How Heavy Should the Tails Be?

Changhe Yuan and Marek J. Druzdzel
Decision Systems Laboratory
Intelligent Systems Program and
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA 15260
{cyuan,marek}@sis.pitt.edu

Abstract

Importance sampling-based inference algorithms have shown excellent performance on reasoning tasks in Bayesian networks (Cheng & Druzdzel 2000; Moral & Salmeron 2003; Yuan & Druzdzel 2005). In this paper, we argue that all the improvements of these algorithms come from the same source, the improvement on the quality of the importance function. We also explain the requirements that a good importance function should satisfy, namely, it should concentrate its mass on the important parts of the target density and it should possess heavy tails. While the first requirement is subject of a theorem due to Rubinstein (1981), the second requirement is much less understood. We attempt to illustrate why heavy tails are desirable by studying the properties of importance sampling and examining a specific example. The study also leads to a theoretical insight into the desirability of heavy tails for importance sampling in the context of Bayesian networks, which provides a common theoretical basis for several successful heuristic methods.

Introduction

Importance sampling is used in several areas of modern statistics and econometrics to approximate unsolvable integrals. It has become the basis for several successful Monte Carlo sampling-based inference algorithms for Bayesian networks (Cheng & Druzdzel 2000; Moral & Salmeron 2003; Yuan & Druzdzel 2005), for which inference is known to be *NP-hard* (Cooper 1990; Dagum & Luby 1993). This paper argues that all the improvements of these algorithms come from the same source, the improvement on the quality of the importance function. A good importance function can lead importance sampling to yield excellent results in a reasonable time. It is well understood that we should focus on sampling in the areas where the value of the target function is relatively large (Rubinstein 1981). Thus, the importance function should concentrate its mass on the important parts of the target density. However, unimportant areas should by no means be neglected. Several authors pointed out that a good importance function should possess heavy tails (Geweke 1989; MacKay 1998). In other words, we should increase the sampling density in those unimportant areas. These two requirements seem to contradict one another. Why heavy tails are important and how heavy they should be are not well understood. This paper addresses these questions by studying the properties of importance sampling and discussing what conditions an admissible importance function should satisfy. We also try to illustrate why heavy tails are important by examining an example in which the conditions can be verified analytically. When analytical verification is impossible, we recommend to use two techniques to estimate how good an importance function is. The study also leads to a theoretical insight into the desirability of heavy tails for importance sampling in the context of Bayesian networks, which provides a common theoretical basis for several successful heuristic methods in Bayesian networks, including *ϵ-cutoff* (Cheng & Druzdzel 2000; Ortiz & Kaelbling 2000; Yuan & Druzdzel 2005), if-tempering (Yuan & Druzdzel 2004), rejection control (Liu 2001), and dynamic tuning (Shachter & Peot 1989; Ortiz & Kaelbling 2000; Moral & Salmeron 2003).

Importance Sampling

In our notation, a regular upper case letter, such as X, denotes a single variable, and x denotes its state. A bold upper case letter, such as $\mathbf{X}$, denotes a set of variables. Their states are denoted by $\mathbf{x}$. Now, let $p(\mathbf{X})$ be a probability density of n variables $\mathbf{X} = (X_1, ..., X_n)$ over domain $\Omega \subset R^n$. Consider the problem of estimating the multiple integral

$$E_{p(\mathbf{X})}[g(\mathbf{X})] = \int_{\Omega} g(\mathbf{X})p(\mathbf{X})d\mathbf{X} \ , \qquad (1)$$

where $g(\mathbf{X})$ is a function that is integrable with regard to $p(\mathbf{X})$ over domain Ω. Thus, $E_{p(\mathbf{X})}[g(\mathbf{X})]$ exists. When $p(\mathbf{X})$ is a density that is easy to sample from, we can solve the problem by, first, drawing a set of i.i.d. samples $\{\mathbf{x}_i\}$ from $p(\mathbf{X})$ and, second, using these samples to approximate the integral by means of the following expression

$$\hat{g}_N = \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i) \ . \qquad (2)$$

