A Blessing of Dimensionality: Measure Concentration and Probabilistic Inference

Pınar Muyan
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Nando de Freitas
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada

Abstract

This paper proposes an efficient sampling method for inference in probabilistic graphical models. The method exploits a blessing of dimensionality known as the concentration of measure phenomenon in order to derive analytic expressions for proposal distributions. The method can also be interpreted in a variational setting, were one minimises an upper-bound on the estimator variance. The results on simple settings are very promising. We believe this method has great potential in graphical models used for diagnosis.

1 INTRODUCTION

Machine learning is plagued with curses of dimensionality, but there are less-known blessings too. The incorporation of these blessings in the design of learning algorithms is a research direction of profound importance (Donoho 2000). In this paper, we exploit one of these blessings, namely the concentration of measure phenomenon, to derive efficient sampling algorithms for inference in probabilistic graphical models.

Various large deviation results state that probability measures often concentrate on small sets, specially in high-dimensions. (In information theory this is know as the asymptotic equipartition property (Cover and Thomas 1991).) These results can be exploited to design optimal sampling algorithms. In the importance sampling scenario, the large deviation theorems allow one to surmount the crucial problem of coming up with good proposal distributions. In particular, Cramér's theorem enables us to derive importance proposal distributions analytically for simple distributions.

Large deviation results for importance sampling appeared initially in the simulation and communications literature (Sadowsky and Bucklew 1990). These results, when applicable, have resulted in powerful samplers (Smith, Shafi and Gao 1997). However, they have been restricted to simple simulation scenarios. In this paper, we present these results to a broader community, while extending them to carry out Bayesian inference in more complex multivariate probabilistic models.

Our method has an elegant variational interpretation. One obtains the proposal distribution for importance sampling (for some restricted scenarios) by analytically minimising an upper-bound on the estimator variance. This provides a more elegant and efficient framework for combining variational and sampling methods than the one proposed originally in (de Freitas, Højen-Sørensen, Jordan and Russell 2001). This new variational perspective applies to sampling schemes such as Markov chain Monte Carlo and particle filtering (Doucet, de Freitas and Gordon 2001, Robert and Casella 1999). However, for ease of presentation, we focus on importance sampling estimators.

2 IMPORTANCE SAMPLING

We begin the paper with a revision of importance sampling (Hammersley and Handscomb 1968, Rubinstein 1981). In statistics, physics and machine learning one is often concerned with solving high-dimensional integrals with respect to a probability distribution $p(x)$ of the form[1]

$$I(f) = \int f(x)\, p(x)\, dx$$

Monte Carlo integration algorithms are easy to implement if one can sample directly from $p(x)$. This, however, turns out to be a hard problem. Importance sampling overcomes it by introducing a proposal distribution $q(x)$ such that its support includes the support

[1] We use continuous distributions to simplify the exposition, but the results also apply in discrete settings.

of $p(x)$. Then we can rewrite $I(f)$ as follows

$$I(f) = \int f(x)\, w(x)\, q(x)\, dx$$

where $w(x) \triangleq \frac{p(x)}{q(x)}$ is known as the importance weight. Consequently, if one can simulate N i.i.d. samples $\{x^{(i)}\}_{i=1}^{N}$ according to $q(x)$ and evaluate $w(x^{(i)})$, a possible Monte Carlo estimate of $I(f)$ is

$$\widehat{I}_N(f) = \sum_{i=1}^{N} f\left(x^{(i)}\right) w(x^{(i)})$$

This estimator is unbiased and, under weak assumptions, the strong law of large numbers applies, that is $\widehat{I}_N(f) \underset{N\rightarrow\infty}{\xrightarrow{a.s.}} I(f)$. It is clear that this integration method can also be interpreted as a sampling method where the posterior density $p(x)$ is approximated by

$$\widehat{p}_N(x) = \sum_{i=1}^{N} w(x^{(i)})\delta_{x^{(i)}}(x)$$

where $\delta_{x^{(i)}}(x)$ denotes the Dirac delta function. Then, $\widehat{I}_N(f)$ is nothing but the function $f(x)$ integrated with respect to the empirical measure $\widehat{p}_N(x)$.

