New Importance Sampling Densities

Hörmann, Wolfgang

DOI:
10.57938/796ff644-690b-4fda-ba3b-8c20922293b3

Published: 01/01/2005

Document Version
Publisher's PDF, also known as Version of record

New Importance Sampling Densities

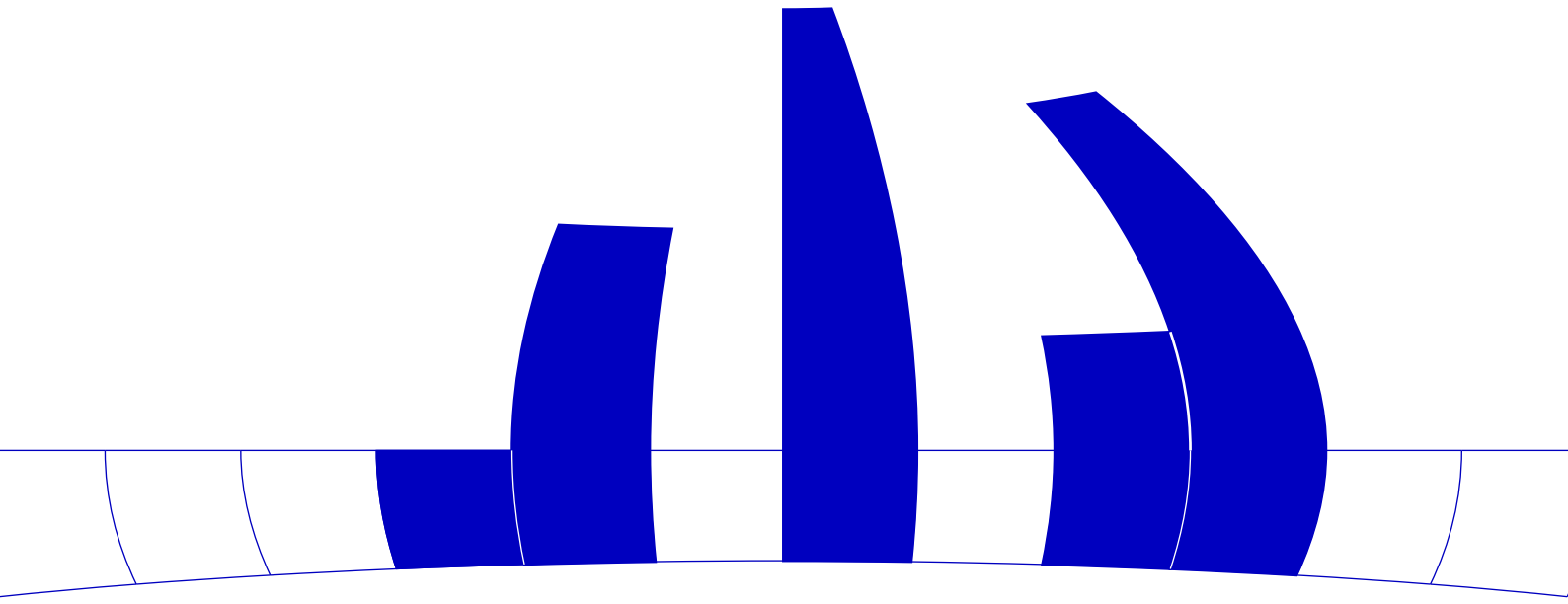
Wolfgang Hörmann

Department of Applied Statistics and Data Processing
Wirtschaftsuniversität Wien

Preprint Series

Preprint 56
May 2005

http://statmath.wu-wien.ac.at/



Journal
Vol. 00, No. 00, Month 200x, 1–17

New Importance Sampling Densities

Wolfgang Hörmann, Bogazici University Istanbul and University of Economics and BA Vienna e-mail hormannw@boun.edu.tr

(Received 00 Month 200x; In final form 00 Month 200x)

To compute the expectation of a function with respect to a multivariate distribution naive Monte Carlo is often not feasible. In such cases importance sampling leads to better estimates than the rejection method. A new importance sampling distribution, the product of one-dimensional table mountain distributions with exponential tails, turns out to be flexible and useful for Bayesian integration problems. To obtain a heavy-tailed importance sampling distribution a new radius transform for the above distribution is suggested. Together with a linear transform the new importance sampling distributions lead to simple and fast integration algorithms with reliable error bounds.

Keywords: Importance sampling, rejection method, Bayesian models

1 Introduction

A standard problem in scientific computing is the evaluation of the expectation of a function $q(x)$ with respect to a multivariate density $f(x)$ that is only known up to proportionality. The integral can be written as

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

where $x = (x_1, x_2, \ldots, x_d)$ denotes a vector in $\mathbb{R}^d$. In parameter estimation problems of Bayesian statistics it is typically necessary to find expectation, variance or several quantiles of some or all marginal distributions of the posterior distribution. For example for $q(x) = x_k$ we obtain the expectation of the k-th marginal. Also for the variances and quantiles the evaluation of $q(x)$ is very cheap whereas the evaluation of the density $f(x)$ can be very expensive.

It is well known and accepted that importance sampling is one useful approach to solve the above integral numerically. A landmark paper that describes importance sampling for Bayesian applications and states mathemat-

This work was supported by the *Fonds zur Förderung der Wissenschaftlichen Forschung,* Proj. 16767-N12.

ical conditions that guarantee finite variance is [6]. Improvements for standard importance sampling, especially adaptive algorithms can be found eg. in [17], [20], [9], [8], [21], [18]. Markov chain Monte Carlo (MCMC) is the second approach widely used for solving the above integration problem. The rapid development of MCMC in the last 15 years could lead to the conclusion that importance sampling is old fashioned or no longer needed. That this is not the case is clearly indicated by quite a few recent publications on importance sampling, especially on a dynamic variant called sequential importance sampling (see eg. [15], [14], [1], [2] and [13]).

