Joint Probabilities

John L. Pollock
Department of Philosophy
University of Arizona
Tucson, Arizona 85721
pollock@arizona.edu
http://www.u.arizona.edu/~pollock

Abstract

When combining information from multiple sources and attempting to estimate the probability of a conclusion, we often find ourselves in the position of knowing the probability of the conclusion conditional on each of the individual sources, but we have no direct information about the probability of the conclusion conditional on the combination of sources. The probability calculus provides no way of computing such joint probabilities. This paper introduces a new way of combining probabilistic information to estimate joint probabilities. It is shown that on a particular conception of objective probabilities, clear sense can be made of second-order probabilities (probabilities of probabilities), and these can be related to combinatorial theorems about proportions in finite sets as the sizes of the sets go to infinity. There is a rich mathematical theory consisting of such theorems, and the theorems generate corresponding theorems about second-order probabilities. Among the latter are a number of theorems to the effect that certain inferences from probabilities to probabilities, although not licensed by the probability calculus, have probability 1 of producing correct results. This does not mean that they will always produce correct results, but the set of cases in which the inferences go wrong form a set of measure 0. Among these inferences are some enabling us to reasonably estimate the values of joint probabilities in a wide range of cases. A function called the Y-function is defined. The central theorem is the Y-Theorem, which tells us that if we know the individual probabilities for the different information sources and estimate the joint probability using the Y-function, the second-order probability of getting the right answer is 1. This mathematical result is tested empirically using a simple multi-sensor example. The Y-theorem agrees with Dempster's rule of combination in special cases, but not in general. The paper goes on to investigate cases in which the Y-theorem cannot be expected to give the right answer, and it is shown that there are generalizations of the Y-theorem that can still be employed.

Keywords: joint probability; Dempster Shafer; nomic probability; objective probability

1. Joint Probabilities

A common problem of information fusion occurs when the individual probabilities prob(A/B) and prob(A/C) are known, but what is wanted is the value of the *joint probability* prob(A/B&C). For instance, there might be two imperfect sensors attempting to detect the occurrence of some remote event. There might be data making it possible to assess the reliability of each sensor individually. That is the probability of the event occurring given a positive reading of the sensor. But what if both sensors produce positive readings? One would naturally expect that to raise the probability of the event, but by how much? And what if one sensor produces a positive reading and the other produces a negative reading? The probability calculus is of no help here. If prob(A/B) = r and prob(A/C) = s, where $0 < r,s < 1$, it is consistent with the probability calculus for the value of the joint probability to be anywhere from 0 to 1.

Various strategies have been devised for estimating joint probabilities. Some strategies are based on classical Bayesian approaches to probabilistic reasoning (e.g., [12], [18]). The most popular general approach is probably that based on the Dempster-Shafer theory [31] (see for example [9], [11], [17]). There are also approaches based on belief-revision models in philosophy and artificial intelligence (e.g., [3], [8]), and approaches based on voting theory (e.g., [18], [34], [35]). Many such strategies are based more on intuitions of reasonableness than on firm mathematics. For example, using weighted voting techniques to resolve conflicts between multiple sensors does not seem unreasonable, but there is no firm mathematical justification for the supposition that this will

produce results that accurately measure the joint probability. Dempster-Shafer theory is so familiar that many people suppose it is based on firm mathematics. Indeed, the mathematics internal to the theory is unassailable, but the justification for using it to assess joint probabilities is based on the intuitive reasonableness of the results, and not everyone agrees that the results are always reasonable (e.g., [17]). In the course of this paper examples will be given in which Dempster-Shafer theory seems clearly to yield undesirable results.

The purpose of this paper is to show that, within the context of a particular approach to probability theory, the problem of estimating joint probabilities has a purely mathematical solution and need not rest on undefended intuitions of reasonableness.

2. Two Kinds of Probability

No doubt the currently most popular theory of the foundations of probability is the subjectivist theory due originally to Ramsey [27] and Savage [30], and developed at length by many more recent scholars. However, the solution to the problem of joint probabilities proferred here begins instead with objective probabilities. Historically, there have been two general approaches to probability theory. The most familiar takes "definite" or "single-case" probabilities to be basic. Definite probabilities attach to specific states of affairs or propositions. For example, one can talk about the probability that it will rain today. In the foundations of probability theory, it is common to use the notation of mathematical logic and symbolize these states of affairs using logical formulas. Definite probabilities will be written using small caps: PROB(P) and PROB(P/Q). To be contrasted with definite probabilities are "indefinite" or "general" probabilities (sometimes called "statistical probabilities"). The indefinite probability of *an A* being a *B* is not about any particular *A*, but rather about the *property* of being an *A*. In this respect, its logical form is the same as that of relative frequencies. Again, it is customary to use logical notation to symbolize properties. For example, the property of being red can be symbolized as "x is red". In this notation, the variable x is said to be *free*. This is to be contrasted with a case in which all variables are bound by quantifiers. For instance, "$(\exists x)x$ is red" means "Some (unspecified) object is red". The latter symbolizes a state of affairs rather than a property. Formulas containing free variables are called *open formulas*. Indefinite probabilities will be symbolized using lower case "prob" and free variables: prob(Bx/Ax).

The reliability of a particular sensor is the general probability of an event of the appropriate sort occurring when the sensor produces a positive reading. This is about positive readings in general — not about any particular instance in which the sensor produces a positive reading. In other words, it is an indefinite probability. If on some particular occasion the sensor produces a positive reading and it is known in some independent way that the remote event is definitely occurring, then the single-case (i.e., definite) probability of the event occurring is 1.0, but that does not alter the general reliability of the sensor.

The distinction between definite and indefinite probabilities is often overlooked by contemporary probability theorists, perhaps because of the popularity of subjective probability (which has no way to make sense of indefinite probabilities). For example, Kolmogorov's axioms are axioms for definite probabilities. But most objective approaches to probability tie probabilities to relative frequencies in some essential way, and the resulting probabilities have the same logical form as relative frequencies. That is, they are indefinite probabilities. The simplest theories identify indefinite probabilities with relative frequencies [4], [15], [29], [33], [35].[1] The simplest objection to such "finite frequency theories" is that probability judgments are often made that diverge from relative frequencies. For example, one can talk about a coin being fair (and so the indefinite probability of a flip landing heads is 0.5) even when it is flipped only once and then destroyed (in which case the relative frequency is either 1 or 0). For understanding such indefinite probabilities, a notion of probability is needed that talks about *possible* instances of properties as well as actual instances. Theories of this sort are sometimes called "hypothetical frequency theories". C. S. Peirce was perhaps the first to make a suggestion of this sort. Similarly, the statistician R. A. Fisher, regarded by many as "the father of modern statistics", identified probabilities with ratios in a "hypothetical infinite population, of which the actual data is regarded as constituting a random sample" ([6], p. 311). Karl Popper [24], [25], [26] endorsed a theory along these lines and called the resulting probabilities *propensities*. Henry Kyburg [16] was the first to construct a precise version of

[1] William Kneale [14] traces the frequency theory to R. L. Ellis, writing in the 1840's, and John Venn [37] and C. S. Peirce in the 1880's and 1890's.

this theory (although he did not endorse the theory), and it is to him that we owe the name "hypothetical frequency theories". Kyburg [16] also insisted that von Mises should be considered a hypothetical frequentist. There are obvious difficulties for spelling out the details of a hypothetical frequency theory. More recent attempts to formulate precise versions of what might be regarded as hypothetical frequency theories are [1], [2], [10], and [19]. This paper takes its impetus from the theory of [19], which will be sketched briefly in section three.

After brief thought, most people find the distinction between definite and indefinite probabilities intuitively clear. However, this is a distinction that sometimes puzzles probability theorists, many of whom have been raised on an exclusive diet of definite probabilities. They are sometimes tempted to confuse indefinite probabilities with probability distributions over random variables. But random variables are not variables at all (in the sense of mathematical logic), but functions assigning values to the different members of a population. Indefinite probabilities have single numbers as their values. Probability distributions over random variables are just what their name implies — distributions of definite probabilities rather than single numbers.

It has always been acknowledged that for practical decision-making what is needed definite probabilities rather than indefinite probabilities. For example, in deciding whether to trust the output of the sensor on some particular occasion, one wants to know the probability of the remote event occurring *in this case*, not the general probability of events of this kind occurring when the sensor produces positive readings. So theories that take indefinite probabilities as basic need a way of deriving definite probabilities from them. Theories of how to do this are theories of *direct inference*. Theories of objective indefinite probability propose that statistical inference gives us knowledge of indefinite probabilities, and then direct inference gives us knowledge of definite probabilities. Reichenbach [28] pioneered the theory of direct inference. The basic idea is that in evaluating the definite probability PROB(Fa), one should look for the narrowest reference class (or reference property) G such that (1) the value of the indefinite probability prob(Fx/Gx) is known, and (2) it is known that Ga. Then PROB(Fa) is identified with prob(Fx/Gx). For example, actuarial reasoning aimed at setting insurance rates proceeds in roughly this fashion. Kyburg [15] was the first to attempt to provide firm logical foundations for direct inference. Pollock [19] took that as its starting point and constructed a modified theory with a more epistemological orientation. [22] and [23] present an updated version of some of the material in [19].

