Journal of Computational Physics **155,** 380–386 (1999)
Article ID jcph.1999.6343, available online at http://www.idealibrary.com on IDEAL®

A New Technique for Sampling Multi-modal Distributions

K. J. Abraham* and L. M. Haines†

**Department of Physics and Astronomy, Iowa State University, Ames Iowa 50011; and †School of Mathematics, Statistics, and Information Technology, University of Natal, Private Bag X01, 3209 Scottsville, Pietermaritzburg, South Africa*
E-mail: abraham@iastate.edu, haines@stat.unp.ac.za

Received January 5, 1999; revised May 21, 1999

In this paper we demonstrate that multi-modal probability distribution functions (PDFs) may be efficiently sampled using an algorithm originally developed for numerical integration by Monte Carlo methods. This algorithm can be used to generate an input PDF which can be used as an independence sampler in a Metropolis–Hastings chain to sample otherwise troublesome distributions. Some examples in one, two, and five dimensions are worked out. We also comment on the possible application of our results to event generation in high-energy physics simulations.

Key Words: Monte Carlo optimisation; Metropolis–Hastings chain; VEGAS algorithm; independence sampler.

The key to solving a wide range of optimisation problems in science and engineering lies in being able to efficiently sample a (possibly very complex) probability distribution function (PDF) in one or more dimensions. In many cases of interest, this requires inverting an integral which may not be possible by analytical or semi-analytical means. In such circumstances, efficient computer algorithms are crucial. The perhaps best known such algorithm is the Metropolis algorithm [1], which can in principle be used to generate an accurate sample from any PDF no matter how complex, by a guided random walk. However, the Metropolis algorithm is potentially inefficient when confronted with a PDF with multiple modes, or peaks, especially if they are well separated. As is well known, a very large number of random steps may be needed to locate a new mode, once one mode has been discovered, leading to a dramatic drop in the efficiency of the scheme. In this paper we will show how this problem can be circumvented in a certain class of problems.

In order to make the subsequent discussion more clear, we will present a brief analysis of the weakness of the Metropolis scheme outlined in the previous paragraph. Let $\mathbf{X}_i$ be some randomly choosen point in the space where the PDF of interest Π (not necessarily normalised) is to be sampled. A new point $\mathbf{X}_f$ at a distance δ from $\mathbf{X}_i$ is choosen and the

0021-9991/99 $30.00

ratio $\Pi(\mathbf{X}_f)/\Pi(\mathbf{X}_i)$ is evaluated. If this ratio is larger than one, then the move $\mathbf{X}_i \to \mathbf{X}_f$ is accepted. Otherwise it is accepted with probability $\Pi(\mathbf{X}_f)/\Pi(\mathbf{X}_i)$. As can be imagined, locating a single peak of Π can be easily accomplished. However, moving from one peak to another separated by a distance which is large compared with the step size δ may require a long succession of steps "against the grain"; the net probability of such a sequence is sometimes so small that a prohibitively large number of trials may be needed in order to establish the existence of the second peak. This, in a nutshell, is the reason for the potential inefficiency of the Metropolis algorithm alluded to earlier.

One plausible remedy, varying δ with each move, has been incorporated into the Metropolis–Hastings algorithm [2], where the sequence of steps is made on the basis of a proposal distribution. If the proposal distribution mimics Π, then all the peaks of Π may be found without difficulty. However, without prior knowledge of the separation between the peaks of Π, it is difficult to make a suitable choice for the proposal distribution. In other words, Π must be mapped out globally in the region of interest *before* it has even been studied. This requirement may appear to present an insurmountable obstacle to the use of the Metropolis–Hastings algorithm; the rest of this paper deals with methodology that we have developed to deal with this problem.

The key to our approach is the observation that the global structure of Π is required for another seemingly different problem, the evaluation of the definite integral of Π over the region of interest. One technique for doing so which is easily adapted to integrands of higher dimensions is adaptive Monte Carlo integration. A number of points are thrown at random along the boundaries of the region of interest (defining a grid) and the function is evaluated between the grid points. This process is repeated; however, the second time around the grid from the first iteration is refined so that it is finer in regions where the function is larger and coarser where the function is smaller. On the third iteration, the grid previously obtained is further refined, and so on. After a suitable number of iterations a reliable estimate of the integral may be obtained for a large class of integrands of interest. Several different variants of this basic algorithm have been developed; we use the VEGAS algorithm [3]. In VEGAS the grid points are used to subdivide the axes into a maximum of 50 bins.[1] Although the bin boundaries are defined along the edges of the region of integration, they may be used to break up the entire region of integration into a number of hypercubes. Ideally, the boundaries of the hypercubes are such that Π integrated over each hypercube gives the same contribution to the definite integral of Π over the region of interest. Smaller hypercubes would then correspond to regions where Π is large, and larger hypercubes to regions where Π is small.

