Lecture notes:
Variance Reduction Techniques

Alex F Bielajew and D W O Rogers
Institute for National Measurement Standards
National Research Council of Canada
Ottawa, Canada
K1A 0R6

Tel: 613-993-2715
FAX: 613-952-9865
e-mail: `alex@irs.phy.nrc.ca`

National Research Council of Canada Report PIRS-0396

On variance reduction:
"First, implement an elegant technique to save computer time.
Then, do it the long way to check that you implemented it correctly."
Martin J Berger

1 Introduction

In this lecture we discuss various techniques which may be used to make calculations more efficient. In some cases, these techniques require that no further approximations be made to the transport physics. In other cases, the gains in computing speed come at the cost of computing results that may be less accurate since approximations are introduced. The techniques may be divided into 3 categories: those that concern electron transport only, those that concern photon transport only, and other more general methods. The set of techniques we discuss does not represent an exhaustive list. There is much reference material available and we only cite a few of them (refs. [1], [2], [3], [4],[5]). An especially rich source of references is McGrath's book [3], which contains an annotated bibliography. Instead we shall concentrate on techniques that have been of considerable use to the authors and their close colleagues. However, it is appropriate to discuss briefly what we are trying to accomplish by employing variance reduction techniques.

1.1 Variance reduction or efficiency increase?

What we really mean to do when we employ variance reduction techniques is to reduce the time it takes to calculate a result with a given variance. Analogue Monte Carlo calculations attempt to simulate the full stochastic development of the electromagnetic cascade. Hence, with the calculated result is associated a variance, s^2. The method by which s^2 is estimated will not be discussed here. Let us assume that it is calculated by some consistent method. If the variance is too large for our purposes we run more histories until our criterion is satisfied. How do we estimate how many more histories are needed? Assuming we can do this, what do we do if it is too expensive to simulate the requisite number of histories? We may need a more subtle approach than reducing variance by "grinding out" more histories.

Let us say we devise some "tricks" that allow us to reduce the variance by, say, a factor of 10 using the same number of histories. Let's also imagine that this new subtle approach we have devised takes, say, 20 times longer on average to complete a particle history. (For example, our variance reduction technique may involve some detailed, expensive calculation executed every particle step.) Although we have reduced the variance by a factor of 10, we take 20 times longer to calculate each particle history. We have actually reduced the efficiency by a factor of two! To add to the insult, we have wasted our own time implementing a technique which reduces efficiency!

We require a reliable figure of merit which we may use to estimate gains in efficiency of a given "variance reduction" technique. It is common to use the efficiency, ϵ, defined by:

$$\epsilon = \frac{1}{s^2 T}, \tag{1}$$

where T is a measure of the computing time used (*e.g.* CPU seconds). The motivation for this choice comes from the following: We assume that mean values of quantities calculated by Monte Carlo methods are distributed normally. It then follows that for calculations performed using identical methods, the quantities $s^2 N$ and $s^2 T$, where N is the number of histories, are constant, on average. This is so because N should be directly proportional to T. By considering the efficiency to be constant, eq. 1 may be used to estimate the total computing time required to reach a given statistical accuracy if a preliminary result has already been obtained. For example, if one wishes to reduce the uncertainty, s, by a factor of 2, one needs 4 times as many histories. More importantly, eq. 1 allows us to make a quantitative estimate of the gain (or loss!) in efficiency resulting from the use of a given "variance reduction" technique since it accounts for not only the reduction in variance but also the increased computing time it may take to incorporate the technique. In the aforementioned example, using eq. 1 we would obtain ϵ(with subtlety)/ϵ(brute force) = 0.5, a *reduction* of 1/2. In the following sections we attempt to present more successful variance reduction techniques!

2 Electron-specific methods

2.1 Geometry interrogation reduction

This section might also have been named "Code optimisation" or "Don't calculate what you don't really need", or something equivalent. We note that there is a fundamental difference between the transport of photons and electrons in a condensed-history transport code. Photons travel relatively long distances before interacting and their transport steps are often interrupted by boundary crossings (*i.e.* entering a new scoring region or element of the geometry). The transport of electrons is different, however. In addition to having its step interrupted by boundary crossings or the sampling of discrete interactions, the electron has other constraints on step-size. These constraints may have to do with ensuring that the underlying multiple scattering theories are not being violated in any way (See the Lecture: Step-size dependencies and PRESTA), or the transport may have to be interrupted so that the equations of transport in an external electromagnetic field may be integrated [6]. Therefore, it is often unnecessary to make repeated and expensive checks with the geometry routines of the transport code because the electron is being transported in an effectively infinite medium for most of the transport steps. The EGS4 code [7], has an option that allows the user to avoid these redundant geometry subroutine calls. With this option switched on, whenever the geometry must be checked for whatever reason, the closest distance to any boundary is calculated and stored. This variable is then decremented by the length of each transport step. If this variable is greater than zero, the electron can not be close enough to a boundary to cross it and the geometry subroutines are not interrogated. If this variable drops to zero or less, the geometry subroutines are called because a boundary crossing may occur.

There is no additional approximation involved in this technique. The gain in transport efficiency is slightly offset by the extra calculation time that is spent calculating the distance to the closest boundary. (This parameter is not always needed for other aspects of the particle transport.) As an example, consider the case of a pencil beam of 1 MeV electrons incident normally on a 0.3 cm slab of carbon divided into twelve 0.025 cm slabs. For this set of simulations, transport and secondary particle creation thresholds were set at 10 keV kinetic energy and we used EGS4 [7] setting the energy loss per electron step at 1% for accurate electron transport [8] at low energies. The case that interrogates the geometry routines on every step is called the "base case". We invoke the trick of interrogating the geometry routines only when needed and call this the "RIG" (reduced interrogation of geometry) case. The efficiency ratio, ϵ(RIG)/ϵ(base), was found to be 1.34, a significant improvement. (This was done by calculating

DNEAR in the HOWFAR routine of a planar geometry code. A discussion of DNEAR is given on pages 256–258 of the EGS4 manual [7].)

