Monte Carlo Integration using Importance Sampling and Gibbs Sampling

Wolfgang Hörmann and **Josef Leydold**

Department of Statistics,
University of Economics and Business Administration Vienna, Austria
hormannw@boun.edu.tr

Abstract- To evaluate the expectation of a simple function with respect to a complicated multivariate density Monte Carlo integration has become the main technique. Gibbs sampling and importance sampling are the most popular methods for this task. In this contribution we propose a new simple general purpose importance sampling procedure. In a simulation study we compare the performance of this method with the performance of Gibbs sampling and of importance sampling using a vector of independent variates. It turns out that the new procedure is much better than independent importance sampling; up to dimension five it is also better than Gibbs sampling. The simulation results indicate that for higher dimensions Gibbs sampling is superior.

1. Introduction

A standard problem in scientific computing is the evaluation of an integral of the form

$$\int_{\mathbb{R}^d} q(x) f(x)\, dx$$

where $x = (x_1, x_2, \ldots, x_d)$ denotes a vector in $\mathbb{R}^d$, $f(x)$ is the multiple of a density function with bounded but often unknown integral and $q(x)$ is an arbitrary real valued function. In typical applications we want to find expectation, variance and some quantiles of all marginal distributions together with all correlations between the variables. If we consider, for example, the expectation it is obvious that we have to use $q(x) = x_k$ to obtain the expectation of the k-th marginal. Also for the variances, quantiles and correlations the evaluation of $q(x)$ is very cheap whereas the evaluation of the density $f(x)$ can be very expensive.

An application for the above setting are for example the parameter estimation problems of Bayesian statistics. Then f is the posterior density that is typically very expensive to evaluate. It is standard to calculate expectation, variance and several quantiles (eg. 2.5%, 50%, and 97.5%) of all or several important one-dimensional marginals. These integrals are calculated as they represent the Bayesian point estimates and confidence intervals for the parameters of the model.

It is important to note that the methods discussed here and the conclusions of this paper are valid for any integral of the form given above as long as the evaluation of $q(x)$ is much faster than the evaluation of $f(x)$.

Using naive Monte Carlo integration random vectors x_i with density f are generated. The average of the $q(x_i)$ is the estimator (i.e. the approximate result) of the above integral. However, often the direct generation of iid. vectors with density f is impossible or very slow. There are three general approaches to evaluate the above integral if direct generation of vectors with density f is not possible:

- Simulation of vectors with density f using the rejection technique.
- Importance sampling.
- Markov Chain Monte Carlo algorithms, in particular Gibbs sampling as the most popular one in Bayesian statistics.

There are quite a few papers on this topic in the literature, certainly an indication that this is a difficult practical problem. Papers mainly concerned with the computing methodology are eg. [6] (among many others) for the Gibbs sampler and [8], [11], [7], [13], [12] and [10] for importance sampling.

In the last years many authors spot that the routine use shifts more and more to Gibbs sampling or other Markov chain Monte Carlo methods. On the other hand it is obvious that it is very easy to parallelize importance sampling that uses independent sampling from the same distribution. In Markov Chain Monte Carlo normally one long Markov chain is generated which makes parallelization more difficult. Nevertheless, it is possible to run a small number of chains in parallel.

We were not able to find papers that compare the performance of these two so different approaches. Thus, the aim of this paper is twofold. We develop a new simple general purpose importance sampling scheme and we compare its performance to the performance of Gibbs sampling.

2. Importance Sampling

For importance sampling we have to find an importance sampling density (also called proposal density) $g(x)$. In most applications it is necessary to find more than one characteristic for several marginal distributions. It is thus in practice not possible to find the optimal importance density for a single $q(x)$. In this situation it is argued in the literature (see eg. [8] and [3]) that $g(x)$ should be roughly proportional to $f(x)$ but with higher tails. The idea of importance sampling is then to sample vectors from the proposal density $g(x)$ and to evaluate the integral

$$\int q(x)w(x)g(x)\,dx \quad \text{with} \quad w(x) = \frac{f(x)}{g(x)} \,.$$

The $w(\cdot)$ is called weight function and $w(x)$ the weight for x. As we do not know the integral of f we cannot use the classic importance sampling estimate $E_g(q(x)w(x))$ but have to use the ratio estimate $E_g(q(x)w(x))/E_g(g(x))$, see [8] for its properties.

