Chapter 2

An Introduction to Rare Event Simulation and Importance Sampling

Gino Biondini*,[1]
*Department of Mathematics, State University of New York at Buffalo, Buffalo, New York, USA
[1]Corresponding author: e-mail: biondini@buffalo.edu

ABSTRACT
This chapter provides a relatively low-level introduction to the problem of rare event simulation with Monte Carlo methods and to a class of methods known as variance reduction techniques that have been devised to deal with this problem. Special emphasis is given to importance sampling, but several other techniques are also presented, including the cross-entropy method, rejection sampling, and Markov chain Monte Carlo methods such as the Metropolis method and Gibbs sampling. A brief discussion is also given about asymptotic efficiency and the connections with large deviations theory.

Keywords: Monte Carlo methods, Rare event simulation, Variance reduction techniques, Importance sampling, Cross-entropy 2000 *MSC:* 65C05, 65B99

1 INTRODUCTION: MONTE CARLO METHODS, RARE EVENT SIMULATION, AND VARIANCE REDUCTION TECHNIQUES

Since its introduction almost 70 years ago (Metropolis and Ulam, 1949) (see Metropolis, 1987 for a historical review), the Monte Carlo (MC) method has been extensively used in engineering and scientific computing. In their most general interpretation, MC methods are a way to compute integrals. They comprise a collection of techniques for generating random samples on a computer as well as their application to solve a variety of problems. In essence, they involve drawing random or pseudo-random samples from a specific distribution and using them to estimate one or more quantities of interest. Such methods are especially

Handbook of Statistics, Vol. 33. http://dx.doi.org/10.1016/B978-0-444-63492-4.00002-2

advantageous over numerical quadrature methods when the dimensionality of the problem is large. As a result, and thanks to their flexibility, such methods have found a wide range of applications (e.g., see Fishman, 1996; Fishman, 2006; Kroese et al., 2011; Landau and Binder, 2000).

A common challenge in MC simulation is that of *rare event simulation*, also referred to as the problem of rare events, where very small probabilities need to be accurately estimated—for example, in reliability analysis, or performance analysis of telecommunication systems. In a nutshell, the problem is that if one needs to quantify the probability of one or more events that occur very rarely, an exceedingly large number of samples are needed even to just produce the desired events, and an even larger number of samples are required to obtain accurate estimates. Other applications that call for rare event simulation are queueing systems (to avoid excessively long waiting times), nuclear physics (avoiding catastrophic accident), security systems (false alarms in radar), material science (technical defects), mathematical science, and insurance.

One approach to overcome the problem of rare events is the use of *variance reduction techniques* (VRTs) (e.g., see the monographs: Bucklew, 2004; Fishman, 1996; Kroese et al., 2011 for general reviews). The general idea behind all of these techniques is to modify the selection of the random samples in such a way that the desired events occur more frequently than they would normally, while simultaneously taking these changes into account in order to obtain unbiased estimates.

Perhaps the most famous VRT is the *importance sampling* (IS) (Fishman, 1996; Kroese et al., 2011; Srinivasan, 2002). The main idea behind IS is to select an appropriate *biasing distribution* (i.e., a change of probability measure) from which to draw the MC samples so that most of the distribution mass falls on the regions of interest. This ensures that many of the MC samples will produce the rare events sought. At the same time, the contribution from each sample is weighted according to the *likelihood ratio*, which ensures that unbiased estimates are obtained.

Of course, for IS to be effective, a good biasing distribution must be chosen. This requires knowledge of which system configurations are likely to produce the rare events of interest. Even though such knowledge is not always available, in many cases it is enough to leverage what is known about the system's behavior in order to guide the choice of biasing distribution, and indeed IS has been used with success in a variety of applications (Biondini et al., 2004; Li et al., 2007; Moore et al., 2008; Smith et al., 1997). (Note that, often, exact knowledge of the most likely failure configurations may not be needed, and an approximate knowledge may be sufficient, since the statistical nature of the MC sampling allows one to take into account the contributions of nearby points in sample space.)

Many other VRTs have also been used with success in various applications, such as multicanonical MC methods (Yevick, 2002), Markov chain Monte Carlo (MCMC) methods (Secondini and Forestieri, 2005), and Gibbs sampling. See

Fishman (1996), Landau and Binder (2000), and MacKay (2003) for a general overview of these methods. The common thread among those VRTs is that they are adaptive. In essence, such methods attempt to find the important regions of sample space numerically. These methods can be applied to problems for which no good choice of biasing distribution is known. When IS is available, however, it is generally advantageous over other methods, because: (i) IS allows one to compute precise error estimates, if desired; (ii) adaptive methods typically require tweaking certain parameters, on which IS is less dependent; and (iii) IS is usually faster than adaptive methods, since, in adaptive methods, a certain portion of numerical simulations needs to be used to look for the most important regions in state space. Indeed, the speed advantage of IS was verified directly in a few cases by a detailed comparison between different methods (Biondini and Kath, 2005; Lima et al., 2005). We should also mention that it is not always necessary to choose between IS and adaptive VRTs. Indeed, yet another technique which has proven to be especially useful in recent years is the cross-entropy method (de Boer et al., 2005; Rubinstein and Kroese, 2004). While it is a useful VRT on its own right, in some cases, the cross-entropy method can also be combined with IS to afford the user the advantages of both IS and those of adaptive techniques.

The remainder of this chapter aims to put the above discussion on a more precise mathematical setting.

2 MC METHODS AND THE PROBLEM OF RARE EVENTS

2.1 MC Estimators

We start with a simple one-dimensional (1D) example. Let X be a random variable (RV) with probability density function (pdf) $p_x(x)$ (Papoulis, 1991). If one defines $Y = y(X)$, where $y(x) = \int_{-\infty}^{x} p_x(x)\,\mathrm{d}x$, it is easy to show that Y is uniform in [0,1]. [To see this, note that $p_y(y)\,\mathrm{d}y = p_x(x)\,\mathrm{d}x$, with $\mathrm{d}y = (\mathrm{d}y/\mathrm{d}x)\,\mathrm{d}x$. But $\mathrm{d}y/\mathrm{d}x = p_x(x)$, so $p_y(y) = 1$.]

Now suppose that we wish to calculate the probability Q that X falls in a range R of interest, namely $Q = \mathbb{P}[\,X \in R\,]$, where $R \subset \mathbb{R}$. We can write Q as

$$Q = \int I_R(x) p_x(x)\,\mathrm{d}x\,. \tag{1}$$

The function $I_R(x)$ is the so-called *indicator function* (or characteristic function) of the set R: namely, $I_R(x) = 1$ if $x \in R$ and $I_R(x) = 0$ otherwise. (Hereafter we will drop the subscript R on I whenever that will not cause ambiguity. Also, integrals without limits are always intended as complete—i.e., over all of sample space—unless specifically noted otherwise.) In particular, we are interested in situations in which it is difficult to compute the above integral analytically.

Making the substitution $x \mapsto y$, we can express Q as $Q = \int I(x(y))\,\mathrm{d}y$. It is therefore natural to try to estimate Q using a frequency count. That is, we draw N independent identically distributed (i.i.d.) random samples $Y_1, \ldots, Y_N$ from a uniform distribution and we write the estimate $\hat{Q}_N = F/N$, where F is the number of samples which fall in the region of interest. More specifically, the above MC estimator is $\hat{Q}_N = (1/N)\sum_{n=1}^{N} I(x(Y_n))$. Equivalently, we can forget about Y and write the above estimator as

$$\hat{Q}_N = \frac{1}{N}\sum_{n=1}^{N} I(X_n)\,, \tag{2}$$

where the i.i.d. random samples $X_1, \ldots, X_N$ are drawn according to the distribution $p_x(x)$. Note that, while Q is a deterministic quantity, $\hat{Q}_N$ is itself a RV. In fact, it is easy to show that

$$\mathbb{E}[\hat{Q}_N] = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}[I(X_n)] = \mathbb{E}[I(X)] = Q\,,$$

where $\mathbb{E}[Z] = \int Z(x)p_x(x)\,\mathrm{d}x$ denotes the expectation value with respect to the pdf $p_x(\cdot)$, which shows that the expectation value of our estimator is indeed the quantity of interest, and

$$\begin{aligned}\mathrm{var}[\hat{Q}_N] &= \mathbb{E}[\hat{Q}_N^2] - \mathbb{E}[\hat{Q}_N]^2 = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} \mathbb{E}[I(X_n)I(X_m)] - Q^2 \\ &= \frac{1}{N}\mathrm{var}[I(X)]\,,\end{aligned}$$

where we used the fact that $\mathbb{E}[I(X_n)^2] = \mathrm{var}[I(X)] + Q^2$ and $\mathbb{E}[I(X_n)I(X_m)] = Q^2$ when $n \neq m$ (because X_n and X_m are statistically independent). Note that the above two results are true more generally, i.e., independently of $I(\cdot)$ being an indicator. For an indicator function, in particular, it is $\mathbb{E}[I(X_n)^2] = Q$ (because $I^2(x) = I(x)$) and therefore

$$\mathrm{var}[\hat{Q}_N] = \frac{1}{N}(Q - Q^2)\,.$$

The above results are easily extended to the multidimensional case. Let $\mathbf{X} = (X_1, \ldots, X_D)^T$ be a vector of RVs with joint pdf $p_{\mathbf{x}}(\mathbf{x})$, and suppose that we are interested in the probability $Q = \mathbb{P}[\, y(\mathbf{X}) \in R\,]$, where $y(\mathbf{x})$ is some real-valued function:

$$Q = \int I_R(y(\mathbf{x}))p_{\mathbf{x}}(\mathbf{x})(\mathrm{d}\mathbf{x})\,, \tag{3}$$

where $(\mathrm{d}\mathbf{x}) = \mathrm{d}x_1 \cdots \mathrm{d}x_D$ is the volume element in $\mathbb{R}^D$. More generally, consider integrals of the type

$$Q = \int f(\mathbf{x}) p_{\mathbf{x}}(\mathbf{x})(\mathrm{d}\mathbf{x}) \,, \tag{4}$$

where $f(\mathbf{x})$ is a generic real-valued function. Situations for which calculating the above integral analytically is practically impossible are very common: The dimensionality of the system might be very large, the function $f(\cdot)$ might be complicated, and/or the region R might be complicated.

By analogy with the 1D case, we can define the MC estimator

$$\hat{Q}_N = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{X}_n) \,. \tag{5}$$

As in the 1D case, we have

$$\mathbb{E}[\hat{Q}_N] = \mathbb{E}[f(\mathbf{X})] = Q \,, \quad \mathrm{var}[\hat{Q}_N] = \frac{1}{N} \mathrm{var}[f(\mathbf{X})] \,. \tag{6}$$

Then, if in particular $f(\mathbf{X}) = I(y(\mathbf{x}))$, we have $\mathrm{var}[\hat{Q}_N] = (Q - Q^2)/N$. The above result implies that the accuracy of a MC estimator is simply proportional to $1/\sqrt{N}$ *independently of the number of dimensions*. This is one of the main advantages of MC methods to compute multidimensional integrals compared to deterministic integration methods.

In passing, we note that

$$\mathrm{var}[f(\mathbf{X})] = \int \big(f(\mathbf{x}) - Q\big)^2 p_{\mathbf{x}}(\mathbf{x})(\mathrm{d}\mathbf{x}) \,.$$

But since in practice we do not know the theoretical variance, we can define a MC estimator for it:

$$\hat{\sigma}_N^2 = \frac{1}{N-1} \sum_{n=1}^{N} \big(f(\mathbf{X}_n) - \hat{Q}_N\big)^2 \,.$$

(The $N-1$ in the denominator is necessary for $\hat{\sigma}_N^2$ to be an unbiased estimator, i.e., so that $\mathbb{E}[\hat{\sigma}_N^2] = \mathrm{var}[f(\mathbf{x})]$.) Note also that an efficient way to compute $\hat{\sigma}_n^2$ is to use the recursion relation

$$(n-1)\,\hat{\sigma}_n^2 = (n-2)\,\hat{\sigma}_{n-1}^2 + \left(1 - \frac{1}{n}\right)\big(f(\mathbf{X}_n) - \hat{Q}_{n-1}\big)^2 \,.$$

Using this formula, one can compute both the sample mean and variance in a single iteration.

As a simple example of an application of MC methods, one can approximate the value of π as follows. The area of the portion of the unit disk in the first quadrant is $\pi/4$. We can write this area as an integral of the form (3), where $\mathbf{x} = (x_1, x_2)$ and $p_{\mathbf{x}}(\mathbf{x}) \equiv 1$ (i.e., x_1 and x_2 are independent uniform RVs in [0,1]), and with $y(\mathbf{x}) = \|\mathbf{x}\|^2 = x_1^2 + x_2^2$ and $R = \{y \in \mathbb{R} : 0 \le y \le 1\}$. We can then estimate this integral with MC methods by simply taking random samples and counting the fraction of samples that fall inside the disk.

2.2 The Problem of Rare Events

While the variance of a MC estimator provides an absolute measure of how widely distributed it is around its mean value, in most cases a relative measure of the relative accuracy of the MC estimator is more useful. Such a measure is provided by the *coefficient of variation* (cv) of a RV Z, which is defined as

$$\mathrm{cv}[Z] = \mathrm{stdev}[Z] \,/\mathbb{E}[Z]\,,$$

where as usual $\mathrm{stdev}[Z] = \sqrt{\mathrm{var}[Z]}$ is the standard deviation. More precisely, the cv gives the number of samples that are necessary on average to achieve a given accuracy.

To apply this concept in our case, suppose $Z = \hat{Q}_N$. Since $\mathrm{var}[\hat{Q}_N] = \mathrm{var}[f(X)]/N$, we have $\mathrm{cv}[\hat{Q}_N] = \mathrm{cv}_Q/\sqrt{N}$, where (with some abuse of notation) we denoted $\mathrm{cv}_Q = \mathrm{stdev}[f]/Q$. Therefore, if we want $\mathrm{cv}[\hat{Q}]$ to be below a target value cv_o, on average we will need $N > (\mathrm{cv}_Q/\mathrm{cv}_o)^2$. In particular, for an indicator function the above calculations yield $\mathrm{cv}[\hat{Q}] = \sqrt{(1-Q)/(NQ)}$.

We can now see the problem of rare event simulation in a more quantitative way: If $Q \ll 1$, the number of samples needed on average to obtain a given value of cv is $N \sim 1/(Q\mathrm{cv}_o^2)$. For example, if $Q \sim 10^{-6}$ and we want a cv of 0.1, we need $N = 10^8$ samples. In other words, the problem is that, if $Q \ll 1$, the events that make $I(y(\mathbf{x})) = 1$ have a very low probability of occurring, and therefore, a large number of samples is needed even to observe one such event, and an even larger number of samples is needed to obtain a reliable estimate.

As mentioned in Section 1, VRTs are a collection of methods aimed at overcoming (or at least alleviating) this problem. In the following, we will look in some detail at two of them, namely IS and the cross-entropy method.