Strictly speaking, this technique may be used for photons as well. For most practical problems, however, the mean free path for the photons in the problem is of the order, or greater than the distance between boundaries. For deep penetration problems or similar problems, this may not be true. However, this technique is usually more effective at speeding up the electron transport part of the simulation.

The extra time required to calculate the distance to the closest boundary may be considerable, especially for simulations involving curved surfaces. If this is so then the efficiency gain may be much less or efficiency may be lost. It is advisable to test this technique before employing it in "production" runs.

2.2 Discard within a zone

In the previous example, we may be just interested in the energy deposited in the planar zones of the carbon slab. We may, therefore, deposit the energy of an electron entirely within a zone if that electron's range is less than the distance to any bounding surface of the zone in which it is being transported. We note that we make an approximation in doing this—we neglect the creation and transport of any bremsstrahlung γ's that may otherwise created. For the worst possible case in this particular example, we will be discarding electrons that have a range that is half of the zone thickness, *i.e.* having a kinetic energy of about 110 keV. The radiative yield of these electrons is only about 0.07%. Therefore, unless we are directly interested in the radiative component of the electron's slowing down process in this problem, the approximation is an excellent one. For the above example, we realise a gain in the efficiency ratio, ϵ(zonal discard + RIG)/ϵ(base), of about 2.3. In this case, the transport cut-off, below which no electron was transported, was 10 keV. If we had used a higher cut-off the efficiency gain would have been less.

Before adopting this technique, the user should carefully analyze the consequences of the approximation—the neglect of bremsstrahlung from the low energy electron component.

2.3 PRESTA!

In a previous lecture, "Step-size dependencies and PRESTA", we discussed an alternative electron transport algorithm, PRESTA. This algorithm, by making improvements to the physical modeling of electron transport, allows the use of large electron steps when one is far away from boundaries. This algorithm may, therefore, be considered to be a variance reduction technique, since it saves computing time by employing small steps only where needed—in the vicinity of boundaries and interfaces. Continuing with the present example, we calculate the gain in efficiency ratio, ϵ(PRESTA)/ϵ(base), to be 6.1. RIG is always switched on with PRESTA, so it is actually fairer to calculate the efficiency ratio, ϵ(PRESTA)/ϵ(RIG), which was found to be 4.6. If we allow zonal discard as well, we calculate the efficiency ratio, ϵ(zonal discard + PRESTA)/ϵ(zonal discard + RIG), to be 3.1. There is a brief discussion in the previous lecture on when PRESTA is expected to run quickly. Basically, the fewer the boundaries and the higher the transport cutoffs, the faster PRESTA runs. A detailed discussion is given in the PRESTA documentation [9].

2.4 Range rejection

As a final example of electron variance reduction, we consider the technique called "range rejection". This is similar to the "discard within a zone" except for a few differences. Instead of discarding (*i.e.* stopping the transport and depositing the energy "on the spot") the electron because it can not reach the boundaries of the geometrical element it is in, the electron is discarded because it can not reach some region of interest. For example, a particle detector may contain a sensitive volume where one wishes to calculate energy deposit, or some other quantity. Surrounding this sensitive volume may be shields, converters, walls *etc.* where one wishes accurate particle transport to be accomplished but where one does not wish to score quantities directly. Electrons that can not reach the sensitive volume may be discarded "on the spot", providing that the neglect of the bremsstrahlung γ's causes no great inaccuracy.

As an example of range rejection, we consider the case of an ion chamber [10]. In this case, a cylindrical air cavity, 2 mm in depth and 1.0 cm in radius is surrounded by 0.5 g/cm^2 carbon walls. A flat circular end is irradiated by 1.25 MeV γ-rays incident normally. This approximates the irradiation from a distant source of ^{60}Co. This is a "thick-walled" ion chamber, so-called because it's thickness exceeds the range of the maximum energy electron that can be set in motion by the incident photons. This sets up a condition of "near charged particle equilibrium" in the vicinity of the cavity. The potential for significant saving in computer time is evident, for many electrons could never reach the cavity. We are interested in calculating the energy deposited to the air in the cavity and we are not concerned with scoring any quantities in the walls. The range rejection technique involved calculating the closest distance to the surface of the cavity on every transport step. If this distance exceeded the CSDA range of the electron, it was discarded. The omission of residual bremsstrahlung photon creation and transport was negligible in this problem. The secondary particle creation thresholds were set at 10 keV kinetic energy as well as the transport cut-off energies. (ECUT=AE=0.521 MeV, PCUT=AP=0.01 MeV, and ESTEPE=0.01 for accurate low energy simulation.) A factor of 4 increase in efficiency was realised in this case.

Range rejection is a relatively crude but effective method. The version described above neglects residual bremsstrahlung and is applicable when the discard occurs in one medium. The bremsstrahlung problem could be solved by forcing at least some of the electrons to produce bremsstrahlung. The amount of energy eventually deposited from these photons would have to be weighted accordingly to keep the sampling game "fair". Alternatively, one could transport fully a fraction, say f, of the electrons and weight any resultant bremsstrahlung photons by $1/f$. The other problem, the one of multi-media discard, is difficult to treat in complete generality. The difficulty is primarily a geometrical one. The shortest distance to the scoring region is the shortest geometrical path only when the transport can occur in one medium. The shortest distance we need to calculate for range rejection is the path along which the energy loss is a minimum. It is not difficult to imagine that finding the "shortest" path for transport in more than one medium may be very difficult. For special cases this may be done or approximations may be made. The "payoff" is worth it as large gains in efficiency may be realised, as seen in the above example.

