LA-UR -93-3336

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: AN ADDITIVE THEORY OF BAYESIAN EVIDENCE ACCRUAL

AUTHOR(S): C. LARRABEE WINTER
MICHAEL C. STEIN

SUBMITTED TO: ADVANCES IN APPLIED MATHEMATICS

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

An Additive Theory of Bayesian Evidence Accrual

Michael C. Stein and C.L. Winter

Analysis and Assessment Division

Los Alamos National Laboratory

Los Alamos, New Mexico 87545

stein@agps.lanl.gov or winter@agps.lanl.gov

(505) 667-3733

Abstract. We derive a theory of data fusion based on an additive approach to Bayesian evidence combination and accrual. Although the additive method can be stated in terms of simple formulae of probability, it is surprisingly rich. It is robust against errors in data, and analysis and numerical simulations indicate that estimated probabilities of hypotheses converge to the expected value of a multiplicative Bayesian update as evidence that is mostly (but not necessarily entirely) correct is accrued. We summarize the method and principal results in the first part of the paper. The method relies on a representation theorem for expected values of uncertain probabilities that is an extension of a theorem of deFinetti's (deFinettii, 1937). The extension states that the expected value of a function of uncertain probabilities can be represented as a weighted sum of exchangeable random variables. We use the extended theorem to show that the additive method approximates the expected value of the ordinary Bayesian posterior, and they are equal in the limit. In the second part of the paper, we sketch proofs of our theorems, derive the additive rule and contrast the additive approach with others, especially multiplicative Bayesian updating on one hand and various consensus-based rules on the other. We show that the additive approach is much less sensitive to anomalous data than is Bayesian updating. The additive method, while similar in spirit to consensus approaches, is not ad hoc.

1.0 Summary.

Many automated systems combine, or accrue, evidence from diverse sources to develop support for hypotheses about a state of nature. In image analysis, for instance, evidence may come from imaging sensors, human intelligence, spatial databases and signal analysis to name a few sources; hypotheses can correspond to the presence or absence of objects of interest at particular locations, and perhaps their disposition. Object recognition systems typically combine the output of multiple feature detectors to support hypotheses about the existence and identification of a target in a single image. Geological survey systems, oil exploration systems for example, sometimes pool the opinions of different experts to estimate the probability that a geological feature will be found in a given region.

In general, accrual methods should be robust against occasional errors in evidence since evidence sources can be unreliable. In particular, they must be robust against errors that cannot be predicted from properties of an ensemble, but are instead peculiar to a given realization. Consensus approaches address this issue directly by forming a consensus among evidence sources. The notion is that while one, or even a few, sources of evidence may be in error, the majority will not be; thus, the effect of outliers can be minimized. We describe a method for accruing evidence to hypotheses that is based on establishing a probabilistic consensus among evidence sources. Our consensus approach is based on a Bayesian theory of additive accrual that is robust against both missing data and bad evidence.

Probabilistic evidence accrual involves inducing the probability of a hypothesis as evidence is accumulated about a particular realization of a random process. Inference is relative to the state of information or knowledge of the assesor (de Finetti, 1937), and should adapt to available data via an inductive process that is not restricted to properties of an ensemble, but is also sensitive to the unique characteristics of a realization. To support stable decision making, it is critical to make the most robust Bayesian decision possible about a particular realization. Inductive probabilities can

be viewed as random variables whose estimates change with respect to the dynamic state of information about a realization.

Our additive rule for accruing evidence to a hypothesis satisfies these criteria. The rule states that when new evidence, E_N, is accrued to a hypothesis, H, that is already supported by existing evidence, E_O, the updated probability is,

$$P(H|E_O \cup E_N) = \frac{P(H)}{P(E_O \cup E_N)} [P(E_O|H) + P(E_N|H)\ (1 - P(E_O|H))] \qquad (1.1)$$

The right side of (1.1) indicates $P(H|E_O \cup E_N)$ is a scaled version of the likelihood of the old evidence, $P(E_O|H)$, supplemented by the likelihood of the new evidence, $P(E_N|H)$, times whatever evidence remains to be accrued, $1-P(E_O|H)$. In (1.1), P(H) is a prior probability whose value is not based on any of the E_i. Usually E_O is itself the union of previous experiments, $E_O = E_1 \cup ... \cup E_{N-1}$. The internal structure of the E_i may be arbitrarily complex. However, a convenient way to view the output of E_i is as 0-1 random variables where $E_i = 1$ corresponds to the statement that "The evidence obtained from the i^{th} source confirms H." We note that the rule, while additive, can lead to decreasing probability for H if $P(E_N|H)$ is small with respect to $P(E_O \cup E_N)$.

To obtain (1.1) we apply Bayes' Rule and only require that the experiments be conditionally independent, i.e. that $P(E_iE_j|H) = P(E_i|H)P(E_j|H)$ if $i \neq j$. However, to fully understand (1.1) it is useful to note that the experiments are exchangeable (deFinetti, 1937) since exchangeability is implied by independence. A collection of events is exchangeable when their joint distribution depends only on the number of events and not on either their order or the specific events. In the case of an infinite sequence of exchangeable 0-1 random variables, $E_1, E_2, \ldots$ de Finetti's Representation Theorem states that there exists a unique probability distribution function Φ on [0,1] such that

$$P_r^n = \binom{n}{r} \int_0^1 p^r (1-p)^{n-r} d\Phi(p) \qquad (1.2)$$

where P^n_r is the probability of obtaining r 1's out of a collection of experiments E_i , i=1,...,n. Therefore, an exchangeable sequence of random variables can be viewed as a mixture of independent random variables each with constant probability p.

To motivate the additive rule, we first extend de Finetti's Theorem to expected values of functions, f(p), of uncertain probabilities, p: We show in Section 2.4 that for $p \in [0, 1]$, continuous f(p) and distribution function $\Phi(p)$,

$$E[f] = \int_0^1 f(p)\, d\Phi(p) = \lim_{n\to\infty} \sum_{r=0}^{n} f(\frac{r}{n}) \binom{n}{r} P^n_r \tag{1.3}$$

where P^n_r is the probability in (1.2). A special case of the extension is when f(p) = p, then we have,

$$E[x] = \lim_{n\to\infty} \sum_{r=1}^{n} \frac{r}{n} \binom{n}{r} P^n_r \,. \tag{1.4}$$

We also show that $P(E_O \cup E_N|H)$ approaches the expected value of the joint distribution of the uncertain evidence sources $E_1, ..., E_n$. Here uncertainty means that the E_i's are random variables, hence so is $P(E_i|H)$. More specifically, we let $\mathcal{E}^n = \{(E_1 \ldots E_n) |$ at least one $E_i = 1\}$; it is a set of random variables whose elements are possible values of evidence combined conjunctively. Let E^n be any element of $\mathcal{E}^n$. Then

$$\lim_{n\to\infty} [E[P(H|E^n)] - P(H|E_1 \cup \ldots \cup E_n)] = 0 \tag{1.5}$$

We calculate the expected value in 2 stages. First, we average over all realizations in $\mathcal{E}^n$. The probability of a given realization is $P(E_1 = 1, E_2 = 0, ...|H) = P(E_1|H)(1 - P(E_2|H)) \ldots$. Second, by considering all possible combinations and increasing n, we get a large sample of the uncertain

joint probabilities. The method implicitly partitions the joint probabilities into histogram bins. We quickly obtain a good approximation of the distribution function of the uncertain probabilities, Φ, through this implicit histogramming.

Before contrasting the additive rule with other methods of evidence combination, we discuss an important distinction between types of evidence. Evidence for a hypothesis can be strong or weak. Strong evidence is always present in data sets that satisfy a hypothesis, while weak evidence may be absent. Evidence used in object recognition systems illustrates this distinction. Suppose the problem is to identify automobiles in some class of images: While it is true that almost all autos have wheels, it is not certain that wheels can be seen in every image of every auto. Wheels may be hidden or unobservable depending on the angle of view. In many practical problems, weak evidence is the dominant form of evidence. In the first place, the ability to observe nearly all evidence is contingent on the realized data set, as the wheel example indicates. Our additive rule reflects that view since the total evidence for a hypothesis is the normalized sum of evidence obtained from many sources. The report of a single source, or even a few sources, is not enough to completely alter $P(H| \cup_i E_i)$ when it contradicts the majority of sources.

We demonstrate that the additive approach to accrual (1.1) is much more stable with regard to step-wise variations in evidence than is a multiplicative approach like standard Bayesian updating (cf. Duda and Hart, 1973). Stability in evidence accrual systems is not just a matter of long-term behavior. In many systems, decisions must be based on interim probabilities. If probabilities swing widely from one update to another, such systems may behave erratically. Robust techniques like the additive rule are inherently stable and appropriate for small data sets. The additive rule is also inherently parallel and opportunistic. It is parallel because exchangeable evidence sources can be processed in any order, and it is opportunistic because evidence can be processed if and when it becomes available.

Contrast the robustness of (1.1) with multiplicative Bayesian updating,

$$P(H|E_O E_N) = \frac{P(E_N|H)}{P(E_N|E_O)}(H|E_O) \tag{1.6}$$

where E_O is now the intersection of the results of previous experiments and we have used conditional independence to simplify Bayes' rule (cf. Duda and Hart, 1973). Obviously $P(H|E_O E_N)$ will be small if $P(E_N|H)/P(E_N|E_O)$ is small. For instance, $P(H|E_1 \ldots E_n) = 0$ if any $P(E_i|H) = 0$. To continue the example given earlier, if E_i is an experiment designed to detect wheels in an image but no wheels are visible, then (1.6) sends $P(H|E_1 \ldots E_n)$ to zero even if the image contains an auto. Note that this can happen when all other experiments, e.g. ones that look for doors, tail lights, bumpers, etc., return $P(E_i|H) = 1$. This seems unreasonable given the contingent nature of data. Furthermore, an experiment can return low $P(E_i|H)$ even if the data on which it is based satisfies H; all that is required is a bad experiment, for instance one that does not identify wheels well under some conditions.

Two related methods for overcoming the brittleness of multiplicative updating are often proposed. One approach amounts to computing the values of all possible realizations of the joint probability $P(H|E_1=e_1 \ldots E_n=e_n)$ where $e_i = 1$ or 0 and using some subset of those probabilities to evaluate the certainty in H. Unfortunately, such an approach can be computationally expensive since it's inherent complexity is on the order of 2^n. The additive rule (1.1), on the other hand has complexity O(n), and furthermore computes the expected value of $P(H|E_1=e_1 \ldots E_n=e_n)$ as we have noted.

The second approach supposes that observational contingencies, for instance the probablity of occlusion in image analysis, can themselves be modeled. Note that a model of occlusion would require 1) completely parametrizing the process of occlusion including the size and shape of the occluding object, its position relative to the target, the size and shape of the target, the imaging geometry and so on, and 2) determining the values of the parameters for a given realization. In the

unlikely event that a complete parameterization of occlusion could be obtained, many models would have to be stored and computed. In most cases, determining the values of the parameters of an observational model requires solving a problem of equivalent complexity to the original problem.