It is quite clear that good weights should be approximately constant, in other words the weights should have a small variance. In [8] it is suggested to use the "effective sample size"

$$n_e = (\sum w_i)^2/(\sum w_i^2) = \frac{n}{V(w_i)+1} \,.$$

Maximizing the effective sample size is equivalent to minimizing the variance of the weights. It is not difficult to show that

$$n_e = \frac{n}{\int f^2(x)/g(x)\,dx} \,.$$

3. Rejection Sampling

For rejection sampling (see [9] or [2] for a detailed explanation) it is necessary to find a hat function $h(x)$ with

$$h(x) \geq f(x) \qquad \text{for} \quad x \in \mathbb{R} \,.$$

The hat function must have bounded integral and should be chosen such that it is not too difficult to generate random vectors with density proportional to $h(x)$. The integral of the hat function should be as small as possible as the expected number of trials α necessary to generate one random vector is

$$\alpha = \frac{\int h(x)\,dx}{\int f(x)\,dx} \,.$$

Note that if we consider the normed densities $\tilde{f}$ and $\tilde{h}$ associated with f and h then we have

$$\alpha = \sup_x \tilde{f}(x)/\tilde{h}(x) \,.$$

3.1. Comparing rejection and importance sampling

Perhaps the most frequently used error calculated in statistics is the mean squared error (MSE). To compare the advantages and disadvantages of the different integration methods it is therefore instructive to consider the MSE when estimating the mean of a marginal distribution. (Without loss of generality we will restrict the discussion to the first marginal distribution.) If we can generate iid. random variates directly it is well known that the MSE for estimating the mean is equal to σ^2/n, where σ^2 denotes the variance of the marginal distribution and n the sample size. As we are focusing on the case that the evaluation of f is very expensive it seems to be of interest to express the MSE in terms of N_f the number of evaluations of the density that are necessary.

If we consider rejection sampling and assume that no squeeze (simple lower bound of the density) is used, which is rarely possible for multidimensional problems, we can express the mean squared error of rejection sampling as function of N_f, the number of evaluations of f:

$$\text{MSE}_{RS}(N_f) = \frac{\alpha\sigma^2}{N_f} \,.$$

For importance sampling we can use the fact that the MSE for the mean of a marginal distribution is approximately equal to σ^2/n_e as n_e is the effective sample size. Thus we get for importance sampling:

$$\text{MSE}_{IS}(N_f) \approx \frac{\int \frac{f^2(x)}{g(x)}\,dx\;\sigma^2}{N_f} = \frac{V(w_i)+1}{N_f} \,.$$

If we compare importance sampling with rejection sampling with the same normed densities $g(x) = \tilde{h}(x)$ we can see (as pointed out in [4]) that we always have

$$\text{MSE}_{IS}(N_f) \leq \text{MSE}_{RS}(N_f)$$

as

$$\int \frac{f^2(x)}{g(x)}\,dx \leq \sup_x \left(\frac{f(x)}{g(x)}\right) \int f(x)\,dx = \alpha \,.$$

Of course the above result does not mean that rejection sampling is useless. Only for the case that we cannot use a squeeze (lower bound for the density) and that the evaluation of the density is much more expensive than the evaluation of $q(x)$ we can see that importance sampling is resulting in a smaller MSE for approximately the same computing effort.

Note: Our setting here is a typical one for Monte Carlo integration. For most discrete event simulations the situation is different. There the evaluation of $q(x)$ involves the whole simulation system and is in most cases much more time consuming than the generation of the random variates.

