Theoretical Analysis and Practical Insights on Importance Sampling in Bayesian Networks

Changhe Yuan *

Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762

Marek J. Druzdzel

Decision Systems Laboratory, Intelligent Systems Program and School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260

Abstract

The AIS-BN algorithm [2] is a successful importance sampling-based algorithm for Bayesian networks that relies on two heuristic methods to obtain an initial importance function: ϵ-*cutoff*, replacing small probabilities in the conditional probability tables by a larger ϵ, and setting the probability distributions of the parents of evidence nodes to uniform. However, why the simple heuristics are so effective was not well understood. In this paper, we point out that it is due to a practical requirement for the importance function, which says that a good importance function should possess thicker tails than the actual posterior probability distribution. By studying the basic assumptions behind importance sampling and the properties of importance sampling in Bayesian networks, we develop several theoretical insights into the desirability of thick tails for importance functions. These insights not only shed light on the success of the two heuristics of AIS-BN, but also provide a common theoretical basis for several other successful heuristic methods.

1 Introduction

Importance sampling is a popular alternative to numerical integration when the latter is hard, and it has become the basis for many importance sampling-based algorithms for Bayesian networks [2,11,16,18,24,25], for which inference

* Corresponding author
Email addresses: cyuan@cse.msstate.edu (Changhe Yuan), marek@sis.pitt.edu (Marek J. Druzdzel).

is known to be *NP-hard* [3,4]. Essentially, these algorithms differ only in the methods that they use to obtain *importance functions*, i.e., sampling distributions. The closer the importance function to the actual posterior distribution, the better the performance. A good importance function can lead importance sampling to yield good convergence results in an acceptable amount of time [15]. It is well understood that we should focus on sampling in the areas where the value of the posterior distribution is relatively large [1,21], and, hence, the importance function should concentrate its mass on the important parts of the posterior distribution. However, unimportant areas should by no means be neglected. Several researchers pointed out that a good importance function should possess thicker tails than the actual posterior distribution [8,15]. However, the desirability of thick tails is much less understood.

This paper tries to address the limitation and develop a better understanding for the importance of thick tails. First, we explain the basic assumptions behind importance sampling and their importance. We then study the properties of importance sampling and discuss what conditions an importance function should satisfy. After that, we specifically study the properties of importance sampling in the context of Bayesian networks, which leads to several theoretical insights into the desirability of thick tails. The insights not only shed light to the success of the AIS-BN algorithm [1], which relies on two heuristic methods to obtain an initial importance function: *ϵ-cutoff*, replacing small probabilities in the conditional probability tables by a larger ϵ, and setting the probability distributions of the parents of evidence nodes to uniform, but also provide a common theoretical basis for several other successful heuristic methods.

The remainder of this paper is organized as follows. The first two section are introductory in nature and mainly review the background material. In Section 2, we introduce the basic theory of importance sampling and the underlying assumptions. We also present the form of the optimal importance function. In Section 3, we discuss what conditions an importance function should satisfy. We also recommend a technique for estimating how well an importance function performs when analytical verification of the conditions is impossible. Section 4 is the core of our paper. We study the properties of importance sampling in the context of Bayesian networks and present our theoretical insights into the desirability of thick tails. We also review several successful heuristics that are unified by the insights.

[1] The authors of the AIS-BN algorithm, Cheng and Druzdzel, received honorable mention in the 2005 IJCAI-JAIR Best Paper Award Awarded to an outstanding paper published in JAIR in the preceding five calendar years. For 2005, papers published between 2000 and 2004 were eligible.