Some proposal distributions $q(x)$ will obviously be preferable to others. An important criterion for choosing an optimal proposal distribution is to find one that minimises the variance of the estimator $\widehat{I}_N(f)$. The variance of $f(x)w(x)$ with respect to $q(x)$ is given by

$$\begin{aligned} \mathrm{var}_{q(x)}\left(f(x)w(x)\right) &= \mathbb{E}_{q(x)}\left(f^2(x)w^2(x)\right) - I^2(f) \\ &\triangleq \eta(q(x)) - I^2(f) \end{aligned} \quad (5)$$

The second term on the right hand side does not depend on $q(x)$ and hence we only need to minimise the first term ($\eta(q(x))$) subject to the constraint that $q(x)$ sums to 1. By forming an appropriate Lagrangian and taking derivatives, we obtain the following optimal importance distribution

$$q^{\star}(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$$

This distribution ensures that $\mathrm{var}_{q(x)}\left(f(x)w(x)\right) = 0$. However, it is not very useful in the sense that it is not easy to sample from $|f(x)|p(x)$. Yet, it tells us that high sampling efficiency is achieved when we focus on sampling from $p(x)$ in the important regions where $|f(x)|p(x)$ is relatively large; hence the name importance sampling.

This result implies that importance sampling estimates can be super-efficient. That is, for a a given function $f(x)$, it is possible to find a distribution $q(x)$ that yields an estimate with a lower variance than when using a perfect Monte Carlo method, *i.e.* with $q(x) = p(x)$. This property is often exploited for simulation in communication networks (Smith et al. 1997).

When the normalising constant of $p(x)$ is unknown, it is still possible to apply the importance sampling method by rewriting $I(f)$ as follows:

$$I(f) = \frac{\int f(x)\, w(x)\, q(x)\, dx}{\int w(x)\, q(x)\, dx}$$

where $w(x) \propto \frac{p(x)}{q(x)}$ is now only known up to a normalising constant. The Monte Carlo estimate of $I(f)$ becomes

$$\widetilde{I}_N(f) = \frac{\frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) w(x^{(i)})}{\frac{1}{N}\sum_{j=1}^{N} w\left(x^{(i)}\right)} = \sum_{i=1}^{N} f\left(x^{(i)}\right) \widetilde{w}(x^{(i)})$$

where $\widetilde{w}(x^{(i)})$ is a normalised importance weight. For N finite, $\widetilde{I}_N(f)$ is biased (ratio of two estimates) but asymptotically, under weak assumptions, the strong law of large numbers applies, that is $\widetilde{I}_N(f) \underset{N\rightarrow\infty}{\xrightarrow{a.s.}} I(f)$. Under additional assumptions a central limit theorem can be obtained (Geweke 1989). In particular, $N\widetilde{\sigma}_N^2 \rightarrow \mathrm{var}_{q(x)}\left(f(x)w(x)\right)$ as $N \rightarrow \infty$, where $\widetilde{\sigma}_N^2$ denotes the sample estimator of the variance:

$$\widetilde{\sigma}_N^2 = \frac{\sum_{i=1}^{N}\left(f(x^{(i)}) - \widetilde{I}_N(f)\right)^2 w^2(x^{(i)})}{\left(\sum_{i=1}^{N} w\left(x^{(i)}\right)\right)^2}$$

Throughout our experiments, we use this estimate of the variance as our measure of performance.

The estimator $\widetilde{I}_N(f)$ has been shown to perform better than $\widehat{I}_N(f)$ in some setups under squared error loss (Robert and Casella 1999).