By the strong law of large numbers, the tractable sum $\overline{g}_n$ almost surely converges as follows

$$\hat{g}_N \rightarrow E_{p(\mathbf{X})}[g(\mathbf{X})] \ . \qquad (3)$$

Copyright © 2005, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

In case that $p(\mathbf{X})$ is a density that we do not know how to sample from but we can only evaluate it at any point, we need to resort to more sophisticated techniques. *Importance sampling* is a technique that provides a systematic approach that is practical for large dimensional problems. The main idea is simple. First, note that we can rewrite Equation 1 as

$$E_{p(\mathbf{X})}[g(\mathbf{X})] = \int_{\Omega} g(\mathbf{X}) \frac{p(\mathbf{X})}{I(\mathbf{X})} I(\mathbf{X}) d\mathbf{X} \qquad (4)$$

with any probability distribution $I(\mathbf{X})$, named *importance function*, as long as $I(\mathbf{X}) > 0$ when $p(\mathbf{X}) > 0$. Therefore, we can choose a density $I(\mathbf{X})$ that is easy to sample from. Let $\{\mathbf{x}_i\}$ be a sequence of i.i.d. random samples that is proportional to $I(\mathbf{X})$. Again, by the *strong law of large numbers*, we have

$$\hat{g}_N = \sum_{i=1}^{N} [g(\mathbf{x}_i) w(\mathbf{x}_i)] \rightarrow E_{p(\mathbf{X})}(g(\mathbf{X})) \ , \qquad (5)$$

where $w(\mathbf{x}_i) = \frac{p(\mathbf{x}_i)}{I(\mathbf{x}_i)}$, under the following weak assumptions (Geweke 1989):

Assumption 1 *$p(\mathbf{X})$ is proportional to a proper probability density function defined on Ω.*

Assumption 2 *$\{\mathbf{x}_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random samples, the common distribution having a probability density function $I(\mathbf{X})$.*

Assumption 3 *The support of $I(\mathbf{X})$ includes Ω.*

Assumption 4 *$E_{p(\mathbf{X})}(g(\mathbf{X}))$ exists and is finite.*

Obviously, importance sampling assigns more weight to regions where $p(\mathbf{X}) > I(\mathbf{X})$ and less weight to regions where $p(\mathbf{X}) < I(\mathbf{X})$ in order to estimate $E_{p(\mathbf{X})}(g(\mathbf{X}))$ correctly.

We do not have much control over what is required in Assumptions 1, 2, and 4, because they are either the inherent properties of the problem at hand or the characteristic of Monte Carlo simulation. We only have the freedom to choose which importance function to use, as long as it satisfies Assumption 3. The apparent reason why this assumption is necessary is avoiding undefined weights in the areas where $I(\mathbf{X}) = 0$ while $p(\mathbf{X}) > 0$. Since we draw samples from $I(\mathbf{X})$ when using Monte Carlo simulation, we bypass the trouble. However, the consequences of the bypass appear in the final result. Let $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$, where Ω_1 is the common support of $p(\mathbf{X})$ and $I(\mathbf{X})$, Ω_2 is the support only of $p(\mathbf{X})$, and Ω_3 is the support only of $I(\mathbf{X})$. When we use the estimator in Equation 5, we have

$$\begin{aligned} \hat{g}_N &= \sum_{i=1}^{N} [g(\mathbf{x}_i) w(\mathbf{x}_i)] \\ &= \sum_{\mathbf{x}_i \in \Omega_1} [g(\mathbf{x}_i) w(\mathbf{x}_i)] + \sum_{\mathbf{x}_i \in \Omega_2} [g(\mathbf{x}_i) w(\mathbf{x}_i)] \\ &\quad + \sum_{\mathbf{x}_i \in \Omega_3} [g(\mathbf{x}_i) w(\mathbf{x}_i)] \qquad (6) \end{aligned}$$

Since we draw samples from $I(\mathbf{X})$, all samples are either in Ω_1 or Ω_3, and no samples will drop in Ω_2. Therefore, the second term in Equation 6 is equal to 0. Also, all the samples in Ω_3 have zero weights, because $p(\mathbf{X})$ is equal to 0. Therefore, the third term is also equal to 0. Finally, we get

$$\begin{aligned} \hat{g}_N &= \sum_{\mathbf{x}_i \in \Omega_1} [g(\mathbf{x}_i) w(\mathbf{x}_i)] \\ &\rightarrow \int_{\Omega_1} g(\mathbf{X}) p(\mathbf{X}) d\mathbf{X} \ , \qquad (7) \end{aligned}$$