Numerical results provide striking evidence for the robustness of the additive update rule versus the multiplicative update rule. Figures 1.1 and 1.2 are plots of the results of choosing at random component probabilities and then inserting them into formula (1.1) for additive updating and formula (1.6) for multiplicative updating. The component probabilities are the probabilities $P(E_O)$ and $P(E_N)$ of getting evidence E_O or E_N, the joint probability $P(E_OE_N)$ of getting E_O and E_N, the joint probabilities $P(E_OH)$ and $P(E_NH)$ of getting E_O and the hypothesis or E_N and the hypothesis, and finally the joint probability $P(E_OE_NH)$ of getting evidence E_O, E_N and the hypothesis.

The component probabilities for two or more evidence gathering experiments and a hypothesis were chosen recursively at random except that they were required to satisfy measure theoretic constraints. The constraints are the following:

$$P(E_OE_N) \leq Minimum(P(E_O), P(E_N)) \tag{1.7}$$

$$P(E_iH) \leq P(E_i), i = O, N \tag{1.8}$$

$$P(E_OE_NH) \leq Minimum(P(E_OH), P(E_NH)) \tag{1.9}$$

The form of the multiplicative and additive rules are equivalent to our previous forms but allow these random probabilities to be input. The form of the multiplicative update rule that we used is

$$P(H|E_OE_N) = \frac{P(E_OE_NH)}{P(E_OE_N)} \tag{1.10}$$

while the form of the additive update rule is

$$P(H|E_O \cup E_N) = \frac{P(E_OH) + P(E_NH) - P(E_OE_NH)}{P(E_O) + P(E_N) - P(E_OE_N)} \tag{1.11}$$

Simulations show the results of iterating the additive rule and multiplicative rules for 3, 17, and 129 experiments; that is, we initialize each simulation with a randomly chosen E_O and then accrue the results of 2, 16 or 128 new experiments whose values are also chosen at random. Values were selected from a uniform distribution on [0,1] with restrictions given by (1.7)-(1.9). Figure 1.1 is a plot of the results from 100,000 runs of the additive rule and Figure 1.2 is a plot using the same number of runs for the multiplicative rule. The central limit theorem-type convergence with increasing number of experiments, which is expected from our theorem (1.5), is clearly evident in Figure 1.1 while no difference is discernible in Figure 1.2. Also, the same tendency toward low values of the iterated update is found for the multiplicative rule while the additive updates converge toward what one would expect for random choices of hypothesis and evidence probabilities.

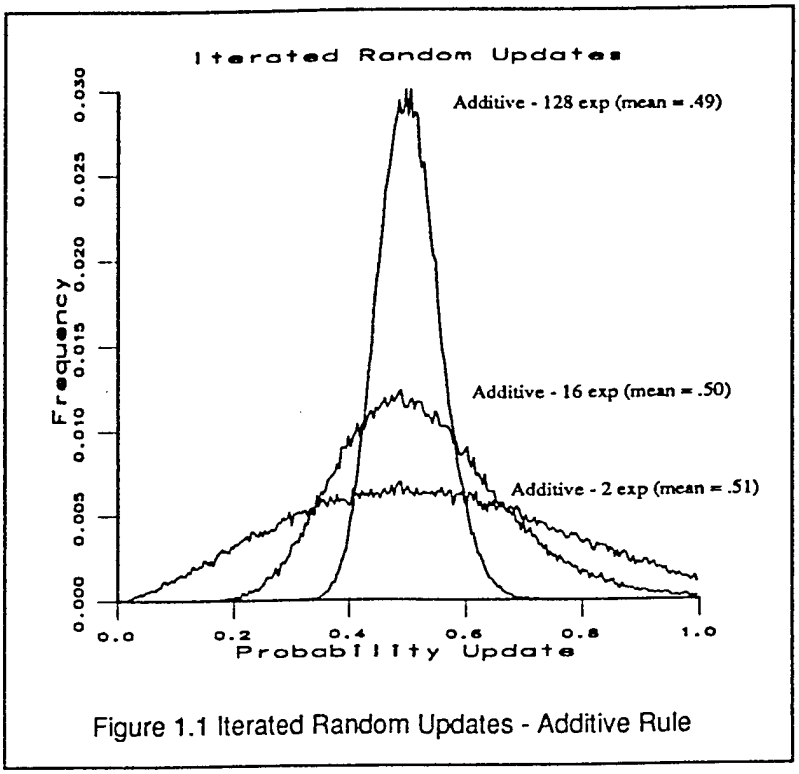


Figure 1.1 Iterated Random Updates - Additive Rule

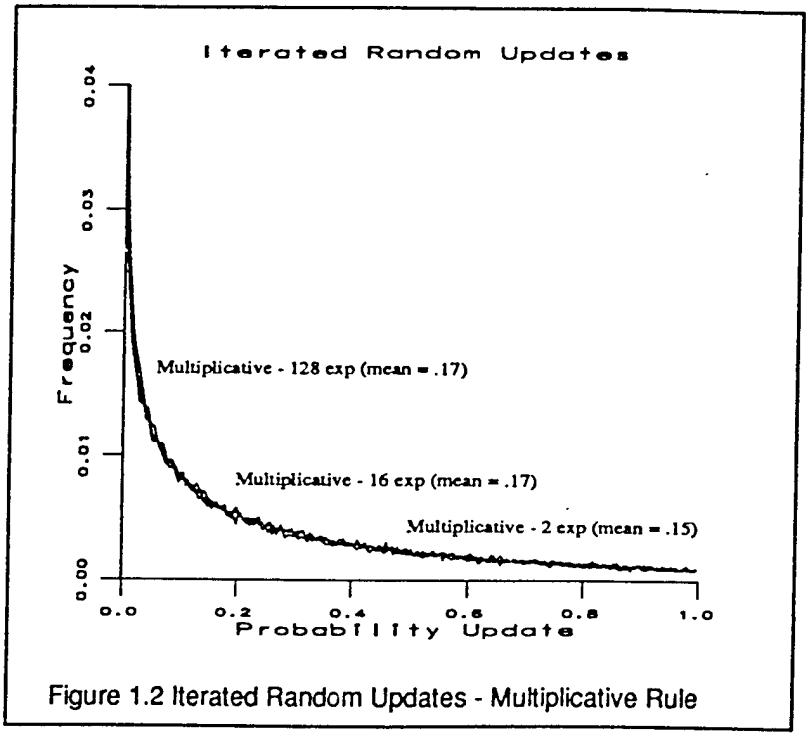


Figure 1.2 Iterated Random Updates - Multiplicative Rule

The conservatism of the additive rule is also evident when we compare likelihood ratios based on the two rules (Figure 1.3). There we contrast the relative effect of new evidence on likelihood ratios,

$$\Lambda^M = \frac{P(E_N|H)\,P(E_O|H)}{P(E_N|\overline{H})\,P(E_O|\overline{H})} \tag{1.12}$$

derived from the Multiplicative Rule and,

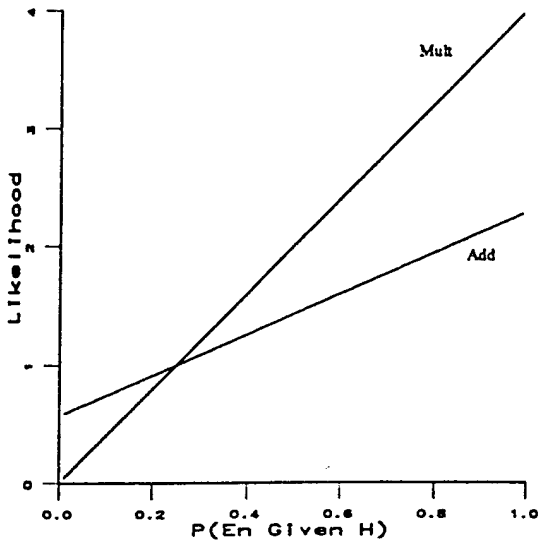
$$\Lambda^A = \frac{[P(E_O|H) + P(E_N|H)\ (1-P(E_O|H))]}{[P(E_O|\overline{H}) + P(E_N|\overline{H})\ (1-P(E_O|\overline{H}))]} \tag{1.13}$$

derived from the Additive Rule. To simplify the discussion, we have dropped the ratio $P(H)/P(\overline{H})$ that multiplies both Λ^M and Λ^A.

To obtain Figure 1.3, we suppose that both rules start with the same prior likelihood,

$$\Lambda_O = \frac{P(E_O|H)}{P(E_O|\overline{H})} \cdot \tag{1.14}$$

We also fix $P(E_O|H) = 0.25$ since the Additive Rule requires those values explicitly, we take $P(E_N|\overline{H}) = 0.25$ and we let $\Lambda_O = 1.0$, i.e. the prior likelihood is indifferent between H and $\overline{H}$. Other values of $P(E_O|H)$, $P(E_N|H)$ and Λ_O yield qualitatively similar results, a fact that we discuss at length in Section 4. Clearly Λ^A is restricted to a narrower range than Λ^M. Furthermore, Λ^M drops to values less than 1 very quickly. Large (small) values of $P(E_N|H)$ lead to Λ^M that is much larger (smaller) than Λ_O, while Λ^A stays closer to Λ_O throughout the range of $P(E_N|H)$. This is consistent with our earlier observation that the Additive Rule gives much less weight to outliers than does the Multiplicative Rule.



	Λ^A	Λ^M
Max	2.27	4.0
Min	0.59	0.0
Slope	1.68	4.0

$\Lambda_O = 1.0$ $P(E_O|H) = 0.25$ $P(E_N|\overline{H}) = 0.25$

Figure 1.3 -- Additive vs. Multiplicative Rule: Low $P(E_O|H)$

Although the additive rule is appropriate in many practical problems, the additive and multiplicative rules can be combined to yield

$$P(H|E_S(E_1 \cup E_2)) = \left[\frac{P(E_S|H)}{P(E_S)}\right] P(H|E_1 \cup E_2) \qquad (1.15)$$

The combined rule looks like Bayes' rule with $P(H|E_1 \cup E_2)$ as a prior. The effect of strong evidence, E_S, is to scale $P(H|E_1 \cup E_2)$ by the Bayesian likelihood of E_S, $P(E_S|H)/P(E_S)$. We give an algorithm for (1.15) below.

Additive approaches to updating are not new. Our current work is related to Jeffrey's rule (Jeffrey, 1965; Diaconis and Zabell, 1982) and can in fact be used to derive his rule when the E_i form a partition of the entire sample space. Our rule is closely related to the updating methods discussed by Winter, Ryan and Hunt (1986), but they neglected the normalization, which is critical if $P(H|\cup_i E_i)$ is to decrease as well as increase as evidence is accrued. The activation of a "neuron" in an artificial neural systems is usually based on a weighted consensus of other neurons, and equation (1.1) can be rewritten to look like the activity of such a neuron. When written in that form, (1.1) suggests an adaptive method for learning the properties of the transformations $P(E_i|H)$. Consensus rules (Berenstein, Kanal and Lavine, 1986; deGroot, 1974) use weighted sums of probabilities to represent support for hypotheses.