As the dimension of the x increases, it becomes harder to obtain a suitable $q(x)$ from which to draw samples. A sensible strategy is to adopt a parameterised $q(x,\theta)$ and to adapt θ during the simulation. *Adaptive importance sampling* (AIS) appears to have originated in the structural safety literature (Bucher 1988), and has been extensively applied in the communications literature (Al-Qaq, Devetsikiotis and Townsend 1995, Remondo, Srinivasan, Nicola, van Etten and Tattje 2000). This technique has also been exploited recently in the machine learning community (de Freitas, Niranjan, Gee and Doucet 2000, Cheng and Druzdzel 2000, Ortiz and Kaelbling 2000). A popular adaptive strategy involves computing the derivative of the first term on the right hand side of equation (5)

$$D(\theta) = \mathbb{E}_{q(x,\theta)}\left(f^2(x)w(x,\theta)\frac{\partial w(x,\theta)}{\partial\theta}\right)$$

and then updating the parameters as follows

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{N} \sum_{i=1}^{N} f^2(x^{(i)}) w(x^{(i)}, \theta_t) \frac{\partial w(x^{(i)}, \theta_t)}{\partial \theta_t}$$

where α is a learning rate and $x^{(i)} \sim q(x, \theta)$. Other optimisation approaches that make use of the Hessian are also possible.

Greedy importance sampling is another interesting strategy (Schuurmans and Southey 2000). Here, one draws the initial positions of several a blocks of samples independently. Then, for each initial sample, one completes the block by taking steps in the direction of maximum $|f(x)|p(x)$ until a local maximum or a specified threshold is reached. The computation of the importance weights in this scenario requires careful thought.

In this paper, we depart from these approximate strategies. We focus on obtaining analytical expressions for $q(x)$ in some restricted, yet important scenarios.

3 MEASURE CONCENTRATION

We focus on the problem of estimating set probabilities

$$P_E \triangleq \Pr(x \in E) = \int \mathbb{I}_E(x) p(x) dx \qquad (9)$$

where $p(x)$ is a distribution that decays exponentially and $\mathbb{I}_E(x) = 1$ if $x \in E$ and 0 otherwise (see Figure 1). This is a ubiquitous problem in the graphical models literature as it is of fundamental importance in medical and industrial diagnosis systems.

In this setting the optimal proposal distribution is $q^\star(x) \propto \mathbb{I}_E(x)p(x)$. Since one cannot sample from $q^\star(x)$, we adopt a variational strategy to find other distributions from which it is easy to sample while minimising the variance of the estimates. In particular, we introduce and minimise an upper-bound on $\eta(q(x))$. This minimisation results in an "optimal" importance distribution $q^b(x)$ from which one can obtain samples easily.

We use the following identity

$$e^{\theta(x-\nu)} \geq \mathbb{I}_E(x) \qquad (10)$$

where θ is a variational parameter chosen so that the exponent is positive. θ is optimised to make the bound as tight as possible. The *rate-point* ν is the point in the set E where $p(x)$ is maximised. This is illustrated with a standard Bayesian linear-Gaussian model in Figure 2.

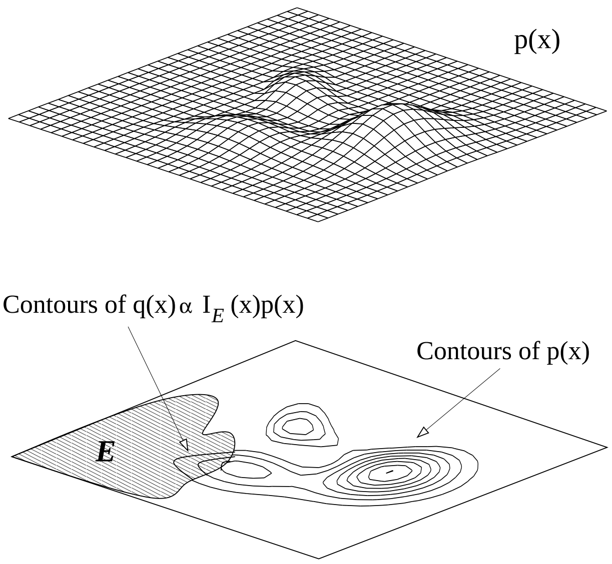

Figure 1: *Importance sampling: one should place more importance on sampling from the state space regions that matter. In this particular example one is interested in computing a tail probability of error (detecting infrequent abnormalities).*