2 Importance Sampling

We start with the theoretical roots of importance sampling. We use uppercase letters for variables and lowercase letters for the states of the variables. We use boldface letters for sets of variables or states. Let $p(X)$ be a probability density of n variables X over domain $\Omega \subset R$, where R is the set of real numbers. Consider the problem of estimating the integral

$$E_{p(X)}[g(X)] = \int_{\Omega} g(X)p(X)dX \ , \tag{1}$$

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

$$\tilde{g}_N = \frac{1}{N}\sum_{i=1}^{N} g(x_i) \ . \tag{2}$$

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

$$\tilde{g}_N \rightarrow E_{p(X)}[g(X)] \ . \tag{3}$$

In case we do not know how to sample from $p(X)$ but can evaluate it at any point up to a constant, or we simply want to reduce the variance of the estimator, we can resort to more sophisticated techniques. *Importance sampling* is a technique that provides a systematic approach that is practical for large dimensional problems. Its main idea is simple. First, note that we can rewrite Equation 1 as

$$E_{p(X)}[g(X)] = \int_{\Omega} g(X)\frac{p(X)}{I(X)}I(X)dX \tag{4}$$

with any probability distribution $I(X)$, named *importance function*, such that $I(X) > 0$ across the entire domain Ω. A practical requirement of $I(X)$ is that it should be easy to sample from. In order to estimate the integral, we can generate samples $x_1, x_2, ..., x_N$ from $I(X)$ and use the following sample-mean formula

$$\hat{g}_N = \sum_{i=1}^{N}[g(x_i)w(x_i)] \ , \tag{5}$$

where the weights $w(x_i) = \frac{p(x_i)}{I(x_i)}$. Obviously, importance sampling assigns more weight to regions where $p(X) > I(X)$ and less weight to regions where $p(X) < I(X)$ in order to estimate $E_{p(X)}(g(X))$ correctly. Again, $\hat{g}_N$ almost surely converges to $E_{p(X)}[g(X)]$.

To summarize, the following weak assumptions are important for the importance sampling estimator in Equation 5 to converge to the correct value [8]:

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

Assumption 2 *$E_{p(X)}(g(X))$ exists and is finite.*

Assumption 3 *$\{x_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random samples, the common distribution having a probability density function $I(X)$.*

Assumption 4 *The support of $I(X)$ includes Ω.*

We do not have much control over what is required in Assumptions 1, 2, and 3, because they are either the inherent properties of the problem at hand or the requirements of Monte Carlo simulation. We only have the freedom to choose an importance function satisfying Assumption 4. The apparent reason why the last assumption is important is to avoid undefined weights in the areas where $I(X) = 0$ while $p(X) > 0$, but such samples will never show up in importance sampling, because we are drawing samples from $I(X)$. Thus, the problem is bypassed. However, the aftermath of the bypass is manifested in the final result. Let Ω^* be the support of $I(X)$. When we use the estimator in Equation 5, we have

$$\hat{g}_N = \sum_{i=1}^{N}[g(x_i)w(x_i)] = \sum_{x_i \in \Omega^* \cap \Omega}[g(x_i)w(x_i)] + \sum_{x_i \in \Omega^* \setminus \Omega}[g(x_i)w(x_i)] \ , \tag{6}$$

where $\setminus$ denotes set subtraction. Since we draw samples from $I(X)$, all samples are in either $\Omega^* \cap \Omega$ or $\Omega^* \setminus \Omega$, and no samples will drop in $\Omega \setminus \Omega^*$. Also, all the samples in $\Omega^* \setminus \Omega$ have zero weights, because $p(X)$ is equal to 0 in this area. Therefore, the second term in Equation 6 is equal to 0. Effectively, we have

$$\hat{g}_N = \sum_{x_i \in \Omega^* \cap \Omega}[g(x_i)w(x_i)] \rightarrow \int_{\Omega^* \cap \Omega} g(X)p(X)dX \ , \tag{7}$$

which is equal to the expectation of $g(X)$ with regard to $p(X)$ only in the domain of $\Omega^* \cap \Omega$. So, the conclusion is that the estimator will converge to a wrong value if Assumption 4 is violated. Figure 1 shows an example of such erroneous convergence. In the example, we integrated the normal distribution

$p(X) \propto N(0, 2^2)$ using importance sampling. Clearly the exact answer should be 1.0. However, we used a truncated normal, $I(X) \propto N(0, 2.1^2)$, $|X| < 3$, as the importance function and converged to 0.8664 instead. The reason is that the support of the importance function did not cover the original density.