3 IMPORTANCE SAMPLING

3.1 Importance-Sampled MC Estimators

As mentioned earlier, the idea behind IS is simple: We want to improve the efficiency of MC methods by pushing (biasing) the simulations to favor the rare events of interest so that they will occur more frequently than they would otherwise. Of course, we must do this in a proper way in order to still have an unbiased estimator (i.e., an estimator whose expectation value is still the quantity of interest).

We do so by introducing a modified density $p_*(\mathbf{x})$, called the *biasing distribution*, and by rewriting the integral in Eq. (4) that defines the quantity of interest as

$$Q = \int f(\mathbf{x}) L(\mathbf{x}) p_*(\mathbf{x}) \,(\mathrm{d}\mathbf{x})\,. \tag{7}$$

The ratio $L(\mathbf{x}) = p_{\mathbf{x}}(\mathbf{x})/p_*(\mathbf{x})$ is called the *importance function* or *likelihood ratio* (or even weight function in some works). An equivalent way to write the integral in Eq. (7) is

$$Q = \mathbb{E}_*[f(\mathbf{X})L(\mathbf{X})]\,, \tag{8}$$

where $\mathbb{E}_*[\cdot]$ denotes expectation values with respect to the density $p_*(\mathbf{x})$. We then define an importance-sampled MC estimator as

$$\hat{Q}_N^* = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{X}_n^*)L(\mathbf{X}_n^*)\,,$$

where the samples $\mathbf{X}_n^*$ are now drawn from the biasing distribution $p_*(\mathbf{x})$. Importantly, note that a necessary requirement in order to carry out the change of measure from (4) to (7) is that the support of $f(\mathbf{x})p_*(\mathbf{x})$ includes that of $f(\mathbf{x})p(\mathbf{x})$ [i.e., $f(\mathbf{x})p_*(\mathbf{x}) \neq 0$ whenever $f(\mathbf{x})p(\mathbf{x}) \neq 0$]. Otherwise $\hat{Q}_N^*$ will not converge to the correct value in general. Conversely, it should be obvious that, as long as this condition is satisfied, $\mathbb{E}_*[\hat{Q}_N^*] = Q$, thanks to Eq. (8). That is, we still have an unbiased estimator for the quantity of interest.

The reason for biasing the sampling distribution can be seen by looking at the variance of our new estimator, namely

$$\mathrm{var}_*[f(\mathbf{X})L(\mathbf{X})] = \int \big(f(\mathbf{x})L(\mathbf{x}) - Q\big)^2 p_*(\mathbf{x})(\mathrm{d}\mathbf{x}) = \int f^2(\mathbf{x})L(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})(\mathrm{d}\mathbf{x}) - Q^2. \tag{9}$$

Thus,

$$\mathrm{var}[f(\mathbf{X})] - \mathrm{var}_*[f(\mathbf{X})L(\mathbf{X})] = \int f^2(\mathbf{x})\big(1 - L(\mathbf{x})\big)\, p_{\mathbf{x}}(\mathbf{x})(\mathrm{d}\mathbf{x})\,. \tag{10}$$

Looking at the integrand in Eq. (9) we see that if $p_*(\mathbf{x}) = f(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})/Q$, we would have $\mathrm{var}_*[f(\mathbf{X})L(\mathbf{X})] = 0$. Thus, in this case our importance-sampled estimator would have zero variance: every sample would always yield the same result, namely the exact value of the quantity Q of interest!

Of course, the above choice of biasing distribution is not practical, because it requires the advance knowledge of the value of Q (which is the desired result). On the other hand, Eq. (10) implies if we can choose $p_*(\mathbf{x})$ so that $p_*(\mathbf{x}) > p_{\mathbf{x}}(\mathbf{x})$ wherever $f^2(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})$ is large and $p_*(\mathbf{x}) < p_{\mathbf{x}}(\mathbf{x})$ wherever $f^2(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})$ is small, the variance of our importance-sampled estimator will be much smaller than the original variance. This corresponds to redistributing the probability mass in accordance with its relative importance as measured by the weight $f^2(\mathbf{x})p_{\mathbf{x}}(\mathbf{x})$. The zero-variance choice is just an ultimate case of this redistribution.

An estimator of the importance-sampled variance can be written using the same methods as before:

$$\hat{\sigma}_N^{*2} = \frac{1}{N-1}\sum_{n=1}^{N}\big(f(\mathbf{X}_n^*)L(\mathbf{X}_n^*) - \hat{Q}_N^*\big)^2\,,$$

which again can be computed recursively as

$$(n-1)\,\hat{\sigma}_n^{*2} = (n-2)\,\hat{\sigma}_{n-1}^{*2} + (1-1/n)\big(f(\mathbf{X}_n^*)L(\mathbf{X}_n^*) - \hat{Q}_{n-1}^*\big)^2 .$$

A case in which the likelihood ratio can be computed particularly easily is the common situation in which the components of both $\mathbf{X}$ and $\mathbf{X}^*$ are statistically independent, for in this case it is $p_{\mathbf{x}}(\mathbf{x}) = \prod_{j=1}^{D} p_{x_j}(x_j)$ and similarly for $p_*(\mathbf{x})$, yielding the likelihood ratio simply as $L(\mathbf{x}) = \prod_{j=1}^{D} p_{x_j}(x_j)/p_{*j}(x_j)$.

Of course, a key question is how to make the choice of a biasing distribution in practice. We will see shortly how IS works in a simple example, but unfortunately there are no general rules that work in all cases, and the task of choosing good biasing distributions is the most difficult step in applying IS. Also note that choosing a bad biasing distribution can make the problem worse and make the variance of the importance-sampled estimator much bigger than the original one. This is why it is occasionally said that IS (like all of computer simulation; Knuth, 2011) is an art. Nonetheless, *there are* general principles that one can follow to select a biasing distribution, and indeed IS has been used with success in a large variety of problems.

3.2 A Simple Example

As an illustration of the concepts discussed above, it will be useful to consider a specific example: a 1D symmetric random walk (RW). Let $\mathbf{X} = (X_1, \ldots, X_D)$ and

$$y(\mathbf{X}) = \sum_{j=1}^{D} X_j ,$$

where, for $j = 1, \ldots, D$,

$$X_j = \begin{cases} +1 & \text{with probability } 1/2 , \\ -1 & \text{with probability } 1/2 . \end{cases}$$

That is, we consider a sequence of D random steps, each one unit to the right or to the left with probability 1/2, and we are interested in computing the final position. In particular, suppose we want to compute the probability that the final position will be to the right of some given threshold:

$$Q = \mathbb{P}[\, y(\mathbf{X}) \geq C \,] .$$

To make things more concrete, suppose $D = 100$ and $C = 70$. This is equivalent to asking the probability that by flipping a coin 100 times we get at least 85 heads.

We can try to estimate the pdf of the final position (and therefore our desired probability) by performing MC simulations. That is, we use Eq. (5) with

$$f(\mathbf{X}) = H(y(\mathbf{X}) - C) ,$$

where $H(\cdot)$ is the Heaviside step function: $H(s) = 1$ for $s > 0$ and $H(s) = 0$ for $s < 0$. The histogram of the final position in a simulation with $N = 100,000$ samples is shown in Fig. 1. The problem is that no samples occurred with final position greater than 50. That, of course, is because our desired event is extremely rare, and therefore, we are very unlikely to see it with a reasonable number of samples.

To obviate this problem, we can simulate a biased RW: given $0 < q < 1$, for $j = 1, \ldots, D$ we take

$$X_j = \begin{cases} +1 & \text{with probability } q\,, \\ -1 & \text{with probability } 1 - q\,. \end{cases}$$

The value $q = 1/2$ reproduces the unbiased case. If $q > 1/2$, however, steps to the right will be more prevalent, which means that we are pushing the final position to the right (which is what we want). The histogram of the final position in a biased simulation with $q = 0.7$ is shown in Fig. 2. The results show we now get a lot more samples with final positions to the right. But of course now we cannot simply take the relative frequency of our event as an estimator of the desired probability, and we need to use the likelihood ratios instead. The individual likelihood ratio for a single step is given by

$$\ell(X_j) = \begin{cases} 1/(2q) & \text{if } X_j = 1\,, \\ 1/[2(1-q)] & \text{if } X_j = -1\,, \end{cases}$$

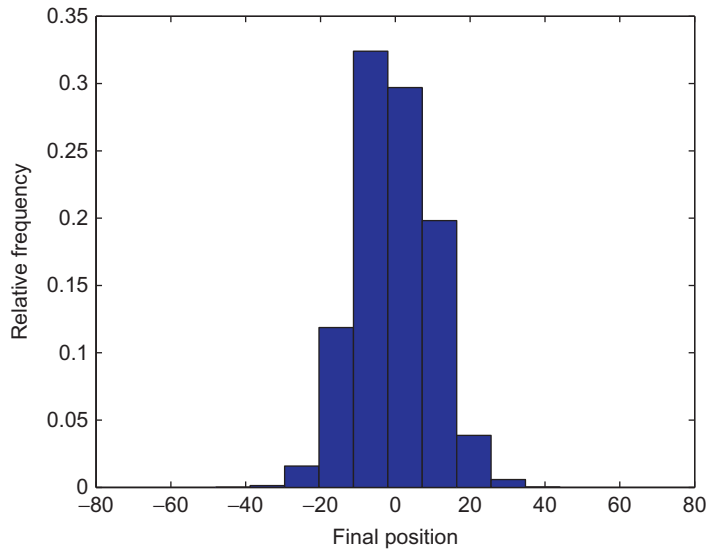

FIGURE 1 Histogram of the final position in an unbiased MC simulation of a symmetric 1D random walk with $N = 100,000$ samples.

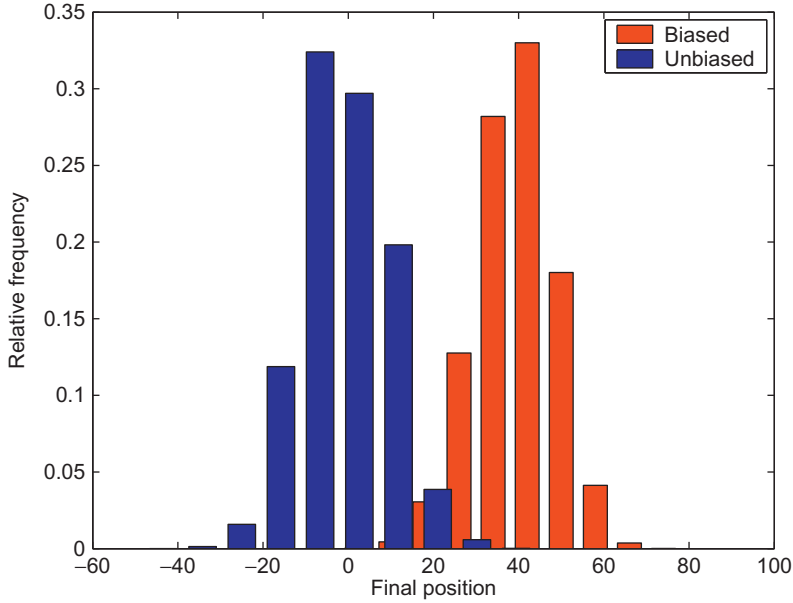

FIGURE 2 Histogram of the final position (in red (light gray in the print version)) in a biased MC random walk with $q = 0.7$, with $N = 100,000$ samples. For comparison, the blue (dark gray in the print version) histogram shows the unbiased case.

and the overall likelihood ratio for a single sample RW is $L(\mathbf{X}) = \prod_{j=1}^{D} \ell(X_j)$. Now recall that if $q > 1/2$, on average there will be more samples for which the final position is to the right. Since $1/(2q) < 1$, we can therefore expect the overall likelihood ratio for those samples to be less than 1 as well. In fact, we will see that the likelihood ratios can get quite small.

The results in Fig. 2 should already demonstrate that use of an appropriate biasing distribution can yield a distinct advantage over standard MC simulations. Roughly speaking, the reason why IS is effective is that, generically, it is much better to estimate a quantity by using a large number of samples each of which contributes a little to the final result (by virtue of the likelihood ratios), rather than using a very small number of result which gives a binary contribution (one or zero). (This is true as long as the contributions are not *too* small, as we will discuss later.) The results, however, also suggest that perhaps one could get an even better result by further increasing the value of q. The natural question is then: What is the optimal value of q? This question raises again the key issue in properly applying IS: How does one choose a good biasing distribution? Usually, this requires some analytical knowledge about the behavior of the system. We turn to this issue next.

3.3 The Optimal Biasing Distribution

The case of a symmetric 1D RW is simple enough that analytical expressions can be derived for the pdf of the final position. Comparison with these analytical results will then provide some insight into the issue of how to best choose a biasing distribution.

It is easy to see that if D is odd and m is even or vice versa, it is $\mathbb{P}[\, y(\mathbf{X}) = m\,] = 0$. If D and m are both even or both odd, instead,

$$\mathbb{P}[\, y(\mathbf{X}) = m\,] = \frac{1}{2^D}\binom{D}{(D+m)/2}.$$

The factor $1/2^D$ arises because we are taking D steps, each of which has a probability 1/2 of being either to the left or to the right. The binomial coefficient arises because, for the final position to equal m, we need a total of $(D+m)/2$ steps to the right and $(D-m)/2$ steps to the left, and there are exactly D choose $(D+m)/2$ ways to arrange this. Taking the sum over all possible results above the threshold, we then simply have

$$Q = \frac{1}{2^D}\sum_{m=C}^{D}{}' \binom{D}{(D+m)/2}, \tag{11}$$

where the prime indicates that the sum should be taken only on even values of m or only on odd values of m depending on whether D is even or odd.

In particular, for $D = 100$ and $C = 70$, we then get $\mathbb{P}[y(\mathbf{X}) \geq C] = 2.4 \times 10^{-13}$. Recalling the discussion about the cv in Section 2.2, we then see that, even in a simple example as this, it would be almost hopeless to try to accurately estimate the desired probability numerically, except perhaps on the fastest supercomputers.

The above discussion, however, does not answer our question of what is the optimal choice of biasing. To answer that question, we need to dig a little deeper. Fortunately, our example is simple enough that we can actually calculate analytically the variance of the biased estimator. Note first that

$$\begin{aligned}\mathrm{var}_*[f(\mathbf{X})L(\mathbf{X})] &= \mathbb{E}_*[f^2(\mathbf{X})L^2(\mathbf{X})] - \big(\mathbb{E}_*[f(\mathbf{X})L(\mathbf{X})]\big)^2 \\ &= \mathbb{E}[f(\mathbf{X})L(\mathbf{X})] - \big(\mathbb{E}[f(\mathbf{X})]\big)^2,\end{aligned}$$

where we used that $f(\cdot)$ is an indicator and we rewrote expectations with respect to $p_*(\mathbf{x})$ as expectations with respect to $p_{\mathbf{x}}(\mathbf{x})$. We then have

$$\mathrm{var}_*[f(\mathbf{X})L(\mathbf{X})] = -Q^2 + \frac{1}{2^D}\sum_{m=C}^{D}{}' \binom{D}{(D+m)/2} \frac{1}{(2q)^{(D+m)/2}\,[2(1-q)]^{(D-m)/2}},$$

where Q is given by Eq. (11). Note that the last part of the sum is precisely the likelihood ratio of a sample with final position m.

