Monte Carlo efficiency improvement by multiple sampling of conditioned integration variables

Sebastian Weitz, Stephane Blanco, Julien Charon, Jérémi Dauchet, Mouna El-Hafi, Vincent Eymet, Olivier Farges, Richard Fournier, Jacques Gautrais

Monte Carlo efficiency improvement by multiple sampling of conditioned integration variables

Sebastian Weitz[a,b,c,d], Stéphane Blanco[e,f], Julien Charon[b,c,d,i,j], Jérémi Dauchet[a,b], Mouna El Hafi[c,d], Vincent Eymet[k], Olivier Farges[c,d,*], Richard Fournier[e,f], Jacques Gautrais[e,h]

[a] *Université Clermont Auvergne, Sigma-Clermont, Institut Pascal, BP 10448, F-63000 Clermont-Ferrand, FRANCE*
[b] *CNRS, UMR 6602, IP, F-63178 Aubière, France*
[c] *Université Fédérale de Toulouse Midi-Pyrénées, Mines Albi, UMR CNRS 5302, Centre RAPSODEE, Campus Jarlard, F-81013 Albi CT Cedex 09, France*
[d] *CNRS, UMR 5302, RAPSODEE, F-81013 Albi, France*
[e] *Université Paul Sabatier, UMR 5213 - Laboratoire Plasma et Conversion d'Energie (LAPLACE), BAt. 3R1, 118 Route de Narbonne, F-31062 Toulouse cedex 9, France*
[f] *CNRS, UMR 5213, LAPLACE, F-31062 Toulouse, France*
[g] *Université Paul Sabatier, UMR 5169 - Centre de Recherches sur la Cognition Animale (CRCA), BAt. 4R3, 118 Route de Narbonne, F-31062 Toulouse cedex 9, France*
[h] *CNRS, UMR 5169, CRCA, F-31062 Toulouse, France*
[i] *Université Perpignan Via Domitia, ED 305, 52 av. Paul Alduy, 66860 Perpignan Cedex 9, France*
[j] *Processus, Materials and Solar energy laboratory, PROMES-CNRS, 7 rue du four solaire, 66120 Font Romeu Odeillo, France*
[k] *MésoStar, F-31062 Toulouse, France*

Abstract

We present a technique that permits to increase the efficiency of multidimensional Monte Carlo algorithms when the sampling of the first, unconditioned random variable consumes much more computational time than the sampling of the remaining, conditioned random variables while its variability contributes only little to the total variance. This is in particular relevant for transport problems in complex and randomly distributed geometries. The proposed technique is based on an new Monte Carlo estimator in which the conditioned random variables are sampled more often than the unconditioned one. A significant contribution of the present Short Note is an automatic procedure for calculating the optimal number of samples of the conditioned random variable per sample of the unconditioned one. The technique is illustrated by a current research example where it permits to increase the efficiency by a factor 100.

Keywords: Monte Carlo integration, Monte Carlo efficiency, Monte Carlo in complex geometry, statistical physics

1. Introduction

Monte Carlo integration is used in many research fields (*e.g.* radiation transport physics, quantum mechanics, financial computing [1, 2]) to evaluate multidimensional integrals that can be written as the expectation $\mathcal{A}$ of a random variable W:

$$\mathcal{A} = E[W] = \int_{\mathcal{D}_X} dx\; p_X(x) \int_{\mathcal{D}_Y(x)} dy\; p_Y(y;x)\; \hat{w}(x,y) \tag{1}$$

where X and Y are (vector) random variables (defined by their domains $\mathcal{D}_X$ and $\mathcal{D}_Y(x)$ as well as their associated probability densities p_X and $p_Y(y;x)$), and W is the random variable defined by the function $\hat{w}$ that to X and Y associates $W = \hat{w}(X,Y)$. Monte Carlo integration permits to evaluate an unbiased estimator of $\mathcal{A}$ by sampling n

*Corresponding author
Email address: olivier.farges@mines-albi.fr (Olivier Farges)

independent and identically distributed (IID) random variables X_i and Y_i (where all the X_i are IID as X, and all the $Y_i(x)$ are IID as $Y(x)$). The plain Monte Carlo estimator A_n is defined by

$$\mathcal{A} = E[A_n] \text{ with } A_n = \frac{1}{n}\sum_{i=1}^{n} \hat{w}(X_i, Y_i). \tag{2}$$