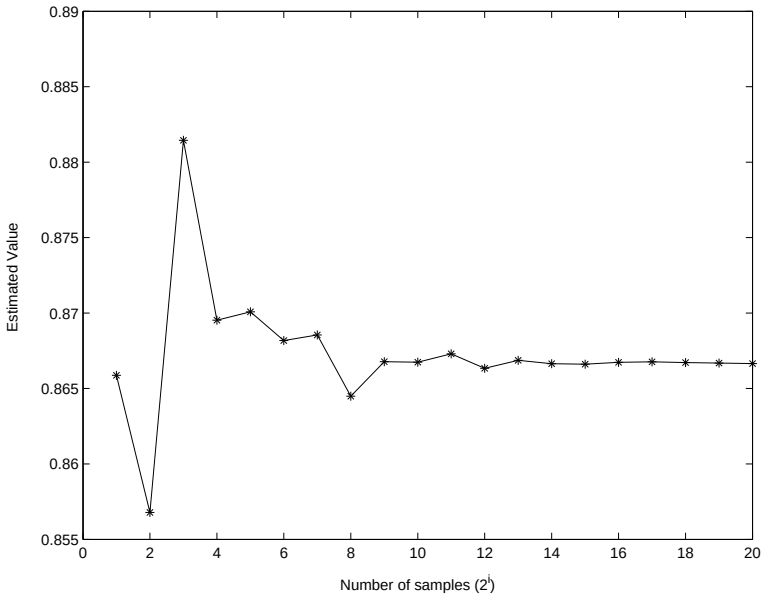

Fig. 1. Convergence results when using a truncated normal, $I(X) \propto N(0, 2.1^2)$, $|X| < 3$, as the importance function to integrate the density $p(X) \propto N(0, 2^2)$. The estimator converges to 0.8664 instead of 1.0.

Standing alone, the assumptions aforementioned 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 $w(x_i)$ that exhibit substantial fluctuations after thousands of replications. To quantify the convergence rate, it is enough to calculate the variance of the estimator in Equation 5, which is equal to

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

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 $I(X)$ and, hence, we only need to minimize the first term. This can be done according to Theorem 1.

Theorem 1 *[21] The minimum of* $Var_{I(X)}(g(X)w(X))$ *over all* $I(X)$ *is equal to*

$$\left(\int_{\Omega} |g(X)|p(X)dX\right)^2 - \left(\int_{\Omega} g(X)p(X)dX\right)^2$$

and occurs when we choose the importance function

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

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

3 Convergence Assessment of Importance Sampling

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 *[8] In addition to assumptions 1-4, suppose*

$$\mu \equiv E_{I(X)}\left[g(X)w(X)\right] = \int_{\Omega} g(X)p(X)dX \ ,$$

and

$$\sigma^2 \equiv Var_{I(X)}[g(X)w(X)] = \int_{\Omega} \left[\frac{g^2(X)p^2(X)}{I(X)}\right] dX - \mu^2 \ .$$

are finite. Then

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

The conditions of Theorem 2 should be satisfied if the result is to be used to assess the accuracy of $\hat{g}_N$ as an approximation of $E_{p(X)}[g(X)]$. However, the conditions in general are not easy to verify analytically in real problems. Geweke [8] suggests that $I(X)$ can be chosen such that either

$$w(X) < w^- < \infty, \forall X \in \Omega, \text{and } Var_{I(X)}[g(X)w(X)] < \infty \ ; \tag{9}$$

or

$$\Omega \text{ is compact, and} p(X) < \overline{p} < \infty, I(X) > \epsilon > 0, \forall X \in \Omega \ . \tag{10}$$

Demonstration of Equation 10 is generally simple. Demonstration of Equation 9 involves comparison of the tail behaviors of $p(X)$ and $I(X)$. One approach is to use the *variance of the normalized weights* to measure how different the importance function is from the posterior distribution [14]. If the distribution $p(X)$ 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* (cv) of the unnormalized weight:

$$cv^2(w) = \frac{\sum_{j=1}^{m} (w(x_j) - \overline{w})^2}{(m-1)\overline{w}^2} \ , \tag{11}$$

where $w(x_j)$ is the weight of sample x_j, $\overline{w}$ is the average weight of all samples, and m is the number of samples.