Any importance sampling approach, also in adaptive or sequential algorithms has to start with a first importance sampling density and the success of all variants also depends on this first choice. We find it therefore astonishing that practically all papers suggest the multi-normal or multi-t distribution or mixtures of them as importance sampling distribution. Of course they are of importance especially for statistical applications but is it really not worth to try other distributions? In this paper we investigate the potential advantage of using new importance sampling distributions, among them distributions that are useful as hat-functions for rejection algorithms.

In Section 2 we introduce importance sampling and will give a simple theoretic argument why importance sampling is more useful than rejection sampling for our setting. In Section 3 we present different one-dimensional densities, Section 4 develops new multivariate distribution families useful for importance sampling. In Section 5 we present a new idea how to select the parameters of the importance sampling distribution and in Section 6 we compare the performance of our new suggestions with standard methods for two examples from Bayesian statistics.

2 Importance Sampling

The direct way to solve our integral with Monte Carlo integration clearly is to generate a sequence of random vectors X_i with density proportional to $f(x)$ and to use the sample mean of $q(X_i)$ as an estimate of the value of the integral. If it is not possible to generate the required X_i directly it is possible to generate random vectors Y_i from a similar density $g(x)$ called importance sampling density (or proposal density) and to calculate a weighted average of $q(Y_i)$ which is again an unbiased estimate of the integral (if the integral of f is known). In many applications we have to evaluate the integral for many different functions $q(x)$ with respect to the same distribution. Then it is not possible to find the optimal importance density for a single $q(x)$. In this situation it is argued in the literature (see eg. [9] and [6]) that $g(x)$ should be roughly proportional to $f(x)$ but with higher tails to obtain bounded variance

estimates. The idea of importance sampling is then to sample vectors from the proposal density $g(x)$ and to evaluate the integral

$$\frac{\int_{\mathbb{R}^d} q(x)w(x)g(x)\,dx}{\int_{\mathbb{R}^d} w(x)g(x)\,dx} \quad \text{with} \quad w(x) = \frac{f(x)}{g(x)}\,.$$

If the integral in the denominator, which is the integral of f, is known the $w(x)$ are the required weights that lead to an unbiased estimate. If that integral is unknown we have to use the ratio estimate which is the sample mean of $q(Y_i)w(Y_i)$ over the sample mean of $w(Y_i)$. (See [9] for properties of the ratio estimate.) In [6] we can find a central limit theorem for the importance sampling estimate. The main conditions are that the support of the importance sampling density contains the whole support of f and that the tails of the importance sampling density are higher than the tails of f, as this leads to a finite variance estimate. In [12] we can find (working out the details of a suggestion of [16]) a test to check if the variance of the IS estimate is finite. In practice a finite variance is of course necessary for a successful IS application but it is not sufficient as a large but finite variance is of little practical use as well.

Importance sampling is often used as a variance reduction technique. Then it is necessary to find a good g for a single function $q(x)f(x)$ of interest. In this paper, however, importance sampling is a technique to numerically calculate expectations of several functions $q(x)$ with respect to a single multivariate density $f(x)$. Therefore the selection of $g(x)$ cannot depend on q. In such situations we need importance sampling densities g that are similar to f as this leads to weights that are approximately constant. Or in other words the weights should have a small variance.

In [17] the squared variation coefficient of the weight function is used as measure for the quality of an importance sampling distribution. To allow for a better direct interpretation we prefer to add one to this measure and define the "relative variance"

$$\mathrm{RV}_{(f,g)} = \frac{V_g(w(x))}{E_g(w(x))^2} + 1 = \frac{\int \frac{f^2(x)}{g(x)}\,dx}{\left(\int f(x)\,dx\right)^2}\,.$$

The variance of the importance sampling estimate is equal to:

$$V_{IS} = \frac{\int (q(x) - E_f(q(x)))^2 \frac{f(x)}{g(x)} f(x)\,dx}{n}\,,$$

where n denotes the sample size. For $q(x)$ the indicator function of a set

with probability (with respect to f) equal to $1/2$ it is easy to see that $V_{IS} = \mathrm{RV}_{(f,g)}/4$; for $g = f$ this variance is $1/4$. Thus $\mathrm{RV}_{(f,g)}$ can be interpreted as the factor by which the variance is increased when using the importance density g instead of the original f when $q(x)$ is the indicator function of a set with probability $1/2$. In other words RV is the factor by which the sample size has to be increased to obtain the same precision as for the direct generation of random vectors with density f. In [6] and [7] the relative numerical efficiency RNE$= 1/RV_{f,g}$ is used as quality measure for g. In [9] it is suggested to use the "effective sample size" $n_e = n/\mathrm{RV}$.

