Journal of Computational Physics 227 (2008) 9463–9476

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Efficient, automated Monte Carlo methods for radiation transport

Rong Kong [a], Martin Ambrose [a], Jerome Spanier [a,b,*]

[a] Claremont Graduate University, 150 E. 10th Street, Claremont, CA 91711, United States
[b] Beckman Laser Institute and Medical Clinic, University of California, 1002 Health Science Road E., Irvine, CA 92612, United States

ARTICLE INFO

Article history:
Received 26 June 2007
Received in revised form 23 June 2008
Accepted 27 June 2008
Available online 23 July 2008

Keywords:
Transport equation
Geometrically convergent Monte Carlo algorithms
Computational efficiency

ABSTRACT

Monte Carlo simulations provide an indispensible model for solving radiative transport problems, but their slow convergence inhibits their use as an everyday computational tool. In this paper, we present two new ideas for accelerating the convergence of Monte Carlo algorithms based upon an efficient algorithm that couples simulations of forward and adjoint transport equations. Forward random walks are first processed in stages, each using a fixed sample size, and information from stage k is used to alter the sampling and weighting procedure in stage $k+1$. This produces rapid geometric convergence and accounts for dramatic gains in the efficiency of the forward computation. In case still greater accuracy is required in the forward solution, information from an adjoint simulation can be added to extend the geometric learning of the forward solution. The resulting new approach should find widespread use when fast, accurate simulations of the transport equation are needed.

1. Introduction

Monte Carlo (MC) simulations have provided a "gold standard" of computational support for many important problems of science and engineering that are modeled using the radiative transport equation (RTE). However, when the method is conventionally applied, according to the central limit theorem, the number of samples processed must be multiplied roughly a hundredfold to gain a decimal digit of accuracy. This slow convergence has encouraged the use of faster, but sometimes considerably less accurate methods (see [1–3]). Many applications areas that rely on transport models would benefit from the availability of faster MC methods.

In this paper, we describe a new MC method, based on sequential application of correlated sampling [4,7], that achieves geometric convergence for very general transport problems. Coupling such simulations of the RTE with simulations of an adjoint RTE leads to an automated, highly efficient MC solution algorithm that "tunes" itself to the specific needs of each RTE problem and requires minimal or no user intervention. In other papers, we will describe a similar algorithm based on importance sampling [8–13,4,5] and we will establish the geometric convergence of each. Here we illustrate the power of our new method by applying it to an example that represents a model tissue problem. Comparison of the numerical results with analytic solutions suggests that gains of computational efficiency by several orders of magnitude in comparison with conventional Monte Carlo should be possible when the algorithms are optimized for more complex, practical problems.

* Corresponding author. Address: Beckman Laser Institute and Medical Clinic, University of California, 1002 Health Science Road E., Irvine, CA 92612, United States. Tel.: +1 949 824 3419; fax: +1 949 824 6969.
E-mail address: jspanier@uci.edu (J. Spanier).

0021-9991/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2008.06.037

2. Forward and adjoint transport equations

The methods we describe in this paper can be applied quite generally to RTE problems involving full spatial, angular, energy and time dependence. However, to simplify both the notation and the exposition, we specialize here to time-independent, single-speed radiation transport for which the accepted model is the integro-differential equation[1]

$$\mathbf{\Omega}\cdot\nabla\Psi(\mathbf{r},\mathbf{\Omega})+\sigma_t(\mathbf{r})\Psi(\mathbf{r},\mathbf{\Omega})=\sigma_s(\mathbf{r})\int_{S^2}\Psi(\mathbf{r},\mathbf{\Omega}')p(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}')\mathrm{d}\mathbf{\Omega}'+Q(\mathbf{r},\mathbf{\Omega}), \tag{1}$$

where the integration is over the sphere of unit direction vectors, S^2. Eq. (1) is assumed to be valid for all vectors $\mathbf{r}$ in the interior of a closed, bounded subregion V of $\mathcal{R}^3$. The solution, $\Psi(\mathbf{r},\mathbf{\Omega})$, describes the radiation intensity at any point $(\mathbf{r},\mathbf{\Omega})$ in the phase space Γ due to a radiation source Q internal to V. In this equation, the coefficient functions $\sigma_t(\mathbf{r}),\sigma_s(\mathbf{r})$ and $p(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}')$ characterize the physical transport, scattering and absorption of radiation in Γ. The RTE simply expresses a balance between radiation arriving at $(\mathbf{r},\mathbf{\Omega})$ (the two terms on the right hand side) and removal (the two terms on the left hand side). Arrivals are due either to the source Q or scattering from direction $\mathbf{\Omega}'$ to direction $\mathbf{\Omega}$ while removal occurs by streaming (pure transport), expressed by the gradient term, and through interactions at $(\mathbf{r},\mathbf{\Omega})$ that produce either absorption there or scattering away from $\mathbf{\Omega}$.

A unique solution $\Psi(\mathbf{r},\mathbf{\Omega})$ is assured for all $\mathbf{r}\in V,\mathbf{\Omega}\in S^2$ when the flux of radiation $\Psi_{\mathrm{inc}}(\mathbf{r},\mathbf{\Omega})$ incident on ∂V from *outside of V* is specified; that is, for unit directions $\mathbf{\Omega}$ for which $\mathbf{\Omega}\cdot\mathbf{n}_{\partial V}<0$, where $\mathbf{n}_{\partial V}$ is the unit outward normal vector on ∂V. As discussed in [6, pp. 20–30], the solution $\Psi(\mathbf{r},\mathbf{\Omega})$ may then be expressed in terms of a volume Green's function, $G[(\mathbf{r},\mathbf{\Omega})\leftarrow(\mathbf{r}_0,\Omega_0)]$, and a surface Green's function, $G_{\partial V}[(\mathbf{r},\mathbf{\Omega})\leftarrow(\mathbf{r}_{\partial V},\mathbf{\Omega}_0)]$:

$$\Psi(\mathbf{r},\mathbf{\Omega})=\int_{V\times S^2}G[(\mathbf{r},\mathbf{\Omega})\leftarrow(\mathbf{r}_0,\mathbf{\Omega}_0)]Q(\mathbf{r}_0,\mathbf{\Omega}_0)\mathrm{d}\mathbf{r}_0\,\mathrm{d}\mathbf{\Omega}_0+\int_{\partial V\times S^2_-}G_{\partial V}[(\mathbf{r},\mathbf{\Omega})\leftarrow(\mathbf{r}_{\partial\mathbf{V}},\mathbf{\Omega}_0)]\Psi_{\mathrm{inc}}(\mathbf{r}_{\partial V},\mathbf{\Omega}_0)\mathrm{d}\mathbf{r}_{\partial V}\,\mathrm{d}\mathbf{\Omega}_0 \tag{2}$$

$$\text{and}\quad S^2_-=\{\mathbf{\Omega}_0:\mathbf{\Omega}_0\cdot\mathbf{n}_{\partial V}<0\},$$

where $Q(\mathbf{r},\mathbf{\Omega})$ is the volume source within V. The second term in Eq. (2) comes from the boundary condition

$$\Psi(\mathbf{r}_{\partial\mathbf{V}},\Omega_{\mathbf{0}})=\Psi_{\mathrm{inc}}(\mathbf{r}_{\partial V},\mathbf{\Omega}_0)\ \text{for}\ \mathbf{r}_{\partial V}\in\partial V\quad\text{and}\quad\mathbf{\Omega}_0\in S^2_-. \tag{3}$$

Introduction of one or more radiation detectors into the physical system is described by a response function $Q^*(\mathbf{r},\mathbf{\Omega})$ that characterizes both the location and physical properties of the detecting instruments. This leads to an adjoint integro-differential equation

$$-\nabla\cdot\mathbf{\Omega}\Psi^*(\mathbf{r},\mathbf{\Omega})+\sigma_t(\mathbf{r})\Psi^*(\mathbf{r},\mathbf{\Omega})=\sigma_s(\mathbf{r})\int_{S^2}\Psi^*(\mathbf{r},\mathbf{\Omega}')p(\mathbf{r},\Omega\leftarrow\mathbf{\Omega}')\mathrm{d}\mathbf{\Omega}'+Q^*(\mathbf{r},\mathbf{\Omega}), \tag{4}$$

whose solution can be interpreted as an importance function in the sense that it is proportional to the response of the detectors from a unit source at $(\mathbf{r},\mathbf{\Omega})$ [4,5,18]. It is also understood that the boundary condition satisfied by Ψ^* on ∂V is dual[2] to that specified for Ψ. Thus, if the region external to V is either nonreentrant or a pure reflector, the boundary conditions for Ψ^* will be dual to those for Ψ (since then either the product $\Psi\Psi^*$ vanishes on ∂V or it satisfies a reflecting boundary condition there). Also, if the boundary segments include points at which both partially reflecting and partially nonreentrant transmission occurs (as is the case for light transport where the Fresnel and Snell's laws are applicable), duality also holds, as was shown by Aronson in [35]. The theory of reciprocity for the RTE makes it clear that either simulations of the forward equation, Eq. (1), or the adjoint RTE, Eq. (4), may be used to estimate the response of the detector Q^* due to the source Q, since this response can be expressed equivalently as a linear functional of either solution:

$$\int_{V\times S^2}\Psi(\mathbf{r},\mathbf{\Omega})Q^*(\mathbf{r},\mathbf{\Omega})\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{\Omega}=\int_{V\times S^2}\Psi^*(\mathbf{r},\mathbf{\Omega})Q(\mathbf{r},\mathbf{\Omega})\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{\Omega}. \tag{5}$$

Approximate solutions of Eq. (4) also play a crucial role in the variance reduction method known as importance sampling, mentioned earlier.