Quite apart from the definite integral, the grid information may also be used to define a PDF $\mathcal{P}$ which roughly mimics Π. Sampling from $\mathcal{P}$ is straightforward; hypercubes are picked at random in such a way that the probability of picking any given hypercube is the same for all hypercubes, and a random number is used to locate a point $\mathbf{X}$ in the hypercube by uniform sampling. $\mathcal{P}$ is defined so that it is the same for all points in a given hypercube, and the value of $\mathcal{P}$ in a hypercube of volume ΔV is $\frac{1}{\Delta V}$. More specifically, in one dimension a random number is used to pick a bin along the x axis in such a way that the probability of picking any bin is the same. Then a second random number is used to pick a point within the bin, all points within the bin being sampled uniformly. ΔV is the bin width, so $\mathcal{P}$ for the point chosen is defined as the inverse of the width of the bin in which the point is located, independent of the precise point choosen in the bin. In two

[1] We have turned off the stratified sampling option in VEGAS, ensuring that there are 50 bins along each axis.

dimensions two random numbers are used to pick an area element, and another two random numbers are used to pick a point in the area element. ΔV is now the area, so $\mathcal{P}$ at the point chosen is defined to be the inverse of the area element. In effect, we have sampled the function globally and have used VEGAS to adaptively construct a PDF $\mathcal{P}$ which is different from Π but which nonetheless mimics Π. This procedure can obviously be generalised to arbitrarily high dimensions. Regions where Π is large (small) correspond to regions where ΔV is small (large) and hence to regions where $\mathcal{P}$ is large (small).

Our strategy for sampling from Π amounts to setting up a Metropolis–Hastings chain using $\mathcal{P}$ as a proposal distribution. From the discussion in the previous paragraph it is clear that regions where Π are large are more likely to be selected than where Π is small. A move $\mathbf{X}_i \to \mathbf{X}_f$ is accepted (rejected) if

$$\frac{\Pi(\mathbf{X}_f)}{\mathcal{P}(\mathbf{X}_f)} \times \frac{\mathcal{P}(\mathbf{X}_i)}{\Pi(\mathbf{X}_i)} > \mathrm{rn}(<\mathrm{rn}), \tag{1}$$

where rn is a random number uniformly distributed between 0 and 1. Essentially, we are using $\mathcal{P}$ as an independence sampler for Π. This method does preserve the condition of detailed balance and the stationary distribution of the resulting Markov Chain does indeed correspond to Π [4]. It is worth mentioning that in calculating sample averages not only accepted but also rejected points need to be taken into account. Note that the fixed step size δ plays no role whatsoever; rather δ varies from move to move tuned to the separation between the peaks of Π. One potential objection to this scheme is that the function must be evaluated a large number of times by VEGAS before a random sample can be drawn from it and it is not obvious whether the number of function evaluations needed is less than would be required in an approach with fixed step size. This objection will be addressed in the examples we consider.

The first and simplest example we consider is a mixture of univariate Gaussians defined on the interval [0, 22]. The precise function Π is given by

$$0.5\{\mathcal{N}(x, 3, 1)\} + 0.2\{\mathcal{N}(x, 14, .025)\} + 0.3\{\mathcal{N}(x, 19, .75)\}, \tag{2}$$

where $\mathcal{N}(x, \bar{x}, \sigma^2)$ denotes a univariate Gaussian with mean $\bar{x}$ and variance σ^2. This function clearly has well-separated multiple peaks; generating a sample from a PDF of this kind is thus liable to be problematic.