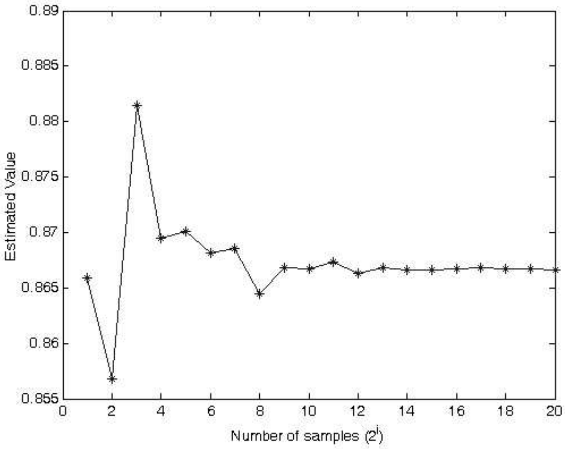

Figure 1: Using a truncated normal $I(\mathbf{X}) \propto N(0, 2.1^2)$, $|X| < 3$ as the importance function to integrate the density $p(\mathbf{X}) \propto N(0, 2^2)$. The result converges to 0.8664 instead of 1.0.

which is equal to the expectation of $g(\mathbf{X})$ with regard to $p(\mathbf{X})$ only in the domain of Ω_1. The conclusion is that the result will inevitably converge to a wrong value if we violate Assumption 3. Figure 1 shows an example of such erroneous convergence.

Optimal Importance Function

Standing alone, the assumptions in the previous section are of little practical value, because nothing can be said about rates of convergence. Even though we do satisfy the assumptions, $\hat{g}_N$ can behave badly. Poor behavior is usually manifested by values of $w(\mathbf{x}_i)$ that exhibit substantial fluctuations after thousands of replications (Geweke 1989). To quantify the convergence rate, it is enough to calculate the variance of the estimator in Equation 5, which is equal to

$$\begin{aligned} & Var_{I(\mathbf{X})}(g(\mathbf{X}) w(\mathbf{X})) \\ &= E_{I(\mathbf{X})}(g^2(\mathbf{X}) w^2(\mathbf{X})) - E^2_{I(\mathbf{X})}(g(\mathbf{X}) w(\mathbf{X})) \\ &= E_{I(\mathbf{X})}(g^2(\mathbf{X}) w^2(\mathbf{X})) - E^2_{p(\mathbf{X})}(g(\mathbf{X})) \ . \qquad (8) \end{aligned}$$

We certainly would like to choose the optimal importance function that minimizes the variance. The second term on the right hand side does not depend on $I(\mathbf{X})$ and, hence, we only need to minimize the first term. This can be done using Theorem 1.

Theorem 1 *(Rubinstein 1981) The minimum of* $Var_{I(\mathbf{X})}(g(\mathbf{X})w(\mathbf{X}))$ *is equal to*

$$Var_{I(\mathbf{X})}(g(\mathbf{X})w(\mathbf{X}))$$
$$= \left(\int_\Omega |g(\mathbf{X})|p(\mathbf{X})d\mathbf{X}\right)^2 - \left(\int_\Omega g(\mathbf{X})p(\mathbf{X})d\mathbf{X}\right)^2$$

and occurs when we choose the importance function

$$I(\mathbf{X}) = \frac{|g(\mathbf{X})|p(\mathbf{X})}{\int_\Omega |g(\mathbf{X})|p(\mathbf{X})d\mathbf{X}} \ .$$

The optimal importance function turns out to be useless, because it contains the integral $\int_\Omega |g(\mathbf{X})|p(\mathbf{X})d\mathbf{X}$, which is practically equivalent to the quantity $E_{p(\mathbf{X})}[g(\mathbf{X})]$ that we are pursuing. Therefore, it cannot be used as a guidance for choosing the importance function.

Why Heavy Tails?

The bottom line of choosing an importance function is that the variance in Equation 8 should exist. Otherwise, the result may oscillate rather than converge to the correct value. This can be characterized by the *Central Limit Theorem*.