We can now look at this variance as a function of q. Even better, we can plot the biased cv; i.e., the ratio $\text{cv}_* = \text{stdev}_*[fL]/Q$. The corresponding results are given in Fig. 3. (Note that these results agree very well with numerical estimates of the variance as a function of q.)

Figure 4 shows that the likelihood ratios when $q > 1/2$ are indeed much smaller than unity, as anticipated. Thus, each sample that ends past the threshold will only give a small contribution to the estimator. (It should be noted that, in our example, the value of the likelihood ratio is the same for all paths that lead to the same value of final position, but this is not true in more general situations.)

From Fig. 3, we see that optimal value of q is 0.85. At that value, the cv is just 2.32, whereas for the unbiased RW ($q = 0.5$) it is 2.04×10^6. Now recall that the number of MC samples needed on average to get a given value of the cv is $N = (\text{cv}_j/\text{cv}_o)^2$ or, for importance-sampled MC, $N = (\text{cv}_*/\text{cv}_o)^2$. Using the optimal value $q = 0.85$, one can therefore obtain a cv of 0.1 using just a few hundred samples. On the other hand, to obtain the same level of accuracy with unbiased MC simulations, one would need over 10^{14} samples. So, in our example IS increases the efficiency of the MC simulations by 10 orders of magnitude! Such a huge increase in efficiency is not a fluke, but has been realized in practical applications (e.g., see Biondini et al., 2004; Marzec et al., 2013; Moore et al., 2008).

The general message that we should take from this example is that the optimal biasing choice is to concentrate the MC samples around the *most likely*

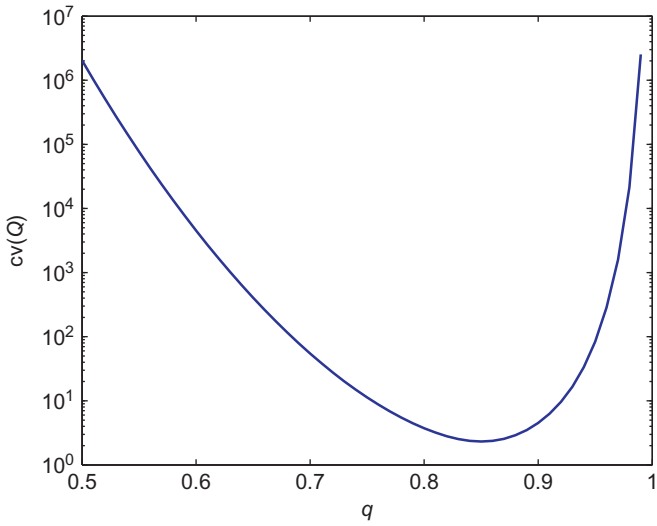

FIGURE 3 The ratio $\text{cv}_* = \text{stdev}_*[Q]/Q$ as a function of q for the 1D random walk.

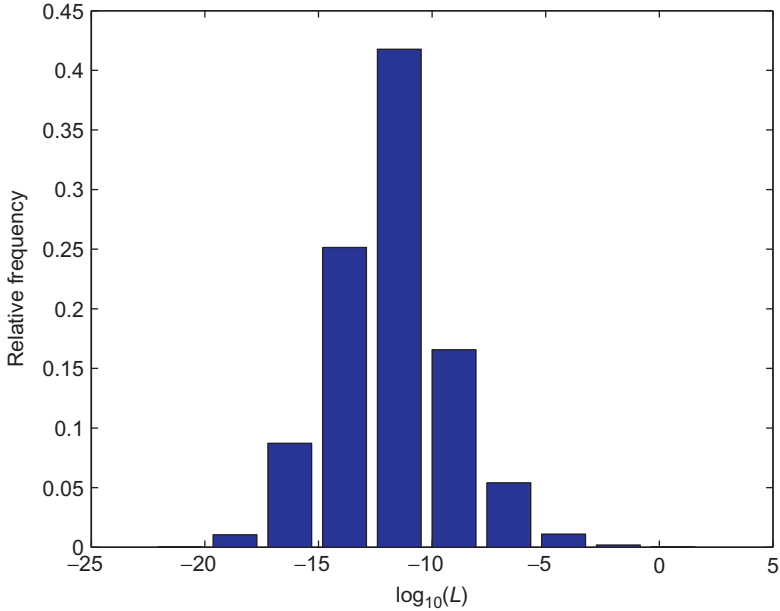

FIGURE 4 Relative frequency of the likelihood ratios in a simulation with $q = 0.85$ and $N = 100,000$.

path that leads to the desired event. The reason why this is so is that the event with the largest value of $p(\mathbf{x})$ among all those for which $I(y(\mathbf{x})) \neq 0$ is the one that provides the dominant contribution to the integral that defines Q. (Note that in many cases of interest, $p(\mathbf{x})$ decays exponentially away from its maximum.) In our example, the most likely way to obtain *at least* 85 heads (the desired event) is to obtain *exactly* 85 heads. So, the optimal biasing choice is to bias the simulations around this value.

3.4 Common Biasing Choices and Their Drawbacks

We next discuss two simple and commonly mentioned approaches to selecting a biasing distribution: variance scaling and mean translation.

With *variance scaling*, one raises the variance of the input RVs in order to increase the chance that some samples will hit the desired event. For example, if the input RVs $\mathbf{X} \in \mathbb{R}^D$ are normal, i.e., $p(\mathbf{x}) = p_\sigma(\mathbf{x})$, with $p_\sigma(\mathbf{x}) = \mathrm{e}^{-\mathbf{x}\cdot\mathbf{x}/2\sigma^2}/(\sqrt{2\pi}\sigma)^D$, one may try to choose $p_*(\mathbf{x}) = p_{\sigma_*}(\mathbf{x})$, with $\sigma_* > \sigma$. In simple situations (such as the 1D case, i.e., $D = 1$), variance scaling can be quite effective. The applicability of the method is rather limited, however, because of its well-known dimensionality problem. Generally speaking, the problem is that, in many situations, the area over which the samples can "spread" grows faster

than that of the region of interest with the number of dimensions increases. Therefore, while it may intuitively seem that increasing the variance would increase the probability of reaching the desired region compared to the unbiased distribution, this probability will in fact decrease. The end result is that, in dimensions larger than one, the best variance is typically the unscaled one—i.e., the unbiased distribution—and all other biasing choices yield worse results than unbiased MC simulations (i.e., the variance of the importance-sampled estimator is larger than that of the standard MC estimator). For this reason, variance scaling has largely been superseded by the mean translation method.

With *mean translation*, one adds a mean to the input RVs in order to increase the chance that some samples will hit the desired event, e.g., with normal RVs, one would choose $p_*(\mathbf{x}) = p_\sigma(\mathbf{x} - \mathbf{m})$, with the vector $\mathbf{m}$ being the mean shift. If $\mathbf{m}$ is chosen correctly, mean translation can be very effective in many situations. This method also has some drawbacks, however. When the dimensionality of the problem is large and/or the indicator function of the desired event has a nontrivial geometry in sample space, the optimal translation point might be impossible to find analytically. In this case, one must resort to hybrid or adaptive methods. Also, problems can arise when the symmetry of the problem leads to degeneracy, e.g., suppose one is interested in the total norm of the sum of the RVs. In this case, there is no single choice of translation point that can lead to the correct result. (In the parlance of large deviations theory, which will be briefly discussed in Section 8, this is an example of a situation in which there are multiple—in this case an infinity of—minimum rate points, and no single dominating point; e.g., see Bucklew, 2004 for a discussion of this issue).

We will return to the problem of selecting a good biasing point in Sections 5 and 8.

4 MULTIPLE IS

In some cases of interest, no single choice of biasing distribution can efficiently capture all the regions of sample space that give rise to the events of interest. In these cases, it is necessary to use IS with more than one biasing distribution. The simultaneous use of different biasing methods (which is similar to the use of a mixture density) is called *multiple importance sampling*.

4.1 Multiple IS: General Formulation

Suppose we want to use J biasing distributions $p_{*1}(\mathbf{x}), \ldots, p_{*J}(\mathbf{x})$, each of which allows us to efficiently reach a given region of sample space. The issue arises of how to correctly weight the results coming from these different distributions. One possible solution to this problem is to assign a weight $w_j(\mathbf{x})$ to each distribution and rewrite Q as:

$$Q = \sum_{j=1}^{J} Q_j = \sum_{j=1}^{J} \int w_j(\mathbf{x}) f(\mathbf{x}) L_j(\mathbf{x}) p_j^*(\mathbf{x}) (\mathrm{d}\mathbf{x}) \,, \tag{12}$$

where $L_j(\mathbf{x}) = p(\mathbf{x})/p_{*j}(\mathbf{x})$ is the likelihood ratio for the jth distribution. Note that the right-hand side of Eq. (12) equals Q for any choice of weights such that $\sum_{j=1}^{J} w_j(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathbb{R}^D$. Each choice of weights corresponds to a different way of partitioning of the total probability.

From (12), a multiply-importance-sampled MC estimator for Q can now be written as

$$\hat{Q} = \sum_{j=1}^{J} \hat{Q}_j = \sum_{j=1}^{J} \frac{1}{N_j} \sum_{n=1}^{N_j} w_j(\mathbf{X}_{j,n}) f(\mathbf{X}_{j,n}) L_j(\mathbf{X}_{j,n}) \,, \tag{13}$$

where N_j is the number of samples drawn from the jth distribution $p_{*j}(\mathbf{x})$, and $\mathbf{X}_{j,n}$ is the nth such sample. Also, one can show that, similar to before, an unbiased estimator of its variance is

$$\hat{\sigma}^2_{\hat{Q}} = \sum_{j=1}^{J} \frac{1}{N_j(N_j - 1)} \sum_{n=1}^{N_j} \left(w_j(\mathbf{X}_{j,n}) L_j(\mathbf{X}_{j,n}) f(\mathbf{X}_{j,n}) - \hat{Q}_j \right)^2 .$$

As before, recursion relations can also be written so that this quantity can be obtained without the need of storing all the individual samples until the end of the simulation:

$$\hat{\sigma}^2_{\hat{Q}} = \sum_{j=1}^{J} \frac{1}{N_j(N_j - 1)} \hat{S}_{j,N_j} \,,$$

with $\hat{Q} = \sum_{j=1}^{J} \hat{Q}_{j,N_j}$ and [in the special case $f(\mathbf{x}) = I(y(\mathbf{x}))$]

$$\hat{Q}_{j,n} = \frac{n-1}{n} \hat{Q}_{j,n-1} + \frac{1}{n} w_j^2(\mathbf{X}_{j,n}) L_j^2(\mathbf{X}_{j,n}) I(y(\mathbf{X}_{j,n})) \,,$$
$$\hat{S}_{j,n} = \hat{S}_{j,n-1} + \frac{n-1}{n} \left(w_j^2(\mathbf{X}_{j,n}) L_j^2(\mathbf{X}_{j,n}) I(y(\mathbf{X}_{j,n})) - \hat{Q}_{j,n-1} \right)^2 .$$

4.2 The Balance Heuristics

Of course, several ways exist to choose the weights $w_j(\mathbf{x})$ when using multiple IS. And the choice of weights is almost as important as the choice of biasing distributions $p_j(\mathbf{x})$. Different weighting functions result in different values for the variance of the combined estimator. A poor choice of weights can result in a large variance, thus partially negating the gains obtained by IS. The best weighting strategies are of course the ones that yield the smallest variance.

The simplest possibility is just to set $w_j(\mathbf{x}) = 1/J$ for all $\mathbf{x}$, meaning that each distribution is assigned an equal weight in all regions of sample space. This choice is not advantageous, however, as we will see shortly. Another simple

choice is that in which the weighting functions are constant over the whole sample space. In this case, one would have

$$Q = \sum_{j=1}^{J} w_j \int I(y(\mathbf{x}))L_j(\mathbf{x})(\mathrm{d}\mathbf{x}) = \sum_{j=1}^{J} w_j \, \mathbb{E}_{*j}[I(y(\mathbf{x}))L_j(\mathbf{x})] \, .$$

The corresponding importance-sampled estimator is then simply a weighted combination of the estimators obtained by using each of the biasing distributions. Unfortunately, the variance of Q is also a weighted sum of the individual variances: $\sigma_j^2 = \sum_{j=1}^{J} w_j \sigma_j^2$, and if any of the sampling techniques is bad in a given region, then Q will also have a high variance. Then, one may be tempted to define the weights according to the actual number of samples from each distribution that fall in a given region. It is important to realize, however, that this is not a good choice, as it does not produce a unbiased estimator (i.e., one whose expectation value is the desired quantity).

On the other hand, there is a relatively simple and particularly useful choice of weights: the *balance heuristics* (Owen and Zhou, 2000; Veach, 1997). In this case, the weights $w_j(\mathbf{x})$ are assigned according to

$$w_j(\mathbf{x}) = \frac{N_j p_{*j}(\mathbf{x})}{\sum_{j'=1}^{J} N_{j'} p_{j'}^{*}(\mathbf{x})} \, . \tag{14}$$

Note that the quantity $N_j p_{*j}(\mathbf{x})$ is proportional to the *expected* number of hits from the jth distribution. Thus, the weight associated with a sample $\mathbf{x}$ with the balance heuristics is given by the relative likelihood of realizing that sample with the jth distribution relative to the total likelihood of realizing that same sample with all distributions. Hence, Eq. (14) weights each $p_{*j}(\mathbf{x})$ most heavily in those regions of sample space where $p_{*j}(\mathbf{x})$ is largest. [Note Eq. (14) can also be written in terms of likelihood ratios, a form which is particularly convenient in Eq. (13).]

The balance heuristics has been shown to be close to optimal in most situations (Veach, 1997). Of course, other strategies are possible, and some of these alternatives do perform better in specific cases (Veach, 1997). It is difficult to tell *a priori* which choice will be best in any given situation, however. Therefore, the balance heuristics is frequently used in practice because of its effectiveness and simplicity.

4.3 Application: Numerical Estimation of Probability Density Functions

In some cases, one is not just interested in one specific probability, but rather would like to numerically estimate the whole pdf of a quantity of interest which is a complicated function of the RVs. As an application of multiple IS, here we briefly discuss the strategy that can be used to set up the numerical simulations.

In our example of a 1D RW, suppose that we want to numerically estimate the pdf of the final position $y(\mathbf{X})$. (Of course, in this case we already did it analytically, but the example will serve to illustrate the procedure.)

