Josef Leydold and Wolfgang Hörmann

Smoothed Transformed Density Rejection

Working Paper

Smoothed Transformed Density Rejection

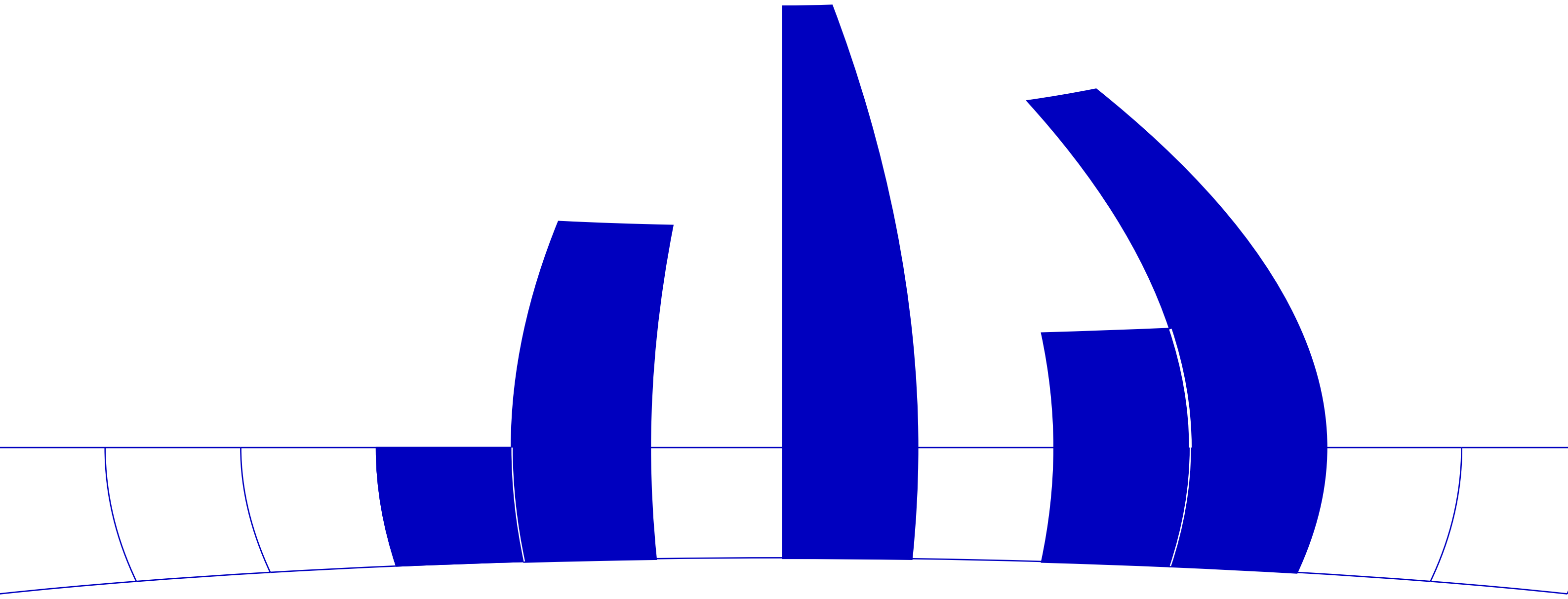
Josef Leydold, Wolfgang Hörmann

Department of Applied Statistics and Data Processing
Wirtschaftsuniversität Wien

Preprint Series

Preprint 51
June 2003

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



Smoothed Transformed Density Rejection [1]

Josef Leydold [a,*] and Wolfgang Hörmann [a,b]

[a] *University of Economics and Business Administration, Department for Applied Statistics and Data Processing, Augasse 2-6, A-1090 Vienna, Austria*

[b] *IE Department, Boğaziçi University Istanbul, 80815 Bebek-Istanbul, Turkey*

Abstract

There are situations in the framework of quasi-Monte Carlo integration where nonuniform low-discrepancy sequences are required. Using the inversion method for this task usually results in the best performance in terms of the integration errors. However, this method requires a fast algorithm for evaluating the inverse of the cumulative distribution function which is often not available. Then a smoothed version of transformed density rejection is a good alternative as it is a fast method and its speed hardly depends on the distribution. It can easily be adjusted such that it is almost as good as the inversion method. For importance sampling it is even better to use the hat distribution as importance distribution directly. Then the resulting algorithm is as good as using the inversion method for the original importance distribution but its generation time is much shorter.