The practical use of Monte Carlo integration is sometimes limited by the prohibitive computational cost required to obtain an estimate with the required precision (the standard deviation σ_{A_n} of the Monte Carlo estimate being inverse proportional to $\sqrt{n}$). This has motivated research to increase the efficiency, which is a quality measure for a Monte Carlo estimator taking into account both its precision and its computational cost [3]:

$$\epsilon_{A_n} = \frac{1}{\sigma^2_{A_n} C_{A_n}} \tag{3}$$

where $\sigma^2_{A_n}$ is the variance of A_n, and C_{A_n} the computational cost required to calculate A_n. Depending on the specific problem, several variance reduction techniques might permit to increase the efficiency (*e.g.*, importance sampling, stratified sampling, control variates and antithetic sampling [2]). The present Short Note presents a technique that increases the Monte Carlo efficiency for problems where the sampling of the unconditioned random variable X is computationally expensive (compared to the sampling of the conditioned random variable Y) whereas the variability of X contributes only little to the variance of W (compared to the variability of Y). This will be quantified in Sec. 2. Such a situation is encountered, *e.g.*, in transport problems in complex geometries where the geometry is statistically distributed (see Sec. 3 for a practical example). The principle is to consider a new Monte Carlo estimator in which Y is sampled more often than X. To our knowledge, despite the simplicity of this technique, it has never been explicitly reported in the Monte Carlo literature. Its formal investigation in the present Short Note permits us in particular to provide an easy-to-implement procedure to automatically compute the optimal number of samples of Y per sample of X (at the end of Sec. 2).

2. Efficiency-optimized Monte Carlo algorithm

We propose to use the new estimator A_{n,n_Y} of $\mathcal{A}$ defined by

$$\mathcal{A} = E[A_{n,n_Y}] \text{ with } A_{n,n_Y} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n_Y}\sum_{j=1}^{n_Y} \hat{w}(X_i, Y_{ij}). \tag{4}$$

where all the $Y_{ij}(x)$ are IID as $Y_i(x)$. A_{n,n_Y} is indeed an estimator of $\mathcal{A}$ since $E[\hat{w}(X_i, Y_{ij})] = E[\hat{w}(X_i, Y_i)]$ for all j. Note that the plain Monte Carlo estimator A_n corresponds to $n_Y = 1$ in Eq. 4. The Monte Carlo algorithm corresponding to Eq. 4 is:

1. repeat n times (for i from 1 to n):
   (a) realize a sample x_i of X_i;
   (b) repeat n_Y times (for j from 1 to n_Y):
      i. realize a sample y_{ij} of Y_{ij};
      ii. calculate $\hat{w}_{ij} = \hat{w}(x_i, y_{ij})$;
   (c) calculate the Monte Carlo weight $\hat{f}_i = \frac{1}{n_Y}\sum_{j=1}^{n_Y} \hat{w_{ij}}$;
2. calculate the Monte Carlo estimate $a_{n,n_Y} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i$ and the standard error $\sigma_{A_{n,n_Y}} = \frac{1}{\sqrt{n-1}}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} \hat{f}_i\right)^2}$.

Let us now determine the efficiency increase permitted by this technique. Therefore we first have to express the contributions of X and Y to the total variance $\sigma^2_{A_{n,n_Y}}$ and the total computational cost $C_{A_{n,n_Y}}$ of the proposed Monte Carlo estimator A_{n,n_Y}. Denoting $\sigma^2_X = Var_X[E_Y[W|X]]$ the explained variance (which is the contribution of X) and $\tilde{\sigma}^2_Y = E_X[Var_Y[W|X]]$ the unexplained variance (which is the contribution of Y) of the random variable W, and then applying successively the law of total variance and the Lindeberg-Levy central limit theorem, leads to

$$\sigma^2_{A_{n,n_Y}} = \frac{1}{n}\left(\sigma^2_X + \frac{1}{n_Y}\tilde{\sigma}^2_Y\right). \tag{5}$$