4 Importance Sampling for Bayesian Networks

Bayesian networks offer a concise and intuitive graphical representation of probabilistic conditional independence relations among the variables in a domain and have proven their value in many disciplines over the last two decades. However, it has been shown that inference in Bayesian networks in general is NP-hard [3,4]. For extremely large or complex models, exact inference is not feasible, and we have to resort to approximate methods. Importance sampling can be easily adapted to solve various inference problems in Bayesian networks, and has become the basis of an important family of approximate methods for Bayesian networks [2,11,16,18,24,25]. In this section, we study the properties of importance sampling in the context of Bayesian networks. The study leads to several theoretical insights into the desirability of thick tails. We also review several successful heuristic methods that are unified by the insights.

4.1 Property of the Joint Probability Distribution

Let $\mathbf{X} = \{X_1, X_2, ..., X_n\}$ be the variables modelled in a Bayesian network. Let us pick an arbitrary scenario of the network, and let p be the probability

of the scenario. Let p_i be the conditional (or prior) probability of the selected outcome of variable X_i, i.e., $p_i = P(X_i|\mathrm{PA}(X_i))$ or $P(X_i)$ if X_i has no parents. We have

$$p = p_1 p_2 \ldots p_n = \prod_{i=1}^{n} p_i \ . \tag{12}$$

Druzdzel [5] shows that p approximately follows the lognormal distribution. Here, we review the main results. Taking the logarithm of both sides of Equation 12, we obtain

$$\ln p = \sum_{i=1}^{n} \ln p_i \ . \tag{13}$$

Since each p_i is randomly picked from the prior or conditional probability distribution of the variable, it can be viewed as a random variable. Therefore $\ln p_i$ 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 sum of the sequence of variances is *divergent*. The variance of $\ln p_i$ is 0 only and only if all values of p_i are the same, i.e., X_i follows a uniform distribution given $\mathrm{PA}(X_i)$. However, in practical models, uniform distributions are uncommon, and, if so, the *Liapounov condition* is satisfied. Even though in practice we are dealing with a finite number of variables, the theorem often gives us a good approximation. In fact, the sum of as few as 12 uniformly distributed random variables is for all practical purpose distributed normally [19]. Therefore, the distribution of the sum in Equation 13 is approximately the following form

$$f(\ln p) = \frac{1}{\sqrt{2\pi \sum_{i=1}^{n} \sigma_i^2}} \exp \frac{-(\ln p - \sum_{i=1}^{n} \mu_i)^2}{2 \sum_{i=1}^{n} \sigma_i^2} \ . \tag{14}$$

Although theoretically each probability in the joint probability distribution comes from a lognormal distribution with perhaps different parameters, Druzdzel 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 [5]. 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. For our incoming analysis, we make the assumption that all probabilities in the joint probability distribution of a Bayesian network come from the same lognormal distribution.

4.2 Desirability of Thick Tails

Based on the preceding discussion, 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 $p(X)$ be the target density and $p(\ln X) \propto N(\mu_p, \sigma_p^2)$. Let $I(X)$ be the importance function and $I(\ln X) \propto N(\mu_I, \sigma_I^2)$. Consider the problem of computing the following integral

$$V = \int_{\Omega} p(X)dX \ . \tag{15}$$

We can use the following estimator

$$\hat{V}_N = \sum_{i=1}^{N} w(x_i) \ , \tag{16}$$

where $w(x_i) = \frac{p(x_i)}{I(x_i)}$. We know that

$$\mu \equiv E_{I(X)}[w(X)] = \int_{\Omega} p(X)dX = 1 \ , \tag{17}$$

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

$$Var_{I(X)}(w(X)) = E_{I(X)}(w^2(X)) - E^2_{I(X)}(w(X)) \ . \tag{18}$$

Plug in the density functions of $p(X)$ and $I(X)$, we obtain

$$\begin{aligned}
&Var_{I(X)}(w(X)) \\
&= \int \frac{p^2(X)}{I(X)}dX - \left(\int p(X)dX\right)^2 \\
&= -1 + \int \frac{\sigma_I}{\sigma_p^2 X\sqrt{2\pi}} \\
&\quad \exp\left(-\frac{(2\sigma_I^2 - \sigma_p^2)\ln^2 X - 2(2\mu_p\sigma_I^2 - \mu_I\sigma_p^2)\ln X + (2\mu_p^2\sigma_I^2 - \mu_I^2\sigma_p^2)}{2\sigma_p^2\sigma_I^2}\right) dX
\end{aligned}$$