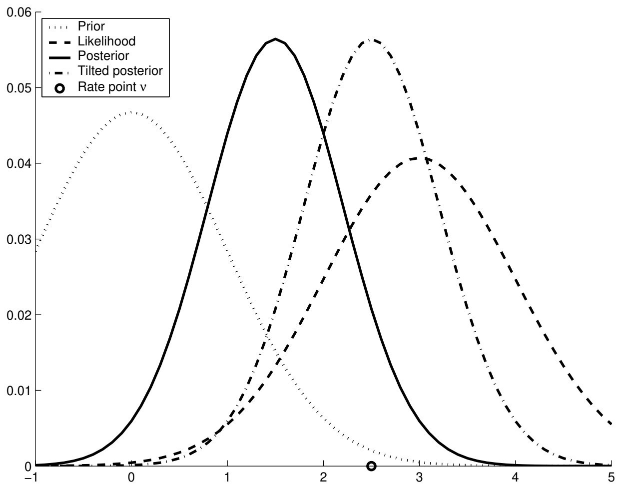

Figure 2: *In this standard Bayesian linear-Gaussian model, we are interested in calculating $P(x > \nu|data)$ for a rate-point $\nu = 2.5$.*

Our identity yields the following upper-bound on the variance:

$$\overline{\eta(q(x))} \quad \triangleq \quad \mathbb{E}_{q(x)} \left(e^{2\theta(x-\nu)} w^2(x) \right) \qquad (11)$$

This bound is minimised by the following proposal distribution:

$$\begin{aligned} q^b(x) &= \frac{p(x)e^{\theta x}}{\int p(x)e^{\theta x}dx} \triangleq \frac{p(x)e^{\theta x}}{M(\theta)} \\ &= p(x)e^{\theta x - \log M(\theta)} \triangleq p(x)e^{\mathcal{I}_\theta(x)} \quad (12) \end{aligned}$$

where $M(\theta)$ is the moment generating function. If $p(x)$ is Gaussian then it is clear that multiplying it by the term $e^{\theta x}$ simply implies a shift in the mean. By optimising the bound we can compute this shift. We do this next.

Since $p(x)$ decays exponentially, the bound should be tight at the point of maximum probability, namely ν. (This is where we are exploiting the concentration of measure blessing.) To achieve this, we optimise θ so that $\mathcal{I}_\theta(x)$ is minimised at $x = \nu$. We carry out this optimisation by differentiating and equating to zero:

$$\frac{\partial}{\partial\theta}\mathcal{I}_\theta(x)\bigg|_{x=\nu} = \frac{\partial}{\partial\theta}\left(\theta x - \log M(\theta)\right)\bigg|_{x=\nu} = 0$$

This derivative gives us

$$\nu = M^{-1}(\theta)\frac{\partial M(\theta)}{\partial\theta} = M^{-1}(\theta)\int xp(x)e^{\theta x}dx = E_{q^b}(x)$$

That is, the proposal distribution, q^b has its mean located at ν. So, for a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the "optimal" proposal distribution to compute $P(x > \nu)$ is given by $q^b(\nu, \sigma^2)$. The results of using this proposal distribution in the Bayesian linear-Gaussian problem are shown in Figure 3.

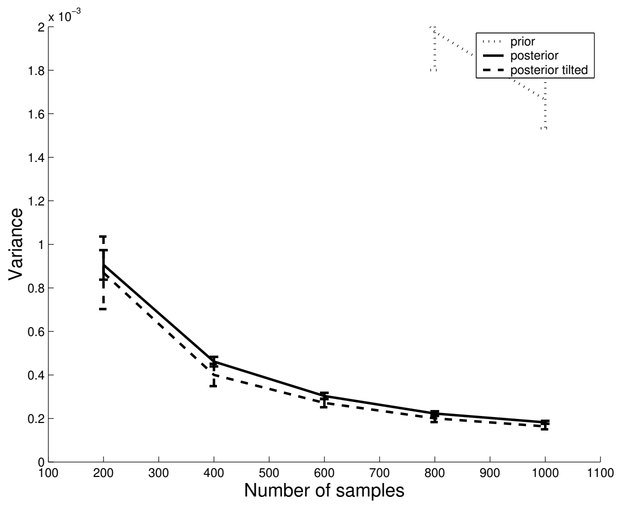

Figure 3: *Estimator variance as a function of the number of samples for the distributions and rate-point shown in Figure 2. The Tilted posterior distribution performs better than the posterior distribution as proposal mechanism. The prior performs poorly.*

The shifted distribution arises in the theory of large deviations in the proof of the lower bound of Cramér's theorem (Bucklew 1986, Dembo and Zeitouni 1993). There it is known as the *tilted distribution.* The function $\sup_\theta \mathcal{I}_\theta(x)$ is known as the *large deviations rate function* or *Legendre-Fenchel transform.* Note that although we focus on Gaussian distributions in this paper, the method applies to other distributions, such as the Laplace distribution.