The integro-differential RTE, Eq. (1), can be converted to an equivalent integral equation

$$\Psi(\mathbf{r},\mathbf{\Omega})=\int_{V\times S^2}\Psi(\mathbf{r}',\mathbf{\Omega}')K(\mathbf{r},\mathbf{\Omega};\mathbf{r}',\mathbf{\Omega}')\mathrm{d}\mathbf{\Omega}'\,\mathrm{d}\mathbf{r}'+S(\mathbf{r},\mathbf{\Omega}) \tag{6}$$

that is more directly linked to the MC probability model on which the simulation is based. The kernel K of (6) describes both the scattering and the transport of radiation and the source S is defined by moving particles from their point $(\mathbf{r},\mathbf{\Omega})$ of origination (as determined by sampling Q) to their first collision location; details may be found in [4,5,18].

We introduce a new dependent variable

[1] Here and throughout the paper we will display only a single integral sign for all integrals, whether they are one- or multi-dimensional. The number and kind of integration variables and the region of integration are shown explicit to avoid possible ambiguity.

[2] Duality here means that the integral $\int_{\partial V\times S^2}n_{\partial V}\cdot\mathbf{\Omega}\Psi(\mathbf{r},\mathbf{\Omega})\Psi^*(\mathbf{r},\mathbf{\Omega})\mathrm{d}r\,\mathrm{d}\Omega=0$.

$$C(\mathbf{r},\mathbf{\Omega}) = \Psi(\mathbf{r},\mathbf{\Omega})\Psi^*(\mathbf{r},\mathbf{\Omega}) \tag{7}$$

and call the product function C an *information density function*. In [29–33], this function is sometimes called a *response* or *contributon function*, and it was used in the early literature to study RTE problems, mainly of nuclear radiation shielding type – problems characterized by their focus on events with low probability outcomes in a simulation.

In the approach presented in this paper, we will use this function to capture the relative "information value" of points in the phase space. The function $C(\mathbf{r},\mathbf{\Omega})$ satisfies the RTE [30]

$$\nabla\cdot\Omega C(\mathbf{r},\mathbf{\Omega}) + \Sigma_s(\mathbf{r},\mathbf{\Omega})C(\mathbf{r},\mathbf{\Omega}) = \int_{S^2}\Sigma(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}')C(\mathbf{r},\mathbf{\Omega}')\mathrm{d}\mathbf{\Omega}' + Q\Psi^* - Q^*\Psi, \tag{8}$$

where

$$\Sigma_s(\mathbf{r},\mathbf{\Omega}) = \int_{S^2}\Sigma(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}')\mathrm{d}\mathbf{\Omega}' \tag{9}$$

and

$$\Sigma(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}') = \sigma_s(\mathbf{r})p(\mathbf{r},\mathbf{\Omega}\leftarrow\mathbf{\Omega}')\frac{\Psi^*(\mathbf{r},\mathbf{\Omega})}{\Psi^*(\mathbf{r},\mathbf{\Omega}')}. \tag{10}$$

Eq. (8) is a radiative transport equation for C that couples the direct and adjoint solutions Ψ,Ψ^*. The boundary conditions for C follow from those that apply for Ψ,Ψ^*. Thus, at a point of nonreentrant transmission clearly $C=0$ while at a point of pure reflection, the normal derivative of C must vanish. The same argument applies if the boundary conditions include points at which *both* reflection *and* nonreentrant transmission occur (as when light is refracted at the boundary).

Eq. (8) implies that there is no absorption of information density,[3] only scattering, and that the scattering depends on *both* the entering and outgoing directions (through the factor $\frac{\Psi^*(\mathbf{r},\mathbf{\Omega})}{\Psi^*(\mathbf{r},\mathbf{\Omega}')}$). This is in contrast with most models of conventional RTE problems for which the scattering is rotationally invariant and therefore depends only on the scalar product between the two directions. Also, there is no loss of information density at the boundaries provided that the adjoint solution, $\Psi^*(\mathbf{r},\mathbf{\Omega})$, satisfies boundary conditions that are dual to those satisfied by $\Psi(\mathbf{r},\mathbf{\Omega})$, as we have assumed. Finally, there is both a source density (in the term $Q\Psi^*$) and a sink (in the term $-Q^*\Psi$).

Each source-detector pair identifies a unique contributon function $C(\mathbf{r},\mathbf{\Omega})$ that characterizes the flow of information from the source to the detector without any losses along the way. These observations will be utilized in the strategy we present later for accelerating the convergence of MC simulations of the RTE.

3. First generation (G1) adaptive zero variance algorithms

In 1962, Halton [27] proposed using Monte Carlo algorithms to solve matrix problems by iteratively applying one of several variance reduction methods, including correlated sampling. More recently, Halton's ideas were extended to the solution of continuous radiation transport problems by researchers at Los Alamos National Laboratory [14–17] and the Claremont Graduate University [19–26]. The basic idea underlying these methods is to process the random walks in batches, called stages, consisting of W independent random walks each and alter the sampling and weighting methods for each new stage by incorporating information "learned" during the previous stage. It has been shown that these methods all produce geometric convergence [19–26], i.e.

$$E_k < \lambda E_{k-1} < \lambda^k E_0, \quad 0<\lambda<1 \; k = \text{stage number}, \tag{11}$$

where $E_k = k$th stage error; e.g. $E_k = \|\Psi(\mathbf{P}) - \widehat{\Psi}^{(k)}(\mathbf{P})\|$,[4] where Ψ = exact transport solution, $\widehat{\Psi}^{(k)} = k$th stage adaptive estimate and $\|\cdot\|$ denotes an appropriate error norm.

These adaptive zero variance MC algorithms for global solutions of transport equations make use of expansions of either the solution $\Psi(\mathbf{P}) = \sum_{i=1}^{\infty} a_i B_i(\mathbf{P})$ or the dual solution, Ψ^*, in orthonormal basis functions, $B_i(\mathbf{P})$, and produce arbitrarily accurate truncated solutions

$$\widehat{\Psi}(\mathbf{P}) \equiv \sum_{i=1}^{M} a_i B_i(\mathbf{P}) \approx \Psi(\mathbf{P}) \tag{12}$$

of Ψ. This is done by estimating the first M expansion coefficients

$$a_i = \int B_i(\mathbf{P})\Psi(\mathbf{P})\mathrm{d}\mathbf{P}, \quad i=1,\ldots,M \tag{13}$$

[3] This follows since the second term on the left-hand side of Eq. (8), which is the removal term due to *either* scattering *or* absorption, expresses just the removal due to scattering alone.

[4] To abbreviate the notation, we have set $\mathbf{P} = (\mathbf{r},\mathbf{\Omega})$.

in adaptive stages of ever-increasing accuracy. Because each coefficient a_i is just a weighted integral of the solution Ψ, it is readily estimated by conventional Monte Carlo methods. Further discussion of the algorithm employed can be found in the Appendix, as well as in [22].

To understand how the information from stage k is linked to that in stage $k+1$, we outline our G1 method based on repeated use of correlated sampling (called Sequential Correlated Sampling or SCS). We introduce the linear integral operator $\mathcal{K}$ defined by

$$\mathcal{K}\Psi(\mathbf{P}) = \int_\Gamma K(\mathbf{P},\mathbf{P}')\Psi(\mathbf{P}')d\mathbf{P}' \tag{14}$$

so that Eq. (6) becomes

$$\Psi(\mathbf{P}) = \mathcal{K}\Psi(\mathbf{P}) + S(\mathbf{P}). \tag{15}$$

Next define a "reduced" source for the adaptive stage $k+1$ in terms of the reduced source at stage k by,

$$S^{(k+1)}(\mathbf{P}) = S^{(k)}(\mathbf{P}) + \mathcal{K}\tilde{\psi}^{(k)}(\mathbf{P}) - \tilde{\psi}^{(k)}(\mathbf{P}), \quad S^{(0)}(\mathbf{P}) = S(\mathbf{P}), \tag{16}$$

where $\tilde{\psi}^{(k)}(\mathbf{P})$ is the approximate correction obtained in stage k to the truncated solution (12).[5] The algorithm suggested by Eq. (16) is implemented by initiating random walks in adaptive stage k using the reduced source function $S^{(k)}$ and selecting all subsequent collision sites using the continuous kernel K. The kth stage correction produced by this method decreases (with probability 1) in magnitude as k increases. Proof of the geometric convergence of this method for the family of problems studied in this paper is found in [23], while a detailed description of the algorithm was published in [22] and is also reviewed in the Appendix.