The first step in our approach is to integrate Π with VEGAS, preserving the grid information generated by VEGAS. In this case the grid information is a set of 50 points in the interval [0, 22]. The points define bins which are such that the contribution to the definite integral from each bin is nearly equal. As expected, the bins are narrow (wide) where the integrand is large (small). Π was evaluated 2500 times for this purpose and a grid reflecting the peaks in Π was used to generate bins of varying widths. These bins were used to define $\mathcal{P}$ in the interval [0, 22] along the lines just described. $\mathcal{P}$ thus obtained has been plotted in Fig. 1; the correspondence between Fig. 1 and Eq. (2) is striking.

The next step is to generate a sample from Π using $\mathcal{P}$ as an independence sampler. The acceptance rate of the Metropolis–Hastings chain is remarkably high, about 80%; i.e., about 80% of the moves were accepted using the criterion defined in Eq. (1). This is desirable from the point of view of minimising CPU time and reflects the accuracy with which $\mathcal{P}$ mimics the underlying distribution Π defined in Eq. (2). In all, Π was evaluated a total of 15,000 times to generate a sample. We have checked that the average value of the random

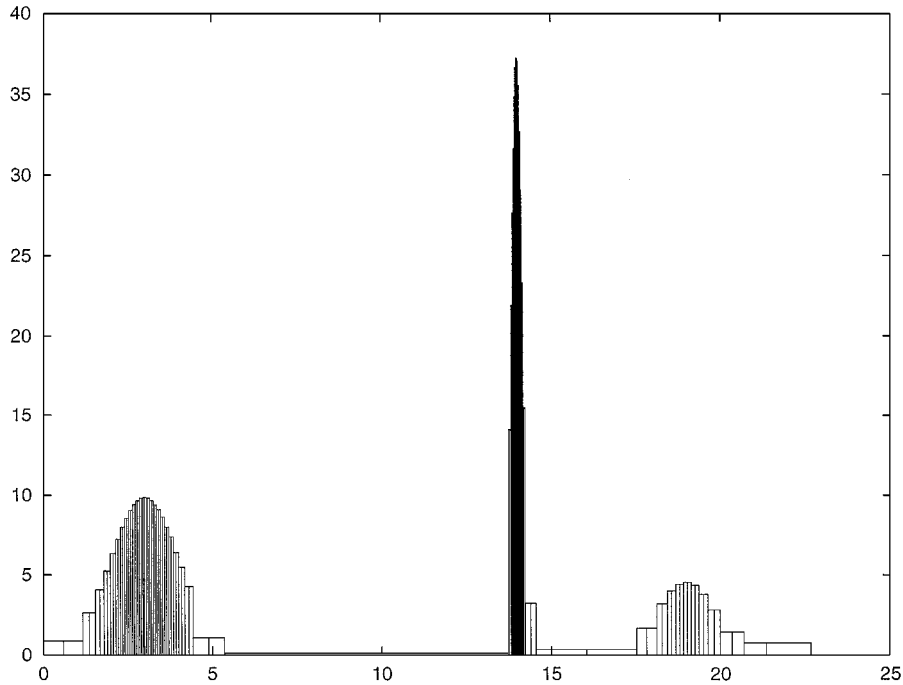

FIG. 1. $\mathcal{P}$ corresponding to Eq. (2).

variable as well as a number of higher moments is correctly reproduced, within statistical error bars. To ensure that this agreement is not a fluke and to check convergence we used a number of independent chains with independent starting points to get a sample of the various moments; for all the moments we checked the sample standard deviation was never more than a few percent of the moments.

This implies not only that all peaks have been discovered but, crucially, that the relative weights of all the peaks have also been correctly reproduced. By way of comparison, we have checked that running a Metropolis chain with Π evaluated over 100,000 times (6 times more than with the independence sampler) with fixed step size does not convincingly reproduce even the first two moments. The advantage of our approach is clear; the additional cost in function evaluations needed to set up the grid is compensated for by the ease with which the different peaks in the distribution are sampled.