Applying this to the problem of joint probabilities, what is known initially is two indefinite probabilities prob(Ax/Bx) and prob(Ax/Cx), and what is sought is the joint probability prob($Ax/Bx\&Cx$). Given the latter, direct inference can be used to determine the single-case definite probability of the remote event in a particular case in which the outputs of both sensors are known.

Focusing attention on indefinite probabilities rather than definite probabilities still does not make it possible to use the probability calculus to compute the values of joint probabilities. However, as now be shown, a particular approach to the theory of indefinite probabilities enables us compute values for the joint probabilities that, although not *guaranteed* to be correct, can usually be expected to be correct.

3. Nomic Probability

It was remarked above that indefinite probabilities are best viewed as being something like relative frequencies in infinite populations of "possible objects". [19] developed a possible worlds semantics for objective indefinite probabilities,[2] and that will be taken that as the starting point for the present theory of probable probabilities. The proposal was that the *nomic probability* prob(Fx/Gx) can be identified with the proportion of physically possible G'*s* that are F'*s*. A *physically possible* G is defined to be an ordered pair $\langle w,x\rangle$ such that w is a physically possible world (one compatible with all of the physical laws) and x has the property G at w. The *subproperty relation* is defined as follows:

$F \preccurlyeq G$ iff it is physically necessary (follows from true physical laws) that $(\forall x)(Fx \rightarrow Gx)$.

Equivalently, $F \preccurlyeq G$ iff the set of physically possible F's is a subset of the set of physically possible G's. One can think of the subproperty relation as a kind of nomic entailment relation (holding between properties rather than propositions).

[2] Somewhat similar semantics were proposed by Halpern [10] and Bacchus [1].

Given a suitable proportion function ρ, it could be stipulated that, where $\mathfrak{F}$ and $\mathfrak{G}$ are the sets of physically possible F's and G's respectively:

$$\text{prob}_x(Fx/Gx) = \rho(\mathfrak{F},\mathfrak{G}).$$

However, it is unlikely that the right proportion function can be selected without appealing to prob itself, so the postulate is simply that *there is* some proportion function related to prob as above. This is merely taken to tell us something about the formal properties of prob. Rather than axiomatizing prob directly, it turns out to be more convenient to adopt axioms for the proportion function. Proportion functions are a generalization of measure functions, studied in mathematics in measure theory.

Note that prob_x is a variable-binding operator, binding the variable x. When there is no danger of confusion, the subscript "x" will be omitted, but sometimes it will be necessary to quantify into probability contexts, in which case it will be important to distinguish between the variables bound by "prob" and those that are left free. To simplify expressions, the variables in the properties will often be omitted, enabling us to write "prob(F/G)" for "prob(Fx/Gx)" when no confusion will result.

It is often convenient to write proportions in the same logical form as probabilities, so where φ and θ are open formulas with free variable x, let $\rho_x(\varphi/\theta) = \rho(\{x \mid \varphi \,\&\, \theta\},\{x \mid \theta\})$. Note that ρ_x is also a variable-binding operator, binding the variable x. Again, when there is no danger of confusion, the subscript "x" will typically be omitted.

Three classes of assumptions about the proportion function will be made here. Let $\#X$ be the cardinality of a set X. If Y is finite, it will be assumed that

$$\rho(X,Y) = \frac{\#X \cap Y}{\#Y}.$$

However, for present purposes the proportion function is most useful in talking about proportions among infinite sets. The sets $\mathfrak{F}$ and $\mathfrak{G}$ will invariably be infinite, if for no other reason than that there are infinitely many physically possible worlds in which there are F's and G's.

The second set of assumptions is that the standard axioms for conditional probabilities hold for proportions. These axioms automatically hold for relative frequencies among finite sets, so the assumption is just that they also hold for proportions among infinite sets.

Finally, three assumptions are adopted that go beyond merely imposing the standard axioms for the probability calculus on proportions. The three assumptions are:

Finite Set Principle:

For any set B, $N > 0$, and open formula Φ,

$$\rho_X\left(\Phi(X) \,/\, X \subseteq B \,\&\, \#X = N\right) =$$
$$\rho_{x_1,\ldots,x_N}\left(\Phi(\{x_1,\ldots,x_N\}) \,/\, x_1,\ldots,x_N \text{ are pairwise distinct } \&\; x_1,\ldots,x_N \in B\right).$$

Projection Principle:

If $0 \le p,q \le 1$ and $(\forall y)(Gy \rightarrow \rho_x(Fx/Rxy) \in [p,q])$, then $\rho_{x,y}(Fx/Rxy \,\&\, Gy) \in [p,q]$.

Crossproduct Principle:

If C and D are nonempty, $\rho(A \times B, C \times D) = \rho(A,C) \cdot \rho(B,D)$.

Note that these three principles are all theorems of elementary set theory when the sets in question are finite. For instance, the crossproduct principle holds for finite sets because $\#(A \times B) = (\#A)\cdot(\#B)$, and hence

$$\rho(A \times B, C \times D) = \frac{\#((A \times B) \cap (C \times D))}{\#(C \times D)} = \frac{\#((A \cap C) \times (B \cap D))}{\#(C \times D)}$$
$$= \frac{\#(A \cap C)\cdot\#(B \cap D)}{\#C \cdot \#D} = \frac{\#(A \cap C)}{\#C} \cdot \frac{\#(B \cap D)}{\#D} = \rho(A,C)\cdot\rho(B,D).$$

The assumption is simply that ρ continues to have these algebraic properties even when applied to infinite sets. This is a fairly conservative set of assumptions.

The objection is often proferred that in affirming the Crossproduct Principle, there must be a hidden assumption of statistical independence. However, that is to confuse proportions with probabilities. The Crossproduct Principle is about proportions — not probabilities. For finite sets, proportions are computed by simply counting members and computing ratios of cardinalities. It makes no sense to talk about statistical independence in this context. For infinite sets one cannot just count members any more, but the algebra is the same. It is because the algebra of proportions is simpler than the algebra of probabilities that it is useful to axiomatize nomic probabilities indirectly by adopting axioms for proportions.

4. The Y-Principle

It is results pertaining specifically to nomic probability that provide a solution to the problem of joint probabilities. Nomic probabilities are proportions among infinite sets of possible objects. The axioms adopted in section three enable one to prove theorems about how proportions among infinite sets work, and how they are related to proportions among finite sets. The details are complex, but they are spelled out and the theorems proven in [22]. For a more informal discussion, see [23]. A brief sketch of the theory will be given here. First, where "$x \underset{\delta}{\approx} y$" means that the absolute value of the difference between x and y is less than or equal to δ, one can use familiar-looking mathematics to prove:

Law of Large Numbers for Proportions:

If B is infinite and $\rho(A/B) = p$ then for every $\varepsilon,\delta > 0$, there is an N such that

$$\rho_X\left(\rho(A/X) \underset{\delta}{\approx} p \;/\; X \subseteq B \;\&\; \#X \geq N\right) \geq 1-\varepsilon .$$

Note that unlike Laws of Large Numbers for probabilities, the Law of Large Numbers for Proportions does not require an assumption of statistical independence. This is because it is derived from the crossproduct principle, and as remarked in section three, no such assumption is required (or even intelligible) for the crossproduct principle.

Given a list of variables $X_1,\ldots,X_n$ ranging over subsets of a set U, Boolean compounds of these sets are compounds formed by union, intersection, and set-complement. So, for example $(X \cup Y)\text{–}Z$ is a Boolean compound of X, Y, and Z. *Linear constraints* on the Boolean compounds either state the values of certain proportions, e.g., stipulating that $\rho(X,Y) = r$, or they relate proportions using linear equations. For example, the condition that $X = Y \cup Z$ generates the linear constraint

$$\rho(X,U) = \rho(Y,U) + \rho(Z,U) - \rho(X \cap Z,U).$$

Given the law of large numbers, the following can be proven:

Limit Principle for Proportions:

Consider a finite set LC of linear constraints on proportions between Boolean compounds of a list of variables $U,X_1,\ldots,X_n$. For any real number r between 0 and 1, if for every $\varepsilon,\delta > 0$, if there is an N such that for any finite set U such that $\#U > N$,

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) \geq 1-\varepsilon ,$$

then for any infinite set U, for every $\delta > 0$:

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) = 1 .$$

This limit principle provides a link between purely combinatorial theorems about the behavior of finite sets in the limit (as their sizes go to infinity) and theorems about proportions in infinite sets.