The desired result now is more than a single number; instead, we are trying to simultaneously estimate all the integrals

$$p_k = \frac{1}{\Delta y_k} \int_{R_k} p_y(y)\, \mathrm{d}y = \frac{1}{\Delta y_k} \int I_{R_k}(y(\mathbf{x})) p_{\mathbf{x}}(\mathbf{x})\, (\mathrm{d}\mathbf{x})\,, \tag{15}$$

for $k = 1, \ldots, K$, where $y_k = y_o + \Delta y_{k-1}$, with $\Delta y_k = y_{k+1} - y_k$, and $R_k = [y_k, y_{k+1}]$. Note that the integrals in Eq. (15) are of the same type as that in Eq. (3). Thus, we can apply the IS techniques presented earlier. It should be clear, however, that no single biasing distribution can efficiently generate the whole range of possible values of y, and therefore, one needs to resort to multiple IS. The procedure is then to:

1. choose a set of J biasing distributions $p_{*1}(\mathbf{x}), \ldots, p_{*J}(\mathbf{x})$;
2. perform a predetermined number N_j of MC simulations for each distribution, keeping track of the likelihood ratio and the weights for each sample;
3. sort the results of all the MC samples into bins and combine the individual samples using one of the weighting strategies presented earlier.

Note that it is not necessary to fix the number of bins and the precise bin locations in advance of the simulations, and one can choose them *a posteriori* to optimize the results.

Figure 5 shows the results obtained from each of three individual importance-sampled MC simulations of the same 1D RW described earlier, together with the corresponding coefficient of variation. Note that, as is often the case in similar situations, one of the biasing distributions was chosen to be the unbiased one, to make sure that the simulations recover the main portion of the desired pdf. As expected, different values of the biasing parameter target different regions of the pdf. (Negative values of final position can obviously be targeted just as easily by choosing $q < 1/2$.) Note how the cvs for each simulations become large near the edges of the region targeted by each simulation, where the expected number of samples is small.

Figure 6 shows the corresponding pdf obtained when the results from the individual simulations are combined into a single multiply-importance-sampled estimator using the balance heuristics. One can see that indeed the combined results have a low cv throughout the range of values desired.

5 THE CROSS-ENTROPY METHOD

As we have seen earlier, in order for IS methods to be effective, it is crucial to choose a good biasing strategy, as poor biasing strategies can lead to incorrect results and/or performance that is even poorer than that of standard MC. In some

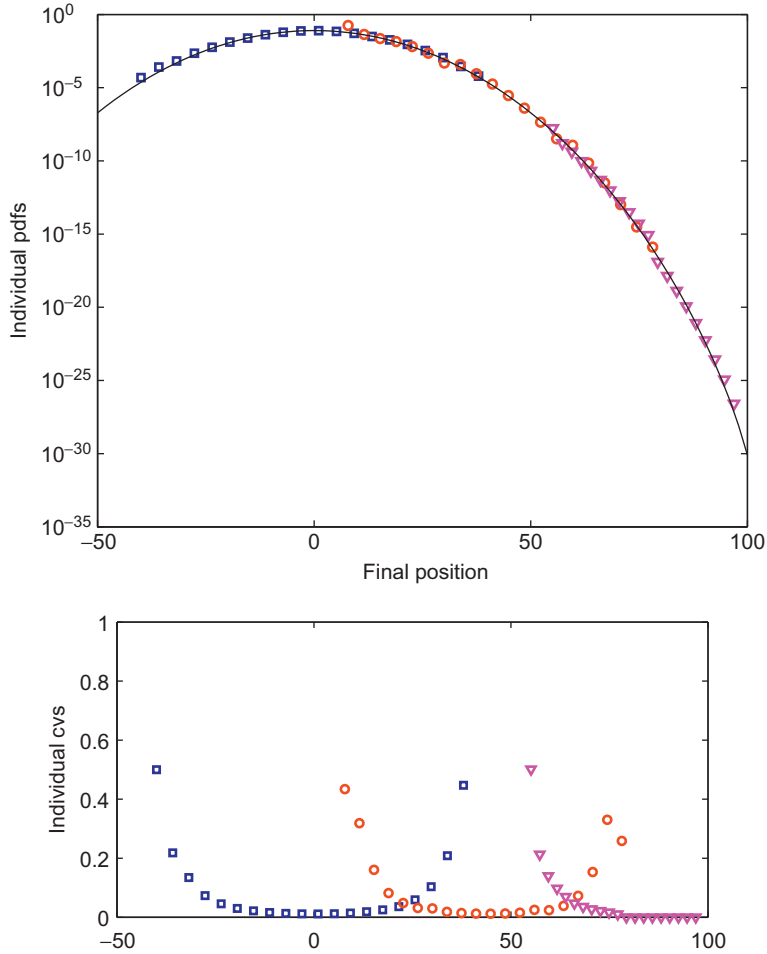

FIGURE 5 The portions of pdf of the 1D random walk as reconstructed from three IS-MC runs with $N = 10,000$ each. Blue (dark gray in the print version): $q = 0.5$; red (gray in the print version): $q = 0.72$; magenta (light gray in the print version): $q = 0.9$. Inset: The cv for each of the simulations.

cases, however, it may be difficult to find such a strategy. A possible alternative in such cases is the use of the *cross-entropy method* (de Boer et al., 2005; Rubinstein and Kroese, 2004).

Recall that the theoretical optimal biasing distribution, $p_{\text{opt}} = I_R(y(\mathbf{x}))\, p(\mathbf{x})/Q$, is not practical, as it requires knowledge of Q in advance. Often, however, one can find a good biasing distribution by requiring it to be "close" to the optimal biasing distribution. This can be accomplished by minimizing the Kullback–Leibler distance (Kullback and Leibler, 1951):

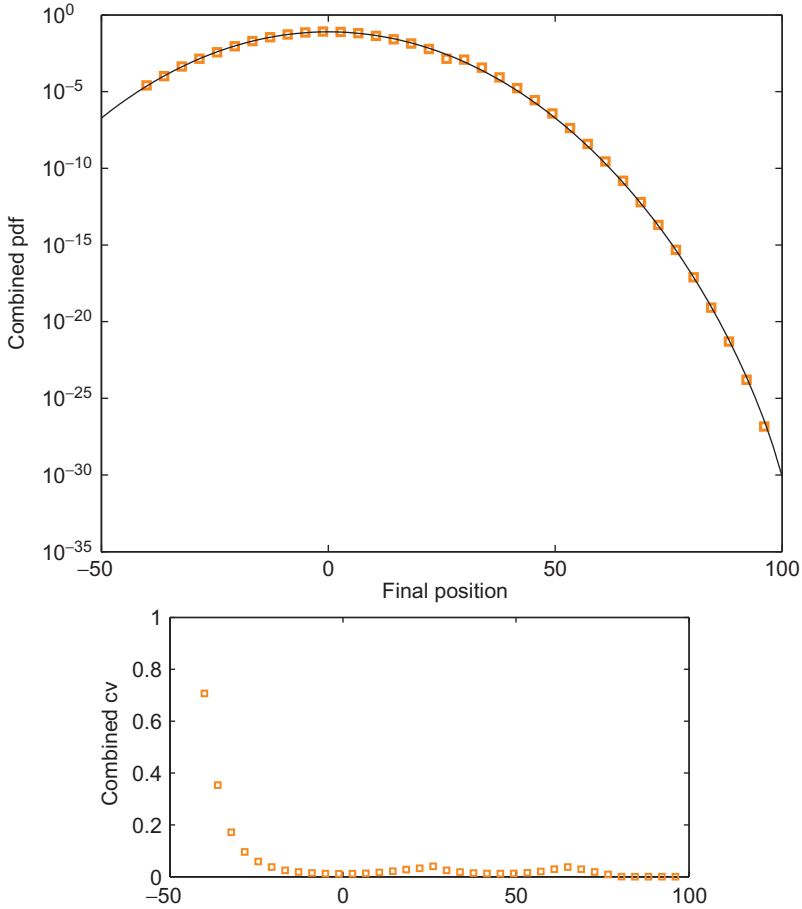

FIGURE 6 The pdf of the 1D random walk as reconstructed by combining the three simulations into a single multiply-importance-sampled run. Inset: The overall coefficient of variation.

$$\begin{aligned}\mathcal{D}(p_{\text{opt}}, p_*) &= \mathbb{E}_{p_{\text{opt}}}\left[\ln \frac{p_{\text{opt}}(\mathbf{x})}{p_*(\mathbf{x})}\right] \\ &= \int \ln(p_{\text{opt}}(\mathbf{x}))\, p_{\text{opt}}(\mathbf{x})(\mathrm{d}\mathbf{x}) - \int \ln(p_*(\mathbf{x}))\, p_{\text{opt}}(\mathbf{x})(\mathrm{d}\mathbf{x})\,, \end{aligned} \tag{16}$$

which is also known as the cross-entropy between two probability distributions. Minimizing $\mathcal{D}(p_{\text{opt}}, p_*)$ is equivalent to maximizing $\int \ln(p_*(\mathbf{x}))p_{\text{opt}}(\mathbf{x})\,(\mathrm{d}\mathbf{x})$. (Note that $\mathcal{D}$ is not a true "distance," as it is not symmetric in its two arguments.) In turn, recalling the expression for p_{opt}, this problem is equivalent to maximizing $\mathbb{E}[I_R(y(\mathbf{x}))\,\ln p_*(\mathbf{x})\,]$.

Suppose that, as is typically the case in practice, the biasing distributions are selected from a family $\{p_*(\mathbf{x};\mathbf{v})\}_{\mathbf{v}\in V}$ parametrized by a vector $\mathbf{v}$, where V is the corresponding parameter space, and suppose $p_*(\mathbf{x};\mathbf{u}) = p(\mathbf{x})$ is the unbiased distribution. Based on the above discussion, one must maximize the integral

$$D(\mathbf{v}) = \int I_R(y(\mathbf{x}))\ln(p_*(\mathbf{x};\mathbf{v}))p(\mathbf{x})\,(\mathrm{d}\mathbf{x})\,. \tag{17}$$

This is usually done numerically. Since the optimal biasing distribution is typically far from the unbiased distribution, however, the region R of interest is generally also far from the region in sample space where the unbiased distribution $p(\mathbf{x})$ is large. Thus, determining the best choice for $\mathbf{v}$ also becomes a rare event simulation.

The solution to this problem is to use a sequence of intermediate regions R_j that reach the desired region iteratively. (For an alternative approach, see Chan and Kroese, 2012.) Let $D_j(\mathbf{v})$ be the integral in Eq. (17) with R replaced by R_j. Starting with the unbiased distribution, one uses MC sampling to minimize the CE distance between the parametrized distribution and the optimal distribution that reaches R_1. This step, which is done by finding the maximum of $D_1(\mathbf{v})$ over this first set of samples, will give a parameter value $\mathbf{w}_2$. One then uses this value to define a biasing distribution and performs an MC simulation with this distribution to minimize the CE distance between the parametrized distribution and the optimal distribution that reaches R_2. Since a biasing distribution is being used, each step of the procedure is an IS simulation of a stochastic optimization. That is, at step j, one must compute

$$\mathbf{w}_{j+1} = \max_{\mathbf{v}\in V} \hat{D}_j(\mathbf{v})\,, \tag{18}$$

where

$$\hat{D}_j(\mathbf{v}) = \frac{1}{M}\sum_{m=1}^{M} I_{R_j}(y(\mathbf{x}^{(m)}))\ \ln(p_*(\mathbf{x}^{(m)};\mathbf{v}))\,L(\mathbf{x}^{(m)})\,, \tag{19}$$

and where $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(M)}$ are i.i.d. samples generated according to $p_*(\mathbf{x};\mathbf{w}_j)$. The optimal biasing distribution can then be adaptively determined by performing the following steps:

1. Set $j = 0$ and the initial parameter $\mathbf{w}_0 = \mathbf{u}$;
2. Generate MC samples according to $p_*(\mathbf{x};\mathbf{w}_j)$;
3. Solve Eq. (18) to find $\mathbf{w}_{j+1}$;
4. If the iteration has converged, stop; otherwise, increase j to $j+1$ and reiterate from step 2.

Once the iteration has converged, one can then perform IS-MC simulations using the biasing distribution $p_*(\mathbf{x};\mathbf{w}_{\text{final}})$.

The regions R_j can be defined in terms of sample quantiles of some quantity of interest (de Boer et al., 2005). A major issue associated with the above

algorithm, however, is how to accomplish step 3. Solving (18) is in general complicated. If $D(\mathbf{v})$ is convex and differentiable, however, the solutions of (18) can be obtained by solving a system of algebraic equations:

$$\frac{1}{M}\sum_{m=1}^{M} I_R(y(\mathbf{x}^{(m)}))\,\nabla_{\mathbf{v}}[\ln p(\mathbf{x}^{(m)};\mathbf{v})]\,L(\mathbf{x}^{(m)};\mathbf{u},\mathbf{w}) = 0\,. \tag{20}$$

In many applications, this equation can be solved analytically. If that is not possible, one can try to find a solution numerically.

The CE method enjoys desirable convergence properties. Specifically, for certain (static) models, under mild regularity conditions the CE method terminates with probability 1 in a finite number of iterations. Moreover, the CE method provides a consistent and asymptotically normal estimator for the optimal reference parameters (see Homem-de-Mello and Rubinstein, 2002). The CE method has been successfully applied to the estimation of rare event probabilities in dynamic models, in particular queueing models involving both light and heavy tail input distributions (de Boer et al., 2004; Kroese and Rubinstein, 2004). Recently, a method that combines IS with the CE method has been developed and used with success to study a specific model of birefringence-induced errors (Marzec et al., 2013; Schuster et al., 2014), and noise-induced perturbations (Donovan and Kath, 2011) of lightwave communication systems. We refer the reader to de Boer et al. (2005) and Rubinstein and Kroese (2004) for further details about the method and its applications.

6 MCMC: REJECTION SAMPLING, THE METROPOLIS METHOD, AND GIBBS SAMPLING

A related simulation problem is that in which the distribution $p_{\mathbf{X}}(\mathbf{x})$ of the RVs $\mathbf{X}$ is not easy to sample from. This might happen for various reasons, e.g., a typical situation is that in which the normalization constant in the distribution is difficult to compute. Another typical situation is that in which the RVs are not independent but are related by complicated nonlinear interdependencies, in which case $p_{\mathbf{X}}(\mathbf{x})$ is a derived density that may be very hard to compute. In these situations, a useful approach could be the use of rejection sampling, the Metropolis–Hastings method (Metropolis et al., 1953), and its variants such as Gibbs sampling. We next give a brief introduction to these methods, referring the reader to Fishman (1996) and MacKay (2003) for further details.