The remainder of the paper is essentially a set of appendices to this summary. It is organized as follows: In Section 2 we discuss assumptions and define a few terms, specifically i) conditional independence, ii) exchangeability, iii) weak and strong evidence. We also restate theorems (1.3) and (1.5) more formally and sketch their proofs. Proofs are given in full in (Stein and Winter, in prep.). In Section 3 we derive (1.1) in 4 different ways because each derivation illustrates a different aspect of the rule. The first derivation depends on straightforward applications of Bayes' Rule and conditional independence. We use the second derivation to show that (1.1) decreases when new evidence does not strongly support H. The third derivation is the basis for the proof of our second theorem (1.5). The fourth derivation relates (1.1) to the expected value of an indicator

function.

In Section 4 we compare the properties of the additive rule and multiplicative rule through numerical results similar to those of Figures 1.1-1.3. The distribution of simulated updates of the additive rule tightens as the number of experiments increases while the distribution of multiplicative updates does not change. We also discuss the effect of new data on likelihoods, Λ^A, and Λ^M. The additional results further confirm Figure 1.3: Λ^M is less stable than Λ^A in the sense that small differences in new data can result in much larger changes in Λ^M than Λ^A. Additionally, we note that Λ^A is affected by the magnitude of $P(E_O|H)$ and thus preserves some information about the absolute goodness of the hypothesis while Λ^M loses such information.

In Section 5 we relate the additive rule to other additive approaches, specifically consensus rules, Jeffrey's rule and neural networks. We indicate the additive rule is identical to Jeffrey's rule when the evidence sources, $\{E_i\}$, constitute a partition of the sample space. Section 6 outlines a few issues for future research.

2 Background.

A word about notation: Where it is not ambiguous we use X to indicate that $X = 1$ and $\overline{X}$ to indicate $X = 0$.

We are interested in problem domains in which a collection of diverse algorithms, or experiments, can report evidence about a hypothesis. In many cases only a subset of experiments may report, and furthermore experiments may report in any order. We assume that experiments are basically good in the sense that they do discriminate between H and $\overline{H}$ in the absence of contingent errors. Specifically, this means $P(\overline{E}|H) << P(\overline{E}|\overline{H})$ and $P(E|\overline{H}) << P(E|H)$. Although we assume experiments are good discriminators, any individual experiment in a realized sequence of experiments may be unreliable. That is, the output of an experiment may not conform to the true state of nature

because of a variety of error sources. On the other hand, we assume that most experiments agree when they are applied to a given event.

2.1 Conditional Independence

Our updating rule is based on the rather weak assumption that experiments are conditionally independent of each other. This amounts to claiming that $P(E_iE_j|H) = P(E_i|H)P(E_j|H)$, or equivalently, that $P(E_i|HE_j) = P(E_i|H)$. In most cases these are reasonable claims about the parameterization of the distribution, P. The second, for instance, says that knowing H suffices to define P, and that additional evidence, E_j, is not useful in parametrizing P.

To get some intuition about conditional independence, consider the case of flipping a coin and trying to predict whether the i^{th} flip will be a head. Suppose H is the statement "The coin is fair," and E_j is a set of flips performed previously. $P(E_i|HE_j)$ amounts to asking, "What is the probability of getting a head (or tail) on the i^{th} flip given that the coin is fair and we have already obtained the sequence of heads and tails contained in E_j?" Clearly the evidence, E_j, adds nothing to the definition of this probability. To determine E_i,all we need know is that the coin is fair, i.e. $P(E_i|HE_j) = P(E_i|H) = 1/2$.

2.2 Exchangeability

A sequence of random variables is exchangeable if the joint probability P satisfies

$$P(E_1 = e_1, E_2 = e_2, \ldots, E_n = e_n) = P(E_{\Pi(1)} = e_1, E_{\Pi(2)} = e_2, \ldots, E_{\Pi(n)} = e_n) \qquad (2.1)$$

where π is a permutation on n indices. This type of probability measure is called symmetric and has been studied by deFinetti (1937, 1964), and was fully treated by Hewitt and Savage (1955). Another way to describe a sequence of exchangeable random variables is to say that the order does not matter to the limiting joint probability distribution.

An important characteristic of many types of evidence sources is that the order of receipt of evi-

dence should not affect the conditional probability of the hypothesis given this evidence. This is important because the order of receipt of evidence may not be the same as the time ordering of the evidence and for certain types of evidence the time ordering is not significant. For example, suppose we are trying to identify an automobile and we receive evidence that it has a convertible top and then we receive evidence that it has wire wheels. It should not make a difference in what order we combine the evidence to the conditional probability that we have a specific kind of automobile given the evidence. Our updating scheme leads us to consider exchangeable random variables.

As shown by de Finetti (1937, 1964) and Hewitt and Savage (1955), a symmetric measure may be represented more simply as a mixture of independent power distributions. The mixture is created by integrating the power distributions over a random probability distribution (see Dubins and Freedman (1967)) on the power distributions. That is,

$$P(E_i \in A_i) = \int_{\Omega} \prod_i P(A_i)\, d\mu(P) \tag{2.2}$$

for all i=1,...,n. Note that $\mu(P)$ is a random probability measure over the set of probability measures on the sample space Ω of the random variables. This result, usually called de Finetti's Representation Theorem or just the Representation Theorem, takes a simpler form in the case of 0-1 or Bernoulli random variables. That is, there exists a unique probability measure on the Borel sets of [0,1] such that

$$P(E_i = e_i)_{i=1,\ldots,k} = \int_{[0,1]} p^j (1-p)^{k-j} \mu(dp) \tag{2.3}$$

where e_i is either 0 or 1 and $j = \Sigma e_i$.

A variant of the Representation Theorem holds even if the sequence is finite. Suppose that k is much smaller than n and $E_1,\ldots,E_k$ is the beginning of a long exchangeable sequence $E_1,\ldots,E_k,E_{k+1},\ldots,E_n$. In that case, (2.3) is approximately true with an error that is essentially on the

order of k/n as shown by Diaconis and Freedman (1980).

2.3 Updating With Weak Evidence

An important distinction of the additive update procedure from the typical update using the multiplicative rule is that we consider union or disjunction of evidence and the typical scheme looks only at the intersection or conjunction. Notice that the union of evidence includes the intersection as well as other regions of the sample space that have not been covered by previous evidence. Evidence comes in one of two forms, and the form is a guide as to whether the update should be done using union or intersection of the new evidence with the old. These two forms we call weak and strong evidence.

Strong evidence is a probabilistic statement about a condition that *must* or *must not* be satisfied by every realization of a random process. For example, when trying to recognize an automobile in imagery, it is useful to remember that they are practically never found in water. Strong evidence, such as the fact that an object is by itself in the middle of a deep lake, should allow us to conclude that it is not an auto. Because strong evidence refers to conditions all of which must be considered, the conjunction of the strong evidence is appropriate and this leads to the normal method for multiplicative Bayesian updating.

Weak evidence is a probabilistic statement about a condition that *may* or *may not* be satisfied. In the auto example, the size of an object may imply that it is a car. But several types of trucks that could be in the scene may be the same size as a car. Also, the possibility of occlusion of a critical component such as wheels requires that we not draw conclusions from the absence of a component. Because weak evidence refers to conditions only some of which may be considered, the disjunction of the evidence is appropriate and this leads to the method of additive updating we discuss in this paper.

Although the two types of updating may be combined (see Section 3.2) we believe that most evidence is weak evidence. The uncertainty associated with information gathering algorithms and processes always allows for the possibility that critical components are missed. Also, as was learned from the knowledge representation activity that went on in AI research for many years, it is very difficult to completely and uniquely identify objects and situations by a reductionist listing of the attributes or components that must make up the object or situation. Thus, a lot of evidence is weak because the object or situation that it is applied to is not uniquely or completely specified by components about which information can be gathered.

Furthermore, evidence can be weak simply because experiments fail. We say a supporting experiment fails if $P(E_i|H=T)$ is small. Supporting experiments can fail for at least 2 reasons. First, a data set may be a member of H yet it may not contain data to support the experiment. This is the problem of missing data. A very common example is the effect of occlusion in image analysis; an experiment designed to recognize human faces may fail on an individual face if the subject wears a stocking cap pulled low on his head, thus obscuring ears and eyebrows. No matter how good an experiment may be, it must fail if the data on which it is based is missing. Second, it must be admitted that experiments can fail just because they are bad, i.e. an experiment may not correctly classify a data set even when the data to support the experiment is available. We call this the problem of systematic error in a supporting experiment.

2.4 Theorems

As noted, we just state our theorems here and sketch their proofs. We give complete proofs in (Stein and Winter, in prep.)

Theorem 1: Extended Representation Theorem. For $p \in [0, 1]$, any continuous function f(p) and a distribution function $\Phi(p)$,

$$E[f] = \lim_{n \to \infty} \sum_{r=0}^{n} f(\frac{r}{n}) \binom{n}{r} P_r^n , \qquad (2.4)$$

where P_r^n is the probability of obtaining r successes in n trials selected from a population of exchangeable random variables.

The theorem states that the expected value of any function of an induced probability can be represented in terms of exchangeable variables. It follows from a few simple facts. First, obviously

$$E[f] = \int_0^1 f(p)\, d\Phi(p) \ . \tag{2.5}$$

Next we can rewrite f(p) in terms of its Bernstein Series,

$$f(p) = \sum_{r=0}^{n} f(\tfrac{r}{n}) \binom{n}{r} x^r (1-x)^{n-r} \ , \tag{2.6}$$

so

$$E[f] = \int_0^1 \lim_{n\to\infty} \left(\sum_{r=0}^{n} f(\tfrac{r}{n}) \binom{n}{r} x^r (1-x)^{n-r} \right) d\Phi(x) \ . \tag{2.7}$$

We apply uniform convergence to exchange the limit and integral in (2.7) and then use deFinetti's Theorem to get

$$\lim_{n\to\infty} \sum_{r=0}^{n} f(\tfrac{r}{n}) \binom{n}{r} \int_0^1 x^r (1-x)^{n-r} d\Phi(x) = \lim_{n\to\infty} \sum_{r=0}^{n} f(\tfrac{r}{n}) \binom{n}{r} P_r^n \ . \tag{2.8}$$

Corollary. If f(p) = p,

$$E[f] = E[p] = \lim_{n\to\infty} \sum_{r=0}^{n} \frac{r}{n} \binom{n}{r} P_r^n = \lim_{n\to\infty} \sum_{r=1}^{n} \frac{r}{n} \binom{n}{r} P_r^n \ . \tag{2.9}$$

Theorem 2: Expected Multiplicative Update. Let $\mathcal{E}^n = \{(E_1 \ldots E_n) |$ at least one $E_i = 1\}$; it is a set of random variables whose elements are possible values of evidence combined multiplicatively. Let E^n be any element of $\mathcal{E}^n$. Then

$$\lim_{n\to\infty} [E[P(H|E^n)] - P(H|E_1 \cup \ldots \cup E_n)] = 0 \tag{2.10}$$

We denote $E_i = 1$ by E_i and $E_i = 0$ by $\overline{E}_i$. The critical term in the additive rule is $P(E_1 \cup \ldots \cup E_N|H)$,

which may be re-written

$$P(E_1 \cup \ldots \cup E_n | H) = \sum_{r=1}^{n} \sum_{\Pi \in S_n} P(E_{i_1} \ldots E_{i_k} \bar{E}_{i_{k+1}} \ldots \bar{E}_{i_n} | H) \tag{2.11}$$

where S_π is the set of all permutations that contain r 1's.