4. A new class of importance sampling densities

In practice the success of importance sampling heavily depends on the used importance sampling density. If we consider the variety of shapes a multi-modal density can have in higher dimensions it is not astonishing that it is impossible to find a general IS density for arbitrary densities in high dimension. There are few general purpose suggestions how to select the proposal density. An exception are adaptive importance sampling schemes (cf. eg. [7] or [13]), that have the disadvantage that they still need an importance sampling density to start with. In addition they use kernel density estimation and are thus very slow.

In this paper we will concentrate on the more tractable but still difficult case of unimodal densities f. One important importance sampling density for a unimodal f with low tails is the split-normal distribution suggested in [3]. This approach assumes that the density f is approximately multi-normal and guesses the variance-covariance matrix of the distribution using a rough numeric estimate of the Hessian matrix in the mode of the log-density. Good results are reported for this method for some Bayesian applications.

In contrast we develop a general approach to construct importance sampling densities for multivariate log-concave distributions. (Log-concave posterior densities are fairly common in Bayesian statistics.) The new method requires only the knowledge of the mode but no estimate of the variance covariance matrix. It was motivated by automatic hat-construction algorithms based on transformed density rejection (TDR) , see [9]. These TDR algorithms construct upper bounds for multivariate densities by using tangent hyperplanes of the log-densities.

For importance sampling densities it is not necessary to obtain upper bounds; only approximations of the densities are required. We thus try to approximate the given log-density $\log f$ by a hyperplane in a region. To do this it is necessary to know the location of the mode and it is convenient to decompose the d-dimensional space into 2^d quadrants using the location of the mode as origin. It is easy to see that a vector of d independent exponential variates has the property that its joint density transformed by the logarithm is a hyperplane. So in each quadrant we can use a vector of independent exponential distributions with different parameters λ_i when we try to approximate the log-density by a hyperplane. The importance sampling density in the first quadrant is the given by:

$$f(x1, x2, \ldots, x_d) = \exp(-\sum_{i=1}^{d} \lambda_i \, x_i)$$

To choose the parameters λ_i for all d coordinates in each quadrant we use the following heuristic formula that is based on the idea to find a hyperplane not too different from the (log-)transformed density:

$$\lambda_i = \frac{3}{\frac{\sqrt{d}}{\log(f(0)) - \log(f(v))} + \frac{\sqrt{d}}{\log(f(0)) - \log(f(e_i))}},$$

where v is the vector of dimension d with all entries equal to 1 and e_i denotes the i-th vector of the canonical basis. If the probabilities p_j of the proposal distribution are selected proportional to the reciprocal value of the product of all λ_i this implies that the importance sampling densities of all quadrants have the same value in the origin. In our numerical experiments we observed that for distributions with high correlations the probabilities for some quadrants are too close to 0. So we decided to correct small probabilities such that for every quadrant the probability is guaranteed to be larger than or equal to $1/(10 \cdot 2^d)$. It should be obvious from our explanations that this proposal density requires a considerable number of parameters: d different λ_i and a probability in every quadrant, which results in a total of $(d+1)2^d$ parameters. Nevertheless, it is straightforward and very fast to calculate these parameters.

5. Gibbs sampling

Gibbs sampling (see eg. [5] or [9]) is the most popular Markov Chain Monte Carlo algorithm. Given an arbitrary starting point the Gibbs sampler generates sequentially variates from the one dimensional full conditional distributions using the current values of the other coordinates. Thus the Gibbs sampler is not generating an iid. sequence of random vectors. Instead it generates a Markov chain of random vectors whose distribution converges to the distribution with density f. It is possible to show that under mild regularity conditions using the vectors x_i of the generated Markov chain the average of the $q(x_i)$ converges to the integral $\int q(x) f(x) \; dx$. Of course the choice of the starting value is of some importance. If the approximate location of the mode is known it is a good starting point. Otherwise the sampler is run for some time till a statistical test indicates that the Markov chain has converged. Then this first part of the chain (the so called burn-in) is discarded and the sampling is continued to generate a vector sample and to calculate the average of the $q(x_i)$. It is possible although not simple to calculate a confidence interval for the integration result using Gibbs sampling.