Theorem 2 *(Geweke 1989) In addition to assumptions 1-4, suppose*

$$\mu \equiv \quad E_{I(\mathbf{X})}[g(\mathbf{X})w(\mathbf{X})] = \int_\Omega g(\mathbf{X})p(\mathbf{X})d\mathbf{X} \ .$$

and

$$\sigma^2 \equiv \quad Var_{I(\mathbf{X})}[g(\mathbf{X})w(\mathbf{X})] = \int_\Omega [\tfrac{g^2(\mathbf{X})p^2(\mathbf{X})}{I(\mathbf{X})}]d\mathbf{X} - 1 \ .$$

are finite. Then

$$n^{1/2}(\hat{g}_N - \mu) \Rightarrow N(0, \sigma^2) \ .$$

The conditions of Theorem 2 should be verified analytically if the result is to be used to assess the accuracy of $\hat{g}_N$ as an approximation of $E_{p(\mathbf{X})}[g(\mathbf{X})]$. We use a normal integration problem as an example. Consider the problem of calculating the integral $\int_\Omega p(X)dX$, where $p(X) \propto N(\mu_p, \sigma_p^2)$, using importance sampling. Let the importance function be $I(X) \propto N(\mu_I, \sigma_I^2)$. We know that

$$\mu \equiv \quad E_{I(X)}[w(X)] = \int_\Omega p(X)dX = 1 \ , \qquad (9)$$

which is obviously finite. We can also calculate the variance as

$$Var_{I(X)}(w(X))$$
$$= \int \frac{p^2(X)}{I(X)}dX - \left(\int p(X)dX\right)^2$$
$$= \frac{(\frac{\sigma_I}{\sigma_p})^2}{\sqrt{2(\frac{\sigma_I}{\sigma_p})^2 - 1}} exp(\frac{(\frac{\mu_I - \mu_p}{\sigma_p})^2}{2(\frac{\sigma_I}{\sigma_p})^2 - 1}) - 1 \ . \qquad (10)$$

The necessary condition for the variance in Equation 10 to exist is that $2(\frac{\sigma_I}{\sigma_p})^2 - 1 > 0$, which means that the variance of the importance function should be greater than one half of the variance of the target density. However, we not only want this quantity to exist, but also want to minimize it. Notice that the quantity $|\frac{\mu_I - \mu_p}{\sigma_p}|$ can be looked on as the standardized distance between μ_I and μ_p with regard to $p(X)$. From the table of the standard normal distribution function, we know that

$$\Phi(X) \cong 1, \ when \ X \geq 3.90 \ , \qquad (11)$$

where $\Phi(X)$ is the cumulative density function of the standard normal distribution. Therefore, when $|\frac{\mu_I - \mu_p}{\sigma_p}| \geq 3.90$, $I(X)$ must be far from close to $p(X)$ in terms of their shapes. For different values of $|\frac{\mu_I - \mu_p}{\sigma_p}|$, we plot the variance as a function of $\frac{\sigma_I}{\sigma_p}$ in Figure 2.

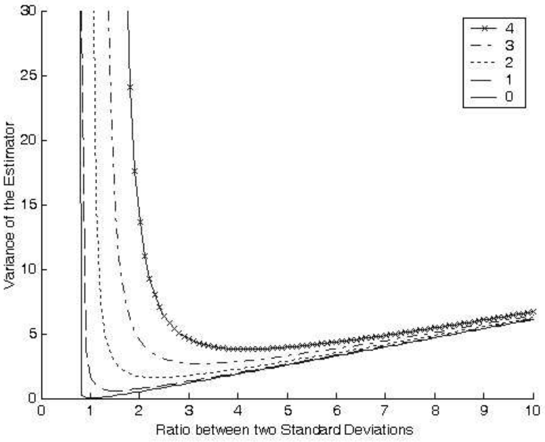

Figure 2: A plot of $\frac{\sigma_I}{\sigma_p}$ against the variance when using the importance function $I(X) \propto N(\mu_I, \sigma_I^2)$ with different μ_Is to integrate the density $p(X) \propto N(\mu_p, \sigma_p^2)$. The legend shows the values of $|\frac{\mu_I - \mu_p}{\sigma_p}|$.

We can make several observations based on this figure.