It turns out that under very general circumstances, combinatorial theorems of the form of the antecedent in the Limit Principle can be proven. In [22], the following theorem is proven:

Probable Probabilities Theorem:

Let $U,X_1,\ldots,X_n$ be a set of variables ranging over sets, and consider a finite set LC of linear constraints on proportions between Boolean compounds of those variables. If LC is consistent with the probability calculus, then for any pair of Boolean compounds P,Q of $U,X_1,\ldots,X_n$ there is a real number r between 0 and 1 such that for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#U > N$, then

$$\rho_{X_1,\ldots,X_n}\left(\rho(P,Q) \underset{\delta}{\approx} r \;/\; LC \;\&\; X_1,\ldots,X_n \subseteq U\right) \geq 1-\varepsilon .$$

In [22] it was shown that there is an algorithm for generating systems of simultaneous equations that characterize the values of the numbers r described in the Probable Probabilities Theorem. When the systems of equations have analytic solutions, one can try to solve them automatically using computer algebra programs. The results presented in this paper were obtained in that way.[3] When the equations do not have analytic solutions, one can still solve them numerically.

The results presented in this paper derive from a few principles of the above form. First, define:

$$\mathrm{Y}(r,s\,|\,a) = \frac{rs(1-a)}{a(1-r-s)+rs}$$

The non-standard notation "$\mathrm{Y}(r,s\,|\,a)$" is used in place of "$\mathrm{Y}(r,s,a)$" because the first two variables turn out to work differently than the last variable. The basic theorem for joint probabilities is then:

Finite Y-Theorem:

For every $0 \leq r,s,a \leq 1$, for every $\varepsilon,\delta > 0$, there is an N such that if U is finite and $\#U > N$, then

$$\rho_{A,B,C}\left(\rho(A,B\cap C) \underset{\delta}{\approx} Y(r,s\,|\,a) \;/\; A,B,C \subseteq U \;\&\; \rho(A,U)=a \;\&\; \rho(A,B)=r \;\&\; \rho(A,C)=s\right) > 1-\varepsilon .$$

By the Limit Principle:

Infinitary Y-Theorem:

For every $0 \leq r,s,a \leq 1$, for every $\delta > 0$, if U is infinite then

$$\rho_{A,B,C}\left(\rho(A,B\cap C) \underset{\delta}{\approx} Y(r,s\,|\,a) \;/\; A,B,C \subseteq U \;\&\; \rho(A,U)=a \;\&\; \rho(A,B)=r \;\&\; \rho(A,C)=s\right) = 1 .$$

Nomic probabilities are proportions among infinite sets of physically possible objects. Thus the Infinitary Y-Theorem implies an analogous principle for nomic probabilities:

Y-Principle:

For every $0 \leq r,s,a \leq 1$, for every $\delta > 0$, for any property U:

$$\mathrm{prob}_{A,B,C}\left(\begin{array}{l}\mathrm{prob}(A/B\,\&\,C) \underset{\delta}{\approx} Y(r,s\,|\,a) \;/\\ A,B,C \preccurlyeq U \;\&\; \mathrm{prob}(A/U)=a \;\&\; \mathrm{prob}(A/B)=r \;\&\; \mathrm{prob}(A/C)=s\end{array}\right) = 1 .$$

Given that $A,B,C \preccurlyeq U$, $\mathrm{prob}(A/B) = r$, $\mathrm{prob}(A/C) = s$, and $\mathrm{prob}(A/U) = a$, it can be expected, with probability 1, that $\mathrm{prob}(A/B\,\&\,C) \underset{\delta}{\approx} Y(r,s\,|\,a)$. This does not mean that the conclusion is

[3] See the Appendix for more information. It turns out that Mathematica is unable to solve these systems of equations. Maple 11 can sometimes solve them, although it is slow and tends to run out of memory even on rather simple problems. To my surprise, my own home-grown computer algebra (written in Common LISP) does significantly better. It is discussed in the appendix.

guaranteed to be true. The probabilities are proportions over infinite sets of possibilities, so even though the probability is 1, there can be an infinite set of exceptions to this expectation. However, the set of exceptions has measure 0. This makes it reasonable to expect the conclusion to be true in the absence of any information to the contrary. This is an example of what, in philosophy, is called "a defeasible inference". It is an inference whose conclusion is not logically guaranteed to follow from the premises, but which is nevertheless a reasonable inference to make. It is generally acknowledged in both philosophy and artificial intelligence that, outside of mathematics, most of the inferences we make are defeasible, and there is an extensive literature on the structure of defeasible inference (see [21] and references contained therein). However, the details are not relevant for present purposes. It suffices to note that defeasible inferences, although not deductively valid, are still reasonable inferences in the absence of contrary information.

If it can be defeasibly expect that $\text{prob}(A/B\,\&\,C) \underset{\delta}{\approx} Y(r,s\,|\,a)$ for every $\delta > 0$, then it can be reasonably expected that prob(A/B & C) = Y(r,s |a). This can be expressed more simply by saying that if A,B,$C \preccurlyeq U$, prob(A/B) = r, prob(A/C) = s, and prob(A/U) = a, then the *expectable value* of prob(A/B&C) = Y(r,s |a). If U represents our background knowledge, prob(A/U) is the *base rate* of A relative to that background knowledge. If the base rate of A is known, and the values of prob(A/B) and prob(A/C) are also known, then it can be reasonably expected that the joint probability prob(A/B&C) has the value Y(r,s |a), even though it does not follow from the probability calculus that it must have this value. This is because, although there are possible exceptions, they constitute a set of measure 0.

To get a better feel for what the Y-Principle implies, it is useful to examine plots of the Y-function. Figure 1 illustrates that Y(r,s |.5) is symmetric around the right-leaning diagonal. Varying a has the effect of warping the Y-function up or down relative to the right-leaning diagonal. This is illustrated in figure 2 for several choices of a.

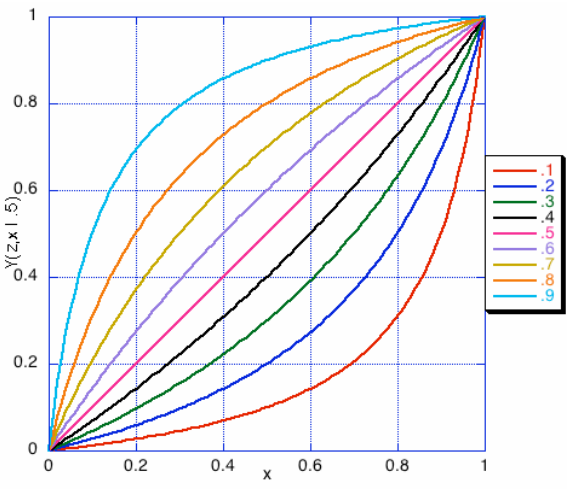

Figure 1. Y(z,x |.5), holding z constant
(for several choices of z as indicated in the key).