3 Photon-specific methods

3.1 Interaction forcing

In problems where the interaction of photons is of interest, efficiency may be lost because photons leave the geometry of the simulation without interacting. The probability distribution for a photon interaction is:

$$p(\lambda)d\lambda = e^{-\lambda}d\lambda, \tag{2}$$

where $0 \le \lambda < \infty$ and λ is the distance measured in mean free paths. It can easily be shown that sampling λ from this distribution can be accomplished by the following formula[1]:

$$\lambda = -\ln(1-\xi), \tag{3}$$

where ξ is a random number uniform on the range, $0 \le \xi < 1$. Since λ extends to infinity and the number of photon mean free paths across the geometry in any practical problem is finite, there is a non-zero and often large probability that photons leave the geometry of interest without interacting. If they don't interact, we waste computing time tracking these photons through the geometry.

Fortunately, this waste may be prevented. We can *force* these photons to interact. The method by which this can be achieved is remarkably simple. We construct the probability distribution,

$$p(\lambda)d\lambda = \frac{e^{-\lambda}d\lambda}{\int_0^\Lambda e^{-\lambda'}d\lambda'}, \tag{4}$$

where Λ is the total number of mean free paths along the direction of motion of the photon to the end of the geometry. This λ is restricted to the range, $0 \le \lambda < \Lambda$, and λ is selected from the equation,

$$\lambda = -\ln(1-\xi(1-e^{-\Lambda})). \tag{5}$$

We see from eq. 5 that we recover eq. 3 in the limit $\Lambda \longrightarrow \infty$. Since we have forced the photon to interact within the geometry of the simulation we must *weight* the quantities scored resulting from this interaction. This weighting takes the form,

$$\omega' = \omega(1-e^{-\Lambda})), \tag{6}$$

where ω' is the new "weighting" factor and ω is the old weighting factor. When interaction forcing is used, the weighting factor, $1-e^{-\Lambda}$, simply multiplies the old one. This factor is the probability that the photon would have interacted before leaving the geometry of the simulation. This variance reduction technique may be used repeatedly to force the interaction of succeeding generations of scattered photons. It may also be used in conjunction with other variance reduction techniques. Interaction forcing may also be used in electron problems to force the interaction of bremsstrahlung photons.

On first inspection, one might be tempted to think that the calculation of Λ may be difficult in general. Indeed, this calculation is quite difficult and involves summing the contributions to Λ along the photon's direction through all the geometrical elements and materials along the way. Fortunately, most of this calculation is present in any Monte Carlo code because it must possess the capability of transporting the photons through this geometry! This interaction forcing capability can be included in the EGS code in a completely general, *geometry independent fashion* with only about 30 lines of code [11]!

The increase in efficiency can be dramatic if one forces the photons to interact. For example, for ion chamber calculations similar to those described in sec. 2.4 and discussed in detail elsewhere [10], the efficiency improved by the factor 2.3. In this calculation, only about

[1] It is conventional to use the expression, $\lambda = -\ln(\xi)$, since both $1-\xi$ and ξ are distributed uniformly on (0,1) but the former expression executes more slowly. However, it has a closer connection to the following mathematical development.

6% of the photons would have interacted in the chamber. In calculating the dose to skin from contaminant electrons arising from the interaction of ^{60}Co (*i.e.* 1.25 MeV γ's) in 100 cm of air [11], the calculation executed 7 times more efficiently after forcing the photons to interact. In calculating the dose from ^{60}Co directly in the skin (a 0.001 cm slice of tissue) where normally only 6×10^{-5} of the photons interact, the efficiency improved by a factor of 2600 [11, 12]!

3.2 Exponential transform, russian roulette, and particle splitting

The exponential transform is a variance reduction technique designed to enhance efficiency for either deep penetration problems (*e.g.* shielding calculations) or surface problems (*e.g.* build-up in photon beams). It is often used in neutron Monte Carlo work and is directly applicable to photons as well.

Consider the simple problem where we are interested in the surface or deep penetration in a simple slab geometry with the planes of the geometry normal to the z-axis. We then scale the interaction probability making use of the following formula:

$$\tilde{\lambda} = \lambda(1 - C\mu), \tag{7}$$

where λ is the distance measured in the number of mean free path's, $\tilde{\lambda}$ is the scaled distance, μ is the cosine of the angle the photon makes with the z-axis, and C is a parameter that adjusts the magnitude of the scaling. The interaction probability distribution is:

$$\tilde{p}(\lambda)d\lambda = (1 - C\mu)e^{-\lambda(1-C\mu)}d\lambda, \tag{8}$$

where the overall multiplier $1 - C\mu$ is introduced to ensure that the probability is correctly normalised, *i.e.* $\int_0^\infty \tilde{p}(\lambda)d\lambda = 1$. For $C = 0$, we have the unbiased probability distribution $e^{-\lambda}d\lambda$. One sees that for $0 < C < 1$, the average distance to an interaction is stretched[2]. For $C < 0$, the average distance to the next interaction is shortened. Examples of a stretched and shortened distribution are given in fig. 1. In order to play the game fairly, we must obtain the appropriate weighting function to apply to all subsequent scoring functions. This is obtained by requiring that the overall probability be unchanged. That is, we require:

$$\omega'\tilde{p}(\lambda)d\lambda = \omega p(\lambda)d\lambda, \tag{9}$$

where ω' is the new weighting factor and ω is the old weighting factor. Solving eq. 9 for ω' yields,

$$\omega' = \omega e^{-\lambda C\mu}/(1 - C\mu). \tag{10}$$

Finally, we require a technique to sample the stretched or shortened number of mean free paths to the next interaction point from a random number. It is easily shown that λ is selected using the formula:

$$\lambda = -\ln(\xi)/(1 - C\mu), \tag{11}$$

where ξ is a random number chosen uniformly over the range, $0 < \xi \leq 1$.