We now go on to two-dimensional examples. Here a complication arises; in dimensions larger than one the VEGAS algorithm implicitly assumes that $\mathcal{P}$ is factorisable; i.e., $\mathcal{P}$ may be accurately represented in the form $\mathcal{P} = p_i(x_i)\,p_j(x_j)\cdots$. For many functions of interest this is a reasonable approximation; however, if the function has a peak along a lower dimensional hypersurface other than a co-ordinate axis, this approximation may be a poor one. In particular, the VEGAS algorithm performs poorly if the function (assumed to be defined in a hypercube) has a peak along a diagonal of the hypercube. However, this does not mean that the distribution $\mathcal{P}$ generated from the VEGAS grid cannot be used to sample from Π. All that happens is that the acceptance rate of the resulting Metropolis chain is lower. To illustrate this point, we consider a mixture of two bivariate Gaussians in a square whose means lie along a diagonal. The precise function is defined below,

$$\Pi = 0.7\{\mathcal{G}(x, y, 4, 4, 1, 1, 0.8)\} + 0.3\{\mathcal{G}(x, y, 12, 12, 1, 1, -0.8)\}, \tag{3}$$

where $\mathcal{G}(x, y, \mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ is defined by

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right].$$

The region of integration is a (16×16) square with one corner at the origin and sides along the positive x and y axes. This function is not well suited to evaluation by VEGAS as both peaks lie along a diagonal of the square, and this is reflected in the fact that the acceptance rate of the Metropolis–Hastings chain is only approximately 23%. However, the grid information does correctly reflect the location of both peaks, as was verified by the fact that both $\langle x\rangle$ and $\langle y\rangle$ were reproduced to within a few percent of the true values. This in itself is significant as both $\langle x\rangle$ and $\langle y\rangle$ for Π defined above are located where Π essentially vanishes. These values thus could not have been reproduced to any reasonable accuracy had only one peak been found. As a further check we have also considered higher moments, i.e., $\langle x^n y^m\rangle$, where $(m+n) \leq 6, m, n = 0, 1, 2, \ldots$. Once again we used multiple independent chains with independent starting points as a cross-check on the moments. The values for the various moments deviated from the true values from a few percent (for the lower moments) to substantially more for higher moments, as would be expected. As a check, we have considered another function,

$$\Pi = 0.7\{\mathcal{G}(x, y, 4, 4, 1, 1, 0.8)\} + 0.3\{\mathcal{G}(x, y, 12, 4, 1, 1, -0.8)\},$$

which differs from the bivariate Gaussian in Eq. (3) in that both peaks now lie along a line parallel to the x axis. Once again, grid information is used to generate a sample from which correct moments can be recovered. This time though, due to the more favourable location of the peaks, the acceptance rate is almost twice as high as previously. We see again that an adaptive Monte Carlo approach can generate an independence sampler for a Metropolis–Hastings chain even when the target distribution Π is two dimensional and has well-separated modes. It is worth pointing out that modifying Π by the introduction of stepping stone distributions [5] has been suggested as a means to facilitate sampling PDFs of this nature; in our approach no such modifications are necessary.

We conclude with a discussion of the relevance of our methods for event generation in experimental high-energy physics simulations, where a sample from a potentially very complicated differential scattering cross section dependent on more than two variables is required. If analytic inversion is not possible (as is often the case), another approach such as rejection sampling is needed. This however requires an enveloping distribution which must be somehow obtained, either by guesswork or possibly by using the VEGAS grid information [6]. While it is relatively easy to make an educated guess for a candidate enveloping distribution, there is no foolproof way to construct an enveloping distribution which envelops the function at *all* points in the region of integration. This can be problematic at points where the function is larger than the proposed enveloping distribution. Rescaling the enveloping distribution at such points would require that all hitherto accepted points be reexamined to see if they would still be accepted with the rescaled distribution. Even if no such points are encountered, in the absence of any rigorous proof that the enveloping distribution truly envelops there is also no rigorous proof that the points in the generated sample are truly representative.

Alternatively, the grid information may be used to construct an importance sampler for a Metropolis–Hastings chain which can be used to generate events. Let us emphasise that

this approach is radically different from that proposed in [6]. In particular, the asymptotic convergence of the Markov Chain towards the function of interest is independent of whether or not the importance sampler is larger than the function of interest over the entire region of integration. Thus there is no need to modify the grid at any time and none of the complications due to an imperfect choice of enveloping function arise. It is also worth pointing out that rejection sampling with an enveloping distribution which does not envelop everywhere can be made rigorous by using an acceptance probability similar to that in the Metropolis–Hastings algorithm [7].