Key words: Monte Carlo method, quasi-Monte Carlo method, nonuniform random variate generation, transformed density rejection, smoothed rejection, inversion
1991 MSC: 65C05, 65C10, 65D30

1 Introduction

There are quite a few situations in the framework of Monte Carlo and quasi-Monte Carlo computation where the application of nonuniform (quasi-) random variates is required. Among them the computation of expected values with respect to some distribution and importance sampling are the most important ones.

[*] Corresponding author. Tel +43 1 313 36–4695. FAX +43 1 313 36–738
Email address: `Josef.Leydold@statistik.wu-wien.ac.at` (Josef Leydold).

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

Preprint submitted to Monte Carlo Methods and Applications 23 December 2003

In the framework of Monte Carlo integration nonuniform variates are generated by transforming uniform pseudo-random numbers. There exists a lot of transformation methods for this task, see [4, 6] for surveys. For all these methods the convergence rate of the estimator is then $\mathcal{O}(N^{-1/2})$. For quasi-Monte Carlo integration so called low-discrepancy sequences (or quasi-random numbers) have to be transformed. Motivated by the Koksma-Hlawka inequality one then expects the convergence rate $\mathcal{O}(N^{-1}\log^d N)$, for an integration problem in $\mathbb{R}^d$. If nonuniform variates are required the *inversion method* is usually used, i.e. uniform (quasi-) random numbers U_i are transformed by means of the inverse of the distribution function F^{-1}, $X_i = F^{-1}(U_i)$, since this method does not change the discrepancy of the sequence. However, the inversion method is often slow and for arbitrary importance distributions only approximate numerical algorithms like Newton's methods or those proposed in [1] or [5] are available. Thus one would like to use more efficient methods. The most powerful of these is the *rejection method.* But in practice one observes that the rate of convergence is then much slower and sometimes even close to that of the Monte Carlo method, i.e. $\mathcal{O}(N^{-1/2})$. This observation is caused by the fact that the rejection method involves the integration of a discontinuous function [8, 9]. To overcome this problem Moskowitz and Caflisch [10] suggested *smoothed rejection* from a constant hat. In this article we discribe this concept and generalize it to non-constant hat functions. We continue with numerical examples and demonstrate that automatic methods like *transformed density rejection* are well suited for quasi-Monte Carlo integration and can lead to more accurate results if exact inversion is impossible or slow.

Throughout this article f and F denote the density and cumulative distribution function, resp., of some distribution of interest; i.e. of the distribution from which we have to draw random samples. Then the *expectation* of some function g with resepect to distribution F is given by the integral

$$A = \mathrm{E}_f(g) = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x}) = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,, \tag{1}$$

for which the following simple estimator can be used:

$$\tilde{A}_N = \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{x}_i)\,, \tag{2}$$

where $\boldsymbol{x}_i \sim F$ is a sample of (pseudo-) random variates with distribution F. *Importance Sampling* is the variance reduction technique most commonly used for Monte Carlo integration. There the following identity is used to compute the integral of some function g over a domain $D \subseteq \mathbb{R}^d$:

$$A = \int_D g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\,f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\,\mathrm{d}F(\boldsymbol{x}) = \mathrm{E}_f(g/f)\ . \tag{3}$$

Thus we get an estimator for this integral by

$$\tilde{A}_N = \frac{1}{N} \sum_{i=1}^{N} \frac{g(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)}\,, \tag{4}$$

where again $\boldsymbol{x}_i \sim F$ is a sample of (pseudo-) random variates with distribution F. Here the distribution F is called *importance distribution*. If it is chosen such that f behaves similar to the function g of interest, the variance of the importance estimator (4) is smaller than that of the naive estimator $\int_D g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \approx \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{u}_i)$, where $\boldsymbol{u}_i$ are uniformly distributed over D.