These "first generation" (G1) adaptive methods and algorithms perform spectacularly well on RTE problems in a few independent variables. We content ourselves here with illustrating this by solving a simply described family of model transport problems in a homogeneous slab of finite thickness T which is infinite in extent in the other two dimensions. The resulting problem, when further specialized to scattering that is either directly forward or backward, becomes tractable analytically and therefore provides an ideal problem with which to test our Monte Carlo algorithms. While sufficiently simple to provide closed form solutions for both the means and the variances of our Monte Carlo estimators [36], by varying the slab thickness, the relative amounts of absorption and scattering, and the fractions of forward and backward scattering, this family of slab problems includes examples that pose severe challenges for conventional Monte Carlo simulations and thus provides useful tests of new ones, like the one described here.

The transport problem we treat can be described as a pair of differential equations

$$\begin{cases} \frac{d\Psi_1}{dx} + \sigma_t\Psi_1 = \sigma_s(p_{11}\Psi_1 + p_{12}\Psi_2) + Q_1(x), & 0 < x < T, \\ -\frac{d\Psi_2}{dx} + \sigma_t\Psi_2 = \sigma_s(p_{21}\Psi_1 + p_{22}\Psi_2) + Q_2(x), & 0 \leqslant x < T, \\ \Psi_1(0) = Q_0, \Psi_2(T) = Q_T, \end{cases} \tag{17}$$

where

$$\begin{aligned} &Q_0, Q_T \geqslant 0, \quad Q_1(x), Q_2(x) \geqslant 0, \\ &\sigma_t \geqslant \sigma_s \geqslant 0, \\ &p_{11}, p_{12}, p_{21}, p_{22} \geqslant 0, \\ &p_{11} + p_{21} = 1, \quad p_{12} + p_{22} = 1. \end{aligned} \tag{18}$$

The constants σ_s, σ_t describe the scattering and total macroscopic cross sections, respectively, and p_{ij} is the probability of scattering from direction j to i, where $i = j = 1$ corresponds to motion from left to right in the slab while $i = j = 2$ corresponds to motion from right to left. The function Ψ_1 thus describes left-to-right-moving radiation and Ψ_2 describes right-to-left-moving radiation. Formulas for the solutions of such problems may be found, for example, in [36].

The problem chosen here describes a tissue optics application in which photons are introduced through a laser light source at one end of 1 cm of tissue with typical physical characteristics that are described by the input data

$$\begin{aligned} &\sigma_a = 0.01/\text{mm}, \quad \sigma_s = 0.99/\text{mm}, \quad T = 10 \text{ mm} \\ &p_{11} = p_{22} = 0.985, \quad p_{12} = p_{21} = 0.015 \\ &Q_0 = 1.0, \quad Q_T = 0, \quad Q_1(x) = Q_2(x) \equiv 0. \end{aligned}$$

[5] Legendre polynomials in each independent variable make an obvious choice of the basis functions $B_i(P)$ to begin our investigations when the variable is defined over any finite interval, just as Laguerre and Hermite polynomials would be appropriate over semi-infinite or doubly infinite intervals, respectively. We have also experimented with other choices of complete orthogonal systems with results similar to those presented in this paper. In fact, no "universal" choice of basis functions will be adequate for all transport problems unless it is essentially constructed from the eigensystem of each RTE problem, which is clearly impractical for all but the simplest transport problems.

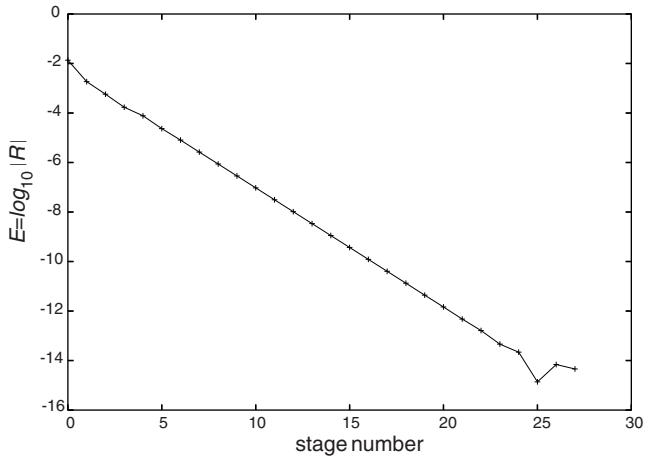

Fig. 1. Geometric convergence for the G1 algorithm.

The scattering of light in tissue is modeled by a very forward-peaked phase function typical for tissue. Here the average cosine of the scattering angle is chosen to be 0.97, for example.

The G1 solution method used in this example represents the approximate solution as a linear combination of the first 25 Legendre polynomials and the simulated measurement is represented in this case by averaging the solution[6] over the final mm of tissue. That is, we chose the function Q^* in Eq. (4) as

$$Q^*(x) = \begin{cases} 1 & \text{for } 9 \leqslant x \leqslant 10 \\ 0 & \text{otherwise} \end{cases}$$

so that Eq. (5) becomes

$$\int_9^{10} [\Psi_1(x) + \Psi_2(x)]\mathrm{d}x = .08030441,$$

which is the average value of the radiation intensity in the final mean free path of tissue. Fig. 1 plots $E = \log_{10}|R|$ versus the number of adaptive stages, where the relative error, R, is

$$R = \frac{\int_9^{10}[\Psi_1(x) + \Psi_2(x)]\mathrm{d}x - \int_9^{10}[\widetilde{\Psi}_1(x) + \widetilde{\Psi}_2(x)]\mathrm{d}x}{\int_9^{10}[\Psi_1(x) + \Psi_2(x)]\mathrm{d}x}.$$

Fig. 1 clearly exhibits a geometric decrease in E as the number of adaptive stages increases. The geometric convergence for this G1 algorithm ceases as the limit $\widehat{\Psi}$ (Eq. (12)) is approached. Here $\widetilde{\Psi}$ is the stochastic approximation to $\widehat{\Psi}$ determined by the SCS algorithm (see Eq. (A.13) in the Appendix).

Values of E indicate the number of significant digits of accuracy in the estimate. The results in Fig. 1 were obtained using 10,000 random walks in each adaptive stage. Thus, the 25 stages required a total of 250,000 random walks to produce an essentially exact result, and all 25 stages required just over 32 min of computation on a 1.6 GHz Xeon E5310 10 quad core processor.

While near machine precision (i.e. relative error on the order of 10^{-15}) is achievable with these G1 algorithms for sufficiently simple transport problems, when more difficult problems (involving more independent variables and/or severe spatial heterogeneity) are solved by these methods, their performance deteriorates. Degradation of the quality of practically achievable G1 adaptive results occurs because: (1) The need to generate expansion coefficients for each independent variable means that the computational burden grows exponentially with $D = \dim(\Gamma)$, Γ = phase space, making problems in 5 or 6 independent variables quite challenging. (2) Severely heterogeneous problems further degrade performance by requiring separate expansions in each homogeneous subregion, further adding to the total number of coefficients needed to describe the solution. (3) The error caused by truncating each expansion after a finite number of terms in each variable and each subregion is very difficult to estimate accurately (and therefore difficult to control). That is, it is difficult to predict in advance how many terms to retain in the expansion of each independent variable in order to achieve a particular precision in the solution. (4) Perhaps the greatest source of unpredictability about the error arises from the need to approximate numerically

[6] In general, the measurement is represented as a weighted integral of the RTE solution. For this simple illustration, we chose the weighting function to be the characteristic function associated with the detector region divided by the volume of that region. The measured quantity then becomes the average value of the radiation solution Ψ in the detector region.

many integrals that arise in the computations. In SCS, these come from the integration of the reduced source, Eq. (16). All quadratures encountered must be carried out with extremely high precision if the final tally estimates are to have essentially unlimited accuracy, and this is clearly not a realistic practical objective.

For general RTE applications, then, either some accuracy must be sacrificed or else the cost of the G1 algorithm will increase dramatically. All of these imperfections taken together mean that the G1 algorithms are not yet useful as practical *general-purpose* tools for solving arbitrary transport problems. Nevertheless, they are very effective methods for solving transport problems in one or two dimensions and they serve as useful computational yardsticks against which to measure the performance of other, more approximate methods for more complex problems.

4. Second generation (G2) adaptive low variance methods

Dramatic gains in efficiency can be accomplished by eliminating the requirement that the solution can be represented globally at every point of phase space by an infinite series of basis functions. The new G2 algorithms achieve this by relaxing the requirement of unlimited precision at every location and angular orientation of the phase space. Instead, only the information most essential to determine *accurate*, but not *perfect*, regionwise weighted *averages* of the solution is sought.

Suppose, then, that interest focuses *only* on estimating weighted integrals such as $\int_{V\times S^2}\Psi(\mathbf{r},\Omega)Q^*(\mathbf{r},\boldsymbol{\Omega})\mathrm{d}\mathbf{r}\,\mathrm{d}\boldsymbol{\Omega}$, of the solution with *high precision* (for example, to within 0.1% relative error) over a fixed, but arbitrary decomposition $\Gamma=\cup_{i=1}^{R}\Gamma_i$ of the phase space. We assume that this decomposition includes one or more regions designated as "true" detector regions,[7] and that the sets $\{\Gamma_i\}$ are pairwise disjoint: $\Gamma_i\cap\Gamma_j=\emptyset, i\neq j$.