$$= -1 + \frac{(\frac{\sigma_I}{\sigma_p})^2}{\sqrt{2(\frac{\sigma_I}{\sigma_p})^2 - 1}} \exp\left(\frac{(\frac{\mu_I - \mu_p}{\sigma_p})^2}{2(\frac{\sigma_I}{\sigma_p})^2 - 1}\right)$$
$$\int \frac{1}{\sqrt{\frac{\sigma_p^2\sigma_I^2}{2\sigma_I^2 - \sigma_p^2}} X\sqrt{2\pi}} \exp\left(-\frac{\ln X - \frac{2\mu_p\sigma_I^2 - \mu_I\sigma_p^2}{2\sigma_I^2 - \sigma_p^2}}{\frac{2\sigma_p^2\sigma_I^2}{2\sigma_I^2 - \sigma_p^2}}\right)^2 dX$$
$$= \frac{(\frac{\sigma_I}{\sigma_p})^2}{\sqrt{2(\frac{\sigma_I}{\sigma_p})^2 - 1}} \exp\left(\frac{(\frac{\mu_I - \mu_p}{\sigma_p})^2}{2(\frac{\sigma_I}{\sigma_p})^2 - 1}\right) - 1 \ . \tag{19}$$

One immediate observation from the above equation is that:

Observation 1 *The necessary condition for the variance in Equation 19 to exist is that* $2(\frac{\sigma_I}{\sigma_p})^2 - 1 > 0$*, which means that the variance of the importance function should be at least greater than one half of the variance of the target density.*

$\frac{\sigma_I}{\sigma_p}$ can be looked on as an indicator of thick tails. The bigger the $\frac{\sigma_I}{\sigma_p}$, the thicker the tails of the importance function $I(X)$ than those of $p(X)$. The quantity $|\frac{\mu_I - \mu_p}{\sigma_p}|$ is the standardized distance between μ_I and μ_p with regard to $p(X)$. It can be looked on as an indicator whether two functions have similar shapes or not. From the table of the standard normal distribution function, we know that

$$\Phi(X) \cong 1, \ when \ X \geq 3.90 \ , \tag{20}$$

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

We can make several additional observations based on this figure.

Observation 2 *Given the value of* $\frac{\sigma_I}{\sigma_p}$*, the variance is monotonically increasing as* $|\frac{\mu_I - \mu_p}{\sigma_p}|$ *increases.*

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 importance sampling performs.

Observation 3 *Given the value of* μ_I *and hence the value of* $|\frac{\mu_I - \mu_p}{\sigma_p}|$*, there is a minimum variance when* $\frac{\sigma_I}{\sigma_p}$ *takes a particular value, say* u*. As* $\frac{\sigma_I}{\sigma_p}$ *decreases*

