# On Virtual Evidence and Soft Evidence in Bayesian Networks

*Jeff Bilmes*
{bilmes}@ee.washington.edu
*Dept of EE, University of Washington*
*Seattle WA, 98195-2500*

UWEE Technical Report
Number UWEETR-2004-0016
May 2004

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

# On Virtual Evidence and Soft Evidence in Bayesian Networks

Jeff Bilmes
{bilmes}@ee.washington.edu
Dept of EE, University of Washington
Seattle WA, 98195-2500

*University of Washington, Dept. of EE, UWEETR-2004-0016*
May 2004

**Abstract**

We provide an in-depth description of virtual evidence and its application to Bayesian networks. As we will see, virtual evidence significantly extends the power of Bayesian networks, and makes them much more widely applicable. It does this by giving Bayesian networks the ability to represent aspects from both directed and undirected graphical models. We describe how both evidence and virtual evidence can be seen as having the evidence nodes in the Bayesian network possess virtual children who are always observed to have value unity, and where the conditional probability table of these children imbue their parents either with a single value (hard evidence) or with a collection of values (soft evidenced), the latter case allowing each such value possessing a weight which gets applied to the probability score. This soft evidence is seen as a straightforward generalization of evidence, where the soft evidence phenomena is interpreted as information provided from outside the context of the process modeled by the Bayesian network (or alternatively, as an extension to the event space that is partitioned normally by the collection of random variables in the Bayesian network). We discuss Pearl's example of soft evidence, interpret hybrid ANN/HMM (artificial-neural network/hidden Markov model) systems as soft evidence, and conclude that while the form of soft evidence depends only on ratios of scores and not on the scores' absolute values, the actual score values may be obtained from any information source so desired. We find that some information sources, however, might be more intuitively justifiable than others.

## 1 Introduction

The use of probability to describe a physical process means that one acknowledges the uncertainty about that process. That uncertainty may be inherent to the process itself (i.e., the process is truly random) or it may instead stem from the lack of complete knowledge of that process. Whatever the reason, probability is only one possible representation of uncertainty that is available, but it is one that is backed up by a rigorous mathematical theory, something that may partially explain probability's appeal.

Probability distributions convey uncertainty about a set of events relevant to the process, but as new knowledge comes in regarding that process, how is one to integrate that information so as to form a new probabilistic model? New information is often termed "evidence", and when we have a probability distribution we often discover that, say, $X = 3$. Given such a probability distribution over a set of random variables, how should that probability distribution be updated when information about some of those random variables becomes available? This begets further questions such as, in what form does this information arrive? Might the information be a statement of certitude about certain random variables ($X = 3$), and if so what should that certitude imply about the rest of the random variables in the distribution.

More interestingly, the information about the process itself contain uncertainty, and only express preferences among certain values of some of the random variables. For example, we discover that either $X = 2$ or $X = 3$ but $X \neq 1$ and $X \neq 0$. We may furthermore discover that there are certain numeric preferences over the two possibilities $X = 2$ and $X = 3$.

This has been the topic of belief updating, where we have some initial belief (represented by a probability distribution) and as evidence comes in, that belief is updated in some way [11]. We do not wish herein to require any cultural or cognitive notion of what "belief" might mean, and therefore we avoid using this term when possible, using instead only the more generic term "probability", and we analyze rules for its updating when presented with different forms of evidence.

The predominant form of probability updating that has been in use today has been Bayes' rule. I.e., if we have a distribution $p(x, y)$ this imparts a marginal probability over $x$ as $p(x) = \sum_y p(x, y)$. If we discover that $Y = 2$, then this changes our belief in $x$ to $p(x|Y = 1) = p(x)\frac{p(Y=1|x)}{p(Y=1)}$. In other words, the update factor applied to $p(x)$ is $p(Y = 1|x)/p(Y = 1)$.

What if there is only uncertain information about the event expressed via variable $Y$? I.e., for each value of $Y$, we only have some numeric expression regarding how certain we are for each. How should this uncertainty be represented? If the event is $Y = y$, should the uncertainty be represented as a distribution over $Y = y$ as in $P(Y = x)$, in such case, shouldn't the joint distribution over $p(X, Y)$ thus thereafter say that the resulting marginals always should correspond to that uncertainty? Are there other notions of uncertainty about an event, one that doesn't require so rigid an evaluation as this? These are the questions we attempt to answer.

In Pearl's classic text [11], he defines the notion of virtual evidence and virtual children. Virtual evidence is defined as a generalization of standard evidence in Bayesian networks. This simple construct provides a significant increase in the power and representational capabilities of Bayesian networks, and solves some of the questions above quite naturally. We define and give an interpretation to such evidence generalizations and uncertainty over evidence, and see how this has been applied. We will see that many statistical models can be described in the Bayesian network framework only with the use of evidence generalization.

## 2 Notation

We use a matlab-like notation $1 : N$ to denote the set of integers $\{1, 2, \ldots, N\}$. A set of $N$ random variables (RVs) is denoted as $X_{1:N}$. We sometimes use $U \equiv 1 : N$ to denote the universe of all possible variables. The word "universe" is used deliberately to mean that $X_{1:N} = X_U$ consists of all of the variables in some underlying physical process that is being modeled by a Bayesian network's corresponding probability distribution. We will use $X \equiv X_{1:N} \equiv X_U$. Given any index subset $S \subseteq 1 : N$ (alternatively $S \subseteq U$), where $S = \{s_1, s_2, \ldots, s_{|S|}\}$, the corresponding subset of random variables is denoted $X_S = \{X_{s_1}, X_{s_2}, \ldots, X_{s_{|S|}}$. In general, we use upper case letters such as $A$, $B$, $S$, $U$, refer to index sets.

Upper case letters, such as $W$, $X$, $Y$, and $Z$, refer to random variables and lower case letters ($w$, $x$, $y$, and $z$) their values. We may refer to $p_X(X = x)$ as $p_X(X = x) = p(X = x) = p(x)$ meaning the probability that the (vector) RV $X$ is equal to the (vector) value $x$. Each random variable can take one of a set of values. We herein are primarily going to refer to discrete RVs for simplicity. While we will discuss observed continuous random variables, hidden continuous random variables require more notational machinery (Lebesgue integration) but are otherwise similar.

The set of values that a RV $X_i$ may take on (sometimes referred to as $X_i$'s domain) is denoted $\mathcal{D}_{X_i}$. The size of the set $\mathcal{D}_{X_i}$ (denoted $|\mathcal{D}_{X_i}|$) is referred to the *cardinality* of the set $\mathcal{D}_{X_i}$ and we therefore also impart a "cardinality" to the random variable $X_i$. We may therefore also refer to the set of values that a vector RV $X_S$ may take on as $\mathcal{D}_{X_S} = \mathcal{D}_{X_{s_1}} \times \mathcal{D}_{X_{s_2}} \ldots \mathcal{D}_{X_{s_{|S|}}}$, which is the Cartesian product of the individual sets $\mathcal{D}_{X_i}, i \in S$. Therefore, we have that $x_S \in \mathcal{D}_{X_S}$. Note also that $\mathcal{D}_{X_U} = \mathcal{D}_X$ is the set of all possible values of the entire universe $X_U$ of random variables.

It is always possible to compute the marginal distribution over any set of variables. I.e.,

$$p(x_S) = \sum_{x_{U \setminus S}} p(x_U)$$

By this equation, the sum means that we sum over all values of all variables in the universe other than the variables $X_S$. We use the notation $U \setminus S$, which means all the members of the set $U$ except for those in $S$. $U \setminus S$ can be read "U except for S", "S removed from U", or even simply "U minus S."

The *event* that the set of all RVs has a specific value is denoted as $\{X_{1:N} = x_{1:N}\}$ or alternatively, $\{X_U = x_U\}$. Such an event is one where the entire universe of variables is seen to have a set of values, $x_i$, for $i \in U$. The probability of the event that a subset of random variables $X_S$ is a particular value $x_S$ is $p_{X_S}(X_S = x_S) = p(x_s)$ for any $S \subseteq U$.

A probability measure space is $(\Omega, \mathcal{F}, P)$ where $\Omega$ is an arbitrary set of points $\omega \in \Omega$, $\mathcal{F}$ is a $\sigma$-field of measurable subsets of $\Omega$ (i.e., a set of subsets of $\Omega$ that are measurable), and $P$ is a probability measure [2]. We will use symbols such as $\eta, \nu \in \mathcal{F}$ for measurable events.

A random variable is defined as a function on a sample space. I.e., a random variable maps measurable events to numeric quantities, therefore we can say that $p(X_U = x_u) = P(\{\omega : X_U(\omega) = x_u\})$. When we have defined a universe of random variables $X_U$, we are saying that we partitioned the space $\Omega$ into measurable events of the form $X_U = x_u$, and from the perspective of the underlying measure $P$, it is only via this partition $p(X_U = x_u)$ that we have all knowledge about the underlying measure $P$, although there can of course exist multiple measures that lead to the same probability distribution, since the partition $X_U = x_u$ need not be the most fine grained.

## 2.1 Bayesian Networks

A Bayesian network (BN) [11, 13, 7, 8] is one type of graphical model [10] where a probability distribution over a set of variable $X_{1:N}$ factorizes with respect to a directed acyclic graph (DAG) in the following way:

$$p(x_{1:N}) = \prod_i p(x_i|x_{\pi_i})$$

where $\pi_i$ are the set of parents, according to a given BN, of variable $x_i$. This is called the directed factorization property (called property **(DF)** in [10]). There are many other (provably equivalent) characterizations of BNs (e.g., d-separation [11, 10]), but this one (DF) suffices for this paper. By a BN, we refer to all distributions that validly factorize with respect to the graph of the BN. A distribution represented by a BN might have more factorization properties than is implied by the BN, but it must not have fewer factorization properties. I.e., if the BN says that $X_1$ is independent of $X_2$ given $X_3$, then this is true all distributions that factorize w.r.t. the BN. The factors $p(x_i|x_{\pi_i})$ in a BN are often called *conditional probability functions* (or CPFs), or conditional probability tables (CPTs).

A random variable $X_i$ is called a *constant random variable* if it is the case that $p(X_i = a) = 1$ for some $x \in \mathcal{D}_{X_i}$. If it is the case that $p(X_i = x_1) > 0$ and $p(X_i = x_2) > 0$ for some $x_1 \neq x_2$, then the variable $X_i$ is not constant and is a true "random" variable. If it is the case that $p(x_i|x_{\pi_i}) > 0$ for all $x_i \in \mathcal{D}_{X_i}$ and $x_{\pi_i} \in \mathcal{D}_{X_{\pi_i}}$ then we say that the conditional probability table $p(x_i|x_{\pi_i})$ is *dense* — we use the word dense since if the table was stored exactly, then all entries would be non-zero. If $p(x_i|x_{\pi_i}) = 0$ for some (but not all) $x_i \in \mathcal{D}_{X_i}$ and $x_{\pi_i} \in \mathcal{D}_{X_{\pi_i}}$ then we say that the conditional probability table $p(x_i|x_{\pi_i})$ is *sparse* — we use the word sparse since the CPT table could be represented as a sparse matrix. If it is the case that $p(x_i|x_{\pi_i}) = \delta(x_i, f(x_{\pi_i}))$ for a particular deterministic function $f : x_{\pi_i} \to x_i$, then we say that the conditional probability table is *deterministic*.

## 2.2 Traditional Evidence and Zero Probability Events

It is often the case that some subset of the variables with index set $E \subseteq 1 : N$ are "evidence nodes" [11, 7] or equivalently are said to be "observed" (or be a *finding*), meaning that we actually know the values of those random variables. This means that by some process, evidence has arrived, and as part of that evidence we are certain about the values of those random variables.

The evidence set can be denoted as $X_E = \bar{x}_E$ or simply just $\bar{x}_E$. All other variables in the network we presumably do not know, and are referred to as *hidden* or *unobserved* variables.