Moreover, denoting C_X and C_Y the computational costs associated, respectively, with a single sampling of X and Y, we can write

$$C_{A_{n,n_Y}} = n\left(C_X + n_Y C_Y\right). \tag{6}$$

Then, the Monte Carlo efficiency ϵ_{A_n} (defined by Eq. 3) can be expressed as function of n_Y using Eqs. 5 and 6, leading to

$$\epsilon_{A_{n,n_Y}} = \frac{1}{\left(\sigma_X^2 + \frac{1}{n_Y}\tilde{\sigma}_Y^2\right)\left(C_X + n_Y C_Y\right)}, \tag{7}$$

which is maximal for

$$n_Y^* = \frac{\tilde{\sigma}_Y}{\sigma_X}\sqrt{\frac{C_X}{C_Y}} = \frac{1}{r_\sigma\, r_C} \tag{8}$$

where we have introduced the ratios $r_\sigma = \frac{\sigma_X}{\tilde{\sigma}_Y}$ and $r_C = \sqrt{\frac{C_Y}{C_X}}$. Therefore, the maximal possible efficiency gain obtained thanks to the proposed technique, that we define as the ratio between the maximal efficiency $\epsilon_{A_{n,n_Y^*}}$ of the new Monte Carlo estimator A_{n,n_Y} and the efficiency $\epsilon_{A_{n,n_Y=1}}$ of the plain Monte Carlo algorithm A_n, is [1]

$$G_{A_{n,n_Y^*}} = \frac{\left(1 + r_\sigma^2\right)\left(1 + r_C^2\right)}{\left(r_\sigma + r_C\right)^2}. \tag{9}$$

Eq. 9 also shows that the here-proposed technique permits to greatly increase the Monte Carlo efficiency in all situations where both $r_\sigma \ll 1$ and $r_C \ll 1$.

We finally propose a procedure to determine n_Y^*, the optimal number of samples of Y per sample of X. In Eq. 8 we have expressed n_Y^* as a function of the ratios r_σ and r_C. These ratios cannot be easily computed directly, but they can be estimated using two runs of the reformulated Monte Carlo algorithm (Eq. 4), one with $n = n_1$ and $n_Y = n_{Y,1}$, the other with $n = n_2$ and $n_Y = n_{Y,2}$. $n_{Y,1}$ and $n_{Y,2}$ must be different, and the choice of n_1 and n_2 is a trade-off between required precision and computation time. [2] The obtained standard deviations σ_1 and σ_2, and the observed total computational times C_1 and C_2, permit to compute σ_X and $\tilde{\sigma}_Y$ (using Eq. 5), C_X and C_Y (using Eq. 6), and finally r_σ and r_C:

$$r_\sigma \approx \sqrt{\frac{\frac{n_1}{n_{Y,2}}\sigma_1^2 - \frac{n_2}{n_{Y,1}}\sigma_2^2}{n_2\sigma_2^2 - n_1\sigma_1^2}} \qquad r_C \approx \sqrt{\frac{n_1 n_{Y,1} C_2 - n_2 n_{Y,2} C_1}{n_2 C_1 - n_1 C_2}}. \tag{10}$$

Note that the number of Monte Carlo samples required to obtain an estimator with a relative standard error σ_r^* can also be easily deduced from Eq. 5 (by replacing $\sigma_{A_{n,n_Y}} = \sigma^*$ and $n_Y = n_Y^*$) [3].

3. Validation of the efficiency increase technique on a test case

Test case. We plan to use the efficiency increase technique presented in this short note to calculate the differential scattering cross section $W_s(\theta_s)$ of complex-shaped particles using the Monte Carlo implementation of Schiff's approximation presented in [4]. The Monte Carlo integral formulation of the problem is given by Eq. D.9 of [4] [4]. This

[1] Note that $G_{A_{n,n_Y^*}} \to_{r_\sigma \to 0, r_C \to 0} \frac{1}{(r_\sigma + r_C)^2}$ in the most favorable situation ($r_\sigma \ll 1$ and $r_C \ll 1$), *i.e.* the efficiency gain cannot be higher than $\frac{1}{r_C^2}$ or $\frac{1}{r_\sigma^2}$. In the example of Sec. 3 $G_{A_{n,n_Y^*}} \approx \frac{1}{r_C^2}$ because $r_\sigma \ll r_C$.