A comparison with rejection sampling

It is possible to evaluate $E_f(q(x))$ using the naive Monte Carlo procedure with rejection sampling to generate vectors from the density f. To compare this approach with importance sampling we consider rejection sampling using a hat-function proportional to the importance sampling density $g(x)$. If we assume for this subsection that f and g have integral one the rejection constant α is $\sup_x f(x)/g(x)$ and $\alpha\, g(x) \geq f(x)$ is valid for every x. As we assume in this paper that the evaluation of the density f is very expensive it is useful to compare the variance of importance and rejection sampling as functions of N_f, the number of evaluations of f. As rejection sampling requires on average α trials (and evaluations of f) to generate one random vector with density f we can express the variance of the Monte Carlo procedure using rejection sampling as

$$V_{RS}(N_f) = \frac{\alpha \int (q(x) - E_f(q(x)))^2\, f(x)\, dx}{N_f}\ .$$

For importance sampling N_f is equal to the sample size. Thus using the formulas for V_{RS} and V_{IS} from above, the definition of α and the trivial inequality

$$\frac{\int (q(x) - E_f(q(x)))^2 \frac{f(x)}{g(x)} f(x)\, dx}{N_f} \leq \frac{\sup_x \left(\frac{f(x)}{g(x)}\right) \int (q(x) - E_f(q(x)))^2 f(x)\, dx}{N_f}$$

we get the result

$$V_{IS}(N_f) \leq V_{RS}(N_f)\ ,$$

also contained in [7]. It shows that when using the same importance sampling density or hat function g and the same number of evaluations of f importance sampling leads to a smaller variance than rejection sampling. Thus we

should use importance sampling and not rejection sampling for the problems considered in this paper. Of course for other stochastic simulations (eg. discrete event simulations) the costs for $q(x)$ and $f(x)$ may be totally different which makes rejection sampling more efficient than importance sampling. For a detailed comparison of importance and rejection sampling see [19] p. 103 and the references given there.

The standard approach for selecting g

The standard suggestion for importance sampling as for example described in [6] and [4] is routinely used in Bayesian applications of importance sampling and is supposed to work for multivariate distributions that are not too different from elliptical contoured. In this approach it is first necessary to find the mode of the original density $\tilde{f}$. Then using the inverse of minus the Hessian matrix of the log-density in the mode it is possible to obtain a rough estimate of the variance-covariance matrix of the distribution. The Cholesky factor of that matrix transforms the distribution with density $\tilde{f}$ into a random vector with approximately uncorrelated components; all of them have mode and expectation zero and approximately unit variance. In the sequel we will write $f(x)$ for the density of that transformed multivariate distribution. Depending on the tail behavior of f the multivariate t-distribution or the multivariate normal distribution are natural candidates that may serve as importance sampling densities for the transformed density f. This procedure of applying a linear transform and then using a multivariate normal or t-density as importance sampling density is what we call the "standard approach".

In applications it turns out that a problem of the standard approach are asymmetric densities f. In [6] we can thus find variants of the standard approach that replace the multivariate normal or t-distributions by the so-called split-normal or split-t distributions. There simply the probabilities for generating positive or negative variates are changed which leads to distributions that are better suited for asymmetric densities. Another practical problem of the standard approach is the choice of the degrees of freedom for the t-distribution or split-t distribution. To be on the safe side we would like to use heavy tails but this often leads to clearly increased variances when using the multi-t or the split-t distributions.

In the rest of this paper we work on improving the standard approach. We will try to replace the multi-normal or multi-t distribution by importance sampling distributions well suited for the asymmetric case and we will introduce a convenient practical way to decide about the tails of the importance sampling distribution.

3 One-dimensional IS distributions

To find importance sampling distributions that can improve the standard approach we start our considerations with the one-dimensional case. It is of little direct practical value but has important implications for the multivariate case. As the standard approach deals with a random vector consisting of approximately uncorrelated components we will try to use the product of one-dimensional importance sampling distributions for multivariate applications. The one-dimensional densities $f(x)$ we are going to use to asses importance sampling densities $g(x)$ are the normal distribution and the Gamma distribution. (The normal distribution as many posterior distributions have shapes close to normal and the Gamma distribution as it is asymmetric.) In the standard approach all distributions are – as explained above – standardized such that they have mode zero and the second derivative of the log density at zero is equal to minus one. For the normal distribution we thus use the standard normal density. To obtain a gamma distribution with shape parameter α, standardized in the above sense, we simply have to use the linear transform $\mathbf{T}(X) = \frac{X-(\alpha-1)}{\sqrt{\alpha-1}}$.

We should select $g(x)$ such that it is simple, allows for easy and fast generation of random variates, has not too low tails and is flexible enough to adopt to the different shapes of non-symmetric bell-shaped curves. As the normal distribution has very low tails it is not suitable. So from the importance sampling distributions suggested in the literature just the t and the split-t distributions remain although both do not have a very simple density or allow for simple random variate generation. In contrast the double-exponential distribution has a very simple density and we can easily use the inversion method for generating random variates. It is also easy to define an asymmetric double-exponential distribution as a mixture of an exponential distribution to the left and an exponential distribution to the right, both with different mean values. If we add the constraint that the density must be continuous we get a two parameter family of distributions that allow for asymmetry; we call it continuous split-exponential distribution. Using the same method we also defined a continuous split-logistic distribution. For these distributions we calculated the minimal RV using numerical integration and a numerical search procedure to find the optimal parameters. The results of Table 1 indicate that all $g(x)$ lead to good results (below 1.1) for the normal distribution. For the slightly skewed Gamma(5) distribution the double exponential distribution is no longer very suitable whereas the different "split distributions" again have RV-values below 1.1. For the heavily skewed Gamma(2) distribution the split-t distribution, that has a density discontinuous at 0, is no longer suitable whereas our new split-distributions, that are all continuous, show again good results. But there

Table 1. RV (relative variance) for different one-dimensional distributions.