We start with the simplest among these methods: *rejection sampling*. Consider for simplicity a 1D case, namely a single RV X distributed according to $p_X(x)$. Suppose that $p_X(x) = \tilde{p}_X(x)/Z$, where $\tilde{p}_X(x)$ is known but Z is not. The idea behind rejection sampling is to use a proposal density $p_*(x) = \tilde{p}_*(x)/\tilde{Z}$ which is known (possibly up to the normalization constant $\tilde{Z}$) and from which we can easily draw samples. Suppose further that we can also find a constant C such

that $C\tilde{p}_*(x) > \tilde{p}_x(x)$ for all x. A single step of the rejection sampling method proceeds as follows:

(i) Generate a RV, X_*, from the proposal density $\tilde{p}_*(x)$.
(ii) Evaluate $C\tilde{p}_*(X_*)$ and generate a uniformly distributed RV u from the interval $[0, C\tilde{p}_*(X_*)]$.
(iii) Evaluate $p_x(X_*)$ and accept or reject the sample X_* by comparing the value of u with the value of $\tilde{p}_x(X_*)$. More precisely, if $u > \tilde{p}_x(X_*)$, then X_* is rejected; otherwise, it is accepted, in which case X_* is added to our set of samples. (The value of u is discarded no matter what.)

The obvious question is why should this procedure generate samples from $p_x(x)$. To answer this, note first that the pair (X_*, u) identifies a point in the two-dimensional xy plane. Moreover, (X_*, u) is selected with uniform probability from the area underneath the curve $y = C\tilde{p}_*(x)$. The above algorithm rejects all points that lie above the curve $y = \tilde{p}_x(x)$. Thus, points (x, u) that are accepted are uniformly distributed over the area under $y = \tilde{p}_x(x)$. This implies that the probability density of the x-coordinates of points that are accepted must be proportional to $\tilde{p}_x(x)$. In turn, this implies that the accepted samples amount to independent samples drawn from $p_x(x)$.

Rejection sampling can be generalized to several RVs in a straightforward way. In many cases, however, it is difficult to produce a proposal density $\tilde{p}_*(x)$ with the desired properties. In some of these cases, the problem can be obviated by the use of the *Metropolis method*. The main idea of the Metropolis method is to create a Markov chain whose transition matrix does not depend on the normalization term. One needs to make sure that the chain has a stationary distribution and such stationary distribution is equal to the target distribution. After a sufficient number of iterations, the chain will then converge to the stationary distribution.

To make these ideas more precise, recall that a (discrete time) Markov chain is a random process $\mathbf{X}_t \in \mathcal{S}$ (where $\mathcal{S}$ denotes sample space) that satisfies the Markov property: $\mathbb{P}[\mathbf{X}_{t+1} \,|\, \mathbf{X}_t, \ldots, \mathbf{X}_1] = \mathbb{P}[\mathbf{X}_{t+1} \,|\, \mathbf{X}_t]$. That is, the process has no memory: the future state of the system only depends on its present state, not on its past. A finite-state Markov chain (namely, one in which the cardinality of $\mathcal{S}$ is finite, $|\mathcal{S}| < \infty$) can be completely specified by the transition matrix $P = (p_{i,j})$ defined by the elements $p_{i,j} = \mathbb{P}[\mathbf{X}_{t+1} = j \,|\, \mathbf{X}_t = i]$. For irreducible chains, the stationary distribution π is the long-term proportion of time that the chain spends in each state. (Such a distribution can be computed noting that $\pi = \pi P$.) The Metropolis method makes use of a proposal density $p_*(\mathbf{X}; \mathbf{X}_t)$ that depends on the current state $\mathbf{X}_t$. More precisely, a single step of the Metropolis method proceeds as follows:

(i) Select a candidate move $\mathbf{X}_*$ generated from the current state $\mathbf{X}_t$ according to the proposal density $p_*(\mathbf{X}_*; \mathbf{X}_t)$.

(ii) Compute the ratio

$$r = \frac{p_{\mathbf{x}}(\mathbf{X}_*)\, p_*(\mathbf{X}_*; \mathbf{X}_t)}{p_{\mathbf{x}}(\mathbf{X}_t)\, p_*(\mathbf{X}_t; \mathbf{X}_*)}\,. \tag{21}$$

(iii) If $r \geq 1$, accept the move. Otherwise accept the move with probability r. (As in rejection sampling, this can be done by drawing a uniform RV u in $[0, 1]$ and accepting the move if $u < r$.)

(iv) If the move is accepted, set $\mathbf{X}_{t+1} = \mathbf{X}_*$. Otherwise remain in the current state (i.e., set $\mathbf{X}_{t+1} = \mathbf{X}_t$).

The approach is similar to rejection sampling, in that a candidate move is generated and then either accepted or rejected with a given probability. Two important differences, however, are that: (a) unlike rejection sampling, here the candidate move depends on the current state and (b) in rejection sampling, rejected points are discarded and have no influence on the list of samples collected, whereas in the Metropolis method a rejection causes the current state to be inserted again into the list of samples. We also note in passing that the original formulation of the method was done for the special case in which the proposal density is symmetric, i.e., $p_*(\mathbf{y}; \mathbf{x}) = p_*(\mathbf{x}; \mathbf{y})$, in which case (21) reduces simply to $r = p_{\mathbf{x}}(\mathbf{X}_*)/p_{\mathbf{x}}(\mathbf{X}_t)$. The more general version of the method described above should be more accurately called the Metropolis–Hastings method.

Unlike rejection sampling, the Metropolis method does not automatically generate samples from $p_{\mathbf{x}}(\mathbf{x})$. Rather, one can show that, for any positive proposal density $p_*(\mathbf{y}, \mathbf{x})$, the density of $\mathbf{X}_t$ tends asymptotically to $p_{\mathbf{x}}(\mathbf{x})$ in the limit $t \to \infty$. Nothing can be said in general about the rate of convergence, however, i.e., about how rapidly the convergence takes place. It is also important to realize that the samples generated by the Metropolis method are not statistically independent (which makes it difficult to compute variances). Indeed, the Metropolis method is our first example of *MCMC* methods, in which a Markov process is used to generate a sequence of states, each state having a probability distribution that depends on the previous state. Since successive samples are dependent, one may need to run the Markov chain for a considerable time in order to generate samples that are effectively independent. Finally, an important caveat is that the Metropolis method relies on diffusion to explore state space. This can be extremely slow and inefficient.

While rejection sampling and the Metropolis method can be used on 1D problems, *Gibbs sampling* (also known as the heat bath method or "Glauber dynamics") is a method for sampling from distributions in dimensions two or higher. The main idea of Gibbs sampling is to use conditional distributions. Consider for simplicity a two-dimensional example, with $\mathbf{X}_t = (X_{1,t}, X_{2,t})^T$. Suppose one has a situation where, while it is complicated to sample from the joint density $p_{\mathbf{x}}(\mathbf{x})$, it is feasible to draw samples from the two conditional

distributions $p_{x_2}(x_2 \mid x_1)$ and $p_{x_1}(x_1 \mid x_2)$. A single iteration of the Gibbs sampling method then proceeds as follows:

(i) Given the current state $\mathbf{X}^t$, generate a new value for X_1 using the conditional distribution $p_{x_1}(x_1 \mid X_{2,t})$.
(ii) Use the new X_1 to generate a new value for X_2 using the conditional distribution $p_{x_2}(x_2 \mid X_1)$ and set $\mathbf{X}_{t+1} = (X_1, X_2)^T$.

One can show that a single iteration of Gibbs sampling can be viewed as a Metropolis method with target density $p_{\mathbf{x}}(\mathbf{x})$, and that this particular implementation has the property that every candidate move is always accepted. Thus, as long as the joint density $p_{\mathbf{x}}(\mathbf{x})$ is reasonably nice, the probability distribution of the samples generated will tend to $p_{\mathbf{x}}(\mathbf{x})$ as $t \to \infty$.

Since Gibbs sampling is a special case of a Metropolis algorithm, it suffers from the same problems. Namely, samples are not independent, and generically speaking state space is explored by a slow RW. On the other hand, Gibbs sampling does not involve any adjustable parameters, and therefore, it is an attractive strategy when one wants to quickly test a new model. Also, various software packages are available that make it easy to set up and simulate a large class of probabilistic models by Gibbs sampling (Thomas et al., 1992).

7 APPLICATIONS OF VRTs TO ERROR ESTIMATION IN OPTICAL FIBER COMMUNICATION SYSTEMS

One of the areas in which IS and other VRTs have recently been applied with considerable success in recent years is the estimation of error probabilities in optical fiber communication systems (Agrawal, 2002; Kaminov and Koch, 1997). As an illustration of the methods discussed in this chapter, we devote this section to a brief review of the problem and of how the techniques that were presented in the previous sections were used in this context.

Errors in optical fiber communication systems are required to be extremely rare, e.g., the bit error ratio [that is, the probability of a transmission error] is required to be 10^{-12} or smaller after error correction. This stringent requirement imposes severe constraints on the design of these systems and creates a need for accurate design tools. On one hand, however, experiments are very expensive (the cost of setting up a fully equipped lab can exceed millions of dollars), and optimizing the system's performance involves selecting precise values for many independent parameters (such as input powers, pulse format, fiber types, and relative section lengths). Therefore, design engineers are in need of accurate mathematical and computational modeling. On the other hand, systems are large and complex, with many physical effects contributing to determine the overall system performance. Moreover, error probabilities are non-Gaussian due to nonlinear interactions. Hence, mathematical methods are alone not sufficient. But precisely because errors are required to be so rare, error probabilities cannot be estimated by standard MC simulations. An approach which has proved to be

successful in this situation is a hybrid one, in which the analytical knowledge of the dominant sources of error is used to design appropriate biasing strategies for IS.

There are two main sources of randomness that contribute to determine the overall system performance in optical fiber transmission systems: fiber disorder, manifesting itself in random birefringence, and amplified spontaneous emission noise from the optical amplifiers that are used to compensate for the attenuation of the signal due to fiber loss (Agrawal, 2002; Kaminov and Koch, 1997). We next briefly describe each of these two problems and the techniques that were brought to bear to study each of them. A further source of variability is the pseudo-randomness of the data stream, which can result in transmission errors through system nonlinearity. For brevity, however, we omit any discussion of this issue, and we refer the reader to Ablowitz et al. (1998), Mecozzi (1998), Sinkin et al. (2007), and references therein for details.

7.1 Polarization-Mode Dispersion

Birefringence arises when the speed of propagation of light in a medium depends on the polarization of the light itself. Although a great deal of effort is devoted to controlling all aspects of the manufacturing of optical fibers, a certain amount of fiber birefringence is always present. The presence of birefringence has the effect that an optical pulse will split into two components, propagating along what are called the fast and slow axes of birefringence. Moreover, the fiber's birefringence (including its strength and the birefringence axes) varies with wavelength, temperature, and time. The random, birefringence-induced perturbations on optical pulses are referred to as *polarization-mode dispersion* (PMD) (Kogelnik et al., 2002).

In most installed systems, PMD-induced impairments are completely determined by the real, three-component first- and second-order PMD vector, denoted, respectively, as $\vec{\tau}$ and $\vec{\tau}_\omega = \mathrm{d}\vec{\tau}/\mathrm{d}\omega$ (where ω is the optical frequency) (Kogelnik et al., 2002). In turn, the growth of PMD with distance is governed by the first- and second-order PMD concatenation equations (Gordon and Kogelnik, 2000), which describe how the first- and second-order PMD vectors of adjoined fiber sections combine with each other to produce the overall behavior of the system. In many cases, after performing an appropriate distance-dependent rotation of the reference frame can be written in the following simplified form (Biondini et al., 2004):

$$\vec{\tau}^{(n+1)} = \vec{\tau}^{(n)} + \Delta\vec{\tau}^{(n+1)}, \quad \vec{\tau}_\omega^{(n+1)} = \vec{\tau}_\omega^{(n)} + \Delta\vec{\tau}^{(n+1)} \times \vec{\tau}^{(n)}. \tag{22}$$

Here $\vec{\tau}^{(n)}$ and $\vec{\tau}_\omega^{(n)}$ are, respectively, the total first- and second-order PMD vectors after the nth fiber section, $\Delta\vec{\tau}^{(n)}$ is the first-order PMD vector of the nth fiber section, and $\Delta\vec{\tau}_\omega^{(n)}$ is the corresponding second-order PMD vector. The rescaled PMD vector $\Delta\vec{\tau}^{(n)}$ of each section can be assumed to be uniformly distributed on the Poincaré sphere; its magnitude $|\tau_n|$ obeys a Maxwellian

distribution with respect to wavelength. Also, for linearly birefringent sections, $\Delta\vec{\tau}_{\omega}^{(n)} = 0$.

The goal of system designers is to estimate the effects of PMD by performing numerical simulations, and in particular to quantify PMD-induced error probabilities. As mentioned before, however, the problem is events in which PMD takes on much larger than average values (resulting in transmission errors) are exceedingly rare. Thus, one would like to have a method to produce large first- and second-order PMD events more frequently than they would occur in practice, and weigh them with correct statistics. We next describe how this can be accomplished using IS. For simplicity, we describe the simplest case in which all fiber sections contribute the same amount of PMD to the total. We emphasize, however, that several other models of PMD generation have been considered in the literature, and a variety of IS and other VRTs have been used with success in all of these cases (Biondini and Kath, 2004, 2005; Biondini et al., 2004; Li et al., 2008, 2010; Lu and Yevick, 2005; Schuster et al., 2014; Secondini and Forestieri, 2005; Yevick, 2002).

It was shown in Biondini et al. (2004) that, when $|\Delta\vec{\tau}^{(n)}|$ is independent of n in (22), the appropriate variables to control in order to monitor the growth of PMD are the relative orientations of the individual sections, $\Delta\vec{\tau}^{(n)}$. To apply IS, one first needs to find the deterministic choices of $\Delta\vec{\tau}^{(n)}$ that maximize the desired combination of first- and second-order PMD. We will label these vectors $\vec{b}^{(n)}$. Once these vectors have been found, one can implement IS by biasing the random MC samples around them. To follow this idea, it is convenient to express the vectors $\vec{b}^{(n)}$ relative to a orthonormal frame of reference formed by the unit vectors $\{\hat{u}_1^{(n)}, \hat{u}_2^{(n)}, \hat{u}_3^{(n)}\}$, where

$$\hat{u}_1^{(n)} = \boldsymbol{\tau}^{(n)}/|\boldsymbol{\tau}^{(n)}|\,, \quad \hat{u}_2^{(n)} = \boldsymbol{\tau}_{\omega,\perp}^{(n)}/|\boldsymbol{\tau}_{\omega,\perp}^{(n)}|\,, \quad \hat{u}_3^{(n)} = \hat{u}_1^{(n)} \times \hat{u}_2^{(n)}\,. \tag{23}$$

Here $\boldsymbol{\tau}_{\omega,\perp}^{(n)}$ is the component of $\boldsymbol{\tau}_{\omega}^{(n)}$ perpendicular to $\boldsymbol{\tau}^{(n)}$. The first of Eq. (22) thus describes a simple 3D RW. Thus, if one only wants to maximize the length of the total first-order PMD vector τ, the best option is to choose $\vec{b}^{(n+1)}$ to be parallel to $\vec{\tau}^{(n)}$ (i.e., to align $\vec{b}^{(n+1)}$ along $\hat{u}_1^{(n)}$). On the other hand, the second of Eq. (22) couples the growth of second-order PMD to that of first-order PMD. Thus, if a nontrivial amount of second-order PMD is desired, one must also take into account the growth of first-order PMD.