One practical problem of the Gibbs sampler is the fact that it is necessary to sample from the full conditional

distributions. If it is known that they are log-concave it is possible to use automatic random variate generation methods for one-dimensional log-concave distributions. For details see [6] and [9].

Despite the popularity of the Gibbs sampler there are hardly any investigation on the size of the mean squared error of integrals evaluated using the Gibbs sampler ([1] is such a rare example). Therefore we tried to quantify the error for a simple but important special case.

5.1. The MSE of the Gibbs sampler for the multi-normal distribution

It is instructive (and comparatively simple) to investigate the properties of the Gibbs sampling scheme for the multi-normal distribution. In the sequel we assume without loss of generality that the variances of all marginal distributions are equal to one.

It is easy to see that for the two dimensional normal distribution the full conditional distribution of X_1 given x_2 is normal with mean $r\,x_2$ and variance $1-r^2$. Thus we have the recursion

$$X_1(1) = rX_2(0) + \sqrt{1-r^2}\,\varepsilon_1$$

$$X_2(1) = rX_1(1) + \sqrt{1-r^2}\,\varepsilon_2$$

where ε_i denotes a sequence of iid. standard normal variates. It is easy to show that the $X_1(i)$ form an AR(1) process with parameter r^2 and variance $1-r^4$ [1].

To calculate the MSE of the mean of the first marginal we can write the sum of the $X_1(i)$ as sum of the ε_i with different coefficients and get:

$$\text{MSE-Gibbs-2dim}(n) = \frac{1+r^2}{n(1-r^2)}\ .$$

For dimensions $d > 2$ the Gibbs sampler of the multi-normal distribution is no longer a simple AR process but if the correlation between all variables are equal it is still not difficult to write the MSE of the expectation of a marginal distribution as a sum of normal variates with different variances and to calculate the variance of this sum numerically. Our results for some dimensions can be found in Table 1. As expected the MSE grows for increasing correlation and of course also grows for increasing dimension.

d	r	MSE Gibbs
2	0.5	1.7
3	0.5	2.2
4	0.5	2.7
6	0.5	3.9
10	0.5	5.9
2	0.9	9.5
3	0.9	17.7
4	0.9	25.9
6	0.9	42.1
10	0.9	74.2
2	0.99	98.8
3	0.99	194.7
4	0.99	289.2
6	0.99	473.6
10	0.99	824.9

Table 1: MSE of the expectation of the first marginal distribution multiplied by 10^4 for the Gibbs sampler with $n = 10000$; f the density of the multi-normal distribution with unit variances and all correlations equal to r.

6. Comparison of Gibbs sampling and importance sampling

It is not easy to make a fair comparison between these two very different methods. For the new importance sampling density proposed above we need the location of the mode, which can be easily found as we assume that the log-density is at least approximately concave. The remaining part of the algorithm is simple; it is also no problem to calculate confidence intervals for the results. However, it must not be overlooked that these confidence intervals can be wrong if the variance of the weights is high which can lead to poor estimates of the standard error. Another clear advantage of importance sampling is that if high precision is needed we can easily parallelize the simulation as we are generating iid. samples.

For Gibbs sampling it is not necessary to know the mode of the distribution but it is necessary to have generation methods available to generate variates from the full conditional distributions of the density f. It is necessary to decide when the chain has converged and it is not easy to calculate confidence intervals for the results as the generated vectors are not independent.

Looking at the above arguments we think that they are slightly in favor of importance sampling. However, most important is the question which of the two methods will lead to more precise results. So we will compare the mean squared error for a standard integration problems to get more objective arguments to decide between these two methods.