For complete generality, one must obey the restriction, $|C| < 1$ since the photon's direction is arbitrary ($-1 \leq \mu \leq 1$). "Path-length stretching" means that $0 < C < 1$, i.e. photons are made to penetrate deeper. "Path-length shortening" means that $-1 < C < 0$, i.e. photons are made to interact closer to the surface. For studies of surface regions, one may use a stronger

[2] Note that the average number of mean free paths to an interaction, $\langle\lambda\rangle$, is given by $\langle\lambda\rangle = \int_0^\infty \lambda\tilde{p}(\lambda)d\lambda = \frac{1}{1-C\mu}$.

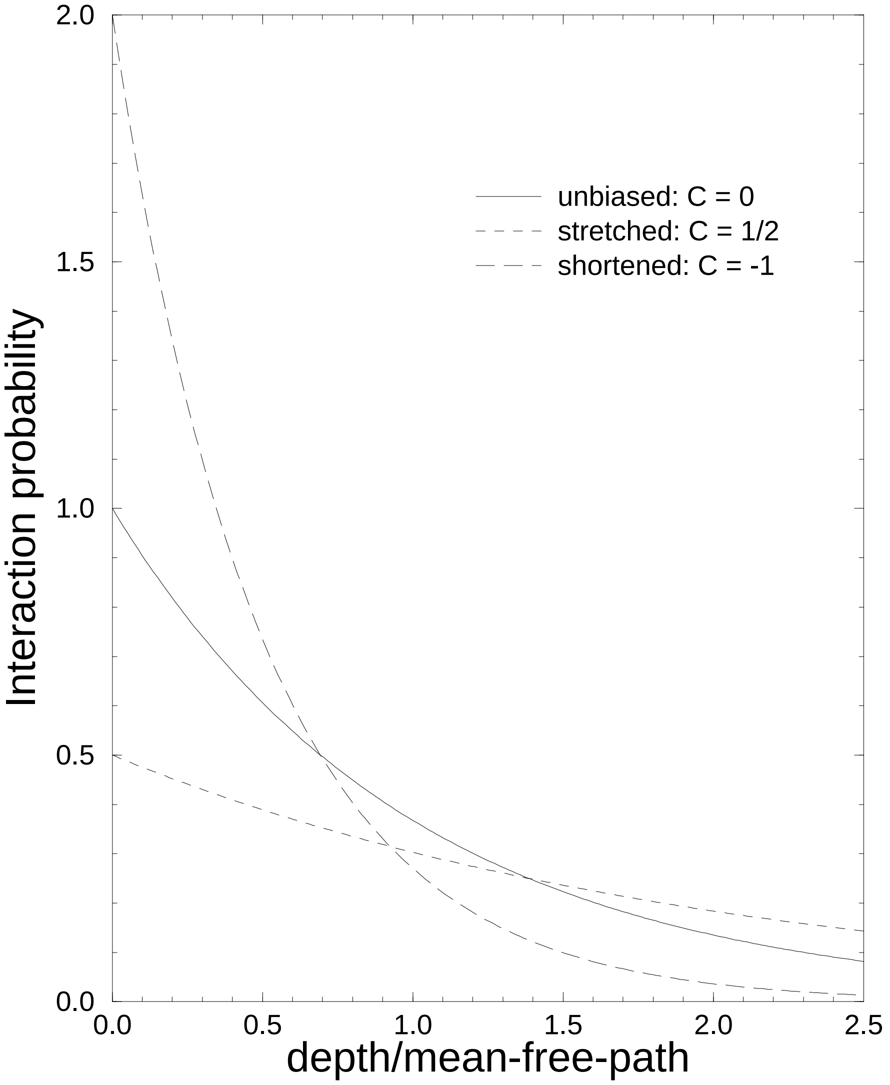

Figure 1: Examples of a stretched ($C = 1/2$) and shortened ($C = -1$) distribution compared to an unbiased one ($C = 0$). In all three cases, $\mu = 1$. For all three curves $\int_0^\infty \tilde{p}(\lambda) d\lambda$ is unity. The horizontal axis is in units of the number of mean free path's (mfp's).

biasing, *i.e.* $C \leq -1$. If one used $C \leq -1$ indiscriminately, then nonsense would result for particles going in the backward direction, *i.e.* $\mu < 0$. Sampled distances and weighting factors become negative. It is possible to use $C \leq -1$ for special, but important cases. (As we shall see in the next section, it is possible to remove all restrictions on C in finite geometries by combining exponential transforms and interaction forcing.) If one restricts the biasing to the incident photons which are directed along the axis of interest (*i.e.* $\mu > 0$) then $C \leq -1$ may be used. If one uses this severe biasing, then as seen in eq. 10, weighting factors for the occasional photon that penetrates very deeply can get very large. If this photon backscatters and interacts in the surface region where one is interested in gaining efficiency, the calculated variance can be undesirably increased. It is advisable to use a "splitting" technique [1], dividing these large weight particles into a N smaller ones each with a new weight, $\omega' = \omega/N$ if they threaten to enter the region of interest. Thresholds for activating this splitting technique and splitting fractions are difficult to specify and choosing them is largely a matter of experience with a given type of application. The same comment applies when particle weights become vary small. If this happens and the photon is headed away from the region of interest it is advisable to play "russian roulette" [1]. This technique works as follows: Select a random number. If this random number lies above a threshold, say α, the photon is discarded without scoring any quantity of interest. If the random number turns out to be below α the photon as allowed to "survive" but with a new weight, $\omega' = \omega/\alpha$, insuring the fairness of the Monte Carlo "game". This technique of "weight windowing" is recommended for use with the exponential transform [13] to save computing time and to avoid the unwanted increase in variance associated with large weight particles.