Once we have received evidence, the question becomes what does it mean? I.e., how should we interpret the fact that we have discovered that $X_E = \bar{x}_E$ and how should that influence the probabilities of the remaining variables.

### 2.2.1 Evidence as a sample from a statistical process

A typical way to interpret evidence is that it is a partial sample of a draw from the underlying probability distribution. I.e., a complete sample has occurred, but it is not fully specified. For example, we might have a data set $\{x_E^{(i)}\}_{i=1}^N$ of size $N$, where each $x_E^{(i)}$ consists of values only of the variables in $E \subseteq U$. For each $i$, some of the variables' values are missing or have not been given. This means that for each $i$, $x^{(i)} \sim p(X)$, so a complete sample has occurred, but some of the variables of that sample $X_U \setminus E$ are unavailable. The subset of variable values that we have been given for each sample is the evidence under that sample. The other random variables, the ones which are not revealed or $X_{U \setminus E}$, are hidden variables, relative to the current sample. A different sample might reveal a different subset, so the evidence and hidden random variables could be different from sample to sample. In practice, however, it is common in such a data set for one fixed set of random variables to consistently be revealed for all samples (exceptions to this, see the semi-supervised learning and co-training literature, google search Jerry Zhu's semi-supervised learning overview).

Note, under this interpretation evidence, it does not mean that the event has probability one. I.e., if we happen to receive evidence that $X_E = \bar{x}_E$, that does **not** mean that $P(X_E = \bar{x}_E) = 1$. Rather, in this interpretation, we have that

$$p(X_E = \bar{x}_E) = \sum_{x_{U \setminus E}} p(x_{U \setminus E}, \bar{x}_E)$$

meaning that we need to marginalize away all other non-evidence random variables to discover the probability of the event.

It might even be the case that we find out that $X_E = \bar{x}_E$, but also that $P(X_E = \bar{x}_E) = 0$, meaning that the event $X_E = \bar{x}_E$ is the impossible event (i.e., an event that receives zero probability). While a zero probability event might seem like something that would never occur (and therefore that you would never encounter or need to worry about), in practice, zero probability events might occur in a number of real-world circumstances. The key reason is that the evidence might not come from a sample of the Bayesian network's distribution itself, rather it comes from some other external information source relative to the BN distribution. This might happen in a number of ways.

First, we must realize that a BN may only be an approximate representation of some underlying true process. By modeling that process stochastically, we are essentially saying that either there is some inherent uncertainty about or within the underlying physical phenomena, or alternatively that there is some peculiarity within the process (e.g., noise) that we wish not, need not, or can not model in a detailed way. We resort to treating that peculiarity as randomness. If there was a true random process in the physical world that we wished to model, such a BN representation might or might not correspond to that truth. A zero probability event, therefore, might occur if there is an inaccuracy in the model specification — i.e., a given BN model might lead to the case where we happen to observe something that the model, as it is specified, says will not happen with non-zero probability.

There are a number of reasons why such an inaccuracy might arise: 1) there might be insufficient information about the true model in training data, or the training set size might be small, and regularization alone might not be enough to remove zero probability events; 2) even when there is plenty of training data, the training data and test data distributions might not be the same (often training and testing environments are different); 3) many parameter learning procedures are iterative, and when in the process of learning the parameters, we will not have yet produced an accurate model – a zero probability event would mean that the parameters have not yet been properly learned; 4) there might be a bug in the specification of the model (i.e., there might be bug when using a toolkit to specify a model); and 5) the BN might utilize concise sparse representations of CPTs where low probability events are floored to zero probability — this can reduce the memory needed to store a model, but might lead to zero probability events. In any case, when this happens the model is said to "score" the evidence with zero probability.

A second reason a zero-probability event might occur follows: during the computation of probabilistic quantities (discussed in more detail below), we might temporarily encounter assignments to random variables (that are set, say, during a summation or search operation) which end up having zero probability (see Section 3.1). These have been refereed to "no-goods" in the constraint-satisfaction and SAT literatures [5]. For example, given a probability distribution $p(X_1, X_2)$ over two variables $X_1$ and $X_2$, we might wish to perform the sum:

$$p(x_1) = \sum_{x_2} p(x_1, x_2)$$

It might be the case that for some particular value pair $\bar{x_1} \in \mathcal{D}_{X_1}$, $\bar{x_2} \in \mathcal{D}_{X_2}$, we have that $p(\bar{x_1}, \bar{x_2}) = 0$. Mathematically, these zero probability events are summed together and, of course, have no effect on the result. But we have encountered zero probability events along the way. If there are many of these zero probability events, summing them naively as implied above is quite wasteful.

Therefore we must be ready to observe events that the BN says are impossible. It might be said that the evidence is obtained separately from any information or knowledge that is contained in the current BN itself. That is, evidence might be seen as coming from an information source completely outside of the current network itself — for example, during parameter learning, the training data does not match the distribution expressed by current parameter assignments, and might never match unless both there is sufficient training data and within the model family spanned by the parameter space lies the true model (e.g., the conditions under which maximum likelihood estimation can yield truth). This idea will be useful when we further discuss evidence uncertainty below.

### 2.2.2 Evidence as probability revision

A second form of evidence incorporation says that the probability of the event itself needs to be revised. In this form, we might find out $X_E = \bar{x}_E$ but here we must update the probability model so that $p(X_E = \bar{x}_E) = 1$. I.e., the distribution $p(X_U)$ needs to be revised to $p'(X_U)$ so that under $p'$ it agrees with that prescribed event probability. One way to do this is as follows:

$$p'(x_U) = p(x_{U \setminus E}|x_E)\delta(x_E, \bar{x}_E) = p(x_U)\frac{\delta(x_E, \bar{x}_E)}{p(x_E)}$$

where $\delta(x_E, \bar{x}_E)$ is one only if $x_E = \bar{x}_E$ and is otherwise zero (see Section 3).

Evidence as a probability revision arises from somewhere outside of the current representation. That is, the finding $X_E = \bar{x}_E$ does not necessarily have anything to do with the current distribution (or its underlying measure). Like in the previous section, a BN and its probability distribution is only a approximate representation of reality, and it assigns probabilities to all possible (zero or non-zero probability) events that can occur in this reality $x_U$. Unlike the previous section, here the evidence can't really be thought of a draw from any probability distribution. Rather, evidence is a specific statement about how the probabilities of a particular event needs to be revised, i.e., so that $X_E$ becomes a constant random variable.

The difference between this form of evidence, and that described in Section 2.2.1 is that here, evidence directly specifies information about the probabilities of some subset of random variables. I.e., the evidence is a direct statement made about the model itself (that we should revise our model so that $p(\bar{x}_E) = 1$ and $p(X_E \neq \bar{x}_E) = 0$). Once we discover this piece of evidence, our probability model should be so adjusted so as to make $\bar{x}_E$ occur with certainty.

In Section 2.2.1, however, the evidence has made a statement not about the probabilities but only about the outcomes of a set of random variables. We discover that $X_E = \bar{x}_E$. If it is the case that we receive a data set in which for each sample we have such evidence, it is possible to combine this together to adjust the model (as in any parameter adjustment method such as maximum-likelihood, or some other optimization procedure used as a form of machine learning). But the evidence itself does not dictate how the model probabilities should be updated.

## 2.3 Probability of evidence

Next, we introduce three frequently needed calculations including: 1) computing the probability of the evidence (this section); 2) computing (posterior) probabilities of the hidden variables given any evidence; and 3) finding the most likely assignment of (all or a subset of) the hidden variables in a conditional distribution (conditioning on the evidence). In all three cases, we assume evidence of the form described in Section 2.2.1, returning to Section 2.2.2 a bit later.

To compute the probability of the evidence, we simply sum out over all hidden variables and compute:

$$p(\bar{x}_E) = \sum_{x_{U \setminus E}} p(x_{U \setminus E}, \bar{x}_E)$$

If we set $H = U \setminus E$, then this notation really means:

$$p(\bar{x}_E) = \sum_{x_{H_1} \in \mathcal{D}_{X_{H_1}}} \sum_{x_{H_2} \in \mathcal{D}_{X_{H_2}}} \cdots \sum_{x_{H_{|H|}} \in \mathcal{D}_{X_{H_{|H|}}}} p(x_{U \setminus E}, \bar{x}_E)$$

Clearly, this computation if done naively would require a cost of $O(|\mathcal{D}_{X_{U \setminus E}}|)$ operations (exponential in the size of $U \setminus E$), so the goal would be to use the fact that $p(x_{U \setminus E}, \bar{x}_E)$ factorizes with respect to a BN, and then distribute (as auspiciously as possible) the sums into the products to minimize computation. This is precisely what the *elimination algorithm* [7] does (sometimes called variable elimination, bucket elimination, and so on). Finding the best way to do this is an NP-hard problem. It is also the case that a portion of the junction tree, that involves moralizing and triangulating the graph, essentially does the equivalent of performing this summation in (hopefully) as computationally cheap a way as possible. In this discussion, however, we do not need get into into the details of probabilistic inference to understand virtual evidence. For now, just assume that probabilistic inference is, as mentioned above, an appropriate distribution of the sums into products of factors to minimize computation.

## 2.4 Posterior Probabilities

Second, we might want to calculate the posterior probability of some subset of variables given some other (evidence) subset. Suppose $E \subseteq U$ is a set of variables that we have evidence for. We might very well be interested in the conditional probability $p(x_S|\bar{x}_E)$ where $S \cap E = \emptyset$. As mentioned above one way of viewing this problem is that of probability (or belief) revision. I.e., the BN has an initial set of beliefs over the set of possible values $\mathcal{D}_{X_S}$ of the set random variables $X_S$, and the belief in $x_S \in \mathcal{D}_{X_S}$ is calculated as $p(x_S) = \sum_{x_{U \setminus S}} p(x_S, x_{U \setminus S})$. Once evidence is introduced, we wish to revise the beliefs in each $x_S \in \mathcal{D}_{X_S}$. A standard way of doing this would be to use Bayes rule, given the posterior belief $p(x_S|\bar{x}_E)$ as above.

If it is the case that $S \cup E = U$, then this means that we must compute the $|\mathcal{D}_{X_S}|$ sized conditional probability table (CPT)

$$p(x_S|\bar{x}_E) = \frac{p(x_S, \bar{x}_E)}{p(\bar{x}_E)} = \frac{p(x_S, \bar{x}_E)}{\sum_{x_S} p(x_S, \bar{x}_E)}$$

If on the other hand it is the case that $S \cup E \subset U$, then we have a partition of $U$ into $S, E$, and $H$, where $H = U \setminus \{S \cup E\}$, and we must calculate

$$p(x_S|\bar{x}_E) = \frac{\sum_{x_H} p(x_S, x_H, \bar{x}_E)}{p(\bar{x}_E)} = \frac{\sum_{x_H} p(x_S, x_H, \bar{x}_E)}{\sum_{x_S} \sum_{x_H} p(x_S, x_H, \bar{x}_E)}.$$

Again, note that $x_U \equiv \{x_S, x_H, x_E\}$. We are often interested in quantities such as

$$p(x_i|\bar{x}_E) = \frac{\sum_{x_{U \setminus (\{i\} \cup E)}} p(x_{U \setminus E}, \bar{x}_E)}{p(\bar{x}_E)}$$

for all $i \in \{U \setminus E\}$. This is the same as above where $S = \{i\}$ for all $i \in U \setminus E$ (so $S$ contains one element and takes turn being all variables in the graph). More generally still, we may partition $S$ into $S = (S_1, S_2, \ldots, S_k)$, and then compute $p(x_{S_i}|\bar{x}_E)$ for $i \in \{1, \ldots, k\}$. This would be needed, for example, in order to perform EM or gradient training of the parameters in the network.