6.1. Comparing the mean squared error

First we have to decide about the rules for our comparison. As we assume that the density f is expensive to evaluate we should use the number of evaluations of f as a main factor. Unfortunately for the Gibbs sampler this number strongly depends on the structure of f. If the full conditional distributions of f are standard distributions no evaluations of f are necessary at all. If we use rejection algorithms for each of the full conditionals it can be necessary to evaluate the full conditionals four times or even more often

		Gibbs		IS indep.		IS new	
d	r	E	V	E	V	E	V
2	0.	2.0	4.0	1.1	1.7	1.1	1.7
3	0.	3.0	6.0	1.2	1.7	1.2	1.7
4	0.	4.0	8.0	1.3	1.8	1.3	1.8
6	0.	6.0	12.0	-	-	1.6	2.0
8	0.	8.0	16.0	-	-	1.8	2.5
2	0.5	3.5	4.6	1.9	3.1	1.2	1.6
3	0.5	6.6	7.5	4.8	9.9	1.7	2.3
4	0.5	11	11	16	42	2.9	4.6
6	0.5	24	22	-	-	32	47
8	0.5	40	32	-	-	210	220
2	0.9	18	17	7.1	13	2.7	5
3	0.9	55	54	80	255	8.4	18
4	0.9	104	100	550	1070	30	73
6	0.9	253	236	-	-	370	760
8	0.9	472	432	-	-	1200	1600

Table 2: The MSE multiplied by 10^4 for expectation (E) and variance (V) of the first marginal. For the multi-normal distribution with unit variances and all correlations equal to r; always using 10^4 evaluations of f; "IS indep." uses independent double exponential random variates whereas "IS new" uses our new procedure of Section 4.

		Gibbs		IS new	
d	m	E	V	E	V
2	0.3	2.3	4.4	1.2	1.7
3	0.3	4.2	6.9	1.4	2.0
4	0.3	5.6	10	1.5	2.6
6	0.3	9.6	12	2.0	2.6
8	0.3	12	19	2.2	3.3
2	0.5	3.4	6.3	1.5	2.3
3	0.5	6.6	7.8	1.9	3.3
4	0.5	10	11	2.3	3.8
6	0.5	16	17	4.1	6.4
8	0.5	20	24	8.0	13
2	0.6	4.4	6.6	2.0	3.5
3	0.6	8.4	10	3.1	5.0
4	0.6	12	15	8.2	14
6	0.6	18	19	36	58
8	0.6	26	29	126	230

Table 3: Same as Table 2 but for a multi-normal mixture distribution with the mixtures centered at $(m, m, \dots)$ and $(-m, -m, \dots)$.

to generate only one coordinate of the vector. And it is not clear if the evaluation of the full conditionals is as time consuming as that of the full density or much cheaper. So there is clearly no general way to make a fair comparison. As a simple assumption, that should not be too unrealistic for many applications, we therefore assume that for dimension d one cycle of the Gibbs sampler (ie. updating all d coordinates) requires d evaluations of the density.

For our experiments with dimension d between two and eight we use the multi-normal distribution with unit variances and equal correlations between all variables. Thus the variance covariance matrix has ones in the diagonal and all other elements are equal to ρ. For ρ we used the values 0, 0.5 and 0.9. We compare the mean squared error of the integral that calculates the mean and of the integral that calculates the variance of the first marginal always making exactly 10^4 evaluations of f. (The mode search for importance sampling and the burn-in for Gibbs sampling are not included). Due to the nature of the methods and our "rule", we have explained above, this means that we have always 10^4 repetitions for importance sampling but only between 5000 ($d = 2$) and 1250 ($d = 8$) repetitions for the Gibbs sampler.

Interpreting the results of Table 2 we can first clearly see that, as expected, our new procedure to select the importance sampling density leads to a clearly smaller mean squared error than using an independent importance sampling density. Another point is that our computational results for the mean squared error of the expectation of the first marginal when using the Gibbs sampler (see Table 1) are confirmed by this simulation study.