Our implementation of the G2 algorithm, named ASCS (= Averaged Sequential Correlated Sampling), finds a piecewise constant approximation Ψ_a of Ψ

$$\Psi_a(\mathbf{P})=\frac{1}{|\Gamma_i|}\int_{\Gamma_i}\Psi(\mathbf{P}')d\mathbf{P}'\equiv\alpha_i\quad\text{provided }\mathbf{P}\in\Gamma_i. \tag{19}$$

Now let $\widetilde{\Psi}_a^{(0)}(\mathbf{P})$ denote an initial estimate of $\Psi_a(\mathbf{P})$ obtained from a conventional Monte Carlo simulation consisting of W random walks. For example, $\widetilde{\Psi}_a^{(0)}(\mathbf{P})$ can be obtained by averaging the total distance traveled by all simulated particles in each subregion Γ_i. In [4] it is shown that this provides an unbiased estimate of the solution integral in each region. More general weighted solution integrals can be handled similarly [28]. Replacement of the continuous function $\widetilde{\Psi}$ by the piecewise constant function $\widetilde{\Psi}_a^{(0)}$ in Eq. (16) then produces an appropriate reduced source for the new G2 adaptive algorithm:

$$S_{\mathrm{red}}^{(1)}(\mathbf{P})\equiv S^{(0)}(\mathbf{P})-\widetilde{\Psi}_a^{(0)}(\mathbf{P})+\int_{\Gamma}K(\mathbf{P},\mathbf{P}')\widetilde{\Psi}_a^{(0)}(\mathbf{P}')d\mathbf{P}', \tag{20}$$

$$S_{\mathrm{red}}^{(0)}(\mathbf{P})=S(\mathbf{P}) \tag{21}$$

and in general, the ASCS algorithm based on this idea is characterized by the iterative scheme

$$S_{\mathrm{red}}^{(k+1)}(\mathbf{P})\equiv S_{\mathrm{red}}^{(k)}(\mathbf{P})-\tilde{\psi}_a^{(k)}(\mathbf{P})+\int_{\Gamma}K(\mathbf{P},\mathbf{P}')\tilde{\psi}_a^{(k)}(\mathbf{P}')\mathrm{d}\mathbf{P}',\quad k=0,1,\ldots \tag{22}$$

where the function $\tilde{\psi}_a^{(k)}$ is the correction from stage k to the approximate solution from previous stages. The G2 adaptive algorithm is also described in the Appendix.

We have implemented this algorithm and confirmed that it converges to estimates of the α_i that depend on the coarseness/fineness of the mesh imposed. The geometric learning ceases when the locally constant approximation $\widetilde{\Psi}_a(P)$ of the transport solution Ψ has been stabilized. Because such an approximate solution, which is discontinuous, cannot satisfy the original RTE pointwise (except in the trivial case that the latter is globally constant), the precision achievable is limited by the overall quality of the decomposition of the phase space, Γ; that is, by the variation of the solution over each subregion of the decomposition.

We applied this G2 algorithm to the same problem described in Section 3. In Fig. 2 we track the convergence obtained when 2000 uniformly space subintervals Γ_i are used to subdivide the phase space [0,10]. The graph plots the error E as a function of the number of G2 adaptive stages. Our G2 algorithm generated 20 adaptive stages (including a conventional MC initial stage) to converge. In this example, with a relatively fine decomposition into 2000 subintervals, we obtain nearly four significant digits of accuracy in the solution with an investment in computer cost of a little more than 22 min.

The geometric learning power of this G2 algorithm alone should make possible accurate solution of many RTE problems not currently accessible by conventional MC. However, we would like to be able to increase this accuracy when it is required. To do this, we need to be able to refine an initial decomposition of the phase space Γ in an intelligent way to achieve the accuracy needed. In other words, in case the precision reached when the G2 geometric learning stops is insufficient, we want to be able to extend it by an *appropriate* refinement of the phase space. What is needed, then, is an *automated* strategy for determining which subregions are most important to refine, and by how much, as well as which subregions should be coalesced for maximal computational efficiency. Such a strategy is described in the following section.

[7] The remaining regions of the decomposition will be those for which average RTE solution values will be determined by the algorithm.

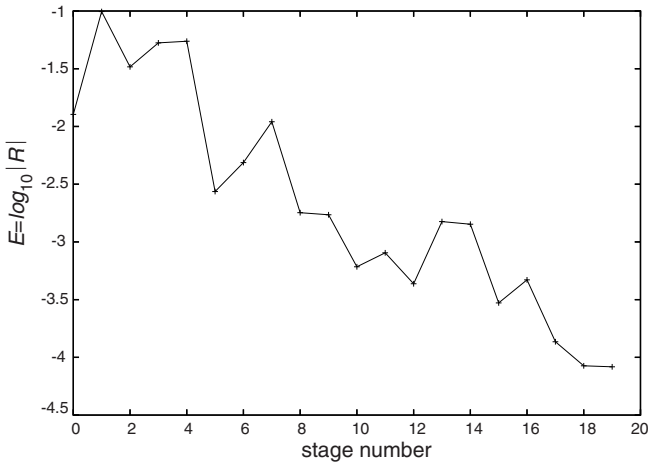

Fig. 2. Geometric convergence for the G2 algorithm.

5. Third generation (G3) phase space refinement methods

The mechanism we propose to exhibit how to refine any phase space decomposition intelligently is to combine information collected from particle trajectories constructed according to the original RTE with information collected from trajectories sampled according to an adjoint RTE. As we pointed out earlier, the function $C(\mathbf{r},\mathbf{\Omega})$ defined by (7)–(10) accurately captures the relative information value of the point $(\mathbf{r},\mathbf{\Omega})$ with respect to the transport of radiation from the source to the detector. Because $C(\mathbf{r},\mathbf{\Omega})$ combines the *intensity* of radiation at $(\mathbf{r},\mathbf{\Omega})$ with the likelihood that radiation at $(\mathbf{r},\mathbf{\Omega})$ *will actually reach* the detector, this function quantifies the data on which intelligent grid refinement should be based.

Of course, obtaining detailed knowledge of the function $C(\mathbf{r},\mathbf{\Omega})$ *pointwise* throughout the phase space poses a daunting problem, even more so than capturing the RTE solution $\Psi(\mathbf{r},\mathbf{\Omega})$ everywhere since $C(\mathbf{r},\mathbf{\Omega})$ obeys a complicated RTE that involves *both* Ψ and Ψ^*. However, we can quite easily estimate *integrals* of $C(\mathbf{r},\mathbf{\Omega})$ over an arbitrary decomposition of Γ by combining information from two G2 algorithm applications, one to obtain an approximate $\Psi(\mathbf{r},\mathbf{\Omega})$ solution and the other an approximate $\Psi^*(\mathbf{r},\mathbf{\Omega})$ solution. These regionwise constant approximations, $\Psi_a(\mathbf{r},\mathbf{\Omega})$ and $\Psi^*_a(\mathbf{r},\mathbf{\Omega})$, can then be multiplied together in each subregion and the resulting approximation $C_a(\mathbf{r},\mathbf{\Omega})$ can be integrated easily to produce the required approximate integrals of $C(\mathbf{r},\mathbf{\Omega})$.

Suppose then that the initial decomposition of Γ consists of R regions: $\Gamma_1,\Gamma_2,\ldots,\Gamma_R$ and let $\Psi_a(\mathbf{r},\mathbf{\Omega})=(\Psi_1,\Psi_2,\ldots,\Psi_R)$ and $\Psi^*_a(\mathbf{r},\mathbf{\Omega})=(\Psi^*_1,\Psi^*_2,\ldots,\Psi^*_R)$ denote the vector solutions obtained by applying the G2 algorithm to the forward RTE (1) and the adjoint RTE (4), respectively. Suppose that the adjoint source Q^* vanishes except in region Γ_i for some unique value of i (so that Γ_i designates the "true" detector position and angular range of this detector). The G2 algorithm then finds a regionwise constant approximation to $\Psi^*(\mathbf{r},\mathbf{\Omega})$: $\Psi^*_a(\mathbf{r},\mathbf{\Omega})\equiv\Psi^*_{i,a}(\mathbf{r},\mathbf{\Omega})=(\Psi^*_{i,1},\Psi^*_{i,2},\ldots,\Psi^*_{i,R})$ whose jth component represents the average value of $\Psi^*_i(\mathbf{r},\mathbf{\Omega})$ over $(\mathbf{r},\mathbf{\Omega})$ in Γ_j. Similarly, the G2 solution for Ψ then finds a regionwise constant approximation $\Psi_a(\mathbf{r},\mathbf{\Omega})=(\Psi_1,\Psi_2,\ldots,\Psi_R)$ whose jth component represents the average value of $\Psi(\mathbf{r},\mathbf{\Omega})$ over $(\mathbf{r},\mathbf{\Omega})$ in Γ_j. The product $C_{i,j}=\Psi_j\Psi^*_{ij}$ may then be integrated over Γ_j and interpreted as an estimate of the average information value of that region when estimating the response of a detector placed in region Γ_i. Thus, the ith row of the "information matrix" $C_{i,j}$ provides the raw data to be used for intelligent grid refinement with respect to a detector in Γ_i.