The tilted distribution is only optimal within the family of upper-bounds that we are minimising. However, since both the probability of the set E and the target distribution $p(x)$ decay exponentially, it is possible to attain asymptotic optimality. This happens when the terms in equation (5) cancel out as the number of samples goes to infinity. A more detailed treatment of asymptotic efficiency is presented in (Sadowsky and Bucklew 1990).

In some situations, we might encounter more than one rate-point. In this case, it seems fairly intuitive to use a mixture of proposal distributions, each with mean at one of the rate-points. In other situations, we might not even know the location of the rate-points. We do not tackle this problem here, but one solution would be to use constrained optimisation methods to find the rate-points.

3.1 DIAGNOSIS NETWORKS

In the remainder of the paper, we shall focus on the problem of probabilistic inference in graphical models used for diagnosis. Figure 4 shows a typical network. This network has the same structure as the well known QMR medical diagnosis network (Jaakkola and Jordan 1999). We are interested in determining the probability that one of the variables (*e.g.* pressure or temperature) exceeds a particular value.

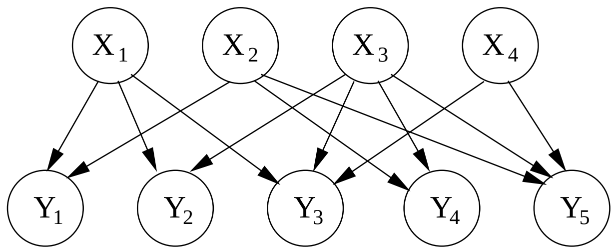

Figure 4: *Diagnosis network: the y nodes correspond to findings while the x nodes can correspond to fault states or diseases.*

For some graphical models, we might not be able to compute the posterior of a particular node of interest (the one with the constraint). However, one can use either the tilted prior distribution or run adaptive importance sampling (ais) to estimate the posterior distribution of the node and then tilt this estimate. In some cases, however, it is possible to compute the posterior of the node of interest analytically, even if we cannot compute the posterior of the entire network. That is, we simply condition on the rest of the network to compute the marginal posterior of interest. Finally, we apply the appropriate tilting to the this marginal posterior.

4 EXPERIMENTS

4.1 TWO-NODE NETWORK

We first consider a network consisting of a single multivariate node x and a multivariate node y. We consider two types of constraints. First, each dimension of x is constrained and we compute $\Pr(x_1 > \nu, x_2 > \nu, \ldots, x_d > \nu | y)$. Second, only the first dimension is constrained and we compute $\Pr(x_1 > \nu | y)$.

We tried 6 methods with different proposal distributions, namely the prior, posterior, approximate posterior, and their tilted counterparts. The approximate posterior was computed with one step of adaptation using an adaptive importance sampling method for graphical models proposed in (Cheng and Druzdzel 2000). We ensure that the means of the tilted distributions correspond to the rate-points ν.

We run each method for different dimensions and rate-points. We set the number of samples to 1000. In each case, we repeated the experiment 20 times to obtain the mean and variance of the estimator variance and the effective number of samples (the number of samples that falls in the set of interest).

We chose a Gaussian prior $p(x) = \mathcal{N}(0, I)$ and a Gaussian likelihood $p(y|x) = \mathcal{N}(Ax, I)$. The posterior was obtained using standard analytical calculations. The evidence was simply $y = 3$.