2 Smoothed Rejection

The standard rejection method has been introduced already by von Neumann [11] in 1951. It requires an integrable nonnegative function called *hat function* that majorizes the density f of the given distribution. It is often written as a multiple $\alpha\, h(\boldsymbol{x})$ of some density function h. Optionally a lower bound s for f, called *squeeze*, is used. We then have

$$0 \le s(\boldsymbol{x}) \le f(\boldsymbol{x}) \le \alpha\, h(\boldsymbol{x})\,. \tag{5}$$

Of course it must be easy to sample from the hat distribution. The basic algorithm itself is rather simple:

1. Generate $\boldsymbol{X} \sim h$.
2. Generate $Y \sim U(0, \alpha\, h(\boldsymbol{X}))$.
3. If $Y \le s(\boldsymbol{X})$ return $\boldsymbol{X}$.
4. If $Y \le f(\boldsymbol{X})$ return $\boldsymbol{X}$.
5. Else try again.

Step 3 is optional and could be skipped. However, it saves some evaluations of the (sometimes) expensive density function. The multiple α above is called the *rejection constant* of the algorithm. It is equal to the expected number of iterations for one random variate. Universal methods like transformed density rejection create hat function and squeeze for the given density automatically. Moreover, sampling from the hat distribution is done by inversion and is typically very fast, see [6, Sect. 4] for an introduction into such methods. Using the rejection method for f can be seen as integration of a discontinuous function:

$$A = \int_{\mathbb{R}^d} g(\boldsymbol{x})\, f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \int_0^{\alpha h(\boldsymbol{x})} g(\boldsymbol{x})\, \chi_{\{y \le f(\boldsymbol{x})\}}\,\mathrm{d}y\,\mathrm{d}\boldsymbol{x}\,, \tag{6}$$

where $\chi_{\{.\}}$ denotes the characteristic function (indicator function). The estimate (2) then reads

$$\tilde{A}_N = \frac{1}{\sum_{i=1}^{N^*} \chi_{\{y_i \leq f(\boldsymbol{x}_i)\}}} \sum_{i=1}^{N^*} g(\boldsymbol{x}_i)\, \chi_{\{y_i \leq f(\boldsymbol{x}_i)\}}\,, \tag{7}$$

where $(\boldsymbol{x}_i, y_i)$ is uniformly distributed in the region $\{(\boldsymbol{x}, y)\colon 0 \leq y \leq \alpha h(\boldsymbol{x})\}$, and N^* is chosen such that $\sum_{i=1}^{N^*} \chi_{\{y_i \leq f(\boldsymbol{x}_i)\}} \approx N$. Notice that N^*/N is (approximately) equal to the rejection constant α. Thus N^* is just the total number of points generated in Step 1 of the rejection algorithm.

However, quasi-Monte Carlo integration does not work very well for discontinuous functions. This is indicated by the fact that the Koksma-Hlawka inequality cannot be applied to such functions. Wang [12] has shown that the integration error of the characteristic function of the rejection method (when using a constant hat function) is given by $\mathcal{O}(N^{-(d+2)/2(d+1)})$.

To overcome this problem Moskowitz and Caflisch [10] suggested to replace the discontinuous characteristic function by some smooth weight function $w(y, f(\boldsymbol{x}))$ such that

$$\int_0^\infty w(y, f(\boldsymbol{x}))\,\mathrm{d}y = f(\boldsymbol{x}) \tag{8}$$

and use the estimator

$$\tilde{A}_N = \frac{1}{\sum_{i=1}^{N^*} w(y_i, f(\boldsymbol{x}_i))} \sum_{i=1}^{N^*} g(\boldsymbol{x}_i)\, w(y_i, f(\boldsymbol{x}_i))\,, \tag{9}$$

where N^* is chosen such that $\sum_{i=1}^{N^*} w(y_i, f(\boldsymbol{x}_i)) \approx N$.