Since at least one $E_i = 1$ in every term on the right of (2.11), we have

$$\sum_{r=1}^{n} \sum_{\Pi \in S_n} P(E_{i_1} \ldots E_{i_k} \bar{E}_{i_{k+1}} \ldots \bar{E}_{i_n} | H) = \sum_{k=1}^{n} \binom{n-1}{k-1} \sum_{l=1}^{n} p_l \prod_{i=1, i \neq l}^{n} p_i \prod_{j=1, j \neq l, j \neq i}^{n} (1-p_j) \tag{2.12}$$

Here we use exchangeability and substitute p_i for $P(E_j = e_i = 1|H)$ and $1-p_i$ for $P(E_j = e_i = 0|H)$.

We continue with

$$\sum_{r=1}^{n} \binom{n-1}{r-1} \sum_{l=1}^{n} p_l \prod_{i=1, i \neq l}^{n} p_i \prod_{j=1, j \neq l, j \neq i}^{n} (1-p_j)$$

$$= \sum_{r=1}^{n} \frac{r}{n} \binom{n}{r} \sum_{l=1}^{n} p_l \prod_{i=1, i \neq l}^{n} p_i \prod_{j=1, j \neq l, j \neq i}^{n} (1-p_j) \, . \tag{2.13}$$

The terms $p_l \prod_{i=1, i \neq l}^{n} p_i \prod_{j=1, j \neq l, j \neq i}^{n} (1-p_j)$ give us various estimates of the probability of obtaining r successes in n trials. After we histogram them into m bins we compute the frequency of each bin ϕ_λ. Then we write

$$\sum_{r=1}^{n} \frac{r}{n} \binom{n}{r} \sum_{l=1}^{n} p_l \prod_{i=1, i \neq l}^{n} p_i \prod_{j=1, j \neq l, j \neq i}^{n} (1-p_j) = \sum_{r=1}^{n} \frac{r}{n} \binom{n}{r} \sum_{\lambda=1}^{m} p_\lambda^r (1-p_\lambda)^{n-r} \phi_\lambda \tag{2.14}$$

where p_λ is the value in the λ^{th} bin. By letting $m \to \infty$ and assuming that the empirical density, ϕ_λ, dgoes to the actual density $d\Phi$, we have (2.9), and so are done.

3 Additive Update Rule

First we state the updating rule, and then derive it in 4 different ways. The second subsection describes how the update rule can decrease belief with new evidence. The third subsection discusses a rule that combines the additive and multiplicative update rules for use in applications with both strong and weak evidence.

3.1 Statement of the Rule and Derivations

We state the rule in a form that satisfies all 4 derivations. In particular we require that experiments be conditionally independent and that they be exchangeable. These fairly weak assumptions will be met by most probabilistic accrual systems. However, individual derivations may actually allow even weaker assumptions. For instance, our first and second derivations do not require exchangeability. We note such points in the remarks following each derivation.

Additive Rule for Weak Evidence. If E_O and E_N are sets of experiments that are independent when conditioned on a hypothesis, H, and if the prior probability $P(H) \neq 0$, then the updated probability of H given that E_O has been supplemented by E_N is

$$P(H|E_O \cup E_N) = \frac{P(H)}{P(E_O \cup E_N)} [P(E_O|H) + P(E_N|H)\,(1 - P(E_O|H))] \qquad (3.1)$$

The rule states that the updated probability, $P(H|E_O \cup E_N)$, depends on the sum of $P(E_O|H)$ with $P(E_N|H)(1-P(E_O|H)) = P(E_N\bar{E}_O|H)$. Although this sum is always positive, $P(H|E_O \cup E_N)$ can decrease through the influence of the scaling factor, $P(H)/P(E_O \cup E_N)$, a point we return to below.

An alternative form of the additive rule

$$P(H|E_O \cup E_N) = \frac{P(H)}{P(E_O \cup E_N)} [P(E_O|H) + P(E_N\bar{E}_O|H)] \qquad (3.2)$$

makes it clear that the value of new evidence depends in large part on how redundant it is with existing evidence. The more E_N overlaps E_O, the smaller is E_N's contribution to (3.2). Although we do not require even the assumption of conditional independence to obtain this form, its computational utility is limited. It will almost never be the case in an application that all possible combinations of E_N with $\bar{E}_O$ can be anticipated, much less modeled. However, (3.2) leads to a statement about experimental design that is probably obvious, but we repeat anyway because we think it useful: unless redundancy is required to assure reliability, it is most cost-effective to keep experiments as uncorrelated as possible.

Derivation 1. We can also obtain (3.1) by simply applying Bayes' Rule and the definition of the probability of the union of 2 sets,

$$P(H|E_O \cup E_N) = \frac{P(E_O \cup E_N|H)\,P(H)}{P(E_O \cup E_N)}$$

$$= \frac{P(H)}{P(E_O \cup E_N)}\,[P(E_O|H) + P(E_N|H) - P(E_O E_N|H)] \qquad (3.3)$$

We obtain (3.1) by applying conditional independence, $P(E_O E_N|H) = P(E_O|H)P(E_N|H)$, and simplifying.

Derivation 2. We can also derive the rule from a difference quotient. This derivation emphasizes the dynamic nature of some accrual systems, and is useful in showing that the additive rule can decrease. First, we define some notation,

$$E^n = \bigcup_{i=1}^{n} E_i \quad = E^{n-1} \cup E_n, \qquad (3.4)$$

which leads to a natural expression of the change in probability of evidence,

$$\Delta P(E) = P(E^n) - P(E^{n-1}) = P(E_n \bar{E}^{n-1}) \qquad (3.5)$$

Defining $\Delta P(H|E)$ to conform with (2),

$$\frac{\Delta P(H|E)}{\Delta P(E)} = \frac{P(H|E^n) - P(H|E^{n-1})}{P(E^n) - P(E^{n-1})}$$

$$= \frac{P(H)}{P(E^{n-1} \cup E_n)}\left[\frac{P(E_n \bar{E}^{n-1}|H)}{P(E_n \bar{E}^{n-1})} - \frac{P(E^{n-1}|H)}{P(\bar{E}^{n-1})}\right] \qquad (3.6)$$

Moving terms around and letting $E_O = E^{n-1}$, $E_n = E_N$,

$$P(H|E_O \cup E_N) = P(H|E_O) + \frac{P(H)}{P(E_O \cup E_N)}\left[\frac{P(E_N \bar{E}_O|H)}{P(E_N \bar{E}_O)} - \frac{P(E_O|H)}{P(E_O)}\right]\Delta P(E) \qquad (3.7)$$

$$= \frac{P(H)}{P(E_O \cup E_N)} \left[P(E_O|H) + P(E_N \bar{E}_O|H) \right] \tag{3.8}$$

Clearly, P(H|E) can decrease when new evidence is added since $\Delta P(H|E) < 0$ when

$$P(E_N|H) < \left[\frac{P(E_O|H)}{1 - P(E_O|H)} \right] \left[\frac{P(E_O \cup E_N)}{P(E_O)} - 1 \right] \tag{3.9}$$

and $P(E_N|H)$ can be arbitrarily small. The numerical results in Section 4 further illustrate this point.

Derivation 3. Our main result is an update formula that successively constructs a probability measure over the hypothesis space. One can view this measure as the limiting probability measure for a sequence of exchangeable random variables or their corresponding events that represent the accruing evidence. We consider that the result of each evidence event E_iH yields a conditionally independent sample $P(E_iH)$ from the closed interval [0,1] and that this sample represents the joint probability of having the evidence and the hypothesis. $P(\bar{E}_iH) = P(H) - P(E_iH)$ represents the joint probability of not having the evidence and the hypothesis.

From Equation (2.1) we must have for each choice of k events out of a total of n events

$$P(E_{i_1}H)P(E_{i_2}H)\ldots P(E_{i_k}H) = P(E_{\pi(i_1)}H)P(E_{\pi(i_2)}H)\ldots P(E_{\pi(i_k)}H) \tag{3.10}$$

where π is a permutation of the integers 1, 2, ... ,n and we have used the conditional independence of the individual experiment events. For this to be true we must have for every k of n evidence events

$$P(E_{i_1}H)\ldots P(E_{i_k}H) = \frac{1}{\binom{n}{k}} \sum_{\pi \in S'_n} P(E_{\pi(1)}H)\ldots P(E_{\pi(k)}H)\, P(\bar{E}_{\pi(k+1)}H)\ldots P(\bar{E}_{\pi(n)}H) \tag{3.11}$$

where it should be noted that the $\pi(i_k)$ are permutations on n letters (e.g. $(\pi(3),\pi(5),\pi(k)) = (7,k+2,n)$) and S'_n denotes only those permutations where $\pi(1)<\pi(2)<\ldots<\pi(k)$. Now we consider only those joint events where we have at least one evidence event and the hypothesis to be

consistent with the fact that we did perform experimentation. The expected value of the probability for all these joint events is

$$\sum_{k=1}^{n} \binom{n}{k} (P(E_{i_1}H)P(E_{i_2}H)\ldots P(E_{i_k}H)) \tag{3.12}$$

and it is easy to show that

$$P(H(\bigcup_{k=1}^{n} E_k)) = \sum_{k=1}^{n} \binom{n}{k} (P(E_{i_1}H)P(E_{i_2}H)\ldots P(E_{i_k}H)) \tag{3.13}$$

It is also easy to show that upon rewriting the right hand side of (3.13) we can obtain our update formula. For the sake of brevity we will demonstrate this for only two evidence sources, but the proof for any number of evidence sources is easily derived from an exact but lengthy computation or by mathematical induction. For convenience let $P(H(E_1 \cup E_2))=p_h$ and $P(E_iH)=p_i$. Combining (3.1.3) and (3.13) for the case of two evidence sources we get

$$p_h = \binom{2}{1}\left(\frac{1}{\binom{2}{1}}(p_1(1-p_2)+p_2(1-p_1))\right)+\binom{2}{2}\left(\frac{1}{\binom{2}{2}}(p_1p_2)\right) \tag{3.14}$$

and after multiplying through and collecting terms we get

$$p_h = p_1+p_2(1-p_1) \tag{3.15}$$

or using the value of p_h, p_1, and p_2

$$P(H(E_1 \cup E_2)) = P(E_1H)+P(E_2H)\,(1-P(E_1H)) \tag{3.16}$$

which is the same as our update formula if the probabilities are rewritten in terms of conditionals.