[2] The user should be able to make a first guess $r_{C,guess}$ concerning r_C and $r_{\sigma,guess}$ concerning r_σ as this is the starting-point of the present note. We then suggest to retain $n_{Y,1} = \frac{1}{r_{\sigma,guess} r_{C,guess}}$ and $n_{Y,2} = \max\left(1, \frac{n_{Y,1}}{100}\right)$, increasing n_1 and n_2 until σ_1 and σ_2 are below 10% of the estimated quantity.

[3] This leads to $n^* = \frac{\tilde{\sigma}_Y^2}{\sigma_r^{*2} A^2} r_\sigma\left(r_\sigma + r_C\right)$, where $\tilde{\sigma}_Y^2 \approx \frac{n_2\sigma_1^2 - n_1\sigma_2^2}{\frac{1}{n_{Y,2}} - \frac{1}{n_{Y,1}}}$ can be deduced from the two Monte Carlo runs already used for r_σ and r_C.

[4] The corresponding computer implementation is freely available online at http://edstar.lmd.jussieu.fr/codes.

computation requires to sample the orientation of the particle, its size and two positions on its projected surface. Here we want to take into account the variability of the shapes of the particles (thinking, *e.g.* of biological cells), which corresponds to additionally sampling the shape. This is associated with a high computational cost (because it requires to mesh the bounding surface of the particle) whereas it contributes only little to the total variance. As simple example we here consider a spheroidal shape [5], that is defined by a single parameter: its elongation R that we assume to be log-normally distributed (median $\bar{R}$ and width parameter s_R). We used the Monte Carlo reformulation (Eq. 4) where the unconditioned random variable is the elongation of the spheroid (*i.e.* $X = R$) and where the conditioned one contains its orientation, its size and two positions on its projected surface (*i.e.* $\vec{Y} = (\Theta_o, R_{eq}, \rho_1, \Phi_1, \rho_2, \Phi_2)$ with the notations of [4]), to compute a Monte Carlo estimator of $W_s(\theta_s)$, demanding a relative standard error $\sigma_r^* = 0.01$.

Results and analysis. Two runs of the Monte Carlo algorithm with $n_1 = 10^3$, $n_{Y,1} = 10^4$, $n_2 = 10^5$ and $n_{Y,2} = 10^2$ permitted to estimate $\tilde{\sigma}_Y \approx 1.3 \cdot 10^3\ \mu m^{-2}\ sr^{-1}$, $r_\sigma \approx 1.2 \cdot 10^{-3}$ and $r_C \approx 0.1$ (Eq. 10), leading to $n_Y^* \approx 8 \cdot 10^3$ and $n^* \approx 2 \cdot 10^4$ (Eq. 8). The expected efficiency gain is therefore $G_{A_{n,n_Y^*}} \approx 100$ (Eq. 9). We then computed the Monte Carlo estimator and found $\hat{W}_s(\theta_s) = 10.5 \pm 0.1\ \mu m^{-2}\ sr^{-1}$ (which is compatible with the imposed relative standard error of 1%). The computation time was 7 min (on a MacBook Pro, 2. GHz Intel Core i5 processor, without using parallelization). Note that, with the plain Monte Carlo algorithm, the same computation would take 11 h. We then carried out a sensitivity study to understand the impact of the precision of the estimated value of r_C (denoted $r_{C,estim}$) on the obtained Monte Carlo efficiency gain $G_{A_{n,n^*_{Y,estim}}}$ [6]. This sensitivity study is interesting because a higher required precision signifies a longer computation time for the two runs of the Monte Carlo algorithm used to compute $r_{C,estim}$. When $n^*_{Y,estim}$ is different from the optimal value n_Y^*, then $G_{A_{n,n^*_{Y,estim}}}$ is lower than the maximal possible efficiency gain $G_{A_{n,n_Y^*}}$ (Eq. 9). However, Fig. 1 (left panel), that displays the ratio of $G_{A_{n,n^*_{Y,estim}}}$ to $G_{A_{n,n_Y^*}}$ as function of the exact r_C, shows that the loss of efficiency is less than 1% if the error on the estimation of r_C is less than a factor 2. Therefore it is not necessary to have a very precise estimation of r_C [7]. We finally applied our technique to a parametric study. In such a context, recalculating n_Y^* for each parameter value can be quite time-consuming. Therefore we have tested the impact of using the same estimated value $n^*_{Y,estim}$ (*e.g.*, the value of n_Y^* for the average value of the parameter) on the resulting Monte Carlo efficiency. We used as parameter the imaginary part κ_r of the refraction index of the spheroid, and chose to take $n^*_{Y,estim} = n_Y^*(\kappa_r = 4 \cdot 10^{-3}) = 8 \cdot 10^3$. Fig. 1 (mid panel) again displays the ratio of $G_{A_{n,n^*_{Y,estim}}}$ to $G_{A_{n,n_Y^*}}$, showing that the loss of efficiency is always lower than 0.5% in the considered parameter range. Therefore it is here pertinent to use the same value of n_Y^* for the whole parameter study. Finally, Fig. 1 (right panel) displays the results of the parametric study with $n = 2 \cdot 10^4$ Monte Carlo samples.