The mean estimator variances obtained in both experiments are shown in Tables 1 and 2. In both cases, the tilted distributions performed better than the other methods. In the multi-constraint case, the tilted prior outperformed the tilted posterior. This is a result of the fact that the set of interest is decaying exponentially as we increase the dimension of x. That is, there are two exponential decays, one due to the Gaussian decay and one due to the fact that the set becomes increasingly more constrained. Since the tilted prior has a higher variance it provides better results. But eventually, this double exponential decay results in a depletion of samples as shown in Figure 5. If the constrained set is held fixed, there is no depletion of samples when using the tilted distributions as shown in Figure 6. This is an encouraging result.

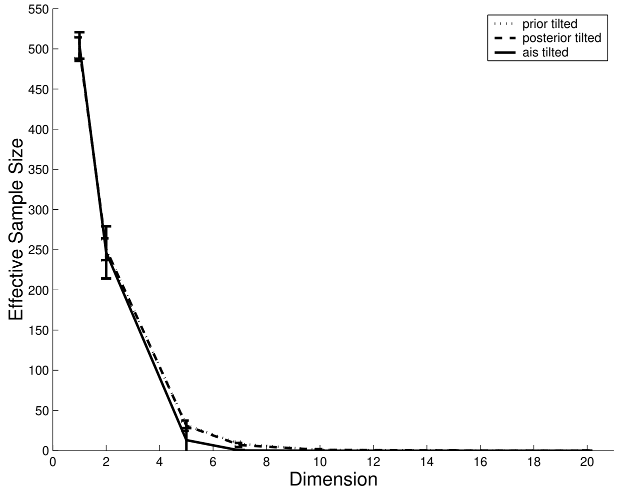

Figure 5: *Effective sample size in multi-constraint case. The fact that both the Gaussian and set of interest are decaying exponentially eventually results in a depletion of samples. Note the prior and posterior are very close.*

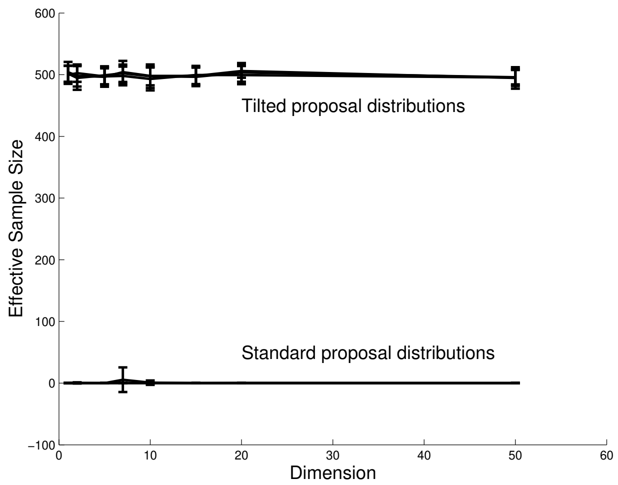

Figure 6: *Effective sample size in single-constraint case. Since the the size of the constrained regions does not decay with increasing dimension, the tilted distributions allow us to sample effectively despite the increasing dimension.*

4.2 MULTI-NODE NETWORK

In this experiment, we adopt the network shown in Figure 4, with bivariate nodes x and y. Our goal is to compute $\Pr(x_{11} > \nu, x_{12} > \nu | y)$. That is, only the first

Table 1: *Mean estimator variance for rate-point $\nu = 4$ and sample size 1000 with a 2-node network and a single constraint. We show the rank of the estimator in brackets. The symbol $-$ indicates that no samples were obtained in the set of interest.*

DIMENSIONS						METHODS
1	5	10	15	20	50	
4.30e-05(**6**)	-	-	-	-	-	prior
2.14e-07(**4**)	-		-	-	-	ais
3.32e-08(**3**)	9.16e-04(**4**)	1.71e-03(**4**)	7.59e-03(**4**)	9.48e-03(**4**)	-	ais tilted
2.00e-06(**5**)	9.99e-07(**2**)	9.99e-07(**2**)	9.99e-07(**2**)	9.99e-07(**2**)	9.99e-07(**2**)	posterior
3.18e-09(**1**)	7.61e-06(**3**)	4.33e-05(**3**)	1.55e-03(**3**)	3.81e-03(**3**)	5.81e-05(**3**)	prior tilted
3.02e-08(**2**)	2.87e-08(**1**)	2.24e-08(**1**)	2.13e-08(**1**)	3.13e-08(**1**)	3.04e-08(**1**)	posterior tilted