Observation 1 *Given the value of* $\frac{\sigma_I}{\sigma_p}$*, as* $|\frac{\mu_I - \mu_p}{\sigma_p}|$ *increases, the variance is monotonically increasing. This observation is consistent with the well understood requirement that* $I(X)$ *should concentrate its mass on the important parts of* $p(X)$*. The more* $I(X)$ *misses the important parts of* $p(X)$*, the worse the importance sampling performs.*

Observation 2 *As* $\frac{\sigma_I}{\sigma_p}$ *increases, the performances of* $I(X)$ *with different* μ_I*s differ less and less. Therefore, in case that we do not know* $|\frac{\mu_I - \mu_p}{\sigma_p}|$*, which means we are not sure if* $I(X)$ *covers the important parts of* $p(X)$ *or not,*[1] *we may want to make the tails of* $I(X)$ *heavier in order to be safe. The results may get worse, but not too much worse.*

Observation 3 *Given the value of* μ_I *and hence the value of* $|\frac{\mu_I - \mu_p}{\sigma_p}|$*, we can always achieve a minimum variance*

[1] We use the term cover to mean that the weight of one density is comparable to that of another density in a certain area.

when $\frac{\sigma_I}{\sigma_p}$ takes some value, say u. As $\frac{\sigma_I}{\sigma_p}$ decreases from u, the variance increases quickly and suddenly goes to infinity. When $\frac{\sigma_I}{\sigma_p}$ increases from u, the variance also increases but much slower.

Observation 4 *The u value increases as $|\frac{\mu_I-\mu_p}{\sigma_p}|$ increases, which means that the more $I(X)$ misses the important parts of $p(X)$, the heavier the tails of $I(X)$ should be.*

The four observations all provide strong support for heavy tails. In practice, we usually have no clue about the real shape of $p(X)$. Even if we have a way of estimating $p(X)$, our estimation will not be very precise. Therefore, we want to avoid light tails and err on the heavy tail side in order to be safe. One possible strategy is that we can start with an importance function $I(X)$ with considerably heavy tails and refine the tails as we gain more and more knowledge about $p(X)$.

It can be shown that similar results hold for several other distributions. Although to generalize from it is hard, we can at least get some idea why in practice we often observe that heavy tails are desirable. Furthermore, we will show later that importance sampling for Bayesian networks has the same results.

The conditions of Theorem 2 in general are not easy to verify analytically. Geweke (1989) suggests that $I(\mathbf{X})$ can be chosen such that either

$$\begin{aligned}&w(\mathbf{X}) < \overline{w} < \infty, \forall \mathbf{X} \in \Omega;\\ &\text{and } Var_{I(\mathbf{X})}[g(\mathbf{X})w(\mathbf{X})] < \infty \;;\end{aligned} \qquad (12)$$

or

$$\begin{aligned}&\Omega \text{ is compact, and}\\ &p(\mathbf{X}) < \overline{p} < \infty, I(\mathbf{X}) > \epsilon > 0, \forall \mathbf{X} \in \Omega \;.\end{aligned} \qquad (13)$$

Demonstration of Equation 13 is generally simple. Demonstration of Equation 12 involves comparison of the tail behaviors of $p(\mathbf{X})$ and $I(\mathbf{X})$. One way is to use the *variance of the normalized weights* to measure how different the importance function is from the target distribution (Liu 2001). If the target distribution $p(\mathbf{X})$ is known only up to a normalizing constant, which is the case in many real problems, the variance of the normalized weight can be estimated by the *coefficient of variation* of the unnormalized weight:

$$cv^2(w) = \frac{\sum_{j=1}^{m}(w(\mathbf{x}_j) - \overline{w})^2}{(m-1)\overline{w}^2} \;. \qquad (14)$$

where $\overline{w}$ is the sample average of all $w(\mathbf{x}_j)$.

Another way is to use *extreme value theory* to study the tail behavior. Smith (1987) shows that if we have an i.i.d. population $\{w(\mathbf{x}_i)\}$ then as a threshold value $u > 0$ increases, the limit distribution of the random variables over this threshold will be a *generalized Pareto* distribution with parameter ξ. The important characteristic of this distribution is that only $\frac{1}{\xi}$ moments exist. Therefore, if we want its variance to exist, ξ should be less than or equal to $1/2$. We can thus find the maximum likelihood estimator$\hat{\xi}$ and form a hypothesis testing problem for deciding between $\xi \leq 1/2$ and $\xi > 1/2$ (Koopman & Shephard 2002). In practice, we may not be satisfied with that $\xi \leq 1/2$, but want ξ to be as small as possible.

