sample* from the space W, in which case $X \cap A$ will be a random sample from A. The expected size of $X \cap A$ is half the cardinality $|A|$ of A, and its standard deviation is $0.5 \times \sqrt{|A|}$. The same holds for the expected size of $X \cap B$ except for the multiplier β. Now if the actual sizes of the 2 sets were the *same as* their expected sizes, then we would get $p(B \cap X | A \cap X)$ to be *equal* to $p(B|A)$. Of course we cannot expect to be so lucky in practice, for actual size can deviate from the expected size. As the sizes of A and B go up, however, the deviation matters less and less, and so the difference between $p(B \cap X | A \cap X)$ and $p(B|A)$ will tend to 0. This gives us our second result.

Theorem 2. *Let $\beta > 0$ be fixed, and let sets A_n and B_n increase monotonically in size with $B_n \subseteq A_n$ and $\lim_{n\to\infty} p(B_n|A_n) = \beta$. Let X_n be randomly chosen subsets of A_n and $\varepsilon > 0$. Then, we have*

$$\lim_{n\to\infty} Pr\left[\left|\frac{|B_n \cap X_n|}{|A_n \cap X_n|} - \beta\right| > \varepsilon\right] = 0.$$

In other words, if our prior probability of B given A was $\beta > 0$, and we received additional information X, then provided A was large, we should expect the posterior probability of B given A to still be close to β, and the probability that it differs by more than ε goes to 0 as $|A|$ goes to infinity.

In the sections below, we will look at the case where A is not merely large but is actually infinite. Investigating that case will involve an excursion into measure theory. It will turn out that the theorem above is (roughly) a corollary of the final result we prove below.

Another rule,

$$\boxed{\text{(M)} \quad \frac{A \mathrel{|}\!\sim B}{\neg B \mathrel{|}\!\sim \neg A}}$$

is not capable of a similar treatment. If our universe consists solely of innumerable pigeons and relatively few penguins, then ‘Most birds fly’ will be true but ‘Most non-flyers are non-birds’ will be false. Indeed, all non-flyers will be birds in our universe!

3. Density and measure.

Notation: Let $[n]$ be the set of natural numbers $\{1, 2, \ldots, n\}$, and for a given set X of natural numbers, let $X[n] = X \cap [n]$. We will follow the notation common among number theorists of denoting by $X(n)$ the *number* of elements of X between 1 and n, that is, $X(n) = |X \cap [n]|$.

Definition 1. *The asymptotic density of set X is defined as the limit of its relative frequency: $d(X) = \lim_{n\to\infty} \frac{1}{n} X(n)$.*

Note that asymptotic density is not defined for all sets of natural numbers, since the limit may not exist, and in fact, the set of subsets on which it is defined is not closed under union

[5] We use β for this probability – a number, as we use $p(\cdot)$ for the probability *function*. We use $\Pr[\cdot]$ to indicate probability for a meta-statement.

or intersection. The asymptotic density *can* be extended to obtain a true measure using an ultrafilter. Nonetheless, we will make do with asymptotic density in this work.

A second issue to address as we move to the infinite case is what it means to choose at random a subset of an infinite set. Clearly the probability of selecting any fixed set must be 0. For concreteness, let a random subset S of N be shorthand for a set such that each $S[n]$ is a uniformly random subset of $[n]$ for each n.

We will initially consider $p(B|A)$ when $A = N$ and then generalize to smaller sets A.

4. Near monotonicity: Infinite setting. Let $B \subseteq N$ have well-defined asymptotic density, with $d(B) = \beta > 0$. That is, we assume

$$\lim_{n\to\infty} \frac{B(n)}{n} = \beta > 0. \tag{1}$$

Let $X \subseteq N$ be chosen uniformly at random. We will now argue that we *almost always* have $p(X) = 1/2$, $p(X \cap B) = \beta/2$, and $p(X \cap B|X) = \beta$. Each of these claims will state that a certain property holds with probability 1 over random choices of subset X. Equivalently, the set of such choices X has measure 1. All probabilities using the notation $\Pr[\cdot]$ are over the random choice of X. More formally, we have the following.