4. Conclusion

The Monte Carlo efficiency technique presented in this Short Note consists in sampling n_Y realizations of the conditioned random variable Y for each realizations of the first, unconditioned random variable X. An automatic procedure to determine the optimal number of samples n_Y is provided. This technique might be extended in different directions:

- Considering n_Y as a function of x (note that the expression of the Monte Carlo estimator A_{n,n_Y} given by Eq. 4 remains valid), or combining the efficiency increase technique with stratified sampling, allows to further increase the efficiency in situations where the convergence is not uniform over $\mathcal{D}_X$ (*e.g.*, in the above example the convergence is slower for strong elongations R, therefore the efficiency will be increased by using a higher value of n_Y in these zones).

[5] The choice of the spheroid is also motivated by the fact that our computational tools permitting to address complex shapes are still under development. For a spheroid, all geometrical computations are analytical and therefore no mesh is required. To obtain realistic values of the Monte Carlo efficiency, the computational cost associated to the generation of the mesh – estimated 100 times greater than the cost associated with the sampling of all the other variables – is simulated by an informatic loop.

[6] Eq. 7 permits to write $G_{A_{n,n^*_{Y,estim}}} = \frac{\epsilon(n^*_{Y,estim})}{\epsilon(n_Y=1)} = \frac{n^*_{Y,estim}\left(1+r_\sigma^2\right)\left(1+r_C^2\right)}{\left(1+n^*_{Y,estim}r_\sigma^2\right)\left(1+n_Y r_C^2\right)}$ where r_C is the exact r_C and $n^*_{Y,estim}$ is the value of n_Y^* calculated with Eq. 8 on the basis of the estimated $r_{C,estim}$.

[7] Note that the sensitivity to an error on the estimation of r_σ is the same as for r_C because n_Y^* is a symmetric function of both ratios (Eq. 8).