Note that general probabilistic inference (i.e., Pearl's belief propagation [11], the sum-product algorithm [1], the junction tree algorithm, Hugin propagation [7], the Shenoy-Shafer algorithm, the generalized distributed law under sum-product, SPI, and so on) are all similar algorithms to compute this quantity exactly. Note that these algorithms (all exact inference methods) all require some form of graph moralization, triangulation, formulation of a junction tree, and sending messages, even if the algorithms do not do this explicitly. It is not necessary to understand these ideas in the context of this discussion on soft evidence. What is important is to realize that these algorithms all correspond to clever ways of distributing the sums into the product of the joint distribution $p(x_U) = \prod p(x_i|x_{\pi_i})$ factored appropriately according to the BN DAG. Some of the algorithms use dynamic programing to re-use some of these factorizations and partial computations. I.e., they reduce computation since for different $i$ values, it is possible to re-use the fact that many of the sums have already been performed (i.e., distributed into products of factors). The NP-complete problem here is to find the optimal way to distribute the sums inside the set of factors.

## 2.5 Most likely (Viterbi) values

Third, it is often necessary to find:

$$x_S^* = \underset{x_S}{\text{argmax}}\, p(x_S|\bar{x}_E) = \underset{x_S}{\text{argmax}}\, p(x_S, \bar{x}_E)$$

where it is the case that $S \cup E = U$. Again, we factor $p(x_S, \bar{x}_E)$ according to the BN, and then distribute the separate "max" operations into the factors as far to to the right as possible (and optimally), in order to get an answer. The optimal way to do this is identical to the optimal way to distribute the sums in the case when we wish to compute $p(\bar{x}_E)$. The algorithm is sometimes called the max-product algorithm (and corresponds to the Viterbi algorithm in HMMs). Here, however, we are distributing "max" operations rather than sums. Again, this is an NP-hard problem, and again the details of how the best way to do this are not necessary to understand virtual evidence.

# 3 Evidence as Smart Sums

In the preceding section, we gave the evidence a different name. Specifically, evidence was set as $X_E = \bar{x}_E$ for some subset $E \subseteq U$.

When we consider the three operations in the preceding section, (where we are using sums and or max operations), we can treat the evidence simply as the application of a delta function within the full summation. For example, we have that:

$$p(\bar{x}_i) = \sum_{x_i} p(x_i)\delta(x_i, \bar{x}_i)$$

where

$$\delta(x_i, \bar{x}_i) = \begin{cases} 1 & x_i = \bar{x}_i \\ 0 & \text{else} \end{cases}$$

is the Dirac delta function. Similarly

$$p(\bar{x}_i, \bar{x}_j) = \sum_{x_i, x_j} p(x_i, x_j)\delta(x_i, \bar{x}_i)\delta(x_j, \bar{x}_j)$$

or generally:

$$p(\bar{x}_S) = \sum_{x_S} p(x_S)\delta(x_S, \bar{x}_S)$$

where

$$\delta(x_S, \bar{x}_S) \triangleq \prod_{k \in S} \delta(x_k, \bar{x}_k)$$

In some sense, we can bring the delta functions into the sums, so that the sums themselves know about the evidence. If we denote such a "smart sum" for variable $x_i$ with evidence $\bar{x}_i$ as

$$\overset{\bar{x}_i}{\sum_{x_i}} f(x_i) \triangleq \sum_{x_i} f(x_i)\delta(x_i, \bar{x}_i)$$

or more generally, for evidence $\bar{x}_E$, we have

$$\overset{\bar{x}_E}{\sum_{x_U}} f(x_U) \triangleq \sum_{x_U \in \mathcal{D}_{X_U}} f(x_U)\delta(x_E, \bar{x}_E) = \sum_{x_U \in \mathcal{D}_{X_U}} f(x_{U \setminus E}, x_E)\delta(x_E, \bar{x}_E) = \sum_{x_{U \setminus E} \in \mathcal{D}_{X_{U \setminus E}}} f(x_{U \setminus E}, \bar{x}_E)$$

In the left sum, we are essentially summing over all possible values of all variables $\mathcal{D}_{X_U}$ and the ones that do not meet the condition that the subset $x_E = \bar{x}_E$ are annihilated by the delta. On the right most sum, we are only summing over the residual, i.e., the variables that remain free to vary and are not guaranteed to produce a zero value if they don't agree with the evidence.

Of course, one would never implement evidence in this fashion (since it would unnecessarily accumulate multiple zeros), but notationally it is very convenient as all evidence can be treated like hidden nodes with the use of smart sums.

When considering the three problems we mentioned above, it turns out that receiving evidence for a set of random variables is equivalent to using a delta function (or smart summation) with respect to the evidence values. In other words, evidence, in the context of the three problems, can be seen as in some sense an external application of a delta function that annihilates the probability score of all random variable values that to not correspond to the received evidence. Lets now look at each of the three problems in this light.

## 3.1 Computing the probability of evidence

The probability of evidence can now be written in a number of ways, such as:

$$p(\bar{x}_E) = \sum_{x_U} p(x_U)\delta(x_E, \bar{x}_E) = \sum_{x_U} p(x_U) \prod_{i \in E} \delta(x_i, \bar{x}_i) = \overset{\bar{x}_E}{\sum_{x_U}} p(x_U)$$

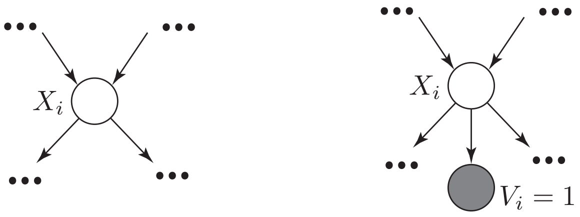

Figure 1: Simple Virtual Evidence. On the left, vertex $X_i$ is connected to the rest of the BN and is in this case hidden. On the right, vertex $X_i$ has one additional single child $V_i$, where it is the case that $V_i = 1$ is always true. The child is not connected to any other vertex in the network. The CPT $p(V_i = 1|X_i = x) = \delta(x, \bar{x})$, where $\bar{x}$ is the desired observed value. This construct is therefore a Bayesian network way of expressing evidence in exactly the same way as was done using delta functions in Section 3.

We see here that any values of variables that do not agree with the evidence are annihilated by the delta function. But for the purposes of computing $p(\bar{x}_E)$ it doesn't matter mathematically if we either: 1) sum over all variable values $\mathcal{D}_{X_U}$ annihilating the ones that do not match the evidence, or 2) sum over only the non-evidence variables keeping the evidence variables fixed in the formula. Of course computationally, it would be better to not sum together many zero values, but again that is beyond the scope of this discussion.

### 3.2 Computing the posterior probability of a set of vars

Computing the posterior $p(x_S|\bar{x}_E)$ where $S \subseteq (U \setminus E)$ and $H = U \setminus (E \cup S)$ can also be easily written in this form:

$$p(x_S|\bar{x}_E) = \frac{p(x_S, \bar{x}_E)}{p(\bar{x}_E)} = \frac{\sum_{x_H} p(x_S, x_H, \bar{x}_E)}{\sum_{x_H, x_S} p(x_H, x_S, \bar{x}_E)} = \frac{\sum_{x_H, x_E} p(x_S, x_H, x_E)\delta(x_E, \bar{x}_E)}{\sum_{x_H, x_E, x_S} p(x_H, x_S, x_E)\delta(x_E, \bar{x}_E)}$$
$$= \frac{{}^{\bar{x}_E}\sum_{x_H} p(x_S, x_H, x_E)}{{}^{\bar{x}_E}\sum_{x_H, x_S} p(x_H, x_S, x_E)} = \frac{{}^{\bar{x}_E}\sum_{x_H} p(x_U)}{{}^{\bar{x}_E}\sum_{x_H, x_S} p(x_U)}$$

### 3.3 Computing the maximum

Again, the same thing occurs, but we essentially substitute a max for a sum[1] (and we could define a smart max if we wished)

$$x_S^* = \operatorname*{argmax}_{x_S} p(x_S, \bar{x}_E) = \operatorname*{argmax}_{x_S} \max_{x_E} p(x_S, x_E)\delta(x_E, \bar{x}_E)$$

# 4 Evidence as virtual children in the graph

Now that we see how it is possible to represent evidence using delta functions and/or smart sums, in this section we'll see another equivalent way in which evidence may be treated. In this way, the evidence nodes, rather than being "observed", are still considered hidden. Each of "hidden evidence" nodes, however have an additional single *virtual observed child* that constrains the hidden evidence node to take on a value with non-zero probability only equal to the evidence for that node. The virtual observed child is indeed observed to have a particular value, but the actual value itself doesn't matter. This is because the virtual child's CPT will do essentially what the delta function of the previous section did, meaning the CPT will be deterministic.

In particular, consider computing $p(\bar{x}_i)$. In the last section, we said we could do this as:

$$p(\bar{x}_i) = \sum_{x_i} p(x_i)\delta(x_i, \bar{x}_i)$$

Now suppose that the random variable $X_i$ had an extra single child $V_i \notin X_U$, and that $V_i$ is connected to no other node in the BN other than its parent $X_i$. Suppose also that it was always the case that $V_i = 1$. This does not mean that

$p(V_i = 1) = 1$ (i.e., we are not saying that $V_i$ is a constant RV). In other words, it does not mean that the BN that we are using would be such that it would designate the event $\{V_i = 1\}$ as certain, where the event probability is:

$$p(V_i = 1) = \sum_{x_U} p(x_U, V_i = 1)$$

Rather the virtual child evidence $V_i = 1$ means that we just happen to observed states of reality in the expanded set of variables (i.e., events over $X_U$ and any virtual children) that only correspond to the cases where the $V_i$ variable has value 1. We only observed events where $V_i = 1$.

The question is, how should we see this model in the context of the generative model interpretation that is easy with Bayesian networks.

One way is to think of this as essentially throwing away, ignoring, or filtering out any events for which $V_i \neq 1$. Ignoring them does not mean they do not happen, but they essentially happen externally to any processing we do, or any consideration of the BN that we use. We consider this as a filtering procedure, where where we sample from $p(x_U, V_i)$ which includes both the cases where $V_i = 0$ and $V_i = 1$ (assuming $\mathcal{D}_{V_i} = \{0, 1\}$is binary), and then a filtering process removes all samples where $V_i = 0$ and then passes the surviving samples on to later processing. The only events considered are those where $V_i = 1$ (so any sum, we are essentially always multiplying by a delta function of the form $\delta(v_i, 1)$.

We stated above that $V_i \notin X_U$. This means that the variable $V_i$ is not contained in the universe of variables comprising our BN, and does not represent a variable that we assume exists within the physical process that we are using the BN to represent. Rather, $V_i$ is a "virtual child", and it is in the BN only to the extent that it allows us to mathematically model the notion of evidence, and the fact that a finding comes from a source outside the information contained within a BN. Specifically, a virtual child is there only for the purposes of constraining its sole parent to be, with non-zero probability, only the given evidence value.