Table 2: *Mean estimator variance for rate-point $\nu = 4$ and sample size 1000 with 2-node network and multiple-constraints. We show the rank of the estimator in brackets. The symbol $-$ indicates that no samples were obtained in the set of interest.*

DIMENSIONS						METHODS
1	2	5	7	10	15	
4.30e-05(**6**)	-	-	-	-	-	prior
2.14e-07(**4**)	-	-	-	-	-	ais
3.32e-08(**3**)	1.52e-12(**3**)	3.28e-16(**3**)	1.13e-11(**3**)	-	-	ais tilted
2.00e-06(**5**)	-	-	-	-	-	posterior
3.18e-09(**1**)	6.30e-15(**1**)	6.20e-31(**1**)	8.17e-44(**1**)	7.66e-50(**1**)	6.90e-94(**1**)	prior tilted
3.02e-08(**2**)	4.22e-13(**2**)	8.20e-25(**2**)	3.25e-34(**2**)	2.19e-49(**2**)	5.74e-70(**2**)	posterior tilted

Table 3: *Mean estimator variance for rate-point $\nu = 3$ and sample size 1000 with diagnosis network as the number of nodes increases. We show the rank of the estimator in brackets. The symbol $-$ indicates that no samples were obtained in the set of interest.*

TOTAL NUMBER OF NODES					METHODS
3	11	21	41	71	
-	-	-	-	-	prior
-	-	-	-	-	ais
5.07e-05(**4**)	5.07e-06(**3**)	1.14e-03(**3**)	-	-	ais tilted
8.42e-06(**3**)	-	-	-	-	posterior
2.47e-06(**2**)	1.36e-08(**1**)	7.19e-08(**1**)	1.13e-05(**1**)	5.34e-03(**1**)	prior tilted
1.95e-06(**1**)	7.07e-08(**2**)	1.71e-07(**2**)	2.14e-05(**2**)	2.08e-06(**2**)	posterior tilted

bivariate node is constrained. This is a more realistic model as often we need to carry out multiple diagnosis while being very careful that one of the variables (say temperature or pressure) does not exceed a specified threshold.

Instead of varying the dimension of the nodes, we kept it constant and varied the number of nodes in the network. The number of x nodes is always one node less than the number of y nodes. The x nodes have the same prior as in the two-node network. The same is true for the likelihood models. Evidence ($y = 3$) was entered on all the y nodes. We adopted a fully connected network.

The mean estimator variances are shown in Table 3. Within variance (not shown), the prior and posterior tilted methods seem to perform similarly. The effective number of samples for both techniques is also very similar, as shown in Figure 7. Once again, it was encouraging to find out that the effective number of samples does not decay with dimensionality when using the tilted distributions. Note that standard proposal distributions failed catastrophically. This is not surprising as we are sampling in high-dimensional settings.

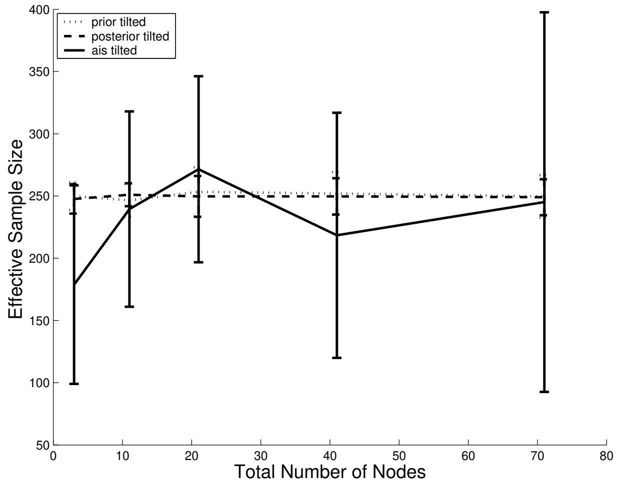

Figure 7: *Effective number of samples for diagnosis network with a varying number of nodes. Again, the effective number of samples for the tilted distributions does not decay as the dimensionality of the problem increases. The tilted prior and posterior distributions tend to exhibit less variance than the tilted adaptive posterior.*

5 CONCLUSIONS

We presented an efficient sampling method for probabilistic inference in an important class of graphical models. The method is based on a blessing of dimensionality known as the concentration of measure phenomenon. The experiments indicate that for some graphical models and sets of interest, the tilted prior and posterior distributions work well despite increasing dimensions. In addition, the number of effective samples seems to remain fairly constant.