Russian roulette and splitting[3] can be used in conjunction with exponential transform, but they enjoy much use by themselves in applications where the region of interest of a given application comprises only a fraction of the geometry of the simulation. Photons are "split" as they approach a region of interest and made to play "russian roulette" as they recede. The three techniques, exponential transform, russian roulette and particle splitting are part of the "black art" of Monte Carlo. It is difficult to specify more than the most general guidelines on when they would be expected to work well. One should test them before employing them in large scale production runs.

Finally, we conclude this section with an example of severe exponential transform biasing with the aim to improve surface dose in the calculation of a photon depth dose curve [11]. In this case, 7 MeV γ's were incident normally on a 30 cm slab of water. The results are summarised in Table 1. In each case the computing time was the same. Therefore, the relative efficiency reflects the relative values of $1/s^2$. As C decreases, the calculational efficiency for scoring dose at the surface increases while, in general, it decreases for the largest depth bin. The efficiency was defined to be unity for $C = 0$ at the for each bin. For the deepest bin there is an increase initially because the mean free path is 39 cm. At first the number of interactions in the 10 cm–30 cm bin increases! Note that as C is deceased the number of histories per given amount of computing time decreases. This is because more electrons are being set it motion, primarily at the surface. These electrons have smaller weights, however, to make the "game" fair.

3.3 Exponential transform with interaction forcing

If the geometry in which the transport takes place is finite in extent, one may eliminate restrictions on the biasing parameter, C, by combining exponential transform with interaction

[3] According to Kahn [1], both the ideas and terminology for russian roulette and splitting are attributable to J. von Neumann and S. Ulam.

Table 1: This series of calculations examines a case where a gain in the computational efficiency at the surface is desired. Each calculation took the same amount of computing time. In general, efficiency at the surface increases with decreased C while efficiency worsens at depth.

C	Histories	Relative efficiency on calculated dose		
	10^3	0–0.25 cm	6.0–7.0 cm	10–30 cm
0	100	≡1	≡1	≡1
-1	70	1	1.0	3.5
-3	55	1.5	1.2	0.6
-6	50	3.5	2.8	0.1

forcing. By using the results of the previous two sections we find the interaction probability distribution to be:

$$p(\lambda)d\lambda = \frac{(1 - C\mu)e^{-\lambda(1-C\mu)}}{1 - e^{-\Lambda(1-C\mu)}}d\lambda. \tag{12}$$

The new weighting factor is:

$$\omega' = \omega\frac{(1 - e^{-\Lambda(1-C\mu)})e^{-\lambda C\mu}}{1 - C\mu}, \tag{13}$$

and the number of mean free paths is selected according to:

$$\lambda = -\frac{\ln(1 - \xi(1 - e^{-\Lambda(1-C\mu)}))}{1 - C\mu}, \tag{14}$$

where ξ is a random number chosen uniformly over the range, $0 < \xi \leq 1$.

In the case $C \to 0$, eqs. 12–14 reduce to the equations of simple interaction forcing given in sec. 3.1. In the case $\Lambda \to \infty$, eqs. 12–14 reduce to the equations of exponential transform given in the previous section. However, the equations of this section permit any value of C to be used irrespective of the photon's direction as long as the geometry is finite, *i.e.* $0 < \Lambda < \infty$. In particular, the strong surface biasing, $C < -1$ need not be restricted to forward directed photons ($\mu > 0$), and penetration problems may use $C > 1$. This latter choice actually causes the interaction probability to *increase* with depth for forward directed photons! Again, as in the previous section, the same comments about particle splitting, russian roulette, and weight windowing apply.

4 General methods

4.1 Secondary particle enhancement

In some applications, one wishes to study the behaviour of secondary particles in an energy regime where they are highly suppressed. For example, X-rays from diagnostic X-ray tubes arise from bremsstrahlung radiation. The bremsstrahlung cross section is much smaller than the Møller cross section in the diagnostic regime (≈70 keV). So, calculating the bremsstrahlung characteristics by Monte Carlo method can be difficult since most of the effort is spent creating knock-on electrons. Another example would be the calculation of the effect of pair production in low-Z materials in the radiotherapy regime, below 50 MeV.

One approach is to enhance the number of these secondary particles by creating many of them, say N, once an interaction takes place and then giving them all a weight of $1/N$ to keep the game "fair". Once the interaction occurs, the secondary energy and directional probabilities can be sampled to produce distributions in energy and angle of the secondary particles emanating from a single interaction point. This method is more sophisticated than "splitting" where N *identical* particles are produced.

It is important that the stochastic nature of the primary particle be preserved. For this reason, the energy deducted from the primary particle is *not* the average of the secondary particles produced. The proper "straggling" is guaranteed by subtracting the entire energy of *one* of the secondary particles. This has the minor disadvantage that energy conservation is violated for the incident particle history that produces the "spray" of secondaries. However, over many histories and many interactions, energy conservation is preserved in an average sense.

The details of the implementation this method for the bremsstrahlung interaction in the EGS4 code is documented elsewhere [14]. Examples of the use of this method in the radiotherapy regime [15] and the diagnostic regime [16] have been published.

4.2 Sectioned problems, use of pre-computed results

One approach to saving computer time is to split the problem into separate, manageable parts using the results of a previous Monte Carlo simulations as part of another simulation. These applications tend to be very specialised and unique problems demand unique approaches. For illustration, we shall present two related examples.