### 4.0.1 Evidence in the probability sample space

We can also view this relative to the underlying sample space $(\Omega, \mathcal{F}, P)$. The original set of random variables partition $\Omega$ into measurable subsets

$$\Omega_{x_A} = \{\omega : X_A(\omega) = x_A\}$$

for each $x_A \in \mathcal{D}_{X_A}$ and for $\emptyset \subset A \subseteq U$, where $\Omega_{x_U} \in \mathcal{F}$. We have that:

$$\mathcal{F}_X \triangleq \bigcup_{\emptyset \subset A \subseteq U} \bigcup_{\Omega_{x_A} : x_A \in \mathcal{D}_{X_A}} \{\Omega_{x_A}\} \subseteq \mathcal{F}$$

meaning that there may be more events in $\mathcal{F}$ that the random variable partitioning does not overlap with. Suppose there was an event $\eta \in \mathcal{F} \setminus \mathcal{F}_X$ that had a conditional relationship with some partitioning $\{\mathcal{E}_{x_E}\}$ of $\Omega$, i.e.,

$$\bigcup_{x_E \in \mathcal{D}_{X_E}} \mathcal{E}_{x_E} = \Omega$$

and where a given $\mathcal{E}_{x_E}$ corresponds to the event $X_E = x_E$. The conditioning relationship would be specified as:

$$p(\eta | \mathcal{E}_{x_E})$$

and if we identify this extra virtual variable $V = 1$ with the event $\eta$ occurring, we have the conditional probability distribution:

$$p(\eta | \mathcal{E}_{x_E}) = p(V = 1 | X_E = x_E)$$

Therefore, these extra virtual children $V$ that we use in the BN merely correspond to probabilistic events in the underlying event space that may not have been designated by the $\Omega$ partition induced by the set of random variables $X_U$. This is what we mean when we add an additional variable $V$ that lies not in the universe $X_U$

On the other hand, it would normally be the case that the event $V_i = 1$ would not preclude (with zero probability) other events in the network as would be needed for a regular evidence event of the form $X_E = \bar{x}_E$ (but see Section 5). We need this extra evidence to ensure that only the event $X_E = \bar{x}_E$ may occur, and this can be done via the CPT $p(v_i | x_i)$. Specifically, we set

$$p(\eta | \mathcal{E}_{x_E}) = p(V = 1 | X_E = x_E) = \delta(x_E, \bar{x}_E).$$

In this case, the event $\eta$ corresponds precisely to $\{V = 1\}$, an event already in $\mathcal{F}_X$. Specifically, we have that $\eta \equiv \{V = 1\} \equiv \{\omega \in \Omega : X_E(w) = \bar{x}_E\} \equiv \{X_E = \bar{x}_E\}$ so that $\eta \in \mathcal{F}_X$.

When each variable $X_i : i \in E$ has its own virtual child $V_i$, we force $V_i$'s parent $X_i$ to be a particular value again using a delta function:

$$p(V_i = 1|X_i = x_i) = \delta(x_i, \bar{x}_i).$$

Once $V_i = 1$ is observed, the only event that explains or allows this is the event $X_i = \bar{x}_i$. The CPT needs only a partial specification, however, because $p(v_i|x_i)$ is never fully used — values in the CPT corresponding only to the case $V_i = 1$ are needed (since we always observe $V_i = 1$). In other words, we need $p(v_i|x_i)$ to be deterministic when $v_i = 1$, but it can be anything for $v_i \neq 1$. For example, we have have $p(V_i = 1|X_i = x_i) = \delta(x_i, \bar{x}_i)$ but $p(V_i = j|X_i = x_i)$ for $j \neq 1$ can otherwise be arbitrary.

With the above CPT, the computation of $p(\bar{x}_i)$, can be performed as follows:

$$p(\bar{x}_i) = p(V_i = 1, \bar{x}_i) = \sum_{x_i, v_i} p(v_i|x_i)p(x_i)\delta(v_i, 1) = \sum_{x_i} p(V_i = 1|x_i)p(x_i) = \sum_{x_i} \delta(x_i, \bar{x}_i)p(x_i)$$

as before.

We see that the full distribution $p(V_i = v_i, X_i = x_i)$ is not needed since many of the values are never used. As we will see below, all that is really needed are the ratios of likelihoods.

## 4.1 Evidence scores other than unity

There is nothing special about the value of unity in the previous section. In fact, we could just as easily have defined the CPT for $V_i$ to be the following:

$$p(V_i = 1|X_i = x_i) = \alpha\delta(x_i, \bar{x}_i).$$

where $\alpha > 0$. In other words, rather than multiplying the probability of assignments of $X_U$ by zero and one, we multiply by zero and $\alpha$. Even when $\alpha \neq 1$, this has only a very trivial (or even no effect) on the three quantities we are interested in (e.g., it does not indicate strength of evidence). Note that this will be true even if $\alpha > 1$ so that it no longer can be interpreted as a probability (see Section 5).

In the following, we will assume that each evidence node $X_i,\ i \in E$ is hidden but has a corresponding virtual child $V_i,\ i \in E$ where each virtual child is observed to always have value $V_i = 1$, and where each $V_i$ CPT implementation has partial specification:

$$p(V_i = 1|X_i = x_i) = \alpha_i\delta(x_i, \bar{x}_i) \text{ for each } i \in E$$

for some fixed set of values $\bar{x}_E$. Also, define:

$$\alpha = \prod_{i \in E} \alpha_i$$

We may now extend the filtering idea from above to this case. If we encounter a set of assignments of the random variables $X_U = x_u$ we multiply the probability of that assignment by $\alpha$ if $x_E = \bar{x}_E$, and multiply the probability of the assignment by zero otherwise.

### 4.1.1 Probability of Evidence

To compute the probability of evidence, we perform:

$$p(\bar{x}_E) = \sum_{x_U} p(x_U) \prod_{i \in E} \alpha_i\delta(x_i, \bar{x}_i) = \sum_{x_U} p(x_U)\alpha\delta(x_E, \bar{x}_E)$$

So we see that the probability of evidence here is just $\alpha$ times the evidence probability computed in Section 3.1 and Section 2.3. It is important to realize that for all possible values of $\bar{x}_E \in \mathcal{D}_{X_E}$, the constant $\alpha$ is the same (i.e., $\alpha$ doesn't change with the values of the vector random variable $X_E$).

### 4.1.2 Posterior Probability

The posterior probabilities, it turns out, are identical regardless of the value of $\alpha$, as long as $\alpha \neq 0$. Using the same notation as in Section 3.2, we have

$$p(x_S|\bar{x}_E) = \frac{p(x_S, \bar{x}_E)}{p(\bar{x}_E)} = \frac{\sum_{x_H,x_E} p(x_S, x_H, x_E)\alpha\delta(x_E, \bar{x}_E)}{\sum_{x_H,x_E,x_S} p(x_H, x_S, x_E)\alpha\delta(x_E, \bar{x}_E)} = \frac{\sum_{x_H,x_E} p(x_S, x_H, x_E)\delta(x_E, \bar{x}_E)}{\sum_{x_H,x_E,x_S} p(x_H, x_S, x_E)\delta(x_E, \bar{x}_E)}$$

which is exactly the same value given in Section 3.2. In other words, we note that $\alpha$ has absolutely no effect on the posterior probability since $\alpha$ cancels out in the numerator and denominator. Therefore, we can't really interpret $\alpha$ (or its constituent $\alpha_i$ values) as a form of strength of evidence at all. The absolute value of $\alpha$ (as long as $\alpha \neq 0$) is irrelevant. Since there is no reason to have $\alpha_i < 0$, we will assume that $\alpha_i > 0$ for each $i$.

### 4.1.3 Most likely assignments

The set of most likely assignments are also not affected by $\alpha_i$, as long as $\alpha_i > 0$ for all $i \in E$ (so that $\alpha > 0$). We have:

$$x_S^* = \operatorname*{argmax}_{x_S} p(x_S, \bar{x}_E) = \operatorname*{argmax}_{x_S} \max_{x_E} p(x_S, x_E)\alpha\delta(x_E, \bar{x}_E) = \operatorname*{argmax}_{x_S} \max_{x_E} p(x_S, x_E)\delta(x_E, \bar{x}_E)$$

# 5 Generalization of Evidence, Uncertain Evidence, and Virtual Evidence

We now address uncertain evidence. There are several forms of uncertain evidence which relate mostly to how numeric scores are associated with different alternate evidence hypotheses [12]. In this section, we describe what has been called *virtual* or *intangible* evidence [11, 12].

The evidence in Section 4.1 will henceforth referred to as *hard evidence*. Recall that the evidence may come from some process external to the current distribution represented by the Bayesian network, and the evidence values in some sense might not at all be associated with the particular current parameter values of the CPTs in the original Bayesian network (i.e., the one without the virtual children defined in Section 4). This process might be an event $\eta$ that was not part of the original sample space partition. The effect is that the probability of every variable assignment $X_U = x_U$ is multiplied by either $\alpha$ or zero depending on if $x_U$ agrees with $\bar{x}_E$. When we consider the pair of values $(\alpha, 0)$ where $\alpha$ is any value $\alpha > 0$, we can view this as a set of external relative weights on the variable assignments. With this particular set of weights, we see that the weight corresponding to any assignment $X_E \neq \bar{x}_E$ is infinitely smaller than the weight corresponding to the assignment $X_E = \bar{x}_E$, because (loosely) $\alpha/\infty = 0$. Since we saw in Section 4.1 that the value of $\alpha$ did not matter, we care here only about relative weighting. That is, we want one assignment weight to have an infinite factor less weight the other assignment weight. Therefore, any value $\alpha > 0$ will do to achieve these ratios.

Virtual evidence arises when we relax the notion that the weight ratios are either finite (for one assignment $\bar{x}_E$) or zero (for the remaining assignments). For each $x_E \in \mathcal{D}_{X_E}$, there is a function $\alpha : \mathcal{D}_{X_E} \to \mathbb{R}^+$, that maps from assignments to variable $x_E$ to non-negative reals. The probability of a variable assignment of $x_A$ gets multiplied by $\alpha(x_E)$ where $E \subseteq A$. As we will see, the specific values of the function $\alpha(\cdot)$ do not really matter, as long as they are non-negative. The only thing that matters is their ratio (as in the case above).

This procedure can be seen as expressing relative degrees of numerical weight regarding the value of $X_E$ taking on different assignments. The evidence that comes in, rather than specifying an event with certainty, instead says that a variety of events could happen, and the numeric weights $\alpha(\cdot)$ provide a numerical coding of this uncertainty. It might be said that these weights express relative "degrees of believe" or "degrees of confidence" in those particular assignments of variables. A key point is that these belief quantities are obtained external to the Bayesian network, just as is hard evidence. In other words, with hard evidence, we gain some information about an event — this information is the discovery that in the event $X_E = \bar{x}_E$. In virtual evidence, we also gain information, but rather than about one particular event, we have numeric scores each for a variety of events.

We can interpret virtual evidence as a sample space event as described in Section 4.0.1. In this case, however, an event $\eta \in \mathcal{F} \setminus \mathcal{F}_X$ occurred since all the events in $\mathcal{F}_X$ correspond to a particular single assignment of random variables $X_A = x_A$. The event $\eta$ (equivalently $V = 1$) is conditionally related to a partition of the event space corresponding to $\mathcal{D}_{X_E}$. That is, we have that

$$p(\eta|X_E = x_E) = p(V = 1|X_E = x_E) \propto \alpha(x_E)$$

Why the likelihood $p(V = 1|X_E = x_E)$ is related only proportionally to $\alpha(x_E)$ is described in Section 5.1. In this case, we have event equivalence $\eta \equiv \{V = 1\}$ where $V$ is the extra virtual random variable. It is important to realize that in making the above statement, we have made a conditional independence assumption about the event $\eta$ and the rest of the BN [12]. I.e., the assumption is that $V \perp\!\!\!\perp X_{U\setminus E}|X_E$. This is perhaps not surprising, however, as this is precisely the assumption given on the right of Figure 1, where $V_i \perp\!\!\!\perp X_{U\setminus\{i\}}|X_i$ (also see the right of Figure 2 for an example of this). The validity of this conditional independence assumption, as does the validity of any conditional independence assumption, depends on the specific application in which evidence is utilized (Section **??**).

As we saw in Section 4.1, the the function $\alpha(\cdot)$ matters only up to a constant of proportionality. In other words, we could substitute $\alpha(x_E)$ with $\alpha'(x_E) = \beta\alpha(x_E)$ for any $\beta > 0$ and the effect would be the same. We will leave a discussion of the interpretation of these ratios to Section 5.2. For now, we define how it is that the ratios are all that matters:

## 5.1 Why do only the ratios matter?

First, let us suppose that $|\mathcal{D}_{X_E}| = M > 0$. We take values $x_E \in \mathcal{D}_{X_E}$ in lexicographic order such that $\alpha^j \triangleq \alpha(x_E)$ if $x_E$ is the $j^{th}$ in this order, for $j = 1 \ldots M$, where $\alpha^j \geq 0$. We generalize the delta functions above to take on more than 0/1 values, specifically:

$$\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \delta(x_E; (\bar{x}_E^1, \alpha^1), (\bar{x}_E^2, \alpha^2), \ldots, (\bar{x}_E^M, \alpha^M)) \triangleq \begin{cases} \alpha^1 & \text{if } x_E = \bar{x}_E^1 \\ \alpha^2 & \text{if } x_E = \bar{x}_E^2 \\ \vdots & \\ \alpha^M & \text{if } x_E = \bar{x}_E^M \end{cases}$$

We see that $\delta(x_E; \{(\bar{x}_E^j, \alpha'^j)\}_{j=1}^M) = \beta\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)$ when $\alpha'^j = \beta\alpha^j \ \ \forall j$.