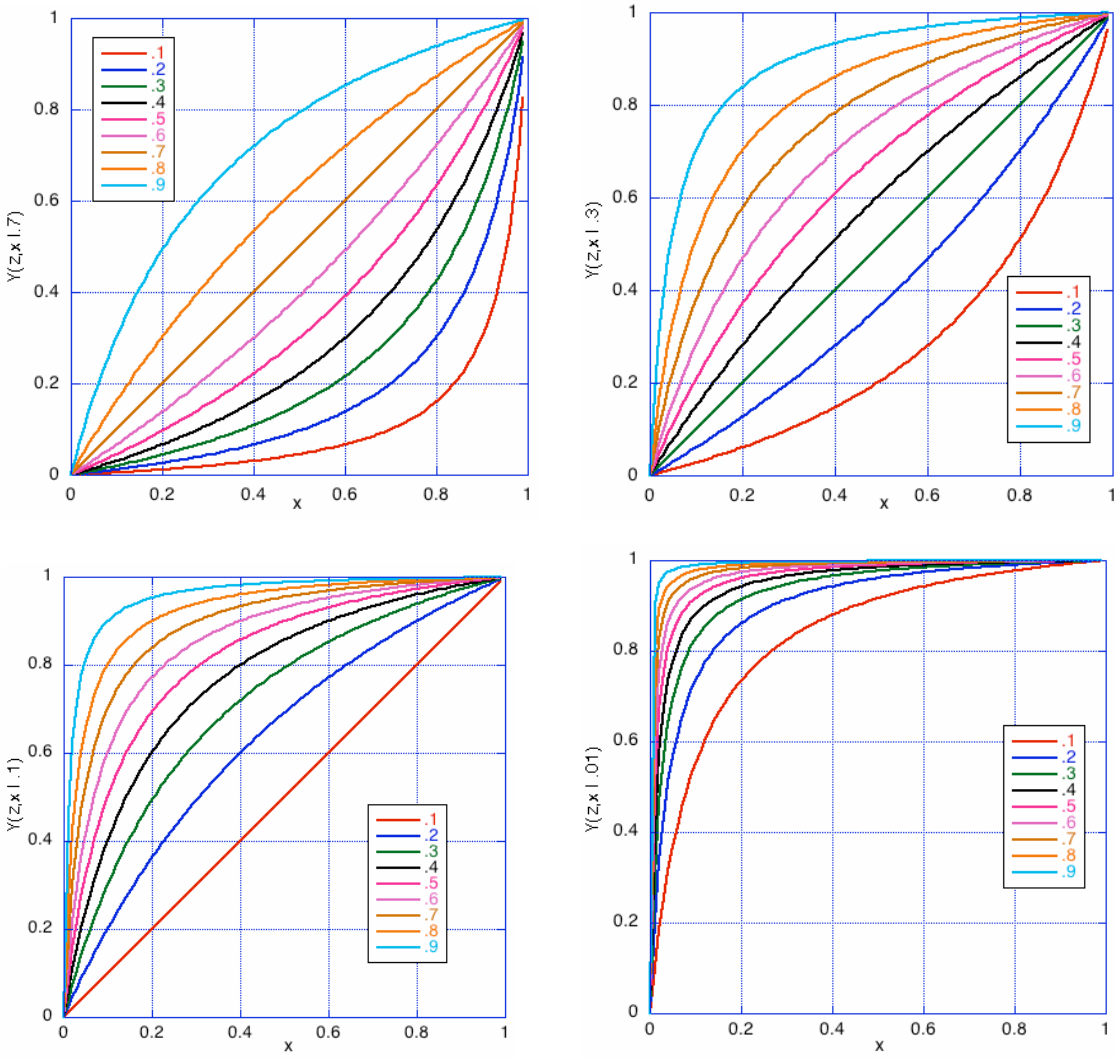

Figure 2. Y(z,x |a) holding z constant (for several choices of z), for a = .7, a = .3, a = .1, and a = .01.

Note that, in general, when $r,s < a$ then Y(r,s |a) < r and Y(r,s |a) < s, and when $r,s > a$ then Y(r,s |a) > r and Y(r,s |a) < s.

The Y-function has a number of important properties.[4] In particular, it is important that the Y-function is associative and commutative in the first two variables:

Theorem 1: Y(r,s |a) = Y(s,r |a).

Theorem 2: Y(r,Y(s,t |a) |a) = Y(Y(r,s |a),t |a).

[4] It turns out that the Y-function has been studied for its desirable mathematical properties in the theory of associative compensatory aggregation operators in fuzzy logic (5], [7], [13]). Y(r,s |a) is the function $D_\lambda(r,s)$ for $\lambda = \frac{1-a}{a}$ [13].

Theorems 1 and 2 are important for the use of the Y-function in computing probabilities. Suppose it is known that prob(A/B) = .6, prob(A/C) = .7, and prob(A/D) = .75, where $B,C,D \preccurlyeq U$ and prob(A/U) = .3. In light of theorems 1 and 2, the first three probabilities can be combined in any order to infer defeasibly that prob($A/B\&C\&D$) = Y(.6,Y(.7,.75 |.3) |.3) = Y(Y(.6,.7 |.3),.75 |.3) = .98. This makes it convenient to extend the Y-function recursively so that it can be applied to an arbitrary number of arguments (greater than or equal to 3):

If $n \geq 3$, $Y(r_1,\ldots,r_n \,|a) = Y(r_1,Y(r_2,\ldots,r_n \,|a) \;|a)$.

Then the Y-Principle can be strengthened as follows:

Generalized Y-Principle:

If $B_1,\ldots,B_n \preccurlyeq U$, prob($A/B_1$) = r_1,…, prob(A/B_n) = r_n, and prob(A/U) = a, the expectable value of prob(A/B_1 &…& B_n & C) = $Y(r_1,\ldots,r_n \,|a)$.

The Y-Principle often constitutes a solution to the problem of joint probabilities. For example, suppose there are two different diagnostic tests for a disease, and Bernard tests positive on both tests. Intuitively, this should make it more likely that Bernard has the disease than if the result of only one test were known. The Y-function can be applied to compute the joint probability. Let U represent our background knowledge, including Bernard's general symptoms other than the results of the test. Suppose the disease is rare, with a base rate of .1, but each positive test individually confers a probability of .4 that the patient has the disease. Two positive tests should increase that probability further. Indeed, Y(.4,.4 |.1) = .8.

Joseph Halpern obserfves (in correspondence) that in the special case in which the base rate is .5, the Y-principle is equivalent to Dempster's "rule of composition" for belief functions [31].[5] However, by ignoring the base rate prob(A/U) (setting it equal to .5 by default), the Dempster-Shafer theory will often give intuitively incorrect results. For example, in the case of the two tests for the disease, two positive tests should increase that probability. However, Y(.4,.4 |.5) = .3, so if the base rate is ignored, two positive tests would lower the probability of having the disease instead of raising it.

Consider the application of the Generalized Y-Principle to a fictional sensor fusion example. Consider have ten sensors attempting to detect remote events of type A. The sensors are binary, i.e., each can give a positive reading "T" or a negative reading "F". Suppose that for each sensor, the probability of A given a positive reading is known, and the probability of $\sim A$ given a negative reading is known. Suppose these values are as listed in table 1. Suppose further that the base rate is prob(A/U) = .15, and the sensor readings are T,T,F,T,F,F,T,T,F,T. What should the probability of there being an A be taken to be? Untutored intuition will not be of much help here, but the Generalized Y-Principle can be used to estimate that it is Y(.73,.68,.12,.88,.09,.07,.90,.82,.03,.86 |.15) = .25.

Sensor number	Prob(A/s = T)	Prob($\sim A/s$ = F)
1	.73	.89
2	.68	.91
3	.91	.88
4	.88	.87
5	.45	.91
6	.78	.93
7	.90	.92
8	.82	.87
9	.78	.97
10	.86	.93

Table 1. Sensor reliabilities

[5] See also [2]. Given very restrictive assumptions, [2] theory gets the special case of the Y-Principle in which a = .5, but not the general case.

5. An Example

The mathematics underlying the Y-Principle is compelling. It makes it reasonable to expect that joint probabilities can usually be computed using the Y-function. However, although this is a reasonable expectation, it is not logically guaranteed to yield the right answer. So it seems wise to test this prediction on a concrete example. To generate a simple database that could be used for this purpose, human subjects were employed as detectors. One hundred paragraphs were generated, each consisting of a list of 100 words. The word "box" occurred in 40 of the paragraphs, and the word "cat" appeared in 35. A subject was shown the 100 paragraphs sequentially, and given 5 seconds to scan each paragraph, looking for the words "box" and "cat". The subjects were used as detectors *for the absence* of these words. If a subject fails to see either word, there will be some probability that the word is absent from the paragraph, and that probability can be estimated on the basis of the data collected. So for each subject S (each detector), prob(not-cat/not-S-see-cat) and prob(not-box/not-S-see-box) can be estimated.

Now consider pairs of subjects (pairs of detectors) S_1 and S_2. All subjects see the same paragraphs, so the data can be used to estimate the joint probabilities prob(not-cat/not-S_1-see-cat & not-S_2-see-cat) and prob(not-box/not-S_1-see-box& not-S_2-see-box) that a word is absent given that neither subject sees it. The base rates of the absence of "box" and "cat" are known, viz., .6 and .65. So, for each pair of subjects S_1 and S_2, the measured joint probability

prob(not-cat/not-S_1-see-cat & not-S_2-see-cat)

can be compared with the predicted joint probability

Y(prob(not-cat/not-S_1-see-cat),prob(not-cat/not-S_2-see-cat) |.65),

and the measured joint probability

prob(not-box/not-S_1-see-box& not-S_2-see-box)

can be compared with the predicted joint probability

Y(prob(not-box/not-S_1-see-box),prob(not-box/not-S_2-see-box) |.6).

Thirteen subjects were used, which generates 78 pairs of subjects, and two data points for each pair of subjects (one for "cat" and one for "box"). The data points consist of measured relative frequencies, from which the corresponding probabilities can be estimated. The mean of the ratio of the measured joint relative frequency to the joint probability predicted by the Y-function was 0.995, with a mean deviation of 0.013. This strongly confirms the correctness of the use of the Y-Principle to estimate the joint probabilities in this example. Note that the Dempster-Shafer theory would give the systematically wrong answer in this example, because the base rates are not .5.

6. Defeaters

The estimation of joint probabilities using the Y-function is a defeasible inference. That is, it is reasonable in the absence of any information to the contrary, but there can be "defeaters" — information in the presence of which this inference ceases to be reasonable. One class of defeaters arises from the following theorems. Define:

B and C are *Y-independent for A relative to U* iff $A,B,C \preccurlyeq U$ and

(a) $\text{prob}(C/B\&A) = \text{prob}(C/A)$

and

(b) $\text{prob}(C/B\&{\sim}A) = \text{prob}(C/U\&{\sim}A)$.