Fluence to dose conversion factors for monoenergetic, infinitely broad electron and photon beams incident normally on semi-infinite slabs of tissue and water have been calculated previously [12, 17]. These factors, called $K_E(z)$, vary with depth, z, and on the energy of the photon beam, E, at the water surface. Dose due to an arbitrary incident spectrum as a function of depth, $D(z)$, is calculated from the following relation:

$$D(z) = \int_{E_{\min}}^{E_{\max}} \Phi(E) K_E(z) dE, \tag{15}$$

where $\Phi(E)$ is the electron or photon fluence spectrum and it is non-zero between the limits of $E_{\min}$ and $E_{\max}$. Each K_E array represents a long calculation. If one uses these pre-calculated factors, one can expect orders of magnitude gains in efficiency. If one is interested in normally incident broad beams only, the calculated results should be quite accurate. The only approximations arise from the numerical integration represented by eq. 15 and associated interpolation errors. However, there are two important assumptions buried in the K_E's—the incident beams are *broad* and incident *normally*. For photons, using narrow beams in this method can cause 10% to 50% overestimates of the peak dose. For narrow electron beams this method is not recommended at all.

Another example is the study of the effects of scatter in a ^{60}Co therapy unit [18]. For the purpose of modeling the therapy unit in a reasonable amount of computing time, it was divided into two parts. First, the source capsule itself was modeled accurately and the phase space parameters (energy, direction, position) of those particles leaving the source capsule and entering the collimator system were stored. About 2×10^6 particles were stored in this fashion taking about 24 hrs of VAX 11/780 CPU time for executing the simulation. This data was then used repeatedly in modeling the transport of particles through the collimators and filters of the therapy head. The approximation inherent in this stage of the calculation is the interaction between the source capsule and the rest of the therapy head. However, since the capsule is small with respect to the therapy head and we are interested in calculating the effects of the radiation somewhat downstream from the therapy head, the approximation is an excellent one. Another aspect of this calculation was that the effect of the contaminant electrons downstream from the therapy head was studied. Again, this part of the calculation was "split off" and done by the method described previously. That is, eq. 15 was used to calculate the depth-dose profiles in tissue.

By splitting the problem into 3 parts, the total amount of CPU time used to simulate the ^{60}Co therapy head [18] required 5–16 hours of CPU time for each geometry. If we had attempted to simulate the problem entirely without "dividing and conquering", the amount of CPU time required would have been prohibitive.

4.3 Geometry equivalence theorem

A special but important subset of Monte Carlo calculations is normal beam incidence on semi-infinite geometries, with or without infinite planar inhomogeneities. The use of a simple theorem, called the "geometry equivalence" or "reciprocity" theorem, provides an elegant technique for calculating some results more quickly. First we prove the theorem.

Imagine that we have a zero radius beam coincident with the z-axis impinging on the geometry described above. We "measure" a response that must have the form $f(z, |\boldsymbol{\rho}|)$, where $\boldsymbol{\rho}$ is the cylindrical radius. This functional form holds true since there is no preferred azimuthal direction in the problem. If the beam is now shifted off the axis by an amount $\boldsymbol{\rho}'$, then the new functional form of the response must have the form, $f(z, |\boldsymbol{\rho} - \boldsymbol{\rho}'|)$, by translational symmetry. Finally, consider that we have a finite circular beam of radius $\rho_{\rm b}$ and we wish to integrate the response over a finite-size detection region with circular radius $\rho_{\rm d}$. This integrated response has the form,

$$F(z, \rho_{\rm b}, \rho_{\rm d}) = \int^{|\boldsymbol{\rho}'| \le \rho_{\rm b}} d\rho' \int^{|\boldsymbol{\rho}| \le \rho_{\rm d}} d\rho \; f(z, |\boldsymbol{\rho} - \boldsymbol{\rho}'|), \tag{16}$$

where $\int^{|\boldsymbol{\rho}| \le \rho_{\rm d}} d\rho$ is shorthand for $\int_0^{2\pi} d\phi \int_0^{\rho_{\rm d}} d\rho$. If we exchange integration indices in eq. 16, then we obtain the reciprocity relationship,

$$F(z, \rho_{\rm b}, \rho_{\rm d}) = F(z, \rho_{\rm d}, \rho_{\rm b}). \tag{17}$$

What eq. 17 means is the following: If we have a circular beam of radius $\rho_{\rm b}$ and a circular detection region of radius $\rho_{\rm d}$, then the response we calculate is the same if we had a circular beam of radius $\rho_{\rm b}$ and a circular detection region of radius $\rho_{\rm d}$! The gain in efficiency comes when we wish to calculate the response of a small detector in a large area beam. If one does the calculation directly, then much computer time is squandered tracking particles that may never reach the detector. By using the reciprocity theorem one calculates the same quantity faster.

In an extreme form the reciprocity theorem takes the form [19],

$$\lim_{\epsilon \to 0} F(z, \rho_{\rm b}, \epsilon) = \lim_{\epsilon \to 0} F(z, \epsilon, \rho_{\rm b}), \tag{18}$$

which allows one to calculate the "central axis" depth-dose for a finite radius beam by scoring the dose in a finite region from a zero-area beam. The gain in efficiency in this case is infinite! The radius, $\rho_{\rm b}$, can even be infinite to simulate a broad beam.

A few remarks about the reciprocity theorem and it's derivation should be made. If the response function, $f(z, |\boldsymbol{\rho}|)$, has a finite lateral extent, then the restriction that the geometry should be semi-infinite may be relaxed as long as the geometry, including the inhomogeneous slabs, is big enough to contain all of the incident beam once the detection region radius and the beam radius are exchanged. Unfortunately, electron-photon beams always produce infinitely wide response functions owing to radiation scatter and bremsstrahlung photon creation. In practice, however, the lateral tails often contribute so little that simulation (and experiments!) in finite geometries is useful. Also, in the above development it was assumed that the detection region was infinitely thin. This is not a necessary approximation but this detail was omitted for clarity. The interested reader is encouraged to repeat the derivation with a detection region of finite extent. The derivation proceeds in the same manner but with more cumbersome equations.