In other words, this new delta function (generalizing the Dirac delta), applies the value $\alpha^j$ when our evidence variables $X_E$ take on particular value $\bar{x}_E^j$.

Suppose now that we define a new quantity:

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) \triangleq \sum_{x_U} p(x_U)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)$$

This quantity generalizes the probability of evidence defined in Section 2.3, but here it may be seen as the score of the (uncertain) evidence. It is not entirely correct to call this a probability unless certain constraints are made on the $\alpha(\cdot)$ weights (described below). We may consider it to be the expected value of the uncertain evidence, however, as we have:

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{x_U} p(x_U)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{x_E} p(x_E)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = E_p[\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)]$$

In this view, standard evidence may be seen as the expected value of a Dirac delta function since $p(\bar{x}_E) = E_p[\delta(x_E, \bar{x}_E)]$, so that the above generalization to uncertain virtual evidence is quite natural.

Relating to the three procedures given earlier, the values of the vector $(\alpha^1, \alpha^2, \ldots, \alpha^M)$ do not really matter up to a constant $\beta > 0$. In other words, we could just as easily use vector $(\alpha'^1, \alpha'^2, \ldots, \alpha'^M)$ where $\alpha'^i = \beta\alpha^i$ for any $\beta > 0$. All that matters for the three procedures is the relative ratios of the various alphas, or $\alpha^i/\alpha^j$, $i, j \in 1 : M$. We define what we mean by "matters" in the next few sections below. In order to be able to generalize standard evidence, we also allow some of the coefficients to be zero (e.g., $\alpha^j = 0$) in which case the ratios might be zero or infinite, as described earlier. We next examine our three procedures:

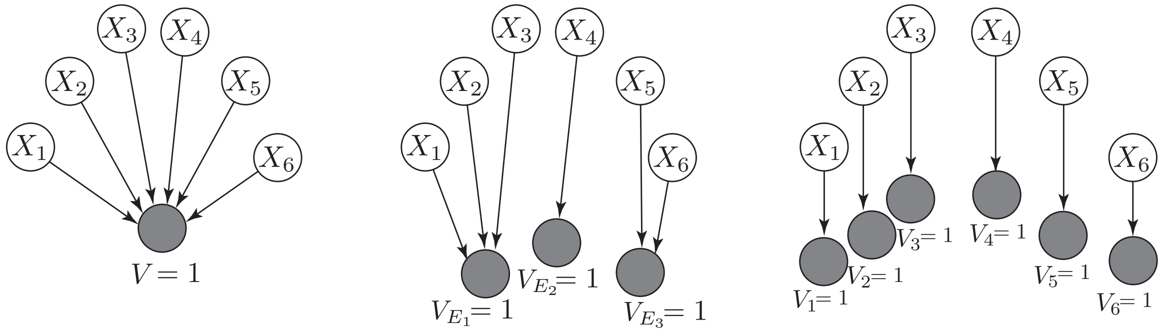

Figure 2: Virtual Evidence variants depending on the factorization of the evidence.

### 5.1.1 Score and Probability of virtual evidence

To compute the score of the virtual evidence, we perform the following:

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{x_U} p(x_U)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{j=1}^M \alpha^j \left( \sum_{x_U} p(x_U)\delta(x_E, \bar{x}_E^j) \right) \tag{1}$$

$$= \sum_{j=1}^M \alpha^j p(\bar{x}_E^j) \tag{2}$$

so the probability of the virtual evidence is simply a weighted sum of the probabilities of each of the individual hard-evidence evidence probabilities, where the weights are just the $\alpha_i$ values. In this form, it is easy to see how if $\alpha^j = \delta(j, \bar{j})$ for a particular value $\bar{j}$, then we have standard evidence.

There are additional points worth making here as well, namely the $\alpha^j$ values might themselves factorize, constraints may be placed on the $\alpha^j$ values to make the virtual evidence score a probability, and if no constraints are given it is only the ratios that matter.

First, the $\alpha^j$ values might themselves factorize. In other words, we can think of

$$\alpha^j = \prod_{i \in E} \alpha_i^j$$

where $\alpha_i^j$ is the weight value applied to the probability of any assignment to $X_U$ whenever $X_i = \bar{x}_i^j$ for $i \in E$. In such case, we can write the virtual evidence score as:

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{j=1}^M \left( \sum_{x_U} p(x_U) \prod_{i \in E} \alpha_i^j \delta(x_i, \bar{x}_i^j) \right) \tag{3}$$

In this case, the evidence variables are free to vary separately with respect to each other in such a way that as long as one of the variables, say $X_i, i \in E$ takes on a particular value $X_i = \bar{x}_i^j$, then the probability of the assignment any $X_U$ with $X_i = \bar{x}_i^j$ will be affected by the factor $\alpha_i^j$, and this is irrespective of the assignments to any other variables $X_{i'}$ where $i' \in E, i' \neq i$.

On the other hand, it might be desirable to have a weight value jointly for each a pair or a group of evidence variable values. We might, for example, partition the evidence set $E$ into disjoint subsets $E = \{E_1, E_2, \ldots, E_L\}$. In this case, we get

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \sum_{j=1}^M \left( \sum_{x_U} p(x_U) \prod_{l=1}^L \alpha_l^j \delta(x_{E_l}, \bar{x}_{E_l}^j) \right)$$

In this case, we have a separate value $\alpha_l^j$ for each set of assignments to the subgroup $X_{E_l}$. Clearly, this approach generalizes and subsumes both Equation 1 (when $L = 1$) hand Equation 3 (when $L = |E|$).

In fact, uncertain evidence in this way (and depending on the factorization of $\alpha^j$) can easily be encoded in a BN, the same way as hard evidence can. Here, any set of evidence nodes $X_E$ that has received virtual evidence possesses a new virtual child node $V$, outside of the universe, that is always observed to have value 1 (unity), and where we set $p(V = 1|X_E = x_E) = \delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)$. Again, this is only a partial specification of the CPT $p(V|X_E)$ as the values for $p(V \neq 1|X_E = x_E)$ need not be specified since they are never used. Of course, the value $V = 1$ is arbitrary, and it could be any observed value for $V$ as long as the corresponding CPT agrees with that value in how it applies the uncertain evidence. Also, the factorization of $\alpha^j$ is expressed depending on the number of virtual children that exist. If there is one global virtual child $V = 1$, then $\alpha^j = \beta p(V = 1|X_E = \bar{x}_E^j)$ does not (necessarily) factorize, as shown on the left in Figure 2, where $\beta > 0$ is a scalar constant of proportionality (which as we will see does not matter). Equation 2 becomes

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \beta \sum_{j=1}^M p(V = 1|X_E = \bar{x}_E^j) p(\bar{x}_E^j)$$

If we wish to have partial factorization, so that say $\alpha^j = \alpha_{E_1}^j \alpha_{E_2}^j \alpha_{E_3}^j$, we express this in a BN as shown in the center of Figure 2. In this case, $\alpha^j = \beta_{E_1} p(V_{E_1} = 1|X_{E_1} = \bar{x}_{E_1}^j) \beta_{E_2} p(V_{E_2} = 1|X_{E_1} = \bar{x}_{E_2}^j) \beta_{E_3} p(V_{E_3} = 1|X_{E_1} = \bar{x}_{E_3}^j)$, where again $\beta_i > 0$ are constants of proportionality. Because of the moralization property of Bayesian networks (namely that once moralized, all parents of an observed child have to appear in at least one clique in the junction tree), such factorization of the virtual evidence can have significant computational complexity reductions. If we wish to have full factorization, then we get the right of Figure 2. Here, $\alpha^j = \prod_i \beta_i p(V_i = 1|X_i = \bar{x}_i^j)$.

Second, what are the constrains on $\alpha^j$ to make the score a true probability. First, we must note that the model in Equation 2 is not a mixture model — i.e., we are not saying that the constraints are such that $\sum_j \alpha^j = 1$ and where $0 \leq \alpha^j \leq 1$. Rather, we require only that $\alpha^j \geq 0$ (or in the factorized case, that $\alpha_i^j \geq 0$ for all $i, j$). When we view virtual evidence as virtual pendant children, connected nothing except for their parents, we see that the scores are viewed as likelihoods $p(V = 1|x_E = \bar{x}_E)$.

Third, we can see again that assuming $\alpha^j > 0 \; \forall j$, given fixed ratios of the $\alpha^j$ values, the virtual evidence score is the same up to a constant factor. In other words, if we were to compute the virtual evidence score (Equation 2) with two sets of virtual evidence values, $\{\alpha^j\}_j$ and $\{\alpha'^j\}_j$ with $\alpha'^j = \beta\alpha^j$, then the final virtual evidence score would also have the same resulting ratio, i.e.,

$$p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \frac{p(\{(\bar{x}_E^j, \alpha'^j)\}_{j=1}^M)}{\beta}.$$

Therefore, we can normalize our virtual evidence score to obtain any desired set of $\alpha^j$ values and only effect the result with a multiplicative constant.

### 5.1.2 Posterior probability of hidden variables

The computation of posterior probabilities (or the "revised belief") shows more precisely how $\beta$ does not matter since the posterior probabilities are identical for all $\beta > 0$. The only thing that matters is the ratios, i.e., $\alpha^j/\alpha^k$. We have

$$p(x_S|\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) = \frac{p(x_S, \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)}{p(\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)} = \frac{\sum_{x_H, x_E} p(x_S, x_H, x_E)\delta(x_E, ; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)}{\sum_{x_H, x_E, x_S} p(x_H, x_S, x_E)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M)} \tag{4}$$

$$= \frac{\sum_{j=1}^M \alpha^j p(x_S, \bar{x}_E^j)}{\sum_{j=1}^M \alpha^j p(\bar{x}_E^j)} \tag{5}$$

The numerator and denominator in the 2nd equality are both themselves scores. The ratio of the scores is a proper probability since any common constant within the $\alpha^j$ values will cancel out. Therefore, it is clear that the posterior probabilities do not at all depend on the $\alpha^j$ values other than their relative relationships $\alpha^j/\alpha^i$. This holds also of course for the factorized versions described in the previous section.

### 5.1.3 The Variable Assignment with the Maximum Score: VEterbi

We can compute the Virtual Evidence Viterbi (VEterbi) assignment of the variables as well.

The set of most likely assignments are also not affected by $\{\alpha^j\}_j$ other than their ratios, as long as $\alpha^j \geq 0$ for all $i \in E$. We have for all $\beta \geq 0$:

$$\begin{aligned} x_S^* &= \operatorname*{argmax}_{x_S} p(x_S, \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) && (6) \\ &= \operatorname*{argmax}_{x_S} \max_{x_E} p(x_S, x_E)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) \\ &= \operatorname*{argmax}_{x_S} \max_{j \in 1:M} p(x_S, \bar{x}_E^j)\alpha^j \\ &= \operatorname*{argmax}_{x_S} \max_{j \in 1:M} p(x_S, \bar{x}_E^j)\beta\alpha^j \\ &= \operatorname*{argmax}_{x_S} \max_{x_E} p(x_S, x_E)\beta\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) \end{aligned}$$

So the maximum assignment is the same regardless of the actual values of $\alpha^j$ as long as the ratios are identical.