The following theorem was proven in [22]:

Y-Theorem:

Let $r = \text{prob}(A/B)$, $s = \text{prob}(A/C)$, and $a = \text{prob}(A/U)$. If B and C are Y-independent for A relative to U then $\text{prob}(A/B\&C) = \text{Y}(r,s\,|a)$.

A slight generalization of the proof of the Y-Theorem produces:

Generalized-Y-Theorem:

Suppose $A,B,C \preccurlyeq U$. Let $r = \text{prob}(A/B)$, $s = \text{prob}(A/C)$, $a = \text{prob}(A/U)$, $\alpha = \dfrac{\text{prob}(C/B\,\&\,A)}{\text{prob}(C/A)}$, and $\beta = \dfrac{\text{prob}(C/B\&\sim A)}{\text{prob}(C/U\&\sim A)}$. Then $\text{prob}(A/B\&C) = \dfrac{1}{1+\dfrac{\beta a(1-r-s+rs)}{\alpha(1-a)rs}}$.

Define:

$$\text{GY}(r,s\,|a,\alpha,\beta) = \frac{1}{1+\dfrac{\beta a(1-r-s+rs)}{\alpha(1-a)rs}}.$$

Trivially:

$$\text{Y}(r,s\,|a) = \frac{1}{1+\dfrac{a(1-r-s+rs)}{(1-a)rs}}.$$

Thus $\text{prob}(A/B\&C) = \text{Y}(r,s\,|a)$ iff $\alpha = \beta$.

In general, $\text{GY}(r,s\,|a,\alpha,\beta) > \text{Y}(r,s\,|a)$ iff $\beta < \alpha$. Therefore:

Y-Relevance-Theorem:

Suppose $A,B,C \preccurlyeq U$. Let $r = \text{prob}(A/B)$, $s = \text{prob}(A/C)$, $a = \text{prob}(A/U)$, $\alpha = \dfrac{\text{prob}(C/B\,\&\,A)}{\text{prob}(C/A)}$, and $\beta = \dfrac{\text{prob}(C/B\&\sim A)}{\text{prob}(C/U\&\sim A)}$. Then:

(a) If $\alpha < 1$ and $\beta \geq 1$ then $\text{prob}(A/B\&C) < \text{Y}(r,s\,|a)$;

(b) If $\alpha > 1$ and $\beta \leq 1$ then $\text{prob}(A/B\&C) > \text{Y}(r,s\,|a)$;

(c) If $\alpha \geq 1$ and $\beta < 1$ then $\text{prob}(A/B\&C) > \text{Y}(r,s\,|a)$.

The Y-Principle derives from the fact that, when $r = \text{prob}(A/B)$, $s = \text{prob}(A/C)$, and $a = \text{prob}(A/U)$, the expectable values of α and β are both 1 (see [22]). Thus, given a reason for believing that one of α or β is not equal to 1, if nothing is known about the other, this constitutes a reason for expecting that $\text{prob}(A/B\&C) \neq \text{Y}(r,s\,|a)$, and so constitutes a defeater for the use of the Y-Principle. There can be defeaters corresponding to each of cases (a), (b), and (c) of the Y-Relevance-Theorem.

In case (a), B is "negatively Y-relevant to C for A", and B is not positively Y-relevant to C for $\sim A$. In this case, $\text{prob}(A/B\&C)$ cannot be computed as $\text{Y}(r,s\,|a)$. $\text{Y}(r,s\,|a)$ provides only an upper bound on $\text{prob}(A/B\&C)$. To illustrate this, distinguish between cases in which $\text{prob}(A/B)$ and $\text{prob}(A/C)$ reflect informational connections and cases in which they reflect causal connections. Examples of informational connections include diagnostic relations in medicine, sensors sensing remote events,

and so forth. In these cases, B does not cause A (e.g., the results of the test do not cause the disease). Rather, the direction of causation is from A to B (the disease causes the test to have certain results). If the connections are informational then it is usually reasonable to expect Y-independence. But in causal cases, this is not a reasonable expectation. Suppose B and C both have a tendency to cause A. For example, poisoning a person and shooting him may both have a tendency to cause his death. In this case, it should not be expected that Y-independence holds. Knowing that a person dies presumably raises the probability of his having been poisoned, but if it is known that he was shot and died, this would raise the probability of his having been poisoned to a lesser degree (if at all). In general, if B and C are probabilistic causes of A, it should be expected that $\text{prob}(C/B\&A) < \text{prob}(C/A)$. On the other hand, it still seems that it should be expected that $\text{prob}(C/B\&{\sim}A) = \text{prob}(C/U\&{\sim}A)$. For instance, a person's not dying lowers the probability that he was poisoned, and it seems to do so just as much if it is known he was shot. So this is a case of type (a), and all that can be reasonably expected is that $\text{prob}(A/B\&C) < Y(r,s\,|a)$.

Informational cases are sometimes of types (b) or (c). For instance, suppose there are two sensors B and C detecting remote events of type A. But suppose there are two subtypes of events of type A — types A_1 and A_2 — where events of type A_1 are more easily detected by the sensors than are events of type A_2. Then if one sensor detects an event, that raises the probability that it is of type A_1, and so raises the probability that the other sensor will also detect it. In other words, $\alpha > 1$. Similarly, there could be a class of cases in which the sensors are more likely to register false positives. In that case, $\beta < 1$.

In all of these cases, Y-independence should not be expected, and so the Y-Principle should not be used to estimate the value of $\text{prob}(A/B\&C)$ directly. However, the Y-Principle can still be used to estimate the value of $\text{prob}(A/B\&C)$ indirectly. Consider the first sort of case. By the probability calculus:

$$\text{prob}(A/B\&C) = \text{prob}(A_1/B\&C) + \text{prob}(A_2/B\&C).$$

If the values of $\text{prob}(A_1/B)$, $\text{prob}(A_1/C)$, $\text{prob}(A_2/B)$, and $\text{prob}(A_2/C)$ are known, the Y-Principle can be used to compute values for $\text{prob}(A_1/B\&C)$ and $\text{prob}(A_2/B\&C)$, and the latter can be summed to estimate $\text{prob}(A/B\&C)$.

Instead of having different kinds of A's, there might be different circumstances in which the reliability of the sensors vary. For instance, if the circumstances can be partitioned into two subcases S_1 and S_2, it can be computed that

$$\text{prob}(A/B\&C) = \text{prob}(A/B\&C\&S_1)\cdot\text{prob}(S_1/B\&C) + \text{prob}(A/B\&C\&S_2)\cdot\text{prob}(S_2/B\&C).$$

Then if the values of $\text{prob}(A/B\&S_1)$, $\text{prob}(A/C\&S_1)$, $\text{prob}(A/B\&S_2)$, and $\text{prob}(A/C\&S_2)$ are known, the Y-Principle can be used to compute values for $\text{prob}(A/B\&C\&S_1)$ and $\text{prob}(A/B\&C\&S_2)$, and those can be used to estimate $\text{prob}(A/B\&C)$.

In this way, Y-independence can often be restored by dividing cases. There can also be cases in which some continuous parameter affects the detectability of an event. For instance, in the "box" and "cat" cases, the amount of time subjects have for scanning a paragraph could be varied. This kind of case can be handled similarly to the above using probability distributions.

7. The Y_0-function

The application of the Y-function presupposes that the base rate $\text{prob}(A/U)$ is known. But suppose it is not. Then what can be concluded about $\text{prob}(A/B\&C)$? It might be suspected that it can be assumed by default that $\text{prob}(A/U) = .5$, and so concluded that $\text{prob}(A/B\&C) = Y(r,s\,|.5)$. That would be interesting because, as remarked above, this is equivalent to Dempster's "rule of composition" for belief functions [31]. As illustrated in section four, if the value of $\text{prob}(A/U)$ is known but ignored, the Dempster-Shafer rule will often yield intuitively incorrect results. But if the value of $\text{prob}(A/U)$ is unknown, can it be assumed to be .5? It turns out that even given ignorance of the base rate, the Dempster-Shafer rule does not give quite the right answer. The difficulty is that knowing the values of $\text{prob}(A/B)$ and $\text{prob}(A/C)$ affects the expectable value of $\text{prob}(A/U)$. Define $Y_0(r,s)$ to be $Y(r,s\,|a)$ where a is the solution to the following set of three simultaneous equations (for variable a, b, and c, and fixed r and s):

$$2a^3-(b+c-2b\cdot r-2c\cdot s-3)a^2$$
$$+(b\cdot c+2b\cdot r-b\cdot cr+2c\cdot s-b\cdot c\cdot s+2b\cdot c\cdot r\cdot s-b-c+1)a-b\cdot c\cdot r\cdot s=0;$$

$$\left(\frac{1-s}{1+(s-a)c}\right)^{1-s}\left(\frac{s}{a-s\cdot c}\right)^{s}=1;$$

$$\left(\frac{1-r}{1+(r-a)b}\right)^{1-r}\left(\frac{r}{a-r\cdot b}\right)^{r}=1.$$

Then the following principle can be proven [22]:

Y_0-Principle:

If prob(A/B) = r and prob(A/C) = s, then the expectable value of prob(A/B & C) = $Y_0(r,s)$.