Moskowitz and Caflisch [10] and Wang [12] construct such weight functions by choosing lower and upper bounds $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$, resp., to the density f, i.e. $0 \leq a(\boldsymbol{x}) < f(\boldsymbol{x}) < b(\boldsymbol{x}) \leq \alpha\, h(\boldsymbol{x})$. Then $w(y, f(\boldsymbol{x}))$ is defined as a continuous piecewise linear function that it is equal to 1 on $[0, a(\boldsymbol{x})]$ and vanishes on $[b(\boldsymbol{x}), \alpha h(\boldsymbol{x})]$, see Fig. 1.

For choosing these functions $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$ we have to keep in mind that on one side we want to reduce the number of evaluations of the density, which is only required if $a(\boldsymbol{x}) < Y < b(\boldsymbol{x})$. On the other hand the resulting weight function w should be "sufficiently" smooth. In both articles considerable improvements of the performance in the framework of quasi-Monte Carlo are reported when smoothed rejection is used whereas there is hardly any effect when using pseudo-random numbers (as one would expect).

However, there are some drawbacks with this approach. In both articles rejection from a constant hat over the unit cube $[0, 1]^d$ is used. This usually has a very poor performance, especially if d is larger than 1. Secondly, if the hat αh

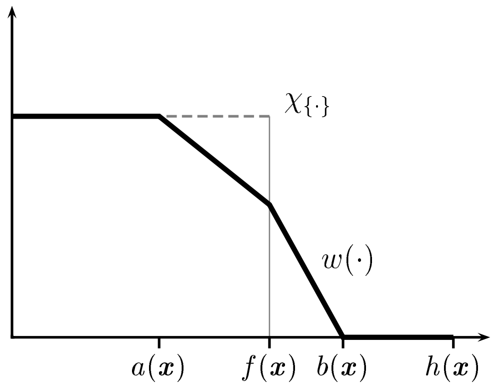


Fig. 1. Characteristic function $\chi_{\{\cdot\}}$ (dashed line) and weight function as introduced in [12] (bold line)

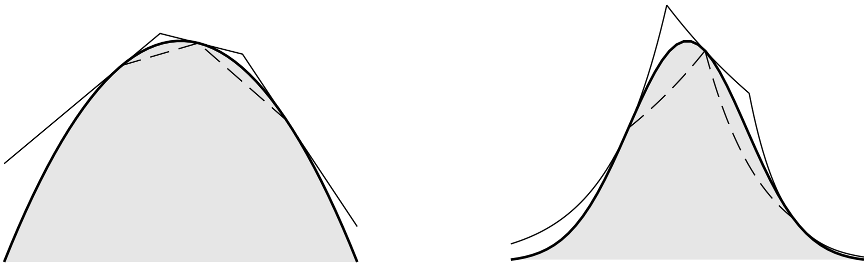


Fig. 2. Transformed density (l.h.s.) and original scale (r.h.s.) with hat and squeeze

is close to the density f then it is not necessary that the weight function vanishes at $\alpha h(\boldsymbol{x})$. This allows for a "smoother" weight function w. The fact that upper and lower bound has to be chosen "manually" can be seen as a third disadvantage of this method. To overcome these drawbacks it is natural to try to apply the idea of smoothed rejection to transformed density rejection. This is done in the next section.

3 Transformed Density Rejection

Transformed density rejection (TDR) is based on the fact that the densities of many (univariate) distributions can be transformed into concave functions by means of a monotone differentiable transformation T, i.e. the transformed densities $T(f(x))$ are then concave. Such densities are called T-*concave* densities; log-concave densities are an example with $T(x) = \log(x)$. Then tangents and secants are used to construct hat and squeeze for the transformed density. The hat is then the minimum of all these tangents. By transforming back into the original scale using T^{-1} we get hat $\alpha h(x)$ and squeeze $s(x)$ for the density; see Fig. 2 for an illustration and [6, Sect. 4] for details. Although the rejection constant is a good measure for the performance of the rejection method, the