Importance sampling density	Normal	Gamma(5)	Gamma(2)
double-exponential	1.08	1.20	1.40
splitnormal	1.00	1.10	1.30
t-distribution (DF=1)	1.25	1.36	1.57
t-distribution (DF=2)	1.10	1.22	1.44
t-distribution (DF=8)	1.01	1.18	1.44
split-t (DF=1)	1.25	1.32	1.46
split-t (DF=2)	1.10	1.16	1.31
split-t (DF=8)	1.01	1.08	1.25
split-logistic	1.01	1.03	1.05
split-exponential	1.08	1.08	1.09
TDR ($c = 0$)	1.02	1.02	1.02
TDR ($c = -1/2$)	1.10	1.10	1.10

exists an important practical problem, the choice of the parameters for the importance sampling distribution. To reach the very good RV results of Table 1 it was necessary to use a slow numerical search algorithm and to evaluate a lot of integrals numerically to find the optimal parameters for the different importance sampling densities $g(x)$ for all but the last two distributions. Of course that is not desirable in practice.

Therefore we need a class of importance sampling densities that allow for an easier choice of the parameters. This is possible using the concept of transformed density rejection (TDR) that was developed to construct hat-functions for rejection algorithms. (See [10] and [11].) The main idea is to find a transformation $T(x)$ such that $T(f(x))$ is a concave function. It is then easy to construct a hat function (upper bound) for $T(f(x))$ by fixing three points of contacts and calculating the point-wise minimum of the tangents in these three points. Transforming back the tangents by the inverse transform T^{-1} results in the desired upper bound for the density f. In [10] the family of transforms

$$T_c(x) = -x^{-c} \text{ for } -1 < c < 0 \text{ and } T_0(x) = \log x$$

was introduced which leads to very simple random variate generation algorithms and good fitting hats. It is of course also possible to use these TDR distributions as importance sampling densities $g(x)$. For the computational easiest special case of $c = 0$ we arrive at a table-mountain distribution with a constant center and two exponential tails; for $c = -1/2$ we obtain a table mountain distribution with heavy tails proportional to $1/x^2$. In general smaller values of c lead to distributions with higher tails.

The advantage of this approach lies in the fact that we can easily find good parameters for these table-mountain distributions by just specifying a left and a right point where a multiple of the density of the table mountain distribution $g(x)$ touches $f(x)$, (the mode of the distribution 0 is always used as point of

contact). The choice of the points of contact is not too critical as for quite different points of contact $g(x)$ still behaves similar to $f(x)$. After some experimentation we observed that it is enough to find a point x_0 on either side of the mode that approximately fulfills $5\,f(x_0) = f(0)$. In Table 1 the results for the table mountains for $c = 0$ and $c = -1/2$ constructed using this very basic rule are given in the lines called TDR; they are so close to optimal that we did not include the optimal results in the table. It is astonishing to see how very low is RV for that simple importance sampling density for $c = 0$ for all three tested distributions. For $c < 0$ the results deteriorate which was to be expected as an importance sampling density with higher than exponential tails is not necessary for the Normal or the gamma distribution. Still the results are very good and underline the favorable properties of the TDR table mountain distributions.

We include here all details of the random variate generation algorithm from the TDR table-mountain distribution with exponential tails (this is the case $c = 0$) as it is an important part of the new multivariate importance sampling algorithm we introduce in Section 4 below. The algorithm utilizes the inversion and the composition method (see [11] for details). In the below algorithm description x_l and x_r denote the left and the right point of contact whereas the third point of contact is always 0 (as this is the mode of the distribution). a_1 and a_5 denote the value of the log-density in x_l and x_r, a_2 and a_6 the approximate derivatives in those points; a_3 and a_4 stand for the border between the left and right tail regions and the center region of the table mountain distribution; a_7, a_8 and a_9 denote the area of the left-tail, the center and the right-tail regions respectively.

Algorithm TDR($c = 0$):

Require: one-dimensional density $f(x)$ with mode at 0; $\delta = 10^{-5}$.
Return: Generates a random variate X of the TDR table mountain similar to f and also returns the value $g(X)$ of the table mountain density in that point.
Set-up: Use a numerical search procedure to find the points of contact $x_l < 0$ and $x_r > 0$ with $f(x_l) \approx f(x_r) \approx f(0)/5$.
Calculate: $a_1 \leftarrow \log f(x_l + \delta)$, $a_2 \leftarrow (a_1 - \log f(x_l))/\delta$,
$a_3 \leftarrow x_l + (\log f(0) - a_1)/a_2$, $a_5 \leftarrow \log f(x_r - \delta)$,
$a_6 \leftarrow -(a_5 - \log f(x_r))/\delta$, $a_4 \leftarrow x_r + (\log f(0) - a_5)/a_6$,
$a_7 \leftarrow f(0)/a_2$, $a_8 \leftarrow f(0)(a_4 - a_3)$, $a_9 \leftarrow -f(0)/a_6$.
Sampling: Generate a $U(0,1)$ random variate and set $U \leftarrow U(a_7 + a_8 + a_9)$.
If $U < a_7$ return $X \leftarrow a_3 + \log(U/a_7)/a_2$ and $g(X) \leftarrow \exp(a_1 + (X - x_l)a_2)$.
Else set $U \leftarrow U - a_7$
If $U < a_8$ return $X \leftarrow a_3 + (a_4 - a_3)U/a_8$