If a is the expectable value of prob(A/U) given that prob(A/B) = r and prob(A/C) = s, then $Y_0(r,s)$ = $Y(r,s\,|a)$. However, a does not have a simple analytic characterization. $Y_0(r,s)$ is plotted in figure 3, and the default values of prob(A/U) are plotted in figure 4. Note how the curve for $Y_0(r,s)$ is twisted with respect to the curve for $Y(r,s\,|.5)$ (in figure 1).

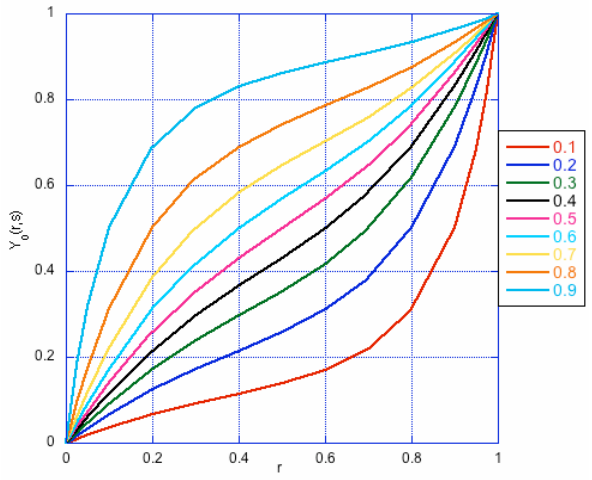

Figure 3. $Y_0(r,s)$, holding s constant (for several choices of s as indicated in the key)

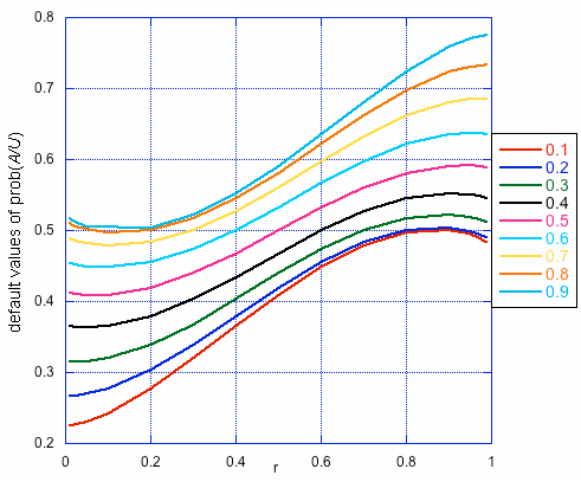

Figure 4. Default values of prob(A/U) (for several choices of s as indicated in the key)

8. Multiple Correlations

Sometimes, diagnostic indicators are correlated with more than one possible diagnosis. For example, the results of a medical test (C) may make it more probable both that one has lung cancer (A) and viral pneumonia (B). Suppose there are two such diagnostic indicators (two medical tests) C and D that both bear upon the same possible diagnoses. How can the information from each be combined? The natural hypothesis is that it can be assumed defeasibly that C and D are Y-independent for both A and B, and then the Y-function can be used to compute both prob(A/C&D) and prob(B/C&D). However, the following theorem of the probability calculus makes it difficult for C and C to be Y-independent for two different properties A and B relative to U:

Theorem 3: Let r = prob(A/C), s = prob(A/D), and a = prob(A/U). If C and D are Y-independent for A relative to U then

$$\text{prob}(D/C) = \frac{\text{prob}(A/C)\cdot\text{prob}(A/D)\cdot\text{prob}(D/U)}{\text{prob}(A/C\,\&\,D)\cdot\text{prob}(A/U)} = \frac{a(1-r-s)+rs}{a(1-a)}\cdot\text{prob}(D/U).$$

An immediate corollary is:

Corollary 1: If C and D are Y-independent for both A and B relative to U, where prob(A/C) = r_1, prob(A/D) = s_1, prob(A/U) = a, prob(B/C) = r_2, prob(B/D) = s_2, and prob(B/U) = b, then $\frac{a(1-r_1-s_1)+r_1s_1}{a(1-a)} = \frac{b(1-r_2-s_2)+r_2s_2}{b(1-b)}$.

If it is known that prob(A/C) = r_1, prob(A/D) = s_1, prob(A/Ux) = a, and prob(B/U) = b, but the values of prob(B/C) and prob(B/D) are unknown, then the expectable values of the unspecified variables are such that it can still be expected that C and D are Y-independent for A relative to U.[6] On the other hand, if it is also known that prob(B/C) = r_2 and prob(B/D) = s_2, where $\frac{a(1-r_1-s_1)+r_1s_1}{a(1-a)} \neq \frac{b(1-r_2-s_2)+r_2s_2}{b(1-b)}$, then it follows from corollary 2 that C and C cannot be Y-independent for both A and B relative to U. If here is no reason to give precedence to one of A and B over the other, it should not be assumed of either that Y-independence holds for it, so this defeats the assumption of Y-independence for both.

Although the Y-Principle cannot be used for this purpose, the expectable values for prob(A/C&D) and prob(B/C&D) can still be computed. Consider two cases. First, suppose that the values of prob(A&B/U), prob(C/U), and prob(D/U) are known. It can be proven that

Theorem 4: If prob(A/C) = r_1, prob(A/D) = s_1, prob(A/U) = a, prob(B/C) = r_2, prob(B/D) = s_2, prob(B/U) = b, prob(C/U) = c, prob(D/U) = d, and prob(A&B/U) = ab, then where abc and abd are the solutions to the following pair of equations:

$$\frac{abc(a-ab+abc-c\cdot r_1)(abc-ab+b-c\cdot r_2)(abc-c(r_1+r_2-1)}{(ab-abc)(abc-c\cdot r_1)(abc-c\cdot r_2)(1-a+ab-abc-b-c(1-r_1-r_2)} = 1$$

$$\frac{abd(a-ab+abd-d\cdot s_1)(abd-ab+b-d\cdot s_2)(abd-d(s_1+s_2-1)}{(ab-abd)(abd-d\cdot s_1)(abd-d\cdot s_2)(1-a+ab-abd-b-d(1-s_1-s_2)} = 1$$

the expectable value of prob(A/C&D) is

$$\frac{(b-ab)(1-a-b+ab)(a\cdot abc\cdot abd-ab(abd\cdot c\cdot r_1+d(abc-c\cdot r_1)s_1))}{\left(\begin{array}{c} a\cdot abc\cdot abd\cdot b(1-a-b)+ab^3(abd\cdot c+d(abc+c(1-r_2(1-s_1)-s_1-r_1(1-s_2)-s_2))) \\ -ab^2\left(\begin{array}{c} abc(abd+d(a+b-s_1+b\cdot s_1-s_2+a\cdot s_2)) \\ +c\left(abd(b-r_1+b\cdot r_1-r_2)+a(abd(1+r_2)+d(1-s_1-r_2(1-s_1+s_2)-r_1(1-s_2)-s_2))\right) \\ +d\left(r_1\cdot s_1+r_2\cdot s_2+b\left(1-r_2-s_2+r_2\cdot s_2-r_1(1+s_1-s_2)+s_2\right)\right) \\ -ab\left(\begin{array}{c} (1-b)b(abd\cdot c\cdot r_1-d(abc-c\cdot r_1)s_1)-a^2\left(abd\cdot c\cdot r_2+d(abc-c\cdot r_2)s_2\right) \\ +a(abc(2abd\cdot b+d(b-s_2))+c\left(abd(b-r_2)+d(b(1-r_2(1-s_1)-s_1-r_1(1-s_2)-s_2)+r_2\cdot s_2)\right)\end{array}\right)\end{array}\right)\end{array}\right)}$$

and the expectable value of prob(B/C&D) is

[6] The proof of this and the other theorems stated below are not included in the paper, but they can be generated automatically using the software mentioned in section four. See the appendix for more details.