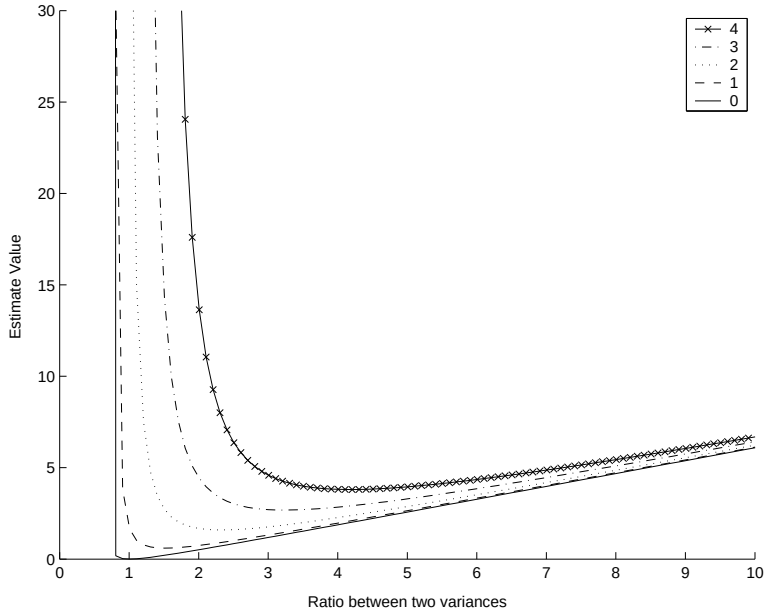

Fig. 2. A plot of the variance of importance sampling estimator as a function of $\frac{\sigma_I}{\sigma_p}$ when using the importance function $I(\ln X) \propto N(\mu_I, \sigma_I^2)$ with different μ_Is to integrate the density $p(\ln X) \propto N(\mu_p, \sigma_p^2)$. The legend shows the values of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

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 *As* $\frac{\sigma_I}{\sigma_p}$ *increases, the performance of* $I(X)$ *with different* μ_I*s differ less and less.*

The above two observations clearly tell us that if we do not know $|\frac{\mu_I - \mu_p}{\sigma_p}|$, i.e., we are not sure if $I(X)$ covers the important parts of $p(X)$ or not,[2] we may want to make the tails of $I(X)$ thicker in order to be safe. You can see that the variance of the estimator becomes larger, but not much larger.

Observation 5 *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 thicker the tails of* $I(X)$ *should be.*

The five observations all provide strong support for thick 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 may not be that precise. Therefore, we want to avoid light tails and err on the thick tail side in order to be safe. One possible strategy is that we can start with an importance function $I(X)$ with considerably thick tails and refine the tails as we gain more and more knowledge about $p(X)$.

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

It can be shown that the above results hold not only for Bayesian networks but also for several well-known distributions, including normal distribution. Although generalizing the results is hard, we can at least get some idea why in practice we often observe that thick tails are desirable.

Furthermore, the theoretical result that the actual posterior distribution is the optimal importance function is derived based on an infinite number of samples. In practice, we can only afford a finite number of samples. In order that the samples effectively cover the whole support of the posterior distribution, we often need to make the importance function possess thicker tails than the posterior distribution. Suppose the mass of the tail area of the posterior distribution is ϵ and we draw totally N samples. In order that the samples cover this area, we need at least one sample dropping in it, the probability of which is

$$p = 1 - (1 - \epsilon)^N \ . \tag{21}$$

In the case that $N\epsilon << 1$, we have

$$p \approx N\epsilon \ . \tag{22}$$

However, since $N\epsilon$ is very small, it is unlikely that any sample will drop in the tail area of $p(X)$. Given the importance of Assumption 4 discussed in Section 3, we may deviate from the correct answer. For the probability to be greater than some value u, we have

$$N > u/\epsilon \ . \tag{23}$$

If we cannot afford the needed number of samples, we can instead increase the sampling density of the importance function in the tail area so that

$$\epsilon > u/N \ . \tag{24}$$

This is exactly why in practice importance functions with thicker tails than the actual posterior distribution often perform better than the latter.

4.3 Methods for Thick Tails

Given that thick 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 posterior distribution. Since Ω is compact and $p(\mathbf{X})$ is finite for Bayesian networks, which satisfy the conditions of Equation 10, we only need to make sure that $I(\mathbf{X}) > 0$ 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 algorithms have been proposed for Bayesian networks. Based on the nature of the methods that the algorithms use to obtain the importance functions, we classify the algorithms into three families. The first family uses the prior distribution of a Bayesian network as the importance function, including the *probabilistic logic sampling* [10] and *likelihood weighting* [6,22] algorithms. The second family resorts to learning methods to learn an importance function, including the *self-importance sampling* (SIS) [22], *adaptive* IS [18], AIS-BN [2], and *dynamic* IS [16] algorithms. The third family directly computes an importance function in the light of both the prior distribution and the evidence, including the *backward sampling* [7], IS [11], *annealed importance sampling* [17], and EPIS-BN algorithms [25]. Although much work has been done in this direction, potential for further work is still huge. Note that the calculation of an importance function is essentially an approximate inference problem for Bayesian networks. There are many deterministic approximate inference algorithms that lack the guarantee to converge to the correct answers but can provide a satisfactory lower/upper bound efficiently, such as *variational methods* [12]. Nothing prevents us from using these methods to quickly get an estimation of the posterior distribution, which may be able to serve as a good importance function.