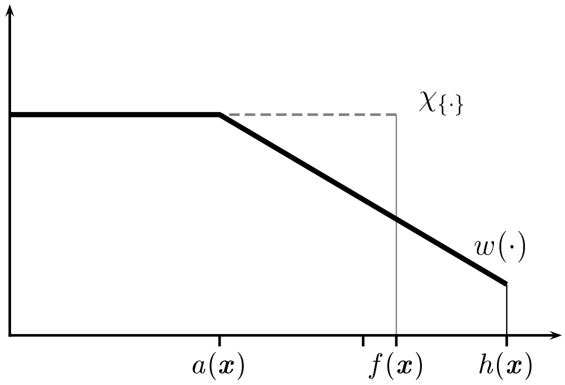


Fig. 3. Characteristic function $\chi_{\{\cdot\}}$ (dashed line) and new weight function (bold line)

ratio $\rho = \int \alpha h(x)\,dx / \int s(x)\,dx$ = area below hat/area below squeeze is easier to obtain in practice. It is a very convenient control parameter for the TDR algorithm. ρ can be made as small as desired. Then TDR is close to the inversion method and the marginal generation speed is fast and depends only on transformation T rather than f. What is most important for us is that TDR automatically delivers an upper (hat) and lower (squeeze) bound for the density which can be used for smoothed rejection. Thus we can construct the following smooth weight function (see Fig. 3): Define

$$a(\boldsymbol{x}) = \max[2\,s(\boldsymbol{x}) - \alpha h(\boldsymbol{x}), 0] \quad \text{and} \quad z(\boldsymbol{x}) = 2\frac{f(\boldsymbol{x}) - a(\boldsymbol{x})}{\alpha h(\boldsymbol{x}) - a(\boldsymbol{x})} - 1 \tag{10}$$

and

$$w(y, f(\boldsymbol{x})) = \begin{cases} 1 & \text{if } y \in [0, a(\boldsymbol{x})], \\ 1 - (1 - z(\boldsymbol{x}))\dfrac{y - a(\boldsymbol{x})}{\alpha h(\boldsymbol{x}) - a(\boldsymbol{x})} & \text{if } y \in (a(\boldsymbol{x}), \alpha h(\boldsymbol{x})] \text{ and } z(\boldsymbol{x}) \geq 0, \\ 1 - \dfrac{y}{2\,f(\boldsymbol{x})} & \text{if } y \in [0, 2\,f(\boldsymbol{x})] \text{ and } z(\boldsymbol{x}) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

It can be easily checked that $z(\boldsymbol{x}) < 0$ can only happen if $a(\boldsymbol{x}) = 0$. Moreover $w(y, f(\boldsymbol{x}))$ is continuous in y for $y \in [0, \alpha h(\boldsymbol{x})]$ and (8) holds. The lower bound $a(\boldsymbol{x})$ is used instead of the squeeze $s(\boldsymbol{x})$ to avoid too steep descents near $\alpha h(\boldsymbol{x})$.

Currently the theory of TDR is developed mainly for the univariate case. There the restriction of T-concavity can even be dropped provided that the inflection points of the transformed density are known. It is also easy to use TDR for importance distributions with independent components, i.e.

$$f(\boldsymbol{x}) = \prod_{i=1}^{d} f_i(x_i)\,. \tag{12}$$

Then we construct hat functions $\alpha_i h_i$ and squeezes s_i for each marginal density f_i and use hat function and squeeze $\alpha\, h(\boldsymbol{x}) = \prod_{i=1}^{d} \alpha_i h_i(x_i)$ and $s(\boldsymbol{x}) = \prod_{i=1}^{d} s_i(x_i)$, where $\alpha = \prod_{i=1}^{d} \alpha_i$. Notice that we also find for the ratio $\rho = \prod_{i=1}^{d} \rho_i$.

It is also noteworthy that many practical integration problems (e.g. expectations with respect to arbitrary multivariate normal distributions) can be formulated as expectations with respect to independent identical components.