In the above, the result $x_S^*$ is the maximum assignment to $X_S$ with respect to the maximum assignment to the virtual evidence variables $X_E$. This is the generalization of the Viterbi assignment, since a Viterbi assignment computes the maximum over all hidden variables jointly. We may instead wish to define the best maximum assignment as:

$$\begin{aligned} x_S^* &= \operatorname*{argmax}_{x_S} \sum_{x_E} p(x_S, x_E)\delta(x_E; \{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M) && (7) \\ &= \operatorname*{argmax}_{x_S} \sum_{j=1}^{M} p(x_S, \bar{x}_E^j)\alpha^j && (8) \end{aligned}$$

Equation 6 corresponds to the Viterbi approximation to the maximum given by Equation 7, which arguably incorporates more of the virtual evidence information. Note that in both cases, again it is only ratios $\alpha^j/\alpha^k$ that matter. The second case, however, can in general be much more computationally difficult.

### 5.2 How can we use $P(V = 1|X_E = \bar{x}_E)$ without having $P(V = 1, X_E = \bar{x}_E)$?

As mentioned in earlier sections, we can think of hard evidence as obtaining partial information about some otherwise unknown sample of the set of random variables $X_U$ described by a BN. In one way or another, this partial information comes from a mechanism external to the BN itself. If the BN reflects truth (meaning the physical object we are representing can perfectly accurately be represented by a BN under a certain parameterization), then evidence can be seen as a sample from the BN where only a subset $X_E$ $E \subseteq U$ of the random variables are revealed. If the BN is only an approximate model of truth, then our assumption is merely that the hidden variables are our best guess as to the process that lead to the observed variables being their current values. In either case, the external evidence says that we should "believe" a sample $X_U = x_U$ infinitely more when the portion $X_E = \bar{x}_E$ than when $X_E \neq \bar{x}_E$.

Virtual evidence is no different, other than the fact that we might obtain a more general form of external information than just hard knowledge of the outcome of some subset of the random variables. Given what we saw about ratios of values above, the safest way to interpret virtual evidence then is to think of it in the following way based on the relative values of the $\alpha^j$ values. That is, the external information we obtain $\{(\bar{x}_E^j, \alpha^j)\}_{j=1}^M$ says only the following:

- There was a sample in the underlying sample space (e.g., $\eta$ or $V = 1$). We have learned from this source something about a subset $X_E$ of the random variables, where $E \subseteq U$. Other than the effect on $X_E$, the source has no *direct* influence on the remaining random variables $X_{U \setminus E}$. All other things being equal (i.e., for all possible values of $X_{U \setminus E}$), we have learned that an assignment to $X_U$ with $X_E = \bar{x}_E^j$ should be believed a factor $\alpha^j/\alpha^i$ more an assignment to $X_U$ with $X_E = \bar{x}_E^i$, for all $1 \leq i, j < M$. More specifically, the probability of any assignment to $X_U$, i.e., $p(X_U = x_U)$ should be multiplied by $\alpha^j$ when $X_E = \bar{x}_E^j$, for $j = 1 \ldots M$.

One can give the above the interpretation of information coming from source external to the BN that provides a form of relative "belief" of different random variable values. The information, however, is only relative, in that it says that one event (i.e., a sub-variable assignment) should be preferred by some constant factor more than another event. This is easily encoded using likelihoods $p(V = 1|X_E = \bar{x}_E^j)$.

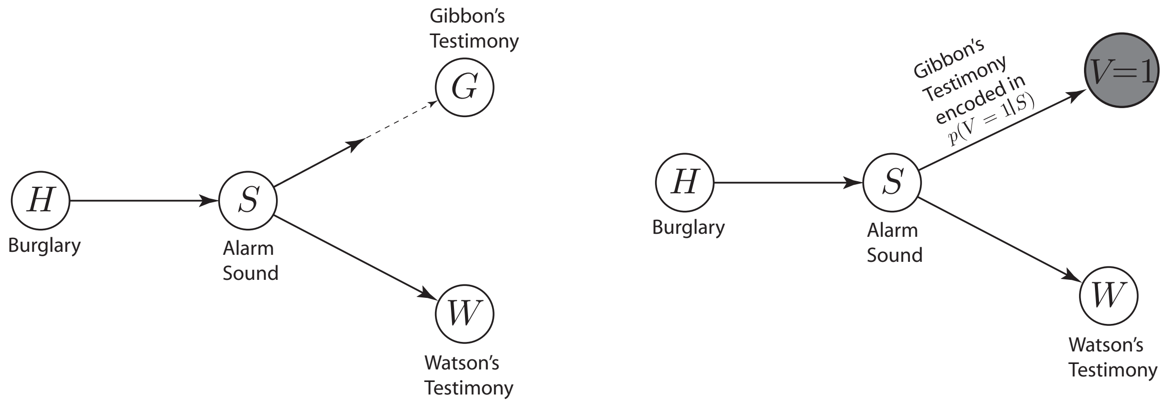

Figure 3: Left: Figure 2.1 from page 43 of Pearl [11]. Right: An interpretation in terms of virtual evidence via the likelihood $p(V = 1|S)$

Typically, the use of the conditional term $p(X_E = \bar{x}_E^j | V = 1)$ means that we have available both the joint $p(X_E = \bar{x}_E^j, V = 1)$ and the prior $p(V = 1)$. I.e., that the event $\{V = 1\}$ is one that has the same status as any of the other events being modeled $\mathcal{F}_X$ (see Section 4.0.1). The event $V = 1$, however, is special in that we only have available information about it in terms of how it relates conditionally to other variables in our system, namely $X_E$. That is, the following assumptions are being made:

- $V \perp\!\!\!\perp X_{U \setminus E} | X_E$, meaning that $V$ only indirectly effects variables other than $X_E$. $X_E$ renders the event $V$ independent of the rest of the model $X_{U \setminus E}$.
- Information in how $V$ effects the model only is available in terms of $P(V = 1|X_E)$, and this is an atomic object (i.e., the virtual evidence is not amenable to deconstruction in terms of Bayes rule to obtain $p(X_E|V = 1)p(V = 1)/p(X_E)$. This means that the joint distribution over the entire model $P(X_U, V)$ is not available.

There are several reasons why this may be valid. First, many statistical models inherently represent only the conditional rather than the joint distribution. For example, discriminative models such as multi-layered perceptrons [3], support-vector machines [17] (when endowed with a distribution), and various conditional maximum entropy models [14] represent only the construct $p(A|B)$ without ever needing any representation of the joint distribution $p(A, B)$. Given observations consisting of both events of the form $V = 1$ and $V = 0$ (for a binary $V$ variable), it would be possible to learn the distribution $p(V = 1|X_E)$ and apply it to the model as above.

Second, there are scenarios where we can reason only about ratios of likelihoods $p(V = 1|\bar{x}_E^j)/p(V = 1|\bar{x}_E^i)$ rather than the absolute values of the likelihoods themselves. Since the ratios are all that matters for computation, we are comfortable when this occurs.

For example, in Pearl's text ([11], page 43), an example is given that consists of a BN with four variables, $B$ for burglary, $S$ for alarm sound, $W$ for Watson's testimony, and $G$ for Gibbon's testimony. The network so given is such that the following factorization holds (Figure 2.1 on page 43 in [11], reproduced in Figure 3 here.).

$$p(H, S, W, G) = p(H)p(S|H)p(G|S)p(W|S) \tag{9}$$

The ultimate goal is to compute the probability of $H$ (a burglary occurred) given knowledge about any of the following: the alarm went off $S = 1$, Watson's testimony $W$, and/or Gibbon's testimony $G$. In the example (section 2.2.2 of [11]), however, it is not known if if the alarm went off, so that the variable $S$ is hidden. The only thing that is known is perhaps some information about $G$ and or $W$.

Virtual evidence arises in [11] when it is not known how precisely the value of $G$ effects the remainder of the network, but rather where only some aspect of $G$ provides us with ratios of preferences about $H$. For example, from $G$ we are able only to ascertain that it is four times more likely that the alarm went off ($S = 1$) than otherwise ($S = 0$). In other words, any probability of the joint set of random variables $p(H, S, W)$ where $S = 1$ (alarm) should be multiplied by a number that is four times as large as when $S = 0$ (no alarm). We are neither able to obtain (nor are we interested in obtaining) the probability of $G$ or the direct relationship between $G$ and the rest of the network. Therefore, rather than encoding this information source as a variable $G$ which is a child of $S$ (as done in Equation 9), we can instead think of the application of a generalized delta function the the joint probability of the three variables remaining after $G$ is removed from the network, i.e., $p(h, s, w)\delta(s; \{(1, 8), (0, 2)\})$. This means that when $s = 1$ the probability is multiplied by $8$ and when $s = 0$ it is multiplied by $2$, a ratio of $4$ to $1$. In other words, the information from $G$ imbues the variable $S$ with virtual evidence favoring $S = 1$ with $4$ to $1$ odds. The universe only consists of the variables $H, S, W$, but there is information from outside the universe giving preferential treatment to certain values of $S$.

While Pearl [11] treats this as a Bayesian network in his Figure 2.1, he uses a different notation for his variable $G$ which is not mentioned in the text explicitly. Namely, he states that $G$ is a child of $S$ only via a multi-patterned edge, where the first part is solid and the second part is dashed. What is meant by this diagram is the notion of virtual evidence, as given on the right in Figure 3. As mentioned above, this could be encoded by a Bayesian network with a variable $V$, always observed so that $V = 1$, with CPT $p(V = 1|S = s) = 8\delta(s, 1) + 2\delta(s, 0)$.

As is mentioned in Pearl's text, this external information source does not really impart a probability. In other words, it is correct neither to say that ($p(S = 1|G = g) = 0.8$,$p(S = 0|G = g) = 0.2$) nor to say ($p(G = g|S = 1) = 0.8$,$p(G = g|S = 0) = 0.2$) for any value of $g$ since this would imply that this evidence is obtained by a finding of a variable within $X_U$. We next quote Pearl[11] directly[1]

> These difficulties arise whenever the task of gathering evidence is delegated to autonomous interpreters who, for various reasons, cannot explicate their interpretive process in full detail but nevertheless often produce informative conclusions that summarize the evidence observed. In our case, Mr. Holmes [an external observer] provides us with a direct mental judgment, based on Mrs. Gibbon's testimony, that the hypothesis *Alarm sound* should be accorded a confidence measure of 80%. The interpretation process remains hidden, however, and we cannot tell how much of the previously obtained evidence was considered in the process. Thus, it is impossible to integrate this probabilistic judgment with previously established beliefs unless we make additional assumptions. [i.e., we cannot compute $p(G, S)$.]
>
> The prevailing convention in the Bayesian formalism is to assume that probabilistic summaries of virtual evidence are produced independently of previous information; they are interpreted as local binary relations between the evidence and the hypothesis upon which it bears, independent of other information in the system. For this reason, we cannot interpret Mr. Holmes's summary as literally stating $P(S|G) = 0.8$. $P(S|G)$ should be sensitive to variations in crime rate information — P(H) — or equipment characteristics — $P(S|H)$. The impact of Gibbon's testimony should be impervious to such variations. Therefore, the measure $P(S|G)$ cannot represent the impact the phone conversation has on the truth of *Alarm sound*.
>
> The likelihood ratio, on the other hand, meets this locality criterion, and for that reason probabilistic summaries of virtual evidence are interpreted as conveying likelihood information. For example, Mr. Holmes's summary of attributing 80% credibility ot the *Alarm sound* event can be interpreted as:
>
> $$P(G|\text{Alarm sound}) : P(G|\text{No alarm sound}) = 4 : 1$$

In Pearl's description, he states that the crux lies in the *local binary relations between the evidence and the hypothesis upon which it bears.* By this, it is meant the evidence $V$ should only influence $S$ and should not directly influence anything else in the network, and that it should influence $S$ only via ratios. These ratios are exactly what may be encoded by the virtual evidence construct given on the right of Figure 3. This information is obtained from a source external to the universe modeled by the BN's probability distribution. In the example, the information was obtained by an interview of Mrs. Gibbon and imparted into the network by the application of weights to the probabilities of the variables within the universe. Another example of this sort is given in [12] and in the references contained therein.