The last step, based on the discussion in the previous section, is to diagnose light tails and try to get rid of them to achieve thick tails. I review several existing heuristic methods for this purpose:

ϵ-cutoff [2,18] defines the tails of the joint probability distribution of a Bayesian network as the states with extremely small or extremely large probabilities. Therefore, it sets a threshold ϵ and replaces any smaller probability in the conditional probability tables in the network by ϵ. At the same time, it compensates for this change by subtracting the difference from the largest probability in the same conditional probability distribution. The purpose is to spread the mass of the joint probability distribution in order to make it more flat. The other heuristic in AIS-BN—setting the probability distributions of the parents of evidence nodes to uniform—also has the similar effect.

IF-tempering [24]: 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

$$I'(X) \propto I(X)^{1/T}, \tag{25}$$

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

Rejection control [14]: When the importance function is not ideal, importance sampling often produces random samples with very small weights. Rejection control adjusts the importance function $I(X)$ in the following way. Suppose we have drawn samples $x_1, x_2, ..., x_N$ from $I(X)$. Let $w_j = p(x_j)/I(x_j)$. Rejection control (RC) conducts the following operation for any given threshold value $c > 0$:

(1) For $j = 1, ..., n$, accept x_j with probability

$$r_j = min\{1, w_j/c\} \ . \tag{26}$$

(2) If the jth sample x_j is accepted, its weight is updated to $w_{*j} = q_c w_j/r_j$, where

$$q_c = \int min\{1, w(X)/c\} I(X) dX \ . \tag{27}$$

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

$$I^*(X) = q_c^{-1} min\{I(X), p(X)/c\} \ . \tag{28}$$

Pruned Enriched Rosenbluth Method (PERM) [9,13,20]: PERM is also a sample population-based method, similar to rejection control. Rejection control is based on the observation that samples with extremely small weights do not play much role in the final estimation, but make the variance of sample weights large. However, there is yet another source of problem: samples with extremely large weights often overwhelmingly dominate the estimator and make other samples less effective. To eschew both problems, PERM assumes that the sample weights are built up in many steps and long range correlations between these steps are often weak. Given the assumption, PERM adjusts the samples for given threshold values $0 < c_- < c^- < \infty$ using the following strategy in each step.

For $j = 1, ..., n$,

(1) If $c_- < w_j < c^-$, accept the sample $\mathbf{x}_j$ and keep its weight intact.
(2) If $w_j < c_-$, accept $\mathbf{x}_j$ with probability 0.5. If the jth sample $\mathbf{x}_j$ is accepted, its weight is updated to $w_{*j} = 2 * w_j$.
(3) If $w_j > c^-$, we split the sample into two samples, each with weight $w_{*j} = w_j/2$.

Effectively, PERM adjusts the importance function so that the new impor-

tance function $I^*(X)$ follows

$$I^*(\mathbf{X}) = q_p^{-1} \begin{cases} 2I(\mathbf{X}), & \Omega_1\text{: } p(\mathbf{X}) > c^- I(\mathbf{X}); \\ I(\mathbf{X}), & \Omega_2\text{: } c_- < p(\mathbf{X})/I(\mathbf{X}) < c^-; \\ I(\mathbf{X})/2, & \Omega_3\text{: } p(\mathbf{X}) < c_- I(\mathbf{X}), \end{cases}$$

where

$$q_p = 2 \int_{\Omega_1} p(\mathbf{X}) d\mathbf{X} + \int_{\Omega_2} p(\mathbf{X}) d\mathbf{X} + (1/2) \int_{\Omega_3} p(\mathbf{X}) d\mathbf{X} \ . \tag{29}$$

Intentionally biased dynamic tuning [2,18]: Dynamic tuning looks on the calculation of importance function itself as a self-improving process. Starting from an initial importance function, dynamic tuning draws samples from the current importance function and then use the samples to refine the importance function in order to obtain a new function. The new importance function improves the old one at each stage. Dynamic tuning has been applied in several learning-based importance sampling algorithms. However, only two of them observe the importance of thick tails [2,18] and apply ϵ-cutoff to try to ensure that property in order to get better convergence rates.