When running our experiments we asked ourselves: Why not use the hat distribution h directly as importance distribution f? In our framework the hat function is a good approximation to the original importance distribution when the control parameter ρ is close to one. The performance is better than for (smoothed) rejection as we can completely skip the rejection step and the calculation of the weight function, respectively. We then have the following procedure: choose a (T-concave) density as a model for the importance distribution. Compute the hat function for TDR and use it for importance sampling.

4 Computational experiences

We have implemented smoothed TDR and want to compare it to original (non-smoothed) TDR and to the inversion method. For TDR we used hat functions with different ratios ρ from rather large (1.34 for each marginal density) to very small values (1.001). First we considered importance sampling examples. The results we obtained for several different experiments were very similar. So we report them only for the following importance sampling problem:

Example 1 *Integrate* $g(\boldsymbol{x}) = \exp\!\left(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\right)/(2\pi)^{d/2}$ *on* $[0,b]^d$ *by means of the importance density* $f(\boldsymbol{x}) = \prod_{k=1}^{d} 1/(\pi(1+x_k)^2)$ *(Cauchy distribution) restricted to* $[0,b]^d$*.*

We run the experiments with $b = 1$, 2, 3, and 5 for dimensions $d = 3$, 5, and 7. As point sets we used a pseudo-random sequence (combined multiple recursive generator `mrg31k3p` by L'Ecuyer and Touzin [7]), a base-2 Niederreiter sequence [3], and a Sobol sequence [2]. As explained above we also tried what happens when the desired importance distribution is replaced by the hat distribution. We applied randomized quasi-Monte Carlo using randomly shifted point sets. We repeated our experiments with $M = 100$ random shifts for various sample sizes and computed empirical root mean square error (rmse) as measurement for average integration error

$$\text{rmse} = \frac{1}{M}\sqrt{\sum_{n=1}^{M} \left(\tilde{A}_{N,n} - A\right)^2} \qquad (13)$$