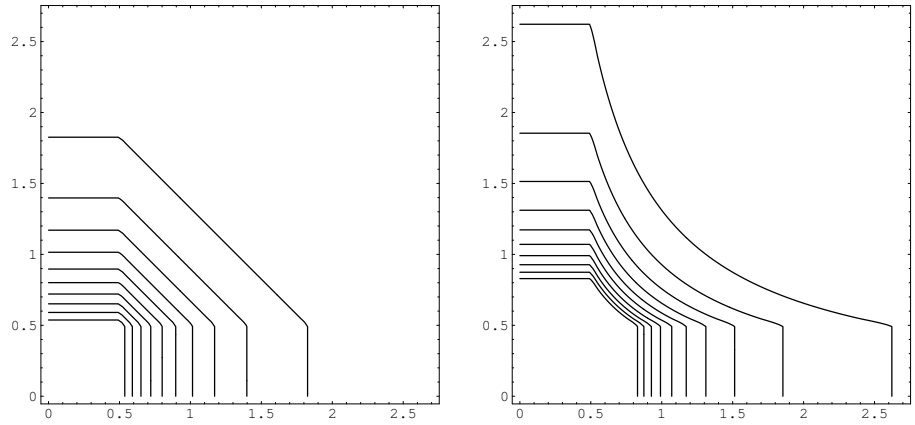


Figure 1. Table mountain distribution with exponential tails (l.h. side) and heavy tails (r.h. side)

$$\text{and } g(X) \leftarrow f(0).$$
$$\text{Else return } X \leftarrow a_4 + \log(1 - (U - a_8/a_9))/a_6$$
$$\text{and } g(X) \leftarrow \exp(a_5 + (X - x_r)a_6).$$

4 Multivariate IS distributions

For multivariate importance sampling it is necessary to find multivariate extensions of the distributions suggested in the above section. Of course this is no problem for the normal and for the t-distribution as there exist well known multivariate versions of these distributions that are also the basis for the split-normal and the split-t distributions (see [6]). For the table-mountain distributions we first tried random vectors of independent one-dimensional table-mountain distributions which means that the joint density is just the product of the one dimensional densities. That is of course simple and leads to importance sampling distribution with convex contour lines for the case $c = 0$ (exponential tails see Figure 1 left-hand side). For independent vectors of heavy tailed distributions (the case $c < 0$) the tail mass tends to concentrate around the axes and leads to clearly non-convex contour lines (see Figure 1 right-hand side).

The same concentration of the tail mass in the direction of the axes can be observed for a vector of independent t-distributions with high tails. The multi-t density has a different behavior, note also that it is, with the exception of the limiting normal distribution, never the product of one-dimensional densities.

That the concentration of the tail mass in the direction of the axes is not due to special properties of the used distribution but occurs for any product of

independent densities with heavy tails follows from the below theorem together with the fact that a distribution with higher than exponential tails cannot be log-concave.

THEOREM 4.1 *Any differentiable not log-concave one dimensional density* $f(x)$ *leads to an independent joint density* $p(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} f(x_i)$ *with non-convex contour lines.*

Proof As f is differentiable and not log-concave we can find an interval $[a, b]$ in which $\log(f(x))$ is strictly convex; (without loss of generality we assume that f is monotonically decreasing in $[a, b]$.

For the case of dimension $d = 2$ it is then enough to prove that $\log(p(a, b)) = \log(p(b, a))$ is larger than $log(p((a + b)/2, (a + b)/2))$ as this shows that the set $\{(x_1, x_2)|p(x_1, x_2) \geq p(a, b)\}$ is not convex. This is easy to see as by definition and the strict convexity of $\log f$ we have:

$$\log(p((a+b)/2, (a+b)/2)) = 2\log f((a+b)/2) < (\log f(a)+\log f(b)) = \log(p(a, b)).$$

For dimension $d > 2$ we can use exactly the same argument for a projection of the density to the subspace described by $x_i = a$ for all $i > 2$. □

The tails of a multivariate importance sampling density are of interest as they guarantee that the density is also scanned in presumable less important regions to check if possibly important details of the density are "hidden" there. It seems obvious that – at least for the "standard approach" – this scanning procedure should not favor special directions too much, a problem that is indicated by non-convex contour lines. The above theorem implies that exponential tails are the highest possible tails that lead to joint distributions with convex contour lines. Thus we may use the independent product of TDR-hats with $c = 0$ but should not use values of $c < 0$ to obtain multivariate importance sampling distributions.

The principle that the the importance sampling distribution should not favor special directions clearly leads to multivariate densities that for dimension d have to decline faster than x^{-d} in any direction. This implies, however, that for higher dimensions no density can have heavy tails in all directions. This is a way to illustrate the problem that importance sampling (like any other integration method) has for increasing dimension.

For importance sampling we need a multivariate distribution with heavy tails that combines convex contour lines with the advantages of table mountain distributions; especially we need the easy choice of the parameters even for non-symmetric distributions. To reach these aims we need a transform on $\mathbb{R}^d$ that maps random vectors of the multivariate table-mountain distribution with exponential tails ($c = 0$) into a heavy-tailed distribution without chang-

ing the distribution in a region around the mode. It is possible to reach this aim with a transform that operates just on the radius $r = ||X||$ of the random vector X where $||.||$ denotes the Euclidean distance. To leave a region around the mode unchanged the transform must leave r unchanged for values of $r < r_0$. To obtain a continuous joint density the transform must be continuously differentiable and to obtain heavy tailed distributions the transform must increase faster than a polynomial. It is easy to see that all these conditions are fulfilled by the family of radius-transforms with parameters $r_0 > 0$ and $k > 0$:

$$t(r, r_0, k) = \begin{cases} \frac{\exp(k(r-r_0))-1}{k} + r_0 & \text{for } r > r_0 \\ r & \text{for } r \leq r_0 \end{cases} ,$$