4.4 Use of geometry symmetry

In the previous section, we saw that the use of some of the inherent symmetry of the geometry realised considerable increase in efficiency. Some uses of symmetry are more obvious, for example, the use of cylindrical-planar or spherical-conical simulation geometries if both the source and target configurations contain these symmetries. Other uses of symmetry are less obvious but still important. These applications involve the use of reflecting planes to mimic some of the inherent symmetry.

For example, consider the geometry depicted in fig. 2. In this case, an infinite square lattice of cylinders is irradiated uniformly from the top. The cylinders are all uniform and aligned. How should one approach this problem? Clearly, one can not model an infinite array of cylinders. If one tried, one would have to pick a finite set and decide somehow that it was big enough. Instead, it is much more efficient to exploit the symmetry of the problem. It turns out that in this instance, one needs to transport particles in only 1/8'th of a cylinder! To see this we find the symmetries in this problem. In fig. 2 we have drawn three planes of symmetry in the problem, planes **a**, **b**, and **c**[4]. There is reflection symmetry for each of these planes. Therefore, to mimic the infinite lattice, any particles that strike these reflecting planes should be reflected. One only needs to transport particles in the region bounded by the reflecting planes. Because of the highly symmetric nature of the problem, we only need to perform the simulation in a portion of the cylinder and the "response" functions for the rest of the lattice is found by reflection.

The rule for particle reflection about a plane of arbitrary orientation is easy to derive. Let $\vec{\mathbf{u}}$ be the unit direction vector of a particle and $\vec{\mathbf{n}}$ be the unit direction normal of the reflecting plane. Now divide the particle's direction vector into two portions, $\vec{\mathbf{u}}_{||}$, parallel to $\vec{\mathbf{n}}$, and $\vec{\mathbf{u}}_{\perp}$, perpendicular to $\vec{\mathbf{n}}$. The parallel part gets reflected, $\vec{\mathbf{u}}'_{||} = -\vec{\mathbf{u}}_{||}$, and the perpendicular part remains unchanged, $\vec{\mathbf{u}}'_{\perp} = \vec{\mathbf{u}}_{\perp}$. That is, the new direction vector is $\vec{\mathbf{u}}' = -\vec{\mathbf{u}}_{||} + \vec{\mathbf{u}}_{\perp}$. Another way of writing this is,

$$\vec{\mathbf{u}}' = \vec{\mathbf{u}} - \mathbf{2}(\vec{\mathbf{u}} \cdot \vec{\mathbf{n}})\vec{\mathbf{n}}. \tag{19}$$

Applying eq. 19 to the problem in fig. 2, we have: For reflection at plane **a**, $(u'_x, u'_y, u'_z) = (-u_x, u_y, u_z)$. For reflection at plane **b**, $(u'_x, u'_y, u'_z) = (u_x, -u_y, u_z)$. For reflection at plane **c**, $(u'_x, u'_y, u'_z) = (-u_y, -u_x, u_z)$. The use of this reflection technique can result in great gains in efficiency. Most practical problems will not enjoy such a great amount of symmetry but one is encouraged to make use of any available symmetry. The saving in computing time is well worth the extra care and coding.

[4] Note that this symmetry applies only to a square lattice, where the spacing is the same for the x and y-axes. For a rectangular symmetry, the planes of reflection would be somewhat different.