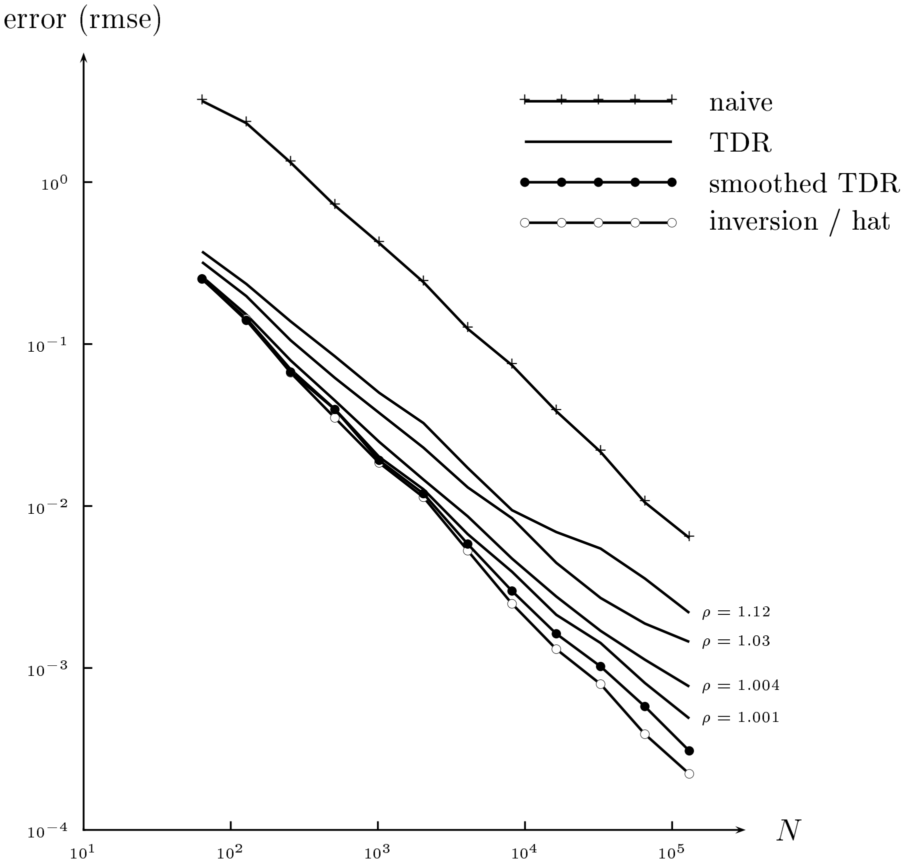


Fig. 4. Result for Experiment 1 with $b = 5$ in dimension $d = 3$ for base-2 Niederreiter sequence. Inversion method is the best method and there is almost no difference whether the hat distribution or the original distribution is used. The performance for smoothed TDR is better than for TDR. It increases when ρ becomes smaller. However the influence of ρ is much smaller in the case of smoothed TDR

where $\tilde{A}_{N,n}$ denotes the results in the n-th run with sample size N. We observed similar tendencies in our experiments although they are influenced by the values of b and d. The different values of b can be seen as integration of different functions whereas d is of course the influence of the dimension. A typical result is shown in Fig. 4. The results show that smoothed TDR can be used for importance sampling applications and not much is lost when compared to inversion. But the smallest errors are reached for inversion and for the case that we use the hat of TDR as importance density. We can say that for small values of ρ the difference between inversion and using the hat of TDR are negligible. If we remember that generating the hat of TDR is very fast, much faster than inversion for most distributions we can say that using the hat of TDR is the best method for importance sampling. This is not astonishing as it simply means that we replace the importance distribution by a distribution that is very similar to the original distribution but can be easily generated by inversion.