When the number of sections is not too small (i.e., larger than 4 or 5), it was found convenient to employ a continuum approximation to find the deterministic biasing directions. Specifically, let $\lim_{\Delta z\to 0} \Delta\vec{\tau}_{n+1}/\Delta z = \vec{b}(z)$. The magnitude of $\vec{b}(z)$ describes the rate at which PMD is added by the birefringent sections. In this limit, one obtains

$$\frac{\mathrm{d}\vec{\tau}}{\mathrm{d}z} = \vec{b}\,, \quad \frac{\mathrm{d}\vec{\tau}_{\omega}}{\mathrm{d}z} = \vec{b} \times \vec{\tau}\,, \tag{24}$$

where z is the longitudinal direction along the fiber. Or, in the frame of reference $\{\hat{u}_1, \hat{u}_2, \hat{u}_3\}$ defined as above,

$$\frac{\mathrm{d}\tau}{\mathrm{d}z} = b_1\,, \quad \frac{\mathrm{d}\tau_{\omega,\parallel}}{\mathrm{d}z} = b_2\,\frac{\tau_{\omega,\perp}}{\tau}\,, \quad \frac{\mathrm{d}\tau_{\omega,\perp}}{\mathrm{d}z} = b_3\tau - b_2\,\frac{\tau_{\omega,\parallel}}{\tau}\,, \tag{25}$$

where (b_1, b_2, b_3) are now the components of $\vec{b}$ with respect to $\{\hat{u}_1, \hat{u}_2, \hat{u}_3\}$. The goal is now to find the function $\vec{b}(z)$ that maximizes second-order PMD or a linear combination of first- and second-order PMD. Fortunately, Eqs. (25) can be solved exactly for any $\vec{b}(z)$:

$$\tau(z) = \int_0^z b_1(\zeta)\,\mathrm{d}\zeta\,, \quad \tau_{\omega,\parallel}(z) = \int_0^z b_3(\zeta)\tau(\zeta)\sin[\beta(z,\zeta)]\,\mathrm{d}\zeta\,,$$

$$\tau_{\omega,\perp}(z) = \int_0^z b_3(\zeta)\tau(\zeta)\cos[\beta(z,\zeta)]\,\mathrm{d}\zeta\,, \tag{26a}$$

$$\beta(z,\zeta) = \int_\zeta^z \frac{b_2(\xi)}{\tau(\xi)}\,\mathrm{d}\xi\,. \tag{26b}$$

The choice of $\vec{b}(z)$ that maximizes the magnitude of second-order PMD (or any combination of first- and second-order PMD) can now be found using calculus of variations. (Detailed calculations can be found in Biondini et al., 2004.) The result is that the maximum growth of second-order PMD is obtained for "in-plane" contributions, namely, $(b_1, b_2, b_3) = b\,(\cos\alpha(z), 0, \sin\alpha(z))$ where the $\alpha(z)$ gradually interpolates between an initial value of 0 (producing pure first-order PMD at first) and a final value of $\pi/2$ (producing pure second-order PMD at the end). In particular, in the case of equal-length sections (namely, for $|\vec{b}(z)| = b$), the angle $\alpha(z)$ has a linearly varying profile: that is, $\alpha(z) = \alpha_{\max}\, z/z_{\max}$, with $\alpha_{\max} = \pi/2$. (The case of nonequal-length sections can be easily obtained from this one by rescaling the independent variable z; see Biondini et al., 2004 for details.) Performing IS-MC simulations with multiple biasing strengths, this biasing choice generates region 3 in Fig. 7.

In many practical situations, however, a more complete coverage of the $|\vec{\tau}||\vec{\tau}_\omega|$ plane is needed. In this case, intermediate biasing choices must also be used in addition to pure first- and second-order biasing. Such choices can be obtained by using calculus of variations to maximize a linear combination of $|\vec{\tau}|$ and $|\vec{\tau}_\omega|$, as obtained from Eqs. (26a). The resulting form of $\vec{b}(z)$ is the same as above, except that the value of the final angle $\alpha_{\max}$ now varies between 0 and π, the particular value depending upon the specific linear combination of first- and second-order PMD being maximized. A selection of angles, together with the resulting regions in the $|\vec{\tau}||\vec{\tau}_\omega|$ plane, is shown in Fig. 7. (Region 1 is the result in the case of biasing for pure first-order PMD.) The advantage of using multiple biasing—as opposed to just pure first- or second-order biasing or no biasing at all—is evident. Each value of $\alpha_{\max}$ generates samples lying in a region that emanates in a roughly radial fashion from the location where the joint pdf is maximum. Together, a set of angles $\alpha_{\max}$ can be used to cover the entire $|\vec{\tau}||\boldsymbol{\tau}_\omega|$ plane. Indeed, Fig. 8 shows the joint pdf of the magnitude of first- and second-order PMD (which is a two-dimensional reduction of the

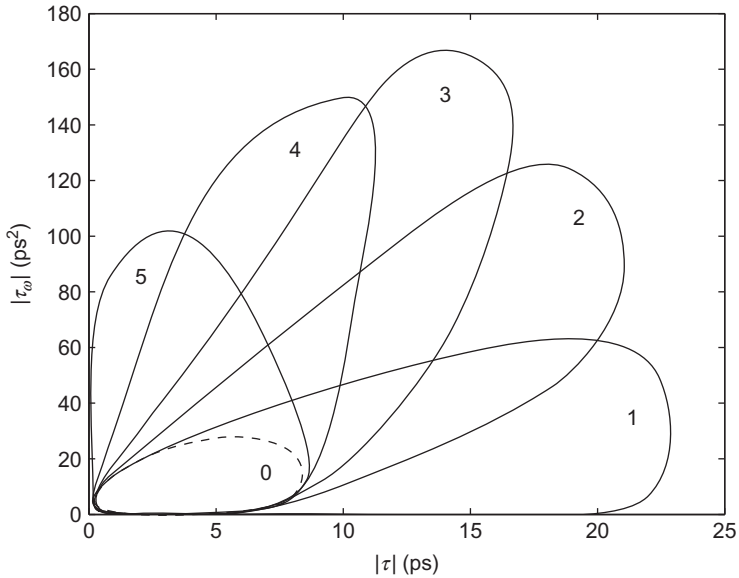

FIGURE 7 The regions of the $|\boldsymbol{\tau}||\boldsymbol{\tau}_\omega|$ plane targeted by the various biasing methods. Region 1 corresponds to pure first-order biasing ($\alpha_{\max} = 0$), region 2 to pure second-order biasing ($\alpha_{\max} = \pi/2$), and regions 3, 4, and 5 to $\alpha_{\max} = \pi/4$, $3\pi/4$, and π, respectively. The dashed line shows the much smaller region obtained with unbiased samples. Fifty birefringent sections with 0.5 ps DGD each were used. *Source: From Biondini et al. (2004).*

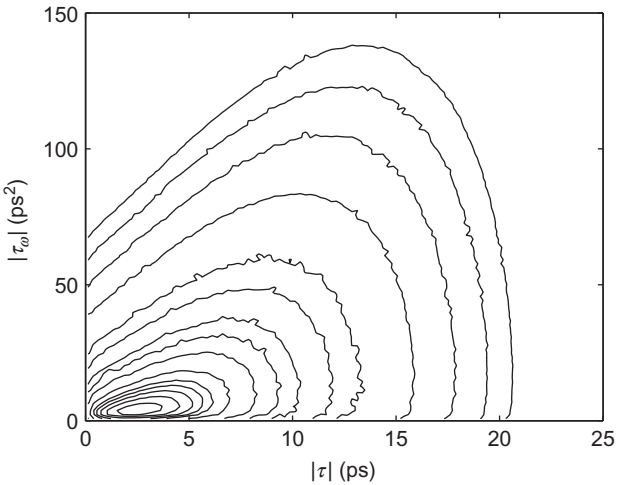

FIGURE 8 Contour plots of the joint pdf of first- and second-order PMD for a concatenation of 50 birefringent sections with 0.5 ps DGD each, as reconstructed from IS-MC simulations. The contours are at 10^{-n} with $n = 1.5, 1.75, 2, 2.25, 3, 4, 5, 6, 8, 10, 15, 20, 25$, and 30. A total of 10^6 Monte-Carlo samples were used. *Source: From Biondini et al. (2004).*

full 3D joint pdf of first- and second-order PMD; Foschini and Poole, 1991) for a system of 50 polarization scramblers, as calculated with the multiple biasing technique described above. In a similar fashion, one can use the same biasing strategies in numerical simulations of pulse transmissions to quantify PMD-induced transmission errors.

7.2 Noise-Induced Perturbations

Together with the invention of the laser in 1960, the birth of optical fiber transmission systems was made possible by the development of low-loss optical fibers, with typical loss coefficients of 0.2 dB/km. Nonetheless, for long-distance communication systems, which span thousands of kilometers, fiber loss remains a serious obstacle, which is compensated by inserting optical fiber amplifiers at various points along the transmission line. Modern optical amplifiers allow the signal to be boosted in the optical domain, avoiding the need for electronic conversion. The downside of this process, however, is the introduction of spontaneous emission photons, which get combined to the signal in the form of additive white Gaussian noise. In addition, since the fiber is weakly nonlinear, the noise interacts with the signal to generate random pulse fluctuations. While these perturbations are not too large on average, they are one of the main sources of errors.

The propagation of optical pulses in fibers is governed by a perturbed nonlinear Schrödinger (NLS) equation with varying coefficients (Agrawal, 2007):

$$\mathrm{i}\frac{\partial q}{\partial z} + \frac{1}{2}d(z)\frac{\partial^2 q}{\partial t^2} + g(z)|q|^2 q = \mathrm{i}S(t,z)\,. \tag{27}$$

Here z is the dimensionless propagation distance, t is the dimensionless retarded time, $q(t,z)$ is the dimensionless slowly varying electric field envelope (rescaled to account for loss and amplification in communication systems), $d(z)$ is the local value of the dispersion coefficient, and $g(z)$ describes the periodic power variations, which are due to loss and amplification. The source term $S(t,z)$ can represent various kinds of perturbations. Here, we focus on the physically interesting case of spontaneous emission noise originating from the optical amplifiers. That is, we consider

$$S(t,z) = \sum_{n=1}^{N_a} v_n(t)\delta(z - nz_a),$$

where N_a is the number of amplifiers, z_a is the dispersion map period, $\delta(z)$ is the Dirac delta distribution, and $v_n(t)$ is white Gaussian noise, satisfying $\mathbb{E}[v_n(t)] = 0$ and $\mathbb{E}[v_n(t)v_{n'}^*(t')] = \sigma^2\delta(t-t')\delta_{nn'}$. In other words, at each amplifier, $z = nz_a$, Eq. (27) is replaced by the jump condition $q(t, nz_a^+) = q(t, nz_a^-) + \sigma v_n(t)$. We note in passing that the numerical simulation of (27) involves a very large (several tens of thousands in practical situations) number of RVs comprised by

the real and imaginary parts of $S(z,t)$ at each of the collocation points in time for each amplifier over the whole transmission line.

In the simplest case of constant dispersion and no gain/loss power variations, without loss of generality one can take $d(z) = g(z) = 1$. In this case, when $S(z,t) = 0$, Eq. (27) is a completely integrable model that admits an infinite number of exact solutions describing elastic interactions among N particle-like objects called solitons (Ablowitz and Segur, 1981; Zabusky and Kruskal, 1965). The simplest case is that of a 1-soliton solution, which is simply the traveling wave solution

$$q(t,z) = A\,\mathrm{sech}[A(t-T)]\,\mathrm{e}^{\mathrm{i}\theta(t,z)}\,, \tag{28}$$

where $\theta(t,z) = V(t-T) + \Phi$ and with $T(z) = Vz + t_o$ and $\Phi(z) = \frac{1}{2}(A^2 + V^2)z + \phi_o$. Note that the 1-soliton solution (28) contains four constant parameters: the amplitude A (which is also its inverse width), the frequency V (which is also the group velocity offset), a temporal offset t_o, and a phase offset ϕ_o.

The case when $d(z)$ and $g(z)$ are not constant but periodic describes a periodic concatenation of fibers with different dispersion properties and is referred to as dispersion management (DM) in the literature. Equation (28) is replaced by a more complicated pulse shape, and the resulting pulses are called dispersion-managed solitons (DMS). Nonetheless, the invariances of the equation imply that DMS still contain the same four pulse parameters. In this case, one can use suitable perturbation methods to derive an equation, called dispersion-managed nonlinear Schrödinger (DMNLS) equation which captures all the essential features of the dynamics as well as the DMS pulse shape (Ablowitz and Biondini, 1998; Spiller and Biondini, 2010).

When noise is present [i.e., $S(t,z) \neq 0$], the nonlinear term in Eq. (27) causes part of the noise to couple to the soliton and induce random deviations of the soliton parameters. One can use perturbation theory on either the NLS or the DMNLS equation to capture the effects of noise on the soliton parameters, obtaining (Li et al., 2007)

$$\frac{\mathrm{d}A}{\mathrm{d}z} = S_A(z)\,, \quad \frac{\mathrm{d}V}{\mathrm{d}z} = S_V(z)\,, \quad \frac{\mathrm{d}T}{\mathrm{d}z} = V + S_T(z)\,,$$

$$\frac{\mathrm{d}\Phi}{\mathrm{d}z} = \frac{1}{2}(A^2 + V^2) + V\,S_T(z) + S_\Phi(z)\,, \tag{29a}$$

where the source terms,

$$S_j(z) = \langle \mathrm{e}^{\mathrm{i}\theta}\bar{y}_j, S\rangle \big/ \langle \bar{y}_j, y_j\rangle\,, \quad j = A, V, T, \Phi, \tag{29b}$$

which are defined in terms of the inner product $\langle f, g\rangle = \mathrm{Re}\int f^*(t)g(t)\,\mathrm{d}t$, are the projection of the noise along the neutral modes y_j of the linearized NLS operator around the soliton solution. Each neutral mode is associated with one of the invariances of the NLS equation as well as with infinitesimal changes in

one of the soliton parameters. Note that since the linearized NLS operator is not self-adjoint, the modes are not mutually orthogonal, and therefore, the projection must be done using the corresponding adjoint modes $\bar{y}_j$. On the other hand, the neutral modes and their adjoints form a biorthogonal basis for the null space of the linearized NLS operator: $\langle \bar{y}_j, y_k \rangle = \langle \bar{y}_j, y_j \rangle \, \delta_{jk}$, where δ_{jk} is the Kronecker delta.