Fact 1 (strong law of large numbers). *Let $Y_1, Y_2, \dots$ be a sequence of independently and identically distributed (IID) random variables with $E[Y_i] = \mu$, and let $S_n = \sum_{i=1}^n Y_i$. Then, we have* $\lim_{n\to\infty} S_n/n = \mu$ *with probability 1.*

Fact 2. *Let X be chosen uniformly at random from N, that is, let X_i be 0 or 1 with equal probability, for each $i > 0$. Then, $Pr[d(X) = 1/2] = 1$.*

Proof. This is simply a restatement of Borel's normal number theorem (Billingsley, 1995), that is, that with probability 1:

$$\lim_{n\to\infty} \frac{X(n)}{n} = 1/2. \tag{2}$$

This also follows from the strong law of large numbers. □

Lemma 1. *With probability 1, we have* $\lim_{n\to\infty} \frac{(X\cap B)(n)}{B(n)} = 1/2$. *(Intuitively, this means* $p(X|B) = 1/2$ *with probability 1.)*

Proof. Let $f(i)$ be the index of the ith element of set B. Since B has positive density, it must be infinite. Let $Y_i = X_{f(i)}$. Then, the random variables $Y_1, Y_2, \dots$ are IID with $E[Y_i] = 1/2$, so by the strong law of large numbers, we have with probability 1 that $\lim_{n\to\infty} S_n/n = 1/2$. Now, what roles do S_n and n play?

First, by definition, $S_n = \sum_{i=1}^n X_{f(i)}$ is the number of elements among the first n members of B which are also in X. $f(n)$ is some index number, say m. Then, S_n is the number of numbers from 1 to m that are in both B and X, that is, $S_n = |B \cap X \cap [m]|$. Second, n is the number of elements we are considering from B, that is, the number of elements of B in the range of 1–m, which is $B(m)$. So we have with probability 1 that

$$1/2 = \lim_{n\to\infty} S_n/n = \lim_{m\to\infty} \frac{(B\cap X)(m)}{B(m)}. \tag{3}$$

Therefore, the result follows. □

We can now state our first result of this section, which suggests that learning X will tend to have little effect on our confidence in B (for well-behaved B).

Theorem 3. *If B has well-defined nonzero density $d(B) = \beta$, then $\frac{d(B\cap X)}{d(X)} = \beta$ with probability 1.*

Proof. Since Equations (2) and (3) each hold with probability 1 (over choices of X), their conjunction does as well. We assume that both hold for the remainder of the proof. Equation (1) holds by assumption. Let $\mathrm{Eq}(i)$ indicate the value of the quantities in Equation (i). Then, we have

$$\mathrm{Eq}(3)\cdot\mathrm{Eq}(1)/\mathrm{Eq}(2) = \left(\lim_{n\to\infty}\frac{(B\cap X)(n)}{B(n)}\right)\cdot\frac{(\lim_{n\to\infty} B(n)/n)}{(\lim_{n\to\infty} X(n)/n)}$$

$$= \lim_{n\to\infty}\frac{(B\cap X)(n)}{B(n)}\cdot\frac{B(n)/n}{X(n)/n} = \lim_{n\to\infty}\frac{(B\cap X)(n)}{X(n)} = \frac{d(B\cap X)}{d(X)}.$$

But we also have

$$\mathrm{Eq}(3)\cdot\mathrm{Eq}(1)/\mathrm{Eq}(2) = 1/2\cdot\beta/(1/2) = \beta,$$

which yields the result. □

The same is true for our confidence, upon learning X, in (well-behaved) $B|A$.

Corollary 1. *If B has well-defined nonzero asymptotic density relative to A, that is, $\lim_{n\to\infty}\frac{(A\cap B)(n)}{A(n)} = \beta$, then $\lim_{n\to\infty}\frac{(A\cap B\cap X)(n)}{(A\cap X)(n)} = \beta$ with probability 1.*