We next describe a simple refinement strategy based on these ideas that we applied to the slab transport problem described earlier (Eqs. (17) and (18)). After completing the G2 phase with 20 stages to compute a piecewise constant approximation $\Psi_a(\mathbf{r},\mathbf{\Omega})=(\Psi_1,\Psi_2,\ldots,\Psi_{2000})$ whose jth component represents the average value of $\Psi(x)$ over the jth subinterval, the G2 algorithm computes a similar piecewise constant approximation $\Psi^*_a(\mathbf{r},\mathbf{\Omega})=(\Psi^*_1,\Psi^*_2,\ldots,\Psi^*_{2000})$ whose jth component represents the average value of $\Psi^*(x)$ over the jth subinterval and the information density function needed for the intelligent mesh refinement strategy is formed from the component-wise product of these two vectors. The G3 algorithm, sketched above and described more fully in the Appendix was applied in this way to produce the output graphed in Fig. 3. The mesh refinement G3 algorithm produced a contributon-based mesh two grid consisting of a total of 7230 nonuniformly distributed subintervals as compared with the 2000 uniform ones making up mesh 1 in the initial G2 phase. Overall, the precision increased by nearly an additional two orders of magnitude.

Comparisons of the efficiency gains that result from the use of the new algorithms are noteworthy for these transport problems. A useful indicator of efficiency in conventional MC implementations is

$$\mathrm{Eff}=\frac{1}{\mathrm{Var}\cdot t}, \tag{23}$$

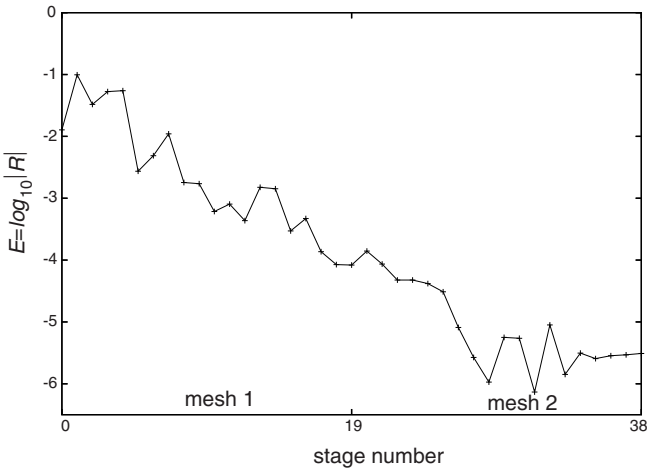

Fig. 3. Geometric convergence for the G3 algorithm.

Table 1
Comparison of efficiencies of each method

Method	*S*	*Est.*	$\|R\|$	σ^2	*t*	*Rel.Eff*
exact	–	.803044	–	–	–	∞
CMC	1	.792834	1.271×10^{-2}	82.549	33	1
G2	20	.803111	8.291×10^{-5}	2.403×10^{-3}	1341	845
G3	39	.803042	3.089×10^{-6}	1.19×10^{-6}	17273	132,529

S = number of adaptive stages, *Est.* = estimate of the average flux in a specified region, $|R|$ = absolute value of the relative error, *t* = simulation run time (s), *Rel.Eff.* = relative efficiency.

where Var is the variance of the estimating random variable and t is the total computer time required to achieve this variance. For conventional MC simulations, this measure[8] is roughly independent of the number N of random samples processed since t is linear in N and Var is inversely proportional to N, and therefore to t. However, our adaptive algorithms are designed to produce variances that decrease *exponentially* with time. Thus, to compare computational efficiency of these new methods with conventional Monte Carlo requires that we examine the amount of error reduction attained by each of the more powerful algorithms and compute the run time that would be necessary for conventional Monte Carlo to achieve this amount of additional error reduction. This last computation is predicated on the assumption that the conventional simulations converge at the rate predicted by the central limit theorem.

To apply these ideas, the time t required for each computation was obtained by careful timing of the runs performed for each of our methods on the same computer: a 1.6 GHz Xeon E5310 10 quad core processor. And since the initial stage of our adaptive algorithms is simply conventional MC (the geometric learning begins with the next stage), the efficiency of conventional MC (CMC) is easily obtained by identifying the time required for this initial stage, observing the error (as measured by the standard deviation) it produced, and extrapolating to the desired precision based on the central limit theorem.

In Table 1 we compare the efficiencies of our G2 and G3 algorithms with that of conventional Monte Carlo (CMC). Thus, for example, to achieve a variance of 2.403×10^{-3} with conventional Monte Carlo when the variance it achieves in 33 seconds is 82.549 would require

$$\frac{82.549}{2.403 \times 10^{-3}} \times 33 \text{ s} = 1.133633 \times 10^{6} \text{ s},$$

whereas this was achieved with the G2 algorithm at a cost of 1341 seconds. The advantage factor of G2 compared with conventional Monte Carlo is thus

$$\frac{1.133633}{1341} \times 10^{6} = 845.$$

The G3 entry in the table was computed similarly.

[8] In MCNP [34] the term "Figure of Merit" (FOM) is used for this quantitative estimate of efficiency.

From the table we see that use of the G2 algorithm alone is sufficient to estimate the detector response in this problem to a fraction of 1% relative error, which is more than adequate for most practical applications. The addition of intelligent grid refinement adds more than an additional order of magnitude accuracy in this problem and to achieve this with conventional Monte Carlo would require more than 50 days of computation on the same platform.

Of course, one need not achieve six or more digits of precision to be successful modeling RTE problems. Taking into account the uncertainties inherent in the mechanisms that supply and detect radiation, 3–4 digits of precision in the simulation would certainly be adequate to distinguish real effects from background noise and to validate approximate RTE solutions. The significance of the accuracy achieved with the G2 and G3 methods, however, is that it is obtained with relatively simple algorithms that do not degrade explosively in higher dimensional problems, as is the case with G1 algorithms. Computational complexity for G2 increases roughly *linearly* with the number of phase space *subdivisions*, and not *exponentially* with growth in the number of phase space *dimensions*. To determine how many subdivisions will suffice for a given accuracy depends, of course, on the variability of the solution and adjoint solution functions over the regions of the decomposition. While more subdivisions may be needed in regions of large solution fluctuation, the G3 strategy guarantees that computing time is not wasted in regions for which the solutions vary only a little.

The simple grid refinement strategy described in this section is only intended to demonstrate the potential of the use of the contributon function for optimizing computational efficiency. We intend to pursue several other strategies based on this general idea to optimize computational efficiency.

6. Summary and conclusions

This paper establishes proof of principle and illustrates the latent power in our proposed methods over conventional Monte Carlo methods and existing adaptive methods. In other papers now in preparation we prove the geometric convergence of the G2 algorithm of this paper under rather general conditions, and we explore more focused applications of the G2–G3 strategy to problems in both neutron transport and photon transport. We have also recently completed a new proof of geometric convergence of our G1 algorithm under conditions more general than those presented in [23]. Each broad application area draws attention to its own specific demands and requirements. For example, in neutron transport, the energy variable merits special attention because the energy dependence of neutron cross sections is both complex and erratic, while for photon transport, directional dependence is of key importance and the contributon function must be used to understand how to vary the decomposition of the unit sphere of direction vectors with spatial location. For electron and other charged particle transport, both energy and angle dependences are important.

We believe, however, that adaptive Monte Carlo algorithms of the sort we have developed here hold the key to making RTE modeling truly practical. We expect that these methods will support accurate RTE modeling even in cases of highly complex geometric heterogeneity and subtleties in angular variation, especially near sources and detectors or in regions of "streaming radiation". At present, effective use of many existing Monte Carlo codes requires skillful user intervention in order to optimize their utility. Moreover, the strategies adopted to achieve optimization for one RTE problem may not easily transfer to another. Because the methods and algorithms developed here are mathematically rigorous and *general* rather than *ad hoc*, they offer real hope of rapid, accurate and automated RTE modeling that "tunes" itself to the specific needs of individual transport problems.

Acknowledgment

This work was partially supported by the National Science Foundation Grant DMS 0712853 and by the Laser Microbeam and Medical Program (LAMMP), NIH grant P41-RR-01192 at the Beckman Laser Institute, University of California, Irvine.

The authors thank the referees for helpful suggestions and Dr. Katherine Bhan for assistance in the preparation of this manuscript.