Equations (29a) are a system of nonlinear stochastic differential equations, which cannot be solved in closed form. (The nonlinearity arises not only from the explicit appearance of A and V in the equations but also, and in a more essential way, on the fact that the source terms depend on the soliton amplitude A.) Useful information can still be extracted from them, however. For the present discussion, it is convenient to employ a continuum approximation of the noise. That is, we consider $S(t, z)$ to be a Gaussian white noise process with zero mean and autocorrelation function $\mathbb{E}[S(t, z)S^*(\tau, \zeta)] = \sigma^2 \, \delta(t - \tau)\delta(z - \zeta)$. As a result, the source terms in Eqs. (29a) become independent white noise processes, with autocorrelation function

$$\mathbb{E}[S_j(z)S_k^*(\zeta)] = \sigma_j^2 \delta_{jk} \delta(z - \zeta) \,, \tag{30}$$

where the source term variances are

$$\sigma_j^2 = \mathrm{var}[S_j(z)] = \mathbb{E}\big[\langle \mathrm{e}^{\mathrm{i}\theta} \bar{y}_j, S \rangle^2 \big/ \langle \bar{y}_j, y_j \rangle^2\big] = \sigma^2 \|\underline{y}_j\|^2 \big/ \langle \bar{y}_j, y_j \rangle^2 \,. \tag{31}$$

In the limit of moderate amplitude deviations, one can approximate Eqs. (29a) by considering the variances of the source terms to be constant. The resulting equations can then be integrated exactly, to obtain

$$A(z) = A_o + W_A(z) \,, \quad V(z) = V_o + W_V(z) \,,$$

$$T(z) = T_o + \int_0^z V(\zeta) \mathrm{d}\zeta + W_T(z) \,, \tag{32a}$$

where for brevity we omitted the expression for $\Phi(z)$, and where

$$W_j(z) = \int_0^z S_j(\zeta) \, \mathrm{d}\zeta \,, \quad j = A, V, T, \Phi, \tag{32b}$$

is a Wiener process with zero mean and autocorrelation function $\mathbb{E}[W_j(z)W_k(\zeta)]$ $= \sigma_j^2 \delta_{jk} \min(z, \zeta)$. The mean values of the soliton parameters at the output $z = L$ are then

$$\mathbb{E}[A(L)] = A_o \,, \quad \mathbb{E}[V(L)] = V_o \,, \quad \mathbb{E}[T(L)] = T_o + V_o L,$$

$$\mathbb{E}[\Phi(L)] = \frac{1}{2}(A_o^2 + V_o^2) \, L + \frac{1}{4}(\sigma_A^2 + \sigma_V^2) \, L^2 \,. \tag{33}$$

Tedious but straightforward stochastic calculus (Papoulis, 1991) also yields the variances of the noise-perturbed output soliton parameters (Spiller and Biondini, 2010):

$$\text{var}[A(L)] = \sigma_A^2 L\,, \quad \text{var}[V(L)] = \sigma_V^2 L\,,$$

$$\text{var}[T(L)] = \sigma_T^2 L + \frac{1}{3}\sigma_V^2 L^3\,, \tag{34}$$

where the expression for $\text{var}[\Phi(L)]$ was again omitted for brevity. (Note how the mean phase is directly affected by the noise, unlike the other soliton parameters.) The cubic dependence of timing and phase jitter on distance (which arise, respectively, as a result of the coupling between carrier frequency and group velocity and as a result of the Kerr effect and Galilean invariance) are well-known in the optics literature and are referred to as Gordon–Haus jitter (Gordon and Haus, 1986) and Gordon–Mollenauer jitter (Gordon and Mollenauer, 1990), respectively.

The above mean variances agree very well with direct numerical simulations of the full NLS equation perturbed by noise. However, their knowledge is not sufficient to accurately estimate noise-induced transmission penalties, for several reasons. First of all, the variances are only correct for small deviations of the pulse amplitude, whereas we are interested in quantifying the probability of large deviations. Second, even though the noise is Gaussian-distributed, the noise-induced changes of the soliton parameters are not necessarily Gaussian. In particular, the variance of each amplitude shift depends on the previous value of the amplitude, which causes the distribution of A to deviate significantly from Gaussian. A Gaussian approximation will therefore only be valid in the limit of small amplitude shifts, and even then only in the core region of the pdf and not in the tails. Finally, even if the noise-induced changes of the soliton parameters were approximately Gaussian-distributed, calculating the probability densities in the tails from the (analytically or numerically obtained) variances would require an exponential extrapolation, and any errors or uncertainties would be magnified correspondingly.

Nonetheless, the information obtained from the above perturbation theory is the key to devise a successful IS for the problem, as we show next. In our case, to successfully apply IS one must find the most likely noise realizations that lead to a desired change of the soliton parameters at the output. As demonstrated in Moore et al. (2003) and Li et al. (2007), one can approach this problem by decomposing it into two logically distinct steps: (i) finding the most likely noise realizations that produce a given parameter change at each amplifier and (ii) finding the most likely way in which individual parameter changes at each amplifier combine to produce a total change at the output. This two-step approach is justified by the fact that the noise at different amplifiers is statically independent. We next briefly describe each of these two steps.

(i) Biasing at a single amplifier. Consider a generic perturbation to the solution at the nth amplifier, $b_n(t)$. Recall from Eqs. (29) that the noise-induced change to a soliton parameter Q (with $Q = A, V, T, \Phi$) is found by taking the projection of the perturbation onto the adjoint mode of the linear DMNLS operator associated

with Q. That is, if $q(t, nz_a^+) = q(t, nz_a^-) + b_n(t)$, the change to parameter Q due to the perturbation $b_n(t)$ is given by

$$\Delta Q_n = \mathrm{Re} \int \bar{y}_Q^* b_n(t) \mathrm{d}t \Big/ \int |\bar{y}_Q|^2 \mathrm{d}t \,. \tag{35a}$$

The problem of finding the optimal biasing at each amplifier is to find the most likely noise realization subject to the constraint of achieving, on average, a desired parameter change at that amplifier. In other words: given a specific parameter change ΔQ_n at the nth amplifier (with $Q = A, V, T, \Phi$), what is the form of $b_n(t)$ that is most likely to produce this prescribed change? For white Gaussian noise, maximizing its probability amounts to minimizing the negative of the log-likelihood, i.e., the negative of the argument of the exponential in the noise pdf. That is, we need to minimize the L_2 norm of the noise,

$$\|b_n(t)\|^2 = \int |b_n(t)|^2 \mathrm{d}t \,, \tag{35b}$$

subject to achieving the desired parameter change ΔQ_n given by Eq. (35a). One can formulate this as a variational problem, whose solution yields the deterministic biasing direction (Moore et al., 2008)

$$b_n(t) = \Delta Q_n \big(\mathrm{Re} \int \bar{y}_Q^* y_Q \mathrm{d}t \Big/ \int |\bar{y}_Q|^2 \mathrm{d}t \big) \, \bar{y}_Q \,. \tag{36}$$

(ii) Biasing across all amplifiers. Next we address the question of how one should distribute the bias for the soliton parameters among all amplifiers in order to achieve a specified parameter change at the output. In other words: what is the most likely set of individual parameter changes $\{\Delta A_n, \Delta V_n, \Delta T_n, \Delta \Phi_n\}_{n=1,\ldots,N_a}$ that realizes a given value of ΔQ_{target} (with Q equal to either A, V, T, or Φ, as before) at the output? For simplicity, we limit our discussion to amplitude deviations, even though the same approach can be used to study variations of all four soliton parameters (Spiller and Biondini, 2010).

We begin by examining the amplitude evolution from one amplifier to the next, namely

$$A_{n+1} = A_n + \Delta A_{n+1} \,. \tag{37}$$

Recall that the most likely noise realization that achieves a given amplitude change at a single amplifier is given by (36), with $Q = A$ in this case. Also recall that the norms and inner products of the linear modes depend on the soliton amplitude and therefore also indirectly on distance. It should be clear that maximizing the probability of obtaining a given amplitude at the output is equivalent to minimizing the sum of the L_2 norm of the biasing functions $b_n(t)$ over all amplifiers. That is, we need to minimize the sum

$$\sum_{n=1}^{N_a} \|b_n\|^2 = \sum_{n=1}^{N_a} |\Delta A_n|^2 / \sigma_A^2 \,, \tag{38a}$$

subject to the constraint

$$\sum_{n=1}^{N_a} \Delta A_n = A_{\text{target}} - A_o \,, \tag{38b}$$

where σ_A^2 is given by Eq. (31). To solve this problem, we consider a continuum approximation. That is, we replace Eq. (37) by the first of Eqs. (29a), with $S(t,z) = b(t,z)$ and $b(t,z)$ given by the continuum analogue of Eq. (36) with $Q = A$, that is: $b(t,z) = (\langle y_A, \bar{y}_A \rangle / \|\bar{y}_A\|^2)\, \bar{y}_A(t)\, \dot{A}$. We then seek a function $A(z)$ that minimizes the continuum limit of Eq. (38a). That is, we seek to minimize the integral from $z = 0$ to $z = L$ of the L_2 norm of $b(t,z)$, namely, the functional

$$J[A] = \int_0^L \dot{A}^2 / \sigma_A^2 \, \mathrm{d}z \,, \tag{39}$$

subject to the fixed boundary conditions $A(0) = A_o$ and $A(L) = A_{\text{target}}$ [which are the continuum limit of (38b)]. Hereafter, the dot denotes total differentiation with respect to z, and L is the total transmission distance as before. After some straightforward algebra, the Euler–Lagrange equation associated with the functional $J[A]$ in (39) can be written as

$$2\ddot{A}\, \frac{1}{\sigma_A^2} + \dot{A}^2 \frac{\partial}{\partial A}\left(\frac{1}{\sigma_A^2}\right) = 0 \,,$$

which is readily integrated to give

$$\dot{A} = c\ \sigma_A \,, \tag{40}$$

where c is an integration constant which determines the total amount of biasing being applied and thereby the value of the amplitude at the output. One can now integrate Eq. (40) to find the optimal path $A(z)$ that realizes a desired amplitude change at the output. Once this path has been obtained, one can then calculate ΔA_n, which was the only unknown in the optimal biasing directions b_n in Eq. (36).

Equation (40) can be solved exactly in the case of constant $d(z)$ and $g(z)$ (that is, for the classical NLS equation). In this case, Eq. (40) reduces to $\dot{A} = c\sqrt{A}$, which is trivially integrated to $A_{\text{nls}}(z) = \left[\left(\sqrt{A_{\text{target}}} - \sqrt{A_o}\right) z/L + \sqrt{A_o}\right]^2$. When $d(z)$ or $g(z)$ are not constant, the functional dependence of σ_A on A is not known explicitly, and therefore, it is not possible to integrate Eq. (40) analytically. Numerical expressions are available for the norms and inner products, however, so one can proceed by numerically integrating $\dot{A}$, obtaining an expression for $z = z(A)$, and then inverting this expression to find the optimal biasing paths. As an example, Fig. 9 shows the results of numerical simulations in which the MC samples were biased along the optimal paths (shown by the thick curves) that produce three given amplitude changes (also indicated in the figure), demonstrating how the random trajectories are indeed closely clustered around these paths. Figure 10 shows the pdf of the output energy as reconstructed

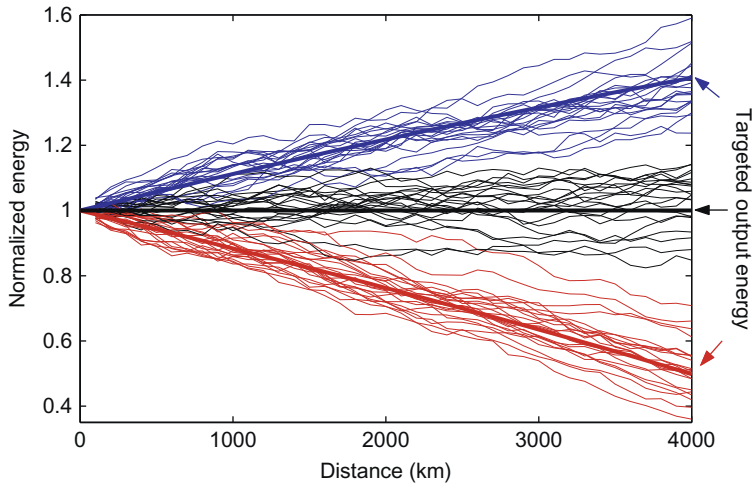

FIGURE 9 Samples from IS-MC simulations of the DMNLS equation. Here, the pulse energy (normalized to input energy) is plotted as a function of time (i.e., distance in physical units). The arrows represent the different targeted output energies: a larger than normal output energy (blue (dark gray in the print version)), a smaller than normal output energy (red (light gray in the print version)), and unbiased energy (black). Also plotted are deterministic paths (thick, smooth curves, with color corresponding to the target) predicted by our perturbation theory. These are the preferential paths around which we attempt to sample by biasing the noise with the adjoint linear modes. For each of three different targeted output energies, a few dozen IS-MC samples are also shown (also colored correspondingly), demonstrating that the actual trajectories indeed follow the predictions of the theory. *Source: From Li et al. (2007).*

from IS-MC simulations of the DMNLS equation using multiple IS and the biasing techniques described above. For comparison purposes, the results of unbiased MC simulation of the original NLS equation (27) with DM and a much larger number of MC samples are also shown, as well as a Gaussian fit to those results, demonstrating that pdf deviates significantly from a Gaussian, and at the same time that IS-MC simulation is an effective to quantify the probability of rare events in the system.

Similar techniques have been recently applied to quantify the effect of noise-induced perturbations in a variety of other system configurations, e.g., see Donovan and Kath, 2011; Li and Kath, 2015; Li et al., 2007; Moore et al., 2003, 2005, 2008; Spiller and Biondini, 2009, 2010; and references therein.

8 LARGE DEVIATIONS THEORY, ASYMPTOTIC EFFICIENCY, AND FINAL REMARKS

A key concept in assessing the effectiveness of a given biasing strategy and/or when using IS to reconstruct a sequence of quantities with decreasing probability (as in the case of the pdf in the example in Section 4.3) is that of asymptotic

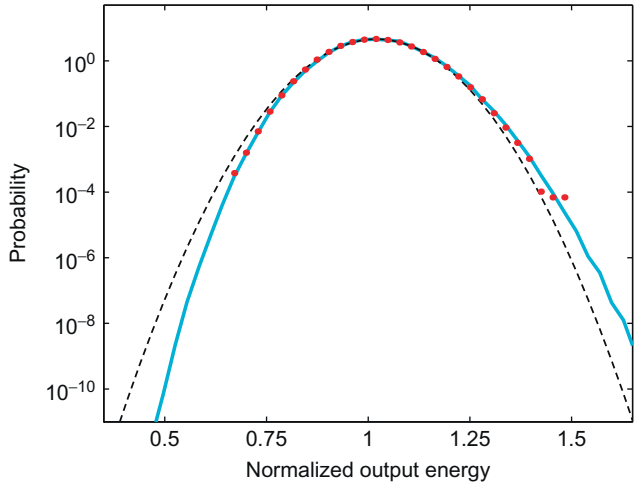

FIGURE 10 pdf of normalized output energy of a dispersion-managed soliton affected by amplifier noise. The solid (cyan (light gray in the print version)) curve shows results from IS-MC simulations of the DMNLS equation with 42,000 samples. The (red (dark gray in the print version)) dots are the results from standard MC simulations of the NLS equation with DM with 1,000,000 samples. The (black) dashed curve is a Gaussian fit to that simulation. Note how unbiased MC simulations of the NLS equation with DM deviate from Gaussian, but agree well with IS-MC simulations of the DMNLS equation as far as down in probability as the unbiased simulations can reach. *Source: From Li et al. (2007).*

efficiency (Glynn and Whitt, 1992; Sadowsky and Bucklew, 1990). The precise definition of asymptotic efficiency is formulated in the framework of large deviations theory (Bucklew, 1990; Dembo and Zeitouni, 1983). Here we will limit ourselves to giving an informal discussion of both of these topics.