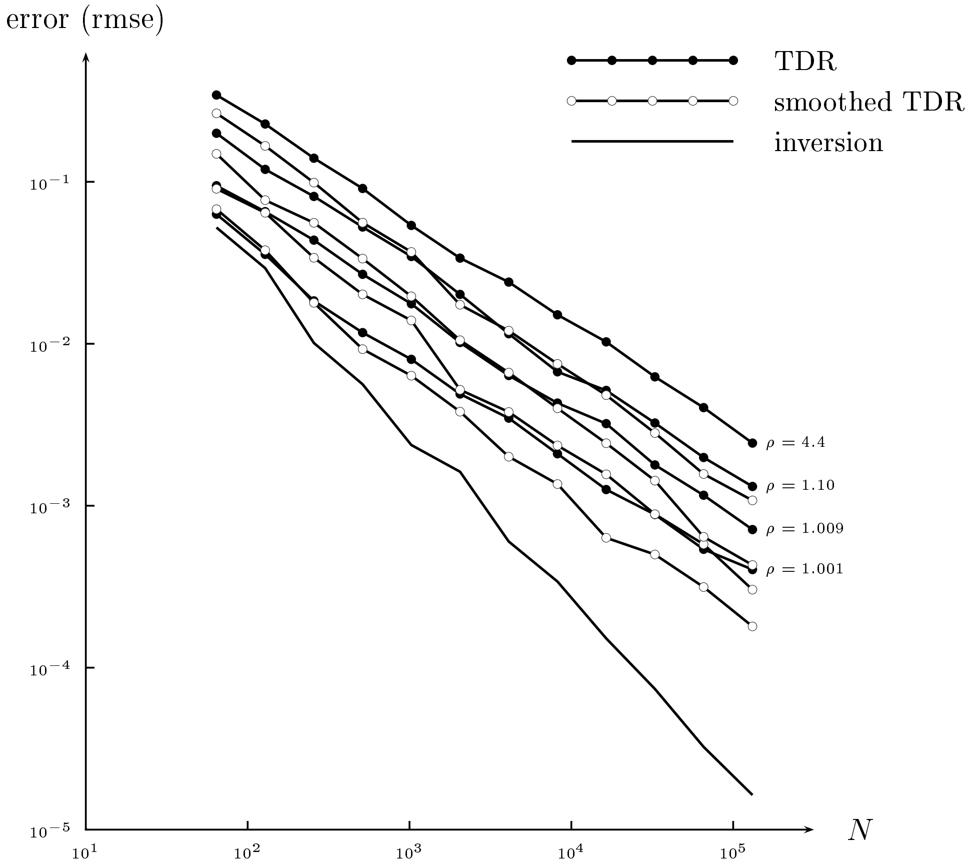


Fig. 5. Result for Experiment 2 in dimension $d = 3$ for base-2 Niederreiter sequence. The performance for smoothed TDR is better than for TDR. In both cases it increases when ρ becomes smaller.

We also tested QMC integration for computing expected values with respect to a certain distribution. Thus we run the following experiment, again in dimensions 3, 5, and 7.

Example 2 *Integrate* $g(\boldsymbol{x}) = \sqrt{\sum_{i=1}^{d} x_i^2}$ *with respect to the standard normal distribution with independent marginal distributions, i.e. with density* $f(\boldsymbol{x}) = \exp\Big(-\frac{1}{2}\sum_{k=1}^{d} x_k^2\Big)/(2\pi)^{d/2}$.

Again it is simple to obtain the exact result. So we have no problems to calculate the exact errors. A typical result is shown in Fig. 5. Here exact inversion is – as expected – best method. For TDR and for smoothed TDR the error depends on the quality of the chosen hat distribution. If ρ close to one (i.e. the hat function is a good approximation of the density) than the results are better than for larger values of ρ. Smoothed TDR is better than (original) TDR, actually it is the best method if exact inversion is not available.

As a conclusion we may say that smoothed TDR is a good method as it is a fast method and its speed hardly depends on the distribution. If the control parameter ρ is set close to 1 it is almost as good as the inversion method and it can even be used to compute integrals with respect to distributions with unknown CDF. For importance sampling it is even better to use the hat distribution as importance distribution immediately. Then the resulting algorithm is as good as using the inversion method for the original importance distribution but its generation time is much faster.

References

[1] J. H. Ahrens and K. D. Kohrt. Computer methods for efficient sampling from largely arbitrary statistical distributions. *Computing*, 26:19–31, 1981.

[2] P. Bratley and B. L. Fox. Algorithm 659: implementing Sobol's quasirandom sequence generator. *ACM Trans. Math. Software*, 14(1):88–100, 1988.

[3] P. Bratley, B. L. Fox, and H. Niederreiter. Implementation and tests of low-discrepancy sequences. *ACM Trans. Model. Comput. Simul.*, 2(3): 195–213, 1992.

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

[5] W. Hörmann and J. Leydold. Continuous random variate generation by fast numerical inversion. *ACM Trans. Model. Comput. Simul.*, 13(4): 347–362, 2003.

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

[7] P. L'Ecuyer and R. Touzin. Fast combined multiple recursive generators with multipliers of the form $a = \pm 2^q \pm 2^r$. In J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, editors, *Proc. 2000 Winter Simulation Conference*, pages 683–689, 2000.

[8] W. Morokoff and R. E. Caflisch. A quasi-Monte Carlo approach to particle simulation of the heat equation. *SIAM Journal on Numerical Analysis*, 30:1558–1573, 1993.

[9] W. Morokoff and R. E. Caflisch. Quasi-Monte Carlo integration. *J. Comp. Phys.*, 122(2):218–230, 1994.

[10] B. Moskowitz and R. E. Caflisch. Smoothness and dimension reduction in quasi-Monte Carlo methods. *Math. Comput. Modelling*, 23(8–9):37–54, 1996.

[11] J. v. Neumann. Various techniques used in connection with random digits. In A. S. Householder et al., editors, *The Monte Carlo Method*, number 12 in Nat. Bur. Standards Appl. Math. Ser., pages 36–38. 1951.

[12] X. Wang. Improving the rejection sampling method in quasi-Monte Carlo methods. *J. Comput. Appl. Math.*, 114(2):231–246, 2000.