To test this in practise, we have considered the example of anomalous single t production in future $\gamma\gamma$ colliders, followed by $t \to bl\nu$ evaluated in the narrow width approximation for the t and W [8]. The five-dimensional phase space has been integrated over with VEGAS and the resulting grid was used as an importance sampler to generate events along the lines of the previous examples. Neglecting the effects of cuts, smearing, and hadronisation, we obtained an acceptance rate of about 75%, even though no attempt whatsoever was made to optimise the grid. In particular, our sampling did not make any use of simplifications resulting either from the use of the narrow width approximation or from the $(V - A)$ structure of weak decays.

However, this example is oversimple as the phase space is factorisable. To see what happens when this is not necessarily true, we have considered radiative production of ν pairs at LEP 2, more precisely $e^+e^- \to \bar{\nu}_e\bar{\nu}_e\gamma$ at $\sqrt{s} = 195$ GeV. The helicity amplitudes for this process (in the limit of zero electron mass) may be found in [9]. In order to avoid singularities we have restricted the phase space to the region where $E_\gamma > 5$ GeV and require that the photon make an angle of at least 20° with the beampipe. We have included the W exchange amplitudes and the full Z propagator, so the phase space is no longer factorisable. There are forward and backward peaks in the photon angle as well as a peak in the invariant mass of the $\bar{\nu}\nu$ pair. Thus, there is a wealth of structure in the distribution from which we wish to sample. As with our previous example, we have set up an independence sampler using the VEGAS grid and have generated events from a Metropolis–Hastings chain. The acceptance probability was about 50%. As we made no use of our prior knowledge of the Breit–Wigner structure of the Z propagator or of the peaks in the photon angle distribution or of the flat distribution in the azimuthal angle about the beam axis, this acceptance probability is probably the minimum that could be acchieved but is still encouragingly high. Fifty chains starting from independent points in the five-dimensional phase space each with 100 events were simulated. To assess the convergence of the chains, we compared the variances of a number of observables in each chain with the variances for the same observables between chains, along the lines of [10]. The estimated potential scale reduction as defined in [10] was never more than 1.1, giving us reason to believe that the chains were overlapping and that stationarity and thus convergence was reached.

This suggests that the methods we have outlined may be worthwhile incorporating into event generators for high-energy physics, at least in instances when the phase space can be integrated over with VEGAS. It is significant that with Markov chain Monte Carlo methods it is feasible to quantify the extent to which the chain is believed to have converged (using the analysis of [10], for example), possibly making precise estimates of Monte Carlo error easier than with more conventional rejection methods. Quite apart from the relative efficiency (or lack of it) of our techniques compared with more conventional rejection schemes we have described a procedure which is complementary to rejection schemes with

independent sources of error, which may thus provide a useful cross-check on the accuracy of rejection schemes.

ACKNOWLEDGMENTS

K.J.A. thanks Krishna Athreya for valuable encouragement and useful discussions, and John Hauptman for reading a preliminary version of the manuscript. In addition, it is a pleasure to acknowledge many valuable clarifications and comments from Hal Stern and Mark Kaiser. L.M.H. thanks the University of Natal and the National Research Foundation, South Africa, for financial support.

REFERENCES

1. N. Metropolis *et al.*, Equations of state calculations by a fast computing machine, *J. Chem. Phys.* **21**, 1087 (1953).
2. W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97 (1970).
3. P. Lepage, A new algorithm for adaptive multidimensional integration, *J. Comput. Phys.* **27**, 192 (1978).
4. L. Tierney, Markov chains for exploring posterior distributions, *Ann. Stat.* **22**, 1701 (1994).
5. N. Sheehan and A. Thomas, On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme, *Biometrics* **49**, 163 (1993).
6. S. Kawabata, A new Monte Carlo event generator for high energy physics, *Comput. Phys. Commun.* **41**, 127 (1986).
7. L. Tierney, Markov chains for exploring posterior distributions, *Ann. Stat.* **22**, 1701 (1994).
8. K. J. Abraham, K. Whisnant, and B.-L. Young, Searching for an anomalous $\bar{t}q\gamma$ coupling via single top quark production at a $\gamma\gamma$ collider, *Phys. Lett. B* **419**, 381 (1998).
9. K. J. Abraham, J. Kalinowski, and P. Sciepko, New probes of anomalous WW_γ couplings at future e^+e^- linacs, *Phys. Lett. B* **339**, 136 (1994).
10. A. Gelman and D. B. Rubin, Inference from iterative simulation using multiple sequences, *Stat. Sci.* **7**, 457 (1992).