Importance Sampling for Bayesian Networks

Bayesian networks model explicitly probabilistic independence relations among sets of variables. In general, inference in Bayesian networks is NP-hard (Cooper 1990; Dagum & Luby 1993). Therefore, people often resort to approximate inference algorithms, such as importance sampling-based algorithms. In this section, we show that the same results as in the previous section hold for importance sampling in Bayesian networks.

Property of the Joint Probability Distribution

Let $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ be variables modelled in a Bayesian network. Let p be the probability of a state of the Bayesian network and p_i be the conditional (or prior) probability of the selected outcome of variable X_i, we have

$$p = p_1 p_2 \ldots p_n = \prod_{i=1}^{n} p_i \;. \qquad (15)$$

Druzdzel (1994) shows that p follows the lognormal distribution. Here, we review the main results. If we take the logarithm of both sides of Equation 15, we obtain

$$\ln p = \sum_{i=1}^{n} \ln p_i \;. \qquad (16)$$

Since each p_i is a random variable, $\ln p_i$ is also a random variable. By *Central Limit Theorem (Liapounov)*, the distribution of a sum of independent random variables approaches a normal distribution as the number of components of the sum approaches infinity under the condition that the sequence of variances is *divergent*. It can be easily shown that any single variance takes the value 0 only if the conditional distribution is uniform. However, in practical models, uniform distributions are uncommon. The *Liapounov condition* is obviously satisfied. Even though in practice we are dealing with a finite number of variables, the theorem gives a good approximation in such circumstances. The distribution of the sum in Equation 16 has the following form

$$f(\ln p) = \frac{1}{\sqrt{2\pi \sum_{i=1}^{n} \sigma_i^2}} \exp \frac{-(\ln p - \sum_{i=1}^{n} \mu_i)^2}{2\sum_{i=1}^{n} \sigma_i^2} \;. \qquad (17)$$

Although theoretically each probability in the joint probability distribution comes from a lognormal distribution with perhaps different parameters, Druzdzel (1994) points out that the conclusion is rather conservative and the distributions over probabilities of different states of a model might approach the same lognormal distribution in most practical models. The main reason is that conditional probabilities in practical models tend to belong to modal ranges, at most a few places after the decimal point, such as between 0.001 and 1.0. Translated into the decimal logarithmic scale, it means the interval between -3 and 0, which is further averaged over all probabilities, which have to add up to one, and for variables with few outcomes will result in even more modal ranges. Therefore, the parameters of the different lognormal distributions may be quite close to one another.

Importance Sampling for Bayesian networks

From now on, we make the assumption that all probabilities in the joint probability distribution of a Bayesian network come from the same lognormal distribution. Therefore, we can look at any importance sampling algorithm for Bayesian networks as using one lognormal distribution as the importance function to compute the expectation of another lognormal distribution. Let $p(\mathbf{X})$ be the original density of a Bayesian network and $p(\ln \mathbf{X}) \propto N(\mu_p, \sigma_p^2)$. Let $I(\mathbf{X})$ be the importance function and $I(\ln \mathbf{X}) \propto N(\mu_I, \sigma_I^2)$. Again, assume that we cannot sample from $p(\mathbf{X})$ but we can only evaluate it at any point. To compute the following multiple integral

$$V = \int_{\Omega} p(\mathbf{X}) d\mathbf{X} \ , \tag{18}$$

we can use the following estimator

$$\hat{V}_N = \sum_{i=1}^{N} w(\mathbf{x}_i) \ , \tag{19}$$

where $w(\mathbf{x}_i) = \frac{p(\mathbf{x}_i)}{I(\mathbf{x}_i)}$. The variance of this estimator is

$$Var_{I(\mathbf{X})}(w(\mathbf{X})) = E_{I(\mathbf{X})}(w(\mathbf{X})) - E^2_{I(\mathbf{X})}(w(\mathbf{X})) \ . \tag{20}$$

After plugging in the density functions of $p(\mathbf{X})$ and $I(\mathbf{X})$, we obtain

$$Var_{I(\mathbf{X})}(w(\mathbf{X})) = \frac{(\frac{\sigma_I}{\sigma_p})^2}{\sqrt{2(\frac{\sigma_I}{\sigma_p})^2 - 1}} e^{\frac{(\frac{\mu_I - \mu_p}{\sigma_p})^2}{2(\frac{\sigma_I}{\sigma_p})^2 - 1}} - 1 \ . \tag{21}$$

which has exactly the same form as in Equation 10. Therefore, we will have the same plots as in Figure 2 and the same observations for importance sampling in Bayesian networks. Therefore, we can conclude that heavy tails are also desirable for importance sampling in Bayesian networks.