In the future, we would like to extend the method and test in real domains. One immediate extension is to adopt the variance expressions for Markov chain Monte Carlo and particle filtering described in (Andrieu and Robert 2002, Doucet and Tadic 2002) and, subsequently, apply our variational upper-bound minimisation scheme.

ACKNOWLEDGMENTS

We would like to thank Arnaud Doucet, Dale Schuurmans and Sekhar Tatikonda for their advice and expertise in preparing this paper. Some of the experiments used the Bayesian Network Toolbox of Kevin Murphy (Murphy 2001).

References

Al-Qaq, W. A., Devetsikiotis, M. and Townsend, J. K. (1995). Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems, *IEEE Transactions on Communications* **43**(12): 2975–2985.

Andrieu, C. and Robert, C. P. (2002). Controlled MCMC for optimal sampling, Department of Statistics, Bristol University.

Bucher, C. G. (1988). Adaptive sampling — an iterative fast Monte Carlo procedure, *Structural Safety* **5**: 119–126.

Bucklew, J. A. (1986). *Large Deviation Techniques in Decision, Simulation, and Estimation*, John Wiley & Sons.

Cheng, J. and Druzdzel, M. J. (2000). AIS-BN: An adaptive importance sampling algorithm for evidential reasoning in large bayesian networks, *Journal of Artificial Intelligence Research*.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, Wiley Series in Telecommunications, New York.

de Freitas, N., Højen-Sørensen, P., Jordan, M. I. and Russell, S. (2001). Variational MCMC, *in* J. Breese and D. Koller (eds), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 120–127.

de Freitas, N., Niranjan, M., Gee, A. H. and Doucet, A. (2000). Sequential Monte Carlo methods to train neural network models, *Neural Computation* **12**(4): 955–993.

Dembo, A. and Zeitouni, O. (1993). *Large Deviations Techniques and Applications*, Jones and Bartlett, Boston.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality, *American Mathematics Society Conference on Math Challenges of the 21st Century*.

Doucet, A. and Tadic, V. (2002). On-line optimization of sequential Monte Carlo methods using stochastic approximation, *American*.

Doucet, A., de Freitas, N. and Gordon, N. J. (eds) (2001). *Sequential Monte Carlo Methods in Practice*, Springer-Verlag.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration, *Econometrica* **24**: 1317–1399.

Hammersley, J. H. and Handscomb, D. C. (1968). *Monte Carlo Methods*, Methuen, London.

Jaakkola, T. and Jordan, M. I. (1999). Variational methods and the QMR-DT database, *Journal of Artificial Intelligence* **10**: 291–322.

Murphy, K. (2001). The Bayes Net Toolbox for matlab, *Computing Science and Statistics*.

Ortiz, L. E. and Kaelbling, L. P. (2000). Adaptive importance sampling for estimation in structured domains, *in* C. Boutilier and M. Godszmidt (eds), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, pp. 446–454.

Remondo, D., Srinivasan, R., Nicola, V. F., van Etten, W. C. and Tattje, H. E. P. (2000). Adaptive importance sampling for performance evaluation and parameter optimization of communications systems, *IEEE Transactions on Communications* **48**(4): 557–565.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.

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

Sadowsky, J. S. and Bucklew, J. A. (1990). On large deviations theory and asymptotically efficient monte carlo estimation, *IEEE Transactions on Information Theory* **36**(3): 579–588.

Schuurmans, D. and Southey, F. (2000). Monte Carlo inference via greedy importance sampling, *in* C. Boutilier and M. Godszmidt (eds), *Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, pp. 523–532.

Smith, P. J., Shafi, M. and Gao, H. (1997). Quick simulation: A review of importance sampling techniques in communications systems, *IEEE Journal on Selected Areas in Communications* **15**(4): 597–613.