Looking at the case $\rho = 0$ we can compare the results with the naive simulation where the MSE is $1/n$ for the expectation and $2/n$ for the variance. The Gibbs sampler is exactly the same as naive simulation in this case. However, as we are reporting the mean squared error for a fixed number of evaluations of f the sample size decreases with dimension which in turn leads to an increased MSE.

Comparing our new importance sampling procedure and Gibbs sampling it is obvious that the importance sampling procedure is clearly better for smaller dimensions and for smaller correlations. For increasing dimension the performance of the importance sampling procedure deteriorates much more quickly than that of the Gibbs sampler so that, depending on the correlation, the break-even point is somewhere around dimension five. For stronger correlations the advantage of the Gibbs sampler in higher dimensions is even more obvious.

Table 3 gives the results for our second simulation experiment. There we used a normal mixture distributions with two equal probable parts: both standard normal vectors with mean vectors $(m, m, m, \dots)$ and $(-m, -m, -m, \dots)$ respectively. It can be clearly seen that the results are very similar to the results of Table 2.

Of course our simulation study is only comparing the performance of the two methods for two simple problems. Nevertheless, it is an indication that the Gibbs sampler that is not using approximations of the density but samples from the exact full conditional distributions performs well for increasing dimension.

7. Conclusions

We have proposed a simple novel method for constructing importance sampling densities. Our simulation results indicate that the new method works fine up to dimension five or six.

We have compared the performance of this importance sampling procedure with the Gibbs sampler. Our results show that for low dimension the performance for importance sampling is clearly better. However, due to the fact that the performance of importance sampling deteriorates much faster with the dimension the Gibbs sampler performs clearly better for higher dimensions (larger than six).

For many densities f the Gibbs sample is more complicated to implement and it is also much more difficult to parallelize. Therefore the simplicity of our general purpose importance sampling approach and the possibility to easily distribute the computations should certainly be utilized for dimensions below six. For higher dimensions importance sampling should only be used if an importance sampling density very close to the real density can be found.

Acknowledgements

This work was supported by the Austrian Science Foundation (FWF), project no. P16767-N12.

References

[1] M.-H. Chen and B. Schmeiser. Performance of the Gibbs, Hit-and-run, and Metropolis samplers. *Journal of Computational and Graphical Statistics*, 2:251–272, 1993.

[2] L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New-York, 1986.

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

[4] J. Geweke. Monte Carlo simulation and numerical integration. In A. Amman, D. Kendrick, and J. Rust, editors, *Handbook of Computational Economics*, pages 731–800. North-Holland, Amsterdam, 1996.

[5] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov chain Monte Carlo in practice*, London, 1996. Chapman & Hall.

[6] W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, 41(2):337–348, 1992.

[7] G. H. Givens and A. E. Raftery. Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *J. Amer. Statist. Assoc.*, 91(433):132–141, 1996.

[8] T. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[9] W. Hörmann, J. Leydold, and G. Derflinger. *Automatic Nonuniform Random Variate Generation*. Springer-Verlag, Berlin Heidelberg, 2004.

[10] S. J. Koopman and N. Shephard. Estimating the likelihood of stochastic volatility model: testing the assumptions behind importance sampling. Technical report, Department of Econometrics, Free University Amsterdam, 2004.

[11] M.-S. Oh and J. O. Berger. Integration of multimodal functions by Monte Carlo importance sampling. *J. Amer. Statist. Assoc.*, 88(422):450–456, 1993.

[12] A. Owen and Y. Zhou. Safe and effective importance sampling. *J. Amer. Statist. Assoc.*, 95(449):135–143, 2000.

[13] P. Zhang. Nonparametric importance sampling. *J. Amer. Statist. Assoc.*, 91(435):1245–1253, 1996.