Derivation 4. As a final alternative derivation we obtain our update formula as an expected value of a certain ratio of random variables. Previously we found that we could write our update formula in the form

$$P(H|\bigcup_{k=1}^{n} E_k) = \frac{P(\bigcup_{k=1}^{n} E_k|H)}{P(\bigcup_{k=1}^{n} E_k)} P(H) = \frac{\sum_{\pi,k} P(E_{\pi(1)} \cdots E_{\pi(k)} \bar{E}_{\pi(k+1)} \cdots \bar{E}_{\pi(n)}|H)}{\sum_{\pi,k} P(E_{\pi(1)} \cdots E_{\pi(k)} \bar{E}_{\pi(k+1)} \cdots \bar{E}_{\pi(n)})} P(H) \tag{3.17}$$

where the intersection events form a partition of the sample space excluding the all evidence complement sets (i.e. $\bar{E}_{\pi(1)} \cdots \bar{E}_{\pi(k)} \bar{E}_{\pi(k+1)} \cdots \bar{E}_{\pi(n)}$) and where π and k are as defined in Derivation 3. Now if we let $\Im_n$ be the σ-algebra generated by the partition and the hypothesis event H we can rewrite this as

$$\frac{\sum_{\pi,k} P(E_{\pi(1)} \cdots E_{\pi(k)} \bar{E}_{\pi(k+1)} \cdots \bar{E}_{\pi(n)}|H)}{\sum_{\pi,k} P(E_{\pi(1)} \cdots E_{\pi(k)} \bar{E}_{\pi(k+1)} \cdots \bar{E}_{\pi(n)})} P(H) = E\left(\frac{\sum_{\pi,k} 1_{(\cap E)H}}{\sum_{\pi,k} 1_{\cap E}} \middle| \Im_n \right) \tag{3.18}$$

where $\cap$EH and $\cap$E are shorthand for the intersection sets of the partition and $1_{\cap EH}$ and $1_{\cap E}$ are their indicator functions. Thus we have rewritten the result of our update formula as an expectation, which by elementary martingale theory implies that the result of our additive update process is a martingale. This allows a lot of powerful theoretical results to be applied to the investigation of the properties of our method. In **Section 5.2** we further rewrite our update rule to relate it to a modern branch of martingale theory about multiplicative random processes.

3.2 Combined Rule

Although most problem domains are based on weak evidence, some contain strong evidence. Thus we note a simple method for combining strong evidence, E_S, with weak and vice versa.

Algorithm for Combining Weak and Strong Evidence. If E_S is independent of E_1 and E_2, and E_1 and E_2 are conditionally independent events, then the probability of H given that E_S must be observed and that we can observe either E_1 or E_2 (or both) is

$$P(H|E_S(E_1 \cup E_2)) = \left[\frac{P(E_S|H)}{P(E_S)}\right] P(H|E_1 \cup E_2)$$

$$= \left[\frac{P(E_S|H)}{P(E_S)}\right]\left[\frac{P(H)}{P(E_1 \cup E_2)}\left[P(E_1|H) + P(E_2|H)\,(1 - P(E_1|H))\right]\right] \qquad (3.19)$$

i.e., the new rule is the product of the Bayesian likelihood ratio of E_S with the Additive Rule.

When additional strong evidence is obtained, it is fused into $P(H|\bullet)$ by applying ordinary multiplicative Bayesian updating. New weak evidence is accrued to $P(H|E_1 \cup E_2)$ by applying our Weak Rule. The basic algorithm is depicted in Figure 3.1. Strong and weak evidence streams are maintained separately and are updated by respectively the multiplicative or additive rules. When new evidence is obtained, it is first accrued to the appropriate stream, and then the streams are combined according to (3.19).

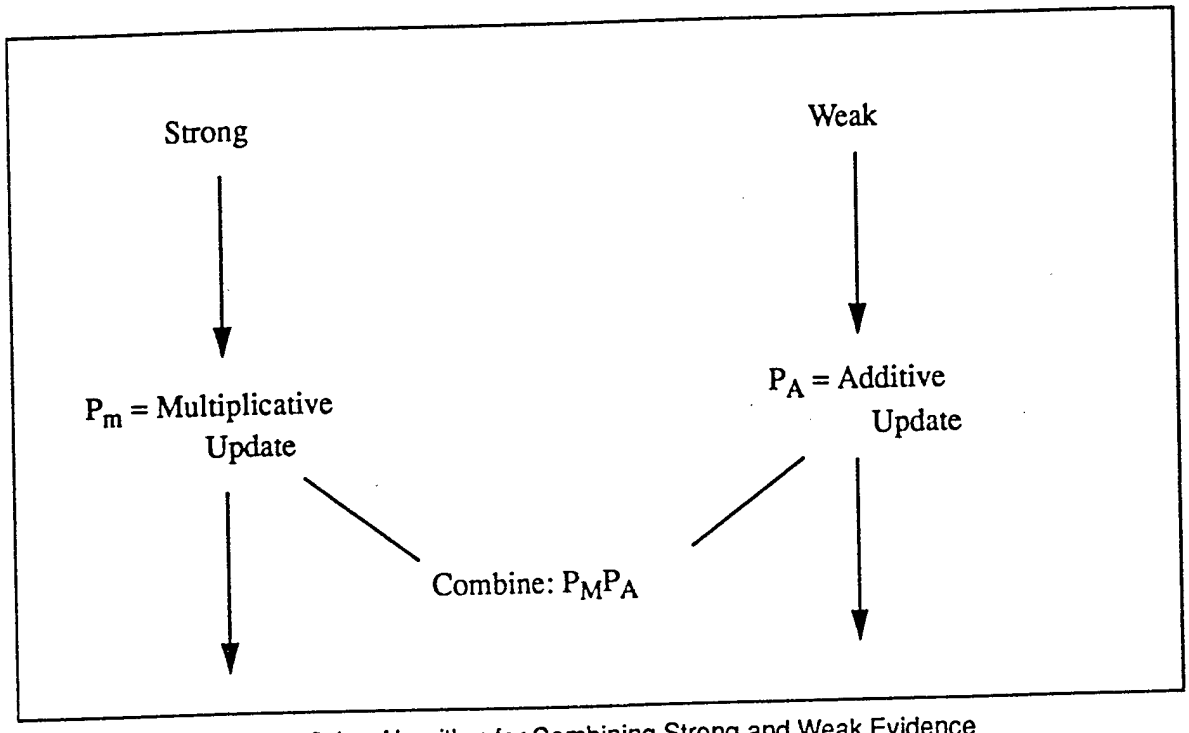


Figure 3.1 -- Algorithm for Combining Strong and Weak Evidence

4 Comparison With Multiplicative Bayesian Updating

In this section we compare the additive rule to several alternative methods of evidence accrual.

We begin by contrasting the additive rule with multiplicative Bayesian updating. The additive rule is Bayesian, but of course it is additive, not multiplicative. Hence it is not as sensitive to anomalous evidence as is ordinary multiplicative Bayesian updating, a fact that we discuss through analytical and numerical results. Furthermore, numerical results indicate that the additive rule converges to the actual value of P(H) as evidence is accrued.

The Multiplicative Rule is based on the notion that all evidence is strong, and therefore, that every experiment can find the data it requires in a given data set. Multiplicative evidence accumulation consists of progressively restricting attention to just those individuals that are strongly supported by all experiments. It is implicitly assumed that individuals that satisfy the hypothesis will have strong support from all experiments. We have already argued that this is unrealistic. Even data sets drawn from objects of interest may not contain data required to support some experiments. Furthermore, the Multiplicative Rule assumes that every experiment, E_i, is good in the sense that if the data required by E_i is in the data set, then $P(E_i|H) >> 0$ and $P(E_i|\overline{H}) << 1$.

4.1 Analysis.

The additive update rule has been previously written with the union of evidence expanded using the inclusion exclusion principle, that is

$$P(\bigcup_{k=0}^{n} E_k) = \sum_i P(E_i) - \sum_{i<j} P(E_i E_j) + \ldots + (-1)^{n+1} P(E_0 E_1 \ldots E_n) \qquad (4.1)$$

Alternatively, we could have expanded the probability of the union of evidence as a partition

$$P(\bigcup_{k=0}^{n} E_k) = \sum_{k=0}^{n} P(\overline{E_0 E_1 \ldots} E_k) \qquad (4.2)$$

Using (4.2) we can rewrite the additive update rule as

$$P(H|\bigcup_{k=0}^{n} E_k) = \frac{P(H)}{P(\bigcup_{k=0}^{n} E_k)} \left[\sum_{k=0}^{n} P(\overline{E_0 E_1 \ldots} E_k|H) \right] \qquad (4.3)$$

For comparison we recall the multiplicative rule in a similar form

$$P(H|\bigcap_{k=1}^{n} E_k) = \frac{P(H)}{P(\bigcap_{k=1}^{n} E_k)}\left[\prod_{k=0}^{n} P(E_k|H)\right] \tag{4.4}$$

These two equations clearly display some primary differences between the two updating schemes. First, the limiting behavior of the additive rule is governed by a sum, which is relatively stable with respect to variations in individual terms, versus the multiplicative rule which is governed by a product that is highly variable due to variations in individual terms. In fact, a worst case for the multiplicative rule is where one of the $P(E_i|H)$ terms is equal to zero forcing all subsequent updates to be equal to zero. As we have pointed out earlier, this conditional probability could be zero or near zero for a variety of reasons and is in fact the reason a more robust update formula is needed. Another obvious difference is in the normalizing terms. The additive rule has a normalizing term which is monotonically increasing and approaching at most the value 1. The multiplicative rule has a normalizing term which is monotonically decreasing and approaching the value 0. Thus variations in the selection and ordering of the evidence experiments E_i may create large fluctuations in the value of the update.

4.2 Probability Update Simulations.

Section 4.1 compared several mathematical properties for the multiplicative and additive rules. In this section we want to present some numerical results that provide striking evidence for the robustness of the additive update rule versus the multiplicative update rule.

Figure 4.1 is a plot of the results of choosing at random the component probabilities and then plugging these into the two formulas for multiplicative and additive updating.