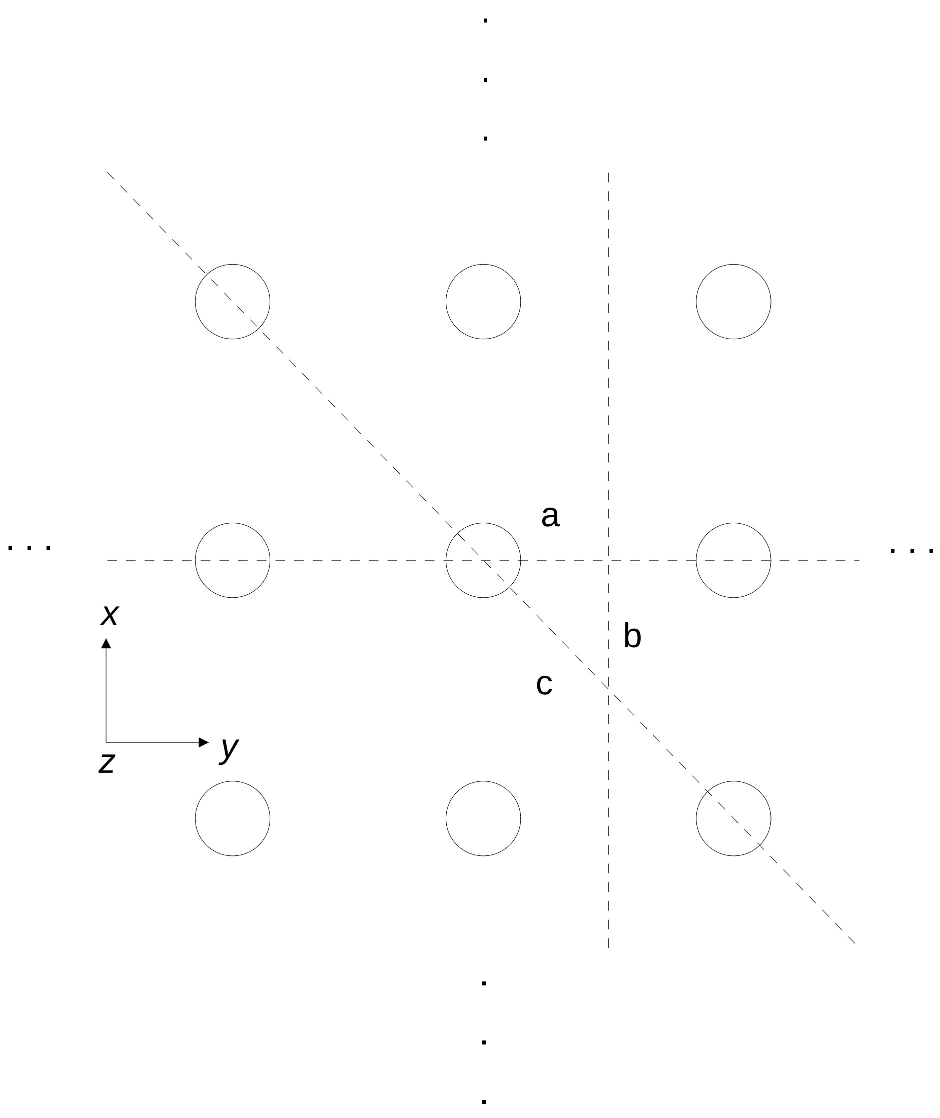

Figure 2: Top end view of an infinite square lattice of cylinders. Three planes of symmetry are drawn, **a**, **b**, and **c**. A complete simulation of the entire lattice may be performed by restricting the transport to the interior of the three planes. When a particle strikes a plane it is reflected back in, thereby mimicking the symmetry associated with this plane.

References

[1] H. Kahn, *Use of different Monte Carlo sampling techniques*, in Symposium on Monte Carlo Methods, (H.A. Meyer ed.) (John Wiley and Sons, New York) 146 – 190 (1956).

[2] J.M. Hammersley and D.C. Handscomb, *Monte Carlo Methods*, (John Wiley and Sons, New York) (1964).

[3] E.J. McGrath and D.C. Irving, *Techniques for Efficient Monte Carlo Simulation, Vols. I, II, and III*, Report ORNL-RSIC-38, Radiation Shielding Information Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee (1975).

[4] G.A. Carlsson, *Effective Use of Monte Carlo Methods*, Report ULi-RAD-R-049, Department of Radiology, Linköping University, Linköping, Sweden (1981).

[5] T. Lund, *An Introduction to the Monte Carlo Method*, Report HS-RP/067, CERN, Geneva (1981).

[6] A.F. Bielajew, *Electron Transport in $\vec{E}$ and $\vec{B}$ Fields*, in "Monte Carlo Transport of Electrons and Photons Below 50 MeV", eds. T.M. Jenkins, W.R. Nelson, A. Rindi, A.E. Nahum and D.W.O. Rogers, (Plenum Press) 421 – 434 (1989).

[7] W.R. Nelson, H. Hirayama and D.W.O. Rogers, *The EGS4 Code System*, Stanford Linear Accelerator Center Report SLAC-265 (Stanford Calif) (1985).

[8] D.W.O. Rogers, *Low energy electron transport with EGS*, Nucl. Inst. Meth. **227** 535 – 548 (1984).

[9] A.F. Bielajew and D.W.O. Rogers, *PRESTA: The Parameter Reduced Electron-Step Transport Algorithm for Electron Monte Carlo Transport*, Nuclear Instruments and Methods **B18** 165 – 181 (1987).

[10] A.F. Bielajew, D.W.O. Rogers and A.E. Nahum, *Monte Carlo Simulation of Ion Chamber Response to ^{60}Co – Resolution of Anomalies Associated with Interfaces,*, Phys. Med. Biol. **30** 419 – 428 (1985).

[11] D.W.O. Rogers and A.F. Bielajew, *The Use of EGS for Monte Carlo Calculations in Medical Physics*, Report PXNR-2692, National Research Council of Canada, (Ottawa, Canada K1A 0R6) (1984).

[12] D.W.O. Rogers and A.F. Bielajew, *Calculated buildup curves for photons with energies up to ^{60}Co* , Med. Phys. **12** 738 – 744 (1985).

[13] J. S. Hendricks and T. E. Booth, *Monte-Carlo Methods and Applications in Neutronics, Photonics and Statistical Physics*, (R Alcouffe, R Dautray, A Forster, G Ledanois, and B Mercier eds.) 83 (1985).

[14] A.F. Bielajew, R. Mohan and C.S. Chui, *Improved bremsstrahlung photon angular sampling in the EGS4 code system*, National Research Council of Canada Report PIRS-0203 (1989).

[15] B.A. Faddegon, C.K. Ross and D.W.O. Rogers, *Forward Directed Bremsstrahlung of 10 – 30 MeV Electrons Incident on Thick Targets of Al and Pb*, Medical Physics **17** 773 – 785 (1990).

[16] Y. Namito, W.R.Nelson, S.M. Seltzer, A.F. Bielajew and D.W.O. Rogers, *Low-energy x-ray production studies using the EGS4 code system* , Med. Phys. **17** (abstract) 557 (1990).

[17] D.W.O. Rogers, *Fluence to Dose Equivalent Conversion Factors Calculated with EGS3 for Electrons from 100 keV to 20 GeV and Photons from 20 keV to 20 GeV*, Health Physics **46** 891 – 914 (1984).

[18] D.W.O. Rogers, G.M Ewart, A.F. Bielajew and G. van Dyk, *Calculation of Electron Contamination in a ^{60}Co Therapy Beam*, In "Proceedings of the IAEA International Symposium on Dosimetry in Radiotherapy" (IAEA, Vienna), Vol 1 303 – 312 (1988).

[19] ICRU, *Radiation Dosimetry: Electron beams with energies between 1 and 50 MeV*, ICRU Report 35, Bethesda MD (1984).