Appendix. Implementation of G1, G2 and G3 algorithms

Problem definition

The problem described by Eqs. (17) and (18) can also be formulated as a system of coupled integral equations:

$$\begin{aligned}\Psi_1(x) &= \Sigma_s\left(p_{11}\int_0^x e^{-\Sigma_t(x-y)}\Psi_1(y)dy + p_{12}\int_0^x e^{-\Sigma_t(x-y)}\Psi_2(y)dy\right) + Q_0 e^{-\Sigma_t x} + \int_0^x e^{-\Sigma_t(x-y)}Q_1(y)dy,\\ \Psi_2(x) &= \Sigma_s\left(p_{21}\int_x^T e^{-\Sigma_t(y-x)}\Psi_1(y)dy + p_{22}\int_x^T e^{-\Sigma_t(y-x)}\Psi_2(y)dy\right) + Q_T e^{-\Sigma_t(T-x)} + \int_x^T e^{-\Sigma_t(y-x)}Q_2(y)dy,\end{aligned} \tag{A.1}$$

or

$$\begin{aligned}\Psi_1(x) &= \Sigma_s p_{11}\int_0^x e^{-\Sigma_t(x-y)}\Psi_1(y)dy + \Sigma_s p_{12}\int_0^x e^{-\Sigma_t(x-y)}\Psi_2(y)dy + S_1(x),\\ \Psi_2(x) &= \Sigma_s p_{21}\int_x^T e^{-\Sigma_t(y-x)}\Psi_1(y)dy + \Sigma_s p_{22}\int_x^T e^{-\Sigma_t(y-x)}\Psi_2(y)dy + S_2(x),\end{aligned} \tag{A.2}$$

where

$$\begin{aligned}S_1(x) &= Q_0 e^{-\Sigma_t x} + \int_0^x e^{-\Sigma_t(x-y)}Q_1(y)dy,\\ S_2(x) &= Q_T e^{-\Sigma_t(T-x)} + \int_x^T e^{-\Sigma_t(y-x)}Q_2(y)dy.\end{aligned} \tag{A.3}$$

Sequential strategy for G1 algorithm

The following algorithm is based on Eq. (A.2) above. We begin by choosing a set of basis functions in $[0,T]$, $\{f_i(x)\}_0^\infty$, for which it is natural to choose the Legendre polynomials, $\{p_i(2x/T-1)\}_0^\infty$, which form a complete orthogonal system on the interval $[-1,1]$. Setting $f_i(x) = p_i(2x/T-1)$, the G1 algorithm finds the solution by truncating the infinite series

$$\begin{aligned}\Psi_1(x) &= \sum_{i=0}^{\infty} a_i f_i(x),\\ \Psi_2(x) &= \sum_{i=0}^{\infty} b_i f_i(x).\end{aligned} \tag{A.4}$$

By orthogonality, we have

$$\begin{aligned}a_i &= \frac{1}{c_i}\int_0^T \Psi_1(x)f_i(x)dx,\\ b_i &= \frac{1}{c_i}\int_0^T \Psi_2(x)f_i(x)dx,\end{aligned}$$

where

$$c_i = \int_0^T (f_i(x))^2 dx.$$

Monte Carlo methods will be applied to estimate the first M pairs of coefficients a_i and b_i and used to reconstruct an approximate solution, $\widehat{\Psi_1} \approx \Psi_1(x)$ and $\widehat{\Psi_2} \approx \Psi_2(x)$. The solution is obtained in stages, in each of which a predetermined number W of random walks is processed conventionally. In the initial, or zeroth stage, we have the system of equations

$$\begin{aligned}\psi_1^{(0)}(x) &= \Sigma_s p_{11}\int_0^x e^{-\Sigma_t(x-y)}\psi_1^{(0)}(y)dy + \Sigma_s p_{12}\int_0^x e^{-\Sigma_t(x-y)}\psi_2^{(0)}(y)dy + S_1(x),\\ \psi_2^{(0)}(x) &= \Sigma_s p_{21}\int_x^T e^{-\Sigma_t(y-x)}\psi_1^{(0)}(y)dy + \Sigma_s p_{22}\int_x^T e^{-\Sigma_t(y-x)}\psi_2^{(0)}(y)dy + S_2(x),\end{aligned} \tag{A.5}$$

where $\psi_1^{(0)}(x)$ and $\psi_2^{(0)}(x)$ are assumed to have an expansion similar to that in (A.4). We then apply Monte Carlo methods to obtain an estimate $\tilde{\psi}_1^{(0)}(x)$ and $\tilde{\psi}_2^{(0)}(x)$ of the exact solution, $\psi_1^{(0)}(x)$ and $\psi_2^{(0)}(x)$. We next set

$$\begin{aligned}\Psi_1(x) &= \psi_1^{(1)}(x) + \tilde{\psi}_1^{(0)}(x),\\ \Psi_2(x) &= \psi_2^{(1)}(x) + \tilde{\psi}_2^{(0)}(x).\end{aligned} \tag{A.6}$$

We substitute (A.6) into (A.2) and obtain a system of equations for $\psi_1^{(1)}(x)$ and $\psi_2^{(1)}(x)$

$$\begin{aligned}\psi_1^{(1)}(x) &= \Sigma_s p_{11}\int_0^x e^{-\Sigma_t(x-y)}\psi_1^{(1)}(y)dy + \Sigma_s p_{12}\int_0^x e^{-\Sigma_t(x-y)}\psi_2^{(1)}(y)dy + S_1^{(1)}(x),\\ \psi_2^{(1)}(x) &= \Sigma_s p_{21}\int_x^T e^{-\Sigma_t(y-x)}\psi_1^{(1)}(y)dy + \Sigma_s p_{22}\int_x^T e^{-\Sigma_t(y-x)}\psi_2^{(1)}(y)dy + S_2^{(1)}(x),\end{aligned} \tag{A.7}$$

where

$$\begin{aligned}S_1^{(1)}(x) &= S_1(x) - \tilde{\psi}_1^{(0)}(x) + \Sigma_s p_{11}\int_0^x e^{-\Sigma_t(x-y)}\tilde{\psi}_1^{(0)}(y)dy + \Sigma_s p_{12}\int_0^x e^{-\Sigma_t(x-y)}\tilde{\psi}_2^{(0)}(y)dy,\\ S_2^{(1)}(x) &= S_2(x) - \tilde{\psi}_2^{(0)}(x) + \Sigma_s p_{21}\int_0^x e^{-\Sigma_t(x-y)}\tilde{\psi}_1^{(0)}(y)dy + \Sigma_s p_{22}\int_0^x e^{-\Sigma_t(x-y)}\tilde{\psi}_2^{(0)}(y)dy.\end{aligned} \tag{A.8}$$

We solve Eq. (A.7) by conventional Monte Carlo methods, and denote the solution by $\tilde{\psi}_1^{(1)}(x)$ and $\tilde{\psi}_2^{(1)}(x)$. Assume that we have obtained solutions through the $k-1$st stage, $\tilde{\psi}_1^{(0)}(x), \tilde{\psi}_1^{(1)}(x), \ldots, \tilde{\psi}_1^{(k-1)}(x)$ and $\tilde{\psi}_2^{(0)}(x), \tilde{\psi}_2^{(1)}(x), \ldots, \tilde{\psi}_2^{(k-1)}(x)$. Set

$$\begin{aligned}\Psi_1(x) &= \tilde{\psi}_1^{(0)}(x) + \tilde{\psi}_1^{(1)}(x) + \cdots + \tilde{\psi}_1^{(k-1)}(x) + \psi_1^{(k)}(x),\\ \Psi_2(x) &= \tilde{\psi}_2^{(0)}(x) + \tilde{\psi}_2^{(1)}(x) + \cdots + \tilde{\psi}_2^{(k-1)}(x) + \psi_2^{(k)}(x),\end{aligned} \tag{A.9}$$

substitute them into (A.2) and obtain equations for the k-th stage correction $\psi_1^{(k)}(x)$ and $\psi_2^{(k)}(x)$:

$$\begin{aligned}\psi_1^{(k)}(x) &= \Sigma_s p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_1^{(k)}(y)\mathrm{d}y + \Sigma_s p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_2^{(k)}(y)\mathrm{d}y + S_1^{(k)}(x),\\ \psi_2^{(k)}(x) &= \Sigma_s p_{21}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_1^{(k)}(y)\mathrm{d}y + \Sigma_s p_{22}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_2^{(k)}(y)\mathrm{d}y + S_2^{(k)}(x),\end{aligned} \tag{A.10}$$

where

$$\begin{aligned}S_1^{(k)}(x) &= S_1(x) - (\tilde{\psi}_1^{(0)}(x) + \tilde{\psi}_1^{(1)}(x) + \cdots + \tilde{\psi}_1^{(k-1)}(x)) + \Sigma_s\Big(p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_1^{(0)}(y) + \tilde{\psi}_1^{(1)}(y) + \cdots + \tilde{\psi}_1^{(k-1)}(y))\mathrm{d}y\\ &\quad + p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_2^{(0)}(y) + \tilde{\psi}_2^{(1)}(y) + \cdots + \tilde{\psi}_2^{(k-1)}(y))\mathrm{d}y\Big),\\ S_2^{(k)}(x) &= S_2(x) - (\tilde{\psi}_2^{(0)}(x) + \tilde{\psi}_2^{(1)}(x) + \cdots + \tilde{\psi}_2^{(k-1)}(x)) + \Sigma_s\Big(p_{21}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_1^{(0)}(y) + \tilde{\psi}_1^{(1)}(y) + \cdots + \tilde{\psi}_1^{(k-1)}(y))\mathrm{d}y\\ &\quad + p_{22}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_2^{(0)}(y) + \tilde{\psi}_2^{(1)}(y) + \cdots + \tilde{\psi}_2^{(k-1)}(y))\mathrm{d}y\Big).\end{aligned} \tag{A.11}$$