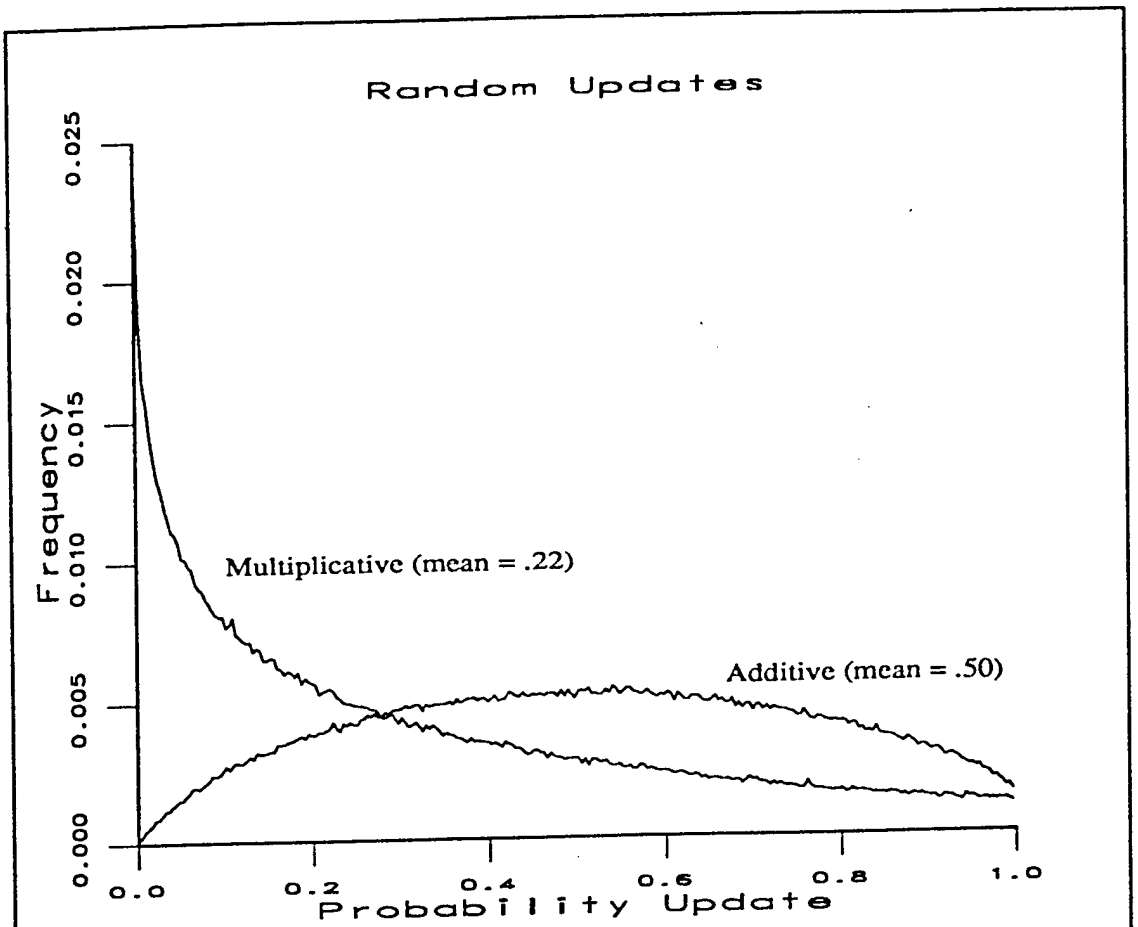


Figure 4.1 Simulation of random updates using the multiplicative and additive rules

The component probabilities for two evidence gathering experiments and a hypothesis are chosen at random with appropriate conditions on some of the probabilities. Specifically these component probabilities are the probabilities $P(E_1)$ and $P(E_2)$ of getting evidence E_1 or E_2, the joint probability $P(E_1E_2)$ of getting E_1 and E_2, the joint probabilities $P(E_1H)$ and $P(E_2H)$ of getting E_1 and the hypothesis or E_2 and the hypothesis, and finally the joint probability $P(E_1E_2H)$ of getting evidence E_1, E_2 and the hypothesis. The conditions are the following:

$$P(E_1E_2) \leq Minimum(P(E_1), P(E_2)) \tag{4.5}$$

$$P(E_1H) \leq P(E_1) \tag{4.6}$$

$$P(E_2H) \leq P(E_2) \tag{4.7}$$

$$P(E_1E_2H) \leq Minimum\,(P(E_1H), P(E_2H)) \tag{4.8}$$

Within the constraints imposed by these conditions the probabilities are chosen at random. The form of the multiplicative and additive rules are equivalent to our previous forms but allow these random probabilities to be input. The multiplicative update rule that we used is

$$P(H|E_1E_2) = \frac{P(E_1E_2H)}{P(E_1E_2)} \tag{4.9}$$

and the additive update rule that we used

$$P(H|E_1 \cup E_2) = \frac{P(E_1H) + P(E_2H) - P(E_1E_2H)}{P(E_1) + P(E_2) - P(E_1E_2)} \tag{4.10}$$

Figure 4.1 clearly shows the stability of the additive update process versus the multiplicative update process. One can view the simulation as taking a prior probability distribution over the prior probability $P(H|E_1)$ which is uniform over [0,1] and transforming it into the posterior distribution over the posterior probabilities $P(H|E_1E_2)$ which is shown in Figure 4.1 for both of the multiplicative or additive update rules. The mean of the posterior probabilities is .22 for the multiplicative case and .50 for the additive case. This shows that the multiplicative update rule on average computes a posterior probability that is about one-half the value of the posterior probability for the additive update rule. Also, the distribution of the posterior probabilities and the distribution of prior probabilities are closer for the additive rule than for the multiplicative rule. Thus the additive update process is much more conservative than the multiplicative process.

Another simulation shows the result of iterating our update rule for 2, 16, and 128 experiments. Figure 4.2 is a plot of the results from 100,000 runs for our additive update rule and Figure 4.3 is a plot using the same number of runs for the multiplicative rule. The central limit theorem-type convergence with increasing number of experiments, which is expected by the martingale property, is clearly evident in Figure 4.2 while no difference is discernible in Figure 4.3. Also, the same tendency toward low values of the iterated update is found for the multiplicative rule while

the additive updates converge toward what one would expect for random choices of hypothesis and evidence probabilities.

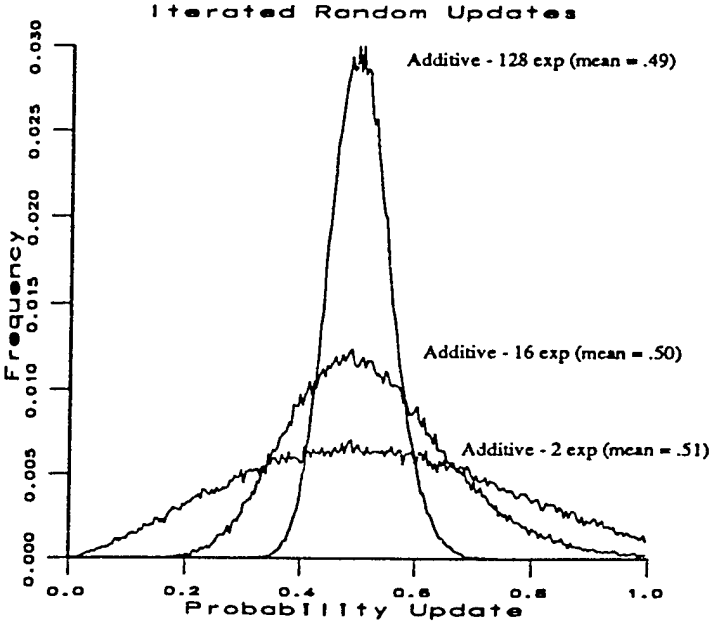


Figure 4.2 Iterated Random Updates - Additive Rule

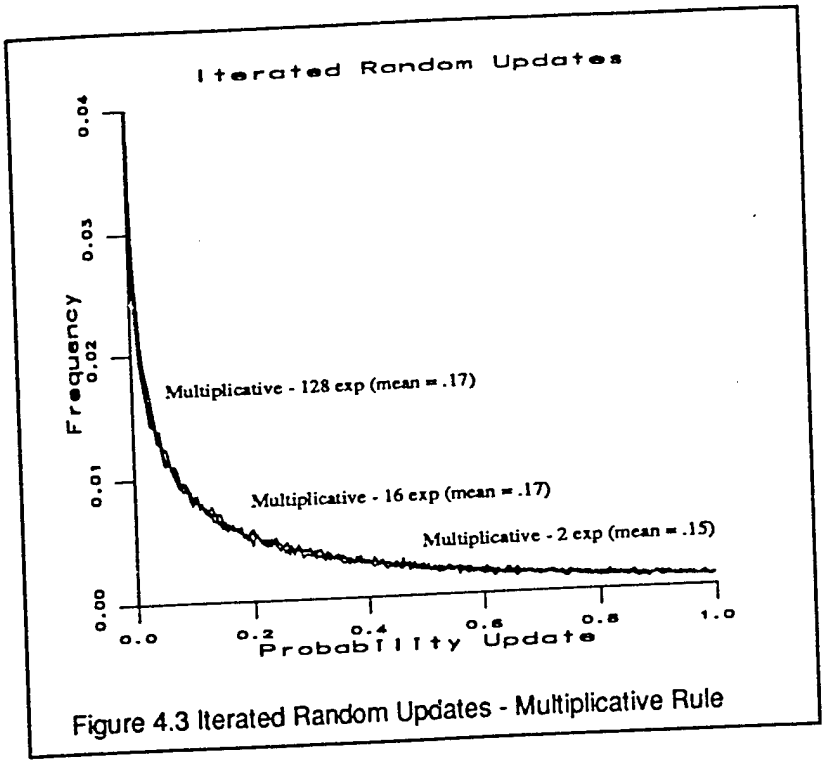


Figure 4.3 Iterated Random Updates - Multiplicative Rule

4.1.3 Likelihood Simulations. We compare the relative effect of new evidence on likelihood ratios,

$$\Lambda^M = \frac{P(E_N|H)\,P(E_O|H)}{P(E_N|\overline{H})\,P(E_O|\overline{H})} \tag{4.11}$$

derived from the Multiplicative Rule and,

$$\Lambda^A = \frac{[P(E_O|H) + P(E_N|H)\ (1 - P(E_O|H))]}{[P(E_O|\overline{H}) + P(E_N|\overline{H})\ (1 - P(E_O|\overline{H}))]} \tag{4.12}$$

derived from the Additive Rule. To simplify the discussion, we have dropped the ratio $P(H)/P(\overline{H})$ that multiplies both L^M and L^A.

To indicate the effect of new evidence we can plot L^A and L^M against $P(E_N|H)$ and $P(E_N|\overline{H})$ (Figures 4.4-4.9). To obtain the figures, we suppose that both rules start with the same prior likelihood,

$$\Lambda_O = \frac{P(E_O|H)}{P(E_O|\overline{H})} \tag{4.13}$$

and we also fix $P(E_O|H)$ since the Additive Rule requires those values explicitly. When $L_O = 1.0$, i.e. when the prior likelihood is indifferent between H and $\overline{H}$, and when $P(E_O|H) = 0.25$, we have Figures 4.4 and 4.5. Other values of $P(E_O|H)$ yield qualitatively similar results. Clearly L^A is restricted to a narrower range than L^M. Furthermore, L^M drops to values less than 1 very quickly.

This is somewhat easier to see if we also fix $P(E_N|\overline{H})$ and plot L^A and L^M against $P(E_N|H)$ (Figures 4.6-4.9). The statistics associated with the figures give maximum and minimum values of L's, the slope of the L curves, and the L^1 distance of the L curves from L_O. The figures indicate that large (small) values of $P(E_N|H)$ can lead to L^M that is much larger (smaller) than L_O, while L^A stays closer to L_O throughout the range of $P(E_N|H)$. This is consistent with our earlier observation that the Additive Rule gives much less weight to outliers than does the Multiplicative Rule.

Comparing Figures 4.6 and 4.7 indicates that the effect of new positive evidence, $P(E_N|H)$, on L^A is reduced if relatively strong evidence ($P(E_O|H) = 0.25$ vs. $P(E_O|H) = 0.50$) has already been accrued. The magnitude of previous evidence has no effect on L^M since it does not depend on $P(E_O|H)$ directly, and thus cannot distinguish cases where prior evidence is negligible from those in which quite a lot of evidence has been accumulated. Figures 4.6 and 4.8 show that high values of new negative evidence ($P(E_N|\overline{H}) = 0.25$ vs. $P(E_N|\overline{H}) = 0.50$) reduce both Λ^A and Λ^M. However, the effect on Λ^M is more pronounced: the slope of the Λ^M curve is reduced by half while the slope of the Λ^A is basically unchanged. Maximum and minimum values of Λ^A and Λ^M show similar effects. Figure 4.9 shows the effect of the prior likelihood, Λ_O. The higher Λ_O, the greater is the effect of new evidence on Λ^M. On the other hand, Λ^A is restricted to a narrower range that is closer to Λ_O.