If we apply this transform to a random vector of independent table mountain distributions with $c = 0$ we obtain a very useful importance sampling distribution that we call radius transform distribution. It is the same as the independent TDR $c = 0$ distributions within the ball of radius r_0 but is transformed into a heavy tailed distribution outside that ball. As the density around the mode is not changed at all the radius transform leads to densities that are smaller than the TDR density for r only slightly larger than r_0 but are clearly larger than the TDR density for r large. Figure 2 shows an example of a radius transform distribution where the constant region of the TDR density is the square $(0, 1/2)^2$; the distribution is transformed by $t(r, r_0, k)$ with $r_0 = 1$ and $k = 1$. In the right hand side of Figure 2 (full conditional density) the density without transform was added to show the influence of the transform. We can see that mass in the region with radius between 1 and 4.5 was transformed into the farer tails.

To calculate the density of a random vector transformed with the radius transform we first write down the coordinate-wise representation of the transform:

$$t_i(x) = x_i \, t(||x||, r_0, k)/||x|| \text{ for } i = 1, \ldots, d \,,$$

where $x = (x_1, x_2, \ldots, x_d)$ denotes a point in $\mathbb{R}^d$. Straightforward algebra shows that the Jacobian is then equal to

$$J(x) = \left(\frac{t(||x||, r_0, k)}{||x||} \right)^{d-1} t'(||x||, r_0, k) \; .$$

Using the inverse of the radius transform $t(r, r_0, k)$ it is possible to find a closed form for this importance sampling density. Assuming that all variables with the exception of the first one are equal to zero it is not difficult to see

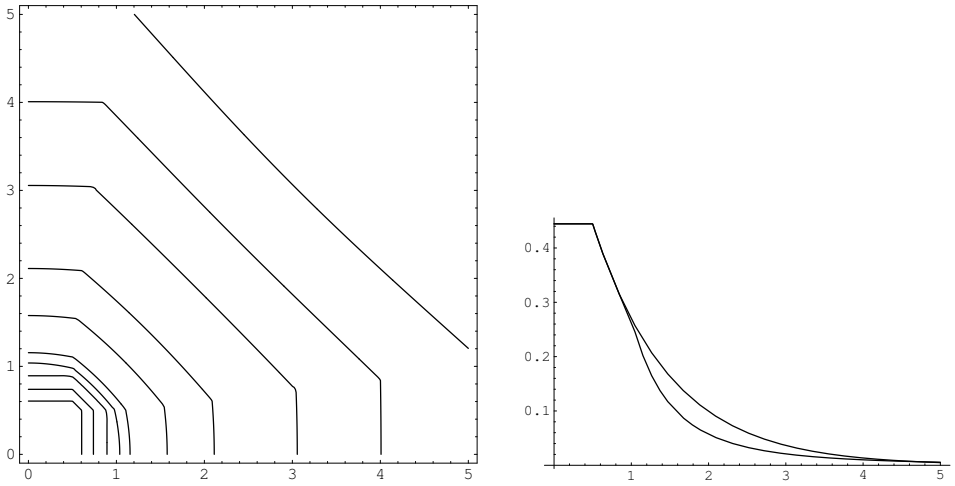


Figure 2. Contour plot of the first quadrant (l.h.s) and conditional density for $x_2 = 0$ (r.h.s) for the radius transform distribution with constant region the square $(0, 1/2)^2$, $r_0 = 1$ and $k = 1$; in the conditional density plot (r.h.s) the density of the untransformed TDR density is shown as well (upper line)

that along the axis the tails behave like $x^{-1/k-d} \log(x)^{d-1}$. Thus the parameter k governs the tail behavior that is similar to the tail behavior of the multivariate t-distribution with $1/k$ degrees of freedom. However, contrary to the t-distribution, the tail behavior is not exactly the same in all directions as the logarithmic term differs according to the direction.

For importance sampling it is not necessary to use the formula of this new importance sampling density; and it is also not necessary to construct a new generation method for that distribution. Instead we first generate a random vector X of independent table mountain distributions with $c = 0$ and calculate its density value. Then we transform the vector X into the vector Y using the radius transform; the density value of the transformed distribution is the density value of the independent table-mountain distribution divided by the Jacobian. The details are given in the below algorithm. (For the densities of the full conditional distributions we use the notation $f_i(x_i) = f(0, 0, \ldots, 0, x_i, 0, \ldots, 0)$.)

Algorithm: Radius-Transform Importance Sampling

Require: d-dimensional density $f(x)$ that has the mode at the origin and consists of approximately uncorrelated components. It is obtained by applying the "standard approach" described in Section 2. Parameters r_0 and k. ($r_0 = 1$ and $k = 1$ may be used if nothing is known about the tail behavior of f.)

Return: This Algorithm returns a vector from the importance sampling distribution and the corresponding weight.

Set-up: For $i = 1$ to d:
Make the setup of Algorithm TDR $c = 0$ for the density $f_i(x_i)$:

Sampling: (1) For $i = 1$ to d:
Generate the variate X_i from the table-mountain distribution and the density value $g_i(X_i)$ using the sampling part of Algorithm TDR $c = 0$.

(2) Calculate $r \leftarrow \sqrt{\sum_{i=1}^{d} X_i^2}$.