Then the pair

$$\begin{aligned}\widetilde{\Psi}_1(x) &\equiv \tilde{\psi}_1^{(0)}(x) + \tilde{\psi}_1^{(1)}(x) + \cdots + \tilde{\psi}_1^{(k)}(x),\\ \widetilde{\Psi}_2(x) &\equiv \tilde{\psi}_2^{(0)}(x) + \tilde{\psi}_2^{(1)}(x) + \cdots + \tilde{\psi}_2^{(k)}(x)\end{aligned} \tag{A.12}$$

is taken as an approximate solution of system (A.2) resulting from the $k+1$ adaptive stages and

$$\widetilde{\Psi}(x) = \widetilde{\Psi}_1(x) + \widetilde{\Psi}_2(x). \tag{A.13}$$

Sequential strategy for G2 algorithm

Assume that the interval $I = [0, T]$ has been subdivided into a disjoint union of subintervals $\{I_i\}_{i=1}^N$,

$$I = \bigcup_{i=1}^N I_i \tag{A.14}$$

and, without loss of generality, we assume that the length of each subinterval is $V_i > 0$. Our goal with G2 algorithm is to estimate the averages of the solution over the subintervals

$$\begin{aligned}a_i &= \frac{1}{V_i}\int_{I_i}\Psi_1(x)\mathrm{d}x,\\ b_i &= \frac{1}{V_i}\int_{I_i}\Psi_2(x)\mathrm{d}x,\end{aligned} \tag{A.15}$$

Mimicking the procedure for the G1 algorithm, we consider the zeroth stage for $\psi_1^{(0)}(x)$ and $\psi_2^{(0)}(x)$

$$\begin{aligned}\psi_1^{(0)}(x) &= \Sigma_s p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_1^{(0)}(y)\mathrm{d}y + \Sigma_s p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_2^{(0)}(y)\mathrm{d}y + S_1^{(0)}(x),\\ \psi_2^{(0)}(x) &= \Sigma_s p_{21}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_1^{(0)}(y)\mathrm{d}y + \Sigma_s p_{22}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_2^{(0)}(y)\mathrm{d}y + S_2^{(0)}(x),\end{aligned} \tag{A.16}$$

where

$$\begin{aligned}S_1^{(0)}(x) &= S_1(x),\\ S_2^{(0)}(x) &= S_2(x).\end{aligned} \tag{A.17}$$

To estimate the integrals,

$$\begin{aligned} a_i^{(0)} &= \frac{1}{V_i}\int_{I_i}\psi_1^{(0)}(x)\mathrm{d}x, \\ b_i^{(0)} &= \frac{1}{V_i}\int_{I_i}\psi_2^{(0)}(x)\mathrm{d}x, \end{aligned} \tag{A.18}$$

we employ conventional Monte Carlo methods. Suppose that the estimated values are $\widetilde{a}_i^{(0)}$ and $\widetilde{b}_i^{(0)}$. Then define the estimated solution as

$$\begin{aligned} \tilde{\psi}_{1,a}^{(0)}(x) &= \widetilde{a}_i^{(0)}, \quad \text{when } x \in I_i, \\ \tilde{\psi}_{2,a}^{(0)}(x) &= \widetilde{b}_i^{(0)}, \quad \text{when } x \in I_i. \end{aligned} \tag{A.19}$$

Continuing to the first stage, we set

$$\begin{aligned} \Psi_1(x) &= \psi_1^{(1)}(x) + \tilde{\psi}_{1,a}^{(0)}(x), \\ \Psi_2(x) &= \psi_2^{(1)}(x) + \tilde{\psi}_{2,a}^{(0)}(x) \end{aligned} \tag{A.20}$$

and substitute (A.20) into (A.2) to produce the equation

$$\begin{aligned} \psi_1^{(1)}(x) &= \Sigma_s p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_1^{(1)}(y)\mathrm{d}y + \Sigma_s p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_2^{(1)}(y)\mathrm{d}y + S_1^{(1)}(x), \\ \psi_2^{(1)}(x) &= \Sigma_s p_{21}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_1^{(1)}(y)\mathrm{d}y + \Sigma_s p_{22}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_2^{(1)}(y)\mathrm{d}y + S_2^{(1)}(x), \end{aligned} \tag{A.21}$$

for the correction $\psi_1^{(1)}(x)$ and $\psi_2^{(1)}(x)$ where

$$\begin{aligned} S_1^{(1)}(x) &= S_1(x) - \tilde{\psi}_{1,a}^{(0)}(x) + \Sigma_s\left(p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\tilde{\psi}_{1,a}^{(0)}(y)\mathrm{d}y + p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\tilde{\psi}_{2,a}^{(0)}(y)\mathrm{d}y\right), \\ S_2^{(1)}(x) &= S_2(x) - \tilde{\psi}_{2,a}^{(0)}(x) + \Sigma_s\left(p_{21}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\tilde{\psi}_{1,a}^{(0)}(y)\mathrm{d}y + p_{22}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\tilde{\psi}_{2,a}^{(0)}(y)\mathrm{d}y\right). \end{aligned} \tag{A.22}$$

From (A.21), we can estimate the integrals

$$\begin{aligned} a_i^{(1)} &= \frac{1}{V_i}\int_{I_i}\psi_1^{(1)}(x)\mathrm{d}x, \\ b_i^{(1)} &= \frac{1}{V_i}\int_{I_i}\psi_2^{(1)}(x)\mathrm{d}x \end{aligned} \tag{A.23}$$

by conventional Monte Carlo methods. In general, having obtained $\tilde{\psi}_{1,a}^{(0)}(x), \tilde{\psi}_{1,a}^{(1)}(x), \ldots, \tilde{\psi}_{1,a}^{(k-1)}(x)$ and $\tilde{\psi}_{2,a}^{(0)}(x), \tilde{\psi}_{2,a}^{(1)}(x), \ldots, \tilde{\psi}_{2,a}^{(k-1)}(x)$, set

$$\begin{aligned} \Psi_1(x) &= \psi_1^{(k)}(x) + \tilde{\psi}_{1,a}^{(0)}(x) + \cdots + \tilde{\psi}_{1,a}^{(k-1)}(x), \\ \Psi_2(x) &= \psi_2^{(k)}(x) + \tilde{\psi}_{2,a}^{(0)}(x) + \cdots + \tilde{\psi}_{2,a}^{(k-1)}(x). \end{aligned} \tag{A.24}$$

This determines a system of equations for $\psi_1^{(k)}(x)$ and $\psi_2^{(k)}(x)$

$$\begin{aligned} \psi_1^{(k)}(x) &= \Sigma_s p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_1^{(k)}(y)\mathrm{d}y + \Sigma_s p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}\psi_2^{(k)}(y)\mathrm{d}y + S_1^{(k)}(x), \\ \psi_2^{(k)}(x) &= \Sigma_s p_{21}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_1^{(k)}(y)\mathrm{d}y + \Sigma_s p_{22}\int_x^T \mathrm{e}^{-\Sigma_t(y-x)}\psi_2^{(k)}(y)\mathrm{d}y + S_2^{(k)}(x), \end{aligned} \tag{A.25}$$

where

$$\begin{aligned} S_1^{(k)}(x) &= S_1(x) - (\tilde{\psi}_{1,a}^{(0)}(x) + \cdots + \tilde{\psi}_{1,a}^{(k-1)}(x)) + \Sigma_s\left(p_{11}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_{1,a}^{(0)}(y) + \cdots + \tilde{\psi}_{1,a}^{(k-1)}(y))\mathrm{d}y\right. \\ &\quad \left. + p_{12}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_{2,a}^{(0)}(y) + \cdots + \tilde{\psi}_{2,a}^{(k-1)}(y))\mathrm{d}y\right), \\ S_2^{(k)}(x) &= S_2(x) - (\tilde{\psi}_{2,a}^{(0)}(x) + \cdots + \tilde{\psi}_{2,a}^{(k-1)}(x)) + \Sigma_s\left(p_{21}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_{1,a}^{(0)}(y) + \cdots + \tilde{\psi}_{1,a}^{(k-1)}(y))\mathrm{d}y\right. \\ &\quad \left. + p_{22}\int_0^x \mathrm{e}^{-\Sigma_t(x-y)}(\tilde{\psi}_{2,a}^{(0)}(y) + \cdots + \tilde{\psi}_{2,a}^{(k-1)}(y))\mathrm{d}y\right). \end{aligned} \tag{A.26}$$

From (A.25), we can estimate the integrals

$$\begin{aligned} a_i^{(k)} &= \frac{1}{V_i}\int_{I_i}\psi_1^{(k)}(x)\mathrm{d}x, \\ b_i^{(k)} &= \frac{1}{V_i}\int_{I_i}\psi_2^{(k)}(x)\mathrm{d}x \end{aligned} \tag{A.27}$$

by conventional Monte Carlo methods. Finally, for each $i(1 \leqslant i \leqslant R)$, we sum over all adaptive stages to obtain the approximation

$$\begin{aligned} \widetilde{a}_i^{(0)} + \widetilde{a}_i^{(1)} + \cdots + \widetilde{a}_i^{(k)} &\to \frac{1}{1V_i}\int_{I_i}\Psi_1(x)\mathrm{d}x, \\ \widetilde{b}_i^{(0)} + \widetilde{b}_i^{(1)} + \cdots + \widetilde{b}_i^{(k)} &\to \frac{1}{V_i}\int_{I_i}\Psi_2(x)\mathrm{d}x \end{aligned} \tag{A.28}$$

for the integral (A.15).