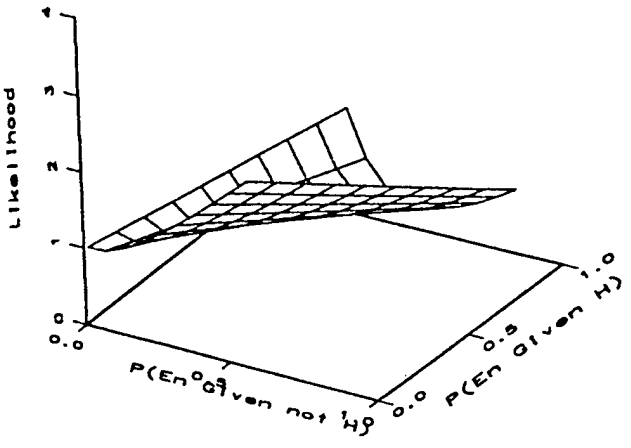


Figure 4.4 -- Updated Additive Likelihood
$\Lambda_O = 1.0$
$P(E_O \mid H) = 0.25$

$\Lambda_{Max} = 3.85$
$\Lambda_{Min} = 0.26$

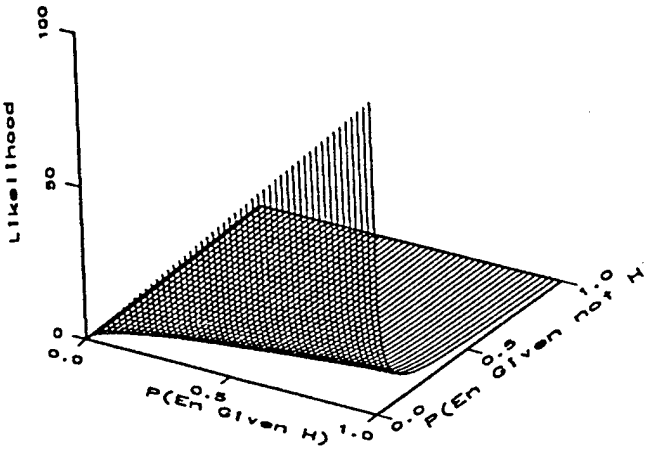
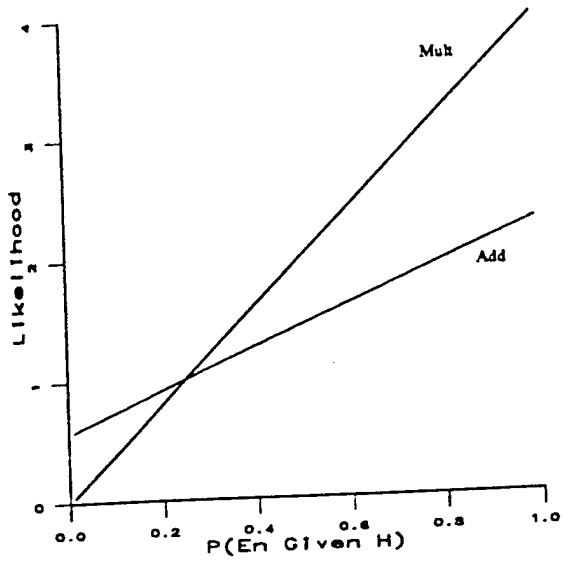


Figure 4.5 -- Updated Multiplicative Likelihood
$\Lambda_O = 1.0$
$P(E_O \mid H) = 0.25$

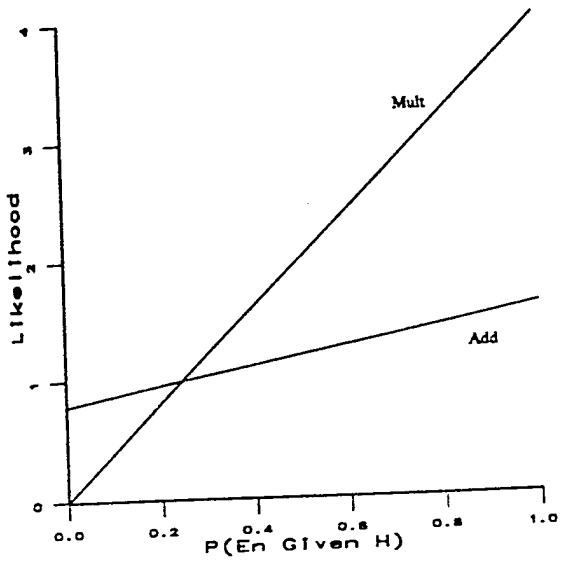
$\Lambda_{Max} = \infty$
$\Lambda_{Min} = 0.0$



	Λ^A	Λ^M
Max	2.27	4.0
Min	0.59	0.0
Slope	1.68	4.0
$\mathcal{L}^1$ distance to L_O	0.64	1.25

$\Lambda_O = 1.0$ $P(E_O|H) = 0.25$ $P(E_N|\overline{H}) = 0.25$

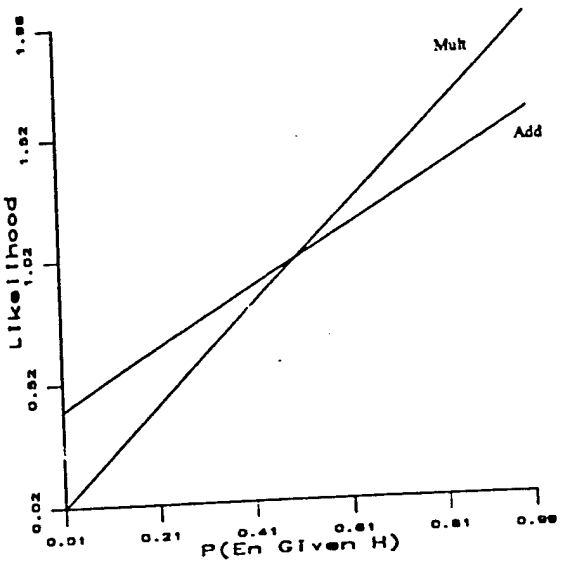
Figure 4.6 -- Additive vs. Multiplicative Rule: Low $P(E_O|H)$



	Λ^A	Λ^M
Max	1.59	4.0
Min	0.81	0.0
Slope	0.79	4.0
$\mathcal{L}^1$ distance to L_O	0.30	1.25

$\Lambda_O = 1.0$ $P(E_O|H) = 0.50$ $P(E_N|\overline{H}) = 0.25$

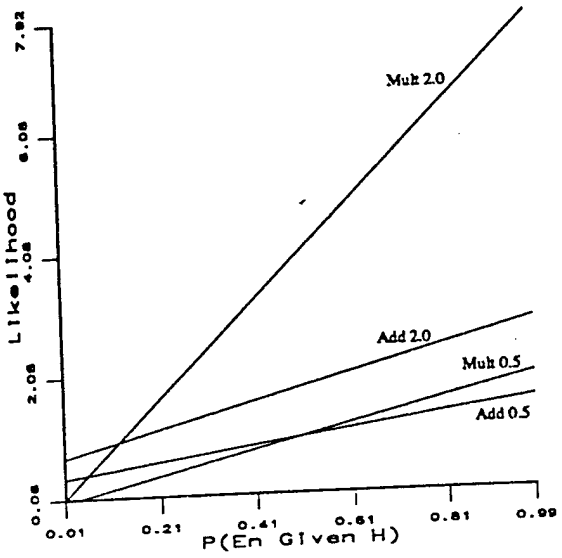
Figure 4.7 -- Additive vs. Multiplicative Rule: High $P(E_O|H)$



	Λ^A	Λ^M
Max	1.59	2.0
Min	0.41	0.0
Slope	1.71	2.0
$\mathcal{L}^1$ distance to L_O	0.34	0.50

$\Lambda_O = 1.0$ $P(E_O|H) = 0.25$ $P(E_N|\overline{H}) = 0.50$

Figure 4.8 -- Additive vs. Multiplicative Rule: High $P(E_N|\overline{H})$



		Λ_O = 0.5	Λ_O = 2.0
Λ^A	max	1.59	0.75
	min	0.42	2.89
Λ^M	max	0.0	0.0
	min	2.0	8.0

$\Lambda_O = 0.5$ and $\Lambda_O = 2.0$
$P(E_O|H) = 0.25$
$P(E_N|\overline{H}) = 0.25$

Figure 4.9 -- Additive vs. Multiplicative Rule: Multiple Λ_O

5 Comparison with Other Additive Rules

We also compare the additive rule to consensus rules and Jeffrey's rule. The additive rule is a kind of consensus rule since it builds up $P(H| \cup_i E_i)$ as a weighted average of evidence sources, but it is derived from simple probability arguments and is not ad hoc. Jeffrey's rule follows from the additive rule when the E_i are a complete partition of the event space. Activation of artificial "neurons" is usually achieved by consensus, and we can write the additive rule so that it looks like a method for activating a neuron. When we do that, we obtain an expression for the weights of an artificial neural system that might be useful in defining learning dynamics.

5.1 Consensus Rules

A branch of applied probability is concerned with combining the opinions of several experts or the subjective probability assessments of several experts. Consensus rules are one general method for combining these opinions or probabilities. These have been explored by a variety of researchers. A modern survey of the necessary properties of general consensus rules and some additional mathematical properties of linear consensus rules is given in Berenstein, Kanal and Lavine (1986). Another good reference is found in DeGroot (1974). An early reference termed the group of opinions an opinion pool. We now describe two types of opinion pools, linear and independent, and mention their relationship to our additive update rule.

The linear opinion pool combines the group of subjective probability distributions in the form

$$P(H| \bigcup_{k=0}^{n} E_k) = \sum_{k=0}^{n} w_i P(H| E_i) \tag{5.1}$$

where the weights w_i are positive and sum to 1. Due to the ad hoc nature of this formula, there is the problem of determining the weights in a probabilistically consistent manner. Even more important is the fact that the formula does not allow the reinforcement of negative evidence as the evidence experiments increase because the sum on the right-hand side of the formula is monotonically increasing.

The independent opinion pool can be written

$$P(H|\bigcup_{k=0}^{n} E_k) = \alpha \prod_{k=0}^{n} P(H|E_k) \tag{5.2}$$

where α is a normalizing constant and the evidence experiments are considered independent. Unless the design and scheduling of experiments is done very carefully the independence assumption may be far from valid. Also, although evidence can negatively reinforce, reinforcement may be unjustifiably extreme (see Berger, 1985).