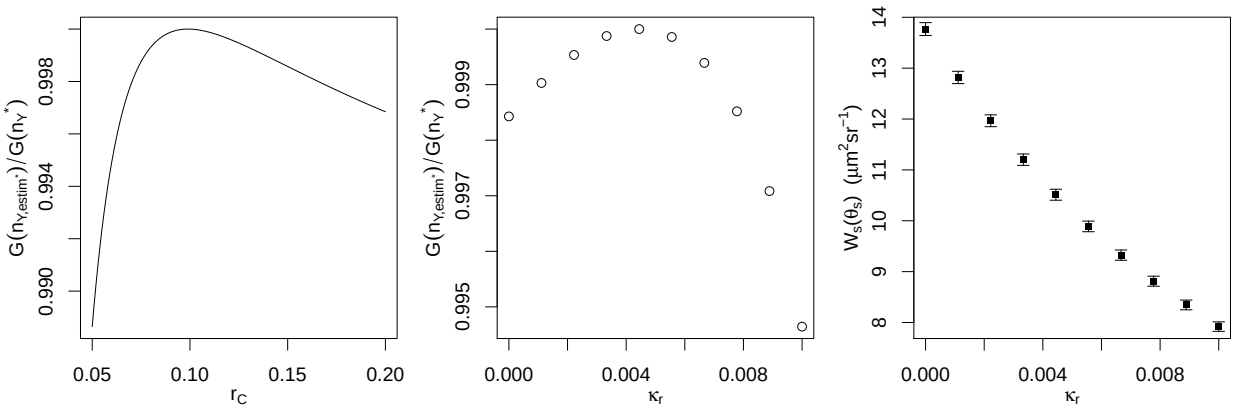

Figure 1: Test case described in Sec. 3. Left panel: ratio of the obtained efficiency gain $G(n^*_{Y,estim})$ (where $n^*_{Y,estim}$ was computed using the estimated value $r_{C,estim} = 0.1$) to the maximal possible efficiency gain $G(n^*_Y)$ (obtained with the exact value r_C) as function of r_C. Mid panel: ratio of the obtained efficiency gain $G(n^*_{Y,estim})$ (where $n^*_{Y,estim} = 8 \cdot 10^3$ was computed for $\kappa_r = 4 \cdot 10^{-3}$) to the maximal possible efficiency gain $G(n^*_Y)$ (obtained with the value n^*_Y corresponding to κ_r) as function of κ_r. Right panel: differential scattering cross section $W_s(\theta_s)$ as function of κ_r (for $n = 2 \cdot 10^4$). Error bars indicate the Monte Carlo standard error σ_A. The used parameter values are $\theta_s = 0.244$, $\bar{R} = 0.5$, $s_R = 1.2$, $\bar{r}_{eq} = 2.38\ \mu m$, $s = 1.18$, $k_e = 14.0\ \mu m^{-1}$, $n_r = 1.08$ and, for the left panel, $\kappa_r = 4.10^{-3}$.

- Distinguishing more than two random variables (*e.g.* X, Y and Z) and sampling Z n_Z times per sample of Y, which is itself sampled n_Y times per sample of X, might permit to obtain even further efficiency increase for calculating higher-dimensional integrals.

- Because of its genericity and simplicity, the technique could be included in software making use of Monte Carlo integration or in common computational libraries (*e.g.*, the GNU Scientific Library (GSL) that already contains Monte Carlo integration routines with automated importance or stratified sampling). In this regard, the computation of n^*_Y (the optimal number of samples of Y per sample of X) can be easily automated (as a pre-computation) because it only requires two runs of the same Monte Carlo algorithm as the one used to evaluate the estimator A_n (only the number of samples n and n_Y changes).

Acknowledgements

This work benefited from fruitful input during collective work at the 2015 meeting of the ZirCon research group in Clermont-Ferrand. Moreover, it has been sponsored by the French government research-program "Investissements d'avenir" through the project ALGUE of the "IDEX Actions Thématiques Stratégiques" (ANR-11-IDEX-002-02), the IMobS3 and SOLSTICE Laboratories of Excellence (ANR-10-LABX-16-01 and ANR-10-LABX-22-01), by the European Union through the program "Regional competitiveness and employment" 2007-2013 (ERDF Auvergne region), and by the Auvergne region. It is also founded by the CNRS through the PIE program PHOTORAD (2010-11) and the PEPS program "Intensification des transferts radiatifs pour le développement de photobioréacteurs a haute productivité volumique" (2012-13).

[1] M. H. Kalos, P. A. Whitlock, Monte Carlo Methods, John Wiley & Sons, 2008.
[2] W. L. Dunn, J. K. Shultis, Exploring Monte Carlo Methods, Elsevier, 2011.
[3] C. Lemieux, Monte Carlo and quasi-Monte Carlo sampling, Springer Science & Business Media, 2009.
[4] J. Charon, S. Blanco, J.-F. Cornet, J. Dauchet, M. El Hafi, R. Fournier, M. Kaissar Abboud, S. Weitz, Monte carlo implementation of schiff's approximation for estimating radiative properties of homogeneous, simple-shaped and optically soft particles: application to spheroidal and cylindrical photosynthetic micro-organisms, Journal of Quantitative Spectroscopy and Radiative Transfer.