(3) If $r \leq r_0$ return vector $X = (X_1, X_2, \ldots, X_d)$
and weight $w = \frac{f(X)}{\prod_{i=1}^{d} g_i(X_i)}$.

(4) Else (case $r > r_0$):
For $i = 1$ to d:
Set $Y_i = X_i \; t(r, r_0, k)/\mathrm{r}$.

(5) Return vector $Y = (Y_1, Y_2, \ldots, Y_d)$
and weight $w = \frac{f(X) t(r,r_0,k)/r)^{d-1} \exp(k(r-r_0))}{\prod_{i=1}^{d} g_i(X_i)}$.

5 Finding good parameters

We have developed different importance sampling densities in the above sections. One important advantage of the TDR approach is that we have a simple procedure to decide about the parameters of the one-dimensional table-mountain. But how should we select r_0 and k? Their choice depends on the knowledge (or belief) we have about the density f. If we are sure that it has sub-exponential tails and is a density for a random vector with independent components we do not need the heavy-tail transform at all and just use the table mountain distributions with exponential tails. If we have doubts about the independence or the tail behavior it is saver to use the heavy-tail transform, r_0 decides about the radius of the ball that should be left unchanged whereas k determines about the size of the tails. Values of k equal or larger than one lead to heavy tailed distributions but this may have the disadvantage that for r only slightly larger than r_0 the density falls very fast in that region (see Figure 2 r.h.s). Of course it would be best to select these parameters such that the RV is as close as possible to one. A similar problem is the selection of the number of degrees of freedom when using the multi-t distribution for importance sampling with the standard approach.

In the literature of adaptive importance sampling we find different suggestions how to find good parameter choices for the importance sampling distribution (see for example [17] for a method involving optimization). In our setting it is of special importance to keep the number of evaluations of f small. Therefore it is desirable to use a pilot run with vectors X_i not only to estimate $RV_{(f,g)}$ for the used importance sampling density g but also $RV_{(f,\tilde{g}_j)}$ for

several other importance sampling densities $\tilde{g}_j$. For these additional estimates only evaluations of all $\tilde{g}_j(X_i)$ but no extra evaluations of f are necessary. We can use the estimate

$$\mathrm{RV}_{(f,\tilde{g}_j)} = \frac{\int \frac{f^2(x)}{\tilde{g}_j(x)}\,dx}{\left(\int f(x)\,dx\right)^2} = \frac{\sum \frac{f(X_i)}{\tilde{g}_j(X_i)} w_i/n}{\left(\sum w_i/n\right)^2}$$

where $w_i = f(X_i)/g(x_i)$ denotes the weights with respect to the original IS density and n the sample size. Note that we can only expect good estimates for RV if $g(x)$ has higher tails than $f(x)$. It is therefore safest to select $g(x)$ such that it has higher tails than all $\tilde{g}_j(x)$. The above formula can thus be used for example with $g(x)$ the multi-t distribution with one degree of freedom and the $\tilde{g}_i(x)$ t-distributions with 2, 4, 8, 16 and 32 degrees of freedom respectively. Or $g(x)$ can be the radius-transformed IS density of the above section with $r_0 = 1$ and $k = 1$ and the $\tilde{g}_i(x)$ are taken from the same distribution with $k = 1/2$, $1/4$, $1/8$, $1/16$ and $1/32$. Especially for high dimensions it is important that the sample size is large e.g. 10^5 or 10^6 as otherwise the calculated estimates for RV can be unreliable. Looking at the results of these estimates we can easily find a parameter value that leads to a small RV.

6 Two Bayesian Examples

To demonstrate the use of the new IS densities in practice we tested them for two Bayesian examples:

The first one models failure data of pumps (see [5]) by a random effect Poisson model that assumes that the Poisson parameters are drawn from a gamma density. The full model consists of two hyper parameters (we used vague Gamma priors for them) and one Poisson parameter for each pump. It is common to use Gibbs sampling for the full model but it is also no problem to integrate out all random effects. Then a posterior distribution with the two hyperparameters remains which is used as density f in the importance sampling procedure. The search for the mode with bad starting values required (using an approximative Newton method) 159 evaluations of the posterior. Using second differences in the mode we estimated the Hessian to obtain the linear transform to obtain approximate independence. For that distribution we applied the different importance sampling densities. The results of Table 2 show that all heavy tailed IS densities work fine for that example. The independent TDR densities with $c = 0$ and the split-normal distributions lead – due to the high tails of the posterior – to very unstable results and were therefore not included. As expected the radius transform distribution is (slightly)

Table 2. RV (relative variance) for 2-parameter posterior for pump-data

radius transform						independent TDR		split-t	
r_0	k		r_0	k		c		DF	
1	1	1.460	1/2	1	1.53	-1/2	1.46	1	2.24
1	1/2	1.345	1/2	1/2	1.31	-1/4	1.39	2	1.64
1	1/4	1.36	1/2	1/4	1.28	-1/8	1.46	4	1.39
1	1/8	1.44	1/2	1/8	1.35	-1/16	unst.[a]	8	1.34
1	1/16	unst.[a]	1/2	1/16	unst.[a]	–	–	16	1.42

[a] short for unstable

better than using independent TDR hats. The split-t distribution also leads to very good results but the choice of the degrees of freedom is critical. The table is organized in a way that the distributions of the same line correspond to approximately the same tail behavior for the radius-transform and the split-t distributions. So we can easily see that for the IS density with the highest tails the split-t distribution leads to worse results than the radius transform. For the independent TDR approach it is not possible to speak of a similar tail behavior as this behavior varies strongly depending on the direction we consider.