Both of the above formulas are ad hoc and require care when choosing appropriate and consistent weights. The linear pool formula does not allow for negative reinforcement and the independent pool formula can be unstable with increasing evidence. Our additive update rule is rigorously and consistently derived from basic probabilistic axioms and models. Also, as shown in Section 3.1, our additive update rule allows for negative reinforcement because the weights are not required to sum to 1 for the result to be consistent as a probability. The negative reinforcement is also shown in the numerical results presented in Section 4. Finally, as discussed previously, our additive update rule changes conservatively with respect to accumulating evidence and thus reinforcement is stable, especially for missing or bad evidence outliers.

5.2 Jeffrey's Rule

Jeffrey (1965) presented a rule that is an alternative to the usual multiplicative update rule based on Bayes rule for revising a probability P to a new probability P* based on new probabilities P*(Ei) on a partition $\{E_i\}^n_{i=1}$. Jeffrey's rule is written

$$P^*(H) = \sum_{i=1}^{n} P(H|E_i) P^*(E_i) \tag{5.3}$$

and is judged applicable if $P^*(H|E_i) = P(H|E_i)$ for all H and i. This condition is satisfied for sequences E_i of exchangeable random variables and in fact Jeffrey's rule is derivable from the

basic formula for total probability and exchangeability. An important property of Jeffrey's rule is that it is the natural rule for revising probability if given a prior P, a partition {Ei}, and a new measure P* on {Ei} one wants to find the "closest" measure to P that agrees with P* on the partition and take this as defining P* on the whole space. This is true for any of several common ways of defining closeness between measures on a countable sample space. See Diaconis and Zabell (1982) for a complete discussion.

Now recalling our additive update rule written in partition form

$$P(H|\bigcup_{k=0}^{n} E_k) = \frac{P(H)}{P(\bigcup_{k=0}^{n} E_k)}\left[\sum_{k=0}^{n} P(\overline{E_0E_1\ldots}E_k|H)\right] \tag{5.4}$$

this can be rewritten as

$$P^*(H(\bigcup_{k=0}^{n} E_k)) = \sum_{k=0}^{n} P^*(H|\overline{E_0E_1\ldots}E_k)\,P^*(\overline{E_0E_1\ldots}E_k) \tag{5.5}$$

where we have replaced P by P* to be consistent with the notation in Jeffrey's rule. Now because we are working with exchangeable random variables and the evidence sets induced by them, we can replace the conditional P with P* to get

$$P^*(H(\bigcup_{k=0}^{n} E_k)) = \sum_{k=0}^{n} P(H|\overline{E_0E_1\ldots}E_k)\,P^*(\overline{E_0E_1\ldots}E_k) \tag{5.6}$$

In this form our additive update rule is directly analogous to Jeffrey's rule and can in fact be used to derive Jeffrey's rule in the case that the $\{E_i\}$ constitute a partition of the entire sample space. This implies that our additive update rule can also be derived as the "closest" measure to P that agrees with P* on the partition.

Section 5.3 Comparison to NN

Weighted consensus building is a common approach to activating the "neurons" that populate artificial neural systems. Our additive rule can be written in a form that is similar to the formula used

to represent activity in artificial neurons,

$$P(H|\bigcup_1^n E_i) = \sum_{i=1}^{n} w_i P(E_i) \ , \tag{5.7}$$

where the process of making a decision regarding the value of the hypothesis is similar to the non-linear thresholding found in artificial neural networks.

Here $P(H|\cup_i E_i)$ is equivalent to the activity in a "goal" neuron that receives input from n "input" neurons, each of which has its own activity, $P(E_i)$. The $P(E_i)$ are prior probalities calculated by earlier portions of the net. The input vector $(P(E_1), ..., P(E_n))$ is filtered through weights $(w_1, ..., w_n)$ that are learned in artificial neural systems. From our additive rule we have

$$w_i = \frac{P(H|E_i)\prod_{k=1}^{i-1} P(\bar{E}_k|H)}{P(\bigcup_1^n E_i)} \tag{5.8}$$

so the additive rule corresponds to a net in which evidence sources compete to activate the goal neuron. When the evidence sources are disjoint, we have $w_i = P(H|E_i)$. Equation (5.8) also relates the additive rule to the kind of linear opinion pooling discussed by Berger (1985). The main technical difference is that linear pooling requires $\Sigma_i w_i = 1$; more important is the fact that the w_i in linear pooling are ad hoc.

Section 6 Future Directions

In this section we describe several aspects of the additive update process that need more research but show promising directions for relating it to several other areas of current research in probability, measure theory and dynamical systems.

Section 6.1 Simulation

The Representation Theorem for sequences of exchangeable random variables that was discussed

in Section 2.3 allows us to simulate the long-term behavior of such sequences and thus understand the long-term properties of evidence sequences and the additive update process. Recalling the theorem

$$P(E_i \in A_i)_{i=1}^{n} = \int_{\Omega} \prod_{i} P(A_i)\, d\mu(P) \tag{6.1}$$

where E_i is the sequence of exchangeable random variables, Ω is the space of all probabiliity measures, and $\mu(P)$ is a measure on the space of probability measures. The theorem states that the probability of a set defined by the conditions $X_i \in A_i$ is given by a mixture of power probabilities weighted by a measure on the space of probability measures. One can also view $\mu(P)$ as a prior on the space of proability measures that gets updated to a posterior probabillity P based on power probabilities on the sets defined by the conditions $E_i \in A_i$. Thus, one can simulate and study the statistical properties of the probabilities P on sets of exchangeable random variables by sampling from Ω using $\mu(P)$. Dubins (1967) discusses methods for sampling random distribution functions using a natural measure on the space of probability measures. In the future we will use this approach to study the long-term and aggregate properties of sequences of exchangeable random variables that represent evidence updates.

Section 6.2 Kahane Martingale

Our additive update formula can be viewed as producing a random variable which is an expectation of sums of random variables (i.e. indicator functions) over a σ-algebra generated by a partition formed from the evidence and hypothesis support sets. In Section 3.1 we derived that

$$P(H \mid \bigcup_{k=1}^{n} E_k) = E\left(\frac{\sum_{\pi,k}^{1} (\bigcap E) H}{\sum_{\pi,k}^{1} \bigcap E} \,\middle|\, \Im_n \right) \tag{6.2}$$

These sums of indicator functions can also be rewritten as

$$E\left(\frac{\sum_{\pi,k}^{1} (\bigcap E) H}{\sum_{\pi,k}^{1} \bigcap E} \,\middle|\, \Im_n \right) = E\left(\frac{\prod_{k=1}^{n} (1_{E_k H} + 1_{\bar{E}_k H})}{\prod_{k=1}^{n} (1_{E_k} + 1_{\bar{E}_k})} \,\middle|\, \Im_n \right) \tag{6.3}$$

This allows us to write our additive update formula in a multiplicative form

$$E\left(\frac{\prod_{k=1}^{n}(1_{E_kH}+1_{\bar{E}_kH})}{\prod_{k=1}^{n}(1_{E_k}+1_{\bar{E}_k})}\middle|\Im_n\right)=E\left(\left(\prod_{k=1}^{n-1}\frac{(1_{E_kH}+1_{\bar{E}_kH})}{(1_{E_k}+1_{\bar{E}_k})}\right)\left(\frac{(1_{E_nH}+1_{\bar{E}_nH})}{(1_{E_n}+1_{\bar{E}_n})}\right)\middle|\Im_n\right) \qquad (6.4)$$

The expected value of the n-th term in the product is 1, because before we perform the experiment we consider the evidence sets to be entirely contained within the hypothesis set H. Thus we may consider the iterated product of indicator function ratios as a special type of martingale discussed by Kahane (1987). These martingales are the basic model for a variety of multiplicative random process applications such as random coverings, certain branching processes, and the cascade processes of Mandelbrot used for modeling turbulence. Kahane studies the limit distribution of the random products and describes their support sets as well as the analytic properties of the random products viewed as an operator on prior measures P(H). We will apply these results to our case and build on them to develop for our case a better understanding of the hypothesis testing theory.

Section 6.3 Logistic Map

If one assumes that $E[P(E_o|H)]= E[P(E_n|H)]=2P_o$ then our additive update rule becomes

$$P(H|E_o\cup E_n)=\frac{P(H)}{P(E_o\cup E_n)}[2P_o+2P_o(1-2P_o)] \qquad (6.5)$$

or

$$P_n=4\alpha_nP_o(1-P_o) \qquad (6.6)$$

where $\alpha_n=P(H)/P(E_o\cup E_n)$ and $P_n=P(H|E_o\cup E_n)$. Now we can see that our update rule is related to the well known quadratic iterator map. Depending on the choice of α_n this mapping may exhibit chaotic behavior. We propose to investigate the behavior of this mapping for α_n in the approriate range for our application. We also will investigate the effect of the mapping if actual random variables are input versus the expected value of the random variables.

References

Carlos Berenstein, Laveen N. Kanal and David Lavine, "Consensus rules," in *Uncertainty in artificial intelligence*, ed. by L.N. Kanal and J.F. Lemmer, Elsevier Science Publ., New York, NY, 1986.

James O. Berger, *Statistical decision theory and Bayesian analysis*, Springer-Verlag, New York, NY, 1985.

B. de Finetti, "Foresight: its logical laws, its subjective sources," in *Studies in subjective probability*, ed. by H.E. Kyburg and H.E. Smokler, John Wiley and Sons, New York, NY, 1964.

B. de Finetti, *Probability, induction and statistics*, John Wiley and Sons, New York, NY, 1972.

Morris H. DeGroot, "Reaching a consensus," J. Am. Stat. Assoc., 69(345), 1974.

P. Diaconis and D. Freedman, "Finite exchangeable sequences," AMS Annals of Prob., 8(4), 1980.

P. Diaconis and S.L. Zabell, "Updating subjective probability," J. Am. Stat. Assoc., 77(380), December, 1982.

L.E. Dubins, "Towards characterizing the set of ergodic probabilities," in *Exchangeability in probability and statistics* ed. by G. Koch and P. Spizzichinio, North-Holland Publ. Co, 1982.

L.E. Dubins and D. A. Freedman, "Random distribution functions," Proc. 5th Berkeley Symp. on Math. Statistics and Probability, ed. by L.M. LeCam and J. Neyman, Univ. California Press, Berkeley, CA, 1967.

L.E. Dubins and D. A. Freedman, "Exchangeable sequences need not be mixtures of independent, identically distributed random variables," Z. Wahrscheinlichkeitstheorie, 48, 1979.

P.O. Duda and P.E. Hart, *Pattern classification and scene analysis*, John Wiley and Sons, New York, NY, 1973.

E. Hewitt and L.J. Savage, "Symmetric measures on Cartesian products," Trans. Am. Math. Soc., 80, 1955.

R. Jeffrey, *The logic of decision*, McGraw-Hill, New York, NY, 1965.

J.P. Kahane, "Positive martingales and random measures," Chin. Ann. of Math., 8 Ser. B, Beijing, Peoples Republic of China, 1987.

M.C. Stein and C.L. Winter, "Additive theory of Bayesian updating," in preparation.

C.L. Winter, T.W. Ryan and B.R. Hunt, "Inference and data structures for image identification," *IEEE Proc. of 5th Phoenix Conf. on Computers and Communication*, IEEE Comp. Soc. Press, Washington, DC, 1986.