Methods for Heavy Tails in Bayesian Networks

Given that heavy tails are desirable for importance sampling in Bayesian networks, we recommend the following strategy when designing an importance function. First, we need to make sure that the support of the importance function includes that of the target distribution. Since Ω is compact and $p(\mathbf{X})$ is finite for Bayesian networks, which satisfy the conditions of Equation 13, we only need to make sure that the $I(\mathbf{X}) > 0$ whenever $p(\mathbf{X}) > 0$. Second, we can make use of any estimation method to learn or compute an importance function. Many importance sampling-based inference algorithms using different methods to obtain importance functions have been proposed for Bayesian networks. Based on the methods, we classify the algorithms for Bayesian networks into three families. The first family uses the prior distribution of a Bayesian network as the importance function, including the *Probabilistic logic sampling* (Henrion 1988) and *likelihood weighting* (Fung & Chang 1989; Shachter & Peot 1989) algorithms. The second family resorts to learning methods to learn an importance function, including the *Self-importance sampling* (SIS) (Shachter & Peot 1989), *adaptive importance sampling* (Ortiz & Kaelbling 2000), AIS-BN (Cheng & Druzdzel 2000), and *dynamic importance sampling* (Moral & Salmeron 2003) algorithms. The third family directly computes an importance function in the light of both the prior distribution and the evidence, including the *backward sampling* (Fung & del Favero 1994), IS (Hernandez, Moral, & Salmeron 1998), *annealed importance sampling* (Neal 1998), and EPIS-BN algorithms.

The last step, based on the discussion in the previous section, is to get rid of light tails and make them heavier. There are several ways of doing this:

ϵ-cutoff (Cheng & Druzdzel 2000; Ortiz & Kaelbling 2000; Yuan & Druzdzel 2005): *ϵ-cutoff* defines the tails in Bayesian networks to be the states with extremely small or extremely large probabilities. So, it sets a threshold ϵ and replaces any smaller probability in the network by ϵ. At the same time, it compensates for this change by subtracting it from the largest probability in the same conditional probability distribution.

if-tempering (Yuan & Druzdzel 2004): Instead of just adjusting the importance function locally, if-tempering makes the original importance function $I(X)$ more flat by tempering $I(X)$. The final importance function becomes

$$I'(X) \propto I(X)^{1/T}, \tag{22}$$

where T ($T > 1$) is the tempering temperature.

Rejection control (Liu 2001): When the importance function is not ideal, one often produces random samples with very small weights when applying importance sampling. Rejection control adjust the importance function $I(\mathbf{X})$ in the following way. Suppose we have drawn samples $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ from $I(\mathbf{X})$. Let $w_j = f(\mathbf{x}_j)/I(\mathbf{x}_j)$. *Rejection control* (RC) conducts the following operation for any given threshold value $c > 0$:

1. For $j = 1, ..., n$, accept $\mathbf{x}_j$ with probability
$$r_j = min\{1, w_j/c\} \ . \tag{23}$$

2. If the jth sample $\mathbf{x}_j$ is accepted, its weight is updated to $w_{*j} = q_c w_j / r_j$, where
$$q_c = \int min\{1, w(\mathbf{X})/c\} I(\mathbf{X}) d\mathbf{X} \ . \tag{24}$$

The new importance function $I^*(\mathbf{X})$ resulting from this adjustment is expected to be closer to the target function $f(\mathbf{X})$. In fact, it is easily seen that

$$I^*(\mathbf{X}) = q_c^{-1} min\{I(\mathbf{X}), f(\mathbf{X})/c\} \ . \tag{25}$$

Dynamic tuning (Shachter & Peot 1989; Ortiz & Kaelbling 2000; Moral & Salmeron 2003): Dynamic tuning looks on the calculation of importance function itself as a self improving process. Starting from an initial importance function, dynamic tuning draws some samples from the current importance function and the use these samples to obtain a new importance function by refining the old one. The new importance function improves the old importance function at each stage. After several iterations, the final importance function is expected to be much closer to the optimal importance function.