Often, for simplicity, the choice of biasing distributions is restricted to a specific family of distributions, usually dependent on one or more parameters, e.g., in a specific situation these could be the mean translation parameters. Now consider a set of probabilities P_n dependent on a parameter n, e.g., P_n could be defined as the probability that the RV $y(\mathbf{X})$ takes values that are larger than n times its mean: $P_n = \mathbb{P}[y(\mathbf{X}) > n\mu]$, with $\mu = \mathbb{E}[y(\mathbf{X})]$. As another example, let $Y_n = (X_1 + \cdots + X_n)/n$ be the mean of n i.i.d. RVs $X_1, \ldots, X_n$. One could ask what is the probability that Y_n deviates more than ϵ from its mean, i.e., $P_n = \mathbb{P}[|Y_n - \mu| > \epsilon]$, where now $\mu = \mathbb{E}[X]$. Furthermore, suppose that the probabilities P_n tend to zero as $n \to \infty$, as is indeed the case in the two examples given. Large deviations theory is concerned with the *rate* at which these probabilities tend to zero. In this sense, it can be thought of as an extension of the law of large numbers.

It is often the case in practical situations that the probabilities P_n decay exponentially as n increases. Loosely speaking, when this happens we say that the sequence $\{P_n\}_{n\in\mathbb{N}}$ satisfies a *large deviations principle*. More explicitly, in

the above example we say that P_n satisfies a large deviations principle with rate function $I(\epsilon)$ if

$$\lim_{n\to\infty} \frac{1}{n} \log P_n = -I(\epsilon) .$$

More precise and comprehensive definitions can be given, which allow one to include a larger class of processes, for some of which the simple requirement above is not satisfied. A large body of work has been accumulated on large deviations theory. Two famous results, namely Cramér's theorem and the Gärtner–Ellis theorem, identify some properties of rate functions. In particular, for the sum of RVs considered above, one can show that the rate function is

$$I(\epsilon) = \sup_{s\in\mathbb{R}} [s\epsilon - \log(M(s))] ,$$

where $M(s) = \mathbb{E}[\exp(sX)]$ is the moment-generating function. For further details, we refer the reader to Bucklew (1990) and Dembo and Zeitouni (1983).

Now let us return to the problem of rare event simulation. It should be clear that the computational cost required for an accurate estimation of P_n with standard MC methods will obviously grow with n. Next, consider a sequence of biasing distributions $p_{*n}(\mathbf{x})$. Roughly speaking, the sequence is said to be *asymptotically efficient* if the computational burden grows less than exponentially fast.

The concept of asymptotic efficiency has important practical consequences. If a family of biasing distributions is asymptotically efficient, the increase in computational efficiency will be larger and larger the further we reach into smaller probabilities. The best-case scenario is that in which the computational cost to reach probability levels of 10^{-n} is independent of n. In that case, the increase in computational efficiency can be arbitrarily large in principle, and in practice is just dictated by how far down in probability we need to reach. We refer the reader to Bucklew (2004) for a discussion of precise conditions that guarantee that a sequence of simulation distributions is asymptotically efficient.

As a final remark, we should comment on the relation between large deviations theory and the study of random dynamical systems. In many cases, one can think of the input RVs as perturbations affecting the behavior of a dynamical system. For example, in the case of optical fiber communication systems, three kinds of randomness are present: (i) the fiber's random birefringence, which depends on distance, time, and wavelength; (ii) the optical amplifiers' quantum noise, which is added to the signal and propagates nonlinearly through the fiber; and (iii) the pseudo-random sequence of information bits. The problem of studying small random perturbations of dynamical systems was first posed in Pontryagin et al. (1933) and has received considerable attention in recent years. In many cases, the most likely configuration of RVs for which the system reaches a given output state can be thought of as a specific path in sample space. In turn, this path can be uniquely identified as the minimizer of the Wentzell–Freidlin action functional (Freidlin and Wentzell, 1984). IS can then

be thought of simply as a numerical (MC) technique to perform an integration in sample space around this "optimal" path. (Note the similarity between this point of view and the path integral formulation of quantum mechanics, e.g., see Weinberg, 1995.) The best-case scenario is of course that in which this optimal path can be identified analytically (e.g., as in Biondini et al., 2004; Moore et al., 2008). In other situations, however, one may be able to solve the minimization problem numerically (as in Spiller and Biondini, 2010). Finally, if this is also not practical, one can avoid the Wentzell–Freidlin formulation altogether and search for it adaptively using the cross-entropy method (as in Donovan and Kath, 2011; Marzec et al., 2013; Schuster et al., 2014).

REFERENCES

Ablowitz, M.J., Biondini, G., 1998. Multiple scale dynamics in communication systems with strong dispersion management. Opt. Lett. 23, 1668–1670.

Ablowitz, M.J., Segur, H., 1981. Solitons and the Inverse Scattering Transform. Society for Industrial and Applied Mathematics, Philadelphia.

Ablowitz, M.J., Biondini, G., Chakravarty, S., Horne, R.L., 1998. On timing jitter in wavelength-division multiplexed soliton systems. Opt. Commun. 150, 305.

Agrawal, G.P., 2002. Fiber optics communication systems. Wiley, New York.

Agrawal, G.P., 2007. Nonlinear Fiber Optics. Academic Press, New York.

Biondini, G., Kath, W.L., 2004. PMD emulation with Maxwellian length sections and importance sampling. IEEE Photon. Technol. Lett. 16, 789–791.

Biondini, G., Kath, W.L., 2005. Polarization-dependent chromatic dispersion and its impact on return-to-zero transmission formats. IEEE Photon. Technol. Lett. 17, 1866–1868.

Biondini, G., Kath, W.L., Menyuk, C.R., 2004. Importance sampling for polarization mode dispersion: techniques and applications. IEEE J. Lightwave Technol. 22, 1201–1215.

Bucklew, J.A., 1990. Large Deviation Techniques in Decision, Simulation and Estimation. Wiley, New York.

Bucklew, J.A., 2004. Introduction to Rare Event Simulation. Springer, New York.

Chan, C.C., Kroese, D.P., 2012. Improved cross-entropy method for estimation. Stat. Comput. 22, 1031–1040.

de Boer, P.-T., Kroese, D.P., Rubinstein, R.Y., 2004. A fast cross-entropy method for estimating buffer overflows in queueing networks. Manag. Sci. 50, 883–895.

de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134, 19–67.

Dembo, A., Zeitouni, O., 1983. Large Deviation Techniques and Applications. Jones & Bartlett, Boston.

Donovan, G.M., Kath, W.L., 2011. An iterative stochastic method for simulating large deviations and rare events. SIAM J. Appl. Math. 71, 903–924.

Fishman, G.S., 1996. Concepts, Algorithms and Applications. Springer-Verlag, Monte Carlo.

Fishman, G.S., 2006. A First Course in Monte Carlo. Thomson, Belmont.

Foschini, G.J., Poole, C.D., 1991. Statistical theory of polarization dispersion in single mode fibers. IEEE J. Lightwave Technol. 9, 1439.

Freidlin, M.I., Wentzell, A.D., 1984. Random Perturbations of Dynamical Systems. Springer-Verlag, New York.

Glynn, P.W., Whitt, W., 1992. The asymptotic efficiency of simulation estimators. Oper. Res. 40, 505.

Gordon, J.P., Haus, H.A., 1986. Random walk of coherently amplified solitons in optical fiber transmission. Opt. Lett. 11, 665–667.

Gordon, J.P., Kogelnik, H., 2000. PMD fundamentals: polarization-mode dispersion in optical fibers. Proc. Natl. Acad. Sci. U.S.A. 97, 4541–4550.

Gordon, J.P., Mollenauer, L.F., 1990. Phase noise in photonic communications systems using linear amplifiers. Opt. Lett. 15, 1351–1353.

Homem-de-Mello, T., Rubinstein, R.Y., 2002. Rare event probability estimation using cross-entropy. In: Yucesan, E., Chen, C.-H., Snowdon, J.L., Charnes, J.M. (Eds.), Proceedings of the 2002 Winter Simulation Conference. pp. 310–319.

Kaminov, I.P., Koch, T.L. (Eds.), 1997. Optical Fiber Telecommunications IIIA. Academic Press, New York.

Knuth, D.E., 2011. The Art of Computer Programming, vols. I–IV. Addison-Wesley, Boston.

Kogelnik, H., Nelson, L.E., Jopson, R.M., 2002. Polarization mode dispersion. In: Kaminow, I.P., Li, T. (Eds.), Optical Fiber Telecommunications IVB. Academic Press, pp. 725–861.

Kroese, D.P., Rubinstein, R.Y., 2004. The transform likelihood ratio method for rare event simulation with heavy tails. Queueing Syst. 46, 317–351.

Kroese, D.P., Taimre, T., Botev, Z.I., 2011. Handbook of Monte Carlo Methods. Wiley Series in Probability and Statistics. Wiley, New York.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22, 79–86.

Landau, D.P., Binder, K., 2000. A Guide to Monte Carlo Simulations in Statistical Physics. Cambridge University Press, Cambridge.

Li, J., Kath, W.L., 2015. Predicting and simulating rare, large deviations in nonlinear lightwave systems. preprint.

Li, J., Spiller, E.T., Biondini, G., 2007. Noise-induced perturbations of dispersion-managed solitons. Phys. Rev. A 75 (053818), 1–13.

Li, J., Biondini, G., Kath, W.L., Kogelnik, H., 2008. Anisotropic hinge model for polarization-mode dispersion in installed fibers. Opt. Lett. 33, 1924–1926.

Li, J., Biondini, G., Kath, W.L., Kogelnik, H., 2010. Outage statistics in a waveplate hinge model of polarization-mode dispersion. IEEE J. Lightwave Technol. 28, 1958.

Lima, A.O., Lima, I.T., Menyuk, C.R., 2005. Error estimation in multicanonical Monte Carlo simulations with applications to polarization-mode-dispersion emulators. IEEE J. Lightwave Technol. 23, 3781–3789.

Lu, T., Yevick, D., 2005. Efficient multicanonical algorithms. IEEE Photon. Technol. Lett. 17, 861–863.

MacKay, D.J.C., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge.

Marzec, Z., Schuster, J., Biondini, G., 2013. On the efficiency of importance sampling techniques for polarization-mode dispersion in optical fiber transmission systems. SIAM J. Appl. Math. 73, 155–174.

Mecozzi, A., 1998. Timing jitter in wavelength-division-multiplexed filtered soliton transmission. J. Opt. Soc. Am. B 15, 152.

Metropolis, N., 1987. The beginning of the Monte Carlo method. Los Alamos Sci. 15, 125–130 (special issue).

Metropolis, N., Ulam, S., 1949. The Monte Carlo method. J. Am. Stat. Assoc. 44, 335–341.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092.

Moore, R.O., Biondini, G., Kath, W.L., 2003. Importance sampling for noise-induced amplitude and timing jitter in soliton transmission systems. Opt. Lett. 28, 105–107.

Moore, R.O., Schafer, T., Jones, C.K.R.T., 2005. Soliton broadening under random dispersion fluctuations: importance sampling based on low-dimensional reductions. Opt. Commun. 256, 439–450.

Moore, R., Biondini, G., Kath, W.L., 2008. A method for the study of large noise-induced perturbations of nonlinear Schrödinger solitons using importance sampling. SIAM Rev. 50, 523–549.

Owen, A., Zhou, Y., 2000. Safe and effective importance sampling. J. Am. Stat. Assoc. 95, 135.

Papoulis, A., 1991. Probability, Random Variables and Stochastic Processes. McGraw Hill, New York.

Pontryagin, L.S., Andronov, A.A., Vitt, A.A., 1933. O statisticheskom rassmotrenii dinamicheskikh sistem. Zh. Eksp. Teor. Fiz. 3, 165–180.

Rubinstein, R.Y., Kroese, D.P., 2004. The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation and Machine Learning. Springer, New York.

Sadowsky, J.S., Bucklew, J.A., 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. IEEE Trans. Inf. Theory 36, 579.

Schuster, J., Marzec, Z., Kath, W.L., Biondini, G., 2014. A hybrid hinge model for polarization-mode dispersion in installed fiber links. J. Lightwave Technol. 32, 1412–1419.

Secondini, M., Forestieri, E., 2005. All-order PMD outage probability evaluation by Markov chain Monte Carlo simulations. IEEE Photon. Technol. Lett. 17, 1417–1419.

Sinkin, O.V., Grigoryan, V.S., Menyuk, C.R., 2007. Accurate probabilistic treatment of bit-pattern-dependent nonlinear distortions in BER calculations for WDM RZ systems. IEEE J. Lightwave Technol. 25, 2959.

Smith, P.J., Shafi, M., Gao, H., 1997. Quick simulation: a review of importance sampling techniques in communications systems. IEEE J. Select. Areas Commun. 15, 597.

Spiller, E.T., Biondini, G., 2009. Phase noise of dispersion-managed solitons. Phys. Rev. A 80 (011805), 1–4.

Spiller, E.T., Biondini, G., 2010. Importance sampling for the dispersion-managed nonlinear Schrödinger equation. SIAM J. Appl. Dyn. Syst. 9, 432–461.

Srinivasan, R., 2002. Importance Sampling: Applications in Communications and Detection. Springer, New York.

Thomas, A., Spiegelhalter, D.J., Gilks, W.R., 1992. Bugs: a program to perform Bayesian inference using Gibbs sampling. In: Bernardo, J., Berger, J., Dawid, A., Smith, A. (Eds.), Bayesian Statistics 4. Clarendon Press, Oxford, pp. 837–842.

Veach, E., 1997. Robust Monte Carlo methods for light transport simulation. Ph.D. thesis, Stanford University, California.

Weinberg, S., 1995. The Quantum Theory of Fields, vol. I. Cambridge University Press, Cambridge.

Yevick, D., 2002. Multicanonical communication system modeling—application to PMD statistics. IEEE Photon. Technol. Lett. 14, 1512–1514.

Zabusky, N.J., Kruskal, M.D., 1965. Interaction of solitons in a collisionless plasma and the recurrence of initial states. Phys. Rev. Lett. 15, 240–243.