Sequential strategy for G3 algorithm

The G3 algorithm is designed to achieve high precision automatically in solving transport problems. The idea is to use the contributon function to reduce computational costs in low contributon regions and uniformize the effort over all subregions based on contributon values. This idea is carried out in practice by two processes: refining the high contributon subregions and recombining subregions in which contributon values might have become too small for optimal efficiency. We describe the G3 strategy for general transport problems and use P, Q to denote generic state space vectors. Assume that the transport problem is expressed by an integral equation

$$\Psi(P) = \int_\Gamma K(P,Q)\Psi(Q)\mathrm{d}Q + S(P), \tag{A.29}$$

where $K(P,Q)$ is the transport kernel and $S(P)$ is the source function. Then the adjoint equation is

$$\Psi^*(P) = \int_\Gamma K^*(P,Q)\Psi^*(Q)\mathrm{d}y + S^*(P), \tag{A.30}$$

where $K^*(P,Q) = K(Q,P)$ and $S^*(P)$ describes the detector. That is, the quantity we are estimating is

$$I = \int_\Gamma S^*(P)\Psi(P)\mathrm{d}P.$$

We denote the approximate solutions produced by the G2 algorithm by $\widetilde{\Psi}(P)$ and $\widetilde{\Psi}^*(P)$, respectively. Then the approximate contributon function is

$$C(P) = \widetilde{\Psi}(P)\widetilde{\Psi}^*(P). \tag{A.31}$$

The G3 algorithm accepts as input a global constant G which roughly indicates the desired precision in estimating the integral I. After linearly ordering the R subregions $\Gamma_1, \ldots, \Gamma_R$, the contributon integrals over all subregions are themselves organized in a linear array. That is, we set

$$C_i = \int_{\Gamma_i} C(P)\mathrm{d}P$$

and reorder the subscripts, if necessary, so that $0 \leqslant C_1 \leqslant C_2 \leqslant \cdots \leqslant C_R$. Provided it is not too small, the value C_1 can be set to G to initiate the G3 algorithm.

Refining the decomposition of the phase space

This step is applied if a subregion contains too much information, as measured by the contributon integrals. The criterion for refining a subregion Γ_i is based on an examination of the integer $r_i = \left[\frac{H_i}{G}\right]$ where

$$H_i = \int_{\Gamma_i} C(P)\mathrm{d}P \tag{A.32}$$

and $\left[\frac{H_i}{G}\right]$ denotes the integer part of $\frac{H_i}{G}$. If $r_i < 2$, we do not subdivide the subregion Γ_i but if $r_i \geqslant 2$, we divide the subinterval Γ_i into r_i smaller subregions with equal volume. This process is continued until all subregions have been tested for refinement.

Recombining

This step is applied if a subregion contains too little information about the contributon function, in which case we consider joining it to an adjacent subregion and retesting the newly recombined union for further possible recombination. Consider the integral of $C(x)$ over the ith subregion

$$H_i = \int_{\Gamma_i} C(P)\mathrm{d}P \tag{A.33}$$

and the ratio $u_i = \frac{H_i}{G}$ (here we do not take the integer part). If $u_i \geqslant \frac{1}{2}$, we ignore the subregion Γ_i for now, but if $u_i < \frac{1}{2}$, we consider combining the subregion Γ_i with one or more adjacent subregions. For example, if $1/2 \leqslant u_{i-1} < 2$ we define a new subregion $\Gamma_{i-1,i} = \Gamma_i \cup \Gamma_{i-1}$ and test this new subregion for possible further recombination with a third subregion. This process is continued until the refinement and recombination logic is completed. The details are too lengthy to spell out fully here. However, the overall objective of the G3 algorithm is to optimize the computational efficiency in estimating the integral I. This relies on increasing the amount of detail needed in computationally important regions while reducing the amount of detail needed in computationally insignificant regions. Convergence of the G3 algorithm occurs when the amount of information extracted from each subregion (as measured by contributon integrals) is approximately uniform across all subregions.

References

[1] R.C. Haskell, L.O. Svaasand, T.T. Tsay, T.C. Feng, M. McAdams, B.J. Tromberg, J. Opt. Soc. Am. A 11 (1994) 2727–2741.
[2] A. Kienle, M.S. Patterson, J. Opt. Soc. Am. 14 (1997) 246–254.
[3] F. Martelli, D. Contini, A. Taddeucci, G. Zaccanti, Appl. Opt. 36 (1997) 4600–4612.
[4] J. Spanier, E. Gelbard, Monte Carlo Principles and Neutron Transport Problems, Addison-Wesley, 1969.
[5] I. Lux, L. Koblinger, Monte Carlo Particle Transport Methods: Neutron and Photon Calculations, CRC Press, 1991.
[6] K.M. Case, P.F. Zweifel, Linear Transport Theory, Addison-Wesley, 1967.
[7] R.Y. Rubinstein, Simulation and the Monte Carlo Method, Wiley, 1981.
[8] M. Leimdorfer, Trans. Chalmers Univ. of Tech., 286, Goteborg, Sweden, 1964.
[9] M. Leimdorfer, Trans. Chalmers Univ. of Tech., 287, Goteborg, Sweden, 1964.
[10] M.H. Kalos, Nucl. Sci. Eng. 16 (1963) 227–234.
[11] H. Greenspan, C.N. Kelber, D. Okrent, Computing Methods in Reactor Physics, Gordon and Breach, 1968.
[12] G. Goertzel, M.H. Kalos, Progr. Nucl. Energy Ser. I 2 (1958) 315–369.
[13] T.E. Booth, LA-10363-MS, a sample problem for variance reduction in MCNP. Technical Report, Los Alamos National Laboratory, 1985.
[14] T.E. Booth, Nucl. Sci. Eng. 127 (1997) 338–345.
[15] T.E. Booth, Nucl. Sci. Eng. 102 (1989) 332–340.
[16] T.E. Booth, LAUR-98-47-55, Adaptively learning an importance function using transport constrained Monte Carlo. Technical Report, Los Alamos National Laboratory, 1998.
[17] T.E. Booth, Nucl. Sci. Eng. 136 (2000) 399–408.
[18] G.I. Bell, S. Glasstone, Nuclear Reactor Theory, Krieger Pub. Co., 1970.
[19] Y. Lai, J. Spanier, in: Monte Carlo and Quasi-Monte Carlo Methods 1998, Springer, 1999, pp. 273–283.
[20] C.K. Hayakawa, J. Spanier, in: Monte Carlo and Quasi-Monte Carlo Methods 1998, Springer, 1999, pp. 214–226.
[21] R. Kong, Transport Problems and Monte Carlo Methods, Ph.D. thesis, Claremont Graduate University, 1999.
[22] R. Kong, J. Spanier, Monte Carlo and Quasi-Monte Carlo Methods 1998, Springer, 1999. pp. 235–251.
[23] R. Kong, J. Spanier, Monte Carlo and Quasi-Monte Carlo Methods 1998, Springer, 1999. pp. 252–272.
[24] J. Spanier, Monte Carlo and Quasi-Monte Carlo Methods 1998, Springer, 1999. pp. 98–113.
[25] J. Spanier, Monte Carlo and Quasi-Monte Carlo Methods 2002, Springer, 2004. pp. 439–449.
[26] R. Kong, J. Spanier, Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer, 2002. pp. 306–317.
[27] J. Halton, Proc. Camb. Phil. Soc. 58 (1962) 57–78.
[28] J. Spanier, Nucl. Sci. Eng. 133 (1999) 73–79.
[29] M.L. Williams, W.W. Engle, Nucl. Sci. Eng. 62 (1977) 92–104.
[30] M.L. Williams, Nucl. Sci. Eng. 108 (1991) 355–383.
[31] C.H. Aboughantous, Nucl. Sci. Eng. 108 (1994) 160–177.
[32] S.N. Cramer, Nucl. Sci. Eng. 124 (1996) 398–416.
[33] I.V. Serov, T.M. John, J.E. Hoogenboom, Appl. Radiat. Isot. 49 (12) (1998) 1737–1744.
[34] Ed. J. Breismeister, MCNP4C, a Monte Carlo N-Particle Transport Code System, Report CCC-660. Technical report, Los Alamos National Laboratory, 1999.
[35] R. Aronson, J. Opt. Soc. Am. A 14 (1997) 486–490.
[36] J. Spanier, SIAM J. Appl. Math. 18 (1970) 172–190.