The second example is a three parameter logistic regression model for binomial probabilities using a constant improper prior. This model is applied to binary dose-response data for beetles analyzed in [3]. For this model the correlations between the regression parameters are so large that it is necessary to center the covariates when Gibbs sampling is applied. In our importance sampling approach this is not necessary as we apply a linear transform anyway. The mode search (from very bad starting values) required 500 evaluations of the posterior. In this example the sample size is large and thus the posterior is very close to multi-normal. Therefore the split-normal distribution leads to close to optimal results (RV=1.05). The other methods are very good as well (all RV$<$ 1.5 for TDR-distributions) only the split-t distribution with one (RV=3.36) and two (RV=2.12) degrees of freedom is clearly worse. To test the performance for a small dataset we randomly selected 49 out of the 481 beetles. The performance of the different IS densities for the resulting posterior are contained in Table 3: Due to the smaller sample size the posterior distribution has higher tails which implies that the split-normal distribution leads to an unbounded variance estimate and also low tail versions of the radius transform distribution and split-t distribution lead to unstable variance estimates. Like for the pump data the results show that the radius transform approach is better than the independent product of TDR hats and we can see that the split-t distribution is clearly worse than our new IS-densities for the heavy tail case (one and two degrees of freedom). For larger degrees of freedom the split-t distribution leads to very good results but again the choice of the degrees of freedom is crucial.

Table 3. RV (relative variance) for small sample three parameter logistic regression

radius transform						independent TDR		split-t	
r_0	k		r_0	k		c		DF	
1	1	1.48	1/2	1	1.72	-1/2	1.75	1	3.19
1	0.5	1.34	1/2	1/2	1.39	-1/4	1.57	2	2.08
1	1/4	1.32	1/2	1/4	1.29	-1/8	1.58	4	1.52
1	1/8	1.38	1/2	1/8	1.32	-1/16	1.7	8	1.30
1	1/16	1.48	1/2	1/16	1.41	-1/32	unst.[a]	16	1.25
1	1/32	unst.[a]	1/2	1/32	unst.[a]	–	–	32	unst.[a]

[a] short for unstable

Our experiments with other data and other Bayesian models showed similar results. So we suggest to use the radius transform distribution with parameters $r_0 = 1$ and $k = 1$ as heavy tailed importance sampling distribution together with the standard approach. One advantage are its heavy tails that should lead to bounded variance estimates for any posterior distribution, a second advantage is the flexibility for asymmetric posterior distributions. Compared with the split-t distribution with one degree of freedom we can, due to the lower RV values, expect savings in the computation time between 50 and 120 percent for two and three parameter models.

The reason that we suggest very heavy tailed importance sampling distributions is that we can trust the variance estimates and thus the confidence intervals calculated from them. As indicated by the results of Tables 2 and 3 using lower tails can decrease RV and thus reduce the variance of the importance sampling estimates but, as indicated by the unstable entries in the tables, it may lead to poor variance estimates and thus incorrect confidence intervals. Note that unstable does not mean that the importance sampling procedure leads to totally wrong results, it just indicates that the variance of the estimates may be clearly larger than estimated.

For the case of high dimensional problems or problems with very expensive density f we advise to use the pilot sample procedure explained in Section 5 to decide about good parameters for the radius transform distribution.

7 Conclusions

We have suggested a new family of multivariate distributions that have properties that are desirable for important sampling experiments used to calculate the expectation of a simple function with respect to a complicated multivariate density. Especially for the case that, due to the unknown tail behavior of the multivariate density, a heavy tailed importance sampling distribution is needed, the new distributions seem to be clearly superior to the multi-t and split-t distributions. We have also suggested a procedure that can help to de-

cide about the parameters of our new distributions. For the case of very high dimensional problems or multimodal densities our importance sampling densities should be useful as starting densities for adaptive or sequential procedures. For repeated parameter estimation of Bayesian standard models importance sampling with the new densities leads to simple and fast integration algorithms with reliable error bounds.

References

[1] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89:539–551, 2002.

[2] N. Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Annals of Statistics*, 32:2385–2411, 2004.

[3] A. J. Dobson. *An Introduction to Statistical Modelling*. Chapman and Hall, London, 1983.

[4] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, Oxford, 2000.

[5] E. George, U. Makov, and A. Smith. Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, 20:147–156, 1993.

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

[7] 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.

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

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

[10] W. Hörmann. A rejection technique for sampling from T-concave distributions. *ACM Trans. Math. Software*, 21(2):182–193, 1995.

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

[12] S. J. Koopman and N. Shephard. Estimating the likelihood of the stochastic volatility model: testing the assumptions behind importance sampling. Technical report, Vrije Universiteit Amsterdam, 2004.

[13] H.-R. Künsch. Recursive monte carlo filters. *Annals of Statistics*, page to appear, 2005.

[14] F. Liang. Dynamically weighted importance sampling in Monte Carlo computation. *J. Amer. Statist. Assoc.*, 97(459):807–821, 2002.

[15] J. S. Liu, R. Chen, and W. H. Wong. Rejection control and sequential importance sampling. *J. Amer. Statist. Assoc.*, 93(443):1022–1031, 1998.

[16] J. Monahan. *Numerical Methods of Statistics*. Cambridge University Press, Cambridge, 2001.

[17] 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.

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

[19] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Verlag, New-York, 2nd edition, 2004.

[20] M. West. Approximating posterior distributions by mixture. *Journal of the Royal Statistical Society B*, 55:409–422, 1993.

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