$$\frac{(a-ab)(1-a-b+ab)\left(ab\cdot c\cdot r_2(abd-d\cdot s_2)-abc(abd\cdot b-ab\cdot d\cdot s_2)\right)}{\left(\begin{array}{c} a\cdot abc\cdot abd\cdot b(1-a-b)+ab^3(abd\cdot c+d(abc+c(1-r_2(1-s_1)-s_1-r_1(1-s_2)-s_2))) \\ -ab^2\left(\begin{array}{c} abc(abd+d(a+b-s_1+b\cdot s_1-s_2+a\cdot s_2)) \\ +c\left(abd(b-r_1+b\cdot r_1-r_2)+a(abd(1+r_2)+d(1-s_1-r_2(1-s_1+s_2)-r_1(1-s_2)-s_2))\right) \\ +d\left(r_1\cdot s_1+r_2\cdot s_2+b\left(1-r_2-s_2+r_2\cdot s_2-r_1(1+s_1-s_2)+s_2\right)\right) \\ -ab\left(\begin{array}{c} (1-b)b(abd\cdot c\cdot r_1-d(abc-c\cdot r_1)s_1)-a^2\left(abd\cdot c\cdot r_2+d(abc-c\cdot r_2)s_2\right) \\ +a(abc(2abd\cdot b+d(b-s_2))+c\left(abd(b-r_2)+d(b(1-r_2(1-s_1)-s_1-r_1(1-s_2)-s_2)+r_2\cdot s_2)\right)\end{array}\right)\end{array}\right)\end{array}\right)}.$$

The equations characterizing *abc* and *abd* do not have analytic solutions, but they can be solved numerically for particular values of *a*, *b*, *ab*, *c*, *d*, r_1, s_1, r_2, and s_2. For example, if $a = .37$, $b = .42$, $ab = .16$, $c = .55$, $d = .53$, $r_1 = .6$, $r_2 = .55$, $s_1 = .45$, and $s_2 = .48$, it can be computed that prob(*A*/*C*&*D*) = .66 and prob(*B*/*C*&*D*) = .59.

If instead the value of prob(*A*&*B*/*U*) is unknown, the following theorem can be proven:

Theorem 5: If prob(*A*/*C*) = r_1, prob(*A*/*D*) = s_1, prob(*A*/*U*) = *a*, prob(*B*/*C*) = r_2, prob(*B*/*D*) = s_2, prob(*B*/*U*) = *b*, prob(*C*/*U*) = *c*, and prob(*D*/*U*) = *d*, then where *abc*, *abd*, and *ab* are the solutions to the following set of simultaneous of equations:

$$\frac{abc(a-ab+abc-c\cdot r_1)(abc-ab+b-c\cdot r_2)(abc-c(r_1+r_2-1)}{(ab-abc)(abc-c\cdot r_1)(abc-c\cdot r_2)(1-a+ab-abc-b-c(1-r_1-r_2)}=1$$

$$\frac{abd(a-ab+abd-d\cdot s_1)(abd-ab+b-d\cdot s_2)(abd-d(s_1+s_2-1)}{(ab-abd)(abd-d\cdot s_1)(abd-d\cdot s_2)(1-a+ab-abd-b-d(1-s_1-s_2)}=1$$

$$\frac{\left(\begin{array}{c}(a-ab)(ab^2-ab\cdot abc-ab\cdot abd+abc\cdot abd)(b-ab)\\ (1-a+ab-abc-b-c(1-r_1-r_2))(1-a+ab-abc-b-d(1-s_1-s_2))\end{array}\right)}{\left(\begin{array}{c} ab(1-a-b+ab)(a-ab+abc-c\cdot r_1)(b-ab+abc-c\cdot r_2)\\ (a-ab+abd-d\cdot s_2)(b-ab+abd-d\cdot s_2)\end{array}\right)}=1$$

then the expectable values of prob(*A*/*C*&*D*) and prob(*B*/*C*&*D*) are as in theorem 4.

If *A* and *B* are mutually exclusive alternatives, things get simpler. For example, Hunter and Liu [11] consider a case in which one information source reports that the probability that the temperature is 8° C is 0.2 and the probability that the temperature is 12° C is 0.8. A second source reports that the probability that the temperature is 8° C is 0.4 and the probability that the temperature is 12° C is 0.6. In this case, the two alternatives (8° C and 12° C) are mutually exclusive possibilities. In this example, they are also treated as exhaustive, because the probabilities sum to 1. That makes the case easy, because trivially, if *B* and *C* are Y-independent for *A* relative to *U*, then *B* and *C* are Y-independent for ~*A* relative to *U*. Hence in this case the Y_0-Principle can be applied to infer that the joint probability that the temperature that the temperature is 8° C is .21. If the base rate is known one can do better. Hunter and Liu use the Dempster-Shafer rule and estimate that the joint probability is .14, which for the reasons given above is probably not a reasonable expectation.

A more interesting case occurs when the probabilities do not sum to 1, or equivalently, when *A* and *B* are mutually exclusive but not exhaustive. In this case the expectable values of the probabilities have analytic characterizations:

Theorem 6: If *A* and *B* are mutually exclusive alternatives, prob(*A*/*C*) = r_1, prob(*A*/*D*) = s_1, prob(*A*/*U*) = *a*, prob(*B*/*C*) = r_2, prob(*B*/*D*) = s_2, prob(*B*/*U*) = *b*, prob(*C*/*U*) = *c*, and prob(*D*/*U*) = *d*, then the expectable value of prob(*A*/*C*&*D*) is

$$\frac{b(1-a-b)r_1 \cdot s_1}{(1-b)b \cdot r_1 \cdot s_1 - a^2 \cdot r_2 \cdot s_2 + a\left(r_2 \cdot s_2 + b(1-r_1-r_2-s_1-s_2+r_2 \cdot s_1 + r_1 \cdot s_2)\right)}$$

and the expectable value of prob(A/C&D) is

$$\frac{a(1-a-b)r_2 \cdot s_2}{(1-b)b \cdot r_1 \cdot s_1 - a^2 \cdot r_2 \cdot s_2 + a\left(r_2 \cdot s_2 + b(1-r_1-r_2-s_1-s_2+r_2 \cdot s_1 + r_1 \cdot s_2)\right)}.$$

For example, making the Hunter and Liu example more realistic, consider two temperature ranges, [5°,8°] and [9°,12°], and supposed it is known that the base rates at which the temperatures fall in those ranges (relative to our background knowledge) is .3 and .4 respectively. If the probability, given one source of information, of the temperature falling in the first range is .2, and the probability of its falling in the second range is .3, and given a second source the probabilities are .15 and .35, then the expectable values for the joint probabilities are .18 and .12 respectively. To understand this result, note that each individual source makes the temperature ranges less probable than the base rates, so collectively the sources make the temperature ranges even less likely. The results can be generalized readily to the case in which there are any number of disjoint intervals.

9. Conclusions

Problems of information fusion would often be simplified if it were known how to estimate the value of the joint probability prob(A/B&C) given the values of prob(A/B), prob(A/C), and prob(A/U). The probability calculus does not enable us to compute a value for the joint probability from the values of the individual probabilities. However, within the theory of nomic probability, it is possible to show that there are inferences which, although not logically guaranteed to yield correct results, nevertheless do so with probability 1. Among these is the Y-Principle, which, in the absence of information to the contrary, allows us to infer defeasibly that prob(A/B&C) = Y(r,s |a).

10. Acknowledgement

This work was supported by NSF grant no. IIS-0412791.

Appendix: Proving the Theorems

The proofs for the theorems stated in section eight have not been included in the paper. These, and many of the other theorems in the paper (whose proofs were included in [22]), can be generated and proven automatically using the software mentioned in section four. This software can be downloaded from http://oscarhome.soc-sci.arizona.edu/ftp/OSCAR-web-page/CODE/Code for probable probabilities.zip. It runs in any implementation of Common LISP, including the free CLISP (http://clisp.cons.org/). For example, in section eight, the following theorem was stated:

Theorem 6: If A and B are mutually exclusive alternatives, prob(A/C) = r_1, prob(A/D) = s_1, prob(A/U) = a, prob(B/C) = r_2, prob(B/D) = s_2, prob(B/U) = b, prob(C/U) = c, and prob(D/U) = d, then the expectable value of prob(A/C&D) is

$$\frac{b(1-a-b)r_1 \cdot s_1}{(1-b)b \cdot r_1 \cdot s_1 - a^2 \cdot r_2 \cdot s_2 + a\left(r_2 \cdot s_2 + b(1-r_1-r_2-s_1-s_2+r_2 \cdot s_1 + r_1 \cdot s_2)\right)}$$

and the expectable value of prob(A/C&D) is

$$\frac{a(1-a-b)r_2 \cdot s_2}{(1-b)b \cdot r_1 \cdot s_1 - a^2 \cdot r_2 \cdot s_2 + a\left(r_2 \cdot s_2 + b(1-r_1-r_2-s_1-s_2+r_2 \cdot s_1 + r_1 \cdot s_2)\right)}.$$