A third reason for the validity of the utilization of $p(V = 1|X_E)$ is that we may wish merely to impose constraints on certain variable configurations or states in our model. Perhaps there are only a subset of values $\mathcal{D}'_{X_E} \subset \mathcal{D}_{X_E}$ that are to be allowed. We may set $p(V = 1|X_E = x_E) = \delta\{x_E \in \mathcal{D}'_{X_E}\}$. Parameter learning can easily proceed, for example, in a model that is subject to these constraints. The semantics of the constraints, moreover, are similar to those used in constraint-networks [5]. In fact, in this way, virtual evidence can be seen to provide BNs the ability to represent hybrid constructs (see Section 5.4).

[1] In [11], a Mr. Holmes is the person who discusses with Mrs. Gibbon's her view of the alarm going off, and gleans from this discussion information external to the universe leading to the 4 to 1 preference in favor of alarm.

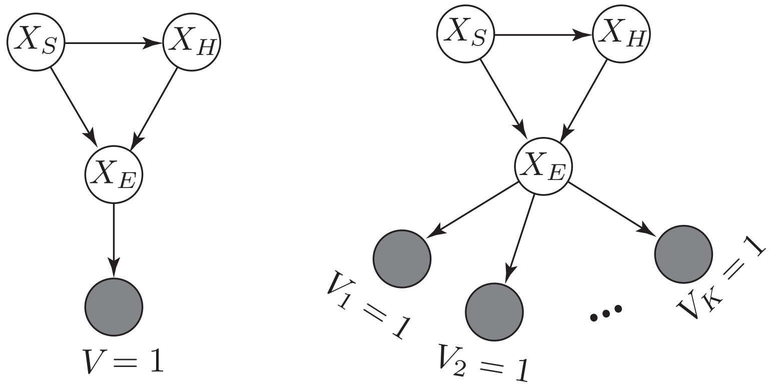

Figure 4: Left: The conditional independence assumption necessary for virtual evidence, namely that $V \perp\!\!\!\perp \{X_S, X_H\} | X_E$. Right, multiple sets of virtual evidence combine together in a natural way by having multiple virtual children.

## 5.3 Virtual Evidence, Bayesian Inference, and Bayesian Reasoning

Bayesian reasoning has a long history in statistics, and it would lead us far astray to even touch the surface of its complexities. We refer the reader to [15, 11] and the multitudinous references in Bayesian reasoning that are available (and that we do not cite).

The essential tenet of Bayesian reasoning is that given prior uncertainty $p(A)$ over an event $\{A\}$, and given some evidence $E$ that is related to $\{A\}$, that prior can be updated based on the likelihood to give posterior uncertainty, i.e., $P(A|E) = \left(\frac{p(E|A)}{p(E)}\right) p(A)$. This principle can be applied in a number of cases, including:

- Bayesian statistical inference, where $A = \theta$ is a set of parameters of a parametric [15] or non-parametric statistical model, and where $E = x_{1:N}$ is data that is drawn from an unknown distribution. The goal is to produce a posterior over parameters $p(\theta|x_{1:n})$ which can be used for future predictions, mixtures, maximum a-posterior estimation, confidence estimation, and so on (see [15]).
- Belief revision, where $A$ represents some event in the world (e.g., “it is raining outside”), and $E$ represents some potentially related event (e.g., “the ground is wet”). In this case, the Bayesian approach simply uses Bayes rule to update prior beliefs about the state of the world to posterior beliefs.

Virtual evidence corresponds to Bayesianism in that it is possible to use Bayes rule to derive the posterior that is given in Equation 5. Specifically, we have:

$$\begin{aligned}
p(x_S|V=1) &= \sum_{x_E}\sum_{x_H} p(x_S, x_E, x_H|V=1) && (10)\\
&= \sum_{x_E}\sum_{x_H} \frac{p(V=1|x_S, x_E, x_H)p(x_S, x_E, x_H)}{p(V=1)} && (11)\\
&= \frac{\sum_{x_E}\sum_{x_H} p(V=1|x_S, x_E, x_H)p(x_S, x_E, x_H)}{\sum_{x'_S}\sum_{x'_E}\sum_{x'_H} p(V=1|x'_S, x'_E, x'_H)p(x'_S, x'_E, x'_H)} && (12)\\
&= \frac{\sum_{x_E}\sum_{x_H} p(V=1|x_E)p(x_S, x_E, x_H)}{\sum_{x'_S}\sum_{x'_E}\sum_{x'_H} p(V=1|x'_E)p(x'_S, x'_E, x'_H)} && (13)
\end{aligned}$$

which is identical to Equation 5. Bayes rule was used to get Equation 11, but it was necessary to utilize the conditional independence assumption that $V \perp\!\!\!\perp \{X_S, X_H\} | X_E$ to obtain Equation 13 (see left of Figure 4).

Therefore, while Bayesian reasoning lies at the heart of both Bayesian inference and Bayesian belief updating, they generally are applied to different random objects.

### 5.3.1 Soft Evidence vs. Virtual Evidence

Another form of belief updating exists as well, including Jeffrey's evidence [12], or what is sometimes called *soft evidence*. Unlike virtual evidence, where one has the likelihoods $p(V = 1|x_E)$, one instead has a probability distribution $p'(x_E)$ that is distinct from $p(x_E) = \sum_{x_{U \setminus E}} p(x_U)$. The goal in this approach is to update $p(x_U)$ to $p'(x_U)$ so that $p'(x_{U\setminus E}|x_E) = p(x_{U\setminus E}|x_E)$. This is done using Jeffrey's update rule

$$p'(x_U) = \sum_{x'_E} p(x_U|x'_E)p'(x'_E)$$

In this notation, we note that $x_U$ includes $x_E$ (since $E \subset U$), and have that $p(x_U|x'_E) = 0$ if $x_E \neq x'_E$.

Soft evidence is fundamentally a different form of evidence than is virtual evidence. Soft evidence is a direct statement about the underlying statements made by a model (e.g., it directly states something about the parameters of a parametric model). In the case above, soft evidence states that the marginal of the distribution over $X_E$ must equal $p'(x_E)$. Virtual evidence, on the other hand, does not prescribe properties to the distribution, but rather is a generalization of evidence regarding some of the random variables (and not their parameters) in the model. Of course, virtual evidence can be used to train the parameters of the model (say using maximum likelihood) but the way in which the evidence influences the parameters is only indirect with virtual evidence.

One way to see the difference between the two approaches is to see how the different forms of evidence combine. Suppose that we have two unequal sets of soft-evidence, i.e., $p'(x_E)$ and $p''(x_E)$. These two pieces of evidence do not combine, since if we apply Jeffrey's rule twice, only the latter application will survive since each application wipes out whatever is the current marginal over $X_E$. If the two sets of soft-evidence do not agree, they are fundamentally incompatible with each other.

Multiple sets of non-agreeing virtual evidence, on the other hand, can be sensibly unified, i.e., the different evidence will combine together to produce a logical combined result. For example, given the two virtual evidence constructs $p(V = 1|x_E)$ and $p(V' = 1|x_E)$ where $V$ and $V'$ are different random variables, the natural generalization of the above is to combine the two by multiplication, leading to the updated posterior:

$$p(x_S|V = 1, V' = 1) = \frac{\sum_{x_E} \sum_{x_H} p(V = 1|x_E)p(V' = 1|x_E)p(x_S, x_E, x_H)}{\sum_{x'_S} \sum_{x'_E} \sum_{x'_H} p(V = 1|x_E)p(V' = 1|x_E)p(x'_S, x'_E, x'_H)} \tag{14}$$

under the appropriate set of conditional independence statements. Generalizing this further to the vector observation $V_{1:K} = (1, 1, \dots, 1)$, we get:

$$p(x_S|V_{1:K} = (1, 1, \dots, 1)) = \frac{\sum_{x_E} \sum_{x_H} p(x_S, x_E, x_H) \prod_{k=1}^K p(V_k = 1|x_E)}{\sum_{x'_S} \sum_{x'_E} \sum_{x'_H} p(x'_S, x'_E, x'_H) \prod_{k=1}^K p(V_k = 1|x_E)} \tag{15}$$

as shown on the right in Figure 4.

## 5.4 Virtual Evidence, Undirected Graphical Models, and Factor Graphs

An undirected graphical model is a representation of any probability distribution that factors with respect to an undirected graph. In particular, if $p(x)$ factors with respect to undirected graph $G = (V, E)$, then we can write:

$$p(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(X_c) \tag{16}$$

where $\mathcal{C}$ are the set of cliques in the graph. This means that any distribution $p(x)$ that factors with respect to graph $G$ must be such that it can be validly written in this way.

Any Bayesian network can be written to factor as an undirected graph by defining cliques corresponding to each child and its parents (although some of the BN's factorizations, in particular V-structures, might be lost in this process when going to an undirected model). I.e., for each variable $i \in U$ and its parents $\pi_i$, we define a clique $c_i = \{i\} \cup \pi_i$, and form the set of cliques as $\mathcal{C} = \cup_i c_i$. Of course in this case, $Z = 1$ since the model already normalizes.

Virtual evidence can thus be seen as applying another clique function on a set of variables corresponding to $X_E$. Starting start with Equation 16, we add a virtual child corresponding to the factor $p(V_i = 1|X_i = x_i) = f(x_i)$. The new distribution becomes

$$p'(x) = \frac{1}{Z} f(x_i) \prod_{c \in \mathcal{C}} \phi_c(X_c) = f(x_i)p(x)$$

Since $X_i$ is only one variable, we can easily absorb the new factor $f()$ into any clique that contains $X_i$ without changing any of the factorization properties of the graph. In this case, the graphical structure does not change and the same graphical model and all of its corresponding factorization properties still applies. On the other hand, the distribution does not any longer properly normalize unless we were to specify the above with a different constant $Z'$. Supposing further that there does not exist a $c \in \mathcal{C}$ such that $E \subseteq c$, then we form a new clique $c' = E$ and define the corresponding clique function $\phi_{c'}(x'_c) = p(V = 1|X_{c'} = x_{c'})$. Here the graph structure would change (i.e., $X_E$ would turn into a sub-clique). Also, $Z$ once again would need to be recomputed to obtain a normalized distribution. However, as shown in the sections above, $Z$ is often not necessary.

A factor graph, on the other hand, is a bipartite graph representation of a probability distribution where factors in the distribution are made explicit. We are given a bi-partite graph $G = (V, F, E)$ where $V$ are the left-hand-side nodes, $F$ are the right-hand-side nodes, and $E \subseteq V \times F$ are the edges that exist only between left and right-hand sides. Each $v \in V$ is associated with a random variable and each $f \in F$ is associated with a factor over some subset of the random variables. For every node $f \in F$, the subset of nodes in $V$ connected to $f$ correspond to the arguments of the factor $f$, and we call this $V_f$. I.e., $f \in F$ corresponds to the factor $f(X_{V_f})$. A factor graph represents any distribution $p(x)$ that can be validly written in the following way:

$$p(x) = \frac{1}{Z} \prod_{f \in F} \phi_f(X_{V_f}) \tag{17}$$

where $\phi_f()$ are arbitrary non-negative functions.

A factor graph can be used to represent the factorization of a BN precisely, since all factors in the BN can be represented by some factor $f \in F$. Adding a virtual child $V_i$ to $X_i$ correspond to adding another factor to the model, one that places a constraint on the possible value of $X_i$. Lets do this for each of an undirected model and a directed model.

A factor graph representation of virtual evidence would add one more node into the right hand side of the graph $F$, the node would only be connected to $X_i$, node's factor function would be the factor $f(x_i)$. Therefore, the factor graph does change in response to a virtual child of a single variable. If $|E| > 1$, we would just add a factor $f'$ such that $X_{f'} = X_E$. The same normalization adjustment would occur here as what occurs in the undirected model.