Proof. We can again use the trick of renaming indices. First, *zoom in* and rewrite $\lim_{n\to\infty}(A\cap B)(n)/A(n)$ as $\lim_{n\to\infty} B_A(m)/m$, where $B_A = A\cap B$ and $m = A(n)$. We can now apply the theorem to obtain, with probability 1:

$$d(B_A\cap X)/d(X) = \lim_{m\to\infty}(B_A\cap X)(m)/X(m) = \beta. \tag{4}$$

Finally, we *zoom back out* and rewrite the latter expression in Eq(4) as

$$\lim_{m\to\infty}(A\cap B\cap X)(m)/X(m) = \lim_{n\to\infty}(A\cap B\cap X)(n)/(A\cap X)(n). \tag{5}$$

The last equality holds by definition of m. □

The final result shows that indeed, B and A are typically well behaved in the needed way. Intuitively, it shows (respectively) that $p(B|X) = p(B)$ and $p(B|A\cap X) = p(B|A)$, with probability 1.

Corollary 2. *For almost all B, we have $\Pr_X\left[\frac{d(X\cap B)}{d(X)} = d(B)\right] = 1$ and $\Pr_X\left[\frac{d(X\cap B)}{d(X\cap A)} = d(B|A)\right] = 1$.*

Proof. We assumed that set B has well-defined asymptotic density to obtain the theorem. Note from Fact 2, however, that *almost all* subsets of N do have asymptotic density (and in particular have density 1/2). □

5. Open questions.

- We proved the result under the assumption that B has asymptotic density, which is almost always true. The asymptotic density can be extended, using an ultrafilter, to obtain a true measure function. Can a similar result be proven for such a measure? Should it?

- Can the result be extended to an uncountable universe?
- Can the result be formulated so as to apply to a nonuniform distribution for selecting members of X, such as proportional to $1/i^2$? In this case, unlike in the uniform distribution, all finite sets will have nonzero probability and moreover, all sets will be measurable.
- We showed that for almost all propositions B, the result holds. Implicitly this is assuming a uniform distribution on the choice of B. What about other distributions? Can we show that the result holds for *almost all distributions* on the choice of B?

6. Acknowledgments. We thank Evan Goris, Haim Gaifman, Dexter Kozen, Roman Kuznets, Sergiu Hart, Ou Liu, David Makinson, Bud Mishra, and Samer Salame for comments.

BIBLIOGRAPHY

Adams, E. (1966). Probability and the logic of conditionals. In Suppes & Hintikka, editors, *Aspects of Inductive Logic*. Amsterdam: North Holland, pp. 265–316.

Arlo-Costa, H., & Parikh, R. (2005). Conditional probability and defeasible inference. *Journal of Philosophical Logic*, **34**, 97–119.

Billingsley, P. (1995). *Probability and Measure* (third edition). New York: Wiley.

Creighton Buck, R. (1946). The measure theoretic approach to density. *American Journal of Mathematics*, **68**(4), 560–580.

Edgington, D. (1995). On conditionals. *Mind*, **104**, 235–329. (She also has an entry in the *Stanford Encyclopedia of Philosophy.*)

Fraassen, B. (1995). Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic*, **24**, 349–377.

Gabbay, D. (1985). Theoretical foundations for nonmonotonic reasoning in expert systems. In Apt, K. R., editor, *Proceedings of NATO Advanced Study Institute on Logics and Models of Concurrent Systems*. Heidelberg: Springer-Verlag, pp. 439–457.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, **44**, 167–207.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, **85**, 297–315.

Parikh, R. (2005). Logical omniscience and common knowledge: WHAT do we know and what do WE know? In Meyden, R., editor, *Proceedings of the Tenth Conference on Theoretical Aspects of Rationality and Knowledge*. Singapore: National University of Singapore Press, pp. 62–78.

Parikh, R. (2006). Review of Sanford (2003). *Essays in Philosophy*, **7**, 1.

Parikh, R. (2007). Some puzzles about probability and probabilistic conditionals. *Symposium on Logical Foundations of Computer Science*. Heidelberg: Springer-Verlag, 449–456.

Sanford, D. (2003). *If P then Q* (second edition). London and New York: Routledge.

DEPARTMENT OF COMPUTER SCIENCE
CUNY GRADUATE CENTER
NEW YORK, NY 10016, USA
E-mail: mjohnson@cs.cuny.edu

DEPARTMENT OF COMPUTER SCIENCE
CUNY GRADUATE CENTER
NEW YORK, NY 10016, USA
E-mail: rparikh@gc.cuny.edu