5 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 posterior probability distribution, it is also highly recommended that the importance function possess thicker tails than the posterior distribution.

In this paper, we provide a better understanding of why thick tails are desirable in the context of Bayesian networks. Our conclusion is somewhat different than the common belief — thick tails are not necessary or better but they are simply safer than light tails. By studying the basic assumptions of importance sampling and its properties in the context of Bayesian networks, we draw several theoretical insights into the desirability of thick tails. The insights not only shed light to the success of the AIS-BN algorithm, which relies on two heuristic methods to obtain an initial importance function: *ϵ-cutoff*, replacing small probabilities in the conditional probability tables by a larger ϵ, and setting the probability distributions of the parents of evidence nodes to uniform, but also provide a common theoretical basis for several other successful heuristic methods.

There is also a lot of future work. Most existing heuristics for thick tails are local methods, i.e., they adjust the importance function locally. We believe that heuristics that are aware of the global structure of an importance function and make global adjustments may bring better performance. Also, it would be interesting to compare the effectiveness of the different heuristic methods across different networks empirically.

6 Acknowledgements

This research was supported by the Air Force Office of Scientific Research grants F49620–03–1-0187 and FA9550–06–1–0243. The initial version of this paper [23] appeared in the proceedings of the 18th International Artificial Intelligence Research Society Conference (FLAIRS-05). We thank several anonymous reviewers for the FLAIRS-05 conference and the International Journal of Approximate Reasoning for several insightful comments that led to improvements in the paper.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 350:5–43, 2003.

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

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

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

[5] M. J. Druzdzel. Some properties of joint probability distributions. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence (UAI–94)*, pages 187–194, Morgan Kaufmann Publishers San Francisco, California, 1994.

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

[7] R. Fung and B. del Favero. Backward simulation in Bayesian networks. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial*

Intelligence (UAI–94), pages 227–234, San Mateo, CA, 1994. Morgan Kaufmann Publishers, Inc.

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

[9] P. Grassberger. Pruned-enriched Rosenbluth method: Simulations of θ polymers of chain length up to 1 000 000. *Physical Review E*, 56:3682–3693, Sept. 1997.

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

[11] L. D. Hernandez, S. Moral, and A. Salmeron. 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, 1998.

[12] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. *An introduction to variational methods for graphical models.* The MIT Press, Cambridge, Massachusetts, 1998.

[13] F. Liang. Dynamically weighted importance sampling in monte carlo computation. *Journal of the American Statistical Association*, 97:807–821, 2002.

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

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

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

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

[18] L. Ortiz and L. Kaelbling. Adaptive importance sampling for estimation in structured domains. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence (UAI–00)*, pages 446–454, Morgan Kaufmann Publishers San Francisco, California, 2000.

[19] W. H. Press. *Numerical Recipes in C.* Cambridge University Press, 1992.

[20] M. N. Rosenbluth and A. W. Rosenbluth. Monte Carlo calculation of the average extension of molecular chains. *Journal of Chemical Physics*, 23(256), 1955.

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

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

[23] C. Yuan and M. Druzdzel. How heavy should the tails be? In I. Russell and Z. Markov, editors, *Proceedings of the Eighteenth International FLAIRS Conference (FLAIRS-05)*, pages 799–804, AAAI Press/The MIT Press, Menlo Park, CA, 2005.

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

[25] C. Yuan and M. J. Druzdzel. Importance sampling algorithms for Bayesian networks: Principles and performance. *Mathematical and Computer Modelling*, 43:1189–1207, 2006.