This theorem was generated automatically by executing the instruction:

```
(analyze-probability-structure
  :subsets '(A B C D)
  :constants '(a b c d r1 s1 r2 s2)
  :subset-constraints '((B subset (- A)))
  :probability-constraints '((prob(A / C) = r1)
                             (prob(A / D) = s1)
                             (prob(B / C) = r2)
                             (prob(B / D) = s2))                                    )
  :probability-queries '(prob(A / (C & D))
                         prob(B / (C & D)))
  :display-infix t)
```

which produces the result:

```
============================================================
Dividing U into 4 subsets A,B,C,D whose cardinalities relative to U are a, b, c, d,
if the following constraints are satisfied:
    prob(A / C) = r1
    prob(A / D) = s1
    prob(B / C) = r2
    prob(B / D) = s2
    (B subset (- A))
and hence
    bd = (* s2 d)
    bc = (* r2 c)
    ad = (* s1 d)
    ac = (* r1 c)
    abd = 0
    ab = 0
    abcd = 0
    abc = 0
and the values of a, b, c, d, r1, s1, r2, s2 are held constant,
grounded definitions of the expectable values were found for all the variables.
------------------------------------------------------------------------------------------
The following definitions of expectable values were found that appeal only to the constants:
----------
abcd = 0
----------
cd = (((((r1 * s2 * a * b) + (r2 * s1 * a * b) + (a * b) + (a * r2 * s2) + (r1 * s1 * b)) -
       ((r1 * a * b) + (r2 * a * b) + (s1 * a * b) + (s2 * a * b) + (r1 * s1 * (b ^ 2)) + ((a ^ 2) * r2 * s2)))
      * c * d)
     / ((1 - (b + a)) * a * b))
----------
bcd = ((r2 * c * s2 * d) / b)
----------
acd = ((r1 * c * s1 * d) / a)
----------
abc = 0
----------
abd = 0
----------
ab = 0
============================================================
Reconstruing a, b, c, d, etc., as probabilities relative to U rather than as cardinalities, the
following characterizations were found for the expectable values of the probabilities wanted:
----------
prob(A / (C & D)) = (((1 - (b + a)) * b * r1 * s1) /
                     (((r1 * s2 * a * b) + (r2 * s1 * a * b) + (a * b) + (a * r2 * s2) + (r1 * s1 * b)) -
                      ((r1 * a * b) + (r2 * a * b) + (s1 * a * b) + (s2 * a * b) + (r1 * s1 * (b ^ 2)) + ((a ^ 2) * r2 * s2))))
----------
prob(B / (C & D)) = (((1 - (b + a)) * a * r2 * s2) /
                     (((r1 * s2 * a * b) + (r2 * s1 * a * b) + (a * b) + (a * r2 * s2) + (r1 * s1 * b)) -
                      ((r1 * a * b) + (r2 * a * b) + (s1 * a * b) + (s2 * a * b) + (r1 * s1 * (b ^ 2)) + ((a ^ 2) * r2 * s2))))
----------
```

==========

If the following line is added to the instruction

```
:display-details t
```

a human-readable proof of the results is produced. The proof presupposes some of the theorems proven in [22].

In section eight it was remarked that although the systems of simultaneous equations that characterize expectable values often lack analytic solutions, they can be solved numerically in specific cases. However, that turns out to be harder than it sounds. Neither Mathematica nor Maple is able to solve these equations numerically. The aforementioned software includes code that can often find numerical solutions. For example, in illustrating theorem 4, it was observed that if $a = .37$, $b = .42$, $ab = .16$, $c = .55$, $d = .53$, $r_1 = .6$, $r_2 = .55$, $s_1 = .45$, and $s_2 = .48$, it can be computed that the expectable value of prob(A/C&D) = .66 and the expectable value of prob(B/C&D) = .59. This result was obtained by executing the instruction:

```
(find-expectable-values
  :args '((a = .37) (b = .42) (ab = .16) (c = .55) (d = .53) (r1 = .6) (r2 = .55) (s1 = .45) (s2 = .48))
  :subsets '(A B C D)
  :probability-constraints '((prob(A / C) = r1)
                             (prob(A / D) = s1)
                             (prob(B / C) = r2)
                             (prob(B / D) = s2))
  :probability-queries '(prob(A / (C & D))
                         prob(B / (C & D)))).
```

References

1. F. Bacchus, *Representing and Reasoning with Probabilistic Knowledge*, MIT Press, 1990.
2. F. Bacchus, A. J. Grove, J. Y. Halpern, D. Koller, "From statistical knowledge bases to degrees of belief", *Artificial Intelligence* 87 (1996), 75-143.
3. R. Booth, "Social contraction and belief negotiation", *Information Fusion* 7 (2006), 19-34.
4. R. B. Braithwaite, *Scientific Explanation*. Cambridge: Cambridge University Press, 1953.
5. J. Dombi, "Basic concepts for a theory of evaluation: The aggregative operator", *European Journal of Operational Research* 10 (1982), 282-293.
6. R. A. Fisher, "On the mathematical foundations of theoretical statistics." *Philosophical Transactions of the Royal Society A*, 222 (1922), 309-368.
7. J. Fodor, R. Yager, A. Rybalov, "Structure of uninorms", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5 (1997), 411 - 427.
8. E. Grégoire, S. Konieczny, "Logic-based approaches to information fusion", *Information Fusion* 7 (2006), 4-18.
9. R. Haenni, S. Hartmann, "Modeling partially reliable information sources: A general approach based on Dempster–Shafer theory", *Information Fusion* 7 (2006), 361-379.
10. J. Y. Halpern, "An analysis of first-order logics of probability", *Artificial Intelligence* 46 (1990), 311-350.
11. A. Hunter, W. Liu, "Fusion rules for merging uncertain information", *Information Fusion* 7 (2006), 97-134.
12. L. R. M. Johansson, R. Suzić, "Bridging the gap between information need and information acquisition", *Seventh International Conference on Information Fusion*, July, 2004, 1202-1209.
13. E. P. Klement, R. Mesiar, E. Pap, "On the relationship of associative compensatory operators to triangular norms and conorms", *Int J. of Unc. Fuzz. and Knowledge-Based Systems* 4 (1996), 129-144.
14. W. Kneale, *Probability and Induction*. Oxford: Oxford University Press, 1949.
15. H. Kyburg, Jr.,*The Logical Foundations of Statistical Inference*, Dordrecht: Reidel, 1974.
16. H. Kyburg, Jr., "Propensities and probabilities." *British Journal for the Philosophy of Science* 25 (1974), 321-353.
17. E. Lefevre, O. Colot, P. Vannoorenberghe, "Belief function combination and conflict management", *Information Fusion* 3 (2002), 149-162.

18. B. Parhami, "A taxonomy of voting schemes For data fusion and dependable computation", *Reliability Engineering and System Safety*, 52 (1996), 139-151.
19. J. L. Pollock, *Nomic Probability and the Foundations of Induction*, New York: Oxford University Press, 1990.
20. J. L. Pollock, *Thinking about Acting: Logical Foundations for Rational Decision Making*, New York: Oxford University Press, 2006.
21. J. L. Pollock, "Defeasible reasoning", in *Reasoning: Studies of Human Inference and its Foundations*, in J. Adler, L. Rips (ed), Cambridge: Cambridge University Press, 2006.
22. J. L. Pollock, "Probable probabilities", OSCAR Project technical report, available at http://oscarhome.soc-sci.arizona.edu/ftp/PAPERS/Probable%20Probabilities.pdf. 2007.
23. J. L. Pollock, "Reasoning defeasibly about probabilities", in M. O'Rourke, J. Cambell (eds.), *Knowledge and Skepticism*, Cambridge, MA: MIT Press, 2008.
24. K. Popper, "The propensity interpretation of probability." *British Journal for the Philosophy of Science* 10 (1956), 25-42.
25. K. Popper, "The propensity interpretation of the calculus of probability, and the quantum theory." In *Observation and Interpretation*, S. Körner (Ed.), 65-70. New York: Academic Press, 1957.
26. K. Popper, *The Logic of Scientific Discovery*, New York: Basic Books, 1959.
27. F. Ramsey, "Truth and probability", in *The Foundations of Mathematics*, ed. R. B. Braithwaite, Paterson, NJ: Littlefield, Adams, 1926.
28. H. Reichenbach, *A Theory of Probability*, Berkeley: University of California Press, 1949. (Original German edition 1935.)
29. B. Russell, *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster, 1948.
30. L. Savage, *The Foundations of Statistics*, Dover, New York, 1954.
31. G. Shafer, *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.
32. L. Sklar, "Is propensity a dispositional concept?" *Journal of Philosophy* 67 (1970), 355-366.
33. L. Sklar, "Unfair to frequencies." *Journal of Philosophy* 70 (1973), 41-52.
34. A. Urken, "Using collective decision system support to manage error in wireless sensor fusion", *Eighth International Conference on Information Fusion*, July, 2005.
35. A. Urken, "Error-resilient collective inference: Theory and preliminary experimental results", *First World Conference on Public Choice*, Amsterdam, March, 2007.
36. B. van Fraassen, *The Scientific Image*. Oxford: Oxford University Press, 1981.
37. J. Venn, *The Logic of Chance*, 3rd ed. London, 1888.