Conclusion

The quality of importance function determines the performance of importance sampling. In addition to the requirement that the importance function should concentrate its mass on the important parts of the target density, it is also highly recommended that the importance function possess heavy tails. While the first requirement is subject of a theorem due to Rubinstein (1981), the second requirement is much less understood, because it seems to contradict the first requirement. By studying the assumptions and properties of importance sampling, we know a good importance function has to satisfy the conditions in Theorem 2. We examine a specific example which shows clearly why heavy tails are desirable for importance sampling. Although the example is in a restricted form, we later show that importance sampling for Bayesian networks has the same results. This insight provides a common theoretical basis for several successful heuristic methods.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research grant F49620–03–1–0187. All experimental data have been obtained using SMILE, a Bayesian inference engine developed at the Decision Systems Laboratory and available at `http://www.sis.pitt.edu/∼genie`.

References

Cheng, J., and Druzdzel, M. J. 2000. BN-AIS: An adaptive importance sampling algorithm for evidential reasoning in large Bayesian networks. *Journal of Artificial Intelligence Research* 13:155–188.

Cooper, G. F. 1990. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42(2–3):393–405.

Dagum, P., and Luby, M. 1993. Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence* 60(1):141–153.

Druzdzel, M. J. 1994. Some properties of joint probability distributions. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence (UAI–94)*, 187–194.

Fung, R., and Chang, K.-C. 1989. Weighing and integrating evidence for stochastic simulation in Bayesian networks. In Henrion, M.; Shachter, R.; Kanal, L.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 5*, 209–219. New York, N. Y.: Elsevier Science Publishing Company, Inc.

Fung, R., and del Favero, B. 1994. Backward simulation in Bayesian networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–94)*, 227–234. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Geweke, J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57(6):1317–1339.

Henrion, M. 1988. Propagating uncertainty in Bayesian networks by probalistic logic sampling. In *Uncertainty in Artificial Intelligence 2*, 149–163. New York, N.Y.: Elsevier Science Publishing Company, Inc.

Hernandez, L. D.; Moral, S.; and Salmeron, A. 1998. A Monte Carlo algorithm for probabilistic propagation in belief networks based on importance sampling and stratified simulation techniques. *International Journal of Approximate Reasoning* 18:53–91.

Koopman, S. J., and Shephard, N. 2002. Testing the assumptions behind the use of importance sampling. Technical report 2002-w17, Economics Group, Nuffield College, University of Oxford.

Liu, J. S. 2001. *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag.

MacKay, D. 1998. *Introduction to Monte Carlo methods*. Cambridge, Massachusetts: The MIT Press.

Moral, S., and Salmeron, A. 2003. Dynamic importance sampling computation in Bayesian networks. In *Proceedings of Seventh European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-03)*, 137–148.

Neal, R. M. 1998. Annealed importance sampling. Technical report no. 9805, Dept. of Statistics, University of Toronto.

Ortiz, L., and Kaelbling, L. 2000. Adaptive importance sampling for estimation in structured domains. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI–00)*, 446–454.

Rubinstein, R. Y. 1981. *Simulation and the Monte Carlo Method*. John Wiley & Sons.

Shachter, R. D., and Peot, M. A. 1989. Simulation approaches to general probabilistic inference on belief networks. In Henrion, M.; Shachter, R.; Kanal, L.; and Lemmer, J., eds., *Uncertainty in Artificial Intelligence 5*, 221–231. New York, N. Y.: Elsevier Science Publishing Company, Inc.

Smith, R. L. 1987. Estimating tails of probability distributions. *Annals of Statistics* 15:1174–1207.

Yuan, C., and Druzdzel, M. J. 2004. A comparison of the effectiveness of two heuristics for importance sampling. In *Proceedings of the Second European Workshop on Probabilistic Graphical Models (PGM'04)*, 225–232.

Yuan, C., and Druzdzel, M. J. 2005. Importance sampling algorithms for Bayesian networks: Principles and performance. *To appear in Mathematical and Computer Modelling*.