When might one use virtual evidence vs. an undirected or factor graph? In some sense, there is no real difference, since any factor in a graph can be represented by an appropriate virtual evidence function $p(V = 1|x_E)$, up to a normalization constant. Sometimes, it is useful to parameterize factors in an undirected model using a log-linear form, so that $\phi_c(x_c) = \exp(\lambda^T g(x_c))$ where $\lambda$ is a vector of coefficients and $g(\cdot)$ is a vector of functions on the variables $x_c$. It would be possible to define a virtual evidence function in the same way, however.

It can be useful at times to specify everything in terms of a log-linear model. On the other hand, it is sometimes quite useful to consider locally normalized factors as in a BN, where $p(x_i|x_{\pi_i})$ can be specified as $x_i$ as a noisy function of its parents. In such a model, having the ability to add virtual evidence factors can be quite useful. The method of training also can have an influence on what to try. Recent work on discriminative training [18, 16, 9] are such that the exponential form is mathematically quite tractable, but such discriminative optimization criterion could just as easily be applied to a BN that utilizes virtual evidence. Therefore, the choice of what to use is really up to the user.

# 6 Applications

We present several applications where the use of virtual evidence has been usefully applied, and these include hybrid neural-network/hidden-Markov model-based speech recognition, backoff-based language models, and in dynamic Bayesian networks.

## 6.1 Virtual Evidence in Hybrid ANN/HMM systems

Hybrid artificial neural network (ANN) - hidden Markov model HMM systems [4] can be seen as utilizing virtual evidence on an underlying Markov chain backbone.

In a typical HMM, there are $T$ discrete random variables $Q_{1:T}$ that form a Markov chain, so that

$$p(q_{1:T}) = \prod_{t=1}^{T} p(q_t|q_{t-1})$$

and a set of continuous or discrete random variables $X_{1:T}$ which represent a transformation of the acoustic speech waveform but embedded in a vector stochastic process. The joint distribution in an HMM factorizes according to a BN as follows:

$$p(q_{1:T}, x_{1:T}) = \prod_{t=1}^{T} p(q_t|q_{t-1})p(x_t|q_t) \tag{18}$$

In a hybrid ANN/HMM system, on the other hand, local estimates of the temporally local posterior probabilities $p_{ANN}(q_t|x_t)$ are produced using a neural network (i.e., a 3-layer multi-layered perceptron which can be shown to approximate posterior probabilities when appropriately trained [3]). More typically, the estimates take the form $p_{ANN}(q_t|x_{t-\tau:t+\tau})$, where typically $\tau = 4$, so that the posteriors may utilize information not just from the current but also from surrounding time frames.

The question is, how to merge the information that the ANN locally garners from the acoustic time window with the stochastic process HMM. One approach, as is argued in [4], is to normalize the posterior by the prior, yielding:

$$p(q_t|x_{t-\tau:t+\tau})/p(q_t) = p(x_{t-\tau:t+\tau}|q_t)/p(x_{t-\tau:t+\tau})$$

and then to use that in place of the typical HMM local likelihood $p(x_t|q_t)$ from Equation 18. It is argued that since this is a scaled likelihood (scaling factor $p(x_{t-\tau:t+\tau})$, it is functionally equivalent to the HMM local likelihood $p(x_t|q_t)$. It can be easily seen, however, that the joint distribution $p(q_{1:T}, x_{1:T})$ does not factorize even proportionally with respect to any Bayesian network into a product of factors consisting of the likes of $p(x_{t-\tau:t+\tau}|q_t)/p(x_{t-\tau:t+\tau})$ [2]

The hybrid ANN/HMM systems, however, can be seen in terms of virtual evidence. In particular, the underlying universe of variables consists only of a Markov chain $Q_{1:T}$. The ANNs are providing external virtual evidence to each variable $Q_t$ in the form of

$$p(q_{1:T}) = \prod_{t=1}^{T} p(q_t|q_{t-1})\delta(q_t, \{(j, \alpha_t^j)\}_{j=1}^{|Q|})$$

and where the weight are such that

$$\alpha_t^j = p(q_t|x_{t-\tau:t+\tau})/p(q_t) = p(x_{t-\tau:t+\tau}|q_t)/p(x_{t-\tau:t+\tau})$$

Since we now know that the external information provides unique information only up ratios, we see that the normalization constant $p(x_{t-\tau:t+\tau})$ is irrelevant to our three main inference problems. Moreover, we may assume that each variable $Q_t$ has a virtual child $V_t = 1$ that is always observed to have value 1. The hybrid ANN/HMM may be given by having the virtual child use CPT with partial specification:

$$p(V_t = 1|Q_t = q_t) = p(q_t|x_{t-\tau:t+\tau})/p(q_t)$$

Let us now generalize this notion to arbitrary Bayesian networks. A (collection of) discrete node(s) $X_E$ $E \subseteq U$ is given virtual evidence via some process entirely separate from the universe of variables $X_U$. This external process is specified via a joint distribution over two sets of variables, $X_E$ and $Z$, and is given by $p(X_E, Z)$. Note that the variables within the distribution $p(X_E, Z)$ have some overlap with $X_U$ (namely $X_E$) but $p(X_E, Z)$ also contains innovation $Z$. The set $(X_E, Z)$ might be called a second *partially overlapping universe* relative to $(X_E, Z)$. Regardless of the name,

[2] Note that in the past, the ANN/HMM material was not explained using the notion of virtual evidence. Given the explanation herein, however, we can see how virtual evidence lends justification to this approach. Note also that the joint distribution will factorize with respect to an undirected graphical model with clique potential functions consisting of $\psi(q_t, x_{t-\tau:t+\tau})$. In particular, the ANN/HMM can be seen as an unnormalized undirected graphical model (or Markov random field), with clique potential functions $\psi$. We consider in this article, however, only a way of fitting a hybrid system into the Bayesian network setting.

it should be clear that the distribution $p(X_E, Z)$ is entirely separate from the distribution represented by the Bayesian network $p(X_U) = p(X_E, X_{U \setminus E})$. The distribution $p(X_E, Z)$ might itself be modeled by a Bayesian network, or might otherwise be modeled by a factors of neural networks, generalized linear models, support vector machines, or any other parametric or non-parametric and linear and/or non-linear form [6].

The question becomes, how do we utilize $p(X_E, Z)$ within $p(X_U)$? More specifically, suppose that $Z$ becomes known, so that in the partially overlapping universe, we find that $Z = z$. The resulting distribution becomes:

$$p(X_E, Z = z) = p(X_E | Z = z)p(Z = z) = p(Z = z | X_E)p(X_E)$$

The portion that overlaps $X_E$ is still unknown in both universes, but given $Z = z$ we have a refined notion of what $X_E$ should be from the 2nd universe. There are several possible ways that we might use the information obtained in the 2nd universe to the 1st universe's benefit.

First, we might apply to $p(X_U)$ a delta function of the following form:

$$p(x_U)\delta(x_E, \{(\bar{x}_E^j, p(\bar{x}_E^j, Z = z))\}_{j=1}^M) \equiv p(x_U)\delta(x_E, \{(\bar{x}_E^j, p(Z = z|\bar{x}_E^j)p(\bar{x}_E^j))\}_{j=1}^M) \quad (19)$$

$$\equiv p(x_U)\delta(x_E, \{(\bar{x}_E^j, p(\bar{x}_E^j|Z = z))\}_{j=1}^M) \quad (20)$$

In other words, we can apply either use full joint distribution $p(\bar{x}_E^j, Z = z)$ or equivalently the likelihood $p(\bar{x}_E^j|Z = z)$ from the 2nd universe as a virtual evidence weight for the set of random variables $X_E$ in the first universe. This is because $p(Z = z)$ is a constant for all values $x_E \in A_E$. Without loss of generality, let us call this the *posterior* case, since we are directly applying the posterior of $x_E$ given $z$.

Alternatively, we might apply the probability of $Z = z$ in the 2nd universe as the first universe's virtual evidence weights as follows:

$$p(x_U)\delta(x_E, \{(\bar{x}_E^j, p(Z = z|\bar{x}_E^j))\}_{j=1}^M) \quad (21)$$

Note that this is not equivalent to applying say $p(\bar{x}_E^j|Z = z)$ as weights since $p(x_E)$ is not necessarily a constant in the second universe. We call this case the *likelihood* case, since we apply as virtual evidence weights to $x_E$ in universe 1 the likelihood of the external data $Z = z$ given $\bar{x}_E^j$. This case corresponds to the hybrid ANN/HMM system mentioned above, since the value $p(\bar{x}_E^j|Z = z)/p(\bar{x}_E^j) = p(Z = z|\bar{x}_E^j)/p(Z = z)$ is proportional to the likelihood $p(Z = z|\bar{x}_E^j)$.

Which form of virtual evidence should we use, the *posterior* form of Equation 19 or the *likelihood* form of Equation 21? The one to use, depends on the application at hand. Examining the right side of Equation 19 and Equation 21, we see that the only difference is in the application of the prior probabilities $p(\bar{x}_E^j)$. These priors reflect some belief regarding the values of $x_E$ in universe two irrespective of the variable $Z$. If it is desirable to encode weights on $x_E$ in universe one only based on the local relationship between $Z$ and $X_E$ in universe two, then the likelihood approach Equation 21 should be used. If on the other hand we have obtained some external prior knowledge about $X_E$ that we wish to apply *in addition* to the local relation between $Z$ and $X_E$ in universe two, then the *posterior* form Equation 19 should be used. es We might even decide to encode and utilize weights on $x_E$ based only on the prior information, irrespective of any local external process, as in:

$$p(x_U)\delta(x_E, \{(\bar{x}_E^j, p(\bar{x}_E^j))\}_{j=1}^M) \quad (22)$$

This, however, is simply a restatement of the form given in Equation 1.

We see, however, that all three of the above forms are correct, since again all we are expressing in the universe are relative weights regarding different external beliefs about the variables $X_E$. And again it is only the ratios of these values that count in universe one.

## 6.2 Virtual Evidence and Backoff-based Language Models

Need to write.

## 6.3 Virtual Evidence and the IBM Machine Translation Models

Need to write.

## 7 Conclusion and Further Reading

Need to write.

## References

[1] S. M. Aji and R. J. McEliece. The generalized distributive law. *IEEE Transactions in Information Theory*, 46:325–343, March 2000.

[2] P. Billingsley. *Probability and Measure*. Wiley, 1995.

[3] C. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[4] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[5] R. Dechter. *Constraint Processing*. Morgan Kaufmann, 2003.

[6] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2001.

[7] F.V. Jensen. *An Introduction to Bayesian Networks*. Springer, 1996.

[8] F.V. Jensen. *Bayesian Networks and Decision Graphs*. Springer, 2001.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[10] S.L. Lauritzen. *Graphical Models*. Oxford Science Publications, 1996.

[11] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2nd printing edition, 1988.

[12] J. Pearl. Jeffrey's rule, passage of experience, and neo-bayesianism. In H. E. Kyburg, R. P. Loui, and G. N. Carlson, editors, *Knowledge Representation and Defeasible Reasoning*, pages 245–266. Kluwer, Boston, 1990.

[13] J. Pearl. *Causality*. Cambridge, 2000.

[14] S.D. Pietra, V.D. Pietra, and J. Lafferty. Inducing features of random fields. Technical Report CMU-CS-95-144, CMU, May 1995.

[15] Christian P. Robert. *The Bayesian Choice*. Springer, 2001.

[16] B. Taskar, C. Guestrin, and D. Koller. Max margin markov networks. In *Neural Information Processing Systems (NIPS)*, 16, Vancouver, Canada, December 2003.

[17] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[18] P.C. Woodland and D. Povey. Large scale discriminative training for speech recognition. In *ICSA ITRW ASR2000*, 2000.